The Analysis of Covariance and Alternatives

Statistical Methods for Experiments, Quasi-Experiments, and Single-Case Studies

Second Edition

BRADLEY E. HUITEMA

Department of Psychology
Western Michigan University
Kalamazoo, Michigan

A JOHN WILEY AND SONS, INC., PUBLICATION

Published by John Wiley & Sons, Inc., Hoboken, New Jersey.
Published simultaneously in Canada

Library of Congress Cataloging-in-Publication Data:

Huitema, Bradley E., 1938-
The analysis of covariance and alternatives / Bradley Huitema. – 2nd ed.
p. cm. – (Wiley series in probability and statistics)
Includes bibliographical references and index.
ISBN 978-0-471-74896-0 (cloth)
1. Analysis of covariance. I. Title.
QA279.H83 2011
519.5′38–dc22

2010054174

Printed in the United States of America

oBook ISBN: 9781118067475
ePDF ISBN: 9781118067451
ePub ISBN: 9781118067468

10 9 8 7 6 5 4 3 2 1

For
Betsy

Craig
Alyssa
Justin

Laura
Kathryn
Brad

Contents

Preface

The purpose of this book is to provide graduate students and research workers in the behavioral and medical sciences with an applied and comprehensive treatment of the analysis of covariance (ANCOVA) and related methods for experiments, quasi-experiments, and single-case designs. Even though covariance analysis was introduced to research workers many years ago (Fisher, 1932), it remains one of the least understood and most misused of all statistical methods. Even well-qualified researchers often apply ANCOVA in situations where such analysis leads to invalid conclusions and fail to employ it when it is appropriate. This was true when the first edition of this book was published 30 years ago and it is still true today.

Although some excellent textbook discussions of the essentials of the topic are now available, many researchers are not exposed to them. They frequently express confusion when attempting to apply ANCOVA. Some of this confusion can be traced to two sources. First, "statistics" courses in some academic departments have become little more than courses on how to run software. Understanding complex methods takes more than generating results. Guidance concerning what to do when assumptions underlying the analysis are not well approximated is routinely omitted from such courses. Similarly, designs other than randomized-group experiments are rarely discussed. Second, researchers often turn to the World Wide Web for guidance; unfortunately, the Internet is littered with misinformation. Misconceptions described in Chapter 29 appeared on the Web.

Many research methodologists now view ANCOVA as an old fashioned label; instead, they describe it as a minor variant of the general linear model that should be estimated using regression routines. The emphasis on the regression approach was a strong theme in the original edition and it continues here, but it does have a downside if insufficient thought is given to the design matrix used in specifying regression models for various versions of ANCOVA. Misinterpretations of regression coefficients are chronic with many variants of ANCOVA. It is a rare researcher who does not misinterpret the coefficient for the treatment dummy variable in a typically specified heterogeneous slopes model. Dedicated ANCOVA routines lead to fewer interpretation errors than do conventional regression routines applied by nonstatisticians.

Although the first part of the book includes an extensive review of applied statistics, it is hoped that all readers have an understanding of basic statistical inference through analysis of variance and regression. Most chapters require almost no knowledge of mathematics, but there are a few exceptions where matrix notation could not be avoided.

This edition is twice the length of the original and the number of chapters has doubled. The expansion is partly explained by a much more thorough review of prerequisite material, but mostly it is a reflection of new developments in the field and added topics such as the analysis of single-case designs. Many new methods introduced throughout the book have not previously appeared in print.

The book is broken down into seven major parts.

Parts I and II contain prerequisite material including analysis of variance for three designs, effect sizes, and multiple regression. Part III introduces simple and multiple ANCOVA using both the original Fisher approach and the general linear model approach. This section also includes assumptions and multiple comparison procedures. Part IV describes robust methods and methods of dealing with assumption departures such as heterogeneous slopes, nonlinear functions, dichotomous dependent variables, and covariates affected by treatments.

Part V introduces analyses for several simple and complex versions of single-case design. Part VI describes power analysis and extends ANCOVA to randomized-block designs, two-factor designs, pretest–posttest designs, and designs with multiple dependent variables. The last part, Part VII, describes measurement error correction and propensity score methods that are useful in the analysis and design of certain quasi-experiments, observational studies, and uncontrolled clinical trials.

Many people have contributed to the development of this edition. Joe McKean was of great help in the construction of tables of critical values for several distributions and for writing a routine that provides *p*-values for the Durbin–Watson statistic. Sean Laraway and Susan Snycerski provided valuable editorial comments on the single-case design chapters and produced many figures. Jessica Urschel produced many figures, delivered important feedback when I strayed from style, and provided detailed copy-editing. Stephanie Stilling provided extensive equation processing on a huge collection of problematic equations, unscrambled copious quantities of uncooperative files, and produced figures. Julie Slowiak scanned the entire first edition and converted files.

I am indebted to the executive director of the Institute of Mathematical Statistics for permission to reprint portions of the table of maximin contrast coefficients from: Abelson, R. P., and Tukey, J. W. (1963) Efficient utilization of non-numerical information in quantitative analysis: General theory and the case of simple order. *Annals of Mathematical Statistics*, 34, 1347-1369. I thank the editors of *Econometrica* for permission to reprint tables of critical values for the Durbin-Watson test from: Savin, N. E., and White, K. J. (1977) The Durbin-Watson test for serial correlation with extreme sample sizes or many regressors. *Econometrica*, 45, 1989-1996.

BRADLEY E. HUITEMA

Kalamazoo, Michigan
October 2010

PART I

Basic Experimental Design and Analysis

CHAPTER 1

Review of Basic Statistical Methods

1.1 INTRODUCTION

Statistical methods are often subsumed under the general headings "descriptive" and "inferential." The emphasis in behavioral and medical science statistics books is frequently on the inferential rather than the descriptive aspects of statistics. This emphasis sometimes occurs because descriptive statistics such as means, variances, standard deviations, correlation coefficients, regression coefficients, and odds ratios can be described and explained in relatively few chapters. The foundations of inferential statistics generally require longer explanations, a higher level of abstract thinking, and more complex formulas. Associated with the large amount of space devoted to inference is a failure on the part of many professionals to appreciate that description is usually the most informative aspect of a statistical analysis.

A perusal of many current journals reveals that the overemphasis on inference (especially tests of significance) and the underemphasis on simple description is widespread; the inferential tail is frequently allowed to wag the descriptive dog. Descriptive statistics are not only underemphasized, they are sometimes completely ignored. One frequently encounters research outcomes reported in terms of only probability values or statements of "significant" or "nonsignificant" with no mention of the size of the difference (e.g., a difference between sample means) associated with the inferential results. The sources of this perversity go beyond statistical training; editorial policies of journals, demands of funding agencies, and criteria established by governmental agencies are also involved. An exploration of the development of this problem is interesting but tangential to this review; it will not be pursued here.

The remaining sections of this chapter begin with a review of conventional hypothesis testing (including elementary statistical decision theory) and interval estimation procedures for the simple randomized two-group experiment. Issues associated with standardized effect sizes, measures of association, generalization of results, and the control of nuisance variation are also presented.

The Analysis of Covariance and Alternatives: Statistical Methods for Experiments, Quasi-Experiments, and Single-Case Studies, Second Edition. Bradley E. Huitema.

1.2 ELEMENTARY STATISTICAL INFERENCE

Research workers generally employ inferential statistics in dealing with the problem of generalizing results based on sample data to the populations from which the subjects were selected. Suppose we (1) randomly select N mentally retarded patients from a Michigan institution, (2) randomly assign these patients to treatments 1 and 2, (3) apply the treatments, and (4) obtain an outcome measure Y on each patient. A useful descriptive measure of the differential effectiveness of the two treatments is the difference between the two sample means. This difference is an unbiased point estimate of the difference between the corresponding population means.

If it turns out that the sample mean difference is large enough to be of clinical or practical importance, the investigator may want to make a statement about the difference between the unknown population means μ_1 and μ_2. Population mean μ_1 is the mean score that would have been obtained if treatment 1 had been administered to all mentally retarded patients in the whole institutional population. Population mean μ_2 is the mean score that would have been obtained if instead the second treatment had been administered to the whole institutional population. Inferential tests and confidence intervals are widely used to evaluate whether there are sufficient sample data to state that there is a difference between unknown population means and (in the case of the confidence interval) to provide an interval within which it can be argued that the population mean difference will be found. A summary of hypothesis testing and confidence interval methods for the two-group design is presented in Table 1.1. An understanding of the conceptual foundation for these methods requires that several crucial distinctions be made among different types of measures and distributions; these are reviewed next.

Recall that the purpose of statistical inference is to make statements about unknown population parameters on the basis of sample statistics. Hence, the purpose of inferential methods is clear only if the distinction between statistics and parameters is understood. A statistic is a measure of a characteristic of a sample distribution. A sample is a subset of a population and a population refers to the entire collection of individuals (or some other unit) about which one wishes to make a statement. When the population size is finite, it may be feasible to obtain data on an outcome measure Y for all members of the population. In this case, summary measures that describe characteristics of the entire population distribution can be computed. Recall that Greek symbols are generally used to denote measures of population distribution characteristics; these measures are known as population parameters. For example, the population mean (μ) and the population standard deviation (σ) describe the distribution characteristics known as central tendency and variability. Just as population parameters are used to describe characteristics of population distributions, sample statistics are used to describe characteristics of sample distributions. Statistics are generally denoted using Roman symbols. Examples of statistics include the sample mean $\bar{Y}$ and the sample standard deviation s. The distinction between statistics and parameters is crucial to understanding statistical inference.

A third type of distribution (in addition to sample and population distributions) provides the basis for both hypothesis tests and confidence intervals. It is known

Table 1.1 Summary of the Computation and Interpretation of the Independent Samples *t*-Test and 95% Confidence Interval for the Case of a Randomized Two-Group Experiment

A. Hypothesis Test

Null hypothesis: H_0: $\mu_1 = \mu_2$.
Test: Independent samples t (assume homogeneous population variances):

$$\frac{\bar{Y}_1 - \bar{Y}_2}{\sqrt{\frac{\sum_{i=1}^{n_1} y_i^2 + \sum_{i=1}^{n_2} y_i^2}{n_1+n_2-2}\left(\frac{1}{n_1}+\frac{1}{n_2}\right)}} = \frac{\bar{Y}_1 - \bar{Y}_2}{s_{\bar{Y}_1-\bar{Y}_2}} = t,$$

where

$\sum_{i=1}^{n_1} y_i^2$ and $\sum_{i=1}^{n_2} y_i^2$ are the within-group sum of squared deviation (centered) scores for groups 1 and 2, respectively;
n_1 and n_2 are sample sizes associated with groups 1 and 2; and
$s_{\bar{Y}_1-\bar{Y}_2}$ is the estimate of the standard error of the difference.

When the t-ratio is obtained from sample data it is called the obtained value of t (denoted as t_{obt}). After it is computed a decision rule is invoked in order to decide whether to reject the null hypothesis. Two forms of decision rules are described below; before either one can be used it is necessary to specify the level of α. Recall that the level of significance α is set before the experiment is carried out. If a directional (one-tailed) test is desired, the level of α may be denoted as α_1. If a nondirectional (two-tailed) test is involved, then the level of alpha may be denoted as α_2.

Decision rule for a nondirectional test using t. If the absolute value of t_{obt} is $\geq t_{cv}$ (where t_{cv} is the critical value), then reject H_0; otherwise retain. The critical value of t is often based on α_2 set at .05; the degrees of freedom $= N - 2$ (where $N = n_1 + n_2$).

Decision rule for a nondirectional test using the p-value. Because most current statistics computer programs provide p-values in addition to obtained t-values, the modern approach is to simply inspect the p-value and compare it with the level of α that has been specified. The decision rule for a nondirectional test is as follows: if the p-value is $\leq \alpha_2$, reject H_0; otherwise retain.

B. Confidence Interval

In the case of a simple two-group independent samples experiment the relevant 95% confidence interval is computed as follows:

$$(\bar{Y}_1 - \bar{Y}_2) \pm [s_{\bar{Y}_1-\bar{Y}_2}(t_{cv})] = (L, U),$$

where

$s_{\bar{Y}_1-\bar{Y}_2}$ is the estimate of the standard error of the difference; and
t_{cv} is the critical value of the t statistic based on $\alpha_2 = .05$ and $N - 2$ degrees of freedom, and L and U are the lower and upper limits of the 95% confidence interval.

as the sampling distribution (not to be confused with the sample distribution). A sampling distribution is a theoretical probability distribution of some statistic. We can conceptualize the sampling distribution of any type of statistic of interest. Examples that you have almost certainly encountered before are the sampling distribution of the mean and the sampling distribution of the difference between two independent sample means.

If we are interested in carrying out statistical inference regarding a mean, we need to know something about the sampling distribution of sample means. This distribution can be conceptualized as follows. Suppose (unrealistically) that we know the mean value (μ) for a specified population; let us say it is 30. We select a sample size, say, $n = 10$. Next, we randomly sample 10 observations from the specified population, compute the mean of these observations, and record this value. Now conceptualize repeating these steps an infinite number of times. The whole distribution of this infinite number of sample means is the sampling distribution of the mean, where each sample mean is based on $n = 10$. We know that the mean of this distribution of sample means will be equal to the mean of the raw scores in the population. The standard deviation of the sampling distribution of the mean is called the standard error of the mean; it is given this name because it can be conceptualized as the standard deviation of the sampling error associated with the sample means included in the sampling distribution. The notation used to describe the standard error of the mean is $\sigma_{\bar{Y}}$. You may recall from elementary statistics that estimating the standard error of the mean is a simple task and that such an estimate (denoted as $s_{\bar{Y}}$) is required when statistical inference is applied to make statements about the unknown population mean in the one-group case.

Similarly, when inferential procedures are needed in the comparison of two sample means (e.g., in the case of a two-group independent sample experiment) it is necessary to have information regarding the sampling error associated with the difference between the two means. In this case the relevant sampling distribution is known as "the sampling distribution of the difference" (which is the shortened term for "the sampling distribution of the difference between two independent sample means"). This type of distribution can be conceptualized as follows.

Imagine two populations, where each one has a mean of 40 (i.e., $\mu_1 = \mu_2 = 40$). Select a sample size to be drawn from each population, say, $n_1 = 15$ and $n_2 = 15$. After 15 observations are randomly selected from the first population and 15 observations are randomly selected from the second population, the sample means $\bar{Y}_1$ and $\bar{Y}_2$ are computed. Next, the second sample mean is subtracted from the first and the difference is recorded. This process is then repeated (hypothetically) an infinite number of times. The distribution of the mean differences that result from this process constitutes the sampling distribution of the difference. The standard deviation of the sampling distribution of the difference is called the standard error of the difference (denoted $\sigma_{\bar{Y}_1 - \bar{Y}_2}$); an estimate of this parameter is required to carry out hypothesis tests (and confidence intervals) on the difference between two sample means. The essential conceptual issues associated with hypothesis testing are subsumed under the topic of statistical decision theory; this topic is briefly reviewed in the next section in the context of testing the equality of two population means (μ_1 and μ_2) using the independent samples *t*-test.

1.3 ELEMENTARY STATISTICAL DECISION THEORY

The concepts of type I error, type II error, and power are central to elementary statistical decision theory. Recall that the decision to reject the null hypothesis regarding the population means is made when (a) the obtained test statistic equals or exceeds the critical (i.e., tabled) value of the statistic or (b) the obtained p-value is equal to or less than the specified level of alpha. If the obtained test statistic is less than the critical value (or if the p-value exceeds alpha), the null hypothesis is retained. This decision strategy will not always lead to the correct decision.

Type I error. If the null hypothesis (i.e., H_0: $\mu_1 = \mu_2$) is true in the case of a two-group experiment, the two population means are equal but we can anticipate that each sample mean will deviate somewhat from the corresponding population mean and that the difference between the sample means will deviate somewhat from zero. Although we expect the difference of an infinite number of sample mean differences (where each mean is based on a random sample) to equal the population mean difference (i.e., $\mu_1 - \mu_2$), we anticipate that, in general, differences between sample means will differ somewhat from the difference between the corresponding population means as a result of sampling fluctuation.

This suggests that differences between two sample means will sometimes be large even though the two populations from which the samples were selected have exactly the same mean. We often employ significance tests to help decide whether an obtained difference between two sample means reasonably can be explained by sampling fluctuation (sometimes called sampling error). If the difference cannot reasonably be attributed to sampling error, we conclude that the population means are different. But if we conclude that the population means are different (i.e., we reject the null hypothesis) when the null hypothesis is true, a type I error is committed. That is, because of sampling error the difference between sample means will sometimes be large enough to result in the decision to reject the null hypothesis even though there is no difference between the corresponding (but unknown) population means. When this type of event occurs (i.e., rejecting the null hypothesis when it is true), we label it as a type I error. A priori (i.e., before the study is carried out) the probability of this happening is known as alpha (α). The researcher can control α by simply selecting the desired α level in either (1) a table of critical values of the test statistic or in (2) appropriate statistical software. For example, if it is decided before the experiment is carried out that it is acceptable for the probability of type I error to be no higher than 5%, the test will be carried out using the critical value associated with $\alpha = .05$. Equivalently, the null hypothesis will be rejected if the observed p-value provided by statistical software is .05 or less.

Type II error. If the difference between sample means is not large enough to yield an obtained t that exceeds the critical value, the null hypothesis is retained. This is a decision error of the second kind if the null hypothesis is actually false. That is, a type II error is committed when the experimenter fails to reject a false null hypothesis. The probability of making a type II error is known as beta (β).

Power. The power of a statistical test refers to the probability that the test will reject the null hypothesis when the null hypothesis is false. Power is equal to $1 - \beta$ (i.e., one minus the probability of making a type II error). Unlike α, which is simply set by the

experimenter before the data are collected, the power of a designed experiment must be computed. In the context of an independent samples two-group t-test, power is a function of (1) the size of the population effect (e.g., the population mean difference), (2) the size of the random error variance (i.e., the within-population variance), (3) the sample size n, and (4) the level specified for α. Power goes up with increases in population effect size, sample size, and α; it goes down as the error variance increases. Charts, tables, and software for estimating power are widely available.

Relationships among the concepts of type I error, type II error, and power are relevant to the appropriate interpretation of hypothesis tests. The issue of power is especially important in understanding how to interpret a result that is statistically nonsignificant. Because there is a widespread misunderstanding of the terms "statistically significant" and "nonsignificant," the next section considers some interpretation issues associated with each of these outcomes.

Interpretation of the Results of Hypothesis Tests and Confidence Intervals

Although the computation of hypothesis tests and confidence intervals of the type shown in Table 1.1 is a trivial issue with the availability of modern statistical software, the interpretation of the outcome of these tests and confidence intervals requires a little thought.

Statistically significant result. When the null hypothesis is rejected it is appropriate to state that there is sufficient information to conclude that the population means are not equal; this is equivalent to stating that the difference between sample means is statistically significant. The finding of statistical significance simply means that there is a strong support for the conclusion that sampling error is not the only explanation for the observed sample difference. Often a difference that is too large to be explained only as sampling error (and is therefore statistically significant) will still be too small to be of any substantive interest.

Recall that a result is declared statistically significant when the p-value is small; the p-value is simply a statement of conditional probability. It refers to the probability of obtaining a sample mean difference at least as large as the one obtained, under the condition that there is no difference between the population means. There is no aspect of this probability statement that says the sample difference is large or important. Hence, there is no justification for claiming that a difference sufficiently large to yield a small p-value (say .05) is necessarily important (unless one is working in an area in which any deviation whatsoever from zero is important).

One determines whether a difference is of practical, clinical, or theoretical importance through knowledge of substantive considerations beyond the inferential test. Credible statements regarding the importance of statistically significant results must rest on knowledge of many nonstatistical aspects of the experiment; these aspects include the type of subjects, the nature of the independent variable, the properties of the dependent variable, previous results, and various contextual variables surrounding the investigation. In short, it is usually necessary to be immersed in the content area of the study in order to understand whether a result is important; a statistical test does not provide this context.

Statistically nonsignificant result. A nonsignificant result is not proof that there is no difference between the population means. Rather, a nonsignificant result should be interpreted to mean that there is insufficient information to reject the null hypothesis. There is a big difference between having proof that the null hypothesis is true and having insufficient evidence to reject it.

There are two major interpretations to keep in mind when a nonsignificant outcome is obtained. First, the nonsignificant result may be obtained because there is very little or no true effect whatsoever. Second, there may be an important true effect but the analysis may have such low power that the effect is not detected. The second interpretation should not be ignored, especially if no information is available regarding the power of the analysis. On the other hand, if (a) the difference between the sample means is trivial and (b) the power of the analysis for detecting small effects is known to be very high, then the second interpretation is less credible.

Interpretation of confidence intervals. Confidence intervals are frequently recommended in place of hypothesis tests as the basic method of statistical inference. Unfortunately, an inspection of many current journals will reveal that this recommendation is not often followed. Perhaps the most important advantage is that a confidence interval is reported in the metric of the dependent variable rather than in a derived measure such as t, F, or a p-value. Because hypothesis tests are so frequently misinterpreted (often because they are not accompanied by appropriate descriptive statistics), the arguments in favor of confidence intervals have substantial weight. A major reason for the misinterpretation of significance tests is a failure to distinguish between point estimation (e.g., the size of the difference between means) and the size of p-values. This confusion is a continuing issue in published research.

I recently reviewed a research paper that was carried out and reported by a well-qualified professional associated with a major school of medicine. Three treatment methods were compared and an inferential statistical analysis based on hypothesis testing was reported. The results section focused exclusively on the outcome of hypothesis tests on three comparisons. There was not a single graphic display or descriptive statistical measure to be found in either the text or the table that accompanied the results section. The only statistics reported were p-values. It was impossible to discover how large the effects were, although it was stated that the effects were large because the p-values were small. Obviously, the author thought that the presentation of small p-values was tantamount to a demonstration of large effects. Unfortunately, this confusion of small p-values with large or important treatment effects is very common. Some journals continue to publish articles that present inferential results unaccompanied by any descriptive measures whatsoever. One should never present inferential test results without the associated descriptive statistics.

If the inferential results in the medical study mentioned above had consisted of confidence intervals rather than hypothesis tests, the failure to provide any descriptive information would have been impossible. This is true because the confidence interval is constructed around the observed sample mean difference. An inspection of the upper and lower limits of the interval provides both descriptive and inferential information because (a) the point at the center of the interval corresponds to the mean

difference, and (b) the width of the interval provides inferential information regarding how much sampling error is present.

When a 95% confidence interval is presented, it is often stated that the unknown population mean difference is contained in the interval and that 95% confidence is attached to this statement. This interpretation needs a little clarification. The interpretation that we are 95% confident that the population mean difference is contained in the computed interval requires that we conceptualize a large number of experiments, not just the one experiment we are analyzing. If the experiments were replicated an infinite number of times, there would be an infinite number of confidence intervals and we would find that 95% of them would contain the population mean difference. Hence, the confidence coefficient (.95 in this example) refers to how confident we are that the one interval computed in this single experiment is one of those that contains the population difference. The values contained in the interval are generally viewed as credible ones to entertain as the true population mean difference; values outside the interval are not.

Hence, a confidence interval provides much more information than does t or p because simple inspection of the interval provides both descriptive and inferential information not provided by a t-value or p-value alone. Although the mean difference and the associated confidence interval supply much useful information, several supplementary approaches have recently become popular. An overview of two of these methods and some of the reasons they are currently being heavily promoted are presented next.

1.4 EFFECT SIZE

The major reason for performing a statistical analysis is to describe and effectively communicate the research outcome. If the description and communication are to be effective, the methods of analysis should be as transparent as possible. Although transparency should be the data analyst's credo, this is not the impression conveyed in many published articles; too often obfuscation appears to be the rule.

Suppose a two-group experiment has been carried out, the data have been plotted, and the means and variances have been reported along with the mean difference on the original metric. Before any hypothesis tests or confidence intervals are considered, it is useful to first ask whether the obtained mean difference is of any practical or theoretical importance. It is often possible for a researcher to state values of differences between means (or some other appropriate descriptive statistic) that fall into one of the following categories: (1) trivial or unimportant, (2) of questionable importance, or (3) of definite practical or theoretical importance. The consideration of what constitutes an important result should affect the emphasis one attaches to inferential results.

If the obtained difference is judged to be unimportant, there is little justification for emphasizing the results of significance tests. If the obtained difference is not trivial, significance tests provide information that may be of interest. This does not mean that significance tests are invalid when applied to trivial differences. Rather, in most

cases it is unnecessary to ask whether a trivial difference is statistically significant (as is often the case with large sample sizes). There is a limited interest in learning that a trivial difference is too large to be explained as sampling error.

If a nontrivial difference is obtained, both the difference between means and the associated *p*-value are likely to be of interest. The failure to distinguish between point estimates of treatment effects and *p*-values associated with point estimates of treatment effects has led to decades of confusion regarding tests of significance. Because researchers so often imply that *p*-values provide information on the size and/or importance of the difference between means there have been recent attempts on the part of research methodologists and some journal editors to encourage researchers to provide more adequate descriptions of results. Consequently, there are now several journals in the behavioral and medical sciences that require the reporting of so-called effect sizes or measures of association.

These editorial policies can be traced back to the *Publication Manual of the American Psychological Association* (*APA*). It states, ". . . it is almost always necessary to include some index of effect size or strength of relationship . . ." (2001, p. 25). Because researchers in many areas have become disenchanted with tests of significance the *APA* recommendation is likely to spread to other areas as well. This movement to focus on the so-called "effect size" is a step forward only if it is understood that such measures should be used to supplement rather than supplant conventional descriptive statistics.

Although the term "effect size" is now well established in the behavioral sciences, it is actually a misnomer. It refers to neither the most natural measure of the size of the effect (viz., the difference between two means) nor to the "effect" parameter, as it is defined in the analysis of variance structural model (i.e., the difference between the population mean for a specific treatment and the combined average of all population treatment means in the experiment). Consequently, the term is often confusing to both statisticians and researchers outside the behavioral sciences. This confusion between the term "effect size" and the actual size of the effect (i.e., the mean difference in the original metric) would be eliminated if precise language were employed. That which is currently called the effect size should be called the "standardized effect size" because it is defined as the following parameter:

$$\frac{\mu_1 - \mu_2}{\sigma_w} = \delta.$$

It can be seen that the difference between the two population means (the effect) is standardized by the common within-population standard deviation σ_w. There are several ways to estimate this standardized effect size parameter. Cohen's d is defined as

$$\frac{\bar{Y}_1 - \bar{Y}_2}{\sqrt{\dfrac{\sum_{i=1}^{n_1} y_i^2 + \sum_{i=1}^{n_2} y_i^2}{n_1+n_2}}} = d.$$

A convenient formula that I prefer is known as Hedges' g; it is defined as

$$\frac{\bar{Y}_1 - \bar{Y}_2}{\sqrt{\dfrac{\sum_{i=1}^{n_1} y_i^2 + \sum_{i=1}^{n_2} y_i^2}{n_1+n_2-2}}} = \frac{\bar{Y}_1 - \bar{Y}_2}{s_w} = g.$$

A frequently recommended modification of this formula that provides slightly less biased estimates of the population effect size is

$$\frac{g(N-3)}{N-2.25}.$$

Although the latter formula provides less biased estimates than are provided by g, the difference is very small in the case of reasonable sample size and it is certainly less intuitive. So I would not bother with the less biased estimator. Regardless of the formula used, the result is interpreted as the difference between means in standard deviation units.

In many cases the standardized effect size can be a useful supplement to the analysis, but in some contexts these measures cloud rather than clarify the outcome of an experiment. Clarity may be lost if the outcome measure is easily conceptualized and is well understood by both the researcher and the audience for the study.

Suppose a researcher is interested in the differential effect of two drugs on litter size in rats and that she has carried out a very large randomized-groups experiment (assume very little sampling error); the number of rats in each litter is used as the dependent variable. The mean number produced under drug I is 4.8 and the mean number under drug II is 8.8; the pooled within-group standard deviation is 2.5. Because the dependent variable is easy to conceptualize, the four-point difference between the means needs little additional explanation. The four-point difference would be more meaningful with the knowledge of typical litter sizes in untreated rats, but even without this additional information the outcome is understandable. It is obvious that litter size is about twice as large under drug II as it is under drug I. On the other hand, if the results are presented in terms of the standardized effect size (i.e., $g = 4/2.5 = 1.6$) instead of the mean difference, the statistical description is likely to be far less transparent for many readers in the audience. The original metric (number of animals in a litter) is easily understood whereas the standardized metric is not. If the research outcome is to be described in a publication intended for the general public, it is a mistake to present only standardized effect sizes. This is not to say that the standardized effect size should never be reported in professional journals. But in most cases results are more effectively communicated in the original metric than in standard deviation units. The typical researcher finds it easier to think (and communicate) in terms of the difference between means on the original metric than in terms of the difference between means in standard deviation units. But there is an exception to the general recommendation to focus on the original metric.

Although the mean difference on the original metric often communicates results more clearly than does the standardized effect size, this is not always so. There are many cases in which the outcome measure is new, unfamiliar, or difficult to conceptualize. In these situations it may be appropriate to focus the description of results on the standardized effect size rather than the difference between group means on the original metric. Hence, mean differences are usually more appropriate for studies that use well-understood dependent variables whereas standardized effect sizes may be preferable for reporting results from studies that use poorly understood outcome measures. The degree to which an outcome measure is understood depends upon both the audience for the study and the measurement conventions of the content area. Both of these should be considered when choosing the metric upon which to focus.

If the original metric seems to be a poor choice because it is poorly understood but the standardized effect size is also poorly understood by the intended audience, there is another choice. I recommend that the standardized effect size be computed and then transformed to a more understandable form. It is possible to make the information contained in the standardized effect size understandable to any audience through the use of appropriate graphics and/or simpler statistics (described below). This is especially true if one of the groups in a two-group experiment is a control. Consider the following example.

Suppose that a randomized-groups experiment is carried out to compare (1) a new method of treating depression with (2) a control condition; an established measure of depression is the dependent variable. If $g = -.75$, the conventional interpretation is that the mean depression score for the treatment group is three-quarters of a within-group standard deviation below the control group mean. This interpretation will not be meaningful to many who read such a statement. But it is possible to transform the information contained in g to a form that is much more understandable by the general reader. It can be determined (under the assumptions of normality and homogeneity of population variances) that 50% of the treated patients have depression scores below that of the average treated patient whereas only 23% of the control patients have depression scores that low. Hence, the treatment increases the percentage from 23% to 50%. The presentation of the results in terms of these percentages is far more meaningful for most readers than is the statement that $g = -.75$. These percentages can be computed by any researcher who is familiar with the process of computing the area of the unit normal distribution below a specified standard score. In this example, we simply determine (using tables or software) the area in the unit normal distribution that falls below a z-score of $-.75$. Alternatively, we could look up the area above a z-score of .75. In this case, we can report that 50% of the control group has depression scores at or above the control group mean whereas only 23% of the treated group has depression scores this high. A graph illustrating the two hypothetical population distributions (treatment and control) and the associated percentages below the treatment-group mean further facilitates an understanding of the outcome. When the distributions are appropriately graphed the advantage of the treatment group over the control group is immediately apparent.

Attempts have been made to provide meaning to standardized effect sizes by labeling them as small, medium, or large (Cohen, 1988). The conventions for absolute

obtained standardized effect sizes are as follows: small = .20, medium = .50, and large = .80. Hence, the example g presented in the previous paragraph would be classified as a medium standardized effect size. As is the case with most conventions of this type, they have taken on a life of their own; much is made of having a result that is "large" whereas "small" results may be denigrated. It should be kept in mind that size and importance is not necessarily the same thing. Small effects can be very important and large effects in some studies may be unimportant. Importance is dictated by substantive issues rather than by the value of g. Unfortunately, the current emphasis on standardized effect sizes has led some researchers to omit the original metric means in published articles. Both the standardized and unstandardized results should be reported even though a choice is made to focus on one or the other in the text describing the experimental outcome.

Perhaps the main reason for reporting standardized effect sizes is to provide an outcome measure for other researchers who are carrying out meta-analytic research. In these situations there is interest in characterizing the effects of certain treatments found in a large collection of studies, where each study provides comparative evidence on the same treatments. If all of the studies that use the same method of measuring a single-outcome construct, the task of summarizing the outcome of the studies is simple. But different studies often use different ways of operationalizing a single outcome construct. In this case a method of converting all the outcome measures to the same metric is desirable; the standardized effect size provides such a metric.

1.5 MEASURES OF ASSOCIATION

A second type of derived descriptive measure of outcome is sometimes called a measure of association; often these measures are described as a type of effect size. Like standardized effect sizes, measures of association are useful when the dependent variable is not a familiar measure. Unlike standardized effect sizes, measures of association are not interpreted as the size of the mean difference on some alternative metric. Rather, measures of association describe the proportion of the variation on the dependent variable that is explained by the independent variable. In a sense these measures describe the extent to which variation in the observed sample behavior is under experimental control.

Although there are several different measures of association, the intent of all of them is essentially the same. The one called the correlation ratio will be reviewed here. The population correlation ratio is denoted as η^2; the sample estimate of this parameter is denoted as $\hat{\eta}^2$. The computation of $\hat{\eta}^2$ is easily accomplished using the output of conventional statistical software that provides t statistics (for the two-group case) or the analysis of variance summary table (for two or more groups). In the case of a two-group design, the sample correlation ratio can be defined as

$$\frac{t_{\text{obt}}^2}{t_{\text{obt}}^2 + (N-2)} = \hat{\eta}^2,$$

where

t_{obt} is the obtained value of the conventional two-group independent samples t-ratio; and
N is the total number of observations in the experiment (i.e., $n_1 + n_2$).

A more general expression that applies to independent sample designs with two or more groups is

$$\frac{SS_B}{SS_T} = \hat{\eta}^2,$$

where

SS_B is the between groups sum of squares (from a one-factor analysis of variance); and
SS_T is the total sum of squares.

When two groups are involved, the t approach and the sum of squares approach for computing the correlation ratio are equivalent. If more than two groups are involved, the t approach is irrelevant and only the sum of squares approach is appropriate. Regardless of the number of groups in the experiment, the interpretation of the correlation ratio is the same. That is, it describes the proportion of the total variation in the experiment that appears to be explained by the independent variable.

It was mentioned earlier that there is a correspondence between standardized effect sizes and measures of association. In the case of two groups, it turns out that

$$\hat{\eta} = \frac{g}{\sqrt{g^2 + 4\left(\frac{N-2}{N}\right)}}.$$

An advantage of reporting correlation ratios in addition to mean differences, significance tests, and/or confidence intervals is that they describe the effectiveness of the treatments in relative rather than absolute units. If the correlation ratio is 0.40, it can be stated that approximately 40% of the total variability in the experiment appears to be under experimental control and 60% of the total variability is due to the sources other than treatment group differences. If *all* variability in the experiment is due to treatments, the correlation ratio is 1.0; in this unrealistic case, the experimenter has complete control of the outcome through manipulating the levels of the independent variable. If there is no difference between the sample means, the correlation ratio is zero; in this case no variability on the dependent variable is explained by the independent variable.

The proportion of variability explained on the dependent variable by the independent variable may be a more meaningful way to describe the outcome of the experiment than is the mean difference on a variable that is completely unknown to the reader of a research report. Like the standardized effect size, the correlation

ratio can be useful in comparing results of different studies, especially if the different studies have used different methods of operationalizing a common outcome construct. Also, as is the case for the standardized effect size, there are conventions regarding labels that are attached to correlation ratios of various sizes. The conventions some methodologists use to label the size of computed values of $\hat{\eta}^2$ are: small = .01, medium = .09, and large = .25; others recommend using .01, .06, and .15, respectively.

Caveats regarding conventions of this type were mentioned previously for the standardized effect size; they also apply to the correlation ratio. Recall that the size of an effect and the importance of an effect are not the same thing. A study demonstrating the effects of aspirin on heart attack in men (Steering Committee of the Physicians' Health Study Research Group, 1988) is frequently cited as an example of data illustrating how elusive the notion of importance can be.

Over 22,000 male physicians were randomly assigned to receive either aspirin or a placebo every other day. After 5 years of treatment, the proportion of heart attacks experienced by the aspirin group was .0094; the placebo group proportion was .0171. The estimate of η^2 was .0011 and the p-value was $< .000001$. Hence, there was no doubt that aspirin therapy decreased the proportion of heart attacks. There was also no doubt that the proportion of variation on the outcome measure explained by the independent variable was very small. Note that the estimate of η^2 is far below the cutoff for declaring that there is even a "small" effect. In spite of these results regarding the small size of the effect expressed as a proportion, the Committee decided that the experiment should be terminated early because the evidence was so convincing. Aspirin therapy to prevent heart attack is now widely recommended for a large proportion of the male population.

But the advantage of the aspirin therapy may not be as impressive as has often been claimed in the popular press. Many people do not consider changing one's probability of heart attack from .0171 to .0094 a massive effect. That is, the change in the probability of heart attack attributed to aspirin therapy was only about three-quarters of a percent. This is the reduction in what is known as absolute risk; it is the reduction that is relevant to the entire population of male physicians. But it is unlikely that this is the measure the Committee had in mind when the statement was made that the effect was large. Instead, it appears that the reason the study was terminated was more influenced by the p-value and the relative reduction in risk. The latter is determined by computing the difference between control and treatment groups on the proportion having a heart attack ($.0171 - .0094$) by the proportion in the control group having one (.0171); this yields a very impressive looking .45. The implication of this is that those in the control condition who will have a heart attack will be 45% less likely to have one if they take aspirin. The main point of this example is that there are different ways of expressing the size of an effect and these different measures are often very inconsistent. A single study can be described as having a trivial effect or a massive effect, depending on the measure chosen.

A computational example of the major statistics described in this chapter is summarized in Table 1.2. Statistics similar to the correlation ratio and standardized effect

size measures reviewed in this section are described in subsequent chapters for more complex designs and for analysis of covariance problems.

1.6 A PRACTICAL ALTERNATIVE TO EFFECT SIZES AND MEASURES OF ASSOCIATION THAT IS RELEVANT TO THE INDIVIDUAL: $p(Y_{TX} > Y_{CONTROL})$

Many clinicians and patients find it less than satisfactory that research results are virtually always reported in terms of group measures and associated p-values. Indeed, one often hears the complaint that group comparisons are essentially irrelevant to an individual patient. This discontent generalizes to effect sizes and measures of association. When attempting to convey the meaning of a two-group trial to patients and others outside the research community, I recommend abandoning conventional group comparison statistics. Instead, provide the probability that a subject randomly selected from the treatment group has a higher outcome score (or a lower outcome score when the treatment is intended to lower the score) than does a subject randomly selected from the control group. This is not a conventionally computed value; I am aware of no standard software that provides it, but the computation is straightforward.

Consider the outcome of the two-group experiment on pain reduction described in Table 1.2. It can be seen that the mean pain for the treatment group is lower than the mean for the control group and that the standardized effect size is -1.15. The approximate probability that a subject randomly selected from the treatment group has a higher outcome score than does a subject randomly selected from the control group is obtained by first computing:

$$\frac{g}{\sqrt{2}} = \frac{-1.15}{\sqrt{2}} = -.8131728 = z.$$

The area below a z-score of $-.8131728$ can be found in a table of the normal distribution to be .21; this value approximates the probability that a subject selected at random from the treatment group has a pain score that is higher than the score of a subject selected at random from the control group. If, instead, we are interested in the probability that a subject selected at random from the treatment group has a pain score that is *lower* than the score of a subject selected at random from the control group, we compute the area above the computed z; in this example that area is equal to .79. When the study is one in which the treatment is expected to increase the score on the dependent variable, it will usually be of greater interest to find the area below the z-score.

Note that the probability value provided using this approach is likely to answer the question a typical patient wants to be answered. It is natural to seek information of this type when contemplating a treatment procedure. Those who think in terms of odds rather than probability values can easily convert from one to the other. Browne (2010a, 2010b) does a nice job of illustrating both.

Table 1.2 Example of Point Estimation, Hypothesis Testing, Interval Estimation, Standardized Effect Size, Correlation Ratio, and $p(Y_{\text{Tx}} < Y_{\text{Control}})$ for a Randomized Two-Group Experiment

Example Data (A Subset of Pain Score Data Analyzed in Wang et al., 2010)

Treatment	Control
6	5
7	10
3	7
7	10
7	6
4	7
4	6
6	10

A. Point Estimate of Treatment Effect

The point estimate of the population difference between treatment and control means on the pain scale is the sample mean difference: $(5.500 - 7.625) = -2.125$.

B. Hypothesis Test (Independent Samples t)

Null hypothesis: $H_0 : \mu_1 = \mu_2$.

$$\text{Computed value of } t: \frac{\bar{Y}_1 - \bar{Y}_2}{s_{\bar{Y}_1 - \bar{Y}_2}} = \frac{5.500 - 7.625}{.92461} = -2.298 = t_{\text{obt}}.$$

Critical value of t (for $\alpha_2 = .05$) based on 14 degrees of freedom $= 2.145$.

Decision. Reject H_0 because $|-2.298| \geq 2.145$ and state that the sample mean difference is statistically significant; infer that $\mu_1 \neq \mu_2$. Equivalently, reject H_0 because the p-value (from *Minitab*) $= .037$, which is less than the value (.05) set for α_2.

Conditional probability interpretation. The p-value .037 is the probability of obtaining an absolute difference between the sample means of at least 2.125 points under the condition that there is no difference between the treatment and control population means (i.e., $\mu_1 = \mu_2$). Because this is an unlikely event under the stated condition, it is reasonable to argue that a better explanation for the sample mean difference is that the population means are different.

Substantive conclusion. There is sufficient evidence to conclude that the mean pain in the treated population is lower than the mean pain in the control population.

C. 95% Confidence Interval for the Difference $(\mu_1 - \mu_2)$

Computation.

$$(\bar{Y}_1 - \bar{Y}_2) \pm [s_{\bar{Y}_1 - \bar{Y}_2}(t_{cv})] = (L, U) = (\bar{Y}_1 - \bar{Y}_2) \pm [s_{\bar{Y}_1 - \bar{Y}_2}(t_{cv})] = (L, U)$$
$$= (-2.125) \pm [.9246(2.145)] = (-4.11, -.142).$$

Interpretation. The population mean difference $(\mu_1 - \mu_2)$ between the treated and control groups on the pain scale is contained in the interval $(-4.11, -.142)$; the confidence coefficient associated with this statement is .95.

Table 1.2 (*Continued*)

D. Standardized Effect Size g

Computation.

$$\frac{\bar{Y}_1 - \bar{Y}_2}{s_w} = \frac{5.500 - 7.625}{1.849} = -1.15 = g.$$

Interpretation. The treatment group mean pain score is -1.15 standard deviation units below the mean pain score of the control group. In addition, because the area in the unit normal distribution below a standard score of -1.15 is .1251, it is estimated that approximately 13% of the control population has pain scores falling below the average pain score in the treatment population whereas 50% of those in the treatment population have scores that low.

E. Correlation Ratio

Computation.

$$\frac{t^2_{\text{obt}}}{t^2_{\text{obt}} + (N-2)} = \frac{-2.298^2}{-2.298^2 + (N-2)} = \frac{5.28}{5.28 + 14} = .27 = \hat{\eta}^2.$$

Interpretation. Twenty-seven percent of the total variation among subjects on the pain score appears to be explained by the independent variable.

F. Probability That a Randomly Selected Treated Subject Has Lower Pain Than Does a Randomly Selected Control Subject

$\frac{g}{\sqrt{2}} = \frac{-1.15}{\sqrt{2}} = -.8131728 = z$; area below z is .21. This is $p(Y_{\text{Tx}} > Y_{\text{Control}})$, but, because the Tx is designed to reduce pain, the interest is in $p(Y_{\text{Tx}} < Y_{\text{Control}})$, which is $(1 - .21) = .79$.

1.7 GENERALIZATION OF RESULTS

A great deal of methodological literature is devoted to the topic of generalizing experimental results (see, e.g., Shadish et al., 2002). Issues such as the methods used to operationalize the independent and dependent variables, the nature of the experimental setting, and the method of selecting subjects are relevant to this topic. Subject selection is a major statistical concern; it plays a major role in defining the population to which the results can be generalized. The selection procedures associated with two versions of the one-factor randomized-groups experimental design are described below.

Case I: Ideal Randomized-Groups Experiment: Random Selection and Random Assignment. An ideal experimental design involves both (a) random selection of N subjects from a defined population and (b) random assignment of the selected subjects to treatment conditions. The results of the ideal experiment (if properly executed) can be generalized to the population from which the subjects were randomly selected.

Case II: Typical Randomized-Groups Experiment: Accessible Selection and Random Assignment. If the subjects for the experiment are not randomly selected from a defined population but are simply a collection of people accessible to the experimenter, it is still appropriate to randomly assign these subjects to treatment conditions and to apply conventional inferential procedures. Although the tests and/or confidence intervals are appropriate, the generalization of results is more ambiguous than with Case I selection. If little is known about the characteristics of the selected subjects, then little can be stated about the population to which the experimental results generalize. The experimenter can only state that the results can be generalized to a population of subjects who have characteristics similar to those who were included in the study. This may not be saying much if little is known about the accessible subjects. Generalization in Case I situations is based on principles of statistical inference; Case II generalization is based on logical considerations and speculation concerning the extent to which the accessible subjects are similar to those in the population to which the experimenter hopes to generalize. Although Case I selection is desirable from the point of view of generalizing results, such selection is frequently impractical; Case II selection is the rule rather than the exception in many areas of behavioral and medical science research.

Knowledge regarding the generality of results from experiments based on Case II selection can be improved by collecting thorough information regarding the characteristics of the accessible subjects. The form of this information may range from general measures of demographic characteristics to very specific medical diagnostic indices. After this information is collected, it is usually reported along with treatment results in order to characterize the type of subjects to which the estimated treatment results apply. The role of the information in this situation is to help provide a context for interpreting the results after the experiment is finished. But there are two other important roles for information regarding subject characteristics.

Subject information can play a role in both the design of the experiment and in the statistical analysis. If the information is appropriately used in the design or analysis, there are dual advantages of doing so. First, the power will be increased relative to what it would have been otherwise; second, the generality of the results may be more clearly defined. The relevance of subject information to these two issues is considered in the next section.

1.8 CONTROL OF NUISANCE VARIATION

The term "nuisance variable" is often applied to variables that are believed to affect scores on the dependent variable but are of no experimental interest. This is an important issue in experimental design because the size of the error term used in computing hypothesis tests and confidence intervals is a function of the amount of nuisance variation. Hence, the power of statistical tests and the width of confidence intervals is also a function of the amount of nuisance variation. The use of many

complex experimental designs and advanced statistical analyses is motivated by the desire to control nuisance variation.

Suppose we are interested in the effects of two methods of teaching reading. Students from a single classroom are randomly assigned to one of the two treatment conditions. If the students in the classroom are very heterogeneous with respect to reading ability (a subject characteristic), we can expect much within-group variation on measures of reading proficiency. The large within-group variation will be reflected in a large estimate for the standard error of the difference. Consequently, the large estimate for the standard error of the difference will lead to a small t-ratio and it will be necessary to conclude that the inferential evidence for an effect is insufficient to reject the null hypothesis. Even if the outcome of the experiment reveals a relatively large difference between the dependent variable means, the t-ratio may be too small to be declared statistically significant; correspondingly, the confidence interval on the mean difference is likely to be very wide.

Some method of controlling (at least partially) this nuisance variable should be employed to increase the power of the analysis and to reduce the width of the confidence interval. Several methods available for this purpose are described below.

Select subjects who are homogeneous with respect to the nuisance variable. One method of controlling nuisance variability involves (a) collecting data on the nuisance variable from a large collection of potential subjects, and (b) selecting only those subjects from the collection who have the same score (or very similar scores) on a measure of the nuisance variable. Obviously, if subjects included in the experiment have been selected in such a way that they all have the same preliminary reading score, the influence of reading skill as a source of within-group variation on the dependent variable has been reduced. If an acceptable number of subjects having the same nuisance variable score can be found, these subjects are randomly assigned and the experiment is carried out as a conventional randomized-groups experiment. The major problem with this method is that it is often very difficult (or expensive) to find a sufficient number of subjects with the same score on the nuisance measure.

Use the nuisance variable as one of the factors in a two-factor design. A frequently used method of dealing with nuisance variability is to employ the nuisance variable to form levels of one of the factors in a two-factor design. Hence, one factor would consist of levels of the nuisance factor and the other factor would consist of levels of the treatment factor.

Use blocking or matching. Matched pair designs and randomized block designs are effective design strategies for increasing power relative to what it would be using a randomized-groups design. Measures on a nuisance variable are used as the matching or blocking variable with these designs.

Use a repeated measures design. In many cases, power can be greatly increased if each subject is exposed to each treatment in a repeated measures design. This is an excellent strategy if there is little chance of carryover effects. Unfortunately, there are many cases in which this is not true.

Use statistical control of nuisance variation: The analysis of covariance. Each of the approaches mentioned above is a design strategy for contending with nuisance

variation in true experiments. The appropriate analysis of data from these designs is likely to have higher power than is associated with the conventional analysis of the typical randomized-groups design. But these alternative designs are not the only path to higher power.

An alternative strategy for contending with nuisance variation (and therefore increasing power) is to use a form of statistical control known as the analysis of covariance (ANCOVA). Unlike the strategies mentioned above, ANCOVA requires no departure from the conventional randomized-groups design with respect to subject selection and assignment. This approach has certain practical and statistical advantages over the alternative procedures mentioned above. Because these advantages are best revealed through direct comparison with conventional alternative methods, a review of matched pair, randomized block, and repeated measures designs is presented in the next two chapters.

1.9 SOFTWARE

Minitab input and output for the analysis of the pain data example are presented below.

As soon as *Minitab* is opened follow this path: Menu bar > Editor > Enable Commands. The *Minitab* prompt will then appear. Type the dependent variable scores in c1 and the associated group number in column c2. Then follow the commands shown below Figures 1.1 and 1.2.

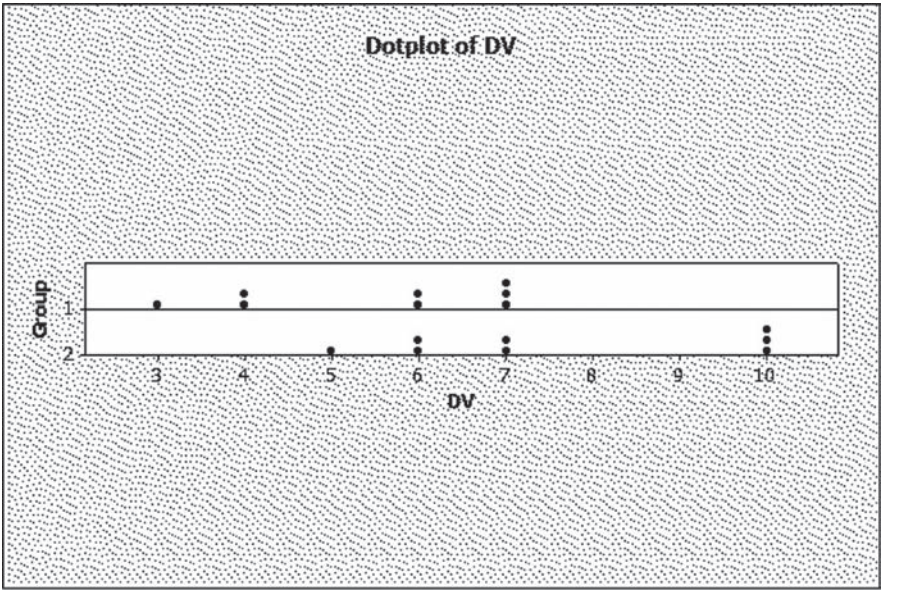


Figure 1.1 Dotplot of pain data.

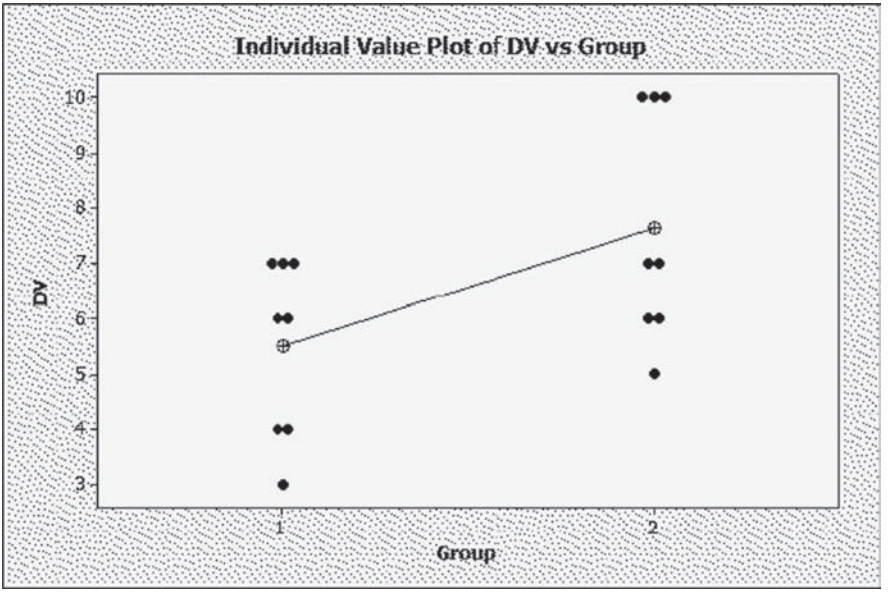


Figure 1.2 Individual value plot of pain data.

```
MTB > print c1 c2

Data Display
Row  DV  Group
  1   6      1
  2   7      1
  3   3      1
  4   7      1
  5   7      1
  6   4      1
  7   4      1
  8   6      1
  9   5      2
 10  10      2
 11   7      2
 12  10      2
 13   6      2
 14   7      2
 15   6      2
 16  10      2

MTB > Dotplot ('DV') * 'Group'.

MTB > TwoT 'DV' 'Group';
SUBC > Pooled;
SUBC > GIndPlot.
```

```
Two-Sample T-Test and CI: DV, Group
Two-sample T for DV

Group  N  Mean  StDev  SE Mean
1      8  5.50   1.60     0.57
2      8  7.63   2.07     0.73

Difference = mu (1) - mu (2)
Estimate for difference:  -2.125

95% CI for difference:  (-4.108, -0.142)

T-Test of difference = 0 (vs not =): T-Value = -2.30
P-Value = 0.037  DF = 14
Both use Pooled StDev = 1.8492
```

1.10 SUMMARY

Both descriptive and inferential statistics are relevant to the task of describing the outcome of a research study and to generalizing the outcome to the population from which the subjects were selected. The major emphasis in evaluating data should be on descriptive statistics, beginning with plots of the original data. Description in the form of easily understood summary statistics such as means, mean differences, and variances (in the original metric) is also essential. Standardized effect sizes and correlation ratios are sometimes helpful additional descriptors. Statistical inference is likely to be of interest when interpreting results, but it should be pursued only after thorough description. Although the reporting of results in terms of statistical significance is conventional practice, confidence intervals are usually more informative. The essentials of elementary statistical decision theory include the concepts of type I error, type II error, and power. Among the goals of a well-designed experiment are low probability of type I error (α), low probability of type II error (β), and high power ($1 - \beta$). More important than these goals is the extent to which the experimental design provides clear results. Random assignment plays an important role in both providing unconfounded estimates of treatment effects and in justifying statistical inference. Complex experimental designs are frequently used to provide higher power than is associated with simple randomized-groups designs. An alternative approach for increasing power is known as the analysis of covariance.

CHAPTER 2

Review of Simple Correlated Samples Designs and Associated Analyses

2.1 INTRODUCTION

The independent samples two-group experiment and the typical analysis associated with this design were reviewed in Chapter 1. This chapter reviews three simple correlated samples designs that are common in behavioral and biomedical science research. The conventional analyses for these designs are also reviewed.

2.2 TWO-LEVEL CORRELATED SAMPLES DESIGNS

The most typical randomized groups experiments and observational studies are based on two independent samples, but it is not unusual to encounter designs that yield correlated samples. The three most popular designs that yield correlated (or dependent) samples are the pretest–posttest study, the matched pairs experiment, and the two-level repeated measures experiment. These designs are similar in that each one has two levels and the most common method of analysis (illustrated in Table 2.1) is the same for all of them. However, there are important differences among these designs regarding the motivation, execution, and the ultimate internal validity of conclusions reached.

One-Group Pretest–Posttest Study

A major motivation for using the one-group pretest–posttest design is that it is easy to implement. Three steps are involved in setting up such a study: pretesting, intervention, and posttesting. Consider these steps in the context of an evaluation of a weight loss program. First, before the program is initiated, a measure of body

The Analysis of Covariance and Alternatives: Statistical Methods for Experiments, Quasi-Experiments, and Single-Case Studies, Second Edition. Bradley E. Huitema.

weight is obtained from each subject in a single group that has been selected for study. This preliminary weight measure is called the pretest because it is obtained before the intervention is applied; the purpose of the pretest is to provide a baseline against which a subsequent measure can be compared. Second, an intervention (the weight loss program) is carried out. Third, after the intervention is completed, posttest measures (body weight) are obtained. The pretest mean establishes the overall baseline weight before the intervention is applied and the posttest mean describes the overall weight after the intervention is applied; it is natural to compare the pretest mean with the posttest mean.

The parametric inferential procedures usually used to evaluate the difference between the pretest and posttest means are the correlated samples t-test and/or the corresponding confidence interval. If the null hypothesis associated with this t-test (i.e., $H_0 : \mu_{\text{Pre}} = \mu_{\text{Post}}$) is rejected, the researcher concludes that the pretest and posttest population means are not equal. Plausible values for the size of the population mean difference are contained in the corresponding confidence interval.

Matched Pairs Experiment

The matched pairs experiment involves carrying out the following steps: (1) randomly select N subjects from a defined population, (2) identify a variable (called the matching variable) that is believed to be correlated with the dependent variable, (3) obtain a measurement from each subject on the matching variable, (4) use the measurements obtained during step 3 to form $N/2$ pairs of subjects having relatively similar scores on the matching variable, (5) randomly assign one of the two subjects within each pair to treatment 1 and the other to treatment 2, (6) apply treatments, and (7) obtain measurements on the dependent variable.

Suppose a researcher is interested in studying the differential effects of two training methods on skill in performing common types of surgery. The matched pair experiment could be implemented as follows:

1. A random sample of 30 is selected from a defined population of medical students.
2. General psychomotor performance is hypothesized to be correlated with surgical skill and, therefore, to be useful for matching purposes.
3. A standardized test of psychomotor skill is administered to each student in the sample.
4. The 30 students are ordered with respect to test scores on the matching variable; those with the two lowest scores constitute the first pair. After the first pair is established, there are 28 students who remain to be matched. The two lowest performing students among those remaining constitute the second pair. This process continues until all $N/2 = 15$ pairs are formed.
5. One of the two students in each pair is then randomly assigned to one of the two training groups; the remaining student is assigned to the other treatment.

6. One method of training is applied to the 15 students in the first group and the second method is applied to the 15 students in the other group.
7. Finally, after the two training methods are completed, all students are evaluated with respect to surgical skill.

The conventional parametric statistical analysis of data obtained using the matched pairs experimental design is the correlated samples *t*-test and/or the associated confidence interval. The null hypothesis associated with this *t*-test is written as follows: $H_0 : \mu_1 = \mu_2$. When this hypothesis is rejected in the context of a well-executed matched pairs experiment, it is inferred that there are causal effects of the treatments; this inference is relevant to the population from which the subjects were initially selected. The corresponding confidence interval provides the plausible range for the difference between population means.

Two-Level Repeated Measures Experiment

The two-level repeated measures experiment is so named because each subject is treated repeatedly (once under each treatment) and is measured on the dependent variable repeatedly (once following the administration of each treatment). Because subjects are treated and measured twice, this design is reserved for experiments where there is justification for assuming that there are essentially no carryover effects from one treatment and measurement to another.

Experiments of this type are designed as follows: (1) n subjects are randomly selected from a defined population, (2) the order of administration of the two treatment conditions is randomized independently for each subject, (3) a treatment (specified in step 2) is applied to each subject, (4) the dependent variable is measured, (5) the alternative treatment is applied, and (6) the dependent variable is measured for the second time.

The conventional statistical analysis of data based on this design is the correlated samples *t*-test and/or the associated confidence interval. The null hypothesis associated with the *t*-test is written as follows: $H_0 : \mu_1 = \mu_2$. If this hypothesis is rejected, it is inferred that the mean for the population exposed to the first treatment condition differs from the mean for that same population exposed to the second treatment. The corresponding confidence interval provides the range of plausible differences between the two population means.

Interpretation Issues Associated with the Three Correlated Sample Designs

Although the appropriate statistical analysis is the same for each type of correlated samples design, the interpretation of results differs for each variant. There are major differences across these designs with respect to both internal and external validity.

The one-group pretest–posttest design is especially susceptible to internal validity threats. If the null hypothesis (i.e., $H_0 : \mu_{\text{Pre}} = \mu_{\text{Post}}$) is rejected, the researcher is justified in stating that the pretest and posttest population means are different. But evidence of a difference between these means is not necessarily a convincing

demonstration that the intervention is the reason for the difference. Plausible alternative explanations for the difference are discussed in depth in many books on research design (see, e.g., Shadish et al. (2002)). The list of these explanations (i.e., threats to internal validity) is long; it includes issues such as history, maturation, mortality, testing, instrumentation, and statistical regression. Each threat can completely invalidate conclusions regarding the effects of interventions.

Unlike the pretest–posttest design, well-executed matched pair and repeated measures designs generally lead to trustworthy estimates of causal effects of treatments. But matched pair and repeated measurement designs are not identical with respect to the generality of results.

Consider the outcome of a recent matched pair experiment that was designed to evaluate the differential effects of two drugs on the amount of pain experienced by postoperative patients. The first mean was based on subjects exposed to drug I, whereas the second mean was based on subjects exposed to drug II. None of the subjects in the first group was exposed to drug II and none of the subjects in the second group was exposed to drug I. The difference between the two sample means can be conceptualized as an estimate of the difference between the mean score for a population that has been exposed to drug I and the mean score for an exactly equivalent population that has been exposed to drug II instead. Recall that the sample subjects were exposed to only one treatment or the other (i.e., this is a type of between-subjects design). There is no aspect of this design in which a subject is exposed to both drugs. Consequently, it can be stated that the sample mean difference is an estimate of the difference between the corresponding population means (where each population is exposed to only one of the two treatments). This interpretation differs from the interpretation of results obtained from a repeated measures experiment.

Results based on the two-condition repeated measures design require acknowledgement of the fact that each subject is exposed to both treatments. The results of experiments of this type are generalized to a population of subjects who have been exposed to both treatments. As is the case with the matched pair design, it can be inferred that the sample mean difference is an estimate of the population mean difference $\mu_1 - \mu_2$. But the interpretation of the population means is not the same in the matched pair and repeated measures designs.

In the case of the repeated measures design, population mean μ_1 refers to the mean of a population of subjects who have all been exposed to the first treatment. But approximately half of this population has been previously exposed to the second treatment as well. Similarly, μ_2 is the mean of a population exposed to the second treatment, but approximately half of this population has been previously exposed to the first treatment. Because many of the subjects in each population have been previously treated, the inference is not the same with repeated measures and matched pairs designs. Does it really matter?

If being exposed to a previously administered treatment and measurement has no effect whatsoever on the response to the currently administered treatment and measurement, then the inference is essentially the same with repeated measurement and matched pair designs. But it is often difficult to predict whether the previous

administration of one treatment will change the effect of another treatment. A nice example of this problem can be found in early experimental psychology work on conditioning. Grice and Hunter (1964) demonstrated that the size of eyelid-conditioning effects investigated using within-subject designs differs dramatically from the size of the effects found using between-subject designs. This is not to say that repeated measures designs should be avoided; rather, the point is that one should be aware that matched pairs and repeated measures experiments may answer different questions.

Correlated Samples *t*-Test for the Analysis of Pretest–Posttest, Matched Pairs, and Repeated Measurement Designs

The most common parametric method of analysis for the three correlated samples designs described above is the correlated samples *t*-test. A computational example of this test, the corresponding confidence interval, the standardized effect size, and the correlation ratio for these designs are presented in Table 2.1.

Table 2.1 Correlated Samples Analysis of the Two-Condition Repeated Measures Design

Subject	Treatment 1	Treatment 2	D = Difference
1	12	9	3
2	7	8	−1
3	5	3	2
4	11	10	1
5	12	10	2
6	3	1	2
7	9	5	4
8	6	3	3
Mean	8.125	6.125	2.000

A. Hypothesis Test

Null hypothesis: $H_0 : \mu_1 = \mu_2$

$$\text{Test statistic}: \frac{\bar{Y}_1 - \bar{Y}_2}{s_{\bar{D}}} = \frac{\bar{D}}{s_{\bar{D}}} = \frac{\bar{D}}{\sqrt{\dfrac{\sum_{i=1}^{n}(D_i - \bar{D})^2}{n(n-1)}}} = \frac{2.00}{0.53452} = 3.742 = t_{\text{obt}},$$

where:

D_i = is $(Y_{i,1} - Y_{i,2})$, the difference between the two raw scores associated with subject i;
$s_{\bar{D}}$ = is the estimate of the standard error of the difference between two correlated means;
n = is the number of differences; and
t_{crit} = is based on $df = n - 1$.

(*Continued*)

Table 2.1 Correlated Samples Analysis of the Two-Condition Repeated Measures Design (*Continued*)

Because t_{obt} (i.e., 3.742) exceeds the critical value ($t_{crit} = 2.365$, $df = 7$) for $\alpha_2 = .05$, the null hypothesis is rejected. Correspondingly, the null hypothesis is rejected because the p-value (.007) is less than α_2 (set at .05).

B. 95% Confidence Interval

$$\begin{aligned}&\left(\bar{Y}_1 - \bar{Y}_2\right) \pm [s_{\bar{D}}(t_{crit})]\\ &= 2.00 \pm [.53452\,(2.365)]\\ &= 2.00 \pm 1.264\\ &= (.74, 3.26)\end{aligned}$$

It is stated that the population mean difference ($\mu_1 - \mu_2$) falls within such an interval; the confidence coefficient associated with the statement is .95.

C. Standardized Effect Size

Pretest–Posttest Version.
Formulas for the standardized effect size differ with the particular version of correlated samples design. In the case of the pretest–posttest version, a useful formula is

$$\frac{\bar{Y}_{\text{Pre}} - \bar{Y}_{\text{Post}}}{\sqrt{\dfrac{\sum_{i=1}^{n} y_{i\text{Pre}}^2}{n-1}}} = \frac{\bar{Y}_{\text{Pre}} - \bar{Y}_{\text{Post}}}{s_{\text{Pre}}} = g,$$

where:

$n =$ is the number of subjects (equal to the number of differences between pretest and posttest measures);

$\sum_{i=1}^{n} y_{i\text{Pre}}^2 =$ is the sum of the squared deviation scores, i.e., $y_{i\text{Pre}}^2 = (Y_{i\,\text{Pre}} - \bar{Y}_{\text{Pre}})^2$ on the pretest; and

$s_{\text{Pre}} =$ is the standard deviation on the pretest measure.

Matched Pairs and Repeated Measures Version
The formula for g for matched pairs and repeated measures designs is shown below; it is essentially the same as in the case of independent samples designs:

$$\text{Standardized effect size} = \frac{\bar{Y}_1 - \bar{Y}_2}{\sqrt{\dfrac{\sum_{i=1}^{n_1} y_i^2 + \sum_{i=1}^{n_2} y_i^2}{n_1 + n_2 - 2}}} = \frac{\bar{Y}_1 - \bar{Y}_2}{s_w} = g.$$

A minor difference exists in the way n_1, n_2, and N are defined for repeated measures designs relative to the way they are defined for independent samples and matched pairs designs. The numbers of observations under the two treatments (i.e., n_1 and n_2) add up to the total number of subjects in the experiment (N) for both independent samples and matched pairs designs, but this is not true in the case of repeated measures designs. In the latter case, the total number of subjects is n and these subjects produce $2n = N$ observations.

Table 2.1 (*Continued*)

The application of this formula to the data from the example repeated measures design yields the following value:

$$\frac{\bar{Y}_1 - \bar{Y}_2}{s_w} = \frac{2}{\sqrt{12.125}} = .57 = g.$$

D. Correlation Ratio

$$\frac{t_{obt}^2}{t_{obt}^2 + (N-2)} = \frac{1.32}{1.32 + (16-2)} = .09 = \hat{\eta}^2,$$

where t_{obt} is the obtained value of t using the independent samples (not the correlated samples) formula applied to the two correlated samples.

The analysis of the example data shown in Table 2.1 can be summarized as follows. The difference between the sample means is two points; further, (A) the hypothesis test leads to the conclusion that this difference is statistically significant, (B) the 95% confidence interval suggests that plausible values for the difference between the population means range from .74 to 3.26, (C) the standardized effect size is a little over half a standard deviation, and (D) the proportion of the total variation in the experiment explained by the treatments is approximately .09.

There are many other methods of analyzing data from correlated samples designs. The most frequently encountered nonparametric analog to the correlated samples t-test is the Wilcoxon signed rank test. The application of this test to the example data yields the same decision, but the p-value is larger. The advantage of the Wilcoxon test is that it is far less affected by outliers than is t. This was clearly not an issue in the example study.

It is interesting to note that if the data from this correlated samples design had been incorrectly analyzed using the independent samples t-test for the inferential analysis, the conclusion would not have been statistically significant. The results of this incorrect analysis yield $p = .27$ and a 95% confidence interval that is much wider (-1.73, 5.73). This interval incorrectly suggests that zero is a credible value to entertain as the difference between population means. Hence, it is obvious that there is a major difference in the inferential statements associated with the correct and incorrect analysis of these data.

The pretest–posttest design, like the other correlated samples designs, is almost always analyzed using the methods described above. It is pointed out in Chapter 30 that the conventional analysis can be very misleading in the case of many versions of the pretest–posttest design and that a more adequate (but more complex) analysis is available.

2.3 SOFTWARE

Minitab input and output for the analysis of two-condition repeated measures and matched pairs designs are shown below. Commands for appropriate graphics, descriptive statistics, correlated samples *t*-test, and the corresponding confidence interval are illustrated for the sample data contained in Table 2.1. The data for treatments 1 and 2 were entered in worksheet columns c1 and c2.

Command to print data:

```
MTB > print c1 c2
Data Display
Row  Treatment I  Treatment 2
  1           12            9
  2            7            8
  3            5            3
  4           11           10
  5           12           10
  6            3            1
  7            9            5
  8            6            3
```

Commands to compute correlated samples t, 95% confidence interval, and histogram of differences:

```
MTB > Paired 'Treatment I' 'Treatment 2';
SUBC> GHistogram.
```

Output:

```
Paired T-Test and CI: Treatment I, Treatment 2
Paired T for Treatment I - Treatment 2
             N   Mean  StDev  SE Mean
Treatment I  8   8.13   3.40     1.20
Treatment 2  8   6.13   3.56     1.26
Difference   8  2.000  1.512    0.535
95% CI for mean difference: (0.736, 3.264)
T-Test of mean difference = 0 (vs not = 0): T-Value = 3.74
P-Value = 0.007
```

2.4 SUMMARY

The one-group pretest–posttest design, the matched pairs design, and the two-condition repeated measures design are similar in certain respects. They all yield two samples of observations, these observations are likely to be correlated, and a

common method of analysis can be used for all of them. But these designs are very different in terms of internal and external validity. The pretest–posttest design is quite vulnerable to many threats to internal validity, whereas the matched pairs and repeated measures designs have high internal validity and are useful alternatives to the randomized groups design. The major advantage of correlated samples designs is that the dependency among the observations leads to analyses having higher power than is obtained using conventional independent samples approaches.

CHAPTER 3

ANOVA Basics for One-Factor Randomized Group, Randomized Block, and Repeated Measurement Designs

3.1 INTRODUCTION

Several types of two-condition experimental designs and associated analyses were reviewed in Chapters 1 and 2. Three direct extensions of those designs and analyses for more than two conditions are described in this chapter. They are labeled as one-factor randomized group, randomized block, and repeated measurement designs.

3.2 ONE-FACTOR RANDOMIZED GROUP DESIGN AND ANALYSIS

The basic two-group experiment reviewed in Chapter 1 is an example of a randomized design with two levels. If more than two types of experimental condition are compared, we have a one-factor experiment with $J > 2$. In this case the treatment groups are formed through random assignment of subjects to treatments (as in a two-group experiment) but the typical method of analyzing the data is the analysis of variance (ANOVA) rather than a collection of independent sample t-tests. The latter approach is usually avoided because most researchers want to control the probability of type I error at α for the whole collection of comparisons in the experiment rather than for each individual comparison.

If individual t-tests are used for all pairwise comparisons it is well known that the probability of type I error for the whole experiment increases rapidly as the number of groups increases. In a three-group experiment, for example, the probability of

The Analysis of Covariance and Alternatives: Statistical Methods for Experiments, Quasi-Experiments, and Single-Case Studies, Second Edition. Bradley E. Huitema.

making at least one type I error in the collection of *t*-tests on pairwise comparisons is approximately .14 when α is set at .05 for each individual test. Correspondingly, for $J = 4$, 5, and 6, the approximate probabilities of at least one type I error are .26, .40, and .54, respectively. If ANOVA is used instead of the set of *t*-tests on all pairwise comparisons, the probability of type I error for the overall *F*-test remains at .05, regardless of the number of groups (given, α set at .05 and the assumptions of the test are met).

Essential Ideas Underlying ANOVA

Consider an experiment in which $J = 3$. Suppose 30 members of a large organization have been selected to participate in an experimental investigation of the effects of three different types of training. These participants are randomly assigned to form three groups of 10. A different treatment is applied to each group and a standardized measure of performance is obtained from each subject. The resulting data are then analyzed to evaluate the differential effects (if any) of the three treatments.

As usual, the analysis begins with a careful inspection of the raw data, the essential descriptive statistics are computed, and finally a formal inferential method is chosen to test the following hypothesis: H_0: $\mu_1 = \mu_2 = \mu_3$. Because this is an omnibus hypothesis stating that all three population means are equal, it is appropriate to use a method that provides an omnibus test. The ANOVA *F* is a test of this type.

The main purpose of the fixed-effects one-factor ANOVA is to test the hypothesis that *J* population means are equal. This omnibus null hypothesis is written as H_0: $\mu_1 = \mu_2 = \cdots = \mu_J$.

The conceptual framework underlying the analysis of variance is the ANOVA model; this model (along with related computation relevant to it) is shown in Table 3.1. It can be seen that the model is a formal explanatory framework that contains three components used to explain dependent variable scores. This framework leads to an interest in computing the total amount of variation in an experiment; this variation is more specifically defined as the total sum of squares (SS_T). After the total sum of squares is computed, it is partitioned into two additive components known as the between-group (also called *among-group* or *treatment)* sum of squares (SS_B) and the within-group (or *error*) sum of squares (SS_W). Hence, $SS_T = SS_B + SS_W$, where the subscripts T, B, and W represent total, between groups, and within groups, respectively. These sum of squares are involved in computing the total, between-group, and within-group estimates of variance; these variance estimates are often called mean squares (MS). They are obtained by dividing sum of squares by the appropriate degrees of freedom, i.e.,

$$\frac{SS_T}{N-1} = \hat{\sigma}_T^2,$$

$$\frac{SS_B}{J-1} = \hat{\sigma}_B^2,$$

$$\frac{SS_W}{N-J} = \hat{\sigma}_W^2,$$

Table 3.1 One-factor Analysis of Variance Model, Definitions of Sum of Squares, Summary Table, MS Expected Values

Model: $Y_{ij} = \mu + \alpha_j + \varepsilon_{ij}$,

where

Y_{ij} = dependent variable score for the ith subject in the jth treatment;
μ = overall population mean (i.e., the mean of the individual population means);
α_j = effect of treatment j; and
ε_{ij} = error component associated with the ith subject in treatment j.

Definitions of Total, Between-Group, and Within-Group Sums of Squares:

Total sum of squares: $\sum_{j=1}^{J} \sum_{i=1}^{n_j} (Y_{ij} - \bar{Y}..)^2 = \mathrm{SS_T}$

Between-group sum of squares: $\sum_{j=1}^{J} n_j (\bar{Y}_j - \bar{Y}..)^2 = \mathrm{SS_B}$

Within-group sum of squares: $\sum_{j=1}^{J} \sum_{i=1}^{n_j} (Y_{ij} - \bar{Y}_j)^2 = \mathrm{SS_W}$

Form of the ANOVA Summary Table:

Source	SS	*df*	MS	*F*
Between-Groups	$\mathrm{SS_B}$	$df_B = J - 1$	$\mathrm{MS_B}$	$\mathrm{MS_B/MS_W}$
Within-Groups	$\mathrm{SS_W}$	$df_W = N - J$	$\mathrm{MS_W}$	
Total	$\mathrm{SS_T}$	$df_T = N - 1$		

Expected Values for $\mathrm{MS_B}$ and $\mathrm{MS_W}$ when (A) H_0 Is True and (B) H_0 Is False:

	A. $\mu_1 = \mu_2$	B. $\mu_1 \neq \mu_2$
$E(\mathrm{MS_b})$	σ_W^2	$\sigma_W^2 + \dfrac{\sum_{j=1}^{J} n_j (\mu_j - \mu)^2}{J - 1}$
$E(\mathrm{MS_W})$	σ_W^2	σ_W^2

where N is the total number of subjects in the experiment, J is the number of treatment groups, and $\hat{\sigma}^2$ is a variance estimate.

If the null hypothesis regarding population means (i.e., H_0: $\mu_1 = \mu_2 = \cdots = \mu_J$) is true, all three variance estimators ($\hat{\sigma}_T^2$, $\hat{\sigma}_B^2$, and $\hat{\sigma}_W^2$) yield unbiased estimates of the common within-population variance (i.e., $\sigma_1^2 = \sigma_2^2 = \cdots = \sigma_J^2 = \sigma_W^2$). The notion of a common within-population variance σ_W^2 is assumed under the ANOVA model. It can be seen in Table 3.1 that two variance estimators ($\mathrm{MS_B}$ and $\mathrm{MS_W}$) appear in the ANOVA summary table under the mean square heading; the ratio of the two is used in testing the null hypothesis. Specifically, the between-group variance estimate ($\mathrm{MS_B}$) is divided by the within-group variance estimate ($\mathrm{MS_W}$) to obtain the F statistic. The notion that this F-ratio provides information on whether there are differences between population means needs explanation.

The logic for using the ratio of the two *variance* estimators to provide information regarding differences between *means* is not difficult if the concept of the expected value of a statistic is understood. Suppose the samples in an experiment have been randomly drawn from populations in which $\mu_1 = \mu_2 = \cdots = \mu_J$; i.e., the null hypothesis regarding population means is true. Consider computing MS_B, MS_W, and the ratio MS_B/MS_W on these sample data; record these values. Then return the sample observations to the populations from which they were sampled and again draw random samples from each population. Once again compute MS_B, MS_W, and the ratio MS_B/MS_W; record these values. Think about repeating these operations an infinite number of times. Then compute the average of all of the MS_B, the average of all the MS_W, and the average of all the MS_B/MS_W ratios. Think of these three averages as the expected value of MS_B, the expected value of MS_W, and the expected value of the F-ratio MS_B/MS_W. The notation for these three expected values is as follows: $E(MS_B)$, $E(MS_W)$, and $E(F)$. These expected values are shown in Table 3.1. Note that when the null hypothesis regarding population means is true, the expected values for both MS_B and MS_W are equal to the common within-population variance. Because each expected value [i.e., $E(MS_B)$ and $E(MS_W)$] is equal to the value of the parameter being estimated (i.e., σ_W^2), both MS_B and MS_W are said to be unbiased estimators. One would think that the expected value of the F-ratio MS_B/MS_W would be exactly 1.0 because both MS_B and MS_W yield unbiased estimates of the parameter σ_W^2 as long as $\mu_1 = \mu_2 = \cdots = \mu_J$. Actually, intuition fails us in this case, but not badly.

Although many elementary statistics books state that the expectation for F is exactly 1.0 when the null hypothesis is true, the exact expectation is actually $\nu_2/(\nu_2 - 2)$ where ν_2 is the denominator degrees of freedom for the F-ratio. (For the technically curious reader, the intuitive expected value of 1.0 is incorrect because the ratio of the two expectations is not the expectation of the ratio.)

When the null hypothesis is false (i.e., at least one population mean differs from the others), the between-group MS has an expected value that is greater than the expected value for the within-group MS. The obtained F will reflect this effect by being large. The obtained F statistic is compared with the tabled value of F (for some specified α level and degrees of freedom $= J - 1$ and $N - J$) to determine whether to reject or retain the null hypothesis. Alternatively, the null hypothesis is rejected when the obtained value of p is $\leq \alpha$.

The reason the F-ratio should be large when the population means are not equal can be understood by inspecting the expected values for MS_B and MS_W in Table 3.1. Note that $E(MS_B)$ and $E(MS_W)$ are not the same when the null hypothesis is false; they are the same only when H_0: $\mu_1 = \mu_2 = \cdots = \mu_J$ is true. It can be seen that the expected value for MS_W is equal to σ_W^2 regardless of whether the population means are equal or not. The situation is very different for $E(MS_B)$. If the population means are equal, the expected value for $MS_B = \sigma_W^2$, but if the population means are not equal, the expected value for MS_B is equal to $\sigma_W^2 + \frac{\sum_{j-1}^{J} n_j(u_j - \mu)^2}{J-1}$. It can be seen in this expression that the right-hand term will equal zero if there are no differences between population means, but it will be nonzero if at least one mean differs from

the others. In the latter case, it is easy to see that MS_B will be positively biased as an estimator of σ_W^2 whenever there are differences among population means and consequently the F-ratio will be larger than it would be if there were no differences among means.

Multiple Comparison Tests

The rejection of the omnibus null hypothesis (H_0: $\mu_1 = \mu_2 = \cdots = \mu_J$) simply indicates that there is sufficient information to conclude that there is at least one nonzero difference between the population means. In some experiments this conclusion may be all that is required. But often there is interest in additional comparisons between individual pairs of means. In this case, so-called *multiple comparison* tests are frequently applied. Many tests of this type have been developed; a simple and powerful one that is useful for all pairwise comparisons is known as the Fisher–Hayter test. This test is carried out only if the omnibus null hypothesis has been rejected using the ANOVA F. The null hypothesis associated with the Fisher–Hayter test is H_0: $\mu_i = \mu_j$, where the subscripts indicate two different populations. The total number of pairwise comparisons available in an experiment with J groups is $\frac{J(J-1)}{2}$.

The Fisher–Hayter test statistic q_{FH} is based on the following formula:

$$\frac{\bar{Y}_i - \bar{Y}_j}{\sqrt{\frac{\text{MS}_W}{2}\left[\frac{1}{n_i} + \frac{1}{n_j}\right]}} = q_{\text{FH}},$$

where i and j indicate the two samples involved in the comparison, MS_W is the within-groups mean square, n_i and n_j are the sample sizes associated with the two means in the comparison, and q_{FH} is the Studentized range statistic.

When actually computed on sample data, the obtained statistic is denoted by $q_{\text{FH obt}}$; the absolute value of this statistic is compared with the critical value of the Studentized range statistic for the specified level of α. The critical q_{FH} is the value found in the Studentized range table (available in the Appendix) at the intersection of the column headed with the value equal to $J - 1$ and the row with degrees of freedom $= N - J$. The null hypothesis is rejected when $|q_{\text{FH obt}}| \geq$ critical q_{FH}. These tests have the claimed desirable properties only when the preliminary ANOVA F is statistically significant. If the F is not statistically significant, the omnibus null hypothesis is retained and it is concluded that there are insufficient data to claim differences among the population means; in this case there is no interest in F–H tests and they should not be computed. The level set for alpha should be consistent for the ANOVA F-test and the Fisher–Hayter tests.

Interpretation of Multiple Comparison Tests

The probability of making one or more type I errors in the set of comparisons is equal to or less than α when using the Fisher–Hayter procedure. It is possible (but

rare) for the ANOVA F to be statistically significant but to then find that no pairwise comparison is significant using Fisher–Hayter tests. (This apparent inconsistency will occur less often with the Fisher–Hayter approach than with other multiple comparison tests such as the popular Tukey, Tukey–Kramer, Bonferroni, and Scheffé methods.) One reason this can occur is that the F-test is sensitive to a larger set of comparisons than the set of pairwise comparisons.

The experiment is conceptualized as all possible comparisons (which is an infinite number when there are at least three groups) under the ANOVA F model. The experiment is viewed as the collection of all pairwise comparisons in the case of the F–H tests. If the F-test is significant but no pairwise comparison is significant when these tests are used, it can be concluded that some complex comparison is significant. There are essentially two types of complex comparison; I label them as (i–j) and non-(i–j).

An (i–j) comparison is of the form where the mean of i means (where i may equal one) is compared with the mean of j different means (where j may equal one). The number of (i–j) comparisons is finite. For example, the total number of (i–j) contrasts in a three-group experiment is six. In addition to the three pairwise comparisons, we could compare the mean of the first group with the average of groups two and three, the means of the second group could be compared with the average of groups one and three, and the mean of the third group could be compared with the average of groups one and two. Comparisons of this type are often written as a linear expression that provides a "contrast," which is simply a formal way of writing a comparison. The general form of the linear expression for a population contrast is

$$c_1(\mu_1) + c_2(\mu_1) + \cdots + c_J(\mu_J) = \psi,$$

where $c_1, c_2, \ldots, c_J$ are the contrast coefficients, $\mu_1, \mu_2, \ldots, \mu_J$ are the population means, and ψ (psi) is the contrast. At least two of the contrast coefficients must be nonzero and the sum of the coefficients must be zero. Although the coefficients can be constructed in many ways, the standard form is to set the sum of the positive coefficients equal to one and the sum of the negative coefficients equal to minus one. The corresponding sample estimate of ψ is defined as $c_1(\bar{Y}_1) + c_2(\bar{Y}_2) + \cdots + c_J(\bar{Y}_J) = \hat{\psi}$.

The six (i,j) population contrasts associated with a three-group experiment are written as follows:

1. $1(\mu_1) - 1(\mu_2) + 0(\mu_3) = \psi$;
2. $1(\mu_1) + 0(\mu_2) - 1(\mu_3) = \psi$;
3. $0(\mu_1) + 1(\mu_2) - 1(\mu_3) = \psi$;
4. $1(\mu_1) - .5(\mu_2) - .5(\mu_3) = \psi$;
5. $-.5(\mu_1) + 1(\mu_2) - .5(\mu_3) = \psi$; and
6. $.5(\mu_1) + .5(\mu_2) - 1(\mu_3) = \psi$.

Note that all three population means are included in each expression even if the coefficient associated with a mean is zero. Also note that the sum of the contrast coefficients is zero for all six contrasts, the sum of the positive coefficients is always one, and the sum of the negative coefficients is always minus one. There are several reasons that these contrast coefficients are of interest. Among them, they are needed in order to compute certain multiple comparison tests and they can be used to easily demonstrate that this set of all (*i–j*) contrasts is not the set of all possible contrasts.

Because the collection of all (*i–j*) contrasts might appear to exhaust all contrasts that the reader can think of, it is reasonable to question my earlier statement that there exists an infinite number of possible contrasts in the three-group case. Consider the following contrast: $1(\mu_1) - .3(\mu_2) - .7(\mu_3) = \psi$. This is not an (*i–j*) contrast; (*i–j*) contrasts involve averages that are based on groups that are given equal weight. For example, the fourth contrast listed above involves the comparison of the first mean with the average of means two and three; the contrast coefficients associated with means two and three result in both of these groups receiving the same weight. But it can be seen in the contrast listed immediately above that the coefficients associated with groups two and three are $-.3$ and $-.7$. This tells us that the contrast is not of the (*i–j)* type; rather it is a non-(*i–j*) contrast. Although it is possible to conceptualize an infinite number of non-(*i–j*) contrasts by applying an infinite number of contrast coefficient combinations, it may seem absurd to actually be interested in non-(*i, j*) contrasts. Surprisingly, the ANOVA *F*-test is almost always a test of the significance of some non-(*i–j*) contrast! It is possible to figure out the exact contrast that is associated with the obtained value of *F* after the fact, but there is little interest in doing so. The main reason for mentioning this property of the *F*-test is to provide an explanation for how it is possible that no pairwise comparison is statistically significant when the ANOVA *F* is significant. Because there are many types of contrast other than pairwise, finding a significant *F* but no significant pairwise comparison implies that some complex comparison is significant.

If the ANOVA *F* is significant there absolutely *must* be a significant contrast. The contrast can be identified and, by applying a multiple comparison procedure known as Scheffé's method, the identified contrast will be significant. There is a mathematical correspondence between the ANOVA *F* and Scheffé's test; indeed, the critical value for Scheffé's method is based on the *F* distribution. If *F* is nonsignificant, there is no contrast that is significant using Scheffé's procedure; if *F* is significant, Scheffé's test on some contrast must be significant. It turns out, however, that Scheffé's procedure is not a good choice in the typical experiment where most interest is focused on pairwise comparisons. I recommend the Fisher–Hayter as the multiple comparison test of choice for all pairwise comparisons. This recommendation is based on computational simplicity and power. The only reason I mention computational simplicity is that most software packages have not yet implemented it. It is more powerful than the major competing procedures for this purpose (such as the Bonferroni, Scheffé, Tukey, and Tukey–Kramer methods).

Example 3.1 Data presented below are from a randomized group experiment with three levels ($J = 3$). The independent variable (i.e., the factor) is a type of training method and the dependent variable is a measure of achievement. Thirty subjects (N) were randomly assigned to produce three groups, each containing 10 independent subjects.

Method I	Method II	Method III
15	20	14
19	34	20
21	28	30
27	35	32
35	42	34
39	44	42
23	46	40
38	47	38
33	40	54
50	54	56

The three group means are $\bar{Y}_1 = 30$, $\bar{Y}_2 = 39$, and $\bar{Y}_3 = 36$. Differences among these means were judged to be of practical importance. The issue of whether data such as these were likely to occur under the condition that the null hypothesis was true was evaluated using a one-factor ANOVA.

ANOVA Summary (Example 3.1)

Source	SS	*df*	MS	*F*
Between groups	420	2	210	1.60
Within groups	3536	27	131	
Total	3956	29		

If α is set at .05, the hypothesis of the equality of the three-population means is retained because the obtained value of F is 1.60 and this does not exceed the critical value of F, which is 3.35 ($df = 2, 27$). Correspondingly, the p-value (obtained from *Minitab* output) is .22, which easily exceeds the α value established for the test. Because the null hypothesis is retained, the Fisher–Hayter tests are not computed. It is concluded that there are insufficient data to claim that the population means associated with the three training methods are different.

Example 3.2 The data in Example 3.1 have been modified to produce the data shown below for Example 3.2. Each subject in the second treatment now has a score that is five points higher than originally shown. The data in treatments I and III are unchanged.

Method I	Method II	Method III
15	25	14
19	39	20
21	33	30
27	40	32
35	47	34
39	49	42
23	51	40
38	52	38
33	45	54
50	59	56

The three-group means are now $\bar{Y}_1 = 30$, $\bar{Y}_2 = 44$, and $\bar{Y}_3 = 36$. The issue of whether data such as these are likely to occur under the condition that the null hypothesis is true is evaluated using a one-factor ANOVA. Note that the between-group sum of squares (and the between-group mean square) is over twice as large in Example 3.2 as in Example 3.1; this is because the differences among the means are larger in Example 3.2. Also note that the within-group sum of squares (and the within-group mean square) is identical in both examples. This illustrates that adding a constant to the scores in any group (and increasing the differences among the means) has no effect on the within-group sum of squares.

ANOVA Summary (Example 3.2)

Source	SS	*df*	MS	*F*
Between groups	987	2	493	3.77
Within groups	3536	27	131	
Total	4523	29		

If α is set at .05, the hypothesis of the equality of the three population means is rejected because the obtained value of F (3.77) exceeds the critical value of F, which, once again, is 3.35 ($df = 2, 27$). Correspondingly, the p-value is .036, which is less than the α value established for the test. Because the omnibus null hypothesis is rejected, individual Fisher–Hayter tests are justified for the three pairwise comparisons. The results of these tests (when α is set at .05) are shown below.

Groups	Sample Mean Difference	Obtained $\lvert q_{F\text{–}H} \rvert$	Critical $q_{F\text{–}H(J-1,\,N-J)=2,\,27}$	Decision
1, 2	$(30 - 44) = -14$	$\left\lvert \dfrac{-14}{\sqrt{\dfrac{131}{2}\left(\dfrac{1}{10} + \dfrac{1}{10}\right)}} \right\rvert = 3.87$	2.905	Reject H_0: $\mu_1 = \mu_2$

Groups	Sample Mean Difference	Obtained $\lvert q_{F-H} \rvert$	Critical $q_{F-H(J-1,\,N-J)=2,\,27}$	Decision
1, 3	$(30-36)=-6$	$\left\lvert \dfrac{-6}{\sqrt{\dfrac{131}{2}\left(\dfrac{1}{10}+\dfrac{1}{10}\right)}} \right\rvert = 1.66$	2.905	Retain H_0: $\mu_1=\mu_3$
2, 3	$(44-36)=8$	$\left\lvert \dfrac{8}{\sqrt{\dfrac{131}{2}\left(\dfrac{1}{10}+\dfrac{1}{10}\right)}} \right\rvert = 2.21$	2.905	Retain H_0: $\mu_2=\mu_3$

It can be seen that the only statistically significant pairwise test is the one involving the comparison of training methods 1 and 2. There is insufficient evidence to claim that there is a difference between population means 1 and 3 or 2 and 3.

Simultaneous Confidence Intervals

An alternative (and often more informative) approach to the multiple comparison problem is to compute simultaneous confidence intervals (SCI). This involves computing a confidence interval for each pairwise comparison. Just as there are many multiple comparison hypothesis testing procedures, there are many procedures that can be used to construct simultaneous confidence intervals. There is not, however, an SCI analog for every multiple comparison hypothesis testing procedure. For example, there is no SCI analog for the Fisher–Hayter procedure. The reason for this is that the Fisher–Hayter is a sequential testing procedure that consists of two stages, where the first stage is the F-test on the omnibus null hypothesis. In general, there is no SCI analog for sequential multiple comparison testing procedures.

The Tukey–Kramer method is an example of a one-stage approach that is appropriate for both hypothesis tests and SCI. The computation of 95% simultaneous confidence intervals using the Tukey–Kramer approach involves constructing a confidence interval for each pairwise difference using

$$\left(\bar{Y}_i - \bar{Y}_j\right) \pm q_{.05,J,N-J}\sqrt{\frac{\mathrm{MS_W}}{2}\left[\frac{1}{n_i}+\frac{1}{n_j}\right]},$$

where i and j indicate the two samples involved in the comparison, $\mathrm{MS_W}$ is the within-group mean square, n_i and n_j are the sample sizes associated with the two means in the comparison, and $q_{.05,\,J,\,N\,-\,J}$ is the 5% critical value of the Studentized range statistic found in the Studentized range table at the intersection of the column headed with the value equal to J and the row with degrees of freedom $= N - J$.

Unlike the Fisher–Hayter hypothesis testing approach, the Tukey–Kramer SCI procedure does not involve a preliminary stage using the ANOVA F-test. It is typical, however, to first compute ANOVA in order to obtain the required MS_W. But the outcome of the F-test is irrelevant to the SCI. Regardless of the value of F, the set of SCI is computed.

Interpretation of Simultaneous Confidence Intervals

The distinction previously made between an experiment defined as one simple contrast and an experiment defined as a collection of contrasts is important when confidence intervals are interpreted. In the case of two-group experiments, there is an interest in constructing one interval that will contain the population mean difference. If, for example, a 95% confidence interval is constructed in each of 100 independent two-group experiments, the difference between population means is expected to be found in all but five of the 100 intervals. In other words, in the long run 95% of the intervals will contain $(\mu_1 - \mu_2)$. It should be kept in mind that the interpretation of confidence intervals is not affected by whether the null hypothesis is true. That is, there is no reason to run hypothesis tests to decide whether to construct confidence intervals. If a confidence interval contains zero, the corresponding hypothesis test will not result in the rejection of the null hypothesis. If the confidence interval does not contain zero, the corresponding hypothesis test will result in the rejection of the null hypothesis. In any event, when the confidence interval approach is employed with two-group experiments, the population difference will be contained in the intervals $100(1 - \alpha)\%$ of the time.

If the experiment contains more than two groups and simultaneous confidence intervals are constructed, the interpretation of each interval is not the same as just described for the two-group case. When three or more groups are involved, the experiment contains more than one comparison and more than one confidence interval. As is the case with multiple comparison hypothesis tests, the experiment is viewed as a collection of comparisons, and the interpretation of the intervals is tied to this collection. Suppose a three-group experiment is carried out and 95% simultaneous confidence intervals are constructed on the three pairwise contrasts by using the Tukey–Kramer approach. We expect the three population contrasts to be contained in the three simultaneous confidence intervals. Specifically, the probability is at least .95 that *all three* population contrasts (i.e., $\mu_1 - \mu_2$, $\mu_1 - \mu_3$, and $\mu_2 - \mu_3$) are contained within the corresponding intervals. The probability statement refers to the whole collection of intervals; it does not refer to an individual interval. For example, the probability that the first interval contains $\mu_1 - \mu_2$ is not .95. Actually, the probability of any one of the population differences falling within the corresponding interval must be greater than .95.

Computational Example of SCI

Simultaneous confidence intervals for the data listed for Example 3.2 (above) are computed as follows:

Contrast $= \psi$ population mean difference	Estimated contrast $= \hat{\psi}$ sample mean difference	SCI $\left(\bar{Y}_i - \bar{Y}_j\right) \pm q_{.05,J,N-J}\sqrt{\frac{\text{MS}_\text{W}}{2}\left[\frac{1}{n_i} + \frac{1}{n_j}\right]}$
$\mu_1 - \mu_2$	$(30 - 44) = -14$	$(-26.70, -1.30)$
$\mu_1 - \mu_3$	$(30 - 36) = -6$	$(-18.70, 6.70)$
$\mu_2 - \mu_3$	$(44 - 36) = 8$	$(-4.70, 20.70)$

Minitab *Input*

Minitab worksheet column c1 was chosen to identify the levels of the factor. Because there are three levels in the example and each level has 10 observations, the column is constructed using the following command line editor path:

```
MTB > set c1

DATA> 10(1) 10(2) 10(3)

DATA> end
```

The dependent variable scores are then entered on the worksheet in column c2. The two columns are labeled as GP and Y; they can be printed using the following command:

```
MTB > print c1 c2
```

Data Display

```
Row  GP   Y
  1   1  15
  2   1  19
  3   1  21
  4   1  27
  5   1  35
  6   1  39
  7   1  23
  8   1  38
  9   1  33
 10   1  50
 11   2  25
 12   2  39
 13   2  33
 14   2  40
 15   2  47
 16   2  49
 17   2  51
```

```
18   2  52
19   2  45
20   2  59
21   3  14
22   3  20
23   3  30
24   3  32
25   3  34
26   3  42
27   3  40
28   3  38
29   3  54
30   3  56
```

Menu commands for the analysis: The menu input path for computing the one-factor ANOVA and the Tukey–Kramer simultaneous confidence intervals using the example data is as follows:

Editor → Enable Commands → set c1 → (enter return) → 10(1) 10(2) 10(3) → (enter return) → [enter data in column c2 on worksheet] → Stat → ANOVA → One-Way → Response:c2 → Factor:c1 → Comparisons → Tukey's, family error rate → OK → Graphs → Individual value plot → OK → OK

Command line editor commands: The ANOVA, the Tukey–Kramer simultaneous confidence intervals, and the plot are produced using the following commands (see Figure 3.1):

```
MTB > Oneway 'Y' 'GP';
SUBC>   Tukey 5;
SUBC>   GIndPlot.
```

Output:

```
One-way ANOVA: Y versus GP
Source  DF    SS   MS     F      P
GP       2   987  493  3.77  0.036
Error   27  3536  131
Total   29  4523

S = 11.44   R-Sq = 21.82%   R-Sq(adj) = 16.02%


                          Individual 95% CIs For Mean Based on
                                 Pooled StDev

Level   N   Mean  StDev  --+--------+--------+--------+------
1      10  30.00  10.87   (--------*-------)
2      10  44.00   9.98                     (-------*-------)
```

```
3      10  36.00  13.23           (--------*--------)
                           --+---------+---------+---------+------
                           24.0      32.0      40.0      48.0

     Pooled StDev = 11.44

     Tukey 95% Simultaneous Confidence Intervals

     All Pairwise Comparisons among Levels of GP

     Individual confidence level = 98.04%

GP = 1 subtracted from:
GP  Lower  Center  Upper ------+---------+---------+---------+--
2    1.30   14.00  26.70                  (---------*---------)
3   -6.70    6.00  18.70            (---------*---------)
                           ------+---------+---------+---------+--
                               -12         0        12        24

GP = 2 subtracted from:
GP Lower  Center  Upper  ------+---------+---------+---------+--
3  -20.70 -8.00   4.70   (---------*---------)
                           ------+---------+---------+---------+--
                               -12         0        12        24
```

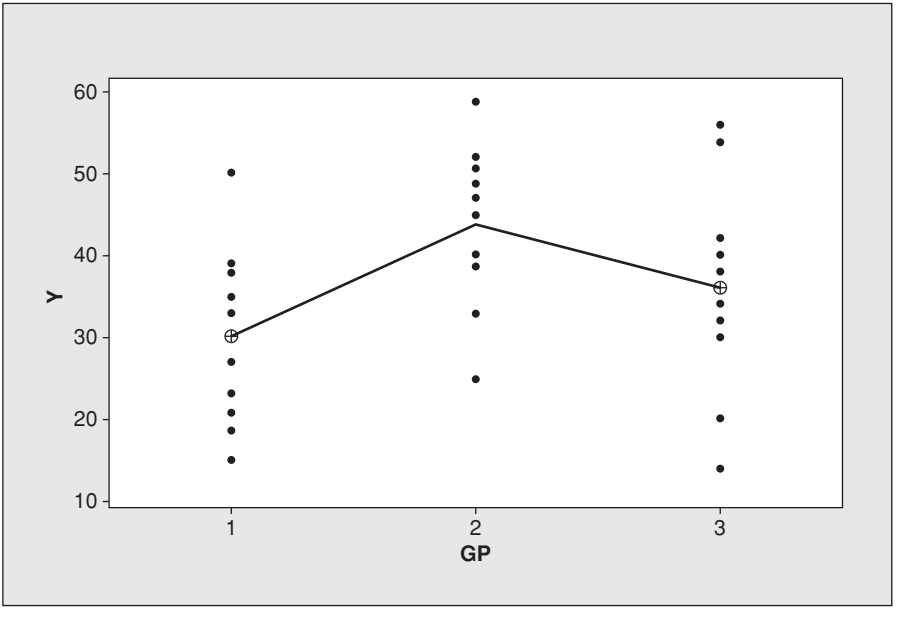


Figure 3.1 Individual value plot of Y vs. GP.

SPSS *Input*

Because the *SPSS* and *Minitab* worksheets are similar, the *SPSS* data input for computing the ANOVA and Tukey–Kramer tests is not shown. The dependent variable column is labeled "y" and the column indicating the treatment group is labeled "gp." The *SPSS* menu path is as follows:

Analyze → Compare Means → One-Way ANOVA →
Dependent list:y → Factor → gp → Post Hoc → Tukey →
Continue → OK

Output:

ANOVA Y

Oneway

	Sum of Squares	df	Mean Square	F	Sig.
Between Groups	986.667	2	493.333	3.767	.036
Within Groups	3536.000	27	130.963		
Total	4522.667	29			

Post Hoc Tests

Multiple Comparisons
Dependent Variable: Y
Tukey HSD

					95% Confidence Interval	
(I) GP	(J) GP	Mean Difference (I-J)	Std. Error	Sig.	Lower Bound	Upper Bound
1.00	2.00	−14.0000	5.11787	.028	−26.6893	−1.3107
	3.00	−6.0000	5.11787	.479	−18.6893	6.6893
2.00	1.00	14.0000	5.11787	.028	1.3107	26.6893
	3.00	8.0000	5.11787	.279	−4.6893	20.6893
3.00	1.00	6.0000	5.11787	.479	−6.6893	18.6893
	2.00	−8.0000	5.11787	.279	−20.6893	4.6893

*The mean difference is significant at the .05 level.

Homogeneous Subsets

Y
Tukey HSD

		Subset for alpha = .05	
GP	N	1	2
1.00	10	30.0000	
3.00	10	36.0000	36.0000
2.00	10		44.0000
Sig.		.479	.279

Means for groups in homogeneous subsets are displayed.
a Uses Harmonic Mean Sample Size = 10.000.

Note that *SPSS* provides much redundant output and that the section called "Post Hoc Tests" actually presents the simultaneous confidence intervals and the probability values associated with the tests, but not the obtained test statistics. Also, no individual value plot is provided and the *p*-values are (incorrectly) labeled "Sig." Ignore the Homogeneous Subsets table.

Both *Minitab* and *SPSS* use the label "Tukey" rather than "Tukey–Kramer" but an inspection of the routines reveals that both actually use Tukey–Kramer. The Tukey procedure is for the case of equal sample sizes only; Kramer (1956) demonstrated, however, that the Tukey procedure can be satisfactorily modified to handle unequal sample sizes by substituting the harmonic mean of the two sample sizes (involved in each pairwise comparison) in place of n (the common sample size) required in the conventional Tukey formula. Although it may not be obvious, this approach is built into the expression described above for the Tukey–Kramer method. Because the Tukey–Kramer reduces to the conventional Tukey in the case of equal sample sizes, there is no reason to learn (or program) two different formulas.

Standardized Effect Sizes and $\hat{\eta}^2$

Recall that the standardized effect size is estimated in a two-group study by dividing the difference between the sample means by the pooled within-group standard deviation. A similar approach applies when there are more than two groups. A standardized effect size may be computed for each pair of mean differences using $\frac{\bar{Y}_i - \bar{Y}_j}{s_W} = g$, where s_W is the square root of the within-groups mean square. *Minitab* provides s_W as a part of the ANOVA output; it appears immediately in the ANOVA summary table and is labeled as "*S*." For example, the standardized effect size for groups two and three in Example 3.2 (above) is 8/11.44 = .70.

The measure of association between the independent and dependent variables is computed using results in the ANOVA summary table. The ratio $\frac{SS_B}{SS_T} = \hat{\eta}^2$ describes the proportion of the sample variation on the dependent variable that is explained by the independent variable. *Minitab* output labels this proportion as R^2 rather than $\hat{\eta}^2$; it is displayed under the ANOVA summary table. It can be seen in the output presented above for Example 3.2 that $SS_B/SS_T = 987/4523 = .218$. About 22% of

the total sample variation on the dependent variable (achievement) is explained by the treatments (training types).

3.3 ONE-FACTOR RANDOMIZED BLOCK DESIGN AND ANALYSIS

The one-factor randomized block design, like the one-factor randomized group design, is employed to test the hypothesis that J population means are equal. Why a second (and more complex) design to test the same hypothesis? There is a major advantage: greater power (or the flip side, smaller sample size required when power is held constant). The randomized group and randomized block approaches differ with respect to both how the design is set up and the method of analysis. Recall that the randomized group design involves simply randomly assigning the available subjects to one of J-treatment conditions. The randomized block design involves three steps. First, information on subjects that can be employed to form blocks is collected. A block is simply a subgroup of subjects who are relatively homogeneous with respect to some variable (called the *blocking variable*) believed to be correlated with the dependent variable. Blocking variables may be continuous quantitative variables (such as age, weight, achievement, experience, and pretest score), ordered categories (such as disease stages), or qualitative (nominal) variables (such as sex, occupation type, classroom, litter, and observer).

Once the blocking-variable information is available for all subjects, the second step can be carried out. This step involves the formation of blocks. If the blocking data are in the form of continuous scores, the blocks are formed by simply (1) ordering the subjects according to their score on the blocking variable and (2) assigning the highest scoring J subjects to block 1, the next highest group of J subjects to block 2, and so on. For example, an experimenter may believe that anxiety as measured by the galvanic skin response (GSR) is related to performance of subjects on a complex motor skills task. Suppose that 25 subjects are available and that the purpose of the experiment is to compare the effectiveness of five different methods of training employees to perform a complex task. First, the 25 subjects are measured on the blocking-variable GSR. Then the subjects are ordered from highest to lowest according to their GSR scores. Because $J = 5$ (i.e., there are five treatments), the subjects associated with the top five GSR scores constitute the first block. The subjects associated with the next highest five scores constitute the second block and so on, so that the subjects associated with the lowest five scores will constitute the last block (block 5).

The third step is random assignment. After the blocks have been formed, the subjects within each block are randomly assigned to treatments. This randomization is carried out independently for each block. For the training example, the five subjects in the first block are randomly assigned to the five treatments. Then the five subjects in block 2 are randomly assigned to the five treatments, and so on. A comparison of the randomization procedures for the randomized group and randomized block designs can be seen in Figure 3.2. In this example, there are five treatments and five

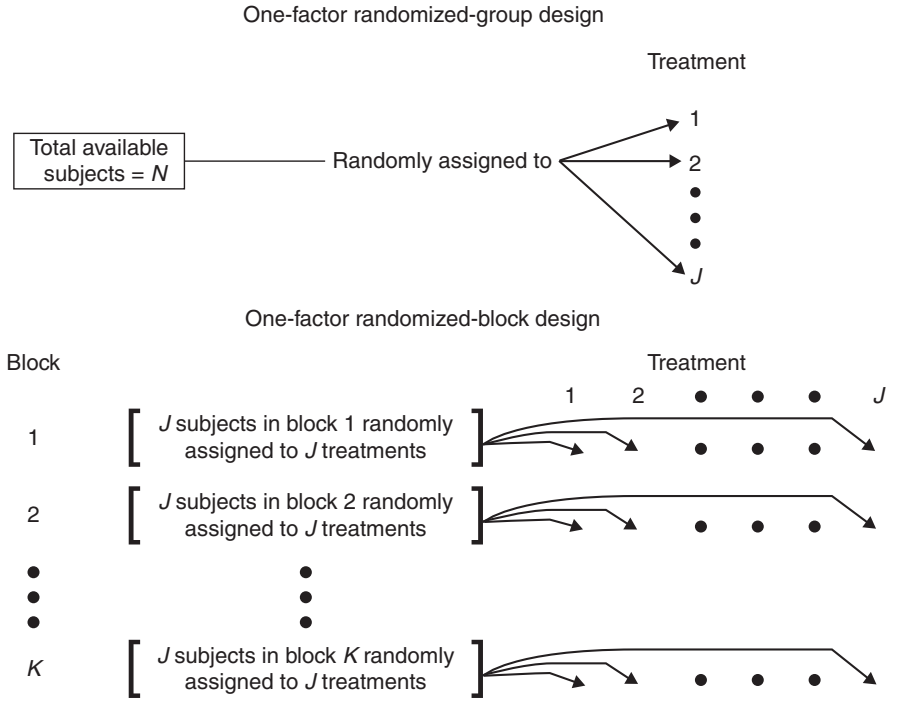


Figure 3.2 Comparison of randomization procedures for one-factor randomized-group and randomized-block designs.

blocks. It is not necessary, however, that the number of treatments and the number of blocks be the same. If 15 subjects are available and five treatments are to be studied, there would be three blocks. Note there is one subject for every combination of block and treatment.

Model

The statistical model for the analysis of the randomized block design can be written as follows:

$$Y_{i,j} = \mu + \kappa_i + \alpha_j + \varepsilon_{ij},$$

where

$Y_{i,j}$ is the dependent variable score for the subject falling in block i and treatment j;
μ is the overall level for the whole experiment;
κ_i is the effect of block i;
α_j is the effect of treatment j; and
ε_{ij} *is* the error.

Table 3.2 Summary for the Randomized Block ANOVA

Source	SS	*df*	MS	*F*
Treatments	SS_{TR}	$df_{TR} = J - 1$	MS_{TR}	MS_{TR}/MS_{Error}
Blocks	SS_{BK}	$df_{BK} = K - 1$		
Error	SS_{Error}	$df_{Error} = (J - 1)(K - 1)$	MS_{Error}	
Total	SS_T	$df_T = N - 1$		

Analysis Procedure

The analysis of variance for the randomized block design involves first computing sum of squares for treatments, blocks, and error. The treatment sum of squares is computed in the same way as was described for computing the between group sum of squares for the randomized group design. The block sum of squares is also computed using the same approach as for the between groups sum of squares, but the blocks play the role of the "groups." The sum of the treatment sum of squares and the block sum of squares is subtracted from the total sum of squares to compute the error sum of squares (which is actually the treatment by block interaction sum of squares). The form of the analysis of variance summary is shown in Table 3.2. The null hypothesis is H_0: $\mu_1 = \mu_2 = \cdots = \mu_J$.

Note that there is no mean square or F-ratio for blocks. Although some software for this design includes an F-test for blocks, there is usually no interest in a test on blocks. After all, if the experimenter has selected a reasonable blocking variable it will be known *a priori* that there are differences between block means. The F-ratio for treatments is of interest. It has $J - 1$ and $(J - 1)(K - 1)$ degrees of freedom.

Multiple Comparison Procedures

Fisher–Hayter multiple comparison tests are based on the following formula:

$$\frac{\bar{Y}_i - \bar{Y}_j}{\sqrt{\dfrac{MS_{Error}}{n}}} = q_{FH},$$

where n is the common sample size associated with each treatment group, which is equal to the number of blocks. The critical value $q_{J-1,(J-1)(K-1)}$ is determined from the table of the Studentized range distribution by entering the column that is headed with the value equal to $J - 1$ and then finding the error of degrees of freedom $= (J - 1)(K - 1)$.

Tukey 95% simultaneous confidence intervals for pairwise comparisons may be constructed using

$$\left(\bar{Y}_i - \bar{Y}_j\right) \pm q_{J,(J-1)(K-1)}\sqrt{\frac{MS_{Error}}{n}},$$

where $q_{J,\ (J-1)(K-1)}$ is based on the 5% critical value of the Studentized range distribution.

Example 3.3 Three treatment methods and four blocks are involved in the example design. The treatments are types of therapy, the blocking variable is a pretreatment measure of anxiety, and the dependent variable (Y) is a measure of depression. The pattern of block means on Y suggests a high correlation between the blocking variable and the dependent variable; the 12 subjects were ordered on anxiety (lowest anxiety scores in block 1 and highest in block 4) and the dependent variable scores (depression) follow the order of the blocks. If there had been no correlation between the blocking and dependent variables, the block means on Y would be similar.

	Treatment			
Block	I	II	III	Block means
1	7	5	8	6.67
2	10	11	13	11.33
3	14	13	17	14.67
4	20	21	24	21.67
Treatment means →	$\bar{Y}_{\mathrm{I}} = 12.75$	$\bar{Y}_{\mathrm{II}} = 12.5$	$\bar{Y}_{\mathrm{III}} = 15.5$	

Analysis of Variance

The 12 rows of input data entered in the *Minitab* worksheet appear below:

Row	Treatment	Block	Depression
1	1	1	7
2	2	1	5
3	3	1	8
4	1	2	10
5	2	2	11
6	3	2	13
7	1	3	14
8	2	3	13
9	3	3	17
10	1	4	20
11	2	4	21
12	3	4	24

The command line editor commands for the one-factor randomized block design are

```
MTB > Twoway 'Depression' 'Block' 'Treatment';
SUBC> Means 'Block' 'Treatment'.
```

The ANOVA output is

```
Two-way ANOVA: Depression versus Block, Treatment

Source       DF        SS        MS        F        P
Block         3   358.250   119.417   159.22   0.000
Treatment     2    22.167    11.083    14.78   0.005
Error         6     4.500     0.750
Total        11   384.917

S = 0.8660   R-Sq = 98.83%   R-Sq(adj) = 97.86%
```

Note that the output is labeled as "Two-way ANOVA" even though the design is a one-factor randomized block. Ignore this label as well as the "S," "R-Sq," and "R-Sq (adj)" output. These values are not appropriate for the computation of effect sizes with this design.

Standardized Effect Sizes and $\hat{\eta}^2$

The standardized effect size statistic g and the measure of association $\hat{\eta}^2$ may be computed for the randomized block designs using

$$\frac{\bar{Y}_i - \bar{Y}_j}{\sqrt{\dfrac{SS_{BK} + SS_{Error}}{N - J}}} = g \quad \text{and} \quad \frac{SS_{Treatment}}{SS_{Total}} = \hat{\eta}^2.$$

Comparison of Randomized Group and Randomized Block ANOVA

It was pointed out that the main reason for preferring the randomized block design to the randomized group design is that the size of the error term in the analysis of variance is likely to be smaller. It may be useful to demonstrate this notion by performing a randomized group ANOVA on the same data that were analyzed (above) using a randomized block design ANOVA. The results of the randomized groups ANOVA are shown below.

```
One-way ANOVA:

Source             DF     SS    MS     F      P
Between groups      2   22.2  11.1  0.27  0.766
Error               9  362.8  40.3
Total              11  384.9
```

It can be seen that the between-group sum of squares and mean square in the randomized group analysis are identical to the treatment sum of squares and mean square in the randomized block analysis. This must be true because the dependent

variable means in the randomized group analysis are identical to the means in the randomized block analysis. But there is a dramatic difference between analyses in terms of F- and p-values. Note that the randomized group analysis yields an F-value of only .27 whereas the randomized block analysis yields a very large F-value of 14.78. The explanation for the difference in the size of F is that the error mean square is much smaller in the randomized block analysis (.75 vs. 40.3). A close inspection of the two summary tables reveals that the sum of squares within groups in the randomized group analysis is the sum of the block sum of squares and the error sum of squares in the randomized block analysis.

If the data within each treatment group are inspected, it can easily be seen that most of the variation is associated with the blocks. This immediately tells us that the randomized group model is a poor representation of the data. Recall that this model says that the variation within treatment conditions is simply unexplainable random fluctuation. But in this example, there clearly is an explanation for most of the variation (viz., blocks). Because we have knowledge of the blocks, the model should incorporate a term to acknowledge them. Correspondingly, the ANOVA should acknowledge blocks so that the error sum of squares is not contaminated by nonrandom variation explained by blocks. This is exactly what the randomized block ANOVA does.

The degree of advantage of the randomized block design over the randomized group design depends mostly on (1) the size of the correlation between the blocking variable and the dependent variable, and (2) the homogeneity of the blocks. The higher the correlation and the more homogeneous the blocks are on the blocking variable, the greater the power advantage of blocking. The next design takes the notion of homogeneous blocks to a logical extreme.

3.4 ONE-FACTOR REPEATED MEASUREMENT DESIGN AND ANALYSIS

Because the power of the randomized block analysis increases as the subjects within blocks increase in similarity, it is logical to attempt to maximize within-block homogeneity by treating a single subject as a block. That is, we can conceptualize each subject in an experiment as a block and then expose each subject to all treatments in random order. If this is done the resulting design is called a repeated measurement design rather than a randomized block design, even though the two designs have much in common.

Recall that there is one subject for each block and treatment combination in the randomized block design. Hence, the total number of subjects (i.e., N) in the randomized block design is J times K; each subject is exposed to only one treatment. Both the randomized group and randomized block designs are classified as between subject designs because each treatment mean is based on a different group of subjects. The repeated measurement design has n subjects and J treatments; although the total number of subjects $= n$, the total number of observations is $nJ = N$.

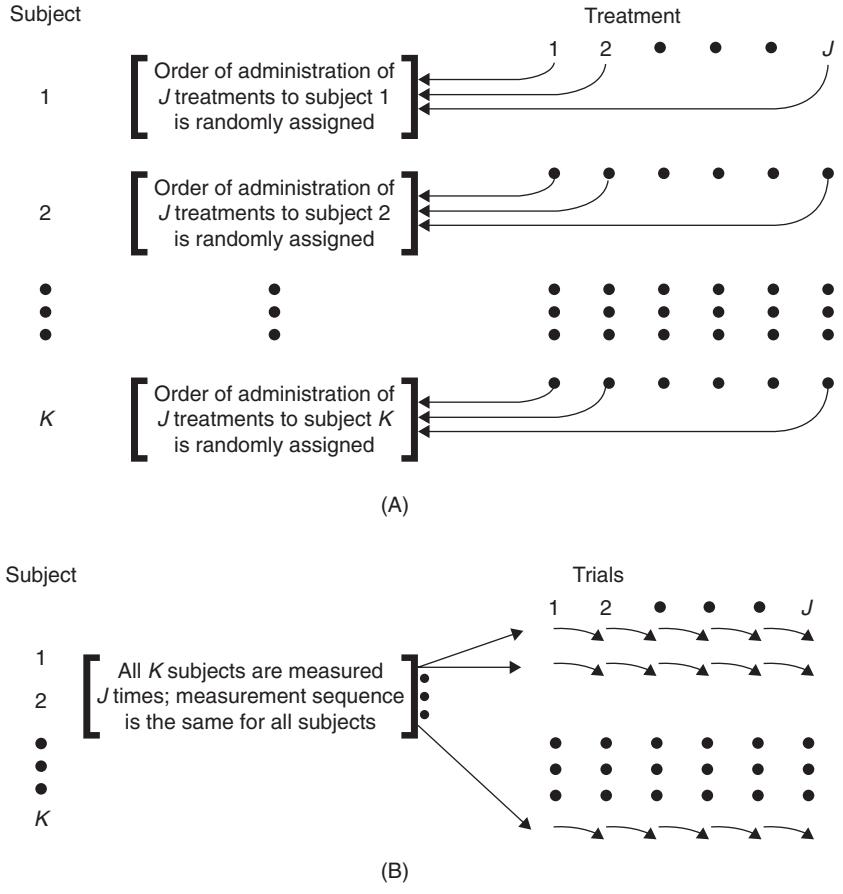


Figure 3.3 Two versions of one-factor repeated measurement design.

Because each subject is exposed to all treatment conditions, this design is classified as a within-subject design; each treatment mean is based on the same group of subjects. This design is sometimes referred to as a "subjects by treatments" design; the randomization procedure for this design version can be seen in Panel A of Figure 3.3. This is a true experimental design that involves both randomization and experimental manipulation of treatment conditions.

It might seem that perfect homogeneity within "blocks" is achieved by using a subject as a block. After all, there is absolutely no heterogeneity of different subjects within each "block." But this appraisal is a little optimistic. We must acknowledge that once a subject is exposed to a treatment, she is not necessarily the same as she was before the exposure. Having been exposed to one or more treatments has implications for both the homogeneity of the subject who defines the "block" and the generality of the results.

Model

The statistical model for the analysis of the repeated measurement design can be written as follows:

$$Y_{i,j} = \mu + \gamma_i + \alpha_j + \varepsilon_{ij},$$

where

$Y_{i,j}$ is the dependent variable score for subject i falling in treatment j;
μ is the overall level for the whole experiment;
γ_i is the effect of subject i;
α_j is the effect of treatment j; and
ε_{ij} *is* the error.

Analysis Procedure

The analysis of variance for the repeated measurement design of the type described above involves first computing sum of squares for treatments, subjects, and error. The treatment sum of squares is computed in the same way as was described for computing the between-group sum of squares for the randomized group design. The subject sum of squares is also computed using the same approach as for the between-group sum of squares, but the subjects play the role of the "groups." The sum of the treatment sum of squares and the subject sum of squares is subtracted from the total sum of squares to compute the error sum of squares (which is actually the treatment by subject interaction sum of squares). The form of the analysis of variance summary is shown in Table 3.3. The null hypothesis is H_0: $\mu_1 = \mu_2 = \cdots = \mu_J$.

Note that there is no mean square or F-ratio for subjects. Although some software for this design includes an F for subjects, it is hardly inspiring to discover that there are individual differences. The F-ratio for treatments is of interest. It has $J - 1$ and $(J - 1)(n - 1)$ degrees of freedom.

The form of the ANOVA shown here (sometimes called the mixed-model solution) is not usually appropriate for other versions of repeated measurement design, such as the one shown in Panel B of Figure 3.3. The latter design is sometimes referred to as a "subjects by trials" design because behavior is simply observed over a collection of equally spaced occasions (trials); different treatment conditions are not introduced. Although this version is also a one-factor repeated measurement design, it is not a true experiment because it involves neither randomization nor experimental manipulation of treatments. Designs of this type are not pursued here.

Table 3.3 Summary for the Repeated Measurement ANOVA

Source	SS	*df*	MS	*F*
Treatments	SS_{TR}	$df_{TR} = J - 1$	MS_{TR}	MS_{TR}/MS_{Error}
Subjects	$SS_{Subjects}$	$df_{Subjects} = n - 1$		
Error	SS_{Error}	$df_{Error} = (J - 1)(n - 1)$	MS_{Error}	
Total	SS_{T}	$df_{T} = N - 1$		

Multiple Comparison Procedures

Fisher–Hayter multiple comparison tests are based on the following formula:

$$\frac{\bar{Y}_i - \bar{Y}_j}{\sqrt{\dfrac{\text{MS}_{\text{Error}}}{n}}} = q_{\text{FH}},$$

where n is the number of subjects. The critical value $q_{J\,-\,1,\,(J\,-\,1)(n\,-\,1)}$ is determined from the table of the Studentized range distribution by entering the column that is headed with the value equal to $J - 1$ and then finding the error degrees of freedom $= (J - 1)(n - 1)$.

Tukey 95% simultaneous confidence intervals for pairwise comparisons may be constructed using

$$\left(\bar{Y}_i - \bar{Y}_j\right) \pm q_{J,(J-1)(K-1)}\sqrt{\frac{\text{MS}_{\text{Error}}}{n}},$$

where $q_{J,\ (J\,-\,1)(n\,-\,1)}$ is based on the 5% critical value of the Studentized range distribution.

Standardized Effect Sizes and $\hat{\eta}^2$

The standardized effect size statistic g and the measure of association $\hat{\eta}^2$ may be computed for the repeated measurement design using

$$\frac{\bar{Y}_i - \bar{Y}_j}{\sqrt{\dfrac{\text{SS}_{\text{Subjects}} + \text{SS}_{\text{Error}}}{N - J}}} = g \quad \text{and} \quad \frac{\text{SS}_{\text{Treatment}}}{\text{SS}_{\text{Total}}} = \hat{\eta}^2.$$

Effects of Design Type on Results

The subjects by treatments version of the repeated measurement design is frequently employed instead of a randomized group or randomized block design because fewer subjects can be employed. This is an important consideration in many research projects, but it should be remembered that this design does not answer the same question that is answered by the randomized group and randomized block designs. Results of between-subject designs (i.e., randomized group and randomized block designs) generalize to a population of subjects who have been exposed to only one treatment. Results of within-subject designs generalize to a population containing a large proportion of subjects who have been exposed to more than one treatment. For example, suppose an experimenter has decided to investigate the effects of four different incentive programs on employee output of an industrial product. If this problem is investigated with a randomized block or randomized group design, each group will be exposed to only one type of incentive.

With the repeated measurement design, each subject will be exposed to all types of incentive. The results of the randomized group and randomized block designs generalize to subjects who have been exposed to only one of the incentive programs. But with the repeated measurement design, every subject has been exposed to all the different incentive conditions. Consequently, the results generalize to a population containing subjects so treated. If a repeated measurement design has been used and incentive number 2 yields a significantly higher mean that the other three, it can be concluded that this incentive type is superior for a population composed of a high proportion of subjects who have been previously exposed to other types of incentives.

Although it is possible that treatment effect estimates from a repeated measurement experiment will be identical to those from a randomized group or randomized block experiment, this is unknown *a priori*. The only way to know whether design type has an effect on results is to collect data using more than one type of design. Examples of unexpected effects of design type can be found in the literature; see Grice and Hunter (1964) for a classic example in psychology.

3.5 SUMMARY

Most one-factor experiments are based on the randomized group design, the randomized block design, or the repeated measurement design. Both randomized group and randomized block designs are between-subject designs; they provide treatment effect estimates that generalize to a population of subjects not previously exposed to the other treatments. The randomized group design is simpler to set up than is the randomized block design, but the latter leads to more powerful tests. The repeated measurement design leads to even more powerful tests than does the randomized block design, but the results generalize to a population of subjects that includes those who have been previously exposed to other treatments in the experiment.

PART II

Essentials of Regression Analysis

CHAPTER 4

Simple Linear Regression

4.1 INTRODUCTION

Regression analysis comes in many different forms and is used for several different purposes in the context of both experimental and nonexperimental research. It may be used at one or more of the three stages of scientific investigations: description, prediction, and control. The most elementary form is known as simple linear regression; it is often used to describe the linear relationship between two quantitative variables and to provide a method of predicting scores on one variable from scores on the other. Common examples include the prediction of job performance from measures of mechanical aptitude, the prediction of degree of clinical depression from the frequency or duration of past traumatic events, and the prediction of LDL cholesterol level from a measure of social stress. These applications may suggest that regression has little in common with ANOVA, but several aspects of these two approaches overlap both conceptually and computationally.

4.2 COMPARISON OF SIMPLE REGRESSION AND ANOVA

There are many similarities and differences between ANOVA and regression analysis. They are similar in that both of them (1) are applied to data collected using either experimental or nonexperimental designs, (2) provide a test for association between independent and dependent variables, and (3) lead to measures describing the relationship between independent and dependent variables. They differ with respect to (1) the type of independent variable to which they apply, (2) the statistical model, (3) the type of descriptive results they provide, (4) the sources of variation considered in partitioning the total sum of squares, and (5) the specific hypotheses tested. Details regarding these and other similarities and differences are explained next.

The Analysis of Covariance and Alternatives: Statistical Methods for Experiments, Quasi-Experiments, and Single-Case Studies, Second Edition. Bradley E. Huitema.

Nature of the Independent Variable

Although both types of analysis focus on the relationship between two variables, one-factor ANOVA usually involves an independent variable that is qualitative in nature (i.e., different types of treatment) and one variable (the dependent variable) that is quantitative in nature. Regression analysis is usually used with two quantitative variables that are approximately continuously scaled.

The two variables in regression analysis are sometimes called predictor (X) and criterion variables (Y) rather than independent and dependent variables if the data are collected using correlational or classic regression designs. Although it is widely believed that simple regression analysis is appropriate only for nonexperimental research, this is not true. Regression analysis is appropriate for data from both nonexperimental and experimental designs of certain forms. This is also true for ANOVA.

Note that in Table 4.1 Panels A and B classify research designs as either nonexperimental or experimental. Subsumed under each of these major design types are more specific design variants. Correlational, classic regression, and observational design variants are subsumed under the nonexperimental category. Two variants of randomized group design are subsumed under the experimental category.

The characteristics that are associated with each specific design variant include (a) the major question of interest, (b) the type of analysis, (c) whether it is necessary to specify one of the two variables as independent and the other as dependent, (d) whether the levels of the independent variable are manipulated, (e) the method of sampling used to obtain the subjects, (f) the assumptions underlying the typical method of analysis, and (g) the populations(s) to which the results are generalized.

Examine the second row (i.e., row *b*) in these two panels. Observe that both ANOVA and regression analysis can be found under each major design category. Next, focus on the first row of each panel. Note that strong causal inferences are justified using both regression analysis and ANOVA in Panel II (experimental designs) but not in Panel I (nonexperimental designs). The message here is that it is *not* the type of statistical analysis that justifies causal interpretations; rather, it is the design.

Although design trumps analysis when attempting to justify causal interpretations, there are distinctions between regression and ANOVA that are important to understand regardless of the design type. Conceptual and computational aspects that differentiate these two analyses are considered next.

Model

A formal way of distinguishing between ANOVA and regression analysis is to compare the associated statistical models. Recall that the one-factor ANOVA model is written as

$$Y_{ij} = \mu + \alpha_j + \varepsilon_{ij},$$

where

Y_{ij} is the dependent variable score for the ith subject in the jth treatment;
μ is the overall population mean (i.e., the mean of the individual population means);
α_j is the effect of treatment j; and
ε_{ij} is the error component associated with the ith subject in treatment j.

This model makes it clear that all variability on Y that is not accounted for by treatment effects is viewed as error. That is, the total variation is conceptualized as being attributable to two components: treatment effects (i.e., the α_j) and error (variation within treatments). Although this is not the only way to conceptualize the total variation in a multiple group design, it is often a useful approach in the typical experiment where the independent variable is qualitative (i.e., levels differing in type rather than amount).

But many studies do not involve the application of different treatments. Often the main research interest is to predict values of some quantitative variable Y from scores on some predictor variable X. In this case, it is typical that the X variable is quantitative in nature rather than qualitative. This situation often leads to the conceptualization of the data analytic problem in terms of the simple linear regression model. This model can be written as

$$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i,$$

where

Y_i is the dependent variable score for individual i;
β_0 is the population Y intercept;
β_1 is the population slope of Y on X;
X_i is the predictor score for individual i; and
ε_i is the population error component associated with individual i.

The simple regression model, unlike the ANOVA model, does not include information concerning treatment group membership; this is indicated by the absence of the subscript j in the various terms of the regression model. The ANOVA and regression models both contain terms to provide explanations of Y, but the explanations are different.

All variation on Y is explained by treatment effects and error in the ANOVA model whereas the regression model explains all variation as a linear function of predictor X and error. If there are no treatment effects in the ANOVA model it is reduced to the following form: $Y_{ij} = \mu + \varepsilon_{ij}$. Similarly, in the case of regression, if there is no linear relationship between Y and X the model is reduced to $Y_i = \beta_0 + \varepsilon_i$. Even though these two reduced models look slightly different, they are actually same because β_0 in the regression model is equal to μ in the ANOVA model when the regression slope is equal to zero. The purpose of inferential tests in ANOVA and regression analysis is to help the researcher decide whether there are treatment effects (in the case of ANOVA) or a nonzero regression slope (in the case of regression analysis).

Table 4.1 Distinguishing Characteristics Associated with Variants of (I) Nonexperimental and (II) Experimental Research Designs and Associated Analyses

	I. Nonexperimental Designs		
Distinguishing Characteristics	Correlational	Classic Regression	Observational
(a) Major question asked	What is the direction and degree of linear relationship between X and Y?	What is the regression function relating Y to X?	Are differences between means associated with observed group categories?
(b) Typical analysis	Correlation	Regression	ANOVA
(c) Necessary to specify one variable as independent and one as dependent?	No	Yes	Yes
(d) Independent variable manipulated?	No	No	No
(e) Sampling method	Simple random from a defined population	Simple random from fixed quantitative levels of X	Simple random from fixed qualitative levels of X
(f) Assumptions underlying inference	Bivariate normality	Y errors nid* with constant variance	Y errors nid* with constant variance
(g) Generalization	Dictated by sampling and design	Dictated by sampling and design	Dictated by sampling and design

Distinguishing Characteristics	II. Experimental Designs	
	Randomized experiment with quantitative independent variable	Randomized experiment with qualitative independent variable
(a) Major question asked	What is the average amount of increase (or decrease) on Y caused by one unit increase on X?	Do different treatments cause differences between population means?
(b) Typical analysis	Regression	ANOVA
(c) Necessary to specify one variable as independent and one as dependent	Yes	Yes
(d) Independent variable manipulated	Yes	Yes
(e) Sampling method	Ideally, random selection from population of interest and then random assignment to quantitative levels of X	Ideally, random selection from population of interest and then random assignment to qualitative levels of X
(f) Assumptions underlying inference	Y errors nid* with constant variance	Y errors nid* with constant variance
(g) Generalization	Dictated by sampling and design	Dictated by sampling and design

*nid = Normally and independently distributed.

4.3 REGRESSION ESTIMATION, INFERENCE, AND INTERPRETATION

Parameter Estimation

The parameters of the regression model are usually estimated from sample data using the method known as ordinary least-squares (OLS). The least-squares estimates of the intercept and slope are denoted as b_0 and b_1, respectively. They are computed as follows:

$$b_0 = \bar{Y} - b_1\bar{X}, \text{ and}$$

$$b_1 = \frac{\sum_{i=1}^{N} xy}{\sum_{i=1}^{N} x_i^2},$$

where $x_i = (X_i - \bar{X})$ and $y_i = (Y_i - \bar{Y})$.

Once these estimates are computed, they can be used to specify the "fitted equation."

Fitted Equation

The fitted equation is the most basic descriptive result of a regression study just as the sample means are the most basic descriptive results associated with ANOVA. The fitted equation (also called the prediction equation) is usually written as follows:

$$\hat{Y}_i = b_0 + b_1 X_i,$$

where

- $\hat{Y}_i$ is the predicted score for sample individual i (also conceptualized as the estimated expected value on Y associated with score X_i);
- b_0 is the sample intercept (i.e., the estimated elevation on Y when X is set equal to zero);
- b_1 is the sample slope (i.e., the average increase on Y associated with a one unit increase on X); and
- X_i is the predictor score for sample individual i.

Once the estimates included in the prediction equation are computed, the regression line should be presented graphically along with the data. The graphic representation of the data including the regression line is an essential aspect of a complete regression analysis; it should not be omitted.

Note that the statistical model differs from the fitted equation in three respects: (1) the model is written using Greek symbols (to define population parameters) whereas the fitted equation uses Roman symbols (to define statistics) because the latter are measures based on sample data, (2) the fitted equation has no error term, and (3) the

Table 4.2 Form of the ANOVAR Summary Table for Simple Linear Regression

Source	SS	*df*	MS	*F*
Regression	SS_{Reg}	$df_{Reg} = 1$	MS_{Reg}	MS_{Reg}/MS_{Res}
Residual	SS_{Res}	$df_{Res} = N - 2$	MS_{Res}	
Total	SS_T	$df_T = N - 1$		

terms in the statistical model explain the *observed Y* scores in the population whereas the terms in the fitted equation define (exactly) the sample *predicted* scores $\hat{Y}_i$.

Analysis of Variance of Regression (ANOVAR)

Just as the ANOVA *F*-test is used to evaluate whether the treatment effects terms (i.e., the α_j) in the ANOVA model are necessary, the ANOVAR *F*-test is used to evaluate whether the slope term (i.e., β_1) in the regression model is necessary to explain the data. The form of the ANOVAR summary table is shown in Table 4.2.

The defining formulas for the various sum of squares in this table are as follows:

$$\text{Regression sum of squares} = SS_{Reg} = \sum_{i=1}^{N} \left(\hat{Y}_i - \bar{Y}\right)^2,$$

$$\text{Residual sum of squares} = SS_{Res} = \sum_{i=1}^{N} \left(Y_i - \hat{Y}_i\right)^2, \quad \text{and}$$

$$\text{Total sum of squares} = SS_T = \sum_{i=1}^{N} \left(Y_i - \bar{Y}\right)^2.$$

These sums of squares may require a little explanation. It can be seen in the ANOVAR summary table that the total variation on the dependent variable (i.e., the total sum of squares) is partitioned into (1) variation accounted for with the use of *X* and a linear prediction rule (i.e., the regression sum of squares), and (2) variation not accounted for with the use of *X* and a linear rule (i.e., the residual sum of squares). The reason for this approach to partitioning is fairly straightforward.

Because $\hat{Y}$ is the value predicted from the fitted linear equation, all of the $\hat{Y}'$s associated with the *X* values in a sample will fall on a straight line. This line is often called the *regression line* associated with the regression of *Y* on *X*. The sum of the squared deviations of the $\hat{Y}'$s around the mean $\bar{Y}$ is the *regression sum of squares*; i.e., $SS_{Regression} = \sum_{i=1}^{N}(\hat{Y}_i - \bar{Y})^2$.

Usually the actual score on *Y* associated with a specific subject (say, subject *i*) does not fall exactly on the regression line and therefore there is some prediction error for this subject. This prediction error is called the *residual* and it is denoted as e_i. That is, $(Y_i - \hat{Y}_i) = e_i$. The residual is the estimate of the error ε_i in the regression model.

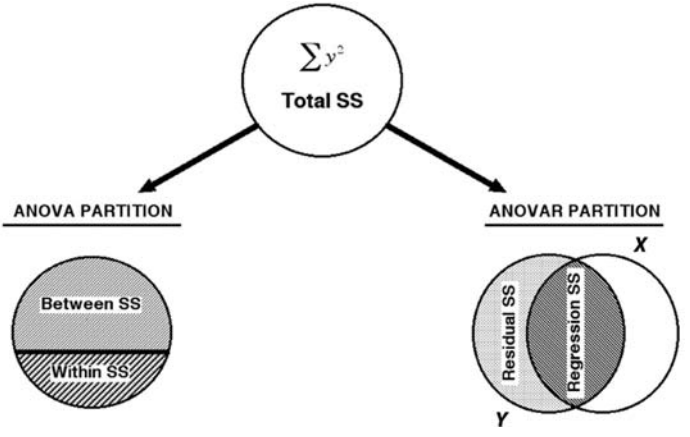


Figure 4.1 Partitioning of the total sum of squares under ANOVA and ANOVAR.

The sum of all of the squared residuals in the sample is $\sum_{i=1}^{N} e_i^2 = \sum_{i=1}^{N} (Y_i - \hat{Y}_i)^2$; this is called the residual sum of squares (SS_{Res}).

Figure 4.1 illustrates the partitioning of the total sum of squares under the two models. Note that the ANOVAR partitioning differs from the approach taken in one-factor ANOVA in that the total sum of squares is partitioned into regression and residual components rather than between-group and within-group components. This difference in partitioning is reflected as a corresponding difference in the way the F-ratio is defined. The ratio of the mean square regression over the mean square residual yields the obtained ANOVAR F-statistic. This statistic is compared with the critical value of F based on 1 and $N - 2$ degrees of freedom. If the obtained value of F is equal to or greater than the critical value, the null hypothesis (i.e., H_0: $\beta_1 = 0$) is rejected and it is concluded that sufficient evidence is available to conclude that the population slope is not zero (i.e., $\beta_1 \neq 0$). That is, the sample b_1 provides convincing evidence that there is some degree of linear relationship between Y and X in the population. Of course, computer software is normally used to carry out the analysis; it will provide the exact probability associated with the obtained value of F under the null hypothesis. The decision rule based on a p-value is reject H_0: $\beta_1 = 0$ if p is equal to or less than α, otherwise retain.

The logic underlying the ANOVAR F-test is the same as in the case of conventional ANOVA. This can be confirmed by inspecting the expected values of the mean squares presented in Table 4.3. Recall that in the case of one-factor ANOVA the between-group and within-group mean squares have the same expected value when the null hypothesis (i.e., H_0: $\mu_1 = \mu_2 = \cdots = \mu_J$) is true. Similarly, in the case of regression analysis, the regression and residual mean squares have the same expected value if the null hypothesis (i.e., $\beta_1 = 0$) is true. Hence, small values of the obtained F are likely in this situation. Alternatively, if the null hypothesis is false in the case of either

Table 4.3 Comparison of Expected Mean Squares Associated with (I) ANOVA and (II) ANOVAR *F*-Tests Under True and False Null Hypotheses

(I) ANOVA	
Null hypothesis true $(\mu_1 = \mu_2 = \cdots = \mu_J)$	Null hypothesis false $(\mu_j \neq \mu$ for at least one $j)$
$E(\text{MS}_\text{B}) = \sigma^2$	$E(\text{MS}_\text{B}) = \sigma^2 + \dfrac{\sum_{j=1}^{J} n_j(\mu_j - \mu)^2}{J-1}$
$E(\text{MS}_\text{W}) = \sigma^2$	$E(\text{MS}_\text{W}) = \sigma^2$
(II) ANOVAR	
Null hypothesis true $(\beta_1 = 0)$	Null hypothesis false $(\beta_1 \neq 0)$
$E(\text{MS}_\text{Regression}) = \sigma^2$	$E(\text{MS}_\text{Regression}) = \sigma^2 + \beta_1^2 \sum_{i=1}^{N} x_i^2$
$E(\text{MS}_\text{Residual}) = \sigma^2$	$E(\text{MS}_\text{Residual}) = \sigma^2$

ANOVA or ANOVAR, the expected value for the numerator mean square is larger than the expected value for the denominator mean square. Note that in the table the expected value for the regression mean square is directly inflated by nonzero values of β_1 whereas the expected value for the residual mean square is unaffected by β_1. This difference in expected values implies large values of F when the null hypothesis regarding the population slope is false.

Example of Simple Linear Regression Analysis

Consider the data in Table 4.4. The scores are measures on a biology aptitude measure (X) and a measure of achievement in biology (Y) based on a sample from a population of undergraduate students.

The slope and intercept estimates are

$$\frac{\sum_{i=1}^{30} xy}{\sum_{i=1}^{30} x^2} = \frac{3022}{5826.67} = .519 = b_1, \text{ and}$$

$$\bar{Y} - b_1\bar{X} = [35 - .519(49.33)] = 9.413 = b_0.$$

Hence the fitted equation is: $\hat{Y}_i = 9.413 + .519(X_i)$. This fitted equation is illustrated in Figure 4.2.

Table 4.4 Scores on Aptitude (X) and Achievement (Y) Measures; $N = 30$

X	Y
29	15
49	19
48	21
35	27
53	35
47	39
46	23
74	38
72	33
67	50
22	20
24	34
49	28
46	35
52	42
43	44
64	46
61	47
55	40
54	54
33	14
45	20
35	30
39	32
36	34
48	42
63	40
57	38
56	54
78	56

The sum of squares required for the ANOVAR F-test are

$$\mathrm{SS}_{\mathrm{Reg}} = \sum_{i=1}^{N} \left(\hat{Y}_i - \bar{Y}\right)^2 = 1567.36,$$

$$\mathrm{SS}_{\mathrm{Res}} = \sum_{i=1}^{N} \left(Y_i - \hat{Y}_i\right)^2 = 2388.641, \quad \text{and}$$

$$\mathrm{SS}_{\mathrm{T}} = \sum_{i=1}^{N} \left(Y_i - \bar{Y}\right)^2 = 3956.00.$$

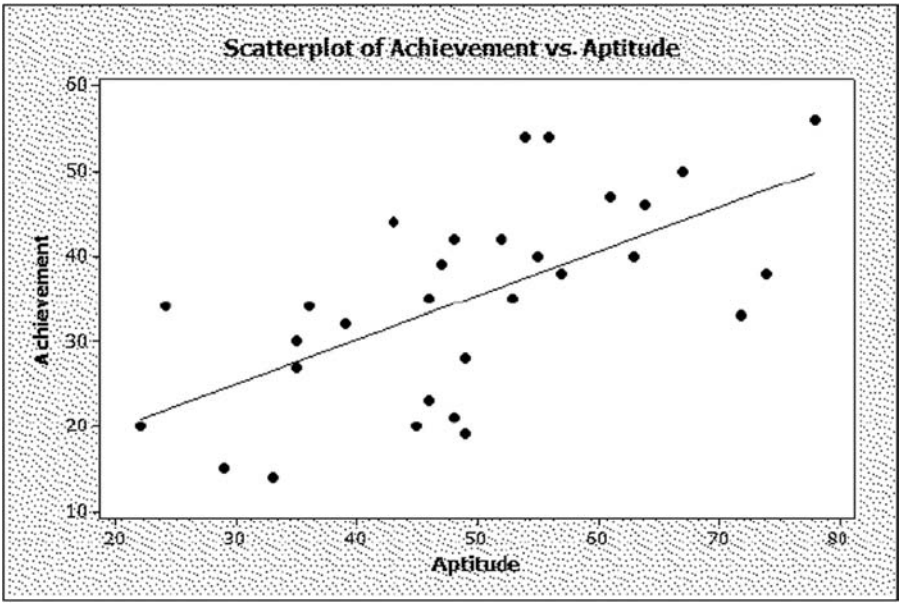


Figure 4.2 Scatterplot and regression line for the regression of achievement (Y) on aptitude (X).

The ANOVAR summary is presented in Table 4.5.

The critical value of F based on 1 and 28 degrees of freedom is 4.20 (for $\alpha = .05$). The p-value associated with the obtained F is .00019. Hence the null hypothesis is rejected and it is concluded that the population regression slope is not zero.

In summary, the general purpose of both ANOVA and ANOVAR F-tests (regardless of the design) is to answer the following general question: Is there a convincing evidence of an association between the independent and dependent variables? But there are differences between the two analyses in terms of the specificity of the question answered. The focus of ANOVAR is on whether there is any *linear* relationship between the independent and dependent variables whereas ANOVA tests whether there is any relationship whatsoever. Hence the results of ANOVA are diffuse whereas those of ANOVAR are focused on a linear relationship. Both analyses involve partitioning the total sum of squares into systematic and error components;

Table 4.5. ANOVAR Summary Table for Simple Linear Regression Example

Source	SS	*df*	MS	*F*
Regression	1567.36	1	1567.36	18.37
Residual	2388.64	28	85.31	
Total	3956.00	29		

the former treats between-group variability as the systematic component whereas the latter treats variability accounted for by a linear function of X as the systematic component.

Additional Descriptive and Inferential Procedures for Simple Regression Problems

The essentials of regression analysis include the fitted regression equation, a graph containing the original data and the regression line, and a test of the hypothesis that the population slope is zero. But additional descriptive and inferential aspects of regression analysis are frequently reported. The additional results of greatest interest are the coefficient of determination, the confidence interval on β_1, the standard error of estimate, and prediction intervals. These useful statistics will be described in this section along with a test that frequently is of no interest whatsoever. The latter will be described in order to convince the reader that it usually should be ignored even though it appears in virtually all regression software.

Coefficient of Determination: r^2. The proportion of the total sum of squares that is explained by the regression of Y on X is known as the "coefficient of determination." Although this is standard statistical terminology, it is easily misunderstood because it has a strong causal connotation. This connotation is typically not justified unless the analysis is performed on data from a true experiment.

The computation of the coefficient of determination involves forming the ratio of the regression sum of squares over the total sum of squares: $\frac{SS_{Regression}}{SS_{Total}} = r^2$. Alternatively, one could compute the Pearson correlation coefficient and then square it; that is, r^2 is simply the squared correlation between X and Y.

This coefficient is frequently defined as the proportion of the variation on Y explained using knowledge of X and a linear prediction rule (i.e., the fitted equation). This definition is sometimes viewed as unsatisfying because the term "explained" is not necessarily clear. Students often ask, "What (exactly) does it mean to say that the dependent variable is explained?"

This can be understood by first selecting one subject from the sample and describing what it means to *not* have variation on Y explained for this subject. Consider the first subject listed in Table 4.4. The dependent variable score for this subject is 15 and the sample mean of this variable is 35. The squared deviation from the mean is $(15 - 35)^2 = 400$. This squared deviation from the mean is unexplained variation: it can be viewed as the quantity that we would like to explain for this subject. Next, use the fitted regression equation to predict Y from this subject's X score. The fitted equation is $\hat{Y}_i = 9.413 + .519(X_i)$ and this subject's X score is 29. Therefore the prediction is $[9.413 + .519(29)] = 24.454$. The difference between the squared deviation of Y from the mean and the squared deviation of the predicted value from the mean is the amount "explained" by X and the linear prediction rule. That is, $(Y_1 - \bar{Y})^2$ is the total variation that is to be explained for subject 1 and $(\hat{Y}_1 - \bar{Y})^2$ is the variation that is actually explained for subject 1 using her X score and the fitted equation.

Hence, we could say that $\frac{(\hat{Y}_1-\bar{Y})^2}{(Y_1-\bar{Y})^2} = \frac{(24.454-35)^2}{(15-35)^2} = \frac{111.22}{400} = .28$ is the proportion of the variation on Y that is explained by X using a linear rule in the case of subject 1. Of course we could do this for every subject in the sample, but there is generally no interest in inspecting all N proportions. Instead there is interest in the overall (over the whole sample) proportion. If we compute the expressions shown above for each subject and then sum the numerator values and the denominator values separately, we get $\sum_{i=1}^{N}(\hat{Y}_i - \bar{Y})^2 = \text{SS}_{\text{Regression}}$ and $\sum_{i=1}^{N}(Y_i - \bar{Y})^2 = \text{SS}_{\text{Total}}$. Both of these sums are listed in the ANOVAR summary shown in Table 4.5. For example, the data ratio is $\frac{\text{SS}_{\text{Regression}}}{\text{SS}_{\text{Total}}} = \frac{1567.36}{3956} = r^2 = .396$. This means that approximately 40% of the total variation among the achievement scores (Y) in the sample is explained using knowledge of aptitude scores (X) and the fitted regression.

Alternative test of $\mathbf{H_0}$: $\boldsymbol{\beta_1 = 0}$. An alternative to the ANOVAR-F test of H_0: $\beta_1 = 0$ is a corresponding t-test. Most regression software routinely produces both of these tests even though they are redundant in the case of simple regression. (They are not redundant in the case of multiple regression.)

The t-test on the slope involves computing the ratio of the sample slope divided by the estimate of the standard error of the slope, using the following formula:

$$\frac{b_1}{s_{b_1}} = t,$$

where the standard error of b_1 is estimated using

$$s_{b_1} = \sqrt{\frac{\text{MS}_{\text{Residual}}}{\sum_{i=1}^{N} x^2}}.$$

The obtained value of t is then compared with the critical value of t based on $N - 2$ degrees of freedom.

There is a simple algebraic correspondence between the obtained t and the obtained ANOVAR-F. It is easy to demonstrate that $t^2 = \text{ANOVAR-}F$. Because these two tests are equivalent, the p-values associated with them are identical and there is no reason to prefer one test to the other.

Example 4.1 The data from the achievement example yield the following statistics:

$N = 30$,
$b_1 = .519$,
$\text{MS}_{\text{Residual}} = 85.31$,
$\sum_{i=1}^{N} x^2 =$ sum of squared deviation scores on $X = 5826.67$,
$s_{b_1} = \sqrt{\frac{85.31}{5826.67}} = .121$, and
$t_{\text{obt}} = .519/.121 = 4.29$.

Because the obtained value of t exceeds t_{cv}, which is 2.048 (for $\alpha_2 = .05$ and degrees of freedom = 28), the null hypothesis is rejected. The corresponding p-value (from *Minitab*) is .00019.

Confidence interval on β_1. An alternative to the hypothesis test just described is the corresponding confidence interval. A persuasive argument can be made that the confidence interval is more informative.

The 95% confidence interval on the unknown parameter β_1 is constructed around the sample slope b_1 as follows:

$$b_1 \pm [s_{b_1}(t_{cv})] = (L, U),$$

where

s_{b_1} is the estimate of the standard error of the slope;
t_{cv} is the critical value of the t distribution based on $\alpha_2 = .05$ and degrees of freedom $= N - 2$; and
(L, U) are the lower and upper limits of the interval.

Example 4.2 The construction of the 95% confidence interval on the population slope parameter β_1 is shown below for the aptitude and achievement data presented in Table 4.4. All of the required previously computed statistics are

$N = 30$,
$b_1 = .519$,
$\text{MS}_{\text{Residual}} = 85.31$,
$\sum_{i=1}^{N} x^2 =$ sum of squared deviation scores on $X = 5826.67$, and
t_{cv} (for $\alpha_2 = .05$ and degrees of freedom = 28) = 2.048.

The estimated standard error of the slope is

$$s_{b_1} = \sqrt{\frac{85.31}{5826.67}} = .121,$$

and the 95% confidence interval is

$$.519 \pm [.121(2.048)] = (.271, .767).$$

It can be stated (using confidence coefficient .95) that the unknown value of the population slope lies within such an interval.

Hypothesis test on β_0. Computer software routinely produces a test of significance on the intercept; the null hypothesis associated with this test is H_0: $\beta_0 = 0$. The test

statistic can be computed using

$$\frac{b_0}{\sqrt{\mathrm{MS}_{\mathrm{Residual}}\left[\frac{1}{N}+\frac{\bar{X}^2}{\sum_{i=1}^{N}x^2}\right]}}=t.$$

The obtained value of t is compared with the critical value of t based on $N-2$ degrees of freedom.

This test is usually of no interest and should be ignored unless there is substantive interest in the value of the intercept. If there is no substantive interest in the intercept (other than as a necessary part of the fitted equation), then surely there should be no interest in the p-value with which it is associated.

Recall that the intercept refers to the level of Y associated with X set equal to zero. Many X variables are scaled in such a manner that no values are near zero, but this does not interfere with the computation of the intercept. For example, a regression analysis was recently computed where Y was shoe size and X was height measured in inches. The value of the intercept was -22 and the associated p-value was far less than α. But there was no substantive interest in knowing the expected value of Y (shoe size) for a subject zero inches tall! Consequently there was no interest in the statistically significant test result on the intercept. This is frequently the case.

As with the shoe size example, it is often substantively impossible for a variable to have a value that is negative or zero. Similarly, it is often impossible for the X variable to have a value of zero. Consequently, it is not surprising that the intercept is often substantively worthless. The intercept is simply one of the two parameter estimates needed to define a straight line; if it does not have substantive meaning there is no interest in the associated test statistic and one should not be impressed if the associated p-value is small.

Example 4.3 The data from the achievement example yield the following statistics:

$N=30$,
$b_0=9.413$,
$\mathrm{MS}_{\mathrm{Residual}}=85.31$,
$\bar{X}=49.33$,
$\sum_{i=1}^{N}x^2=$ sum of squared deviation scores on $X=5826.67$, $s_{b_1}=\sqrt{85.31\left[\frac{1}{30}+\frac{(49.3\bar{3})^2}{5826.67}\right]}=6.203$, and
$t_{\mathrm{obt}}=9.413/6.203=1.52$.

Because the obtained value of t is less than t_{cv}, which is 2.048 (for $\alpha_2 = .05$ and degrees of freedom = 28), the null hypothesis is retained. The p-value is .14. In this example, it is possible for X (aptitude test score) to be zero; hence the intercept and the associated inference may be of some interest, but it is unlikely.

Standard Error of Estimate

The prediction error associated with a prediction is the difference between the actual value of Y and the value based on the prediction equation. That is, $(Y_i - \hat{Y}_i) = e_i$. If the predicted values equal the actual values for all subjects in a sample, there is no prediction error. But this is unrealistic; prediction error is always present in a realistic statistical model. It is useful to have an overall measure of prediction error based on the entire sample. The square root of the mean square residual is such a measure; this is an estimate of the parameter known as the standard error of estimate. Common notation for this estimate is $s_{Y|X}$, but I prefer to denote it as s_e. The subscript e specifies the variable on which the standard deviation estimate has been computed.

$$s_e = \sqrt{\frac{\sum_{i=1}^{N} \left(Y_i - \hat{Y}_i\right)^2}{N-2}} = \sqrt{\frac{\sum_{i=1}^{N} e_i}{df_{\text{Residual}}}} = \sqrt{\frac{\text{SS}_{\text{Residual}}}{df_{\text{Residual}}}} = \sqrt{\text{MS}_{\text{Residual}}}.$$

(Note that the standard deviation clearly has not been computed on the Y values conditional on X as is implied inappropriately when denoting the statistic as $s_{Y|X}$.)

The comparison of the standard error of estimate s_e with the standard deviation on Y (i.e., s_Y) reveals the extent to which the variation around the regression line (i.e., the prediction errors based on the regression equation) is smaller than the variation around the mean on Y. If there is no linear relationship between X and Y in the population, then the standard deviation on Y is exactly the same as the standard error of estimate. If there is a very strong linear relationship between X and Y, the standard error of estimate must be much smaller than the standard deviation on Y. Because standard deviations are familiar to many nonresearchers, the reporting of both s_e and s_Y often can be a useful way to convey the meaning of regression results.

Example 4.4 The quantities required in the computation of both s_e and s_Y are obtained from the ANOVAR summary table; they are shown below for the academic achievement example:

$$s_e = \sqrt{\text{MS}_{\text{Residual}}} = \sqrt{85.31} = 9.24,$$

$$s_Y = \sqrt{\frac{\text{SS}_{\text{Total}}}{df_{\text{Total}}}} = \sqrt{\frac{3956}{29}} = 11.68.$$

Because the standard error of estimate is substantially smaller than the standard deviation on Y, it can be seen that knowledge of X (aptitude) and the fitted equation is useful in accounting for much variation on Y (achievement).

Prediction Intervals

Suppose a sample drawn from the population of applicants to your organization was tested on selection test X, all applicants were hired, and after an adequate period, job performance evaluations (Y) were obtained. Further, suppose a regression analysis was performed and the conclusion was that there is a linear relationship between test scores and job performance in the sampled population.

The prediction equation from this analysis can be used to predict job performance for any future applicant providing a score on X. An important question is: How accurate should the prediction be considered to be? The purpose of a so-called prediction interval is to provide an answer this question.

The prediction interval is constructed using the following expression:

$$\hat{Y}_k \pm s_e \left({}_{\alpha_2}t_{N-2}\right) \sqrt{1 + \frac{1}{N} + \frac{\left(X_k - \bar{X}\right)^2}{\sum_{i=1}^{N} x^2}} = (L, U),$$

where

$\hat{Y}_k = b_0 + b_1 X_k$;
s_e is the standard error of estimate;
X_k is the predictor score for the kth subject (who is not in the original sample used to estimate the prediction equation);
N is the sample size used in estimating the prediction equation; and
${}_{\alpha_2}t_{N-2}$ is the critical value of t associated with α_2 (alpha is set at .05 (nondirectional) for a 95% prediction interval) and $N - 2$ degrees of freedom.

One can state (using confidence coefficient .95) that the unknown value of Y that will be obtained for future subject k will fall between the upper and lower limits of the interval.

Example 4.5 Suppose we are interested in predicting the biology achievement score for student k; this student obtained a biology aptitude test score of 60. Keep in mind that this subject is not in the sample that was used to estimate the prediction equation. It can be seen that there is no subject with an aptitude test score of 60 in Table 4.4. Each of the subjects listed in the table has a Y score; this is not true for new subject k; this subject has not yet been admitted to the university. The purpose of the prediction interval is to provide limits within which subject k's achievement score will fall, if he enrolls.

The point estimate for subject k is $9.413 + .519(60) = \hat{Y}_k = 40.53$ and the 95% prediction interval is

$$40.53 \pm 9.236(2.048)\sqrt{1 + \frac{1}{30} + \frac{(60 - 49.3\bar{3})^2}{5826.67}} = (21.12,\ 59.95).$$

Hence, it is predicted (using confidence coefficient .95) that subject k's future achievement score (Y_k) will fall within the limits of the interval that ranges from approximately 21 to 60.

4.4 DIAGNOSTIC METHODS: IS THE MODEL APT?

The adequacy of the fitted regression equation as a description of the relationship between X and Y depends upon the actual form of the relationship between these variables. Suppose a simple regression analysis is computed on data where the relationship between X and Y is strongly nonlinear in form. These data will be misrepresented by the linear equation and most aspects of the analysis will be invalid. For example, predictions will be invalid using many (if not most) values of X and the prediction intervals will be very misleading. At some values of X, the prediction intervals will be much narrower than they should be and at other values of X they will be wider than they should be. In the case of severe nonlinearity some of the prediction intervals may contain no future values at all. Consequently, it is important to determine whether the linear model is a reasonable representation of the data. Although many formal tests for nonlinearity and other departures from the assumptions of the simple linear model exist, they are often impotent in the case of small samples and unnecessary in the case of large samples. Instead, I recommend that three or four simple graphs be used to provide the essential diagnostic information that is required.

The required graphs display the plot of Y on X (with the regression line included) as shown earlier in Figure 4.2, the plot of the residuals e on X, as shown in Figure 4.3, and the univariate plot of the residuals, as shown in Figure 4.4. The first graph should be included routinely in reporting results of a regression analysis. This graph provides an overall view of the nature of the data. It effectively displays linear trend, but it does not provide a very sensitive display of departures from linearity.

The plot of e on X (Figure 4.3) can clearly reveal departures from linearity, and other issues including independence of errors, heteroscedasticity, and outliers. The third graph (Figure 4.4) is the univariate plot of e; it illustrates the shape of the residual distribution. Plots of this type usually do not look normal when small samples are involved even though the sampled population distribution is normal; but they are often useful for identifying obvious departures from a reasonable approximation to normality. The appearance of the residuals displayed in Figure 4.4, for example, does not seem to conform to a normal distribution, but because the sample is small the departures would not be considered extreme. A more sophisticated evaluation of departures from normality involves a so-called normal probability plot, which is shown in Figure 4.5.

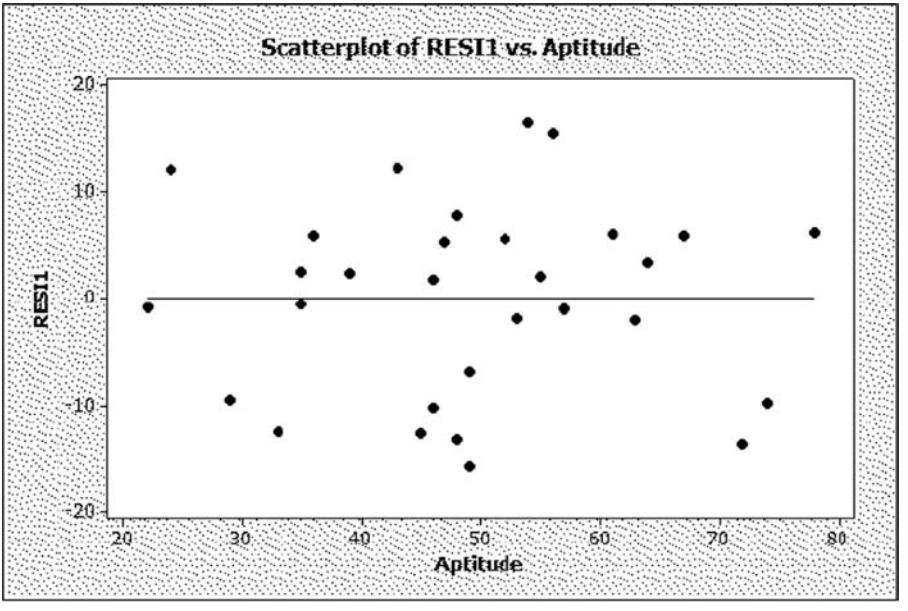


Figure 4.3 Scatterplot of residuals (Y) and aptitude (X).

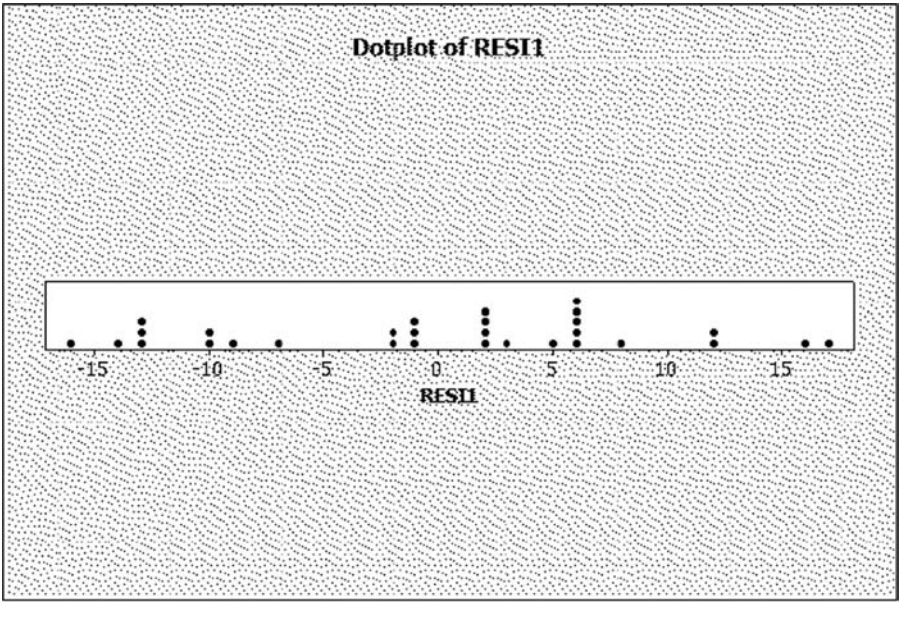


Figure 4.4 Univariate plot of residuals from the regression of achievement (Y) on aptitude (X).

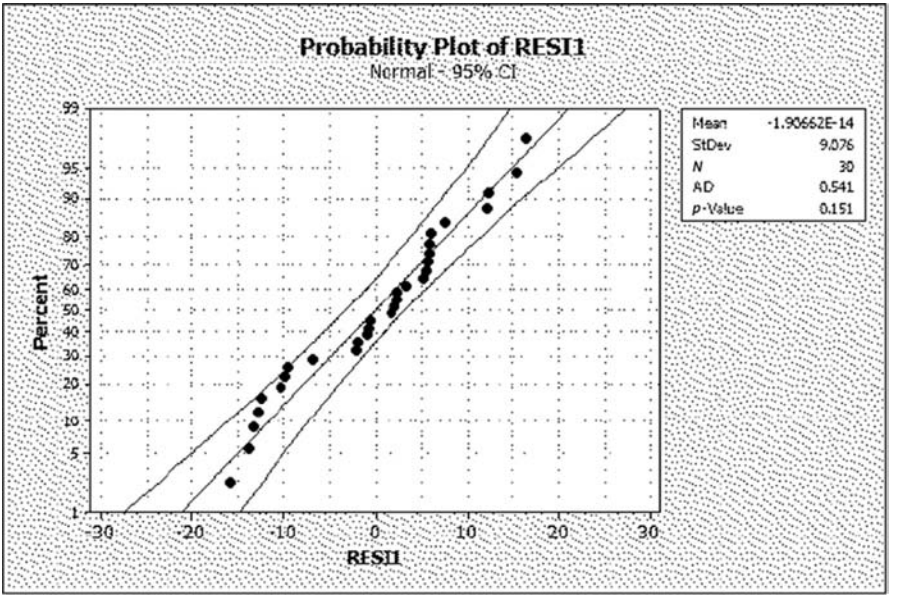


Figure 4.5 *Minitab* probability plot of the residuals from the regression of achievement (Y) on aptitude (X).

The straight line shown in this plot is not the regression line. Rather, this line indicates where the residuals will fall if they are exactly normally distributed (which never happens). Hence, departures from the line indicate departures from exact normality. If the residuals fall fairly close to the line so that they are contained within the curved bounds, they are considered to conform to approximate normality. The box included to the right of the plot contains several statistics; the one at the bottom is the p-value associated with a formal test for normality known as the Anderson–Darling (A–D) test. Small values of p indicate significant departures from normality. Neither the probability plot nor the A–D test is necessarily more informative than the simple univariate plot of the residuals. Univariate residual plots are usually the most direct way of capturing the distribution shape.

The three recommended plots for the achievement example lead to the following conclusions: (1) the relationship is linear (note the clear linear trend apparent in the plot of Y on X in Figure 4.2), (2) there are no obvious departures from linearity or homoscedasticity, and no outliers are present (see Figure 4.3), and (3) there are moderate departures from normality (see Figure 4.4). Figure 4.5 confirms that the departures from normality are not extreme because all residuals are within the bounds and the A–D test p-value is above conventional levels for alpha.

4.5 SUMMARY

Various aspects of the linear relationship between X and Y are described using the correlation coefficient, the slope, the coefficient of determination, and the standard

error of estimate. The correlation coefficient describes the direction and degree of linear relationship, the slope describes the average increase on Y associated with a one unit increase on X, the coefficient of determination describes the proportion of the variation on Y that is explained using knowledge of X and a linear prediction rule, and the standard error of estimate describes the standard deviation of the prediction errors.

Once the parameters of the regression model are estimated, the prediction scheme (i.e., the linear prediction rule defined by the fitted equation) can be specified and predictions can be made. The accuracy of a specific prediction based on the regression equation is conveyed by the width of the prediction interval.

In some cases researchers want to move beyond description and prediction. If there is interest drawing causal conclusions regarding the relationship, one usually must make strong (and probably unrealistic) assumptions or collect data based on an experimental design. That is, a design in which one randomly assigns subjects to levels of X and then experimentally manipulates those levels. In this case, it will be the design rather than the simple regression analysis that justifies strong causal statements.

CHAPTER 5

Essentials of Multiple Linear Regression

5.1 INTRODUCTION

The previous chapter described the simple linear regression model; recall that this model contains only one predictor variable. The present chapter describes extensions of regression analysis that accommodate multiple predictor variables. Although many applications of multiple regression analysis are similar to those of simple regression, other applications are considerably more complex. Multiple regression is often used to (1) predict Y from two or more predictor variables, (2) clarify the predictive role of a single predictor variable in a set of predictor variables, (3) partially control for confounding, (4) fit curves to data when the relationship between X and Y is not linear, and (5) serve as a general explanatory system describing how Y is related to a set of predictor variables.

As with simple regression, a single variable must be chosen as the outcome variable to be explained. For example, in describing a study of hypertension one might state that, "Blood pressure was regressed on age and saturated fat consumption." This phrase conveys the idea that blood pressure is the dependent variable and the other two measures (age and saturated fat consumption) are the predictor variables. The dependent variable is always "regressed on" one or more predictor variables. I begin this overview with a gradual introduction to the basics of multiple regression analysis in the context of the two-predictor model. Once this model is understood, the extension to models with more than two predictors is straightforward.

The Analysis of Covariance and Alternatives: Statistical Methods for Experiments, Quasi-Experiments, and Single-Case Studies, Second Edition. Bradley E. Huitema.

5.2 MULTIPLE REGRESSION: TWO-PREDICTOR CASE

Two-Predictor Model

A regression problem with two predictors is often represented using the following model for the population data:

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \varepsilon_i,$$

where

Y_i is the dependent variable score for individual i;
β_0 is the Y intercept;
β_1 is the first partial regression coefficient;
β_2 is the second partial regression coefficient;
X_{1i} is the score on the first predictor for individual i;
X_{2i} is the score on the second predictor for individual i; and
ε_i is the error component associated with individual i; that is,

$$\varepsilon_i = [Y_i - (\beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i})].$$

It is useful to compare directly the one- and two-predictor models, which are repeated below:

Simple regression model:	$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i,$ and
Two-predictor multiple regression model:	$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \varepsilon_i.$

Note that the two-predictor model differs from the simple regression model only in that it contains an additional coefficient (β_2) and the associated predictor variable (X_2). Recall that the two coefficients (β_0 and β_1) in the simple regression model define the least-squares regression *line*. Correspondingly, the three coefficients (β_0, β_1, and β_2) in the two-predictor multiple regression model define the regression *plane*.

The β_0 coefficient is called the intercept in both models. In the simple model, it is defined as the average elevation on Y when X is set equal to zero; in the two-predictor model it is defined as the average elevation on Y when both X_1 and X_2 are set equal to zero. In the two-predictor model, both β_1 and β_2 are called "partial regression coefficients" rather than slopes; this terminology suggests that a partial regression coefficient differs from a slope. Unfortunately, it is not obvious from visual inspection of the two models that β_1 has one meaning under the simple regression model and another meaning under the multiple regression model.

Interpretation of Partial Regression Coefficients

Recall that in the case of simple regression, the coefficient β_1 is interpreted as the average increase on Y when X increases by one unit. In the case of a two-predictor multiple regression model, β_1 is the average increase on Y associated with a one-unit

increase on X_1, *holding constant* X_2. Although the standard notation for the two-predictor model has been used above, I have rewritten it in an embellished manner (below) in order to clarify the nature of all three of these population parameters:

$$Y_i = \beta_{0|X_1,X_2} + \beta_{1|X_2}X_1 + \beta_{2|X_1}X_2 + \varepsilon_i.$$

Note the vertical line that appears in the subscript after each coefficient. Read this line as "conditional on." Hence, the subscript on the first partial regression coefficient $\beta_{1|X_2}$ informs us that this coefficient is conditional on X_2. Similarly, the second partial regression coefficient is conditional on X_1. The same notion is relevant in the case of the fitted equation.

Fitted Equation

The fitted equation is normally written as $\hat{Y}_i = b_0 + b_1X_{1i} + b_2X_{2i}$, where the coefficients b_0, b_1, and b_2 are described as the sample estimates of the corresponding population parameters β_0, β_1, and β_2. The coefficients in this equation can be embellished to produce the following form of the fitted equation: $\hat{Y}_i = b_{0|X_1,X_2} + b_{1|X_2}X_1 + b_{2|X_1}X_2$. The latter form is more cumbersome but it warns the reader that the various coefficients are conditional. The extensive computation required to obtain the parameter estimates is not described here; details are presented in Section 5.3. Good routines for estimating multiple regression models are available in all major statistics software packages. I focus the remainder of this section on an understanding of partial regression coefficients.

In order to understand what it means for the sample partial regression coefficient b_1 to be conditional on X_2, it is necessary to recall two concepts from simple regression: (1) the residual and (2) the residual sum of squares. Remember that the residual e_i is simply the actual observed score on Y minus the value of Y predicted using the fitted equation. Specifically, $e_i = (Y_i - \hat{Y}_i)$, where the predicted value is computed using $\hat{Y}_i = b_0 + b_1X_i$. Variation among a sample of residuals is linearly independent of scores on X. That is, all variation in the sample of Y scores (i.e., the total sum of squares) that is predictable using a linear function of X (i.e., the regression sum of squares) is subtracted from the total sum of squares in order to compute the residual sum of squares. Hence, the correlation between e_i and X must be zero.

The idea of residualizing one variable with respect to another variable can be used to understand partial regression coefficients. Suppose a two-predictor analysis is of interest. First we regress Y on X_2 and label the residuals of this regression as "Y residuals." Next, we regress X_1 on X_2 (i.e., X_1 is treated as the dependent variable and X_2 is used as the only predictor) and label the residuals of this regression as "X_1 residuals." Two sets of residuals are now available. Because both Y and X_1 were residualized with respect to X_2, each set of residuals is linearly independent of X_2. That is, there is no variation in either set of residuals that can be predicted from X_2 using a linear function. Therefore, it can be argued that both sets of residuals are free of linear association with X_2 (as measured). This implies that the linear association between the two sets of residuals is linearly independent of X_2.

If the "Y residuals" are now regressed on the "X_1 residuals" (using simple regression), we will discover that the slope resulting from this simple regression is identical to the first partial regression coefficient from the multiple regression of Y on both X_1 and X_2. Because the first partial regression coefficient in a two-predictor model is exactly the same value as the slope based on the simple regression of the Y residuals on the X_1 residuals, we can see exactly what is meant when the X_2 variable is said to be "held constant." In this situation, variable X_2 (as measured) is said to be partialed out of the association between variables Y and X_1.

Correspondingly, the second partial regression coefficient (i.e., $\beta_{2|X_1}$) involves the association between Y and X_2, holding constant X_1. The value of this coefficient is equal to the value of the slope that would be obtained if one were to residualize both Y and X_2 with respect to X_1, and then carry out a simple regression of the Y residuals on the X_2 residuals. That is, the value of the slope from this analysis is equal to the value of the second partial regression coefficient in the multiple regression of Y on both X_1 and X_2. This coefficient describes the average increase on Y associated with a one-unit increase on X_2, holding constant X_1. This interpretation makes sense because variation in both Y and X_2 that is a linear function of X_1 is removed before estimating the coefficient. In this sense it is argued that X_1 has been held constant statistically.

Example 5.1 Suppose we have a regression problem in which there is interest in investigating the relationship between two predictor variables and a measure of heart disease. The first predictor is a measure of psychological stress (X_1) obtained 5 years before the outcome variable is measured; the second predictor is a measure of obesity (X_2), also measured 5 years before the dependent variable is measured. The dependent variable (Y) is a new measure of the severity of heart disease that is based on a combination of coronary calcium measurements and the percentage of narrowing of coronary arteries. Data from an unreasonably small sample ($N = 10$) is presented in Table 5.1. I encourage the reader to carry out all of the analyses described below.

As with any other type of analysis, the first step involves plotting the raw data. The recommended plots include a univariate plot of each variable, three separate scatterplots (to describe the relationship between each pair of variables), and an approximation to a three-dimensional scatterplot of all three variables. Plots of the last type are available in the interactive three-dimensional scatterplot option in *SPSS;* this option provides representations of the data that appear to be in three dimensions. These representations can be twisted and turned in all directions; visual inspection of these representations is helpful in gaining an understanding of the relationships among all three variables simultaneously. Outliers and major departures from linearity of the relationships among all variables are usually apparent in these plots.

If we carry out a multiple regression analysis on the data in Table 5.1 (i.e., we regress Y on both X_1 and X_2), the numerical descriptive part of the output from typical regression software (in this case *Minitab*) is of the following form:

The regression equation is

$$Y = -21.5 - .074(X_1) + 0.854(X_2).$$

Table 5.1 Data for a Two-Predictor Multiple Regression Analysis Where Psychological Stress and Obesity Measures Are Predictors and Heart Disease Is the Dependent Variable

Patient	X_1 = Stress	X_2= Obesity	Y = Heart Disease
1	19	43	12
2	25	50	23
3	31	52	19
4	30	54	20
5	19	42	13
6	26	54	22
7	31	58	27
8	28	52	25
9	22	50	21
10	26	54	19

Note that the output indicates that the "regression equation" is presented in the second line. There is a notation problem with this output that is typical of regression analysis software. Recall that there is a difference between the statistical model and the corresponding fitted equation. In the present example, the two-predictor multiple regression *model* (using conventional rather than embellished notation) is written as

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \varepsilon_i,$$

whereas the corresponding *fitted equation* is

$$\hat{Y}_i = b_0 + b_1 X_{1i} + b_2 X_{2i}.$$

Computer output provides the coefficients for the latter not the former. Therefore, the output symbol on the left should be $\hat{Y}$ rather than Y because a fitted equation provides the prediction of the value on Y, *not* the actual value on Y.

Inspect the coefficients in the fitted equation; the intercept estimate $b_0 = -21.5$, the first partial regression coefficient estimate $b_1 = -.0735$, and the second partial regression coefficient estimate $b_2 = .8544$. These estimates are interpreted as follows: The average score on Y (heart disease) decreases by .0735 points with each one-unit increase on X_1 (psychological stress) holding constant X_2 (obesity). Similarly, heart disease increases an average of .8544 points with each one-unit increase on X_2 (obesity), holding constant X_1 (psychological stress). These coefficients describe the sample regression plane. The fitted equation is the basic descriptive result of the regression analysis. When the purpose of the study is purely descriptive, there is no need for the inferential portion of the output that is provided by virtually all regression software. But when the sample has been randomly selected from a defined population and there is interest in generalizing the sample results to that population, inferential methods are relevant. These methods are described next.

Inferential Methods

Evaluating the Combination of Predictors

Before inferring that the combination of predictor variables in the equation is helpful in predicting heart disease in the population, it may be useful to carry out a formal test to evaluate whether the sample data justify such an inference. The specific question is: Can population heart disease scores be predicted with less error using the optimum linear combination of variables X_1 and X_2 (in the population equation) than would be found if the predictors were ignored? In other words, is it worthwhile to include these predictors in the *population* model? We already know the set of predictors is useful in predicting the Y scores in the observed *sample* (because the sample partial regression coefficients are not equal to zero), but this is not the issue. What we want to know is whether these meager sample data ($N = 10$) are sufficient for us to infer that the same is true in the population from which the sample was drawn.

This question can be formalized as a null hypothesis and answered using a model comparison test. The null hypothesis is: H_0: $\beta_1 = \beta_2 = 0$. The two models to be compared are

$$\begin{aligned} &\text{Model I:} \quad Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \varepsilon_i . \\ &\text{Model II:} \quad Y_i = \beta_0 + \varepsilon_i . \end{aligned}$$

If the two partial regression coefficients in the first model are set equal to zero, the result is the second model. A rejection of the null hypothesis implies that the model including the predictors is more adequate than the model without them. We can test the hypothesis by comparing the residuals from fitting the two-predictor regression model (model I) with the residuals from fitting the model that ignores information on the two predictors (model II).

The residuals from the two models are defined as follows:

$$\begin{aligned} &\text{Model I residual:} \quad e_i = (Y_i - \hat{Y}_i), \text{ where } \hat{Y}_i = b_0 + b_1 X_{1i} + b_2 X_{2i} . \\ &\text{Model II residual:} \quad (Y_i - b_0) = (Y_i - \bar{Y}) . \end{aligned}$$

Note that the mean on Y is the value that is predicted for each subject using model II (the intercept is equal to the mean on Y when there are no predictors in the model). For the example data we find that the residual SS using model II is $\sum_{i=1}^{N} (Y_i - \bar{Y})^2 = 202.90$.

The residual SS under model I is $\sum_{i=1}^{N} (Y_i - \hat{Y}_i)^2 = 60.04$, where the predicted values are based on the prediction equation $\hat{Y}_i = -21.5 - .074(X_1) + 0.854(X_2)$.

The F-test has the following form:

$$\frac{(\mathrm{SS}_{\mathrm{Res\,II}} - \mathrm{SS}_{\mathrm{Res\,I}})/(df_{\mathrm{Res\,II}} - df_{\mathrm{Res\,I}})}{\mathrm{SS}_{\mathrm{Res\,I}}/(df_{\mathrm{Res\,I}})} = F,$$

where $\mathrm{SS}_{\mathrm{Res\,I}}$ and $\mathrm{SS}_{\mathrm{Res\,II}}$ are the residual sum of squares from fitting models I and II, respectively, and $df_{\mathrm{Res\,I}}$ and $df_{\mathrm{Res\,II}}$ are the residual degrees of freedom associated with

models I and II, respectively. Because residual degrees of freedom are equal to the number of observations minus the number of parameters estimated, $df_{\text{Res I}} = N - 3$ and $df_{\text{Res II}} = N - 1$. Consequently, the obtained value of F is evaluated using the F-distribution based on 2 and $N - 3$ degrees of freedom.

The application of this formula to the example data yields $\frac{(202.90-60.04)/2}{60.04/7} = 8.33$. The p-value associated with this F is .014. Therefore, the null hypothesis is rejected and it is concluded that the population partial regression coefficients β_1 and β_2 are not both equal to zero. Consequently, a strong argument can be made that population predictions will be more accurate using model I than using model II.

Although the approach described above is not difficult to apply once the two sets of residuals are computed, this is not the form of the test that is usually carried out; there is an easier but equivalent way. It turns out that this model comparison test is equivalent to the conventional ANOVAR F-test that is automatically provided in output produced by virtually all regression computer routines. The differences between the two forms of the test can be found in the labels used and the arrangement of the arithmetic. Specifically, the regression SS in the conventional ANOVAR summary table is equivalent to the difference between the residual SS from fitting model I and the residual SS from fitting model II, the residual SS in the ANOVAR table is the residual sum of squares from fitting model I, and the total SS in the ANOVAR table is the residual sum of squares from fitting model II.

The inferential portion of *Minitab* regression output is shown below for the example data:

```
Predictor      Coef      StDev      T        p
Constant    -21.50      11.88    -1.81     .113
X1           -.0735       .4430   -.17      .873
X2             .8544      .4008   2.13      .071

s = 2.929   R-Sq = 70.4%  R-Sq (adj) = 62.0%

Analysis of Variance

Source           DF       SS        MS      F      p
Regression       2   142.863     71.432  8.33   .014
Residual Error   7    60.037      8.577
Total            9   202.900
```

Ignore the top half of this output for now; concentrate on the bottom half. Note that the ANOVAR F is exactly the same as was found using the model comparison approach. This implies that the ANOVAR F-test is a model comparison method.

Evaluating Individual Predictors

The ANOVAR F-test and the associated correlational statistics such as R^2 (discussed subsequently) provide no information regarding the individual predictors. Rather, these statistics are relevant in evaluating the two predictors as a set. Information

regarding the individual predictors requires analyses of two different types because there usually is interest in two types of descriptions.

First, the researcher may want a description of the relationship between Y and each predictor variable separately. Two separate simple regression analyses (i.e., heart disease regressed on stress and heart disease regressed on obesity) provide these descriptions. These separate analyses are shown below.

```
Regression Analysis: Heart Dis versus Stress
The regression equation is
Heart Dis = 0.79 + 0.751 Stress

Predictor       Coef    SE Coef      T       P
Constant       0.794      6.756     0.12   0.909
Stress         0.7512     0.2593    2.90   0.020

S = 3.51792   R-Sq = 51.2%   R-Sq(adj) = 45.1%

Analysis of Variance
Source           DF      SS       MS      F      P
Regression        1   103.89   103.89   8.39  0.020
Residual Error    8    99.01    12.38
Total             9   202.90



Regression Analysis: Heart Dis versus Obesity

The regression equation is
Heart Dis = - 20.4 + 0.796 Obesity

Predictor     Coef      SE Coef    T      P

Constant    -20.434     9.357    -2.18  0.06
Obesity      0.7964     0.1830    4.35  0.002

S = 2.74484   R-Sq = 70.3%   R-Sq(adj) = 66.6%

Analysis of Variance

Source            DF      SS       MS       F      P
Regression         1   142.63   142.63   18.93  0.002
Residual Error     8    60.27     7.53
Total              9   202.90
```

The individual coefficients of determination associated with stress and obesity are .512 and .703, respectively. These values describe the percentage of variation on heart disease explained by stress (alone) and by obesity (alone). The p-values associated with the t-statistics for the slopes (which also are tests for the corresponding correlation coefficients and coefficients of determination) are .020 and .002 for stress

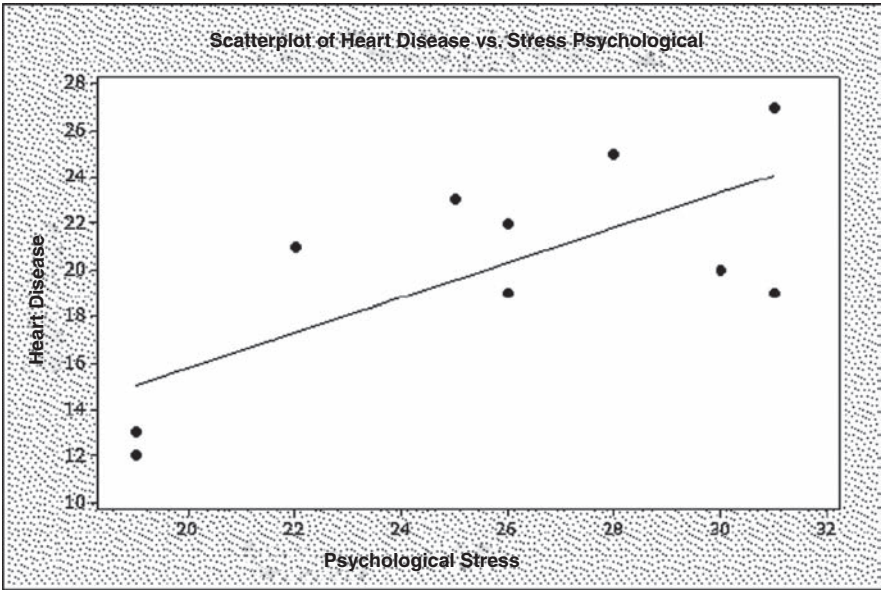


Figure 5.1 Scatterplot of heart disease on psychological stress.

and obesity, respectively. These statistics confirm the substantial linear relationships that can be seen in the plots of heart disease on the individual predictors, which are shown in Figures 5.1 and 5.2.

Hence, the data in this example provide strong evidence for the conclusion that there is a linear relationship between heart disease and each of the individual predictors. Keep in mind, however, that these conclusions are not based on the multiple regression analysis presented earlier. The multiple regression analysis provides a different type of information regarding the two predictors. The first half of the *Minitab* inferential output for the multiple regression analysis (previously ignored) is repeated below.

```
Predictor      Coef      StDev      T        p
Constant    -21.50      11.88    -1.81    .113
X1            -.0735      .4430    -.17    .873
X2             .8544      .4008    2.13    .071

s = 2.929   R-Sq = 70.4%  R-Sq (adj) = 62.0%
```

Note that there is a row for the intercept (labeled as the "constant") and each predictor (X_1 and X_2). It can be seen that the partial regression coefficient associated with stress is $-.0735$. Recall that a partial regression coefficient provides information on the relationship between two residualized variables; in this example both heart disease and stress have been residualized with respect to obesity. A visual display of these two residualized variables can be seen in Figure 5.3. The residuals from

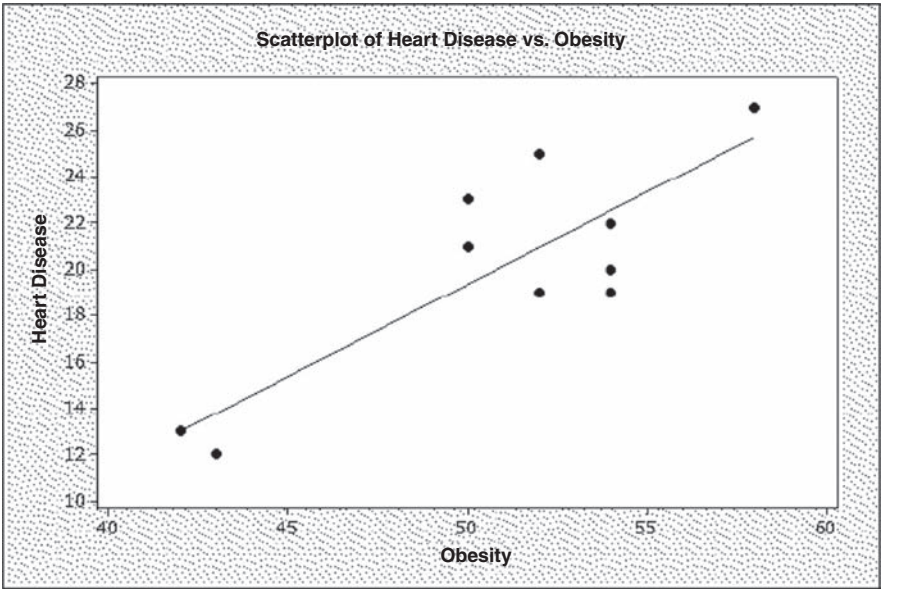


Figure 5.2 Scatterplot of heart disease on obesity.

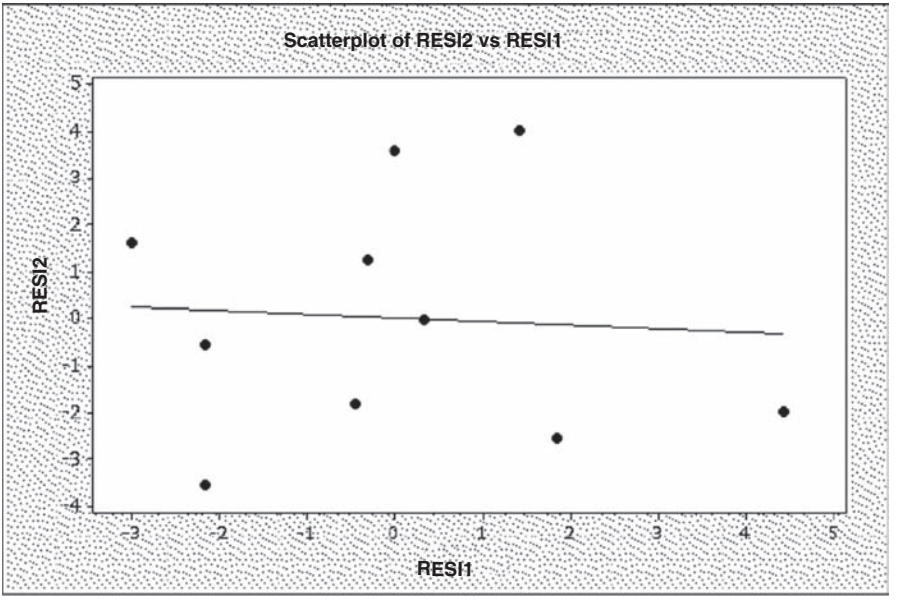


Figure 5.3 Scatterplot of heart disease residuals (RES2) on stress residuals (RES1). Both heart disease and stress were residualized with respect to obesity.

the regression of heart disease on obesity are labeled in the figure as RESI2 and the residuals from the regression of stress on obesity are labeled as RESI1.

It can be seen that this plot differs greatly from the one shown in Figure 5.1. The slope illustrated in Figure 5.1 is .7512 ($p = .020$) and the first partial regression coefficient illustrated in Figure 5.3 is $-.0735$ ($p = .873$). Hence, the relationship between heart disease and stress very much depends on whether obesity is in the model.

Because the relationship between heart disease and stress essentially disappears when obesity is controlled, the need to retain stress in the model should be questioned. It can be argued that the model using only obesity as the predictor is likely to be more predictive than the model that includes both predictors. Two statistics, the adjusted R^2 and s (found half way down the multiple regression output shown above), support this argument. The adjusted R^2-statistic is described in detail in Section 5.3. Briefly, it is a reduced-bias estimator of the proportion of population variation on Y that is explained by the predictor(s) in the model.

The value of the adjusted R^2 under the two-predictor model is .62, whereas the one-predictor (obesity only) value is .666. This implies that the one-predictor model is more effective than the two-predictor model in predicting heart disease and that including stress as a second predictor is likely to actually reduce the proportion of variation explained in the population. Correspondingly, the estimates of the standard error of estimate (denoted as "s" in *Minitab* output) are 2.929 and 2.745 for the two-predictor and one-predictor models, respectively. Hence, the estimated prediction errors are expected to be smaller with the one-predictor model than with the two-predictor model.

This outcome may seem counterintuitive; it does not appear reasonable to argue that the optimum combination of stress and obesity is less predictive than obesity alone. This intuition is correct as long as we are dealing with the sample situation. That is, it is not possible for the multiple regression sum of squares to be smaller than the regression sum of squares for either the first predictor (alone) or the second predictor (alone). The same is true of the sample coefficient of determination R^2. Note that the coefficient of determination is larger for the two-predictor model ($R^2 = .704$) than for either the one-predictor model that uses stress as the predictor ($r^2_{yx_1} = .512$) or the one-predictor model that uses obesity as the predictor ($r^2_{yx_2} = .703$). But, as shown in the previous paragraph, the situation changes when using the adjusted R^2-values, which are less biased estimates of the unknown population coefficients of determination. The adjusted values are more appropriate than the unadjusted values for inferential purposes.

If unnecessary predictors are included in a regression analysis, they are likely to introduce noise to the fitted equation. In most cases, the coefficient associated with an important predictor in the equation will be estimated with less precision if the model includes an irrelevant predictor than if the model does not. This is acknowledged in the size of the standard error that is used as the denominator in the t-test for the partial regression coefficient. Hence, the precision of the coefficients and ultimately the accuracy of the $\hat{Y}$'s provided by the prediction equation are increased when unnecessary predictors are removed from the equation. Therefore, tests on the partial regression coefficients are important because they can identify superfluous

predictors. Unfortunately, the computation of the standard errors required for these *t*-tests (shown in computer output) is neither intuitive nor simple; for this reason it will not be pursued here. On the other hand, *F*-tests that are equivalent to the *t*-tests are much more intuitive and straightforward; I describe these in the next subsection.

***F*-tests on Partial Regression Coefficients**

The purpose of this subsection is to clarify the nature of the *t*-tests on partial regression coefficients that are shown above in the *Minitab* output. These tests are provided in output from virtually all regression routines. Although these *t*-values are algebraically equivalent to the *F*-values to be described (i.e., $t^2 = F$), there are two reasons to pursue the *F* approach. First, the logic behind the *F* can be understood without knowledge of matrix algebra or variance-covariance matrices. Second, it involves the computation of quantities that have direct application in the estimation of advanced correlation measures to be described subsequently.

Recall that the first partial regression coefficient (b_1) is conditional on the second predictor being in the model. Just as b_1 is conditional on X_2, there is a corresponding regression sum of squares for X_1 that is conditional on X_2 being in the model. In order to compute this conditional sum of squares, we need the results of two regressions: (1) the regression of Y on both X_1 and X_2 and (2) the regression of Y on X_2 alone.

The example *Minitab* ANOVAR summary tables from the two regression analyses are repeated below:

Two-predictor (X_1 = Stress, X_2 = Obesity) ANOVAR:

Source	DF	SS	MS	F	p
Regression	2	142.863	71.432	8.33	.014
Residual Error	7	60.037	8.577		
Total	9	202.900			

One-predictor (X = Obesity) ANOVAR:

Source	DF	SS	MS	F	P
Regression	1	142.63	142.63	18.93	0.002
Residual Error	8	60.27	7.53		
Total	9	202.90			

Note that the regression sum of squares explained by X_1 and X_2 in combination is 142.863. Next, note that regression sum of squares explained by obesity alone (i.e., X_2) is 142.63. If we subtract the sum of squares for X_2 (alone) from the sum of squares explained by both X_1 and X_2 in combination, we can determine the sum of squares uniquely explained by X_1. Hence, the unique contribution of X_1 is $(142.863 - 142.63) = .233$. This value is then entered as the SS for the unique contribution of X_1 in the first row of a new ANOVAR summary table that is shown below. The associated *df* is the difference between the regression *df* for the two-predictor model and the regression *df* for the one-predictor model (i.e., $[2 - 1] = 1$). The entire second line

of this table is taken from the previously computed ANOVAR summary table for the two-predictor model.

Source	SS	df	MS	F	p
Unique contribution of X1	.233	1	.233	.0272	.87
Two-predictor Residual	60.037	7	8.577		

The F-ratio is defined as $\frac{\text{MS}_{\text{Unique contribution of } X_1}}{\text{MS}_{\text{Two-predictor residual}}}$ and the associated degrees of freedom are 1 and $N - 3$. The p-value associated with the obtained value of F (based on 1 and 7 degrees of freedom) is determined using the *Minitab* calculator for the F-distribution. Note that the p-value associated with this test agrees with the p-value found for the t-test on the first partial regression coefficient shown earlier for the two-predictor analysis.

The F-test on the second partial regression coefficient follows a similar pattern. In this case, we need the *Minitab* ANOVAR results from the previously computed simple regression of Y on X_1 alone as well as the results of the two-predictor analysis. These are repeated below:

Two-predictor (X_1 = Stress, X_2 = Obesity) Analysis:

Source	DF	SS	MS	F	p
Regression	2	142.863	71.432	8.33	.014
Residual Error	7	60.037	8.577		
Total	9	202.900			

One-predictor (X_1 = Stress) Analysis:

Source	DF	SS	MS	F	P
Regression	1	103.89	103.89	8.39	.020
Residual Error	8	99.01	12.38		
Total	9	202.90			

The focus now is on the unique contribution of X_2 beyond that which is explained by X_1. Therefore, we subtract the regression SS for X_1 alone (103.89) from the regression SS for X_1 and X_2 in combination (143.863). The difference between these two describes the SS for the unique contribution of X_2; i.e., $(142.863 - 103.89) = 38.973$ = SS unique contribution of X_2. As before, there is one *df* for the unique contribution, and the entire second line is from the two-predictor ANOVAR. The summary table for testing the unique contribution of X_2 is laid out as follows:

Source	SS	df	MS	F	p
Unique contribution of X2	38.973	1	38.973	4.5369	.071
Two-predictor Residual	60.037	7	8.577		

The square root of the obtained F is equal to the absolute value of t on the second partial regression coefficient that was shown earlier for the two-predictor analysis.

Correspondingly, the p-value on this t is exactly the same as the p-value for the F-statistic computed here.

Note that neither the first nor the second partial regression coefficient is statistically significant (if α is set at .05) because the p-values for these coefficients are .87 and .07, respectively. Remember that this does not mean the two predictors are useless or that both of them should be removed from the model. After all, two other previously described types of analysis conclude that the predictors are related to heart disease. First, the ANOVAR F-test based on the two-predictor model convincingly demonstrates that the combination of both predictors is more effective than not using the set of predictors in the model (the p-value is .014). Second, the two simple regression analyses demonstrate that each predictor used alone is useful in predicting heart disease (the p-values are .002 and .02). Clearly, the partial regression coefficients and the associated tests are telling us something different. These tests tell us that (1) the first predictor (stress) does not make a statistically significant improvement in prediction when it is added to the model using only X_2 (obesity) as the predictor, and (2) X_2 (obesity) does not make a statistically significant improvement in prediction when it is added to the model using only X_1 (stress) as the predictor. Each of these analyses provides information that differs from that provided by the ANOVAR F-test; they cannot contradict this test because different hypotheses are involved.

Multiple Correlation Coefficient: Two-Predictor Case

Although I have specified the null hypothesis associated with the ANOVAR F-test as $H_0 : \beta_1 = \beta_2 = 0$, there is an equivalent null hypothesis that is based on a correlational framework. Just as the simple correlation coefficient (r_{yx}) is a measure of the degree of linear relationship between variables Y and a single variable X, the multiple correlation coefficient (R) is used when there are multiple predictors. R is interpreted as the correlation between Y and the optimum linear combination of the predictors. (Subscripts are sometimes added to R to specify the predictors.) The optimum linear combination is defined by the fitted regression equation. So in the case of two predictor variables, the simple correlation computed between Y and $\hat{Y}$ (where $\hat{Y}_i = b_0 + b_1 X_{1i} + b_2 X_{2i}$) is equal to R_{yx_1,x_2}. Alternatively, R_{yx_1,x_2} is often computed as $\sqrt{\frac{\text{SS}_{\text{Multiple regression}}}{\text{SS}_{\text{Total}}}}$. Both R and its square (R^2) are widely used descriptive measures that are helpful in the interpretation of regression results.

The ANOVAR F in the case of two predictors is a test of the significance of R_{yx_1,x_2}. The null hypothesis in this case is written as $H_0 : \rho_{yx_1,x_2} = 0$. (In other words, the population multiple correlation coefficient is equal to zero.) If this null hypothesis is rejected, it is concluded that the population multiple correlation is not zero. This turns out to be equivalent to concluding that the model containing two predictors (model I) explains more variation than the model without predictors (model II). Hence, the ANOVAR F-statistic tests both $H_0 : \beta_1 = \beta_2 = 0$ and $H_0 : \rho_{yx_1,x_2} = 0$. If both of the population partial regression coefficients are equal to zero, then the population multiple correlation coefficient must also be equal to zero.

The square of the multiple correlation coefficient (R^2) is often called the coefficient of multiple determination. This is a very useful descriptive statistic that has been referred to earlier in this chapter and will be used often in subsequent chapters. Recall that it is interpreted as the proportion of the total variation on Y that is explained by the optimum linear combination of the predictors. It can be computed as the ratio of the multiple regression sum of squares over the total sum of squares; these sum of squares appear in the ANOVAR summary table associated with multiple regression problems. (Computer output does not include the term "multiple" before the term "regression" on output listing the ANOVAR table; it is supposed to be understood that regression sum of squares is based on multiple predictors whenever the fitted model has two or more predictors.) Note in the summary table for the example data (above) that the multiple regression sum of squares is 142.863 and the total sum of squares is 202.900. The ratio is $.704 = R^2$. Also note that the *Minitab* output displays R^2 expressed as a percentage. Hence, 70% of the sample variation on heart disease is explained by the optimum linear combination of the psychological stress and obesity measures.

An additional coefficient displayed in *Minitab* regression output to the right of R^2 is the adjusted R^2. As mentioned previously, this statistic exists because it is known that R^2 is a positively biased estimator of the population coefficient of determination. The adjusted value is a less biased estimate of the unknown population parameter $\rho^2_{yx_1,x_2}$ than is the unadjusted R^2-value. When large samples are used, there will be little bias and little difference between the unadjusted and adjusted R^2-values. In the present small N example there is a substantial difference; the adjusted $R^2 = .62$ whereas the unadjusted value is .70. Although the unadjusted statistic is the more biased estimator, it provides an appropriate description of what has been found in the sample. Additional detail regarding the computation and properties of unadjusted and adjusted coefficients is described in Section 5.3.

Partial Correlation Coefficients

A common concern in the interpretation of research based on correlational designs is the so-called third variable problem. Consider the example data. Suppose the researcher's major interest is in describing the degree of linear relationship between heart disease and psychological stress. The simple correlation found between these variables is .72; this is an interesting finding. But the researcher may suspect that this relationship is confounded by obesity. She may argue that the correlation is high because both heart disease and psychological stress are likely to be caused by obesity (the third variable). In this case, she may decide that the most obvious and direct way to clarify the relationship between heart disease and stress is to carry out an additional study in which she will compute the correlation between these two variables on a large sample consisting of participants who have exactly the same score on obesity. This is a good idea because she could then convincingly argue that variation on obesity has been directly controlled in the design by selection of a sample with essentially no variation on obesity.

But the implementation of this strategy requires that she (1) selects a very large initial sample of participants, (2) obtains an obesity score from each participant, (3) selects a subsample of participants (from the initial sample) who have exactly the same obesity score, and (4) computes the correlation on the subsample. The correlation computed between heart disease and psychological stress would then appear to be free of confounding by obesity because the observed obesity scores would not differ from one participant to another in the subsample. A major practical problem with this approach to controlling for a confounding third variable is that it may require more resources than are available. A very large initial sample may be required in order to identify a homogeneous subsample of adequate size. When this approach is not feasible an alternative is to use statistical control.

The First Partial Correlation Coefficient: $r_{yx_1 \cdot x_2}$

The partial correlation coefficient is a statistical attempt to control for the third variable problem; it does not require the collection of data on additional subjects. Recall that the first partial regression coefficient in a two-predictor multiple regression analysis is actually the slope computed on two variables (Y and X_1) after they have been residualized with respect to X_2. If we correlate the two sets of residuals (using simple Pearson correlation), the result is described as the "partial correlation between Y and X_1, holding constant X_2." This coefficient is denoted as $r_{yx_1 \cdot x_2}$, where the dot is followed by the variable that is statistically controlled. The hypothesis test for a partial *regression* coefficient applies to the corresponding partial *correlation* coefficient as well. Hence, there is no need to learn a new hypothesis test on partial correlation coefficients; we simply inspect the hypothesis test and p-value associated with the first partial regression coefficient in the two-predictor multiple regression analysis. The first partial correlation coefficient and associated p-value for the example data are $r_{yx_1 \cdot x_2} = -.06$ and $p = .87$.

Both the value of the simple correlation between heart disease and stress (.72) and the value of the partial correlation between those variables after the association with obesity is controlled (−.06) are of interest. Note that the relationship essentially disappears when controlling for obesity. Although this outcome is clear, the reason that controlling for obesity has eliminated the relationship is not. Some researchers believe that an outcome of this type proves that there is no causal association between X_1 and Y. This is only one possible explanation among many. If we are to claim that we understand the development of heart disease, it is important to know whether the role of obesity is as a confounder, a mediator, or as some other type of variable. Unfortunately, a correct "causal" interpretation depends upon knowledge we probably do not have. Some of the relevant issues can be clarified using causal diagrams to illustrate the nature of the variable used to control other variables in the analysis.

Inspect Figure 5.4. Panel A illustrates causal relationships among the three variables when X_2 (obesity) is a confounding variable. Under this causal scenario, it can be seen that obesity causes both X_1 (psychological stress) and Y (heart disease). Hence, the reason stress and heart disease are correlated is that obesity causes both of them. In this case, the partial correlation $r_{yx_1 \cdot x_2}$ is straightforward to interpret because the variable after the dot truly is the confounding variable.

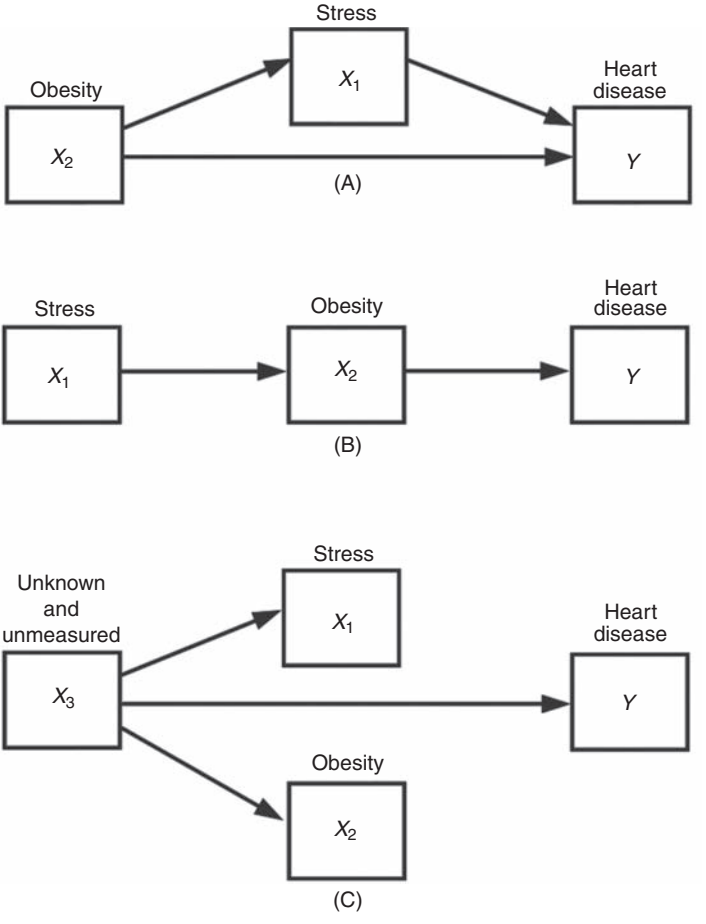


Figure 5.4 Illustration of X_2 playing three roles: as a confounder (Panel A), as a mediator (Panel B), and as a marker for a causal variable (Panel C).

Panel B illustrates a very different causal scenario. Here psychological stress causes obesity and then obesity causes heart disease. Hence, there is a causal chain beginning with X_1 (psychological stress), continuing through X_2 (obesity), and ending with Y (heart disease). In a situation like this it is common to label X_2 as a mediator of the effects of X_1 in causing Y; X_2 is viewed as the mechanism through which X_1 operates in causing Y. In this case, obesity is clearly not a confounding variable, even though the pattern of the correlations among the variables is consistent with the pattern found with confounding. Unfortunately, the data analysis alone does not clearly identify the role of obesity. It may be a confounding variable, a mediating variable, a marker for some additional variable(s) that causes both stress and obesity, or it may have some other role. Scenarios A (confounding) and B (mediating) are both expected to have low values of $r_{yx_1 \cdot x_2}$ and relatively high values of r_{yx_1}, $r_{x_1x_2}$,

and r_{yx_2}. Elaborate statistical methods have been proposed to shed light on issues of this type, but they are usually less than convincing.

A third scenario is illustrated in Panel C. Here the actual variable causing heart disease (X_3) is unknown and unmeasured, but both obesity and stress are markers for this variable. This means that these measured predictors are correlated with each other, although they do not cause each other. They are also correlated with heart disease, but the partial correlation $r_{yx_1 \cdot x_2}$ is low. Once again, if the researcher observes that r_{yx_1}, $r_{x_1x_2}$, and r_{yx_2} are high and the value of $r_{yx_1 \cdot x_2}$ is very low, she will probably conclude that obesity is the causal variable.

The point of the diagrams in Figure 5.4 is that the finding of a high value for r_{yx_1} and a low value for $r_{yx_1 \cdot x_2}$ may support for the interpretation that X_2 is a confounding variable, but it cannot be counted on. The potential causal scenarios are not limited to those illustrated in the figure; for example, heart disease may cause both obesity and stress, or heart disease and stress (or obesity) may be reciprocal. Popular contemporary statistical methods (viz., structural equation models) may be helpful for casting doubt on the plausibility of certain postulated causal scenarios, but this is unlikely without thorough knowledge of the properties of the variables under study and a well-supported causal theory. Better clarification is likely to be provided by longitudinal experiments that measure change across time.

Unfortunately, there often is no basis upon which to decide the causal role of the various predictors in the model. In this case, the researcher should acknowledge the fact that the model is not explanatory in a causal sense and that it does not necessarily represent a contribution to understanding. In some applications of regression analysis the major goal is not to understand. Rather, the focus may be the prediction of an outcome. It is not essential to identify the role of the predictor variables in this case, although it is helpful to understand the nature of the variables even in this case.

The Second Partial Correlation Coefficient: $r_{yx_2 \cdot x_1}$

Just as the first partial correlation coefficient ($r_{yx_1 \cdot x_2}$) estimates the simple correlation between Y and X_1 after controlling for X_2, the second partial correlation coefficient ($r_{yx_2 \cdot x_1}$) estimates the simple correlation between Y and X_2 after controlling for X_1. The same approach to understanding this coefficient can be applied. That is, it can be interpreted as the correlation between two columns of residuals (viz., Y residualized by X_1, and X_2 residualized by X_1). The residuals required for both partial correlations are displayed in Table 5.2.

The intercorrelations among these four columns are shown below:

```
        RESI1  RESI2  RESI3
RESI2  -0.063
RESI3  -0.873  0.055
RESI4  -0.548  0.811  0.627
```

The correlation between RESI1 and RESI2 is equal to the first partial correlation $r_{yx_1 \cdot x_2}$ and the correlation between RESI3 and RESI4 is equal to the second partial correlation $r_{yx_2 \cdot x_1}$. (The other correlations are not of interest here.) Comparisons of the

Table 5.2 Residuals from Regressing X_1 on X_2, Y on X_2, X_2 on X_1, and Y on X_1

Row	RESI1	RESI2	RESI3	RESI4
1	−0.45798	−1.80880	−1.43292	−3.06681
2	0.01112	3.61672	−0.22433	3.42586
3	4.43086	−1.97599	−4.01575	−5.08148
4	1.85060	−2.56870	−1.05052	−3.33026
5	0.33215	−0.01245	−2.43292	−2.06681
6	−2.14940	−0.56870	2.81043	1.67463
7	−0.30992	1.24589	1.98425	2.91852
8	1.43086	4.02401	−1.12004	3.17219
9	−2.98888	1.61672	2.67137	3.67952
10	−2.14940	−3.56870	2.81043	−1.32537

The four columns of residuals in this table are defined as follows:
RESI1 = residuals from the regression of X_1 on X_2,
RESI2 = residuals from the regression of Y on X_2,
RESI3 = residuals from the regression of X_2 on X_1, and
RESI4 = residuals from the regression of Y on X_1.

previously computed simple correlations and the corresponding partial correlations between Y and each predictor are summarized below.

Simple correlations: $r_{yx_1} = .72$ and $r_{yx_2} = .84$.
Partial correlations: $r_{yx_1 \cdot x_2} = -.06$ and $r_{yx_2 \cdot x_1} = .63$.

Note that the first partial correlation is close to zero, but the second partial correlation is much larger. It may be of interest to carry out formal hypothesis tests on these coefficients. Although *Minitab* (and most other software packages) computes p-values for simple correlation coefficients, they have not been requested here because such values are incorrect when the variables are residuals. They do not acknowledge the additional source of sampling error introduced by the variable that has been held constant. Appropriate tests are described next.

Tests on Partial Correlation Coefficients

The hypothesis tests on partial correlation coefficients are identical to the tests on the corresponding partial regression coefficients. Because the conventional computer output for a two-predictor regression analysis contains tests for the partial regression coefficients, there is no need for additional tests for the partial correlation coefficients. If a partial regression coefficient is equal to zero, the corresponding partial regression coefficient must also be equal to zero. If a partial regression coefficient is statistically significant, the corresponding partial correlation coefficient must also be statistically significant and have the same sign.

Semipartial (Part) Correlations: $r_{y(x_1 \cdot x_2)}$ and $r_{y(x_2 \cdot x_1)}$

An additional type of coefficient that is sometimes of interest in a two-predictor problem is known the semipartial correlation or part correlation. Just as there are

two simple correlations and two partial correlations in two-predictor models, there are also two part correlations. These coefficients are denoted as follows: $r_{y(x_1 \cdot x_2)}$ and $r_{y(x_2 \cdot x_1)}$. The first coefficient is the correlation between the original Y variable and the residuals in X_1 that remain after regressing X_1 on X_2. Unlike the first partial correlation coefficient, which is the correlation between two residualized variables, the part correlation is the correlation between Y and one residualized predictor variable. The subscript indicates that the residualizing operation is applied to only the variable in parentheses in front of the dot; because the Y variable is outside the parentheses this indicates that it has not been residualized.

In the case of the example data, the correlation between the heart disease variable and the variable labeled RESI1 (see Table 5.2) produces the first part correlation, $r_{y(x_1 \cdot x_2)}$. The value of this coefficient is $-.034$; it describes the degree of linear relationship between heart disease and the aspect of the stress variable that is linearly independent of scores on the obesity variable. Correspondingly, the coefficient $r_{y(x_2 \cdot x_1)}$ can be computed by correlating the heart disease variable and the variable labeled RESI3. The value of this coefficient is .438; it describes the degree of linear relationship between the heart disease variable and the aspect of the obesity variable that is linearly independent of scores on the stress variable.

In most cases, part correlations are squared. In this case, they describe the proportion of the total variation on Y that is uniquely explained by one of the predictor variables beyond that which is explained by the other predictor used alone. This idea can be re-expressed as the following ratio: $\frac{\text{SS}_{\text{Unique contribution of } X_i}}{\text{SS}_{\text{Total}}}$.

For the example data, $r^2_{y(x_1 \cdot x_2)} = \frac{\text{SS}_{\text{Unique contribution of } X_1}}{\text{SS}_{\text{Total}}} = \frac{.233}{202.9} = .001$ is the proportion of the total variation on heart disease that is uniquely explained by psychological stress. The hypothesis test for the first partial regression coefficient is also a test of $H_0{:}\rho^2_{y(x_1 \cdot x_2)} = 0$. The p-value associated with this test is .87; hence, there is no convincing evidence that variation on heart disease in the population is explained by psychological stress, beyond that which is explained by obesity.

The same approach applies to the computation of the second part correlation and its square. That is, $r^2_{y(x_2 \cdot x_1)} = \frac{\text{SS}_{\text{Unique contribution of } X_2}}{\text{SS}_{\text{Total}}} = \frac{38.973}{202.9} = .192$. The hypothesis test for the second partial regression coefficient is also the test for the second semipartial correlation coefficient and its square. The null hypothesis in this case may be written as $H_0{:}\rho^2_{y(x_2 \cdot x_1)} = 0$. The p-value associated with the test of this hypothesis is .07; if α had been set at .05, it would be stated there is insufficient evidence to claim a nonzero correlation in the population between heart disease and obesity residualized with respect to psychological stress.

Relevance of Squared Part Correlation to R^2

Because a squared part correlation describes the proportion of the total variation that is explained by one predictor beyond that which is explained by the other predictor used alone, two equalities exist: $R^2_{yx_1x_2} = r^2_{yx_1} + r^2_{y(x_2 \cdot x_1)}$ and $R^2_{yx_1x_2} = r^2_{yx_2} + r^2_{y(x_1 \cdot x_2)}$. The squared coefficients for the example data are $R^2_{yx_1x_2} = .704$, $r^2_{yx_1} = .512$, $r^2_{y(x_2 \cdot x_1)} = .192$, $r^2_{yx_2} = .703$, and $r^2_{y(x_1 \cdot x_2)} = .001$. It can be seen that the coefficient of multiple determination (.704) is equal to $.512 + .192$. Alternatively, it is equal to $.703 + .001$.

5.3 GENERAL MULTIPLE LINEAR REGRESSION: m PREDICTORS

General Linear Regression Model

The general linear regression model has m predictors. It is usually written as follows:

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \cdots + \beta_m X_{mi} + \varepsilon_i.$$

The embellished form is

$$Y_i = \beta_{0|X_1,X_2,\ldots,X_m} + \beta_{1|X_2,\ldots,X_m} X_{1i} + \beta_{2|X_1,X_3\ldots,X_m} X_{2i} + \cdots + \beta_{m|X_1,X_2\ldots,X_{m-1}} X_{mi} + \varepsilon_i.$$

The purpose of the additional notation in the embellished form is to make clear the meaning of the coefficients. Each coefficient is conditional on the variables that appear after the vertical line in the subscripts. For example, if $m = 1$, this model reduces to the simple regression model. In this case, the intercept is conditional on the predictor being in the model and the slope is conditional on no other predictors. When $m \geq 2$, each coefficient is conditional on different predictor variables. For example, if three predictors are in the model, the intercept is conditional on all three, the first partial regression coefficient is conditional on the second and third predictors, the second partial regression coefficient is conditional on the first and third predictors, and the third partial regression coefficient is conditional on the first and second predictors.

Fitted Equation

As with simple regression, the fitted sample equation is the initial descriptive result of interest in a multiple regression analysis. This equation is usually written as follows:

$$\hat{Y}_i = b_0 + b_1 X_{1i} + b_2 X_{2i} + \cdots + b_m X_{mi},$$

where

- $\hat{Y}_i$ is the predicted dependent variable score for individual i;
- b_0 is the intercept; the estimated elevation on Y when X_1 through X_m are set equal to zero;
- b_1 is the first partial regression coefficient; the average increase on Y associated with a one unit increase on X_1, holding constant $X_2, X_3, \ldots, X_m$;
- b_2 is the second partial regression coefficient; the average increase on Y associated with a one unit increase on X_2, holding constant $X_1, X_3, \ldots, X_m$,
- b_m is the mth partial regression coefficient; the average increase on Y associated with a one unit increase on X_m, holding constant $X_1, X_2, \ldots, X_{m-1}$;
- $X_{1i}, X_{2i}, \ldots, X_{mi}$ are the predictor scores for the ith individual in the sample; and
- m is the number of predictor variables (not including the intercept column).

It can be seen that this is a straightforward extension of the two-predictor equation.

Overview of Parameter Estimation

The parameters of the general multiple regression model are estimated using the method of ordinary least-squares (OLS). This method produces estimates of the model coefficients that meet the least-squares criterion; recall that this is the same criterion used in the estimation of the coefficients of the simple and two-predictor regression models. Hence, in the case of m predictor variables, no set of estimates of the parameters of the model can be found that will yield a smaller residual sum of squares than will the least-squares estimates.

A brief overview of the computation involved to produce the OLS estimates is presented below. This is provided for the interested reader who is (1) curious about how regression software works and (2) has been exposed to matrix algebra. Otherwise, skip down to the subsection on the interpretation of partial regression coefficients.

Let $\mathbf{X}$ be a design matrix where the first column consists of N ones and subsequent columns contain N scores on predictor variables 1 through m. That is,

$$\mathbf{X} = \begin{bmatrix} 1 & X_{1,1} & X_{2,1} & \dots & X_{m,1} \\ 1 & X_{1,2} & X_{2,2} & \dots & X_{m,2} \\ \cdot & \cdot & \cdot & \dots & \cdot \\ \cdot & \cdot & \cdot & \dots & \cdot \\ \cdot & \cdot & \cdot & \dots & \cdot \\ 1 & X_{1,N} & X_{2,N} & \dots & X_{m,N} \end{bmatrix}.$$

Let $\mathbf{Y}$ be a vector of N scores on the dependent variable, i.e.,

$$\mathbf{Y} = \begin{bmatrix} Y_1 \\ Y_2 \\ \cdot \\ \cdot \\ \cdot \\ Y_N \end{bmatrix}.$$

The vector of least-squares estimates can be produced using the following expression:

$$(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{Y} = \begin{bmatrix} b_0 \\ b_1 \\ b_2 \\ \cdot \\ \cdot \\ \cdot \\ b_m \end{bmatrix},$$

where $\mathbf{X^T}$ is the transpose of the design matrix and the superscript $^{-1}$ indicates the inverse. The least-squares estimates $b_0, b_1, \dots, b_m$ are the coefficients in the fitted multiple regression equation.

Computations for Example Data

$$\mathbf{X} = \begin{bmatrix} 1 & 19 & 43 \\ 1 & 25 & 50 \\ 1 & 31 & 52 \\ 1 & 30 & 54 \\ 1 & 19 & 42 \\ 1 & 26 & 54 \\ 1 & 31 & 58 \\ 1 & 28 & 52 \\ 1 & 22 & 50 \\ 1 & 26 & 54 \end{bmatrix} \quad \mathbf{Y} = \begin{bmatrix} 12 \\ 23 \\ 19 \\ 20 \\ 13 \\ 22 \\ 27 \\ 25 \\ 21 \\ 19 \end{bmatrix}$$

$$\mathbf{X^T} = \begin{bmatrix} 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 \\ 19 & 25 & 31 & 30 & 19 & 26 & 31 & 28 & 22 & 26 \\ 43 & 50 & 52 & 54 & 42 & 54 & 58 & 52 & 50 & 54 \end{bmatrix}$$

$$(\mathbf{X^T X}) = \begin{bmatrix} 10 & 257 & 509 \\ 257 & 6789 & 13259 \\ 509 & 13259 & 26133 \end{bmatrix}$$

$$(\mathbf{X^T X})^{-1} = \begin{bmatrix} 16.4434 & .332255 & -.488847 \\ .3323 & .022886 & -.018083 \\ -.4888 & -.018083 & .018734 \end{bmatrix}$$

$$\mathbf{X^T Y} = \begin{bmatrix} 201 \\ 5304 \\ 10410 \end{bmatrix}$$

$$(\mathbf{X^T X})^{-1}\mathbf{X^T Y} = \begin{bmatrix} -21.5016 \\ -.0735 \\ .8544 \end{bmatrix} = \begin{bmatrix} b_0 \\ b_1 \\ b_2 \end{bmatrix}.$$

Interpretation of Partial Regression Coefficients when $m \geq 3$

The interpretation of the partial regression coefficients in the case of $m \geq 3$ is similar to the interpretation with two predictors, but now each coefficient is conditional on all of the remaining predictors in the model. For example, if there are five predictors, the first partial regression coefficient is interpreted as the average increase on Y associated with a one-unit increase on X_1, holding constant predictors X_2 through X_5. Similarly, the second partial regression coefficient is interpreted as the average increase on Y associated with a one-unit increase on X_2, holding constant predictors X_1 and X_3 through X_5. The same pattern continues for all remaining partial regression coefficients.

To confirm this interpretation, we could go through the following steps. First, regress Y on X_2 through X_5, and save the residuals. Second, regress X_1 on X_2 through

X_5, and save the residuals of this second regression. Because we have now residualized both Y and X_1 with respect to X_2 through X_5, we can see that both sets of residuals are linearly independent of X_2 through X_5. Therefore, there is no variation in either the first or the second set of residuals that is predictable using a linear function of X_2 through X_5. This means that if we were to regress the residualized Y on the residualized X_1 using simple regression, the slope would equal the first partial regression coefficient in the five-predictor model. Regardless of the number of predictors in the multiple regression model, each partial regression coefficient can be conceptualized as a simple slope computed on two variables that have been residualized with respect to all but one of the predictors in the model.

ANOVAR: *m* Predictors

When more than two predictors are included in the model, the null hypothesis is similar to the one described previously for the two-predictor case, but it includes m partial regression parameters rather than two. That is, H_0: $\beta_1 = \beta_2 = \cdots = \beta_m = 0$. This null hypothesis states that all of the population partial regression coefficients are equal to zero. If this hypothesis is rejected, it is concluded that at least one of them does not equal zero. The null hypothesis can also be stated in correlational form as follows: H_0: $\rho_{yx_1,x_2,\ldots,x_m} = 0$. The form of the ANOVAR summary that is used in the case of m predictors is shown in Table 5.3.

The defining formulas for the various sum of squares in this table are as follows:

$$\text{Regression sum of squares} = \text{SS}_{\text{Reg}} = \sum_{i=1}^{N} \left(\hat{Y}_i - \bar{Y}\right)^2$$

$$\text{Residual sum of squares} = \sum_{i=1}^{N} \left(Y_i - \hat{Y}_i\right)^2$$

$$\text{Total sum of squares} = \text{SS}_{\text{T}} = \sum_{i=1}^{N} \left(Y_i - \bar{Y}\right)^2,$$

where $\hat{Y}_i = b_0 + b_1 X_{1i} + b_2 X_{2i} + \cdots + b_m X_{mi}$.

The obtained value of F is evaluated using the F-distribution with m and $N - m - 1$ degrees of freedom.

Table 5.3 Form of the ANOVAR Summary Table for Multiple Linear Regression Models with *m* Predictors

Source	SS	*df*	MS	*F*
Regression	SS_{Reg}	$df_{\text{Reg}} = m$	MS_{Reg}	$\text{MS}_{\text{Reg}}/\text{MS}_{\text{Res}}$
Residual	SS_{Res}	$df_{\text{Res}} = N - m - 1$	MS_{Res}	
Total	SS_{T}	$df_{\text{T}} = N - 1$		

Rationale for the ANOVAR F-test

As with simple regression, the rationale for the test is based on the expectations of the regression and residual mean squares. Specifically, the expected values are

$$E(\text{MS}_{\text{Reg}}) = \sigma^2 + \boldsymbol{\beta}^{\mathbf{T}}\mathbf{x}^{\mathbf{T}}\mathbf{x}\boldsymbol{\beta}$$
$$E(\text{MS}_{\text{Res}}) = \sigma^2,$$

where $\boldsymbol{\beta}$ is the vector of m population partial regression parameters, $\mathbf{x}$ is the centered (i.e., deviation score) design matrix not including the unity column, and $\mathbf{T}$ is the transpose. If at least one of the population partial regression coefficients differs from zero, the expected value for the regression mean square exceeds the expected value for the residual mean square. Because the ratio of the mean square regression over the mean square residual defines the obtained F, it can be seen that, in general, F is inflated by nonzero values of the partial regression coefficients. Large observed values of F are usually a reflection of nonzero elements in the vector of population partial regression parameters.

R, R^2, and Adjusted R^2

In the general case of m predictor variables the notation for the sample multiple correlation coefficient is $R_{yx_1,x_2,\ldots,x_m}$; this coefficient may be defined as the correlation between Y and the optimum linear combination of predictor variables X_1 through X_m. The corresponding R^2-statistic describes the proportion of the sample variation on the dependent variable that is explained by the optimum linear combination of the set of m predictor variables. The fitted equation defines the linear combination; it is optimum in the sense that it meets the least-squares criterion. $R^2_{yx_1,x_2,\ldots,x_m}$ is a positively biased estimator of the corresponding population parameter $\rho^2_{yx_1,x_2,\ldots,x_m}$; the degree of bias is a function of the number of predictors (m) and sample size (N).

If there is no linear relationship (in the population) between Y and the predictors this implies that $\beta_1 = \beta_2 = \cdots = \beta_m = 0$ and equivalently $\rho^2_{yx_1,x_2,\ldots,x_m} = 0$. In this case, the expected value of the sample coefficient R^2 is *not* zero. Instead, the expectation is $E(R^2) = \frac{m}{N-1}$. For example, if there are 10 predictors and 21 subjects the expected value of $R^2 = .50$ (and the expected R is approximately .71) even though the population $\rho^2_{yx_1,x_2,\ldots,x_m} = 0$. The adjusted R^2 is used in place of the unadjusted R^2 when there is interest in estimating the population value.

The adjusted $R^2 = 1 - \frac{(1-R^2)(N-1)}{N-m-1}$. The adjusted value is always lower than the unadjusted value, but the difference between them will be trivial when m is small and N is large. The application of this formula to the example in the previous paragraph where N is very small and m is large (i.e., $N = 21$ and $m = 10$) yields $1 - \frac{(1-.50)(21-1)}{N-10-1} = 0$. In this instance, it can be seen that the value of the adjusted R^2 is equal to the value of the unadjusted R^2 minus the expected value under the null hypothesis. That is, $(R^2 - \frac{m}{N-1}) = (.50 - \frac{10}{21-1}) = 0$, which is exactly the value of the parameter $\rho^2_{yx_1,x_2,\ldots,x_{10}}$ in this contrived example. Of course, the adjusted estimator

does not always provide an estimate that equals the value of the underlying parameter, but it can be expected to provide a less biased estimate than is provided by the unadjusted estimator.

In summary, both unadjusted and adjusted coefficients are useful. If the major interest is in simply describing what has been found in a particular sample, the unadjusted value is appropriate. The unadjusted values are often used in the computation of more complex statistics that are a function of two different unadjusted R^2s. The adjusted R^2 is appropriate when the major interest is in estimating the value of the population coefficient $\rho^2_{yx_1,x_2,\ldots,x_m}$.

Standard Error of Estimate when $m \geq 3$

Recall that in the case of one predictor variable the standard error of estimate (usually denoted as $\sigma_{Y|X}$) refers to the standard deviation of Y around the regression line. Because the elevation of the regression line associated with a specific point on X is equivalent to $\hat{Y}$, the deviation around the regression line is defined as $(Y - \hat{Y})$. This interpretation also holds in the case of multiple regression, but $\hat{Y}$ is then based on two or more predictors. When two predictors are involved the standard error of estimate is denoted as $\sigma_{Y|X_1,X_2}$; this refers to the standard deviation of Y around the regression plane. When three or more predictor variables are involved the standard error of estimate is denoted as $\sigma_{Y|X_1,X_2,\ldots,X_m}$; in this case it refers to the standard deviation of Y around the regression hyperplane. Regardless of the number of predictors, $(Y - \hat{Y})$ is the residual e and the sample standard error of estimate provides an estimate of the standard deviation of the population errors ε. Hence, a more general way of denoting the standard error of estimate, regardless of the number of predictor variables, is to use σ_ε for the population parameter and s_e for the sample statistic. Regression software such as *Minitab* simply uses s with no subscript; unfortunately this is often misinterpreted as the standard deviation of the raw scores because s is the notation usually used for this purpose. Regardless of the number of predictors in the model, the standard error of estimate can be estimated using $\sqrt{\text{MS}_{\text{Res}}}$, where MS_{Res} is the mean square residual shown in the ANOVAR summary table.

Assumptions

The assumptions associated with the general linear regression model are essentially the same as in the case of simple regression. The errors (ε_i) of the model are assumed to be independent of each other, the mean of the error distribution associated with each combination of predictor variable scores is assumed to be normal, and the mean of each of these error distributions is assumed to be zero.

The assumption that each error distribution has a mean of zero is often misunderstood to mean that the average of all the residuals in the sample is zero. The mean of all of the residuals in the sample is always zero, but this is not the assumption. The assumption refers to the individual error distributions associated with each combination of predictor variables. For example, in a three-predictor problem, suppose that three predictor scores associated with case i are 17, 26, and 152. Conceptualize

a whole subpopulation of cases, where each one has predictor scores 17, 26, and 152. The distribution of the subpopulation errors ε_i around $\hat{Y}$ for this particular set of predictor scores is assumed to be zero. Now consider other combinations of predictor scores that are associated with other cases. A separate error distribution can be conceptualized for each combination of predictor scores that appears in the population; each one is assumed to have a mean of zero. When this assumption is met it implies that a linear relationship exists between Y and $\hat{Y}$. When the means of the subpopulation error distributions are nonzero, this demonstrates that the functional form of the relationship between Y and $\hat{Y}$ is nonlinear.

Diagnostics

Plots of the data should be visually examined for departures from assumptions. The plots should include Y against each predictor, the residuals against the fitted values, each predictor against the other predictors, and either a univariate plot or a normal probability plot of the residuals. Two of these plots (based on the heart disease data described earlier) are shown in Figures 5.5 and 5.6. If available, three-dimensional plots of Y against two predictors at a time may also be helpful.

Because the sample is so small in this example, departures from the assumptions are difficult to identify unless they are very large. A crude evaluation of the homoscedasticity assumption involves splitting the residuals into two subgroups based on the upper and lower halves of the $\hat{Y}$ distribution and then comparing the variation

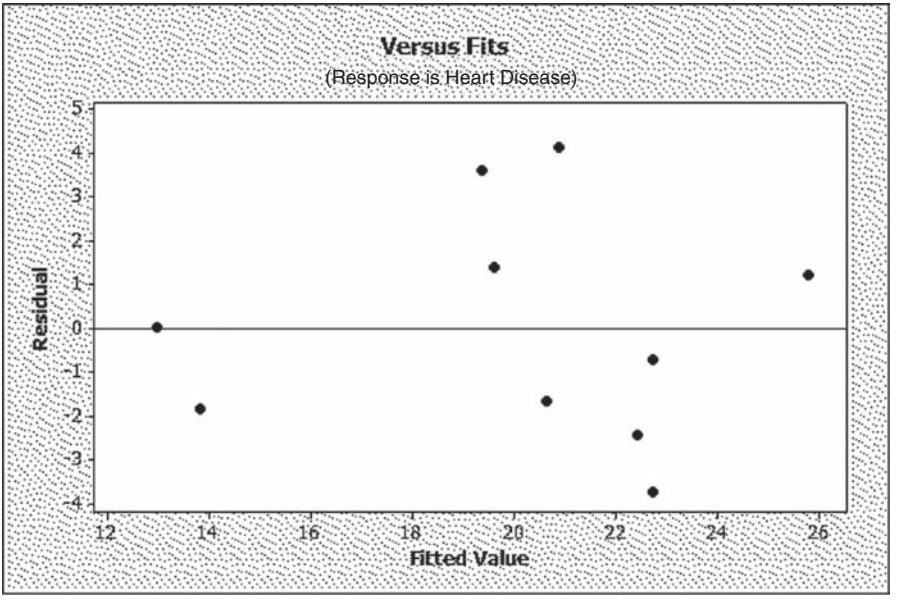


Figure 5.5 Scatterplot of heart disease residuals against the fitted values ($\hat{Y}$).

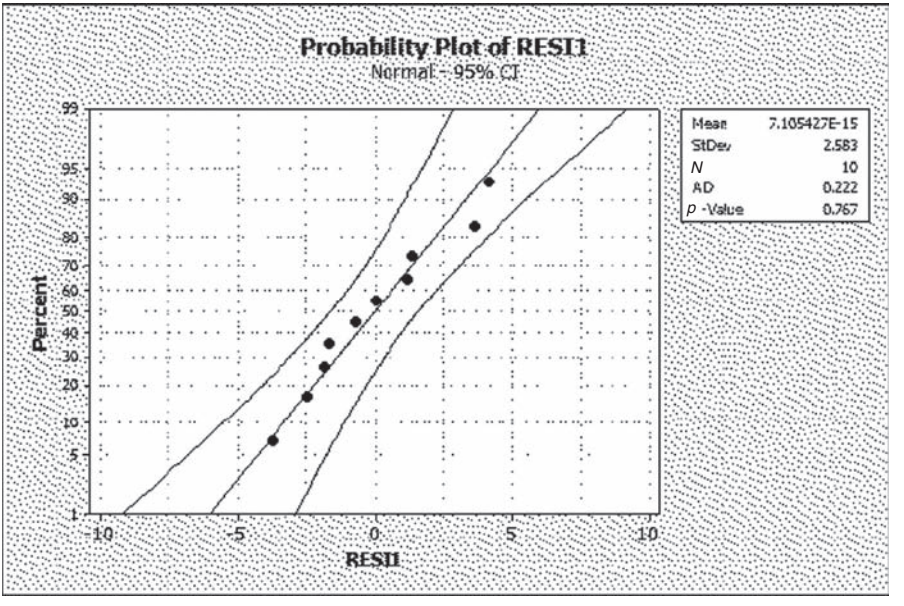


Figure 5.6 Normal probability plot of heart disease residuals.

of the residuals for these subgroups. This can be done easily through visual inspection of Figure 5.5; note that the variation of the residuals appears not to be greatly different for the two subgroups formed in this manner. (For those who insist on formal tests, the Bartlett and Levene tests of homogeneity of variance each yield a p-value of .56.) The normal probability plot illustrated in Figure 5.6 reveals no major departures from normality. Although many additional diagnostic methods exist, a careful inspection of the recommended plots is almost always adequate for identifying departures from the assumptions of the model.

Many of the formal diagnostic tests that are available in regression software appear to me to be analytic overkill, especially in evaluating homoscedasticity and normality assumptions. These tests are frequently insensitive with small samples (where departures from assumptions are often of greatest concern) and overly sensitive with large samples. There are, however, two situations in which formal diagnostic testing is often helpful.

First, it is useful to know when one or a few observations have an unusually strong influence on the regression coefficients. A statistic known as Cook's distance (often denoted as D_i) is available in most software packages. The details of this statistic are not described here (they can be found in any modern regression analysis textbook). The D_i-statistic considers the difference between the set of fitted regression coefficients based on the complete sample and the set of fitted regression coefficients based on the complete sample minus one case. If the coefficients in the two sets of estimates are the same, the value of $D_i = 0$. Large values of D_i are associated

with large differences between the two sets of estimates. There are N-values of D_i in the analysis; one value is associated with each case. Each value is compared with the critical value of F using $\alpha = .50$ (not .05) and $df = m + 1$ and $N - m - 1$. If the value of D_i exceeds the critical F, it is concluded that case i has a large influence on the fitted equation based on all N subjects. This does not necessarily mean that there is something wrong with the equation, but it does suggest that the data should be carefully inspected to be sure that the predictor and outcome values recorded for case i are credible. Occasionally, a large value of D_i will identify a subject for whom the data were incorrectly entered or a subject who does not belong to the population of interest.

A second situation in which formal diagnostic tests are useful occurs when analyzing data that were collected from a single sampling unit at many points in time (rather than from many sampling units measured only once). A formal test of the independence of error assumption is useful here. Because dependency of errors occurs for many different reasons, a test for detecting it should be viewed as a general diagnostic method for model misspecification in time-series designs. Tests of this type are described in Chapter 18.

Remedial Procedures

Recall that the major reason for carefully inspecting residuals and applying assumption tests is to help decide whether the fitted model is apt. If a major departure from the assumptions is identified, it is likely that one (or more) of the following approaches will lead to a more appropriate analysis: (1) a simple transformation of the original data, (2) a more complex OLS model, or (3) a complex model that cannot be estimated using OLS. Brief descriptions of the first two of these approaches are in the remainder of this section; the third approach is described in the subsequent section.

Transformations for Nonlinear but Monotonic Functions

When the relationship between the predictor and dependent variables is not adequately described using a linear function, some remedial action is called for. It is likely that a simple transformation can be found for either X or Y (or both) that will linearize the relationship. In this case, the relationship is said to be intrinsically linear; this implies that simple linear regression analysis applied to appropriately transformed variables results in an adequate analysis. But this approach only works when the form of the nonlinearity is monotonic.

A distinguishing characteristic of a monotonic function is that it never changes direction. A monotonic increasing relationship is one in which increases on X are associated with increases on Y; the curve never changes direction so that increases on X are associated with decreases on Y. Similarly, a monotonic decreasing relationship is one in which increases on X are associated with decreases on Y; the curve never changes direction so that increases on X are associated with increases on Y. Hence, it is easy to identify a monotonic relationship in a plot of Y on X. If it changes direction, it is not monotonic.

Often a transformation that effectively removes nonlinearity also improves conformity with homoscedasticity and normality. The most common procedure for

identifying an appropriate transformation is trial and error. The most useful transformations are those in the family of so-called power transformations; this terminology refers to the power to which the original variable is raised. The most common transformations (e.g., square root, log, and reciprocal) can be expressed as a variable raised to a specific power. For example, if X is the variable to be transformed and all values are greater than one, we might try raising X to .5; in this case $= X^{.5} = \sqrt{X} =$ the square root transformation. Other powers of X yield other transformations. When the power is less than one, the upper end of the distribution is squeezed and positive skew is reduced; when the power is greater than one, the upper end of the distribution is spread out and negative skew is reduced. When Y is plotted against an appropriately transformed X (or when an appropriately transformed Y is plotted against X), the observed function will no longer be curved. A sophisticated computationally intensive approach for identifying the most appropriate power transformation is available in some software packages; it is known as the Box–Cox procedure. It turns out, however, that in most cases square root and log transformations are adequate.

It should be kept in mind that the motivation for transforming data is to linearize the relationship so that a simple linear model will fit. My preference is to first seek a transformation for the predictor, especially when the measurement procedure for the dependent variable is well established. When there is a strong tradition of using a specific untransformed response measure, a transformation may create confusion when presenting results to some audiences. Another problem with transforming Y may occur if the transformation is applied in cases where the untransformed data have homoscedastic errors; in this case the transformation may introduce heteroscedasticity. On the other hand, if the untransformed Y data are both heteroscedastic and skewed, a single transformation of Y may kill three birds with one stone; that is, it may linearize the relationship, remove heteroscedasticity, and eliminate skewness.

Polynominal Regression for Nonmonotonic Nonlinear Functions

Although most nonlinear relationships are monotonic in form, occasionally it will be found that the direction of the relationship changes with increases on X. For example, increases on X may be accompanied by increases on Y only until a certain point on X is reached; further increases on X are then accompanied by decreases on Y. Situations like this are easily identified visually in a plot of Y on X. Functions of this type are described as nonmonotonic. In this case, a simple transformation is of no use whatsoever.

An easy solution is to fit a curve to the data using a polynomial regression model. The general form of this model is $Y_i = \beta_0 + \beta_1 X_i + \beta_2 X_i^2 + \cdots + \beta_m X_i^m + \varepsilon_i$. Because curves with a single bend are most common, the quadratic version of the polynomial model is a very popular approach for fitting curves.

The quadratic model is written as follows: $Y_i = \beta_0 + \beta_1 X_i + \beta_2 X_i^2 + \varepsilon_i$.

The β_1 and β_2 parameters of this model are described as the linear and curvature coefficients, respectively. Surprisingly, this model (like all polynomial regression models) can be estimated using any multiple linear regression routine. This involves using X as the first "predictor" and X^2 as the second "predictor," although there is only

one actual predictor variable in the study. In a sense we pretend that X^2 is the second predictor and then estimate the model by simply regressing Y on variables X and X^2.

It may seem strange to state that a "linear model" can fit curves, but there is a straightforward explanation. Although it is often stated that the simple linear regression model is appropriate when the relationship between X and Y is linear, and inadequate when the relationship is nonlinear, statisticians use the general term "linear model" to refer to something other than the simple linear regression model. It does not refer to the functional form revealed in a bivariate plot of Y on X. Instead, "linear model" refers to the way the terms in the model are combined.

It can be seen above that the various terms in the quadratic model are simply added together to form a linear combination in a manner that is identical to what is done using a two-predictor model. So technically, a polynomial regression model is a linear model; it is said to be "linear in the parameters." This differs from another class of models for nonlinear data, which are known as "nonlinear regression models." An exponential regression model is an example of a nonlinear regression model; models of this type are not linear in the parameters. Exponential models will not be discussed in this chapter, but other types of nonlinear regression model are mentioned in the next section.

Occasionally a researcher uses polynomial regression in the case of monotonic nonlinearity instead of using a transformation. This is acceptable if the goal is simply to fit a curve that describes the data. But transformations have two distinct advantages with monotonic data. Suppose the researcher is interested in predicting Y for a new person (who was not in the sample used in fitting the equation) and this person has a value of X that lies outside the range of X included in the sample. The quadratic model should be avoided in situations like this; extrapolations often lead to inaccurate, counterintuitive, and substantively impossible predictions. Simple linear regression using a reasonably transformed variable will almost always provide far more accurate predicted values in this situation. The approach also has the advantage of fewer parameter estimates than are required in polynomial regression. This leads to both higher power and simpler interpretation.

5.4 ALTERNATIVES TO OLS REGRESSION

Although simple transformations and polynomial regression methods solve most assumption departures, there are situations in which this is not true. Sometimes heteroscedasticity cannot be transformed away, dependency of the errors exists for unknown reasons, or the errors are not approximately continuous. A brief overview of several methods used to solve these problems is presented in this section.

Weighted Least-Squares: Accommodating Heteroscedasticity

Transforming data to reduce heteroscedasticity is sometimes viewed as undesirable. A common alternative when the dependent variable is continuous is to instead fit the model using a procedure known as weighted least-squares. The general idea behind

this approach is to heavily weight observations falling at levels of X that are associated with low variability on Y and to assign moderate weight to those observations falling at levels of X having high variability on Y. The results based on this approach differ from those based on ordinary least-squares in that the value of the coefficients may be slightly different and the size of the standard error estimates for the coefficients are often much smaller.

Multilevel Analysis: Accommodating Dependency Among Errors Caused by Within-Group Similarities

Sometimes a regression analysis is carried out on data from participants who have been sampled from a large number of different intact populations. It is typical for subjects within groups to have substantial social interaction with each other and to not interact with participants from other groups. Suppose nurses are sampled from 40 cardiac units (each unit associated with a different hospital) and data are collected from each nurse regarding his/her degree of preference for a specific treatment procedure (Y). It is likely that the responses will be more similar within cardiac units than between them. Similarly, the regression of Y on one or more predictors (say, age, sex, and experience) is likely to produce residuals that are dependent. When the assumption of independent errors is violated, the hypothesis tests and confidence intervals are often invalid.

Rather than pooling the data from all the hospitals into one total sample, we could compute a separate regression analysis within each hospital. This should solve the dependency problem mentioned previously, but a more efficient approach known as multilevel analysis is sometimes preferable. Multilevel analyses represent a compromise between pooling the data and analyzing each group separately. If there are no between-group differences, the multilevel analysis is essentially equivalent to a conventional OLS analysis computed on pooled data. If there are large differences among groups, the multilevel analysis provides information similar to that provided by separate analyses for each group.

Multilevel analysis is useful when there is interest in (a) using a group-level measure (e.g., the mean level of experience for a group) to predict variation on a group-level dependent variable (e.g., the mean preference score), (b) predicting outcomes for individual subjects within groups, and (c) studying the interaction between a group-level predictor and an individual-level predictor.

This type of analysis has become very popular during the last decade, especially among researchers working in nonexperimental areas. The multilevel approach is not pursued in subsequent chapters because the major focus of this book is on experimental research. Multilevel models have little to add to the typical experiment that has only two or three groups.

Time-Series Regression: When the Errors Are Time Dependent

Data collected from a single sampling unit at many equally spaced time points is sometimes called a time series. If the observations from these designs are modeled

using regression procedures, the errors may be dependent. When this occurs the associated hypothesis tests and confidence intervals are likely to be invalid. The best solution to this problem is to include additional predictors that model the explanations for the dependency. But if appropriate predictors are unmeasured or unknown, this solution cannot be implemented. In this case, it may be necessary to expand the model to include one or more autoregressive parameters that describe dependency among the errors. Such a model is described as a regression model with autoregressive errors. Examples of time-series models of this type are described in Chapter 18.

Generalized Linear Models: When *Y* Is Not Approximately Continuous

It is not unusual for the dependent variable to be measured in such a manner that the observations are not approximately continuous. This problem always occurs when the dependent variable is dichotomous; it occurs often when the dependent variable is either the amount of time until an event occurs or a count of the occurrence of some type of event within a certain time period. Several forms of the generalized linear model that handle these types of response measure are mentioned below.

Logistic Regression: When the Dependent Variable Is Binary (Dichotomous)

Sometimes the dependent variable of interest consists of only two categories such as alive versus dead or diseased versus healthy. In this case, the two dependent variable categories are identified using scores of zero and one. The purpose of the analysis is to describe the probability of being a "one," given a score on one or more predictors. If conventional regression modeling is attempted in situations like this it will fail.

The residuals of a conventional regression analysis cannot conform approximately to all of the assumptions of the model when using a binary (dichotomous) response measure. If the dependent variable scores consist of only zeros and ones the distribution of such scores cannot be normal. Consequently, the errors of the model cannot be normally distributed and homoscedasticity is highly unlikely.

Further, a conventional regression analysis will lead to predicted scores that will be less than zero for some values of X and to values greater than one for other values of X. Of course, these predictions must be wrong because they are outside the range of possible probability values. If the purpose of the analysis is to provide the probability of a subject falling in category "one," an appropriate analytic procedure must provide predicted values that are bounded to fall in the interval 0–1. Several approaches have this property, but the most popular is the logistic regression model. It turns out that the probability of falling in category one is a nonlinear function of the predictor(s).

Because the logistic model is nonlinear, the parameters cannot be estimated using conventional multiple linear regression software; there is no closed-form solution. But efficient estimation routines can be found in all of the major software packages. Applications of logistic models are described in several subsequent chapters.

Ordinal Logistic Regression for Ordinal Categorical Dependent Variables

The response measure is sometimes a collection of ordered categories rather than an approximately continuous variable. For example, five disease stages may be conceptualized as ordered categories of disease severity. In this case, there may be no way to argue that the difference in severity between categories one and two is the same as the difference between categories two and three or any other pair of adjacent categories. If it cannot be argued that the difference in severity is similar between adjacent categories, it is questionable to analyze the outcome using a conventional linear model. In this case, an ordinal logistic regression model may be more appropriate.

On the other hand, the underlying dependent variable may be continuous and approximately normally distributed. If this type of variable is measured somewhat crudely using, say, five ordered categories, a conventional linear model is likely to be satisfactory. An important advantage of using a conventional linear model analysis in this case is that the results are much easier to communicate than are the results of an ordinal logistic regression. I recommend the use of a conventional linear model unless it is quite obvious that the underlying differences between various adjacent categories are large. Because "quite obvious" and "large" are subjective terms, there often will be ambiguity regarding the analytic choice. In this case, it is reasonable to carry out and report both analyses.

Survival Analysis: When the Dependent Variable Is the Time to an Event

There is often interest in studying predictors of the amount of time that passes before an important event occurs. This time period is often called the survival time, which is appropriate if the event is death. There is major interest in knowing which variables are predictive of when the event will occur. Conventional regression models are not usually satisfactory for analyzing data from such designs because survival time is likely to be "right-censored" for many subjects. That is, by the end of the data collection period we know which subjects have not experienced the event of interest, but we do not have information on how much time will pass before the event occurs. Hence, this information is said to be right-censored because the unknown time of the event occurs in the right-hand end of the time distribution of events. Survival analysis (also called event history analysis) methods are ideal for analyzing data of this type. The most popular method for performing a survival analysis is known as the Cox proportional hazards regression model; software for implementing it is widely available.

Negative Binomial Regression: When the Dependent Variable Consists of Counts

The dependent variable often consists of the number of times some event occurs within a specified time interval. For example, one might record the number of cigarettes smoked during a 1-week interval for each participant in a sample. Suppose the predictor is the amount of psychological stress experienced by each participant during the same 1-week interval. It may be found that the variance of the residuals is much higher for high stress participants than for low stress participants. This type of heteroscedasticity can often be satisfactorily dealt with using a square root

transformation on the counts. An alternative approach is to use the negative binomial regression model. Although a better known strategy is to use Poisson regression, this model unrealistically assumes that that the variance is the same as the mean. The negative binomial regression model more realistically models the variance as a quadratic function of the mean.

Tobit Regression Model: When a Substantial Proportion of the Y Distribution Is Censored

Occasionally the measurement method fails to discriminate adequately at the lower or upper end of the response continuum when a continuous dependent variable is involved. This can result in an observed distribution that has a large proportion of scores piled up at either the lowest or the highest point on the scale. One possible solution is known as tobit regression. This model is likely to provide better predictions than are provided by OLS estimation, but there are conditions under which OLS and other models (such as the probit model) may be preferred because of higher power, more adequate software, and fewer problems of interpretation. A useful reference on the tobit model is Long (1997).

5.5 SUMMARY

Multiple regression analysis involves the simultaneous use of two or more variables in predicting an outcome variable. It is used to clarify the predictive role of a single predictor variable in a set of predictor variables, to partially control for confounding, to serve as a general explanatory system, and to fit curves to data when the relationship between X and Y is not linear. The essential aspects of a multiple regression analysis include the fitted regression equation, tests on each of the parameters of the model, the estimate of the standard error of estimate, an analysis of variance for the full-regression model, and a measure of the proportion of the total variation on the dependent variable explained by the set of predictors included in the model.

Several additional measures of association are associated with multiple regression analyses. Among them are the partial correlation coefficient, the part correlation coefficient, and the multiple correlation coefficient. Partial correlation measures the degree of linear relationship between two variables after statistically controlling for one or more other variables. The squared part correlation describes the proportion of the total variation on the dependent variable explained by a specific variable above and beyond that which is explained by one or more other variables in the multiple regression model. The multiple correlation coefficient is a measure of the degree of linear relationship between the dependent variable and the optimum linear combination of all predictors in the model. Nonmonotonic curved relationships can be modeled using a minor variant of multiple regression analysis known as polynomial regression.

Several special purpose regression approaches are available when the assumptions of the conventional multiple regression model are seriously violated. Weighted

regression is useful when heteroscedasticity is present in models of continuous variables. Time-series regression models with autoregressive errors and multilevel models are appropriate for the analysis of designs that have dependent errors. Logistic regression is appropriate for many types of research that use binary dependent variables. Ordinal regression is available for studies that use a dependent variable that consists of ordered categories, and tobit regression may be useful when many of the dependent variable scores fall at the minimum or maximum points of the distribution.

PART III

Essentials of Simple and Multiple ANCOVA

CHAPTER 6

One-Factor Analysis of Covariance

6.1 INTRODUCTION

Similar to the analysis of variance, the analysis of covariance (ANCOVA) is used to test the null hypothesis that two or more population means are equal. ANCOVA always involves at least three variables: an independent variable, a dependent variable, and a covariate. The covariate is simply a variable likely to be correlated with the dependent variable. The main advantages of including the covariate in *randomized experiments* are: (1) generally greater power, (2) a reduction in bias caused by chance differences between groups that exist before experimental treatments are administered, and (3) conditionally unbiased estimates of treatment effects.

The power of the ANCOVA test is central in true experiments where the design involves random assignment of subjects to treatments. In this case the increase in power (relative to ANOVA) is the major payoff in using the analysis of covariance. At the same time, ANCOVA includes an adjustment of the treatment effect estimate; this adjustment reduces bias that may be caused by chance pretreatment differences between groups. This type of bias is generally small in randomized designs.

There are various types of covariates. Common types are *baseline measures* (i.e., pretest measures based on the same instrumentation used to measure the dependent variable) and variables other than baseline measures that are correlated with the dependent variable, including *organismic characteristics* (such as age, blood pressure, body size, and sex), and *environmental characteristics* (both physical and social). Covariates are often approximately continuously scaled, but other types of scales qualify as well.

In *observational studies* (i.e., designs in which random assignment is not employed to form treatment groups) the main motivation for applying ANCOVA is to reduce confounding between independent and dependent variables. Indeed, the intent here is to make comparisons of nonequivalent groups more fair by adjusting for potential confounding variables. Although observational studies are commonly analyzed using

The Analysis of Covariance and Alternatives: Statistical Methods for Experiments, Quasi-Experiments, and Single-Case Studies, Second Edition. Bradley E. Huitema.

ANCOVA (or an equivalent regression model), interpretation problems are likely and other approaches are sometimes preferable.

Conditional Versus Unconditional Inference

The motivation for the use of ANCOVA instead of ANOVA in the case of randomized experiments requires some explanation. Consider a small two-group randomized design where the independent variable consists of two types of drug, age (X) is a variable (called the covariate) that is measured before treatments are applied and is believed to be correlated with the dependent variable, and a measure of complex psychomotor performance is the dependent variable Y. It is known that random assignment will produce groups that are probabilistically equivalent (pretreatment) on all variables, including X and Y. Hence, in the absence of treatments the expected value on X and Y is the same for the two groups. This is simply saying that if the process of random assignment is performed an infinite number of times to form two groups and the mean for each group is computed for each randomization, the average of the two group means (over all randomizations) will be the same. So "in expectation" random assignment provides groups that are exactly equivalent.

It follows that the difference between sample means on Y (after treatment) is an unbiased estimate of the treatment effect (i.e., the difference between population means). The notion of an unbiased treatment effect refers to the long run expectation of the mean difference over an infinite number of randomizations. In this sense one can argue that sample mean differences on Y are unconditional treatment effect estimates.

But in the case of a single experiment the two groups will not be exactly the same on a continuous variable before treatments are applied. If we randomly assign our available pool of subjects to form two small experimental groups we should not be surprised to discover that the two group means on X (age) or any other variable are somewhat different only as a result of sampling error. Suppose the second sample mean on X is five points higher than the first sample mean. If the experiment is carried out using these samples the effect estimate (say, an eight-point difference on performance measure Y) is conditional on the five-point difference on X. After all, the experiment was carried out under the condition that the groups differed by five points on X. This means that the eight-point difference on Y is a *conditionally biased* estimate of the treatment effect because the groups were not exactly equivalent before treatment. This does not mean that an ANOVA F-test and the associated effect estimate are "wrong." But it is possible to do better.

ANCOVA incorporates the information available on the X variable and provides a *conditionally unbiased* (i.e., conditional on X) estimate of the treatment effect. Information regarding the 5-year difference between groups on X is acknowledged in the effect estimation procedure; the result is an estimate of the mean difference on Y that would have been obtained if the two sample means had been exactly the same on X before treatments were applied. Hence, in the case of a randomized-group

experiment the ANCOVA treatment effect estimate is said to be conditionally unbiased whereas the estimate provided using ANOVA is conditionally biased. The long run average of these two types of effect estimates are the same under random assignment, but the ANCOVA estimates are more precise. An explanation for why this is true is described next.

General Ideas Associated with ANCOVA

The major differences between ANOVA and ANCOVA can be illustrated fairly easily using a two-group experiment. Again consider the study evaluating the comparative effectiveness of two drugs on scores obtained on a complex psychomotor test. Subjects are randomly assigned to the two treatments, treatments are applied, and psychomotor test data are obtained. Suppose a statistically nonsignificant ANOVA F-value is obtained. The experimenter then points out that there is a great deal of age variability within the two groups and that age is well known to be highly related to scores on the psychomotor test used in the experiment. This suggests that some of the error variation in the ANOVA should not be viewed as unexplainable random variation, even though this is the way it is conceptualized in the ANOVA model. If ANCOVA is computed (using age as the covariate) the analysis will statistically control or, more correctly, partition the effect of the covariate measure from the relationship between the treatments and the dependent variable. In this study age is partitioned from the relationship between drug type and psychomotor test score. The ANCOVA F-test is more likely to identify a statistically significant treatment effect than is ANOVA. The reason for the contradictory results of the two analyses can most easily be understood by inspecting the error distributions associated with each analysis (see Figure 6.1).

The major distinction between the two analyses is that the ANOVA error term is based on variation of Y around individual group means, shown as distributions A and B, whereas the ANCOVA error term is based on variation of Y scores around pooled within group regression lines, distributions C and D. For example, the error for subject i (in group 1) is estimated as $\left(Y_{ij} - \bar{Y}_1\right)$ under the ANOVA model whereas the error in the ANCOVA model is estimated using $\left(Y_{ij} - \hat{Y}_{ij}\right)$. The elevation of the regression line illustrated for group 1 at the X-value associated with subject i is $\hat{Y}_{ij}$. Because every subject has a score on X, every subject has an estimated error. The effect of the smaller within-group variation associated with ANCOVA is an increase in the power of the analysis. Note that the ANOVA distributions overlap more than do the ANCOVA distributions.

The second reason for employing ANCOVA rather than ANOVA—that of bias reduction—is also illustrated in the figure. Examine the positions of the two bivariate distributions. The drug 1 distribution is slightly to the right of the drug 2 distribution; correspondingly, the mean age score for drug 1 subjects is slightly above the mean age score for drug 2 subjects. Because age is positively correlated with test performance, we would expect the mean test score to be slightly higher for those subjects assigned to receive drug 1 simply because they are slightly older. Even if treatments are not

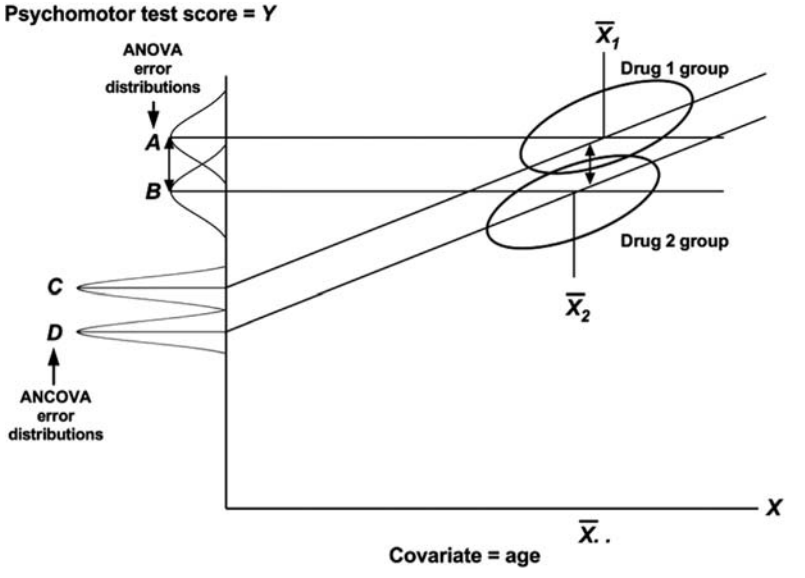


Figure 6.1 Distributions of estimated errors associated with ANOVA (A and B) and ANCOVA (C and D).

applied, we would predict that the group with the higher mean age will also have superior performance on the psychomotor test; in other words, a slight amount of bias is present because the older group is favored from the start. In this example the analysis of covariance adjusts for this bias by slightly decreasing the mean of group 1 and slightly increasing the mean of group 2. This is because the adjustment procedure changes each treatment mean to the value that is predicted under the condition that all treatments (drug groups in this example) have the same covariate (age) mean. More specifically, the adjusted means are *adjusted to the level that would be expected if all group covariate means were equal to the grand covariate mean* $\bar{X}..$ (i.e., the mean on the covariate for all subjects in the experiment).

In this example the adjustment causes the difference between the adjusted means of distributions *C* and *D* to be smaller than the difference between the unadjusted means of distributions *A* and *B*. But if the ANCOVA *F* is computed on these data the smaller difference will be associated with a larger *F*-value than with ANOVA; this is because the error term of the ANCOVA is much smaller than the error term of the ANOVA. Again note that distributions *C* and *D* overlap less than do distributions *A* and *B even though the mean difference between C and D is smaller than the mean difference between A and B*.

Quick Review of ANOVA and ANOVAR

Because the details of ANCOVA can be most easily explained by building on the rationale for the ANOVA and ANOVAR *F*-tests, it is helpful to keep in mind the

essentials of these two simpler tests. Recall from Chapter 4 that the analysis of variance and regression are two seemingly different approaches to data analysis. Whereas ANOVA is generally used to test the equality of J population means, simple ANOVAR is used to test the hypothesis of zero population slope. Both analyses involve the partitioning of the total sum of squares into variation explained by two different sources. Under the ANOVA model the total sum of squares is partitioned into the between-group sum of squares and the within-group (error) sum of squares. Similarly, under the regression model, the total sum of squares is partitioned into the regression sum of squares and the residual (error) sum of squares. An F-ratio is formed in ANOVA by dividing the MS between groups by the MS within groups. Because the former may contain one source of variability that cannot affect the latter (due to differences between means, such as those produced by treatment effects) large values of F are attributed to differences between means. In ANOVAR the MS regression may contain variability that is due to a nonzero slope. The MS residual does not contain this source of variability. Hence the F-ratio based on MS regression over MS residual will tend to be large when the population slope is nonzero. It turns out that both of these models can be integrated into the analysis of covariance model.

Although the traditional approach to conceptualizing the ANCOVA is to view it as an integration of ANOVA and ANOVAR, an alternative approach is to view it as a minor variant of multiple regression analysis. Because each approach has advantages both of them are presented; the traditional approach is presented in this chapter and the regression approach is described in Chapter 7.

6.2 ANALYSIS OF COVARIANCE MODEL

It was pointed out in Chapter 4 that the ANOVA model treats between-group variation as systematic and within-group variation as unexplainable random variation, whereas the regression model treats variation accounted for by regression as systematic and deviations from the regression as unexplainable random variation. The ANCOVA model treats *both* between-group and regression-variation as systematic (nonerror) components. The statistical model for the analysis of covariance is

$$Y_{ij} = \mu + \alpha_j + \beta_1 \left(X_{ij} - \bar{X}..\right) + \varepsilon_{ij},$$

where

Y_{ij} is the dependent variable score of ith individual in jth group;
μ is the overall population mean (on dependent variable);
α_j is the effect of treatment j;
β_1 is the linear regression coefficient of Y on X;
X_{ij} is the covariate score for ith individual in jth group;
$\bar{X}..$ is the grand covariate mean; and
ε_{ij} is the error component associated with ith individual in jth group.

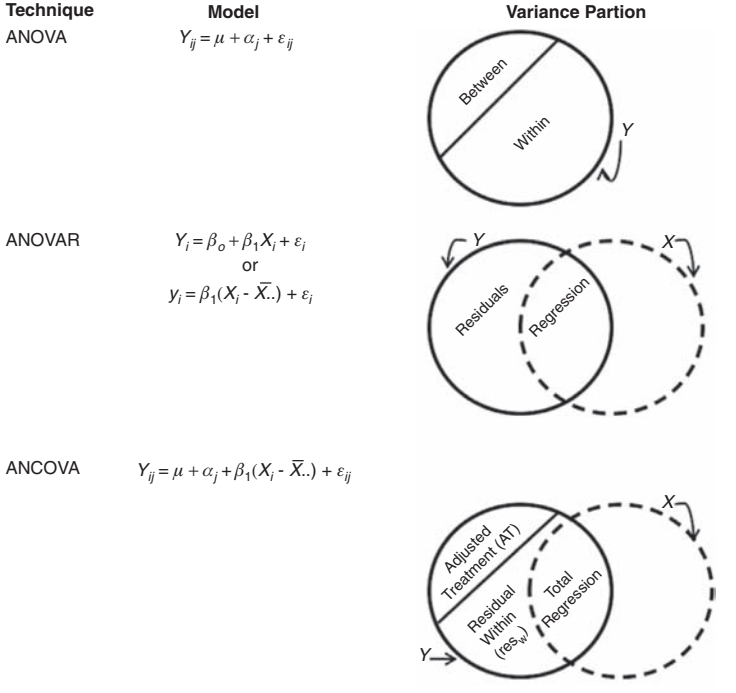


Figure 6.2 Comparison of ANOVA, ANOVAR, and ANCOVA models and partitioning.

A comparison of the models for ANOVA, ANOVAR, and ANCOVA and illustrations of the general type of partitioning involved for each are presented in Figure 6.2.

It can be seen that when the regression term is added to the ANOVA model the result is the ANCOVA model. If the assumptions for the ANCOVA are met for a given set of data, the error term will be smaller than in ANOVA because much of the within-group variability will be accounted for and removed by the regression of the dependent variable on the covariate. Because the result of the smaller error term is an increase in power, it is quite possible that data analyzed by using ANCOVA will yield statistically significant results where ANOVA yields nonsignificant results. The opposite is possible under certain conditions, but unlikely. The justification for these statements should become clear as the computation and rationale for the analysis are explained.

6.3 COMPUTATION AND RATIONALE

Suppose we have data from a three-group experiment where the dependent variable is a measure of achievement, the independent variable is type of training, and the

covariate is a measure of aptitude that is believed to be correlated with achievement. The starting point for ANCOVA is exactly the same as for ANOVAR or ANOVA; the total sum of squares is computed.

The total sum of squares in this experiment can be viewed as containing variability in achievement scores *(Y)* resulting from:

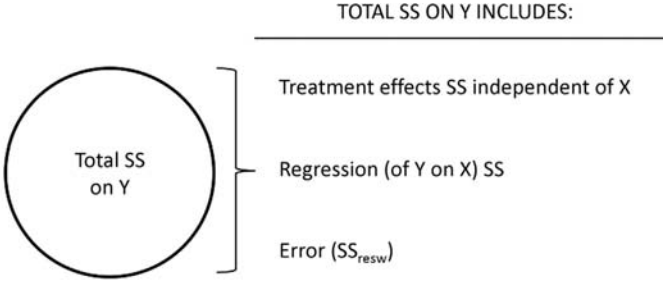
1. Treatment effects (effects of being exposed to different types of training) that are independent of covariate measures X.
2. Differences in achievement among subjects that can be predicted from aptitude X.
3. Differences among subjects that are neither due to treatment effects nor are predicted from covariate X (i.e., error).

After the total sum of squares is computed, the second step is to remove the sum of squares due to the regression of Y on X from the total sum of squares. That is, those differences in achievement (Y) that can be predicted by using the covariate *(X)* are subtracted from the total variability on *Y.* These and the remaining ANCOVA steps are illustrated in the remainder of this section, using n_1, n_2, and n_3 as the number of subjects in groups 1, 2, and 3 respectively, N as the total number of subjects in the experiment, X as a raw score on the covariate, x as a deviation from the grand mean (i.e., $X - \bar{X}..$) on the covariate, Y as a raw score on the dependent variable, and y as a deviation score on the dependent variable.

Step 1. Computation of total sum of squares.

$$\text{Total SS} = \sum y_t^2 = \sum Y_t^2 - \frac{\left(\sum Y_t\right)^2}{N}.$$

The total sum of squares includes SS due to (1) treatment effects independent of *X,* (2) differences in achievement predictable from test X (i.e., variability accounted for by regressing Y on *X),* and (3) differences among subjects that are not due to treatment effects and cannot be predicted from test X (i.e., error).

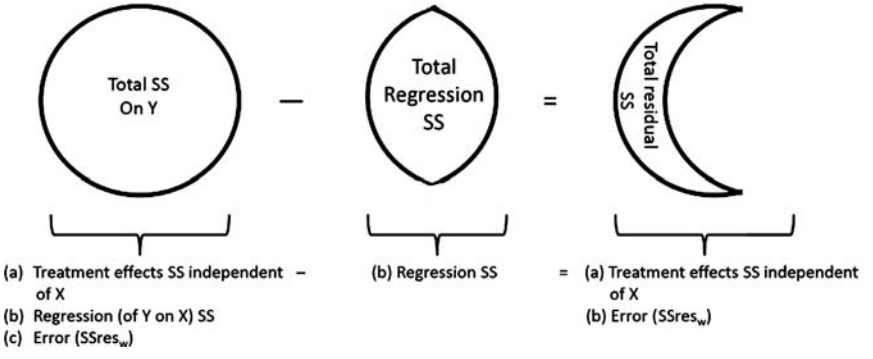


Step 2: Computation of total residual SS.

$$\text{Total residual SS } (\text{SSres}_t) = \sum y_t^2 - \frac{\left(\sum x y_t\right)^2}{\sum x_t^2}$$

$$= \left[\sum Y_t^2 - \frac{\left(\sum Y_t\right)^2}{N}\right] - \left[\frac{\left[\sum XY_t - \frac{\left(\sum X_t\right)\left(\sum Y_t\right)}{N}\right]^2}{\sum X_t^2 - \frac{\left(\sum X_t\right)^2}{N}}\right].$$

Because the total SS contains treatment effects independent of X, variability due to the regression of Y on X, and error, it follows that the total residual SS contains only variability due to treatment effects independent of X and error because the regression SS has been removed; that is:



Step 3. Computation of within-group sum of squares.

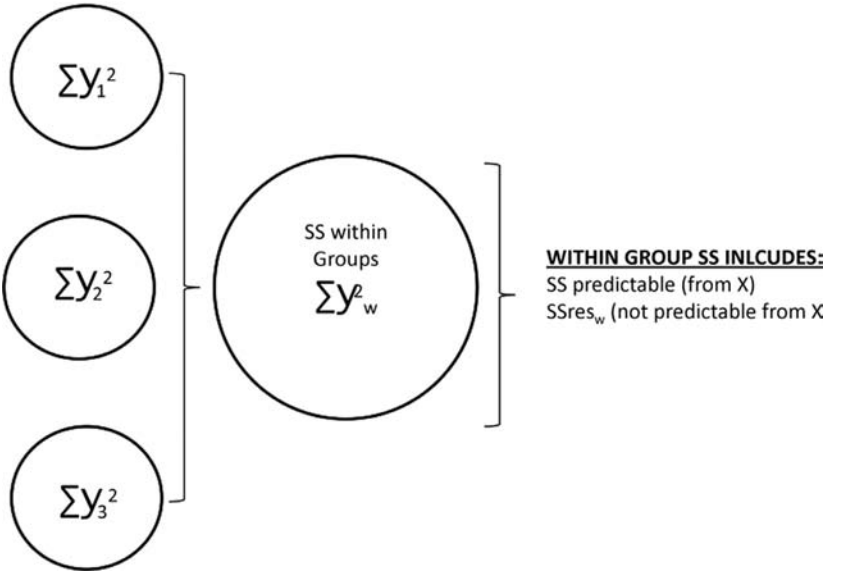
$$\text{Within-group SS} = \sum y_w^2 = \sum y_1^2 + \sum y_2^2 + \sum y_3^2$$

$$= \left[\sum Y_1^2 - \frac{\left(\sum Y_1\right)^2}{n_1}\right] + \left[\sum Y_2^2 - \frac{\left(\sum Y_2\right)^2}{n_2}\right]$$

$$+ \left[\sum Y_3^2 - \frac{\left(\sum Y_3\right)^2}{n_3}\right].$$

The within-group sum of squares may be obtained by computing $\sum y^2$ for each group and then pooling the results of the three separate computations. The

within-group SS is not, of course, influenced by treatment or between-group differences. The within-group sum of squares includes SS due to differences predictable from X and differences not predictable from X (error).



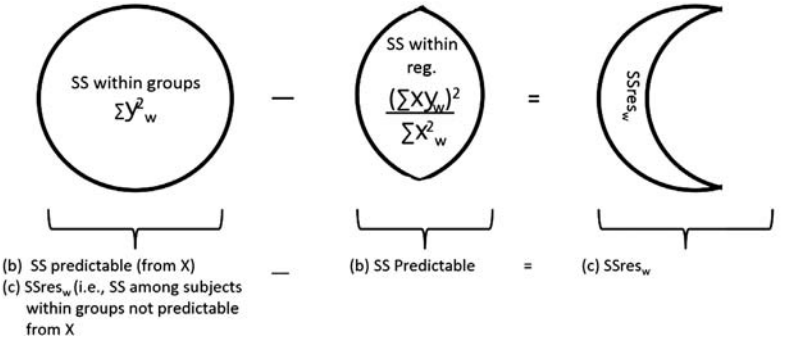
The within-group deviation cross products and sums of squares on X required in the next step are computed using the following formula:

$$\begin{aligned}\sum xy_{\mathrm{w}} &= \sum xy_1 + \sum xy_2 + \sum xy_3 \\ &= \left[\sum XY_1 - \frac{\left(\sum X_1\right)\left(\sum Y_1\right)}{n_1}\right] + \left[\sum XY_2 - \frac{\left(\sum X_2\right)\left(\sum Y_2\right)}{n_2}\right] \\ &\quad + \left[\sum XY_3 - \frac{\left(\sum X_3\right)\left(\sum Y_3\right)}{n_3}\right]\end{aligned}$$

and

$$\begin{aligned}\sum x_{\mathrm{w}}^2 &= \sum x_1^2 + \sum x_2^2 + \sum x_3^2 \\ &= \left[\sum X_1^2 - \frac{\left(\sum X_1\right)^2}{n_1}\right] + \left[\sum X_2^2 - \frac{\left(\sum X_2\right)^2}{n_2}\right] + \left[\sum X_3^2 - \frac{\left(\sum X_3\right)^2}{n_3}\right].\end{aligned}$$

Step 4. Computation of within-group residual SS (or error SS). By subtracting SS due to predictable differences among subjects within groups (sometimes called *within-group regression SS*) from SS within groups, we obtain residual sum of squares within (i.e., $SSres_w$), which is used as the error SS in ANCOVA. Differences contributing to this sum are not predictable from X (using a linear rule) and are not accounted for by treatment differences.



Step 5. Computation of adjusted-treatment effects (AT) SS. By subtracting the SS residual within (step 4) from the total residual SS (step 2), the adjusted-treatment SS is obtained. This quantity was described as the sum of squares due to "treatment effects independent of X" in the previous steps. It is more conventional, however, to refer to this portion of the partition as the adjusted-treatment sum of squares.

Step 6. Computation of F-ratio. Step 5 involves the partitioning of the total residual SS into the sum of squares residual within (i.e., $SSres_w$) and the adjusted treatment SS. The latter two correspond directly to within- and between-group (or treatment) SS in a simple analysis of variance. Thus the F-ratio can be obtained by dividing mean square adjusted treatment (MS_{AT}) by mean square error ($MSres_w$). Degrees of freedom are computed in essentially the same way as in ANOVA except that an additional degree of freedom is lost from the error MS for each covariate (or covariate polynomial) employed in the analysis. The ANCOVA summary table has the following form:

Source	SS	*df*	MS	*F*
Adjusted treatment (AT)	SS_{AT}	$J - 1$	$\frac{SS_{AT}}{J-1}$	$\frac{MS_{AT}}{MSres_w}$
Error (res_w)	$SSres_w$	$N - J - C$	$\frac{SSres_w}{N-J-C}$	
Total residual (res_t)	$SSres_t$	$N - 1 - C$		

The sums of squares appearing in the first three rows (SS_{AT}, $SSres_w$, and $SSres_t$) are obtained in steps 5, 4, and 2, respectively. C is the number of covariates (which is one in this chapter), N is the total number of subjects, and J is the number of groups.

Measure of Association

A measure of the proportion of the total variability (on Y) in the sample explained by the differences between adjusted treatment means is: $SS_{AT}/SS_{Total} = \hat{\eta}^2_{adj}$.

Adjusted Standardized Effect Size

The standardized effect size associated with the comparison of groups i and j may be computed as the ratio $\frac{\bar{Y}_{i\text{ adj}} - \bar{Y}_{j\text{ adj}}}{\sqrt{MS_{within}}} = \frac{\bar{Y}_{i\text{ adj}} - \bar{Y}_{j\text{ adj}}}{s_w} = g_{adj}$. The computation of the adjusted means $\bar{Y}_{i\text{ adj}}$ and $\bar{Y}_{j\text{ adj}}$ is described in the next section. The denominator is simply the square root of the mean square within groups from an ANOVA on Y.

6.4 ADJUSTED MEANS

An adjusted mean is a predicted score. It is the mean dependent variable score that would be expected or predicted for a specified group of subjects if the covariate mean for this group were the same as the grand covariate mean. When subjects are assigned to treatments, the various samples usually will not have *exactly* the same covariate mean even though a randomization procedure may have been employed. Because these group mean differences on the covariate are likely to exist and the covariate is (if properly chosen) highly related to the dependent variable, the investigator may wonder whether the dependent variable mean for one group is higher than for another group because differences between the groups existed *before* the experimental treatments were administered. In the long run (i.e., across many such experiments) the group means will be equal (before treatment administration) on the covariate, the dependent variable, and all other variables if random assignment to treatments has been employed. In a particular experiment, however, there will likely be *some* difference among sample means due to sampling fluctuation.

Suppose a randomized two-group experiment with equal sample sizes is conducted to investigate the effects of two different types of desensitization treatment of snake phobia. A behavioral avoidance test is administered before treatment and again after treatment. The pretreatment scores are employed as the covariate and the posttreatment scores are employed as the dependent variable. Let us say that the covariate means for the two groups are 25 and 30; this five-point difference has occurred even though randomization was employed in establishing the treatment groups. After the treatments have been administered and the dependent variable means are available, it would be reasonable for the investigator to wonder what the two-treatment means on the dependent variable would have been if the two groups had *exactly* the same covariate means rather than covariate means that differ by five points. We attempt to answer this question by computing the adjusted means. In the present example the adjusted means are the predicted dependent variable means expected to occur under the two treatments when the covariate means for both groups are 27.5, which is the covariate grand mean. Very simply, we are answering a "what if" question. *What* would the dependent variable means be *if* the covariate means were exactly 27.5

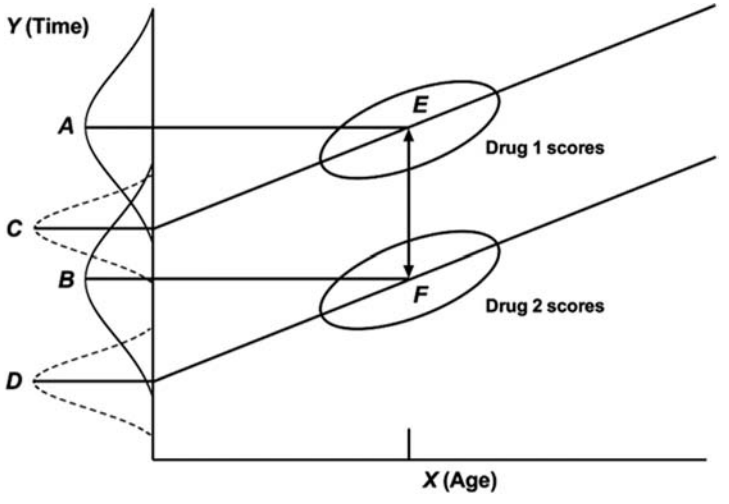


Figure 6.3 Equivalence of difference between intercepts and difference between adjusted means.

(i.e., equal to the covariate grand mean) rather than 25 and 30? If a randomized design is involved and assumptions underlying ANCOVA are met, the adjusted means provide a precise answer to this question. A discussion of the assumptions is provided in Chapter 8.

It has been stated that the purpose of ANCOVA is to test the null hypothesis that two or more adjusted population means are equal. Alternatively, we could state that the purpose is to test the equality of two or more regression intercepts. Under the assumption of parallel regression lines, the difference between intercepts must be equal to the difference between adjusted means. This equality can be seen in Figure 6.3.

Note that the difference between points C and D (the intercepts) is exactly the same as the difference between points E and F (the adjusted means). It also turns out that the difference between points A and B (the unadjusted means) is equal to the difference between the adjusted means in *this particular example*. Only when the covariate means for all groups are the same, or the slope b_w is zero, will the unadjusted means be identical to the adjusted means.

Adjusted means should be reported as a standard part of any covariance analysis because inferential statements concerning the results of ANCOVA refer to differences among adjusted means (or intercepts). It is critical that the adjusted means be computed and closely inspected before the results of a covariance analysis are reported because it is often difficult to determine the relative size of adjusted means by simply inspecting the unadjusted means. This point should become clear when the details of the adjustment procedure are described.

The formula for the computation of adjusted means is

$$\bar{Y}_j - b_w \left(\bar{X}_j - \bar{X}..\right) = \bar{Y}_{j\text{adj}},$$

where

$\bar{Y}_{j\ \text{adj}}$ is the adjusted mean for jth treatment group;
$\bar{Y}_j$ is the unadjusted mean for jth treatment group;
b_w is the pooled within-group regression coefficient;
$\bar{X}_j$ is the covariate mean for jth treatment group; and
$\bar{X}..$ is the grand covariate mean.

Suppose a two-group ANCOVA is computed, and the following data are obtained:

Treatment 1		Treatment 2
$\bar{X}_1 = 50$		$\bar{X}_2 = 10$
$\bar{Y}_1 = 120$		$\bar{Y}_2 = 80$
	$\bar{X}_{..} = 30$	
	$b_w = 0.70$	

The adjusted means for the two groups are

$$\begin{aligned} 120 - [0.70(50 - 30)] &= 120 - 14 \\ &= 106 \\ &= \bar{Y}_{1\ \text{adj}} \end{aligned}$$

and

$$\begin{aligned} 80 - [0.70(10 - 30)] &= 80 - [-14] \\ &= 94 \\ &= \bar{Y}_{2\ \text{adj}} \end{aligned}$$

An examination of the adjustment formula indicates that the magnitude of the adjustment is a function of (1) the difference between the treatment-group covariate mean $\bar{X}_j$ and the grand covariate mean $\bar{X}..$, and (2) the pooled within-group regression coefficient. It can be seen that the pooled within-group slope is used to predict the extent to which between-group regression toward the grand mean is expected to occur.

The effects of varying the degree and direction of the difference between covariate means can be seen in Figure 6.4, where each situation is drawn with the same slope b_w and the same difference between unadjusted means. Situation I may be used as a reference against which the others may be compared. Note that the adjusted and unadjusted differences are the same in this situation. Situation II*a* indicates what happens to the adjusted means when the group with the higher unadjusted mean also has a somewhat higher covariate mean—the difference becomes smaller. Situation II*b* indicates that the adjustment procedure may result in no difference between adjusted means, and situation II*c* indicates that adjustment can actually reverse the order of the adjusted means relative to the unadjusted means. That is, the unadjusted mean for treatment 1 is higher than the unadjusted treatment mean 2, but adjusted treatment

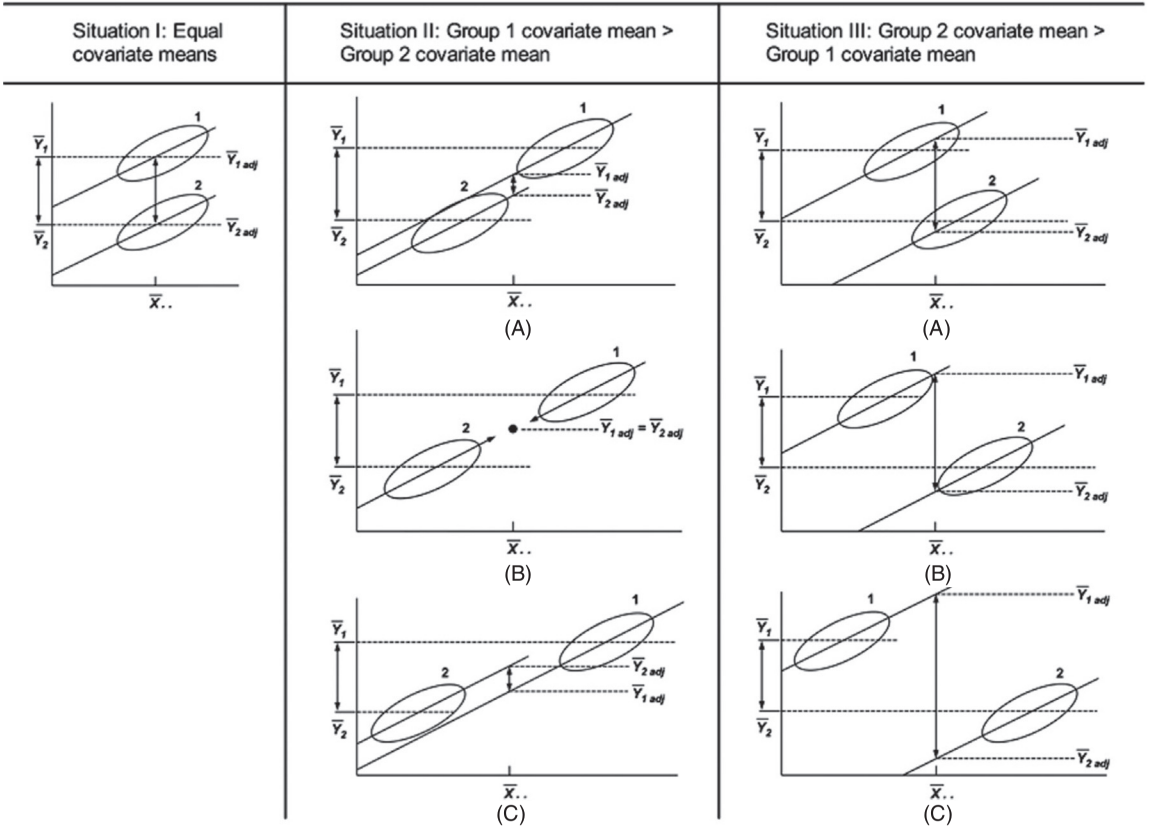


Figure 6.4 Effects of different covariate means on adjusted means.

mean 1 is lower than adjusted mean 2. Situation III diagrams indicate that if the treatment 1 covariate mean is lower than the treatment 2 covariate mean, the difference between the adjusted means will increase as the $\bar{X}_1 - \bar{X}_2$ difference increases.

The general rule to keep in mind is that "winners lose and losers win." That is, the group with the higher covariate mean will lose the most (i.e., the downward adjustment will be the greatest), and the group with the lower covariate mean will gain the most. This will be true as long as the slope b_w is positive. Suppose that the following data have been obtained:

	$\bar{X}$	$\bar{Y}$		
Group 1	30	50	$n_1 = 30$	
Group 2	20	40	$n_2 = 30$	$b_w = 0.80$

Because group 1 has the higher covariate mean we know that the group 1 adjusted mean will be lower than 50 and that the group 2 adjusted mean will be higher than 40. Using the adjustment formula, we find

	$\bar{Y}_{adj}$
$50 - 0.80(30 - 25) =$	46
$40 - 0.80(20 - 25) =$	44

If the covariate means were equal we would find the following situation:

	$\bar{Y}$	$= \bar{Y}_{adj}$
Group 1	50	50
Group 2	40	40

because

$$50 - 0.80(0) = 50$$
$$40 - 0.80(0) = 40$$

In contrast, if the difference between covariate means were 20 points, we would find

	$\bar{Y}$	$\bar{Y}_{adj}$
Group 1	50	42
Group 2	40	48

because

$$50 - 0.80(10) = 42$$
$$40 - 0.80(-10) = 48$$

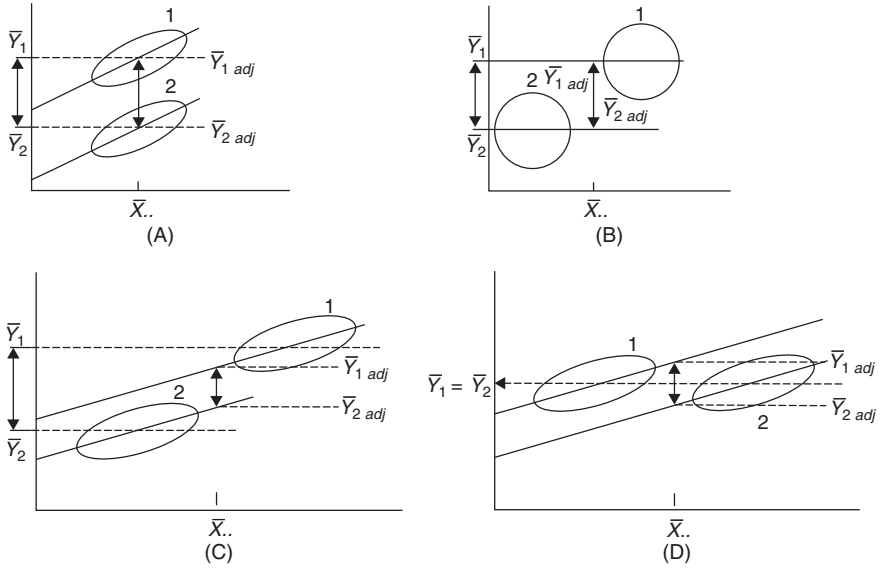


Figure 6.5 Effects of varying slope b_w and covariate means on adjusted means.

The effect of the slope on the adjusted means is illustrated in Figure 6.5. Situation *a* in Figure 6.5 indicates that no adjustment is involved, even though the slope is quite steep, when the covariate means are equal. Situation *b* indicates that no adjustment is possible when the slope is equal to 0.0 regardless of the difference between covariate means. Situation *c* involves a combination of the slope of *a* with the covariate mean difference of *b*. A large adjustment is evident in this situation. Situation *d* indicates that it is possible to have no difference between unadjusted means and yet a relatively large difference between adjusted means when the slope is steep and a substantial difference exists between covariate means.

Sample sizes are equal in the adjustment examples described up to this point. If the absolute size of the adjusted means rather than the difference between them is of primary interest, the relative number of subjects in the groups should be considered in the interpretation. When the number of subjects is the same in the two-group situation, the magnitude of the adjustment is the same for both groups. For example,

	$\bar{X}$	$\bar{Y}$	$\bar{Y}_{adj}$
Group 1	5	10	11.25
Group 2	10	10	8.75

$$\sum X_1 = 250 \qquad n_1 = 50$$
$$\sum X_2 = 500 \qquad n_2 = 50$$
$$\sum X_3 = 750 \qquad b_w = 0.5$$
$$\bar{X}.. = 7.5$$

$$\bar{Y}_{1\ adj} = 10 - [0.5(5 - 7.5)] = 11.25$$

$$\bar{Y}_{2\ adj} = 10 - [0.5(10 - 7.5)] = 8.75$$

Note that the Y-values have been adjusted for both groups by the same number of units.

When the number of subjects in the groups is not the same, the grand mean $\bar{X}..$ is more heavily weighted by the group with the larger n, and, as the following contrived example shows, the Y means of the two groups are adjusted by different amounts:

	$\bar{X}$	$\bar{Y}$	$\bar{Y}_{\text{adj}}$
Group 1	5	10	10.05
Group 2	10	10	7.55

$$\sum X_1 = 490 \qquad \sum X_2 = 20 \qquad \sum X_3 = 510$$

$$n_1 = 98 \qquad n_2 = 2 \qquad b_w = 0.5 \qquad \bar{X}.. = 5.1$$

$$\bar{Y}_{1\text{ adj}} = 10 - [0.5(5 - 5.1)] = 10.05$$

$$\bar{Y}_{2\text{ adj}} = 10 - [0.5(10 - 5.1)] = 7.55$$

The two previous examples of adjustment indicate once again that it is possible for adjusted treatment means to differ when the unadjusted values are identical. Hence it should be clear that the adjusted means must be computed to discover the values that are being compared in the ANCOVA.

A more extreme example of the necessity for computing the adjusted means is as follows:

	$\bar{X}$	$\bar{Y}$	$\bar{Y}_{\text{adj}}$
Group 1	115	90	78
Group 2	85	85	97

$$n_1 = n_2 \qquad b_w = 0.8 \qquad \bar{X}.. = 100$$

$$\bar{Y}_{1\text{ adj}} = 90 - [0.8(115 - 100)] = 78$$

$$\bar{Y}_{2\text{ adj}} = 85 - [0.8(85 - 100)] = 97$$

If nothing beyond the unadjusted Y means and the ANCOVA F-value were to be inspected, it would be easy to misinterpret the results completely because the order of the means would be reversed in the adjusted and unadjusted situations. Thus, it can be seen that the adjusted means can be smaller than, larger than, in the same rank order as, or in a different rank order than, the corresponding unadjusted means. It is also true that the ANCOVA F-value can be smaller, larger, or equal to the ANOVA F-value for given sets of data. This does not mean, however, that a small difference between adjusted means is necessarily associated with a small F-value. The use of ANCOVA often results in a dramatic reduction in the size of the error term (relative to the ANOVA error term), and a small difference between adjusted means may be associated with a relatively large F-statistic. In most experiments the ANCOVA F exceeds the ANOVA F. Procedures

for comparing differences between specific pairs of adjusted means are discussed in Chapter 9.

6.5 ANCOVA EXAMPLE 1: TRAINING EFFECTS

The computational details associated with ANCOVA are described in this section for the training study previously mentioned in Section 6.3. Recall that the independent variable is type of training (three levels), aptitude (X) is the covariate, and achievement is the dependent variable (Y). The conventional data layout appears in Table 6.1. Table 6.2 shows the same data rearranged in the format required for input to most ANCOVA software routines.

Step 1. Computation of total sum of squares.

$$\begin{aligned}\text{Total SS} &= \sum y_t^2 \\ &= \sum Y_t^2 - \frac{\left(\sum Y_t\right)^2}{N} \\ &= 40{,}706 - \frac{1{,}102{,}500}{30} \\ &= 3956.00\end{aligned}$$

This quantity contains variability predictable from the aptitude test (X), variability due to differences produced by the three types of behavioral objective, and error (i.e., differences unrelated to either the aptitude test or the treatments).

Table 6.1 Data Layout for Three-Group ANCOVA

Treatment (Type of Training)					
1		2		3	
X	Y	X	Y	X	Y
29	15	22	20	33	14
49	19	24	34	45	20
48	21	49	28	35	30
35	27	46	35	39	32
53	35	52	42	36	34
47	39	43	44	48	42
46	23	64	46	63	40
74	38	61	47	57	38
72	33	55	40	56	54
67	50	54	54	78	56

Table 6.2 Three-Group Data Layout Required for Input to Most ANCOVA Software Routines

Group	*X*	*Y*
1	29	15
1	49	19
1	48	21
1	35	27
1	53	35
1	47	39
1	46	23
1	74	38
1	72	33
1	67	50
2	22	20
2	24	34
2	49	28
2	46	35
2	52	42
2	43	44
2	64	46
2	61	47
2	55	40
2	54	54
3	33	14
3	45	20
3	35	30
3	39	32
3	36	34
3	48	42
3	63	40
3	57	38
3	56	54
3	78	56

Step 2. Computation of total residuals.

$$SS_{res_t} = \sum y_t^2 - \frac{\left(\sum x y_t\right)^2}{\left(\sum x_t^2\right)}$$

$$= \sum y_t^2 \text{ (from step 1)} - \frac{\left[\sum XY_t - \frac{\left(\sum X_t\right)\left(\sum Y_t\right)}{N}\right]^2}{\sum X_t^2 - \frac{\left(\sum X_t\right)^2}{N}}$$

$$= 3956.00 - \frac{\left[54{,}822 - \dfrac{(1480)(1050)}{30}\right]^2}{78{,}840 - \dfrac{2{,}190{,}400}{30}}$$

$$= 3956.00 - 1567.36 = 2388.64$$

Because the portion of the total variation on the dependent variable that is predictable from the aptitude test scores is the total regression sum of squares (i.e., 1567.36), the remaining sum of squares (total residuals) contains variability due to treatment effects (different types of training) and variability not due to different treatments. To determine how much of the total residual variability is due to treatments and how much is not due to treatments, we perform steps 3, 4, and 5.

Step 3. Computation of within-group sum of squares. We must compute the quantity $\sum y^2$ for each treatment group separately and then add the results to obtain the within-group sum of squares.

$$\left.\begin{aligned}
\sum y_1^2 &= \sum Y_1^2 - \frac{\left(\sum Y_1\right)^2}{n_1} = 10{,}064 - \frac{90{,}000}{10} = 1064.00 \\
\sum y_2^2 &= \sum Y_2^2 - \frac{\left(\sum Y_2\right)^2}{n_2} = 16{,}106 - \frac{152{,}100}{10} = 896.00 \\
\sum y_3^2 &= \sum Y_3^2 - \frac{\left(\sum Y_3\right)^2}{n_3} = 14{,}536 - \frac{129{,}600}{10} = 1576.00
\end{aligned}\right]$$

$$3536.00 = \sum y_{\mathrm{w}}^2$$

The within-group sum of squares includes variability predictable from the aptitude test (X) and variability not predictable from X. Note that between-group or treatment variability is of no concern here because the computations were carried out separately for each group; this, of course, eliminates the possibility of any variability due to treatments affecting the within-group sum of squares.

Step 4. Computation of within-group residual sum of squares. Because the within-group sum of squares contains variability predictable from the aptitude test, it is necessary to remove the predictable variability in order to obtain the portion of the total residual sum of squares that is unrelated to either the treatments or the aptitude test. This is accomplished by computing the within group regression sum of squares and subtracting it from the within group sum of squares; the result is called the

within-group residual sum of squares.

$$\left.\begin{aligned}\sum xy_1 &= \sum XY_1 - \frac{\left(\sum Y_1\right)\left(\sum Y_1\right)}{n_1} = 16{,}603 - \frac{(520)\,(300)}{10} = 1003.00\\ \sum xy_2 &= \sum XY_2 - \frac{\left(\sum Y_2\right)\left(\sum Y_2\right)}{n_2} = 19{,}241 - \frac{(470)\,(390)}{10} = 911.00\\ \sum xy_3 &= \sum XY_3 - \frac{\left(\sum Y_3\right)\left(\sum Y_3\right)}{n_3} = 18{,}978 - \frac{(490)\,(360)}{10} = 1338.00\end{aligned}\right]$$

$$3252.00 = \sum xy_{\mathrm{w}}$$

$$\left.\begin{aligned}\sum x_1^2 &= \sum X_1^2 - \frac{\left(\sum X_1\right)^2}{n_1} = 29{,}054 - \frac{270{,}400}{10} = 2014.00\\ \sum x_2^2 &= \sum X_2^2 - \frac{\left(\sum X_2\right)^2}{n_2} = 23{,}888 - \frac{220{,}900}{10} = 1798.00\\ \sum x_3^2 &= \sum X_3^2 - \frac{\left(\sum X_3\right)^2}{n_3} = 25{,}898 - \frac{240{,}100}{10} = 1888.00\end{aligned}\right]$$

$$5700.00 = \sum x_{\mathrm{w}}^2$$

The within-group regression coefficient (to be used later in adjusting the means) is

$$b_{\mathrm{w}} = \frac{\sum xy_{\mathrm{w}}}{\sum x_{\mathrm{w}}^2} = \frac{3253}{5700} = 0.5705,$$

and the within-group regression sum of squares is

$$\frac{\left(\sum xy_{\mathrm{w}}\right)^2}{\sum x_{\mathrm{w}}^2} = \frac{(3252)^2}{5700} = 1855.35 = \mathrm{SS}_{\mathrm{reg_w}}.$$

By subtraction of the within-group regression sum of squares from the previously computed within-group sum of squares, we obtain the within-group residual sum of squares:

$$\sum y_{\mathrm{w}}^2 - \frac{\left(\sum xy_{\mathrm{w}}\right)^2}{\sum x_{\mathrm{w}}^2} = 3536 - 1855.35 = 1680.65 = \mathrm{SS}_{\mathrm{res_w}}$$

We now know how much of the total residual sum of squares is error variability unrelated to either treatments or aptitude test scores.

Step 5. Computation of adjusted treatment effects. By subtracting the within-group residual sum of squares from the total residual sum of squares, we obtain the adjusted treatment (AT) sum of squares. This quantity represents differences among treatment groups that are not predictable from the aptitude test scores.

$$\begin{aligned}&\text{Total residuals (step 2)} = 2388.64 = SS_{res_t}\\ &\text{– Within residuals (step 4)} = 1680.65 = SS_{res_w}\\ \hline &\text{Adjusted treatment effect (AT)} = 707.99 = SS_{AT}\end{aligned}$$

Step 6. Computation of F-ratio.

Source	SS	*df*	MS	*F*
Adjusted Treatment (AT)	707.99	2	354.00	5.48
Error (res_w)	1680.65	26	64.64	
Total residuals (res_t)	2388.64	28		

The study involves one covariate, three groups, and 30 subjects. The degrees of freedom are $J - 1$ or 2 for treatments, $N - J - C$ or $30 - 3 - 1 = 26$ for error, and $N - 1 - C$ for total residuals. The obtained F is then compared with the critical value of F with 2 and 26 degrees of freedom; $F_{(.05,\,2,\,26)}$ is 3.37, and the null hypothesis of no treatment effect is rejected. A summary of the computations involved in the ANCOVA and the adjustment of means is presented in Table 6.3.

Adjusted means:

$$\begin{aligned}&\textit{Group 1}.\quad 30 - 0.57(52 - 49.33) = 28.48 = \bar{Y}_{1\text{ adj}}\\ &\textit{Group 2}.\quad 39 - 0.57(47 - 49.33) = 40.33 = \bar{Y}_{2\text{ adj}}\\ &\textit{Group 3}.\quad 36 - 0.57(49 - 49.33) = 36.19 = \bar{Y}_{3\text{ adj}}\end{aligned}$$

The proportion of total variability on Y accounted for by adjusted treatment effects is:

$$SS_{AT}/SS_t = 707.99/3956 = .18.$$

Because it may be misleading to compare adjusted means if slopes are not homogeneous, the homogeneity of regression test should be considered as a necessary adjunct to the ANCOVA summary. The details of this test are presented in the next section.

6.6 TESTING HOMOGENEITY OF REGRESSION SLOPES

An assumption underlying the correct usage of ANCOVA is that the population regression slopes associated with the treatment populations are equal. The

Table 6.3 Summary of Sums of Squares Computation for ANCOVA

Total	Group 1	Group 2	Group 3	Within
		Sums of Squares on Y		
$\sum y_t^2 = 3956.00$	$\sum y_1^2 = 1064.00$	$\sum y_2^2 = 596.00$	$\sum y_3^2 = 1576.00$	$\sum y_w^2 = 3536.00$
		Sums of Squares on X		
$\sum x_t^2 = 5826.67$	$\sum x_1^2 = 2014.00$	$\sum x_2^2 = 1798.00$	$\sum x_3^2 = 1888.00$	$\sum x_w^2 = 5700.00$
		Sums of Cross Products		
$\sum xy_t = 3022.00$	$\sum xy_1 = 1003.00$	$\sum xy_2 = 911.00$	$\sum xy_3 = 1338.00$	$\sum xy_w = 3252.00$
		Regression Weights		
$\frac{\sum xy_t}{\sum x_t^2} = 0.5187 = b_1^t$	$\frac{\sum xy_t}{\sum x_1^2} = 0.4980 = b_1^{(\text{group 1})}$	$\frac{\sum xy_2}{\sum x_2^2} = 0.5067 = b_1^{(\text{group 2})}$	$\frac{\sum xy_3}{\sum x_3^2} = 0.7087 = b_1^{(\text{group 3})}$	$\frac{\sum xy_w}{\sum x_w^2} = 0.5705 = b_w$
		Sums-of-squares Regression		
$\frac{\left(\sum xy_t\right)^2}{\sum x_t^2} = 1567.36 = SS_{reg_t}$	$\frac{\left(\sum xy_1\right)^2}{\sum x_1^2} = 499.51$	$\frac{\left(\sum xy_2\right)^2}{\sum x_2^2} = 461.58$	$\frac{\left(\sum xy_3\right)^2}{\sum x_3^2} = 948.22$	$\frac{\left(\sum xy_w\right)^2}{\sum x_w^2} = 1855.35 = SS_{reg_w}$
		Sums-of-squares Residuals		
$\sum y_t^2 - \frac{\left(\sum xy_t\right)^2}{\left(\sum x_t^2\right)} = 2388.64 \; SS_{res_t}$	$\sum y_1^2 - \frac{\left(\sum xy_1\right)^2}{\sum x_1^2} = 564.49$	$\sum y_2^2 - \frac{\left(\sum xy_2\right)^2}{\sum x_2^2} = 434.42$	$\sum y_3^2 - \frac{\left(\sum xy_3\right)^2}{\sum y_3^2} = 627.78$	$\sum y_w^2 - \frac{\left(\sum xy_w\right)^2}{\sum y_w^2} = 1680.65 = SS_{res_w}$
		$1626.69 = SS_{res_t}$		

consequences of violating this assumption are described in detail in Chapter 11. Briefly, the problem is that the adjusted means are inadequate descriptive measures of the outcome of a study if the size of the treatment effect on Y (i.e., the vertical distance between the regression lines) is not the same at different levels of the covariate X. If the slopes are heterogeneous, the treatment effects differ at different levels of the covariate; consequently, the adjusted means can be misleading because they do not convey this important information. When the slopes are homogeneous, the adjusted means are adequate descriptive measures because the treatment effects are the same at different levels of the covariate. A method of testing the assumption of homogeneous regression slopes is presented in this section. When the results of this test suggest that the slopes are heterogeneous, the procedures described in Chapter 11 should be employed.

If the slopes for the treatment populations in an experiment are equal, that is, $\beta_1^{(\text{group } 1)} = \beta_1^{(\text{group } 2)} = \cdots = \beta_1^{(\text{group } J)}$ the best way of estimating the value of this common slope from the samples is by computing an average of the sample b_1 values. In the example study the sample regression slopes for the three-treatment groups are

$$b_1^{(\text{group } 1)} = 0.4980$$
$$b_1^{(\text{group } 2)} = 0.5067$$
$$b_1^{(\text{group } 3)} = 0.7087$$

It can be seen in Table 6.2 that these values were obtained by dividing $\sum xy$ by $\sum x^2$ for each group. If the three separate values $\sum xy_1$, $\sum xy_2$, and $\sum xy_3$ are summed to obtain $\sum xy_{\text{w}}$ and divided by the sum of $\sum x_1^2$, $\sum x_2^2$, and $\sum x_3^2$, which is $\sum x_{\text{w}}^2$, we have the pooled within-group regression slope b_{W}, which is the weighted average of the three separate within-group slopes. (For consistency, the symbol for the pooled within-group slope should be $b_1^{(\text{w})}$ rather than b_{W}; the latter, however, is employed consistently in the ANCOVA literature.) The slope b_{W} is our best estimate of the population slope β_1, which is the slope assumed to be common to all treatment populations. As long as $\beta_1^{(\text{group } 1)} = \beta_1^{(\text{group } 2)} = \beta_1^{(\text{group } 3)} = \beta_1$, the estimate b_{W} is a useful statistic to employ. But if the population slopes are not equal, it no longer makes sense to obtain an average or pooled slope b_{W} to estimate a single population parameter because the separate sample values are not all estimates of the same parameter.

Now the problem is to decide whether the treatment populations have the same slope. More specifically, in terms of the example problem, the question is, "Are the three sample values 0.4980, 0.5067, and 0.7087 all estimates of a single parameter β_1 or are the differences among these values so large that it is unlikely that they have come from populations with the same slope?"

If we have reason to believe that the sample values differ more than would be expected on the basis of sampling fluctuation, we may conclude that the population slopes are not equal. The homogeneity of regression F-test is designed to answer the question of the equality of the population slopes. The null hypothesis associated with

this test is

$$H_0 : \beta_1^{(\text{group 1})} = \beta_1^{(\text{group 2})} = \cdots = \beta_1^{(\text{group J})}$$

The steps involved in the computation of the test are described next.

Steps 1 and 2. Computation of within-group sum of squares and within-group residual sum of squares ($SSres_w$). Steps 1 and 2 were described and carried out earlier (in Section 6.5) as the third and fourth steps in computing ANCOVA. The example data summary values are repeated here.

$$\sum y_w^2 = 3536.00$$

$$\frac{\left(\sum xy_w\right)^2}{\sum x_w^2} = 1855.35$$

$$\sum y_w^2 - \frac{\left(\sum xy_w\right)^2}{\sum x_w^2} = 1680.65 = \text{SS}_{\text{res}_w}$$

Step 3. Computation of individual sum of squares residual (SS_{res_i}). The third step involves computation of the sum of squares residual for each treatment group separately and then pooling these residuals to obtain the pooled individual residual sum of squares (SS_{res_i}). The difference in the computation of SS_{res_w} and SS_{res_i} to keep in mind is that SS_{res_w} involves computing the residual sum of squares around the single b_w value whereas SS_{res_i} involves the computation of the residual sum of squares around the b_1 values fitted to each group separately.

$$\left.\begin{aligned}
\sum y_1^2 - \frac{\left(\sum xy_1\right)^2}{\sum x_1^2} &= \text{SS residual for group 1} = 564.49 \\
\sum y_2^2 - \frac{\left(\sum xy_2\right)^2}{\sum x_2^2} &= \text{SS residual for group 2} = 434.42 \\
\sum y_3^2 - \frac{\left(\sum xy_3\right)^2}{\sum x_3^2} &= \text{SS residual for group 3} = 627.78
\end{aligned}\right]$$

$$1626.69 = \text{SS}_{\text{res}_i}$$

Step 4. Computation of heterogeneity of slopes sum of squares. The discrepancy between SS_{res_w} and SS_{res_i} reflects the extent to which the individual regression slopes are different from the pooled within-group slope b_w; hence, the heterogeneity of slopes

SS is simply $SS_{res_w} - SS_{res_i}$. The rationale for this computation is straightforward. Note that SS_{res_i} is less than SS_{res_w} for the example data. It turns out that SS_{res_i} can never be larger than SS_{res_w} just as, in an ordinary ANOVA, the sum of squares within can never be larger than the sum of squares total. There is only one explanation for SS_{res_w} being larger than SS_{res_i}—the individual within-group slopes must be different. Within-group SSres around the average least-squares regression line b_w will be equal to the residuals around the individual within-group slopes if and only if $b_1^{(\text{group 1})} = b_1^{(\text{group 2})} = \cdots = b_1^{(\text{group J})}$. Obviously, when the sample slopes are all equal, the sum of squares for the heterogeneity of the slopes is zero, because when the individual slopes are the same they are also equal to b_w, and SS_{res_i} then must equal SS_{res_w}.

When the individual within-group slopes differ, the single slope b_w cannot have residuals as small as those around the separate slopes. And when large differences between individual slopes exist, SS_{res_w} is much larger than SS_{res_i}. A single regression slope simply cannot fit different samples of data as well as a separate slope for each sample, unless there are no differences among the slopes. The heterogeneity of slopes SS for the example data is $(1680.65 - 1626.69) = 53.96$.

Step 5. *Computation of F-ratio*. The summary table for the F-test is as follows:

Source	SS	*df*	MS	*F*
Heterogeneity of slopes	SS_{het}	$J - 1$	$SS_{het}/(J-1)$	MS_{het}/MS_{res_i}
Individual residual (res_i)	$SSres_i$	$N - (J2)$	$SS_{res_i}/(N-(J2))$	
Within residual (res_w)	$SSres_w$	$N - J - 1$		

If the obtained F is equal to or greater than $F_{[\alpha, J-1, N-(J2)]}$, the null hypothesis $\beta_1^{(\text{group 1})} = \beta_1^{(\text{group 2})} = \cdots = \beta_1^{(\text{group J})}$ is rejected. For the example data, the summary of the analysis is:

Source	SS	*df*	MS	*F*
Heterogeneity of slopes	53.96	2	26.98	0.40
Individual residual (res_i)	1626.69	24	67.78	
Within residual (res_w)	1680.65	26		

The critical value using $\alpha = .05$ is $F_{(.05, 2, 24)} = 3.40$; hence the null hypothesis is not rejected. It is concluded that there is insufficient evidence to claim a violation of the homogeneous regression assumption and that the ANCOVA model is justified. The individual slopes 0.4980, 0.5067, and 0.7087 appear to differ because of sampling fluctuation only.

6.7 ANCOVA EXAMPLE 2: SEXUAL ACTIVITY REDUCES LIFESPAN

Partridge and Farquhar (1981) studied the physiological cost of increased reproduction on lifespan. They assigned approximately 25 male fruit flies (*Drosophilia*

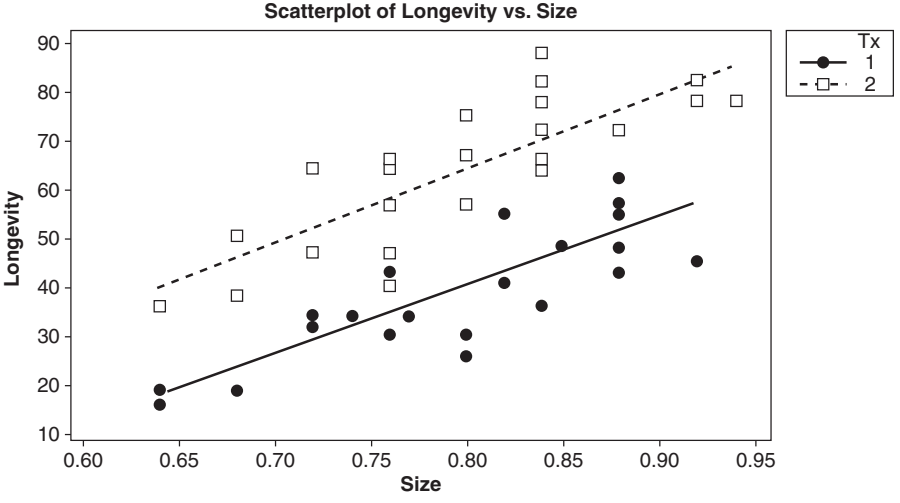


Figure 6.6 Scatterplot of longevity on size. Longevity is measured in days, size is measured as thorax length. Subjects: *Drosophilia melanogaster.* (Data from Partridge and Farquhar, 1981, Figure 2c.)

melanogaster) to each of several experimental and control conditions. Data based on two experimental conditions are analyzed here. Sexual activity was manipulated by supplying each male in one group with eight virgins per day, and supplying each male in the other group with eight inseminated females per day. Because newly inseminated females will not usually re-mate for at least 2 days, there was a clear difference between the two groups in the anticipated amount of sexual activity. These conditions continued until the male died. The dependent variable was longevity. The results are displayed in Figure 6.6. ANOVA indicates that there was a statistically significant treatment effect (as can be seen in the software section), but the preferred analysis is ANCOVA.

It was known from previous research that there is a correlation between male body size and longevity; therefore, body size was used as the covariate in an ANCOVA. The ANOVA and ANCOVA results are presented in the software section shown below in section 6.8 (Example 2 analysis using *Minitab*), along with the test of homogeneous regression slopes.

It can be seen that the unadjusted mean difference from ANOVA (−23.61) is essentially the same as the difference between the ANCOVA adjusted means (−23.75), but the ANOVA $F = 33.61$ whereas the ANCOVA $F = 97.89$. This is not surprising if the error mean squares from these analyses are compared. The ANCOVA error term is about a third the size of the ANOVA error term. The p-value from the test for homogeneity of regression slopes is .732, which supports the ANCOVA model.

The adjusted standardized effect size is

$$\frac{\bar{Y}_{1\text{ adj}} - \bar{Y}_{2\text{ adj}}}{\sqrt{\text{MS}_\text{w}}} = \frac{-23.75}{14.39} = -1.65 = g_{\text{adj}}$$

The probability that a randomly selected male from the group kept with virgins will have a shorter lifespan than a randomly selected male from the other group is easily computed. The required statistic is $\frac{g_{adj}}{\sqrt{2}} = \frac{-1.65}{\sqrt{2}} = -1.17 = z$; the area in the standard normal distribution above this z is .88, which is the probability estimate of interest.

6.8 SOFTWARE

Analysis of Training Example Using *Minitab* There are several ways to compute ANCOVA and the homogeneity of regression slopes test using *Minitab*. The cleanest ANCOVA output is provided using the ANCOVA commands (shown below) that are entered after opening the command line editor. This approach is applied to the example data shown in Table 6.1. First, the data from all three groups are stacked to produce three columns and 30 rows. The first column contains the treatment group number (1, 2, or 3) that is associated with each subject; name this column TX. The second column is the covariate X, and the third column is the dependent variable Y. Residuals are always of interest; we will assign them to column 6. Open the command line editor and proceed as shown below.

Input

```
MTB > ancova Y=TX;
SUBC> covariate X;
SUBC> means TX;
SUBC> Residuals c6.
```

Output

```
ANCOVA: Y versus TX
Factor  Levels  Values
TX           3  1, 2, 3
Analysis of Covariance for Y
Source      DF   Adj SS        MS      F          P
Covariates   1  1855.35   1855.35  28.70      0.000
TX           2   707.99    354.00   5.48      0.010
Error       26  1680.65     64.64
Total       29  3956.00
S = 8.03992   R-Sq = 57.52%   R-Sq(adj) = 52.61%
Covariate    Coef       SE Coef      T        P
X          0.5705       0.106     5.357  0.000
Adjusted Means
TX   N       Y
1   10     28.479
2   10     40.331
3   10     36.190
```

Note that the ANCOVA summary table is not in the format I have used earlier in this chapter. This is typical of all the major software routines. The first line is a test of the significance of the covariate, which is not usually necessary. The second line is labeled "TX" but it is actually the adjusted treatment effect line that I have previously labeled as AT; the remaining values on this line display the ANCOVA F- and p-values. The total residual sum of squares is not shown (but it can be computed by adding the sums of squares for the second and third lines). The line labeled "error" refers to the pooled within-groups error, so the sum of squares for this line is SS_{res_w}. The "S" is the square root of the mean square error and the next line presents another test on the significance of the covariate. In the case of a single covariate the square of this t is equal to the F-value in the first line; the p-values for the two tests are the same. In the case of multiple covariates the two tests are not the same.

The homogeneity of regression slopes test can be performed using the general linear model commands shown below:

Input

```
MTB > glm Y=TX X TX*X;
SUBC> covariate X;
SUBC> resid c4.
```

Output

```
General Linear Model: Y versus TX
Factor  Type   Levels  Values
TX      fixed       3  1, 2, 3

Analysis of Variance for Y, using Adjusted SS for Tests
Source     DF      Seq SS    Adj SS    Adj MS      F        P
TX          2      420.00     80.16     40.08   0.59    0.561
X           1     1855.35   1855.24   1855.24  27.37    0.000
TX*X        2       53.96     53.96     26.98   0.40    0.676
Error      24     1626.69   1626.69   67.78
Total      29     3956.00

S = 8.23278   R-Sq = 58.88%   R-Sq(adj) = 50.31%

Term         Coef  SE Coef      T      P
Constant    6.855    5.586   1.23  0.232
X          0.5711   0.1092   5.23  0.000
X*TX
1       -0.0731   0.1521  -0.48  0.635
2       -0.0645   0.1565  -0.41  0.684
```

Once again the output format does not match the format discussed previously for the homogeneity of slopes test. Recall that we want the mean square for heterogeneity

of slopes and the mean square for the individual residuals. The third and fourth lines provide these mean squares; the F-value in the third line (i.e., .40) and its p-value are of interest; ignore all of the other output.

The output shown below includes information that will usually be of interest during a thorough inspection of the data. The input data should be inspected to be sure it agrees with the original data, the residuals Res_w (from ANCOVA) and Res_i (from the homogeneity of slopes analysis) should be plotted and inspected, the variance estimates (for each group) around the pooled within-group slope and around the individual slopes should be similar, and Cook's distance (described in Chapter 5) may be helpful in identifying unusually influential observations. Additional diagnostics are described in Chapter 8.

Row	TX	X	Y	RESw	RESi	VARi	VARw	Cook
1	1	29	15	-1.8779	-3.5457	62.7213	63.8980	0.004036
2	1	49	19	-9.2884	-9.5060	48.2689	49.0834	0.041992
3	1	48	21	-6.7179	-7.0079	69.7531	73.7573	0.022292
4	1	35	27	6.6989	5.4662			0.036262
5	1	53	35	4.4295	4.5020			0.009388
6	1	47	39	11.8526	11.4901			0.070708
7	1	46	23	-3.5768	-4.0119			0.006587
8	1	74	38	-4.5516	-2.9563			0.022301
9	1	72	33	-8.4105	-6.9603			0.067609
10	1	67	50	11.4421	12.5298			0.095370
11	2	22	20	-4.7368	-6.3331			0.029125
12	2	24	34	8.1221	6.6535			0.075499
13	2	49	28	-12.1411	-12.0133			0.070987
14	2	46	35	-3.4295	-3.4933			0.005628
15	2	52	42	0.1474	0.4666			0.000011
16	2	43	44	7.2821	7.0267			0.026194
17	2	64	46	-2.6989	-1.6135			0.005886
18	2	61	47	0.0126	0.9066			0.000000
19	2	55	40	-3.5642	-3.0534			0.006918
20	2	54	54	11.0063	11.4533			0.064031
21	3	33	14	-12.8716	-10.6610			0.126994
22	3	45	20	-13.7179	-13.1653			0.092953
23	3	35	30	1.9874	3.9216			0.002740
24	3	39	32	1.7053	3.0869			0.001698
25	3	36	34	5.4168	7.2129			0.019423
26	3	48	42	6.5705	6.7087			0.020658
27	3	63	40	-3.9874	-5.9216			0.011028
28	3	57	38	-2.5642	-3.6695			0.003581
29	3	56	54	14.0063	13.0392			0.103693
30	3	78	56	3.4547	-0.5519			0.020182

Analysis of Training Example Using SPSS The *SPSS* menu commands for computing ANCOVA, residuals, and Cook's distance are shown below. The worksheet columns are named tx for treatments, x for the covariate, and y for the dependent variable.

Analyze → General Linear Model → Univariate → Dependent Variable: *y* → Fixed Factor(s): *tx* → Covariates(s): *x* → Save: Residuals(unstandardized) → Cook's Distance → Continue → Options → Display Means for: *tx* → Continue → OK

SPSS ANCOVA Output

Tests of Between-Subjects Effects

Dependent Variable *y*

Source	Type III Sum of Squares	*df*	Mean Square	*F*	Sig.
Corrected model	2275.352*	3	758.451	11.733	.000
Intercept	102.056	1	102.056	1.579	.220
x	1855.352	1	1855.352	28.703	.000
tx	707.992	2	353.996	5.476	.010
Error	1680.648	26	64.640		
Total	40706.000	30			
Corrected total	3956.000	29			

*R-Squared = .575 (Adjusted R-Squared = .526).

Tx

Dependent Variable: y

Tx	Mean	Std. Error	95% Confidence Interval	
			Lower Bound	Upper Bound
Group 1	28.479*	2.558	23.220	33.737
Group 2	40.331*	2.555	35.080	45.582
Group 3	36.190*	2.543	30.964	41.417

*Covariates appearing in the model are evaluated at the following values: $x = 49.3333$.

The menu commands for computing the homogeneity of regression slopes test are:

Analyze → General Linear Model → Univariate → Dependent Variable: *y* → Fixed Factor(s): *tx* → Covariates(s): *x* → Model → Factors & Covariates: *tx* → Build term(s) → *tx*(*F*) → *x*(*C*) → Build term(s) → *tx*(*F*)**x*(*C*) → Build term(s) → Continue → OK

SPSS Homogeneity of Regression Output

Tests of Between-Subjects Effects					
Dependent Variable: *y*					
Source	Type III Sum of Squares	*df*	Mean Square	*F*	Sig.
Corrected Model	2329.310*	5	465.862	6.873	.000
Intercept	102.072	1	102.072	1.506	.232
Tx	80.157	2	40.078	.591	.561
x	1855.238	1	1855.238	27.372	.000
tx × *x*	53.959	2	26.979	.398	.676
Error	1626.690	24	67.779		
Total	40706.000	30			
Corrected Total	3956.000	29			

*R-Squared = .589 (Adjusted R-Squared = .503).

Analysis of Partridge and Farquhar (1981) Data Using *Minitab*

```
MTB > print c1 c2 c3
Data Display
Row Tx  Size  Longevity
1    1  0.64         16
2    1  0.64         19
3    1  0.68         19
4    1  0.72         32
5    1  0.72         34
6    1  0.74         34
7    1  0.76         30
8    1  0.76         43
9    1  0.76         43
10   1  0.77         34
11   1  0.80         26
12   1  0.80         30
13   1  0.82         41
14   1  0.82         55
15   1  0.84         36
16   1  0.84         36
17   1  0.85         48
18   1  0.85         48
19   1  0.88         43
20   1  0.88         48
21   1  0.88         55
22   1  0.88         57
23   1  0.88         62
24   1  0.88         72
25   1  0.88         55
26   1  0.92         45
27   2  0.64         36
```

```
28   2  0.68         38
29   2  0.68         50
30   2  0.72         47
31   2  0.72         64
32   2  0.76         40
33   2  0.76         47
34   2  0.76         57
35   2  0.76         64
36   2  0.76         66
37   2  0.80         57
38   2  0.80         67
39   2  0.80         75
40   2  0.84         64
41   2  0.84         66
42   2  0.84         72
43   2  0.84         78
44   2  0.84         82
45   2  0.84         88
46   2  0.88         72
47   2  0.92         78
48   2  0.92         78
49   2  0.92         82
50   2  0.94         78
```

ANOVA:

Minitab Commands

```
MTB > Name c4 "RESI1"
MTB > Oneway 'Longevity' 'Tx';
SUBC>   Residuals 'RESI1';
SUBC>   GFourpack.
```

Output

```
One-way ANOVA: Longevity versus Tx
Source  DF     SS    MS      F      P
Tx       1   6956  6956  33.61  0.000
Error   48   9936   207
Total   49  16892

S = 14.39   R-Sq = 41.18%   R-Sq(adj) = 39.95%


Individual 95% CIs For Mean Based on
Pooled StDev
Level   N   Mean  StDev  ----+--------+--------+--------+---
1      26  40.81  13.80  (----*----)
2      24  64.42  15.00                  (----*-----)
           -----+---------+---------+---------+----
40        50        60        70

Pooled StDev = 14.39
```

ANCOVA:
Commands

```
MTB > ancova Longevity=Tx;
SUBC> covariate size;
SUBC> means Tx;
SUBC> residuals c4.
```

Output

```
ANCOVA: Longevity versus Tx
Factor  Levels  Values
Tx           2  1, 2
Analysis of Covariance for Longevity
Source      DF   Adj SS      MS      F      P
Covariates   1   6556.4  6556.4  91.18  0.000
Tx           1   7038.6  7038.6  97.89  0.000
Error       47   3379.4    71.9
Total       49  16892.0

S = 8.47957   R-Sq = 79.99%   R-Sq(adj) = 79.14%
Covariate   Coef  SE Coef      T      P
Size       145.6     15.2  9.549  0.000
```

Adjusted Means

```
Tx   N  Longevity
1   26     40.741
2   24     64.489
```

Homogeneity of Regression Test:
Minitab Commands

```
MTB > glm c3=c1 c2 c1*c2;
SUBC> covariate c2.
```

Output

```
General Linear Model: Longevity versus Tx
Factor  Type   Levels  Values
Tx      fixed       2  1, 2
Analysis of Variance for Longevity, using Adjusted SS
for Tests
```

Source	DF	Seq SS	Adj SS	Adj MS	F	P
Tx	1	6956.1	27.6	27.6	0.38	0.543
Size	1	6556.4	6558.9	6558.9	89.51	0.000
Tx*Size	**1**	**8.7**	**8.7**	**8.7**	**0.12**	**0.732**
Error	46	3370.7	3370.7	73.3		
Total	49	16892.0				

S = 8.56021 R-Sq = 80.05% R-Sq(adj) = 78.74%

6.9 SUMMARY

Performance on a response variable is conceptualized in three different ways under the ANOVA, ANOVAR, and ANCOVA models. Under the ANOVA model the total variability in an experiment is viewed as a function of treatment effects (mean differences) and random fluctuation. Under the ANOVAR model the total variability is viewed as a function of the level of performance on a predictor variable and random fluctuation. Under the ANCOVA model the total variability is viewed as a function of performance on a predictor variable (or covariate), treatment effects that are independent of the predictor variable, and random fluctuation. Hence the ANCOVA model is an integration of the ANOVA and ANOVAR models.

Because one purpose of ANCOVA is to estimate and test differences among adjusted means, it is very important to recognize the factors that affect the adjustment. The relative size of the samples, the pooled within-group regression coefficient, and the mean difference on the covariate all play a part in the adjustment process. There is generally little difference between adjusted and unadjusted means when randomized-group experiments are employed. This is because small differences on the covariate can generally be expected with these designs. When nonrandomized designs are employed, the mean differences on the covariate and hence the degree of adjustment may be large.

The interpretation of ANCOVA and the associated adjusted means relies very heavily on the assumption of homogeneous regression slopes for the various treatment groups. If this assumption is not met the ANCOVA F-test and the adjustment process can lead to highly misleading results. For this reason the homogeneity of slopes test should be carried out whenever ANCOVA is employed.

CHAPTER 7

Analysis of Covariance Through Linear Regression

7.1 INTRODUCTION

The similarity among analysis of variance (ANOVA), analysis of variance of regression (ANOVAR), and analysis of covariance (ANCOVA) models was described in Chapter 6. Because these analyses are all based on linear models, the reader familiar with multiple linear regression will not be surprised to discover that analysis of variance and analysis of covariance problems can be computed with any typical multiple regression computer program.

7.2 SIMPLE ANALYSIS OF VARIANCE THROUGH LINEAR REGRESSION

Simple (one-way) analysis of variance problems can be computed through regression analysis by regressing the dependent variable scores (Y) on so-called dummy variable(s). The dummy variables, which are used to identify group membership, are easily constructed. Suppose we have a simple two-group analysis of variance problem with five subjects in each group. The 10 dependent variable scores are regressed on a column of dummy scores arranged as follows:

The Analysis of Covariance and Alternatives: Statistical Methods for Experiments, Quasi-Experiments, and Single-Case Studies, Second Edition. Bradley E. Huitema.

	D Dummy Variable	Y Dependent Variable	
Group 1 dummy variable scores	1	Y_1	Group 1 dependent variable scores
	1	Y_2	
	1	Y_3	
	1	Y_4	
	1	Y_5	
Group 2 dummy variable scores	0	Y_6	Group 2 dependent variable scores
	0	Y_7	
	0	Y_8	
	0	Y_9	
	0	Y_{10}	

If a subject belongs to the first-treatment group, he or she is assigned a dummy variable score of one. If a subject does not belong to the first-treatment group, he or she is assigned a dummy variable score of zero. Hence, the column of dummy scores simply indicates whether a subject belongs to the first-treatment group. The analysis involves nothing more than regressing the dependent variable scores on the dummy variable scores. The output of the typical regression program will provide the correlation between the two variables, the fitted regression equation, a test of the significance of the regression slope, and/or an ANOVAR.

If the computer output does not include ANOVAR or any other test statistics, the correlation between the dummy variable(s) and the dependent variable provides enough information to carry out the analysis of variance. This is true, regardless of the number of treatment groups.

In the two-group case, the simple correlation between the dummy variable and the dependent variable (r_{yD}) can be tested for significance as follows:

$$\frac{r_{yD}^2/1}{\left(1-r_{yD}^2\right)/(N-2)} = F.$$

Critical value $= F_{(\alpha,\,1,\,N-2)}$.

This F-test is a test of the significance of the point-biserial correlation between the dummy variable and the dependent variable. It turns out that this test is equivalent to an ANOVA F-test on the difference between the means of the two treatments, which, in turn, is equivalent to the independent sample t-test.

Because (1) the F-test of the statistical significance of the regression slope, (2) the F-test for the point-biserial correlation coefficient, (3) the ANOVA F-test on the difference between two means, and (4) the independent sample t on the difference between two means are all equivalent tests, it follows that any one of these tests can

Table 7.1 Example Illustrating Similarities of Tests on Slope, Correlation Coefficient, and Mean Difference

Data:	D Dummy Variable	Y Dependent Variable
	1	3
	1	4
	1	3
	1	2
	1	3
	0	3
	0	12
	0	14
	0	7
	0	9

be substituted for any other one. Likewise, any of these test statistics can be converted into a point-biserial correlation coefficient as follows:

$$\sqrt{\frac{\text{ANOVAR } F}{\text{ANOVAR } F + (N-2)}} = \sqrt{\frac{\text{Point biserial } F}{\text{Point biserial } F + (N-2)}}$$

$$= \sqrt{\frac{\text{ANOVA } F}{\text{ANOVA } F + (N-2)}}$$

$$= \sqrt{\frac{t^2}{t^2 + (N-2)}}$$

$$= r_{\text{point biserial}}$$

It may be helpful to consider why these various tests are equivalent. Suppose the data presented in Table 7.1 were collected in a study of the effectiveness of vigorous exercise in the reduction of angina pain in patients who have had one heart attack. The experimental group is exposed to a carefully monitored exercise routine that involves maintaining the heart rate at 140 for 25 minutes every day for 12 months. The control group is composed of subjects who continue their normal living patterns without an exercise program. Subjects are assigned to the two conditions using a table of random numbers. (The very small sample sizes employed here are not recommended in practice.) The dependent variable is the number of angina attacks experienced by the subjects during the last 2 months of the program.

Analyses

1. ANOVAR on regression slope

Source	SS	*df*	MS	*F*
Regression	90	1	90.0	9.474
Residual	76	8	9.5	
Total	166	9		

$$b_0 = 9.00, b_1 = -6.00 \text{ Regression equation : } \hat{Y} = 9.00 - 6.00(D)$$

2. Point-biserial correlation *F*-test

$$r_{\text{point biserial}} = r_{yD} = 0.73632$$

$$\frac{r_{yD}^2/1}{\left(1 - r_{yD}^2\right)/(N-2)} = \frac{0.54217}{(1 - .54217)/10 - 2} = 9.474 = F_{\text{obt}}$$

3. ANOVA *F*-test

Source	SS	*df*	MS	*F*
Between	90	1	90.0	9.474
Within	76	8	9.5	
Total	166	9		

4. Independent sample *t*-test

$$\frac{\bar{Y}_1 - \bar{Y}_2}{\sqrt{\frac{\sum y_1^2 + \sum y_2^2}{n_1+n_2-2}\left[(1/n_1) + (1/n_2)\right]}} = \frac{3 - 9}{\sqrt{[(2+74)/(5+5-2)]\left(\frac{1}{5} + \frac{1}{5}\right)}}$$

$$= -3.078$$

$$\text{and} \quad t_{\text{obt}}^2 = 9.474 = F_{\text{obt}}$$

The purpose of illustrating the relationships among the tests on the slope, the point-biserial correlation coefficient, and the mean difference is to give the reader some feel for the appropriateness of employing correlation and regression procedures in testing mean differences. It is likely, however, that questions remain concerning *why* the regression of the dependent variable on a dummy variable gives the same *F* as a conventional ANOVA *F*-test. The answer to this question is fairly simple if one of the central concepts in regression analysis, the least-squares criterion, is kept in mind.

Recall that when Y is regressed on X (or D, as we have labeled it here), the intercept (b_0) and the slope (b_1) are fitted to the data in such a way that the sum of the squared discrepancies between the observed and the predicted scores is minimum; that is,

$$\sum_{i=1}^{N} \left(Y_i - \hat{Y}_i\right)^2 \text{ is minimum,}$$

where $\hat{Y}_i = b_0 + b_1 D_i$ and D_i is one or zero, depending on whether the ith observation falls in group 1 or group 2. There is no way that the sum of the squared discrepancies or errors can be less if knowledge of values of D and a linear prediction rule are used. This sum of squared errors is generally called the *sum of squares residuals* in the ANOVAR summary table. Now think about the quantity called the *sum of squares within* in the ANOVA. The sum of squares within is also based on a least-squares criterion.

Suppose you are assigned the task of selecting *one* value for each of the two groups of scores; each value must have the property that the sum of the squared differences between the value selected and the actual scores is minimum. The value you should select for each group is the mean of that group. Any value other than the group mean (in the case of each group) will yield a sum-of-squared-difference scores that is larger than the sum obtained by using the group mean.

Now if (1) the sum-of-squares residuals in a regression analysis is the minimum sum of squared differences between Y and $\hat{Y}$, where $\hat{Y}$ is based on information on the predictor variable D which indicates group membership, and (2) if the sample means provide the minimum sum of squares within each group in a two-group ANOVA, what values do you think the regression equation will predict for each group when Y is regressed on the dummy variable D?

The mean of treatment group 1 is the predicted value when $D = 1$, and the mean of the second-treatment group is the predicted value when $D = 0$; in other words,

$$\bar{Y}_1 = b_0 + b_1 (1)$$
$$\bar{Y}_2 = b_0 + b_1 (0).$$

Note that these formulas do in fact yield the means for the data in Table 4.1, where it can be seen that the prediction equation is

$$\hat{Y} = 9 - 6(D).$$

Because the value of D for the first group is one, the predicted value is

$$\hat{Y} = 9 - 6(1) = 3.$$

Hence, $\bar{Y}_1 = 3$. Likewise, the predicted value associated with the second group, which has a D-value of zero is

$$\hat{Y} = 9 - 6(0) = 9.$$

Hence, $\bar{Y}_2 = 9$. Clearly, the regression of Y on D yields the same information as the ANOVA. The sum-of-squares residual in the ANOVAR is equivalent to the within-group sum of squares in the ANOVA; the sum-of-squares regression in the ANOVAR is equivalent to the sum of squares between groups in the ANOVA. Since the degrees of freedom are the same for these two analyses, it follows that the F also must be the same. If the equivalence of these two analyses is understood, it is not difficult to grasp why the F-test for the point-biserial correlation coefficient is also equivalent.

Recall from elementary correlation analysis that the point-biserial correlation is used to provide a measure of the relationship between a continuous variable and a dichotomous variable. Also recall that a squared correlation coefficient is called a *coefficient of determination*. In the analysis of the data in Table 7.1, the point-biserial correlation between the dependent variable (number of pain attacks, a continuous variable) and the dummy variable D (group membership, a dichotomous variable) is .736. The square of this value is approximately .54, which is the coefficient of determination. This is interpreted as the proportion of the variability in pain frequency that is accounted for on the basis of knowledge of group membership. This notion of the proportion of variability on Y explained on the basis of information on group membership D, can be easily understood if the concept of prediction error is kept in mind.

Suppose we know nothing about which treatment group is associated with the 10 dependent variable scores in Table 7.1. Next, suppose that the list of 10 scores is lost but we know that the grand mean of the 10 lost scores is 6. If our task is to guess the score associated with each one of the 10 subjects, our best guess for each is the grand mean 6. This guess is the "best" in a least-squares sense. There is no *single* value that we can guess that will provide a smaller sum of squared differences between the value guessed (predicted) and the actual score. This sum of squared prediction errors is, of course, the total sum of squares in ANOVA and ANOVAR. In the example, the total sum of squares is

$$\sum_{i=1}^{10} \left(Y_i - \bar{Y}_{..}\right)^2 = 166.$$

Hence, if we predict that each of the 10 subjects has a score of 6 and then later locate the 10 actual obtained scores, we will discover that the total sum of squares of the obtained scores is the same as the sum of the squared prediction errors we have made by predicting the grand mean for each subject.

At this stage we might ask how much better our prediction would be if we had knowledge of the treatment to which each subject had been exposed. If the information concerning the treatment exposure is available, we simply include that information in a regression analysis in the form of the group membership dummy variable D.

If there is a difference between the two sample means, the sample slope b_1 will not be zero and the value predicted by the regression equation will not be the grand mean; rather, the predicted value will be the mean of the treatment group to which an individual belongs. As long as the treatment means are different, the sum of the

prediction errors or residuals will be smaller if the values predicted by the regression equation (i.e., the group means) rather than the grand mean are used.

This difference between the total sum of squares and the residual sum of squares in the regression analysis yields the regression sum of squares; that is,

$$SS_{total} - SS_{Res} = SS_{Reg}.$$

Because $SS_{total} = SS_{Reg} + SS_{Res}$, it follows that the ratio SS_{Reg}/SS_{total} is *the proportion of the variation on Y that is accounted for by group membership D*. This ratio, then, is interpreted in the same way as the squared point-biserial correlation coefficient; the two are equivalent:

$$\frac{SS_{Reg}}{SS_{total}} = r^2_{\text{point biserial}}.$$

Alternatively, it turns out that

$$\left(1 - r^2_{\text{point biserial}}\right) SS_{total} = SS_{Res}$$
$$\left(r^2_{\text{point biserial}}\right) SS_{total} = SS_{Reg}.$$

Hence, it can be seen that the ANOVA, the ANOVAR, and the point-biserial correlation F-tests are all equivalent. It was also pointed out that the independent sample t-test yields the same information as do the F-tests mentioned earlier. Because most elementary statistics texts contain a proof that $t^2 =$ ANOVA F in the case of two groups, this correspondence is not pursued here.

ANOVA Through Regression with Three or More Groups

If more than two groups are involved, the basic changes in the analysis procedure described above are that (1) there are more dummy variables, and (2) the correlation between the dummy variables and the dependent variable is a multiple correlation rather than a simple correlation. The number of dummy variables, regardless of the number of treatments, is $J - 1$. Hence, the number of dummy variables in a three-group problem is two. The data layout for a three-group analysis of variance through multiple regression is provided in Table 7.2 for the first ANOVA example data set mentioned in Chapter 3.

The steps followed in arranging the data were

Step 1. All dependent variable scores were entered under the Y column heading.

Step 2. The number of dummy variables required was computed ($3 - 1 = 2$), and column headings D_1 and D_2 were entered next to the Y column heading.

Step 3. A "one" was entered in column D_1 for each dependent variable score associated with the first-treatment group; a "zero" was entered in this column for each dependent variable score not associated with the first-treatment group.

Table 7.2 Data Layout for ANOVA Through Multiple Regression Analysis

D_1	D_2	Y	
1	0	15	Group 1
1	0	19	
1	0	21	
1	0	27	
1	0	35	
1	0	39	
1	0	23	
1	0	38	
1	0	33	
1	0	50	
0	1	20	Group 2
0	1	34	
0	1	28	
0	1	35	
0	1	42	
0	1	44	
0	1	46	
0	1	47	
0	1	40	
0	1	54	
0	0	14	Group 3
0	0	20	
0	0	30	
0	0	32	
0	0	34	
0	0	42	
0	0	40	
0	0	38	
0	0	54	
0	0	56	

Step 4. A "one" was entered in column D_2 for each dependent variable score associated with the second-treatment group; a "zero" was entered in this column for each dependent variable score not associated with the second-treatment group.

Note that the first 10 rows in the D_1 column contain "ones." This is because the first 10 subjects belong to the first-treatment group. Since the next 20 subjects do not belong to the first-treatment group, they are each assigned a "zero." The first 10 rows in column D_2 are "zeros" because the first 10 subjects do not belong to the

second-treatment group. Subjects 11 through 20 are assigned "ones" in column D_2 because they do belong to the second-treatment group. Hence, the "ones" in column D_1 indicate which subjects belong to treatment 1. The "ones" in column D_2 indicate which subjects belong to treatment group 2. Obviously, if a subject has "zeros" in both columns D_1, and D_2, he belongs to group 3.

Each dummy variable is treated as a predictor variable in a multiple regression analysis; that is, the dependent variable scores are regressed on the dummy variables. The output of the typical multiple regression program will contain, as a minimum, the multiple correlation coefficient (R) and the fitted multiple regression equation. The analysis of variance can be carried out by employing the following general formula for the ANOVA F test:

$$\frac{R^2_{yD_1,D_2,\ldots,D_{J-1}}/m}{\left(1-R^2_{yD_1,D_2,\ldots,D_{J-1}}\right)/(N-m-1)} = F.$$

Critical value $= F_{(\langle, m, N-m-1)}$,

where

$R^2_{yD_1,D_2,\ldots,D_{J-1}}$ is multiple correlation coefficient between the dummy variables and the dependent variable;
m is the number of dummy variables; and
N is the total number of subjects.

The rationale for this formula is quite straightforward. Recall that the basic sum-of-squares partition in the analysis of variance is between-group + within-group = total. The squared multiple R provides the proportion of the total variability in Y that is accounted for by the dummy variables. This turns out to be equivalent to the proportion of the total variability that is between-group variability. It follows that $1 - R^2$ must provide the proportion of the total variability that is within-group or error variability. The F-ratio then becomes

$$\begin{aligned}\frac{R^2/\text{between-group degrees of freedom}}{\left(1-R^2\right)/\text{within-group degrees of freedom}} &= \frac{R^2/(J-1)}{\left(1-R^2\right)/(N-J)}\\ &= \frac{R^2/m}{\left(1-R^2\right)/(N-m-1)}\\ &= F\end{aligned}$$

If we multiply the total sum of squares (SST) by R^2 and $1 - R^2$, we obtain the between-and within-group sum of squares, respectively. If these sum of squares are computed, the equivalence of the conventionally computed ANOVA and ANOVA through multiple regression may become more obvious. The ANOVA summary using this approach is tabulated as follows:

Source	SS	*df*	MS	*F*
Between	R^2(SST)	$J-1$	$\frac{R^2\,(\mathrm{SST})}{J-1}$	$\frac{\mathrm{MS}_b}{\mathrm{MS}_w}$
Within	$(1-R^2)$SST	$N-$ J	$\frac{\left(1-R^2\right)\mathrm{SST}}{N-J}$	
Total	(1)SST	$N-1$		

In addition to the multiple correlation coefficient, the fitted multiple regression prediction equation is provided by the computer output. Recall that this equation is generally written as

$$\hat{Y} = b_0 + b_1X_1 + b_2X_2 + \cdots + b_mX_m,$$

where

$\hat{Y}$ is the predicted score;
b_0 is the multiple regression intercept;
$b_1 - b_m$ are the partial regression coefficients associated with $X_1 - X_m$; and
$X_1 - X_m$ are the predictor or independent variables.

Because I have labeled the dummy variable columns with *D*s rather than *X*s, the equation can be written as

$$\hat{Y} = b_0 + b_1D_1 + b_2D_2 + \cdots + b_{J-1}D_{J-1},$$

where

$\hat{Y}$ is the predicted score (which turns out to be mean of treatment identified by dummy variables);
b_0 is the multiple regression intercept;
$b_1 - b_{J-1}$ are the partial regression coefficients associated with the dummy variables; and
$D_1 - D_{J-l}$ are the $J-1$ dummy variables required to identify group membership.

The information contained in the multiple regression equation can be used to compute the sample means as follows:

$$\begin{aligned}
\bar{Y}_1 &= b_0 + b_1 \\
\bar{Y}_2 &= b_0 + b_2 \\
&\cdot \quad \cdot \quad \cdot \\
&\cdot \quad \cdot \quad \cdot \\
&\cdot \quad \cdot \quad \cdot \\
\bar{Y}_{J-1} &= b_0 + b_{J-1} \\
\bar{Y}_J &= b_0
\end{aligned}$$

The rationale for this multiple regression approach leading to the same results in multiple-group experiments as are obtained with a conventional analysis of variance is the same as was described previously for two-group experiments. Recall that in the two-group case, the dependent variable is regressed on one dummy variable that indicates membership in one of the two groups. The residual sum of squares associated with this simple regression is equivalent to the within-group sum of squares. Likewise, when three or more groups are involved, the dependent variable is regressed on the collection of dummy variables that is required to indicate group membership.

Keep in mind that the parameter estimates in the multiple regression equation (i.e., the intercept and the partial regression coefficients) are computed in such a way that the sum of the squared prediction errors is a minimum; i.e.,

$$\sum_{1}^{N} \left(Y_i - \hat{Y}_i\right)^2 \text{ is a minimum,}$$

where Y_i is the observed or actual dependent variable score for the ith subject, $\hat{Y}_i$ is the score predicted for the ith subject using the fitted multiple regression equation, and N is the total number of subjects in the experiment.

Just as the regression of Y on a single dummy variable in the two-group case yields a prediction equation that predicts the mean of the group to which a subject belongs, the multiple regression of Y on a collection of dummy variables yields a multiple regression equation that predicts the mean of the group to which a subject belongs. The only difference between the two-group case and the case of more than two groups is that the former involves one dummy variable and simple regression where the latter involves more than one dummy variable and multiple regression. Because the least-squares criterion is associated with both simple and multiple regression, it makes sense that group means are predicted regardless of the number of groups in an experiment. This is true because there is no single value other than the mean associated with a given group that will meet the least-squares criterion. Another parallel between the case where $J = 2$, and the case where $J > 2$ can be seen in the correlation values associated with the simple and multiple regressions.

Just as r^2 in the two-group case yields the proportion of the total variation on Y accounted for by knowledge of group membership, the coefficient of multiple determination (R^2) yields the proportion of the total variation on Y accounted for by knowledge of group membership in the case of more than two groups. Similarly, the product

$$r^2(\text{SST}) = \text{between-group SS for 2-group experiments}$$

and the product $R^2(\text{SST}) =$ between or among group SS for experiments with > 2 groups.

The regression analysis on the data presented above results in the following:

$$R = .32583$$
$$R^2 = .10617$$
$$b_0 = 36$$
$$b_1 = -6$$
$$b_2 = 3$$

The regression intercept (b_0) and the partial regression coefficients (b_1 and b_2) provide the constants required in the computation of the predicted scores. As was just mentioned, the predicted score for any subject will be equal to the mean of the group to which the subject belongs. Hence, the predicted score (or group mean) for any subject is obtained by entering the fitted multiple regression equation $\hat{Y} = b_0 + b_1 D_1 + b_2 D_2 = 36 - 6D_1 + 3D_2$ with the relevant predictor values. Because each subject in the experiment has a D_1 score and a D_2 score, we simply enter the equation with these scores to obtain the predicted score for that subject (or to obtain the mean of the group to which the subject belongs). For example, the first subject in the first group has a D_1 score of one and a D_2 score of zero. This means that the predicted score for this subject, or any subject in the first group, is

$$b_0 + b_1 (1) + b_2 (0) = 36 - 6 (1) + 3 (0) = 30.$$

This reduces to

$$b_0 + b_1 = 36 - 6 = 30.$$

Hence, the mean of the first group is 30. Because the subjects in the second group have $D_1 = 0$ and $D_2 = 1$, the predicted score is

$$b_0 + b_1 (0) + b_2 (1) = 36 - 6 (0) + 3 (1) = 39.$$

This reduces to $b_0 + b_2 = 36 + 3 = 39$; therefore, the mean of the second group is 39. The dummy scores for subjects in group 3 are $D_1 = 0$ and $D_2 = 0$. The predicted score is

$$b_0 + b_1 (0) + b_2 (0) = 36 - 6 (0) + 3 (0) = 36.$$

This reduces to $b_0 = 36$.

Thus, it can be seen that the three-group means are

$$\bar{Y}_1 = b_0 + b_1$$
$$\bar{Y}_2 = b_0 + b_2$$
$$\bar{Y}_3 = b_0.$$

Table 7.3 Actual, Predicted, Deviation, and Squared Deviation Scores for Example Data

Subject	Actual Y	Predicted $\hat{Y}$	Deviation $Y - \hat{Y}$	(Deviation)^2 $(Y - \hat{Y})^2$		
1	15	30	−15	225		
2	19	30	−11	121		
3	21	30	−9	81		
4	27	30	−3	9		
5	35	30	5	25	1064	
6	39	30	9	81		
7	23	30	−7	49		
8	38	30	8	64		
9	33	30	3	9		
10	50	30	20	400		
11	20	39	−19	361		
12	34	39	−5	25		
13	28	39	−11	121		
14	35	39	−4	16		
15	42	39	3	9		
16	44	39	5	25	896	3536
17	46	39	7	49		
18	47	39	8	64		
19	40	39	1	1		
20	54	39	15	225		
21	14	30	−22	484		
22	20	30	−16	256		
23	30	30	−6	36		
24	32	30	−4	16		
25	34	30	−2	4		
26	42	30	6	36	1576	
27	40	30	4	16		
28	38	30	2	4		
29	54	30	18	324		
30	56	30	20	400		

The predicted scores ($\hat{Y}$), actual scores (Y), prediction errors ($Y - \hat{Y}$), and squared prediction errors $(Y - \hat{Y})^2$ for all subjects are presented in Table 7.3. Note that the sum of the squared deviation scores is exactly the same as the within-group sum of squares that was computed using a conventional analysis of variance procedure in Chapter 3.

The analysis is summarized as follows:

Source	SS		*df*	MS	*F*
Between-group	(0.10617)3956	= 420.00	2	210.00	1.60
Within-group	(1 − 0.10617)3956	= 3536.00	27	130.96	
Total		3956.00			

or

$$\frac{0.10617/2}{(1-0.10617)/27} = 1.60 = F.$$

7.3 ANALYSIS OF COVARIANCE THROUGH LINEAR REGRESSION

Once the regression approach to ANOVA problems is mastered, the analysis of covariance can be easily conceptualized as a slight extension of the same ideas. As the models in Table 7.4 indicate, ANCOVA differs from ANOVA only in that it contains a term for the regression of Y on the covariate (X). This means that we can carry out the analysis of covariance by using a regression program by regressing the Y scores on both the dummy variable scores (necessary to designate group membership or treatments) and the covariate scores. The coefficient of multiple determination $\left(R^2_{yD_1,\ldots,D_{J-1},X}\right)$ resulting from regressing Y on dummy variables and the covariate yields the proportion of the total variability accounted for by both group membership and covariate scores. Then Y is regressed on the covariate alone to obtain the proportion of the total variability accounted for by X (i.e., r^2_{yX}). The difference between the two coefficients of determination indicates the unique contribution of the dummy variables to the first computed coefficient.

This procedure has been applied in the analysis of the ANOVA Example 1 data described previously. The results appear in the right-hand column of Table 7.4 in two forms. The F-values obtained under the two forms of the analysis presented in this table are the same as those obtained using the traditional solution (see Table 3.2). Table 7.4 also contains the predictor variables and summaries for the ANOVA on Y and the ANOVAR. The reader should carefully compare the similarities and differences between the predictors and computations associated with the three models.

The Y columns are of course the same under the three models because the purpose of this example is to illustrate how a given set of dependent variable scores is analyzed under ANOVA, ANOVAR, and ANCOVA. The first two columns under the ANOVA are the dummy variables, which have already been explained (Section 7.2). Regressing Y on the dummy variables results in an ANOVAR summary table with regression and residual sum of squares and degrees of freedom identical to the between- and within-groups sums of squares and degrees of freedom in a conventional ANOVA on Y. This ANOVA is not of particular interest here, but the dummy variables are. Note that these dummy variables again appear under the ANCOVA model. Also note that the X column under the ANOVAR is identical to the X column under the ANCOVA model.

The regression approach to ANCOVA on these data proceeds as follows:

1. Y is regressed on X and r^2_{yX} is obtained. The $r^2_{yX} = .396198$.
2. Y is regressed on the two dummy variables and X. $R^2_{yD_1,D_2,X} = .575164$.

Model								
ANOVA $Y_{ij} = \mu + \alpha_j + \varepsilon_{ij}$			ANOVAR $Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i$		ANCOVA $Y_{ij} = \mu + \alpha_j + \beta_1 \left(X_{ij} - \bar{X}..\right) + \varepsilon_{ij}$			
Data Layout for Example Problem								
D_1	D_2	Y	X	Y	D_1	D_2	X	Y
1	0	15	29	15	1	0	29	15
1	0	19	49	19	1	0	49	19
1	0	21	48	21	1	0	48	21
1	0	27	35	27	1	0	35	27
1	0	35	53	35	1	0	53	35
1	0	39	47	39	1	0	47	39
1	0	23	46	23	1	0	46	23
1	0	38	74	38	1	0	74	38
1	0	33	72	33	1	0	72	33
1	0	50	67	50	1	0	67	50
0	1	20	22	20	0	1	22	20
0	1	34	24	34	0	1	24	34
0	1	28	49	28	0	1	49	28
0	1	35	46	35	0	1	46	35
0	1	42	52	42	0	1	52	42
0	1	44	43	44	0	1	43	44
0	1	46	64	46	0	1	64	46
0	1	47	61	47	0	1	61	47
0	1	40	55	40	0	1	55	40
0	1	54	54	54	0	1	54	54
0	0	14	33	14	0	0	33	14
0	0	20	45	20	0	0	45	20
0	0	30	35	30	0	0	35	30
0	0	32	39	32	0	0	39	32
0	0	34	36	34	0	0	36	34
0	0	42	48	42	0	0	48	42
0	0	40	63	40	0	0	63	40
0	0	38	57	38	0	0	57	38
0	0	54	56	54	0	0	56	54
0	0	56	78	56	0	0	78	56

Table 7.4 ANOVA, ANOVAR, and ANCOVA as General Linear Models (*Continued*)

ANOVA					ANOVAR					ANCOVA				
Source	SS		*df*	MS	*F*	Source	SS	*df*	MS	*F*	Source	SS	*df*	MS	*F*
Between	$= \left(R^2_{yD_1,D_2}\right)$ SST $= (0.106165)\,3956$ $= 420.00$		2	210.00	1.60	Regression	$= \left(r^2_{yX}\right)$ SST $= (0.396195)\,3956$ $= 1567.36$	1	1567.36	18.37	AT	$= \left(R^2_{yD_1,D_2,X} - r^2_{yX}\right)$ SST $= (0.575164 - .396198)\,3956$ $= 707.99$	2	354.00	5.48
Within	$= \left(1 - R^2_{yD_1,D_2}\right)$ SST $= (1 - 0.106165)\,3956$ $= 3536.00$		27	130.96		Residuals	$= \left(1 - r^2_{yX}\right)$ SST $= (1 - 0.396195)\,3956$ $= 2388.64$	28	85.31		Res_w	$= \left(1 - R^2_{yD_1,D_2,X}\right)$ SST $= (1 - 0.575164)\,3956$ $= 1680.65$	26	64.64	
Total	$= 3956.00$		29			Total	$= 3956$	29			Res_t	$= \left(1 - r^2_{yX}\right)$ SST $= (0.603802)\,3956$ $= 2388.64$	28		
Abbreviated Summary															
$\dfrac{R^2_{yD_1,D_2,X}/2}{\left(1 - R^2_{yD_1,D_2,X}\right)/27} = \dfrac{0.0531}{0.0331} = 1.60 = F$						$\dfrac{r^2_{yX}/1}{\left(1 - r^2_{yX}\right)/28} = \dfrac{0.3962}{0.02156} = 18.37 = F$					$\dfrac{R^2_{yD_1,D_2,X} - r^2_{yX}/(J-1)}{\left(1 - R^2_{yD_1,D_2,X}\right)/(N-J-1)} = \dfrac{0.089483}{0.016340} = 5.48 = F$				

Table 7.5 Data Layout for Homogeneity of Regression Test

D_1	D_2	X	D_1X	D_2X	Y
1	0	29	29	0	15
1	0	49	49	0	19
1	0	48	48	0	21
1	0	35	35	0	27
1	0	53	53	0	35
1	0	47	47	0	39
1	0	46	46	0	23
1	0	74	74	0	38
1	0	72	72	0	33
1	0	67	67	0	50
0	1	22	0	22	20
0	1	24	0	24	34
0	1	49	0	49	28
0	1	46	0	46	35
0	1	52	0	52	42
0	1	43	0	43	44
0	1	64	0	64	46
0	1	61	0	61	47
0	1	55	0	55	40
0	1	54	0	54	54
0	0	33	0	0	14
0	0	45	0	0	20
0	0	35	0	0	30
0	0	39	0	0	32
0	0	36	0	0	34
0	0	48	0	0	42
0	0	63	0	0	40
0	0	57	0	0	38
0	0	56	0	0	54
0	0	78	0	0	56

3. The value r^2_{yX} is subtracted from $R^2_{yD_1,D_2,X}$. Because r^2_{yX} is the proportion of the total variation on Y that is accounted for by the covariate, and $R^2_{yD_1,D_2,X}$ is the proportion accounted for by both the covariate and the independent variable (i.e., the dummy variables), the difference between these two coefficients of determination must reflect the proportion of the total variation that is uniquely accounted for by the independent variable. Hence, $0.575164 - 0.396198 = 0.178966$, which is the AT (adjusted treatment) effect expressed as a proportion. The sum of squares for adjusted treatments (SSAT) can be obtained by multiplying the difference $\left(R^2_{yD_1,D_2,X} - r^2_{yX}\right)$ times the total sum of squares (SST). In this case $(0.178966)3956 = 707.99$.
4. The proportion of the total variation on Y that is not explained by the independent variable or the covariate is $1 - R^2_{yD_1,D_2,X}$. We can obtain the within-group residual sum of squares (SS_{Res_w}) by multiplying $(1 - R^2_{yD_1,D_2,X})$ times the total sum of squares. Hence, $(0.424836)3956 = 1680.65$.
5. The sum of the AT sum of squares and the within-group residual sum of squares is the total residual sum of squares. As a computational check, the SS_{Res} can also be computed by multiplying $1 - r^2_{yX}$ times the total sum of squares (SST). In the example, $707.99 + 1680.65 = 2388.64$ and $(0.603802)3956 = 2388.64$.

As is the case with ANOVA and ANOVAR, results of ANCOVA are generally presented in conventional summary tables even though regression procedures are employed. It is possible to compute the F-ratio by using only R^2 coefficients and degrees of freedom as shown at the bottom of Table 7.4. Thus, the general form for the summary of the simple (one factor one covariate) ANCOVA is

Source	SS	*df*	MS	*F*
Adjusted treatment	$(R^2_{yD_1,\ldots,D_{J-1},X} - r^2_{yX})$SST	$J - 1$	MS_{AT}	MS_{AT}/MS_{Res_w}
Within residual	$(1 - R^2_{yD_1,\ldots,D_{J-1},X})$ SST	$N - J - 1$	MS_{Res_w}	
Total residual	$(1 - r^2_{yX})$ SST	$N - 2$		

or simply

$$\frac{\left(R^2_{yD_1,\ldots,D_{J-1},X} - r^2_{yX}\right)/(J-1)}{\left(1 - R^2_{yD_1,\ldots,D_{J-1},X}\right)/(N-J-1)} = F,$$

where

$R^2_{yD_1,\ldots,D_{J-1},X}$ is the coefficient of multiple determination obtained by regressing the dependent variable on all group membership dummy variables and covariate;
r^2_{yX} is the coefficient of determination obtained by regressing the dependent variable on covariate;
J is the number of groups; and
N is the total number of subjects and obtained F is evaluated with $F_{(\alpha, J-1, N-J-1)}$.

7.4 COMPUTATION OF ADJUSTED MEANS

The equation associated with the regression of Y on the dummy variables and the covariate provides the information required for the computation of the adjusted means. The adjusted mean for any group can be shown to be equal to $b_0 + b_1(D_1) + \cdots + b_{J-1}(D_{J-1}) + b(\bar{X}..)$. A formal proof can be found elsewhere (Watcharotone et al., 2010).

For the example data, the regression of Y on D_1, D_2, and X results in the following regression intercept and weights:

$$b_0 = 8.0442099$$
$$b_1 = -7.71158$$
$$b_2 = 4.14105$$
$$b_3 = 0.570526.$$

The grand mean on X is 49.33. Hence, $\bar{Y}_{j\text{adj}} = 8.0442099 - 7.7\ 1158(D_1) + 4.14105(D_2) + 0.570526(\bar{X}..)$. The adjusted means for the three treatment groups are

$$\bar{Y}_{1\text{adj}} = 8.0442099 - 7.71158(1) + 4.14105(0) + 0.570526(49.33) = 28.48;$$
$$\bar{Y}_{2\text{adj}} = 8.0442099 - 7.71158(0) + 4.14105(1) + 0.570526(49.33) = 40.33; \text{ and}$$
$$\bar{Y}_{3\text{adj}} = 8.0442099 - 7.71158(0) + 4.14105(0) + 0.570526(49.33) = 36.19.$$

These adjusted means are the same as those presented in Table 6.2, which were based on the conventional adjusted mean formula.

7.5 SIMILARITY OF ANCOVA TO PART AND PARTIAL CORRELATION METHODS

The reader familiar with multiple correlation analysis will recognize the similarities among ANCOVA, multiple correlation, part correlation, and partial correlation procedures. The ANCOVA procedure described in the previous section is the same as the procedure for testing the difference between simple and multiple correlation coefficients based on the same sample. It follows, then, that ANCOVA can be conceptualized as a variety of part or partial correlation. That is, for a two-group case, the ANCOVA F test is the same as the F test of the significance of the partial correlation $r_{yX_2 \cdot X_1}$. If X_2 is viewed as the dummy variable, we can rewrite the partial correlation between the dependent variable and the dummy variable, holding constant the covariate (X) as $r_{yX_2 \cdot X_1}$. Also, an analogy can be drawn between ANCOVA and the part correlation $r_{y(D \cdot X)}$ as shown in Figure 7.1.

Hence, in the two-group situation (1) the conventional ANCOVA F-test, (2) the F-test of significance of the difference between $R_{yD,\,X}$ and r_{yX}, (3) the F-test of the significance of the difference between the partial correlation $r_{yD \cdot X}$ and zero, and (4)

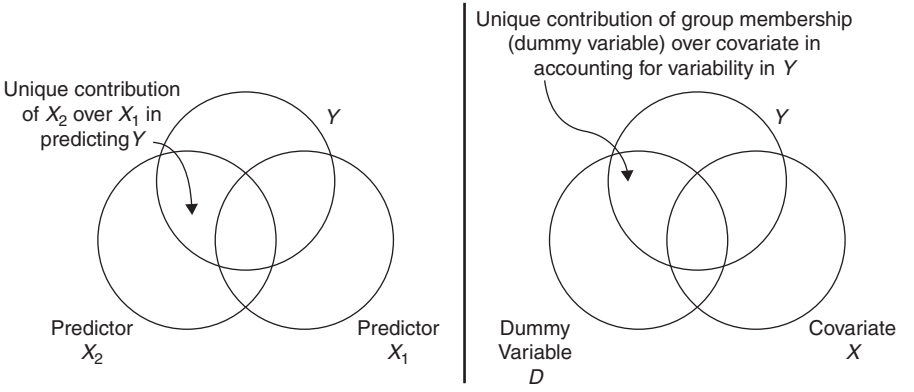


Figure 7.1 Comparison of partitioning for part correlation analysis and ANCOVA through regression.

the F-test of the significance of the difference between the part $r_{y(D\cdot X)}$ and zero, are all equivalent.

When more than two groups are involved, the number of dummy variables will be greater than one, and the following F-tests are equivalent:

1. ANCOVA F.
2. F-test of difference between multiple $R^2_{yD_1,\ldots,D_{J-1},X}$ and r^2_{yX}.
3. F-test of significance of squared multiple partial correlation $R^2_{yd_1,\ldots,d_{J-1}\cdot x}$.
4. F-test of significance of squared multiple part correlation $R^2_{y(d_1,\ldots,d_{J-1}\cdot x)}$.

Because all these tests are equivalent, it seems reasonable to learn only one. It will be seen in Chapter 10 that the F-test of the difference between $R^2_{yD_1,\ldots,D_{J-1},X}$ and r^2_{yX} that has been used in this chapter, i.e.,

$$\frac{\left(R^2_{yD_1,\ldots,D_{J-1},X} - r^2_{yX}\right)/(J-1)}{\left(1 - R^2_{yD_1,\ldots,D_{J-1},X}\right)/(N-J-1)} = F$$

is easily generalized to the case in which multiple covariates are employed.

7.6 HOMOGENEITY OF REGRESSION TEST THROUGH GENERAL LINEAR REGRESSION

The predictors shown in Table 7.4 allow us to compute ANOVA on X, ANOVA on Y, ANOVAR, and ANCOVA through the general linear regression approach. The homogeneity of regression analysis is not included on this list. This analysis can be accomplished through a simple extension of the ANCOVA procedure just described. An expanded version of the upper portion of Table 7.4 is presented in Table 7.5. The

additional columns are labeled D_1X and D_2X. These columns are simply the products of the values of the associated columns. The first subject has a D score of 1 and an X score of 29. Hence, her D_1X score $= 1(29) = 29$. Her D_2 score is zero; therefore, the D_2X score is zero. After the predictor columns are filled in, Y is regressed on all of them. The coefficient of multiple determination $R^2_{yD_1,\ldots,D_{J-1},X,D_1X,\ldots,D_{J-1}X}$ resulting from this regression gives us the proportion of the total variation on Y that is explained by the dummy variables, the covariate, and the interaction of dummy variables with the covariate. It turns out that the interaction of dummy variables with the covariate is the heterogeneity of the regression slopes. Thus, the difference between (1) the R^2 based on dummy variables, the covariate, and the interaction of dummy variables with the covariate, and (2) the R^2 based on only dummy variables and the covariate must reflect the extent to which there are heterogeneous regression slopes. The homogeneity of regression test statistic can be written as

$$\frac{\left(R^2_{yD_1,\ldots,D_{J-1},X,D_1X,\ldots,D_{J-1}X} - R^2_{yD_1,\ldots,D_{J-1},X}\right)/(J-1)}{\left(1 - R^2_{yD_1,\ldots,D_{J-1},X,D_1X,\ldots,D_{J-1}X}\right)/(N-[2J])} = F,$$

where $R^2_{yD_1,\ldots,D_{J-1},X,D_1X,\ldots,D_{J-1}X}$ is the coefficient of multiple determination obtained by regressing the dependent variable on the group membership dummy variables, the covariate, and the products of group membership dummy variables times the covariate. Henceforth this coefficient is written as $R^2_{y\mathbf{D},\mathbf{X},\mathbf{DX}}$. The coefficient of multiple determination $R^2_{yD_1,\ldots,D_{J-1},X}$ is obtained by regressing the dependent variable on the group membership dummy variables and the covariate. Henceforth this coefficient is written as $R^2_{y\mathbf{D},\mathbf{X}}$.

The total number of subjects is N, the number of groups is J, and the obtained F is evaluated using the critical value $F_{(\alpha,\,J-1,\,N-[2J])}$. For the example data, the required R^2-values are $R^2_{y\mathbf{D},\mathbf{X},\mathbf{DX}} = R^2_{yD_1,D_2,X,D_1X,D_2X} = 0.588800$ and $R^2_{y\mathbf{D},\mathbf{X}} = R^2_{yD_1,D_2,X} = 0.575164$. The difference is 0.01364, and the test is

$$\frac{0.01364/2}{0.4112/24} = \frac{0.00682}{0.01713} = 0.398 = F.$$

The obtained F is less than the critical value of $F_{(.05,\,2,\,24)}$, and the null hypothesis is retained; there are insufficient data to conclude that the regression slopes are not homogeneous. Note that this result is in perfect agreement with the result obtained in Section 6.5 that was based on the conventional computation procedure.

7.7 SUMMARY

The analysis of covariance, the analysis of variance, and simple regression analysis are all special cases of the general linear model. Multiple regression software can be conveniently employed to carry out the analysis of covariance and the homogeneity of regression slope tests, if dedicated routines for ANCOVA are not available. It

is necessary to construct appropriate dummy variables if the multiple regression approach is to be utilized.

Dummy variables are used to identify the group to which an individual belongs. Because group membership is the independent variable in an analysis of variance problem, it is possible to carry out the ANOVA F-test by regressing the dependent variable on the dummy variables. The conventional F-test associated with such a regression analysis is equivalent to a conventional ANOVA F-test. Likewise, with covariance analysis, the regression approach can be followed.

The dependent variable is regressed on both the dummy variables and the covariate as the first step in ANCOVA. One result of this regression is an estimate of the proportion of the variability that is accounted for by group membership and the covariate combined. The next step is to regress the dependent variable on the covariate alone. An estimate of the proportion of the variation on the dependent variable that is accounted for by just the covariate is obtained from this regression. The difference between the proportion found in the first regression (on the dummy variables and the covariate) and the proportion found in the second regression (on the covariate only) yields the proportion of the total variation that can be attributed to the treatment groups independent of the covariate. The F-test on this difference in proportions is equivalent to the ANCOVA F-test on differences between adjusted means.

The homogeneity of regression slopes test can also be computed using regression. This involves the regression of the dependent variable on the dummy variables, the covariate, and the products of the dummy variables and the covariate. The difference between the proportion of the total variation that is accounted for by this regression and the proportion that is accounted for by regressing on the dummy variables and the covariate yields the proportion accounted for by covariate-treatment interaction. This turns out to be the proportion due to heterogeneity of regression slopes. Hence, the F-test on this proportion is equivalent to the conventional homogeneity of regression F-test.

CHAPTER 8

Assumptions and Design Considerations

8.1 INTRODUCTION

Statistical assumptions refer to the properties under which a model is mathematically derived; they are not the same thing as properties of the design but they are often related. If the statistical assumptions of the ANCOVA model are met, one can be confident that the ratio of the mean square for the effect over the mean square for error is distributed as F under the null hypothesis. Hence, the test should have the claimed statistical properties. But this does not necessarily mean that the analysis has answered the researcher's question. The adequacy of an application of a statistical analysis depends upon both (1) the conformity of the data with the assumptions of the model, and (2) the alignment of the parameters of the model with the parameters actually of interest to the researcher.

The match between the estimated parameters (i.e., the estimates provided in by the analysis) and the parameters of interest (i.e., what the researcher wants to know) should be the top priority in an adequate analysis. If they are not aligned, the analysis is likely to be misleading (and perhaps useless) irrespective of how well the data conform to the statistical assumptions of the model. Whereas assumption departures are routinely identified using standard diagnostic procedures for regression models, the identification of invalidating design properties requires attention to the research question, methods of data collection, and the nature of the data actually collected. Additional distinctions between assumptions and design considerations are discussed subsequently.

The Analysis of Covariance and Alternatives: Statistical Methods for Experiments, Quasi-Experiments, and Single-Case Studies, Second Edition. Bradley E. Huitema.

8.2 STATISTICAL ASSUMPTIONS

An examination of the linear model $Y_{ij} = \mu + \alpha_j + \beta_1(X_{ij} - \bar{X}..) + \varepsilon_{ij}$ underlying the one-factor ANCOVA makes it clear that dependent variable scores are conceptualized as the sum of the terms of the model. That is, each observation (Y_{ij}) is viewed as the sum of four components: (1) the grand mean (μ) of all dependent variable observations in the population, (2) the effect of the jth treatment ($\alpha_j = \mu_j - \mu$), (3) the effect of the covariate [$\beta_1(X_{ij} - \bar{X}..)$], and (4) the error, where the error is the deviation of the observation from the adjusted mean of the population to which it belongs, i.e., $\varepsilon_{ij} = (Y_{ij} - \mu_{j\,\text{adj}})$. Because both treatment and covariate effects are in the model simultaneously, each effect is statistically independent of the other. The statistical assumptions for the "normal error" ANCOVA model are relatively straightforward because it is simply another linear model. It is assumed that

1. The errors are independent.
2. The within-group slopes are homogeneous.
3. The errors have a mean of zero regardless of the treatment or the level of X.
4. The errors are normally distributed.
5. The variance of the errors is the same regardless of either the treatment level or X.
6. The treatment levels and the X are fixed.

Independent Errors

The errors ε_{ij} in the ANCOVA model are assumed to be independent. Independence is well approximated in most experiments where the subjects are treated independently, but certain types of applied research are likely to yield dependent errors. The issue here is deciding whether the subjects within treatment groups are responding independently of each other. A lack of independence is usually apparent in an examination of the residuals because the pattern does not appear to be random.

This is important because dependence of the errors can have drastic effects on the F-test. Suppose two methods of teaching calculus are being evaluated, and a two-group randomized ANCOVA design with 20 subjects in each group is employed using a pretest as the covariate and a final exam as the dependent variable. If 12 of the 20 students in the first group are copying all their answers from one of the remaining eight students, several problems arise. First, it should be clear that there are not 40 subjects responding independently of each other. One student is completely controlling the responses of 12 other students. The degrees of freedom for error in this design would normally be $40 - 2 - 1$ (i.e., $N - J - 1$), but the formula for computing degrees of freedom is based on an N of 40. Actually, the effective N is only $40 - 12 = 28$. The first group contributes only eight independent scores. Hence, the errors are not independent. This lack of independence would be quite apparent in a residual plot because 13 scores would stand out as a subgroup.

If subjects are randomly assigned to treatments, it is more likely that the errors will be independent than if subjects are assigned according to a nonrandom procedure. The random assignment to treatments does not, however, guarantee independence. For example, if all subjects within a given randomly formed treatment group are exposed to the treatment at the same time rather than independently, the errors may be correlated. This may occur because certain extraneous variables may affect many or all subjects within one or more groups, or within-group social interaction during the experiment may introduce dependency. Because it is often difficult to know a priori if subgroups have generated dependency, measures have been developed to measure it; alternative methods of analysis are available to accommodate dependency when it is present. Suppose a two-group experiment is designed and that each group contains 60 subjects and that the residuals of fitting the model reveals fairly distinct clusters of six subgroups of 10 within each treatment condition.

If the reason for the clustering can be identified and measured, it should be added to the model in the form of dummy variables (if categorical) or an additional covariate (if continuous). The model should be re-estimated and the residuals from this model should not appear to have clustered values. But if there is no explanation for the clustered subgroups of residuals, it may be worthwhile to compute a measure of dependency to confirm the visual identification of dependency.

A useful measure of error dependency in a one-factor study is a statistic called the *intraclass correlation coefficient*. (Actually there are many different intraclass coefficients, only the one most relevant for the purpose of measuring dependency among errors is described here.) It is a measure of the extent to which the errors within the subgroups are more similar than are the errors between subgroups. The range of possible values for this coefficient is $\frac{-1}{n_{\text{sg}}-1}$ through 1, where n_{sg} is the sample size for the subgroups within the treatment groups. If the value approaches positive 1, it is concluded that there is high dependency within subgroups; negative values imply higher similarity between subgroups than within subgroups. When the errors are independent, the intraclass correlation will be near zero. The coefficient can be computed as follows:

1. Carry out ANOVA on the original Y_s but use the subgroups (suspected as causing dependency) as the levels of the factor. Label the MS error from this analysis as MS_a.
2. Carry out ANOVA where the mean for each subgroup (not the original Y) is entered as the response for each subject; use the original treatment conditions as the levels of the independent variable. Label the sum of squares error from this analysis as SS_b. Let $S =$ number of subgroups and $J =$ number treatment groups. Label the ratio $\text{SS}_b/(S - J)$ as MS_b.
3. The intraclass correlation is computed using $\frac{\text{MS}_b - \text{MS}_a}{\text{MS}_b + \text{MS}_a(n_{\text{wsg}}-1)} = r_{\text{Intraclass}}$, where n_{wsg} is the number of subjects within each subgroup.

Note that both MS_a and MS_b are measures of error variance. One is based on variation of Y within the subgroups and the other is based on variation of subgroup

means within the levels of the treatments. If the value of the coefficient is close to zero, the conventional analysis is satisfactory. A test of significance is sometimes applied to this coefficient by forming the ratio of the larger over the smaller of the error mean squares. Because the focus is usually on positive values of the intraclass coefficient (because this implies that type I error will be excessive), the ratio of most interest is $\frac{MS_b}{MS_a} = F$, where the degrees of freedom are $S - J$ and $N - S$; as usual, N is the total number of subjects in the experiment. When this approach is used, I recommend that α be set at .10 or .20. Obtained values of F exceeding the critical value lead to the rejection of the hypothesis that the errors are independent.

When positive dependency is identified (regardless of the identification method), the conventional ANCOVA p-value is suspect (because it is almost certainly too small) and some alternative should be considered. The alternative analysis should either not be affected by the subgroup dependency or the dependency should be modeled.

The simplest alternative is to use the subgroup X and Y means as the covariate and dependent variable scores (rather than the individual X and Y scores) in a conventional ANCOVA. When a large number of subgroups are available, this approach is usually satisfactory. Although this strategy is easily carried out using conventional software, a more modern solution is to use a multilevel approach that combines aspects of individual- and group-level analyses into a single analysis. The multilevel approach is known under other names (e.g., random regression coefficient model and hierarchical linear model) and has a relatively long history in a few areas such as education. Excellent references on these models are Gelman and Hill (2007), Kreft and de Leeuw (1998), and Raudenbush and Bryk (2002).

In some cases, an experiment is designed from the beginning with an expectation that dependency among individuals will be a problem. So rather than looking for dependency in subgroups after the data are collected, the experiment is designed by randomly assigning clusters (i.e., subgroups) of subjects rather than randomly assigning individual subjects. These designs are gaining popularity in medical research and are called cluster-randomized trials. An excellent description of these designs and associated analyses is available in Hayes and Moulton (2009).

Homogeneity of Within-Group Regression Slopes

It is assumed that the regression slopes associated with the various treatment groups are the same. The reason for this becomes clear when the β_1 coefficient in the ANCOVA model and the sample b_w coefficient in the adjustment formula are examined. Recall that b_w is the pooled within-group estimate of the population parameter β_1. The fact that β_1 has no superscript to identify different treatments indicates that one common slope is assumed to fit the data. It makes sense to pool the data from the different groups to obtain the single estimate (b_w) only if there is a single parameter (i.e., β_1). If different population slopes are associated with different treatments, a common slope parameter does not exist; the pooled estimate (b_w) cannot be an appropriate estimate of the different population slopes in this case.

Testing the Homogeneity of Regression Slopes Assumption
A test of the null hypothesis

$$H_0 : \beta_1^{\text{group 1}} = \beta_1^{\text{group 2}} = \cdots = \beta_1^{\text{group } J}$$

is described in Chapters 6 and 7. If this hypothesis is rejected, the conclusion is that the population slopes associated with the various treatments are not the same. This means that the conventional ANCOVA model does not fit the data; another approach should be considered in this situation. Two alternatives described as picked-points analysis and Johnson–Neyman analysis are presented in Chapter 11. Reasons for considering these approaches are described next.

Consequences of Violating the Homogeneity of Regression Assumption
A treatment-slope interaction is said to exist when the slopes are not parallel. Suppose a two-group experiment employing one covariate yields the slopes shown in Figure 8.1A.

Because the adjusted means are the same for both treatments, the ANCOVA leads to the conclusion that there is no treatment effect. It is true that those individuals at the mean on X do not differ on Y from group to group. But it is true only for those individuals with X scores falling at the grand mean ($\bar{X}..$) on the covariate. Note that the regression lines suggest that Y scores are quite different for individuals in the two

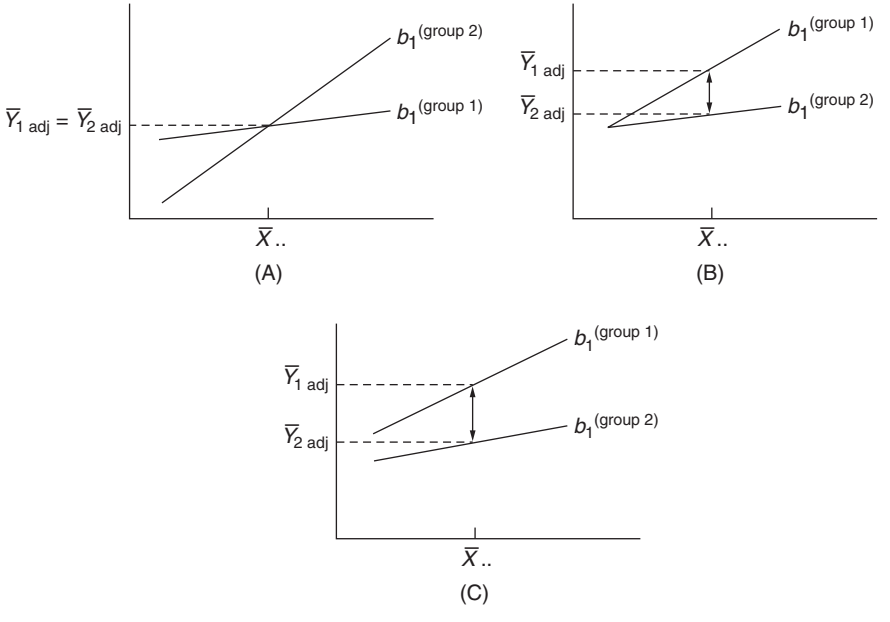


Figure 8.1 Examples of heterogeneous regression slopes.

groups when we look only at those with high or low covariate scores. A complete description of the situation will include the statement that the effect of the treatments depends on the level of X. Subjects with high X scores appear to perform better under treatment 2 than under treatment 1, low X subjects appear to perform better under treatment 1 than under treatment 2, and average X subjects do not appear to differ in performance under the two treatments. Because different interpretations are associated with different levels of X, the use of covariance analysis in the case of heterogeneous within-group slopes should be questioned.

The heterogeneous regression slopes interpretation problem can be seen to be equivalent to the problem of interpreting main effects in a two-factor ANOVA with interaction present. An appropriate procedure for dealing with the two-factor ANOVA is to (a) test for $A \times B$ interaction, and if interaction is present, (b) ignore the main effect tests, and (c) test for simple main effects. Similarly, with ANCOVA, an appropriate procedure is to (1) test for heterogeneity of regression, and if heterogeneity is present, (2) ignore the ANCOVA adjusted treatment effects test, and (3) compute and report the Johnson–Neyman procedure or picked-points analysis (described in Chapter 11).

In addition to the logical problem of comparing adjusted means when slope heterogeneity is present is the statistical problem of the distribution of the F-statistic. The effects of heterogeneous regression slopes on the ANCOVA F-test have been the subject of a number of Monte Carlo investigations (e.g., Hamilton, 1976; Hollingsworth, 1980; Levy, 1980; Sullivan and D'Agostino, Sr., 2002; Wu, 1984). These studies used different criteria for robustness, different degrees of slope heterogeneity, different degrees of unbalance, and different definitions of the ANCOVA model, but we can focus on the case of (a) equal sample size, (b) heterogeneity of slopes of the size likely to be encountered in practice (i.e., differ by less than .4 standardized units), and (c) distributions that are normal, uniform, exponential, chi-square, or double exponential. In these situations, the conventional ANCOVA type I error is reasonably close to α or is conservative. That is, when homogeneity of slopes is not present, the probability of type I error is likely to be lower than the nominal value. The rationale for this finding is essentially as follows.

The error sum of squares in ANCOVA (SS_{Res_w}) is the pooled within-group sum of squared deviations around the pooled within-group slope, b_w. If the sample slopes are not homogeneous, the sum of squares around the pooled slope b_w will be larger than would be obtained if the sum of squares around the individual slopes were computed. Hence, as the degree of heterogeneity of slopes increases, the slope b_w will become an increasingly less adequate fit of the data points. That is, the error sum of squares (i.e., squared deviations around b_w) must increase as the degree of slope heterogeneity increases. It should make sense, then, that heterogeneity of slopes will yield smaller ANCOVA F-values (smaller than with homogeneous slopes) because the error mean square (the denominator of the F-ratio) is an overestimate of the population conditional variance. Small F-ratios are, of course, associated with large probability values.

This general finding of reasonably accurate or conservative p-values in the case of heterogeneous slopes and equal sample size is consistent with the mathematical

conclusions reached by Atiqullah (1964) and more recent simulation results of Sullivan and D'Agostino, Sr. (2002) who investigated the effects of heterogeneity of slopes in the presence of data distorted from normality by floor effects.

One study (Hollingsworth, 1980) contradicts this general finding. Hollingsworth concluded that the probability of type I error increases with slope heterogeneity; this conclusion is based on the assumption that the differences between the covariate means are not zero. In randomized studies, however, the expected mean difference between groups is equal to zero for all variables (i.e., dependent variables and covariates) unless the treatment affects the covariate. Hence, only in nonrandomized observational studies and randomized experiments where the treatment affects the covariate is it likely that slope heterogeneity will increase type I error.

In summary, the three consequences of violating the assumption of homogeneity of regression leads to (1) difficulty in interpreting the meaning of a retained null hypothesis (i.e., uncertainty as to whether overall mean effects are masking treatment differences associated with specific levels of X), (2) biased F-tests where the bias increases with heterogeneity of slopes and discrepancies in sample size, and (3) reasonably accurate or conservative tests when equal sample sizes are used. Because a complete interpretation of results requires information concerning the slopes, it is suggested that the data be plotted and the test for the homogeneity of regression be carried out routinely in any situation where the analysis of covariance is employed. When the slopes are clearly heterogeneous, the investigator will probably be more interested in examining the regression functions for each treatment group separately than in ANCOVA. This problem and appropriate methods of analysis are discussed in Chapter 11.

Linearity of Within-Group Regression Function

It is assumed that the mean of the conditional error distribution within each group at each point on X is equal to zero; this will be true under the ANCOVA model when the within-group relationship between X and Y is linear. If instead the function describing the XY relationship within groups is nonlinear, the errors of the linear model will not be zero at most points on X. Hence, to state that the errors of the conditional Y distributions are zero at all levels of X is tantamount to stating that the within-group regression functions are linear. Of course, it is quite possible to have zero means for error distributions when the relationship is nonlinear, if a nonlinear function is used to describe the data. But the model would no longer be the conventional ANCOVA model as defined earlier.

Testing the Assumption of Linearity

Several procedures are available for testing linearity in conventional regression problems. Because the residuals of the fitted model should always be computed as a first step in the diagnosis of any linear model, it is natural to simply inspect them for departures from linearity. This is usually an adequate evaluation. A more formal approach is to compute a model comparison test where linear and higher order (usually

quadratic) polynomial forms of the ANCOVA model are contrasted. A description of this approach is presented in Chapter 12.

Consequences of Violating the Linearity Assumption

Recall from elementary correlation and regression analysis that if a simple linear model is employed where a higher order polynomial model fits the data, the degree of relationship between X and Y is underestimated. That is, the simple correlation coefficient will be smaller than a correlation measure based on fitting an appropriate function. Correspondingly, in ANCOVA the reduction of the total and within-group sum of squares after adjustment for X will be too small if the linear model is inappropriate. The most obvious consequence of the underadjustment is that the utility of the covariate will be diminished because the adjustment of the means (if any) and the gain in power of ANCOVA over ANOVA depends on the degree of linear relationship between the covariate and the dependent variable. A mathematical treatment of the bias problem resulting from the use of a linear model when a quadratic model is appropriate can be found in Atiqullah (1964). Elashoff (1969) provides a somewhat simplified explanation of Atiqullah's paper.

Fixed Covariate Measured Without Error

The interpretation of any type of statistical analysis is related to the nature of the measurement model underlying the variables employed. For example, the values of the covariate can be conceptualized as being (1) fixed with no measurement error, (2) fixed with measurement error, (3) random with no measurement error, or (4) random with measurement error. If all the population values of X about which inferences are going to be made are included in the sample, the covariate can be conceptualized as a fixed variable. In this case, additional samples drawn from the same population will include the same values of X. If the covariate is an attitude scale and the X-values obtained (in an unreasonably small sample) are 27, 36, 71, and 93, the covariate would be considered to be fixed if (1) the values of X in future samples drawn from the same population include only 27, 36, 71, and 93, and (2) the experimenter is interested in generalizing the results to a population having these four values. This, of course, is not very realistic. If the range of possible scores on the attitude scale is zero through 100, for example, it is very unlikely that future samples will include only the four values obtained in the first sample. It is also unlikely that the experimenter would be interested in limiting the generalization of ANCOVA to a population having covariate values of 27, 36, 71, and 93. Future samples will almost certainly contain different values of the covariate. Because subjects are randomly sampled from the population, it is realistic to view the X-values in a given experiment as a random sample of X-values from a population of values that range (in this example) from zero to 100. One reason, then, why future samples will generally have different values of X is because X is a random variable.

Another issue to be considered in the conceptualization of the covariate is measurement error. Almost all characteristics are imperfectly measured. If additional measures of the covariate (either repeated measures with the same measuring

instrument or alternative measures of the same characteristic) are employed with the same subjects who originally obtained the X scores of 27, 36, 71, and 93, it is unlikely because of errors in measurement that exactly the same values will be obtained once again. Hence, "measurement error" refers to imperfections in the measurement process employed to collect data; it does not refer to sampling error. This problem of measurement error or unreliability in X is one of the central issues in measurement theory. Carroll et al. (2006) should be consulted for a thorough discussion of these topics.

Because the covariate is likely to be a random variable in most studies (virtually all studies in the behavioral and medical sciences) and measurement error is almost always present, it is realistic to view X as a random variable measured with error. Interestingly, the analysis of covariance was derived under the assumption that the X variable is fixed and measured without error.

Consequences of Violating the Assumption of Fixed Covariate Values

The consequences of violating the assumption of fixed error-free X-values depend on the nature of the violation and the type of research design employed. Violations of this assumption are not important with most applications of ANCOVA, but there are some cases in which serious interpretation errors are likely.

It appears that when the covariate is fixed but measured with error (1) there is little discrepancy between the actual and nominal type I error rate, (2) the estimates of the population slope and intercept (or adjusted mean) are not seriously affected, and (3) power is slightly lower. (Keep in mind that this case of X fixed but measured with error is not the typical case encountered in most research applications.) Similarly, when X is random but measured without error, ANCOVA is appropriate (Scheffé, 1959).

Most applications of ANCOVA involve the use of covariates that are considered random and measured with error. In certain situations this case can present problems of interpretation. In general, when a true randomized experiment is involved and subjects are randomly assigned to treatment groups, the reliability of the covariate is not critical; the sensitivity (power) of the analysis is, however, lowered (DeGracie and Fuller, 1972). It is when ANCOVA is used to adjust means in nonrandomized studies that the measurement error problem may be of concern. The following classic example, presented by Lord (1960), suggests the nature of the problem.

Suppose that we have two populations with different means on both X and Y. Also suppose that the reliability of X is perfect and that the common slope obtained by regressing Y on X is 1.0. Now draw a sample from each population and plot the data. The data in Figure 8.2A illustrate what such a plot would look like. Note that the group means on the covariate as well as on the dependent variable are quite different. Keep in mind that the mean difference on the covariate is not a chance difference; these sample data are from populations having different covariate means. This is clearly not a randomized-group experiment. Rather, it is an observational study in which there is interest in employing ANCOVA to reduce bias in the difference between the Y means that can be accounted for by the covariate X. There is nothing unreasonable about attempting to remove bias in Y that can be accounted for by the covariate. Indeed, in Figure 8.2 (Panel A) it can be seen that the adjustment has worked perfectly in the

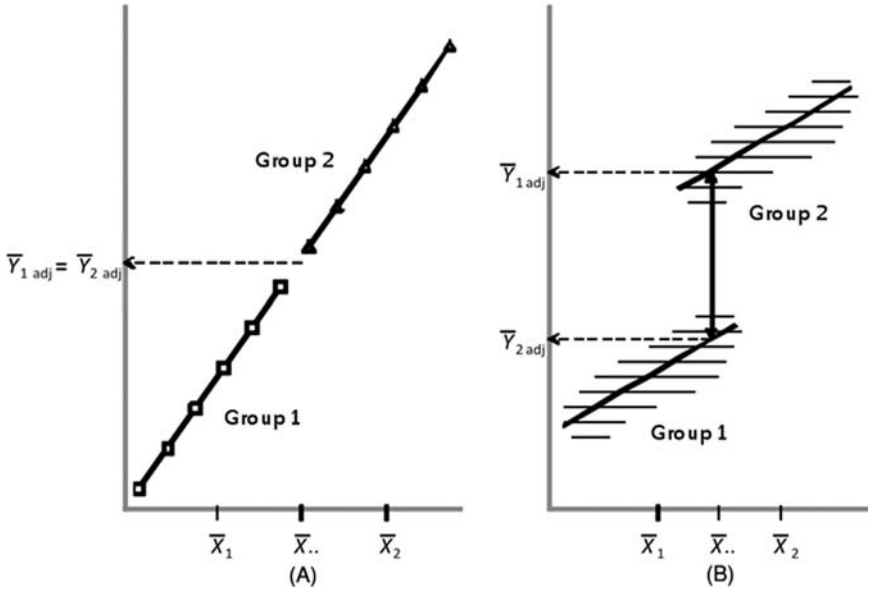


Figure 8.2 Effect of measurement error in X on adjusted means when samples are drawn from populations with different means. Panel A depicts no difference between adjusted means when X and Y are perfectly correlated and X is measured without error. Panel B shows less steep slopes after the introduction of measurement to X; consequently a large difference between adjusted means is estimated.

sense that there are no differences on Y that cannot be predicted from X. This is as it should be if X and Y are both measures of the same characteristic. But keep in mind that the reliability of X is unrealistically set at 1.0 in this example; that is, there is no measurement error.

Let's now take the data of Figure 8.2 (Panel A) and introduce measurement error into the covariate. Because error has been added to the X variable, the scores are spread horizontally around the regression line and the correlation between X and Y can no longer be perfect; thus, the regression slope must change as indicated in Panel B.

The extent to which measurement error affects the slope can be seen in the following equation:

$$\hat{\beta}_{\mathrm{e}} = \beta_{\mathrm{p}} \left(\frac{\sigma^2_{\mathrm{tr}}}{\sigma^2_{\mathrm{obs}}} \right),$$

where

$\hat{\beta}_{\mathrm{e}}$ is the expected population slope when measurement error is present in covariate;
β_{p} is the population slope based on a perfectly reliable covariate;
σ^2_{tr} is the true-score variance on X (i.e., variance of X scores that, hypothetically, have been measured without error); and
σ^2_{obs} is the observed variance of X, i.e., (σ^2_{tr} + variance due to measurement error = σ^2_{obs}).

If there is no measurement error, the true-score variance and the observed score variance are the same, which means that in this case,

$$\hat{\beta}_e = \beta_p(1) = \beta_p.$$

When measurement error is present, the term $(\sigma^2_{tr}/\sigma^2_{obs})$ is less than one. Therefore, the slope associated with scores that contain measurement error is less steep than the slope based on X measured without error (see Carroll et al., 2006; Cochran, 1968, and Fuller and Hidiroglou, 1978, for additional details on the effects of measurement error). Because measurement error is virtually always present, the regression slopes in ANCOVA are virtually always underestimated relative to what they would be with a perfectly reliable covariate. Hence, the major consequence of carrying out ANCOVA in the case of observational studies is that the adjusted mean difference may be misinterpreted.

Note in Figure 8.2, the difference between the slopes under the error and no-error situations and the effect this difference has on the adjusted means. Without error in X the adjusted Y means are identical; with error in X the adjusted means are quite different. (Introducing error into Y does not affect the slope.) It is important to note that outcome discrepancy has occurred even without change in the mean scores on X or Y in the two situations. The point of the example is that *if the covariate means are not equal, the difference between the adjusted means is partly a function of the reliability of the covariate*. That said, it should be pointed out that the difference between adjusted means shown in Panel B in Figure 8.2 has been appropriately adjusted for the *observed* values of X. That is, the computed adjusted difference is indeed conditional on the observed values of X. This is all that is claimed for ANCOVA.

Figure 8.3 extends the situations shown in Figure 8.2 and shows that the difference between adjusted means does not necessarily increase with the introduction of measurement error. Rather, the adjusted difference can be too small *or even in the wrong direction* (relative to what it would be with perfect measurement) when the covariate contains measurement error. Note in Figure 8.3A that the adjusted mean for group 1 exceeds the group 2 adjusted mean when true scores are available as the covariate. In this case of groups from two different populations, neither the X nor the Y means are equal. Group 1 is superior to group 2, but if X is measured with moderate measurement error, as depicted in Figure 8.3B, the adjusted means will be equal even though there actually is a group effect. In Figure 8.3C, the covariate is measured with greater error than in Figure 8.3B and the adjusted mean difference leads to the conclusion that group 2 is superior to group 1.

Several comments are in order before the reader gets the impression that ANCOVA yields incorrect results. First, underadjusted means are not a concern with all research designs. Remember that the design here was one in which samples were drawn from two *different* populations having different covariate means. It was a simple observational study where X was employed to remove bias in comparison of the two groups on Y. The hypothetical situations were designed in such a way that the true (perfectly reliable) X variable completely accounted for bias in Y, but the unreliable X variable did not account for all bias in comparison of the two groups on Y. Actually,

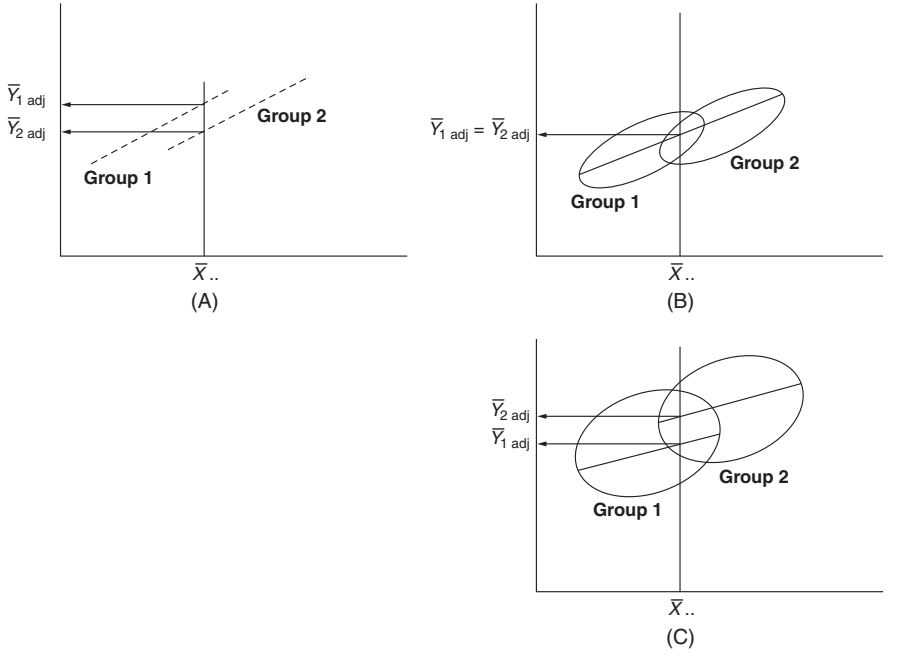


Figure 8.3 Effect of varying degrees of measurement error on adjusted means. (A) Situation with no error in *X* and group 1 with the larger adjusted mean. (B) Situation with moderate amount of error introduced into *X* that results in the conclusion that the adjusted means are equal. (C) Situation with large amount of error introduced into *X* that results in the conclusion that group 2 adjusted mean is larger.

the ANCOVA performed what it was supposed to do in both cases; it can be argued that the problem is not with the analysis but with the interpretation.

Recall that one way of interpreting an adjusted mean difference is to state that it is the difference on *Y* that would be expected if the groups all had the same mean scores on *X*. This statement holds for the case of unreliable as well as reliable *X* scores and for both randomized and observational studies. But the insidious problem here is that when there are *no* treatment effects, the mathematical expectation for the adjusted mean difference is *not* always zero!

When subjects in the various groups are not from the same population, the expected adjusted mean differences are generally *not* zero when there are no treatment effects. Hence, a significant ANCOVA *F* and the associated adjusted means should *not* be interpreted as a reflection of treatment effects when the groups have not been selected from a single population. This interpretation problem is not peculiar to ANCOVA. The problem is a general one; it is inappropriate to damn ANCOVA (as many have done) when the same problem is associated with even the simplest matching designs and analyses—ANCOVA is innocent; measurement error and other reasons for a lack of *XY* correlation and nonequivalent group studies are culpable. It is pointed out in a subsequent chapter that it is sometimes possible to compute a measurement error corrected ANCOVA. The motivation for this is that the researcher is sometimes

interested in what the adjustment would be if there were no measurement errors in the covariate.

It was mentioned earlier that reliability of the covariate is not a major consideration with most designs that are based on subjects selected from a single population. The reason for this is that the effect of measurement error with these designs is primarily one of a reduction in power. Because measurement error attenuates the correlation between X and Y, the error term in ANCOVA using a very unreliable covariate may not be much smaller than with ANOVA. But bias is not introduced to the adjusted means by measurement error with these designs as long as the treatment does not affect the covariate. Because the error mean square is likely to be somewhat smaller with ANCOVA (even when reliability of X is not high) than with ANOVA, some adjustment of the means will occur as long as $b_w \neq 0.0$ and covariate mean differences exist; there is no reason to avoid the use of covariance analysis. This is true regardless of the reliability of the covariate as long as (1) it is known that the subjects have been randomly assigned to treatments from a single population or assigned on the basis of the covariate score and (2) the treatment does not affect the covariate.

Homogeneity of Conditional Variances

There are two cases of homogeneous variance:

Case 1. The variance of the Y scores is constant across treatments for specified X-values, but X is related to the variance of Y. Note in Figure 8.4 that the overall variance for the two distributions is the same but the variance for Y (conditional on individual X-values) increases as X increases.

Case 2. The treatment group conditional variances on Y are the same within each individual group regardless of the level of X, but the overall conditional variances differ between groups. An example of Case 2 violation is illustrated in Figure 8.5.

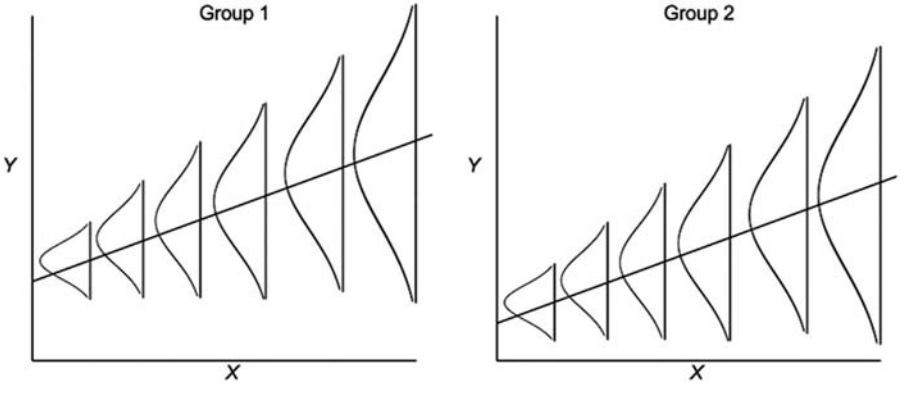


Figure 8.4 Conditional Y distributions with variances increasing with X.

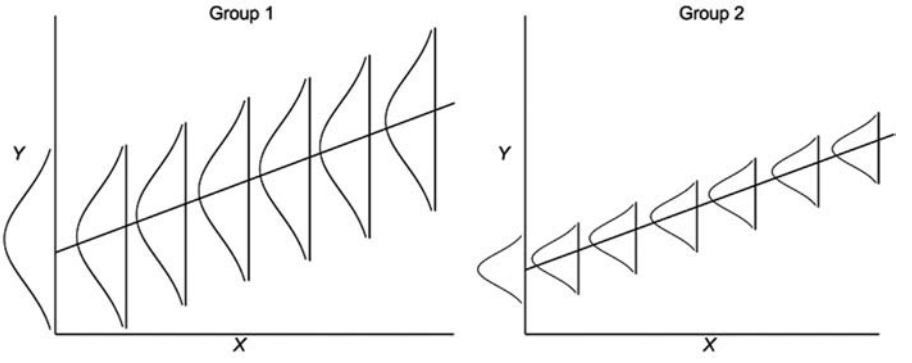


Figure 8.5 Heterogeneous variances across groups.

Tests of Homogeneity of Conditional Variances Assumption

An examination of the plotted residuals of the fitted ANCOVA model is the most direct way of evaluating homogeneity of variance assumption. Conventional tests of homogeneity of variance may be applied to the residuals of the fitted ANCOVA model, but I see little need for formal tests. However, two formal tests provided by *Minitab* are shown below for those who prefer them. Although these tests can be applied to both cases, they are shown below for only the second case. The reason for this is described after the example.

Computational Example 8.1 The data summarized in Table 6.2 are employed in the example of the homogeneity of variance test. The hypothesis tested is

$$H_0 : \sigma^2_{y_1|x} = \sigma^2_{y_2|x} = \cdots = \sigma^2_{y_J|x}.$$

In the case of ANCOVA, the well-known Levene test is simply an ANOVA applied to the absolute values of the residuals of the fitted ANCOVA model. It can be carried out using *Minitab*. A treatment column, TX, is constructed using 1s, 2s and 3s to identify the treatment associated with the subjects. The next column, X, contains the covariate scores, and the third column contains the dependent variable scores. The commands for computing (1) the ANCOVA residuals and (2) the homogeneity of conditional variance tests (for Case 2 homogeneity) are as follows:

```
MTB > ancova Y=TX;
SUBC> covariate X;
SUBC> means TX;
SUBC> Residuals c6.

              ANCOVA: Y versus TX
              Factor  Levels  Values
              TX           3  1, 2, 3
```

```
Analysis of Covariance for Y
Source       DF   Adj SS       MS       F       P
Covariates    1  1855.35  1855.35  28.70  0.000
TX            2   707.99   354.00   5.48  0.010
Error        26  1680.65    64.64
Total        29  3956.00

S = 8.03992   R-Sq = 57.52%   R-Sq(adj) = 52.61%

Covariate    Coef  SE Coef      T      P
X          0.5705    0.106  5.357  0.000

Adjusted Means
TX   N       Y
1   10  28.479
2   10  40.331
3   10  36.190
```

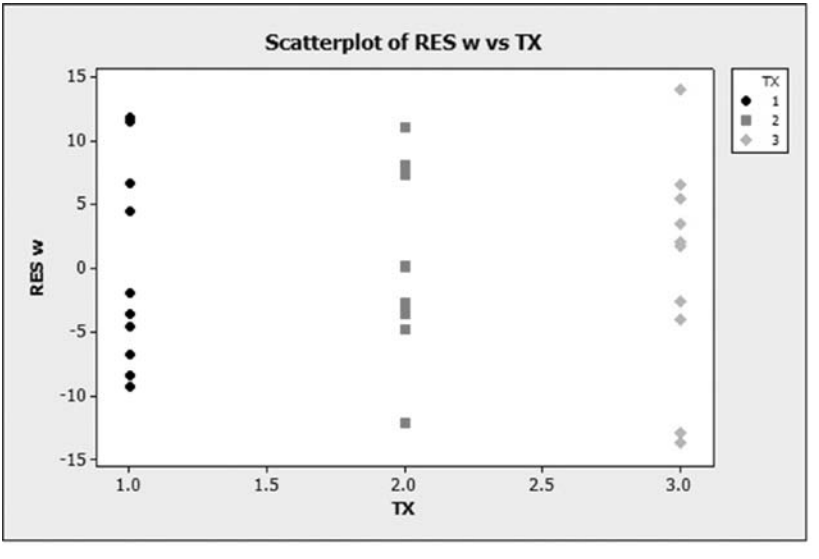
```
MTB > Vartest C6 'TX';
SUBC>   Confidence 95.0;
SUBC>   Title "Example 1 Homogeneity of Variance Test";
SUBC>   Variances 'VARS2'.

Test for Equal Variances: C6 versus TX
95% Bonferroni confidence intervals for
standard deviations

TX   N    Lower    StDev    Upper
 1  10  5.09224  7.99362  17.0111
 2  10  4.46306  7.00595  14.9092
 3  10  5.47102  8.58821  18.2764

Bartlett's Test (Normal Distribution)
Test statistic = 0.36, p-value = 0.836

Levene's Test (Any Continuous Distribution)
Test statistic = 0.15, p-value = 0.858
```

Note that when typing the commands to perform ANCOVA, residuals are requested and assigned to a column; in this example they are assigned to column 6. After the model is fitted the residuals are available and, as can be seen immediately below the adjusted means, the commands for requesting the homogeneity of variance tests are shown. *Minitab* automatically provides results for both Bartlett's and Levene's homogeneity of variance tests. Bartlett's test is often considered to be overly sensitive to nonnormality, so many researchers prefer Levene's test. In this example, the two

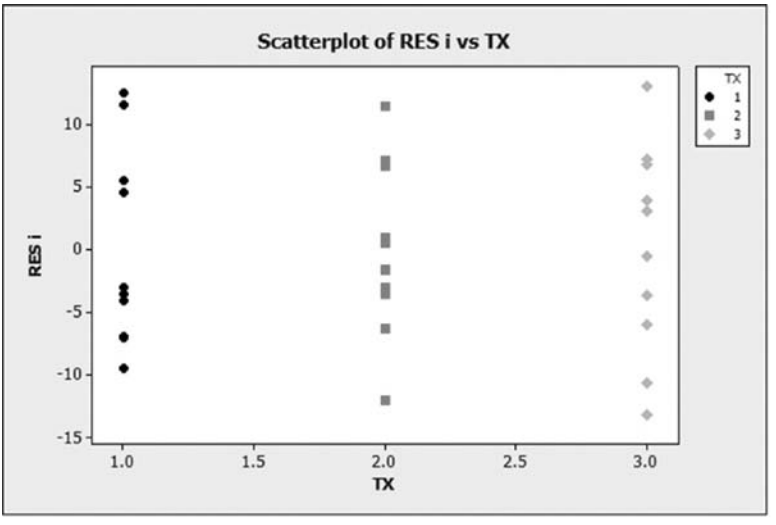
tests yield almost identical p-values (.84 for Bartlett and .86 for Levene). If the obtained p-value from the selected homogeneity of variance test is less than α, the hypothesis of equal conditional variances (i.e., $H_0 : \sigma^2_{y_1|x} = \sigma^2_{y_2|x} = \cdots = \sigma^2_{y_j|x}$) is rejected. Obviously that has not occurred in this example. The evidence supports the ANCOVA model.

This is not surprising because it can be seen below in the plots of the residuals that the variation is similar for the three groups; the standard deviations of the residuals from fitting the ANCOVA model are 7.99, 7.01, and 8.59 for groups 1, 2, and 3, respectively. Two different plots are shown here. The first one illustrates the residuals (Res_w) based on the ANCOVA model; recall that these residuals refer to deviations around the pooled within-group regression slope b_w. The second plot illustrates the residuals (Res_i) associated with the model that fits an individual regression coefficient to each group. The two plots indicate that there is very little difference in the residuals produced by the two fitted equations. This is expected because the homogeneity of regression test indicates that there is little evidence for heterogeneity of slopes; this implies that the pooled slope describes the data for all three groups about as well as individual slopes do, and therefore the two types of residuals must be similar.

The second set of residuals (i.e., Res_i) is needed in evaluating heterogeneous regression analyses (i.e., the Johnson–Neyman technique and picked-points analysis described in Chapter 11) that are recommended as alternatives to ANCOVA. These tests assume homogeneous error variances for the individual treatment population regressions. Hence, Levene's test may also be applied to the residuals of the individually fitted regression lines. These residuals are routinely provided when testing the homogeneity of slope test using the *Minitab* general linear model routine.



Residuals around the pooled within group slope b_w



Residuals around individual within group slopes b_i

Consequences of Violating the Homogeneity of Variance Assumption

The first form of heterogeneous variances (illustrated in Figure 8.4) appears not to be a critical issue regarding the robustness of ANCOVA whether the sample sizes are equal or unequal (Shields, 1978). The second form of heterogeneity (illustrated in Figure 8.5) can be important. Kocher (1974), Shields (1978), and Rheinheimer and Penfield (2001) carried out simulation work that leads to the conclusion that violation of the assumption of homogeneity of error variances is not likely to lead to serious discrepancies between actual and nominal type I error rates unless sample sizes are unequal. However, as in the case of nonnormality, the effects of heterogeneous conditional variances are related to the distribution of the covariate. When the covariate is approximately normally distributed, it is unlikely that the ANCOVA F-test is affected enough by heterogeneous variances to be of practical concern as long as the design is balanced (i.e., has equal sample sizes). The pattern of effects of heterogeneity of variance on the performance of ANCOVA is essentially the same pattern as has been found for ANOVA. When variance sizes and sample sizes differ, the F is conservative if the larger variances are associated with the larger sample sizes and the smaller variances are associated with the smaller sample sizes. When the smaller variances are associated with the larger sample sizes, the bias is liberal (i.e., the true α is greater than the nominal α). So, if the differences among sample sizes are large and the conditional variances are not homogeneous, the ANCOVA F may be biased in either a conservative or a liberal direction. With an unbalanced design and a high degree of nonnormality in the distribution of X, heterogeneity of variance may lead to substantial bias in the distribution of F.

Remedial measures and alternative analyses include (1) Box–Cox transformations (usually log or square root), (2) weighted least-squares regression, (3) a robust

regression approach based on Wilcoxon estimation (see McKean, 2004), (4) a robust method designed to accommodate heterogeneity in linear models proposed by Dixon and McKean (1996), (5) a new method proposed by Abebe et al. (in preparation), and (6) a robust alternative proposed by Wilcox (2005, pp. 531–532). I lean toward weighted least-squares as the preferred method (where the reciprocal of the individual group conditional variances are used as the weights) because transformations often introduce interpretation problems and robust methods are less familiar and require somewhat less accessible software. Weighted least-squares routines are available in all major software packages. I do not recommend popular rank based nonparametric procedures, such as the Conover–Inman (1982) rank transform ANCOVA, because these procedures are not (contrary to myth) robust with respect to heterogeneity of conditional variances in the unequal sample size case (Olejnik and Algina, 1985); like most robust methods they are, however, effective in greatly reducing the influence of outliers.

Fixed Treatment Levels

It is assumed that the treatment levels are fixed by the experimenter. That is, the treatment levels included in the experiment are not selected by randomly sampling the population of possible treatment levels. Rather, the levels selected are the specific levels of interest to the experimenter, and the generalization of the results of the experiment is with respect to these levels. Obviously, no test of this assumption is possible—the experimenter must decide how treatments are selected.

Consequences of Violating the Assumption of Fixed Treatment Levels

If the experimenter randomly selects treatment levels, a random effects ANCOVA model is required; the fixed effects approach is no longer relevant. The required analysis differs considerably from that of the conventional fixed effects model because results are reported in terms of variance components rather than adjusted means. Milliken and Johnson (2002) provide a thorough description of this model for both one and two factor designs. They also describe *SAS* routines for the appropriate analysis.

Normality

The errors of the ANCOVA model are assumed to be normally distributed. The conditional error distribution at each level of X, as shown in Figure 8.6, is assumed normal. Consequently, the marginal distribution on Y (not shown) is also normal.

Testing the Assumption of Normality

As with other assumptions, the most direct diagnostic for normality is to inspect the plotted residuals of the ANCOVA model. A second graphic approach is to inspect the normal probability plot (described in Chapter 4) to determine if the points fall near the straight line that defines the expected proportions under normality. The *Minitab*

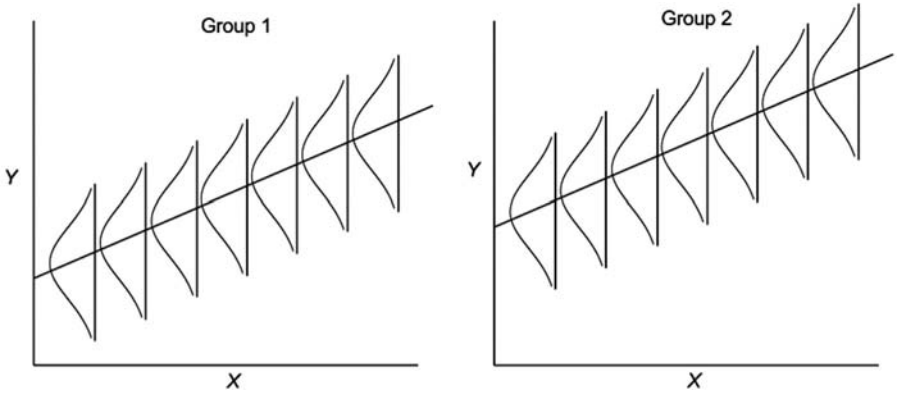


Figure 8.6 Illustration of normality of conditional distributions.

version of this plot includes the results of the Anderson–Darling (*A–D*) normality test. The null hypothesis associated with this test is that the distribution is normal; a small value of *p* is interpreted as a statistically significant departure from normality. An example of this plot and the *A–D* test is shown in Figure 4.5 for a simple regression problem.

Consequences of Violating the Normality Assumption

Analysis of covariance may be more sensitive to departures from normality than is the case with ANOVA. Atiqullah (1964) has shown that the extent to which ANCOVA is biased by nonnormality is largely determined by the distribution of the covariate. Even though normality of the covariate is not an assumption of the ANCOVA model, the bias introduced into the analysis by nonnormality of the dependent variable is greater when *X* is not normally distributed. Moderate departures from normality combined with reasonable sample size (say, 20) appear to have little effect on *F* when equal sample size is involved. This is fortunate because nothing real is precisely normally distributed; after all, normality is a mathematical abstraction. So, the question is whether one has data that are sufficiently close to normality to have trustworthy tests. In the case of large groups with equal sample sizes and homogeneous variances, there is little evidence in the statistical literature of major inferential problems. Large discrepancies in sample size and variance combined with severe kurtosis are almost certainly a problem.

Alternative Methods for Severe Nonnormality

Many procedures have been proposed to handle nonnormality and data that may contain one or more outliers (e.g., Bathke and Brunner, 2003; Conover and Inman, 1982; McKean, 2004). Two of these methods are described in Chapter 14.

8.3 DESIGN AND DATA ISSUES RELATED TO THE INTERPRETATION OF ANCOVA

Random Selection and Random Assignment

The ideal experiment involves both random selection and random assignment. Random selection is involved when subjects are randomly selected from some defined population. Random assignment means that the available subjects (whether randomly selected or not) are randomly and independently assigned to the treatment groups.

Random Selection

If subjects are randomly selected from a defined population, the researcher has a strong foundation for stating that the results generalize to that population. If random selection is not used and the characteristics of the subjects involved in the experiment are not known, there is ambiguity regarding the population to which the results generalize. In the latter case, the slippery default statement is that the results generalize to a population having characteristics like those in the experiment. Of course, this statement is vacuous and it implies that there was little reason to employ inferential methods in the study. Realistically, however, most research is based on selecting subjects using neither random selection nor a complete absence of information regarding subject characteristics. Indeed in most medical investigations, a great deal is known about subject characteristics because extensive selection criteria are used to define the population from which subjects are to be sampled.

Random Assignment

Simple random assignment may be either unrestricted (as is the case when neither the number of subjects available for the experiment nor the number of subjects per group is known before the experiment begins) or based on random allocation where both the total N and the individual sample size n are known before the experiment begins. The former type is frequently used in clinical trials whereas the latter type is typical of small basic experiments.

Random assignment of subjects to treatments is a key component for drawing causal conclusions from an experiment. It yields treatment groups that are probabilistically equivalent on all variables before the experiment begins, and increases the likelihood that the errors of the ANCOVA model will be independent and normally distributed. Although there are conditions under which causal conclusions are justified without random assignment, these conditions are usually unrealistic. For example, under the condition called covariate sufficiency (Stone, 1993) the study includes all covariates that affect the dependent variable. If this is true, random assignment is not required for causal inference. So just how likely is it that one will (1) know which covariates are required, (2) have measurements on them all, (3) know the correct functional form, and (4) know the measurement error variances for all of them? If all of this information were available there would hardly be a reason to run the study. There appears to be a misconception, especially in social science

research, that simply using a "causal" (structural equation) model exempts one from the covariate sufficiency requirement; it does not.

Interestingly, a realistic but rarely used nonrandomized design has been discussed in the behavioral sciences for decades that justifies causal inference when analyzed correctly. Subjects with scores at or below a specific predetermined covariate point are assigned to one treatment and those with scores above the point are assigned to another treatment. Because the key feature of this design is that treatments are assigned *solely* on the basis of the observed covariate scores, I refer to it as an intentionally biased-assignment design. The common label for this design is "regression-discontinuity" but, as often happens in statistics, this confuses the design with the analysis. The study is *designed* by intentionally assigning subjects in a biased manner that guarantees different means on the covariate; the *analysis* involves looking for a discontinuity in the regression of Y on X. The measure of the discontinuity is often estimated using the conventional ANCOVA adjusted mean difference. Aspects of this design and analysis appear in a different guise in Chapters 18–21 (single-case designs) where the covariate is time.

Consequences of Not Randomly Assigning

If we exclude the biased assignment and single-case designs, two problems are likely to occur when random assignment is not employed: (1) the equivalence of the populations from which the samples have been obtained is suspect, and (2) the assumption of independent errors is less likely to be met if the subjects are not independently treated. The first problem is discussed in the remainder of this section; the second problem is discussed in an earlier section of this chapter.

When the subjects are *not* randomly assigned to treatments, the ANCOVA F-*t*est and the adjusted means are likely to be biased as estimates of the causal effect. If the subjects *are* randomly assigned to treatments and are treated as assigned, the ANCOVA F-test and the adjusted means should *not* be biased. The situation in which ANOVA on X is significant even though treatments have been randomly assigned is sometimes described as "unhappy randomization" and can be expected to occur $\alpha(100)\%$ of the time in randomized experiments. That is, because significant results are expected to occur $\alpha(100)\%$ of the time when the null hypothesis is true, it is not unheard of for significant differences on the covariate to occur when random assignment has been employed. Indeed, this is a situation in which ANCOVA clearly yields less ambiguous results concerning treatment effects than does ANOVA.

In the typical "observational study" ANCOVA cannot be counted on to adjust the means in such a way that they can be interpreted in the same way as with a randomized design. Although ANCOVA is often employed in this situation, a covariate should not be expected to eliminate the bias that goes along with comparing means from populations that differ in many ways.

In the case of a clean randomized design, the difference between adjusted means is an unbiased estimate of what the difference between group means would be if each group had a covariate mean equal to the grand covariate mean. ANCOVA works

here because the pooled within-group slope b_w used in adjusting means is a good basis for predicting between-group regression effects. With observational studies (where selection into groups is based on unknown characteristics), the difference between adjusted means is *not* generally an unbiased estimate of what the mean treatment difference would be if each group had a covariate mean equal to the grand covariate mean.

The foundation for the adjustment process is that there is a common covariate mean in the population. Adjustment to a common covariate value (i.e., $\bar{X}..$) makes sense for a randomized study because $E(\bar{X}_1 - \bar{X}_2) = 0$, but the expectation is unknown in the case of observational studies. In the absence of random assignment, the expected covariate mean difference is unlikely to be zero and therefore there is no common mean toward which regression will occur. This is one reason that adjusted treatment effects provided through ANCOVA cannot be counted on to be unbiased in observational studies in the same sense that they can be in randomized designs. But this does not mean that the treatment-group means on the covariate necessarily need to be balanced (i.e., equal) in order for ANCOVA to yield unbiased effect estimates, as was pointed out above in the case of the biased assignment design.

Testing for Covariate "Balance" in Basic Experimental Research and Clinical Trials

Most well-designed and executed basic experimental research involves a random assignment process that is carried out appropriately. Rarely is there question regarding the adequacy of the specific assignment approach and its implementation. In these experiments there is no reason whatsoever to check to see if the means on the covariate are balanced. This point is eloquently made in an entertaining article by Senn (1994) that must be read, if only to be reminded of Dante and to learn how to perform the "Devil's Algorithm" to stealthily bias an experiment without getting caught.

A well-designed and implemented experiment should be analyzed as it is designed. If a covariate has been included in the design, it should be included in the analysis. To include a covariate in the research plan is to imply that there is interest in estimating the treatment effect conditional on the covariate. Obviously, such an estimate will not be provided unless the covariate is indeed included in the estimation procedure. When the research question is conditional, the outcome of tests on either covariate balance or the strength of the covariate is irrelevant.

The frequently stated need for balance on X is apparently based on the belief that this property is required for unbiased estimation. This is false, and has pointed out as such in the literature for decades, perhaps most persistently in recent years by Senn (1994, 1995, 2005); still the misunderstanding persists. I will attempt to illustrate that the ANCOVA solution is no more controversial than nonorthogonal analysis of variance, but the argument is essentially the same as has been used by Senn. Readers

familiar with nonorthogonal ANOVA should have no problem understanding why balance on X is not a prerequisite for unbiased treatment effect estimation.

Suppose two researchers have set up a one-factor design where the two levels of the independent variable are two treatments for pain and the dependent variable is a 35-point pain scale. A sample of 24 subjects is randomly selected from a pain clinic population; after random allocation to the two conditions each group consists of 12 subjects. Data regarding personal characteristics, including scores on a crude optimism scale, are collected before treatments are administered. The scale is of interest because one of the researchers is convinced that there is a strong relationship between optimism and perceived pain. For this reason the research plan specifies that the statistical analysis will be a one-factor ANCOVA using the optimism scale as the covariate. The first researcher predicts that the first treatment will be less effective than the second and therefore the average pain will be lower for the second group. The second researcher favors the first treatment and predicts a lower mean score for this group.

The experiment is carried out and a one-factor analysis of variance is carried out by one of the two investigators. The following outcome (Y) data are obtained:

Treatment 1	Treatment 2
2	11
26	31
24	15
20	13
22	35
26	13
24	12
23	12
4	14
23	14
20	13
22	13
$\bar{Y}_1 = 19.67$	$\bar{Y}_2 = 16.33$

The statistical consultant initially provides results from a one-factor ANOVA. The ANOVA $F = 1.05$ ($p = .32$), which leads to the conclusion that the difference between the two treatment means (19.67 and 16.33) is not statistically significant, but the direction of the difference is consistent with the prediction of the first researcher. Then the one-factor ANCOVA is computed on the same outcome data. The covariate scores based on the crude measure of optimism are shown below; a score of 1 is reported for subjects who tend to be optimistic and a score of -1 is reported if they tend to be pessimistic.

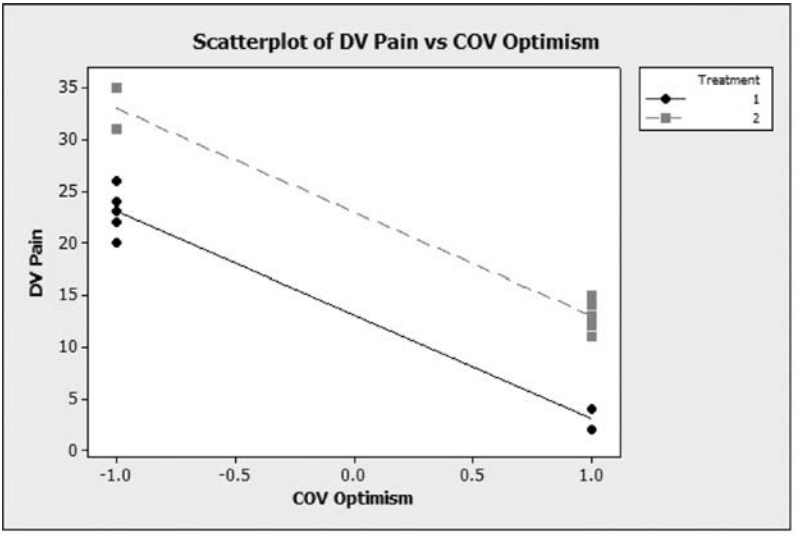
Treatment 1	Treatment 2
1	1
−1	−1
−1	1
−1	1
−1	−1
−1	1
−1	1
−1	1
1	1
−1	1
−1	1
−1	1
$\bar{X}_1 = -.67$	$\bar{X}_2 = .67$

Note that the treatment groups are "unbalanced" on the covariate. Because the groups were randomized we do not expect to find unbalance of this size by chance very often; but it occurred here so we will call it an instance of unhappy randomization. So now the question is, "Do we care that the groups are unbalanced on the covariate?" Of course we do. If the analysis of the outcome is no more sophisticated than the one-factor ANOVA (or the equivalent *t*-test) shown above, there is every reason to question the adequacy of the effect estimate. The claim that covariate balance is not required for unbiased effect estimation does not apply to analyses of this type unless one refers to "bias in expectation." Across an infinite number of replications of the randomized experiment, the conventional estimator (i.e., the sample mean difference on Y) is *not* biased in expectation because $E(\bar{Y}_1 - \bar{Y}_2) = (\mu_1 - \mu_2)$. But in the individual study we care a lot about imbalance; fortunately we can find an effect estimator that is much better than $\bar{Y}_1 - \bar{Y}_2$.

It makes sense in the current example to ask, "What would be the difference between treatment means if the groups were balanced with respect to optimism?" One way to answer this is the analysis of covariance. If we apply ANCOVA to the data shown above we find:

```
Analysis of Covariance for DV Pain

Source       DF   Adj SS       MS        F       P
Covariates    1  1333.33  1333.33   451.61   0.000
Treatment     1   333.33   333.33   112.90   0.000
Error        21    62.00     2.95
Total        23  1462.00

Adjusted Means
Treatment    N  DV Pain
1           12   13.000
2           12   23.000
```

Note that the adjusted means ($\bar{Y}_{1\,\text{adj}} = 13$ and $\bar{Y}_{2\,\text{adj}} = 23$) are very different than the unadjusted means (19.67 and 16.33) in both direction and magnitude. The initial conclusion based on ANOVA is reversed after the ANCOVA adjustment. It is now concluded that the first treatment is much more effective than the second because the adjusted pain score is much lower for the first treatment. Further, the p-value for the treatment effect is very small ($p < .001$) and, as can be seen in the plot of these data, the effect is consistent regardless of the covariate score. Conditioning on the covariate makes a profound difference on the interpretation of this experiment. This is an example of Simpson's Paradox.



A third way of analyzing the data is to reconceptualize the covariate as a factor in a two-factor analysis of variance. If we do so, it is obvious from inspection of the tabulated data (shown below) that this is a nonorthogonal design because the cell frequencies are different. The appropriate way to analyze designs of this type is to estimate the effects of each factor conditional on the other factor(s) in the design. In the case of a two-factor design having the same number of subjects in each cell, the effect estimates for factor A, factor B, and the interaction are independent of each other. We want to maintain this same desirable property (i.e., independence of effect estimates) even though the cell sample sizes are not the same; this is accomplished by conditioning each effect estimate on the others in the design.

	B_1	B_2	
A_1	2, 4	11, 15, 13, 13, 12 12, 14, 14, 13, 13	$\bar{Y}_{A_1} = 11.33$
A_2	26, 24, 20, 22, 26 24, 23, 23, 20, 22	31, 35	$\bar{Y}_{A_2} = 24.67$
	$\bar{Y}_{B_1} = 19.67$	$\bar{Y}_{B_2} = 16.33$	

The results of this analysis (using the *Minitab* general linear model routine with type III sum of squares) is revealing.

```
MTB > GLM 'Y' = a b a*b;
SUBC>   Brief 2;
SUBC>   Means a b a*b.

Analysis of Variance for Y, using Adjusted SS for Tests

Source  DF   Seq SS   Adj SS   Adj MS       F      P
A        1  1066.67  1333.33  1333.33  430.11  0.000
B        1   333.33   333.33   333.33  107.53  0.000
A*B      1     0.00     0.00     0.00    0.00  1.000
Error   20    62.00    62.00     3.10
Total   23  1462.00

Least Squares Means for Y

        Mean  SE Mean
1      8.000   0.6819
2     28.000   0.6819

B
1     13.000   0.6819
2     23.000   0.6819

A*B
1  1   3.000   1.2450
1  2  13.000   0.5568
2  1  23.000   0.5568
2  2  33.000   1.2450
```

Because factor A represents the covariate and factor B represents the two treatment levels, we are interested in the effects of factor B. Note that two means associated with this factor are 13 and 23. These are identical to the adjusted means provided by ANCOVA. The sum of squares for treatments and error are also the same, but there is a slight difference in error degrees of freedom (20 for GLM and 21 for ANCOVA). This difference in degrees of freedom is the reason that the two F-values are not identical. But if the GLM routine is repeated using only factor A (which corresponds to the covariate) and factor B (treatments) in the model (omitting the interaction because there are no interaction effects), the two F-values are identical (i.e., 112.90), as can be seen by comparing the ANCOVA results with those shown below for the reduced model (no interaction) two-factor nonorthogonal ANOVA.

```
MTB > GLM 'Y' = a b;
SUBC>   Brief 2;
SUBC>   Means a b.
```

```
General Linear Model: Y versus A, B
Analysis of Variance for Y, using Adjusted SS for Tests

Source  DF   Seq SS   Adj SS   Adj MS       F      P
A        1  1066.67  1333.33  1333.33  451.61  0.000
B        1   333.33   333.33   333.33  112.90  0.000
Error   21    62.00    62.00     2.95
Total   23  1462.00
```

The ANCOVA and this GLM solution are identical even though the labeling is different. In both cases, the treatment effect estimate and the sum of squares for treatment and error are conditional on the covariate. This is what we want because these analyses are not confounded by the imbalance in the covariate that causes the inadequacy of the ANOVA and the associated unadjusted means. The bottom line is that imbalanced covariate means *are* a problem and that the solution is ANCOVA in an appropriately randomized experiment.

When Covariate Balance Can Be a Problem in Randomized Designs

The argument that ANCOVA yields unbiased effect estimates even when covariates are not balanced, rests on the assumption that appropriate randomization has occurred. This is not always the case.

Bungled Randomization

Flaws in the randomization process are not likely to occur in tightly monitored basic experimental designs of moderate size. But clinical trials are another story. A large (11,000 patients) hypertension clinical trial (known as the CAPP Study) alluded to by Senn (2005) is a prime example of things going wrong. Two covariates were systolic and diastolic blood pressure; the *t*-values comparing the pretreatment means on these covariates were 5.8 and 8.9, respectively! Apparently a satisfactory explanation for the problem has never been provided.

Intentional Card Stacking (Fraud)

Sometimes clinical trials are appropriately randomized during design, but the process is intentionally sabotaged (Berger and Weinstein, 2004; Dyer, 2003; Farthing, 2004; Greenhouse, 2003; Schultz, 1995).

The problem of selection bias (not necessarily intentional) has drawn the attention of statisticians who have developed tests to detect selection bias (Berger, 2005a; Berger and Exner, 1999). These tests are more sophisticated than simple comparisons of covariate means. The definitive work in this area is a groundbreaking book by Berger (2005b) that covers a treasure trove of examples of selection bias, explanations of how subversion occurs, the impact it has, how to prevent it, how to detect it, and how to attempt to adjust for it. The last topic on this list is the most controversial among statisticians. Senn's (2005) take is that no testing and correction procedure is protection against ruthless criminals and that the only way to be convincing in

a controversial research area is to arrange to have skeptics supervise the trial. This actually occurred in an evaluation of homeopathy. The world's most famous magician and skeptic, "The Amazing Randi," (who belongs on the list of heroes for the rational) supervized both randomization and blinding (Maddox et al., 1988).

Selecting Covariates: NonClinical Experiments Versus Clinical Trials

The typical experiment has one or two covariates that are selected before analyzing the data; there is no particular problem with covariate selection in this case. Some researchers, upon learning that covariate selection is a controversial topic, are surprised. What they do not understand is that a huge collection of covariates is available in the case of clinical trials. Indeed, the number of covariates measured in the typical clinical trial is on the order of 100 (Beach and Meier, 1989).

This becomes a problem for two reasons. First, the number of available and measured covariates often exceeds the number of cases studied. This means there are more covariates than degrees of freedom; conventional estimation is impossible in this case. Second, the specific set of covariates chosen for the analysis affects the treatment effect estimate. This means that post hoc selection of covariates allows the researcher to pick and choose various combinations of covariates until the most pleasing result is obtained. Computer routines can be set up to effectively facilitate this process.

Types of Covariate Screening

There are essentially two types of covariate screening.

Covariate Balance

The first version involves testing differences between covariate means before deciding to include the covariate in the analysis. The belief appears to be that a covariate is worth including only if the covariate means are significantly different. Otherwise the covariate is not included in the model. This strategy has been attacked using arguments of two types: (1) philosophical/logical and (2) empirical (using both simulation data and published clinical data).

The philosophical/logical argument is that testing for balance makes no sense. Senn (1994), in one incisive and cleaver paragraph, cuts to what he describes as

> "The two incontrovertible facts about a randomized trial:
>
> **1** Over all randomizations the groups are balanced;
> **2** For a particular randomization they are unbalanced."

He goes on to argue that no "significant imbalance" can cause the first fact to be untrue and no lack of significant balance can cause the second fact to be false. Thus, it seems that the only reason to test the difference in covariate means is to evaluate the randomization process itself. A significant result implies either incompetence in the process or deception in the allocation.

The second argument is that two major inferential problems are associated with screening for covariate balance and these have been demonstrated repeatedly. The problems are:

(a) ANCOVA tests for treatment effects that are based on adjusting only for covariates identified by pretesting for covariate imbalance yield *p*-values that are too large, making ANCOVA conservative (Bancroft, 1964; Beach and Meier, 1989; Begg, 1990; Forsythe, 1977; Permutt, 1990; Schluchter and Forsythe, 1985; Senn, 1994).

(b) Prognostic covariates (usually the most useful covariates) are excluded from the analysis when they appear to be balanced.

Covariate-Dependent Variable Relationship

The second type of covariate screening focuses on the relationship between the covariate(s) and the dependent variable. The test of significance of the relationship between the covariate and the outcome is the screening hurtle for inclusion. In the case of many covariates, a conventional variable selection procedure such as stepwise regression is sometimes used. Simulation evaluations of this practice indicate the ANCOVA error variance is underestimated. Attempts to fix this problem include bootstrap error variance estimation and adjustment approaches, but remediation of estimation after collecting data seems to miss the point that there would be no problem to be remediated if screening were avoided in the first place.

In fairness to researchers who routinely screen on the basis of the *p*-value on the covariate, it is quite understandable given current popular software and associated documentation for ANCOVA. ANCOVA software routines present output in a format that inappropriately encourages an emphasis on the covariate-outcome relationship. It is not surprising that there is an overemphasis on this test; it appears in the first line of the typical ANCOVA summary table. The simpler ANCOVA summary format recommended in Chapter 6 avoids a display of this frequently misleading aspect of the analysis. (Some worry that the sums of squares in the simpler table do not up to the total sum of squares. This is as it should be; all sources of variation appearing in the simpler table are adjusted for the covariate. It is similar for nonorthogonal ANOVA.)

Covariate Not Affected by the Treatment

There is no statistical assumption that the covariate is not affected by the treatment. It is, however, a condition that certainly should exist for the most unambiguous interpretation of ANCOVA applied to an experimental study. In a randomized two-group experiment, for example, the adjusted mean difference can be interpreted as an unbiased estimate of what the mean difference on Y would be if both groups had exactly the same covariate mean.

The mean difference on X will generally be small in this case (because random assignment has been employed) as long as the treatment has not affected the covariate. Because the mean difference on X is due to sampling fluctuation, the interpretation of the adjusted mean difference as the difference that would be expected if both groups

had the same covariate mean presents no logical problem. That is, the adjusted difference can be interpreted as the mean difference that would be expected if a matching design had been employed and subjects selected to have X scores equal to $\bar{X}..$ had been randomly assigned to treatments.

When the treatment affects the covariate the results are interpreted in essentially the same way as when the treatment is independent of the covariate, but two problems may arise. First, treatments may produce covariate means that, when averaged, yield a grand covariate mean that has no counterpart in reality. There may be no subjects in existence who do (or even could have) covariate scores equal to the grand covariate mean.

Suppose an investigation is being carried out to evaluate the effects of two different on-the-job methods of assembling a recently designed complex electronic component. Employees are randomly assigned to two groups, and one of the two experimental assembly methods is assigned to each group; the dependent variable is the number of hours of training required to produce a component with no defects. The assembly tasks are performed in various areas of the plant; each area is subjected to different levels of uncontrolled noise from various sources that are unrelated to the assembly task. It is believed that the performance of the employees on the assembly task is affected by noise.

Because the noise level is not the same at the various work sites associated with the different employees, it could affect the mean difference between the two assembly methods. In an attempt to remove the possible contamination effect of noise from the experimental comparison, noise levels associated with each employee's work site are measured and used as the covariate in ANCOVA. If the treatments (methods of assembly) do not affect the noise level, ANCOVA can be expected to provide an unbiased estimate of what the treatment effect would be if all employees were subjected to the same ($\bar{X}..$) noise level.

If, however, the characteristics of the methods of assembly of the components constitute the cause of the noise, the treatment causes variation on the covariate. In this case it would be questionable to employ noise as the covariate. Figure 8.7 illustrates the difference in the size of the adjusted treatment sum of squares when the treatments do not affect the covariate (left-hand panel) and when the treatments affect the covariate (right-hand panel). Note on the right that adjusting the treatment sum of squares on Y due to the covariate (noise) removes part of the treatment effect.

In this case, the hypothesis of equal adjusted population means has little meaning. This is because one of the outcomes of the treatments is the noise. It does not make sense to adjust an outcome for another outcome in most cases. If it is impossible to have two treatments without different noise levels, the ANCOVA is comparing two treatments that do not exist in reality. Suppose the noise levels for the two assembly methods are as follows:

Assembly Method	Mean (db)	Range (db)
1	20	10
2	90	10

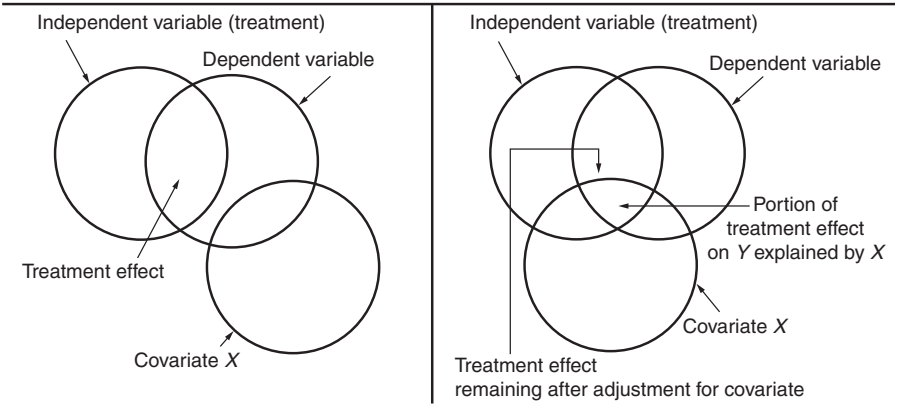


Figure 8.7 Treatments uncorrelated with the covariate (left panel) and treatments correlated with the covariate (right panel).

The grand mean on *X* is 55, but neither method is associated with a noise level near 55. It is meaningless to interpret the outcome of ANCOVA; it provides an answer to a question that has no relevance. No one cares about what the estimated difference in mean performance would be for subjects who are exposed to a noise level of 55 db. It is not useful to try to estimate this difference because, first, neither method 1 nor method 2 is associated with a 55-db noise level. The research question (viz., "What is the difference in performance under the two methods?") is not answered because one aspect of each treatment is the noise level. Second, even if someone thought it would be useful to know the difference given a noise level of 55 db, there are no observations in the experiment anywhere near 55 db and hence no data to provide an answer. The ANCOVA could be computed, but it is not clear why an investigator would want to estimate and interpret adjusted means that refer to levels of *X* that are associated with treatments and subjects that are nonexistent and could not exist in the future.

The preceding discussion should not be interpreted to mean that there are no cases where ANCOVA is appropriate when the treatment affects the covariate. Indeed, one of the more useful applications of ANCOVA is with studies that employ each of several response measures (in several analyses) as the covariate to clarify how the treatment affects behavior. For example, in studying the effects of two types of exercise on resting heart rate and frequency of psychosomatic complaints, it might be of interest to use resting heart rate as the covariate and frequency of psychosomatic complaints as the dependent variable. If the difference between treatment groups on frequency of psychosomatic complaints is largely eliminated when resting heart rate is employed as the covariate, it would be reasonable to *speculate* that the treatment effect on the dependent variable is mediated by the covariate. It could not be *concluded* that *X* causes *Y* to vary because there are many other variables that could cause variability on *Y*. That is, the treatment may have affected a third variable that causes both complaints and heart rate to vary.

At this point in the discussion of treatments affecting the covariate, the reader may wonder why ANCOVA was dismissed as nonmeaningful in the earlier example (where noise was the covariate) but useful in the second case (where heart rate was the covariate). This is because the covariate (noise level) was an integral part of the treatment in the first case, but this was not true in the second case. It is reasonable to employ resting heart rate as a covariate because the treatments do not limit the possible resting heart rate values, whereas the treatments did directly control noise level.

Suppose the mean resting heart rates for the exercise groups 1 and 2 were 75 and 85, respectively. It is not difficult to argue that an experiment could be carried out by randomly assigning subjects (each having a resting heart rate of 80) to two different exercise programs. Recall that it did not make sense to consider assigning subjects associated with 55-db noise levels to treatments in the other experiment.

Testing the Assumption of Independence of Treatment and Covariate

It is possible to test the assumption of the independence of the treatment and the covariate by performing an ANOVA on the covariate. In a randomized experiment in which the covariate has been measured before the treatments are administered, there is no reason to report the result of such a test since the treatment cannot have an effect in this case. If the covariate is measured during or after the treatment, the ANOVA on the covariate may be relevant because it provides information on the effect of the treatments.

For most nonrandomized studies, the ANOVA on the covariate should be carried out regardless of whether treatments have been applied. If treatments have not been applied but the ANOVA F is significant, this is a warning that the ANCOVA F and adjusted means are almost certain to be biased as reflections of the treatment effects. If the ANOVA on the covariate is not significant, the ANCOVA F and adjusted means are still subject to problems of bias, but the degree of bias, although unknown, may not be large.

There is one experimental design, the biased assignment design (described earlier) that is not a randomized design, but it is not subject to the problems of bias in the ANCOVA F and adjusted means. If a nonrandomized design other than the biased assignment design is employed and the covariate is measured after treatments are administered, the ANOVA on the covariate will be difficult to interpret because treatment effects and pretreatment differences among the populations will be confounded.

Summary of Consequences of Treatments Affecting the Covariate

If the treatment affects the covariate in a randomized-group experiment, the hypothesis tested is not, in general, the same with ANCOVA as with ANOVA. That is, the adjusted and unadjusted population means are not generally the same when the population covariate means are different. The covariate may either remove or introduce differences on Y that may be misinterpreted as reflections of treatment effects. Also, the adjusted mean difference may be impossible to interpret meaningfully because such a difference may be the predicted mean difference for nonexistent treatment

conditions and subjects. An alternative analysis should be considered whenever the treatment affects the covariate.

Alternative Analysis When the Treatment Affects the Covariate

The almost universal recommendation for analyzing data where the treatment has affected the covariate is to simply ignore the covariate information and perform ANOVA. I do not recommend following this common practice for reasons described in Chapter 13, where a preferable approach is presented.

8.4 SUMMARY

If the experimenter has appropriately assigned treatments to subjects and has treated subjects within each group independently, the most important assumption should be satisfied. The homogeneity of regression assumption is easily tested by using the *F*-test designed for this purpose. If the covariate is affected by the treatment and an unbiased treatment effect is the major interest, an alternative to conventional ANCOVA described in Chapter 13 should be considered. ANCOVA may be useful when there is interest in describing how the mean effect estimate changes when the covariate is added to the model. It is possible that the reason for change is that the mechanism through which the treatment works is mediated by the covariate; this interpretation is speculative because there are many other plausible explanations for the role of the covariate.

The reliability of the covariate is critical in nonrandomized observational studies, but not in most biased assignment and randomized-group experiments. The reason for this is that unreliability in nonequivalent group studies can lead to statements concerning adjusted means that are very misleading. With most randomized designs, unreliability simply results in ANCOVA *F*-tests with somewhat lower power than corresponding tests with highly reliable covariates. Measurement error corrected ANCOVA is recommended as an alternative to conventional ANCOVA as a method of correcting for unreliability in some nonequivalent designs.

The effect of moderate nonlinearity is a slight conservative bias in the ANCOVA *F*-test. When nonlinearity is suspected, an alternative method such as nonlinear ANCOVA should be considered. The effects of not meeting the assumptions of normality and homogeneity of variance of the conditional *Y* scores are similar to the effects of violations of corresponding assumptions with ANOVA. It is known, however, that there is an additional consideration with ANCOVA; the effect of nonnormality and heterogeneity of variances on bias in the ANCOVA *F*- test is related to the distribution of the covariate. It appears unlikely that departure from these assumptions result in serious bias of the *F*-test when balanced designs are employed. Robust linear model ANCOVA is recommended when either or both of these assumptions are obviously violated in randomized experiments.

CHAPTER 9

Multiple Comparison Tests and Confidence Intervals

9.1 INTRODUCTION

The issues of error rate and methods of specifying contrasts associated with multiple comparison tests and simultaneous confidence intervals are discussed in Chapter 3 for one-factor ANOVA. The same issues are relevant in the case of ANCOVA. Analogs to the multiple comparison methods discussed in Chapter 3 are presented in this chapter along with two additional methods.

Recall that when three or more groups are employed, the level of α associated with the F-test does not refer to the probability that a specific comparison will be incorrectly declared significant. Rather, it refers to the probability that one or more of all possible comparisons will be incorrectly declared significant. Because the collection of all possible comparisons defines the experiment, this probability of type I error is known as the experimentwise or familywise error rate. Multiple comparison tests are designed to maintain the familywise error rate at the nominal value specified for α.

9.2 OVERVIEW OF FOUR MULTIPLE COMPARISON PROCEDURES

Four multiple comparison procedures are presented in this chapter. They include the following methods: (1) Fisher–Hayter (F–H), (2) Tukey–Kramer (T–K), (3) Bonferroni, and (4) Scheffé.

The choice among these procedures should be based on the type of comparisons of greatest interest to the investigator and whether simultaneous confidence intervals are desired. General recommendations are as follows:

1. Use the Fisher–Hayter procedure when the comparisons of interest are limited to hypothesis tests on any or all pairwise comparisons and simultaneous confidence intervals are not desired.

The Analysis of Covariance and Alternatives: Statistical Methods for Experiments, Quasi-Experiments, and Single-Case Studies, Second Edition. Bradley E. Huitema.

2. Use the Tukey–Kramer when most or all pairwise comparisons are of interest and simultaneous confidence intervals are desired.
3. Use the Bonferroni procedure for hypothesis tests and simultaneous confidence intervals on a small number of planned pairwise and/or complex (*ij*) comparisons.
4. Use the Scheffé procedure when a large number of planned or unplanned complex comparisons of any type is of interest.

Note that recommendations 2 through 4 contain the ambiguous words "most," "small," and "large". This is unavoidable because the best choice (in terms of power or width of simultaneous confidence intervals) among the last three procedures is a function of several factors, including the number of groups, the number of comparisons, and the degrees of freedom. Fortunately, it is quite acceptable to compute simultaneous confidence intervals by using more than one procedure and to then select the procedure that yields the shortest intervals. This approach presumes that the decision concerning which contrasts are of interest has been made a priori.

9.3 TESTS ON ALL PAIRWISE COMPARISONS: FISHER–HAYTER

The most frequently encountered multiple comparison situation is the one in which the researcher is interested in all pairwise comparisons. Many multiple comparison tests can be used in this situation but the Fisher–Hayter test generally has higher power than the popular competing procedures.

Two steps precede the computation of any multiple comparison procedure used in ANCOVA: (1) the ANCOVA and (2) the ANOVA on the *covariate*. Quantities from both analyses are employed in the multiple comparison formulas. Results of applying these analyses to the three-group achievement study described in previous chapters (i.e., Table 6.1 in Chapter 6) are tabulated as follows for reference throughout the chapter.

ANCOVA (Data from Table 6.1)

Source	SS	*df*	MS	*F*
Adjusted treatment	707.99	2	354.00	5.48
Residual within	1680.65	26	64.64	
Total residuals	2388.64	28		
Critical value $= F_{(.05,\,2,26)} = 3.37$				

The adjusted means are

$$\text{Group } 1 = 28.48 = \bar{Y}_{1\,\text{adj}};$$
$$\text{Group } 2 = 40.33 = \bar{Y}_{2\,\text{adj}};\ \text{and}$$
$$\text{Group } 3 = 36.19 = \bar{Y}_{3\,\text{adj}}.$$

ANOVA on Covariate (Data from Table 6.1)

Source	SS	*df*	MS	*F*
Between	126.67	2	63.33	.30
Within	5700.00	27	211.11	
Total	5826.67	29		

Critical value $= F_{(.05,\,2,27)} = 3.35$

The covariate means are

$$\text{Group } 1 = 52.00 = \bar{X}_1;$$
$$\text{Group } 2 = 47.00 = \bar{X}_2; \text{ and}$$
$$\text{Group } 3 = 49.00 = \bar{X}_3.$$

Fisher–Hayter Test Procedure

Stage 1. Perform the ANCOVA F-test at level α. If this test is nonsignificant, retain the hypothesis that $\mu_{1\,\text{adj}} = \mu_{2\,\text{adj}} = \cdots = \mu_{J\,\text{adj}}$, conclude that the evidence is insufficient claim any effects whatsoever, and terminate the analysis. If this test is significant, proceed to stage 2.

Stage 2. The Fisher–Hayter test on each pairwise difference between adjusted means is carried out using the same α as was employed during stage 1. The formula for evaluating the difference between two adjusted means, say, i and j, is as follows:

$$\frac{\bar{Y}_{i\,\text{adj}} - \bar{Y}_{j\,\text{adj}}}{\sqrt{\dfrac{\text{MS}_{\text{Res}_w}\left[\left[\dfrac{1}{n_i} + \dfrac{1}{n_j}\right] + \dfrac{\left(\bar{X}_i - \bar{X}_j\right)^2}{\text{SS}_{w_x}}\right]}{2}}} = q,$$

where

MS_{Res_w} is the analysis of covariance mean square error term obtained from the ANCOVA summary table;

n_i, n_j are the sample sizes for the ith and jth groups, respectively;

$\bar{X}_i, \bar{X}_j$ are the covariate means for the ith and jth groups, respectively; and

SS_{w_x} is the sum of squares within groups on X variable; this is obtained from summary table for ANOVA on covariate.

The critical value against which the absolute obtained q values are compared is the Studentized range statistic $q_{J-1,\,N-J-1}$. This value is found in the Studentized range table by finding the column heading that is equal to $J - 1$ and degrees of freedom equal to $N - J - 1$.

Interpretation of Fisher–Hayter Tests

Recall the meaning of the rejected null hypothesis in analysis of covariance problems. If three or more groups are involved, the rejection of the hypothesis $H_0 : \mu_{1\,\text{adj}} = \mu_{2\,\text{adj}} = \cdots = \mu_{J\,\text{adj}}$ leads us to conclude that some linear combination of adjusted means results in a population contrast that is not zero. Each significant Fisher–Hayter test leads to the conclusion that the associated pairwise difference between adjusted population means is not zero. The researcher can conclude that the probability of obtaining one or more significant differences (using *F–H*) in the experiment by chance alone is not greater than α.

Numerical Example of F–H Tests

The analysis of covariance on the example data yields a significant F; it is appropriate to carry out *F–H* tests. The test on the difference between adjusted means 1 and 2 is

$$\frac{28.48 - 40.33}{\sqrt{\dfrac{64.64\left[\left(\frac{1}{10} + \frac{1}{10}\right) + \dfrac{(52-47)^2}{5700}\right]}{2}}} = \frac{-11.85}{2.57} = -4.61 = q_{\text{obt}}.$$

The obtained q is 4.61 (ignoring sign) and the critical value against which this obtained value is compared is $q_{2,\,26} = 2.91$. The difference between the two adjusted means is significant at the .05 level.

The test on the difference between adjusted means 1 and 3 is

$$\frac{28.48 - 36.19}{\sqrt{\dfrac{64.64\left[\left(\frac{1}{10} + \frac{1}{10}\right) + \dfrac{(52-49)^2}{5700}\right]}{2}}} = \frac{-7.71}{2.55} = -3.02 = q_{\text{obt}}.$$

Because $|-3.02|$ is greater than 2.91, the null hypothesis H_0: $\mu_{1\,\text{adj}} = \mu_{3\,\text{adj}}$ is rejected at the .05 level.

The test on the difference between adjusted means 2 and 3 is

$$\frac{40.33 - 36.19}{\sqrt{\dfrac{64.64\left[\left(\frac{1}{10} + \frac{1}{10}\right) + \dfrac{(47-49)^2}{5700}\right]}{2}}} = \frac{4.14}{2.55} = 1.63 = q_{\text{obt}}.$$

The null hypothesis $H_0 : \mu_{2\,\text{adj}} = \mu_{3\,\text{adj}}$ is retained.

Three pairwise comparisons have been computed and two of these were found to be significant using $\alpha = .05$. The probability that one or more of the comparisons has been incorrectly declared significant because of sampling error is less than .05.

9.4 ALL PAIRWISE SIMULTANEOUS CONFIDENCE INTERVALS AND TESTS: TUKEY–KRAMER

The Tukey–Kramer method is appropriate for all pairwise tests and/or simultaneous confidence intervals. This will come as a surprise to those who have read the first edition of this book. At the time that edition was written recent theoretical results cast doubt on the adequacy of the Tukey approach when used with random rather than fixed covariates (Bryant and Brunvold, 1980; Bryant and Paulson, 1976; Scheffé, 1959; Thigpen and Paulson, 1974). It appeared that it was necessary to make allowances for random (rather than fixed) covariates(s) in order to maintain type I error at the nominal level. After much theoretical work, tables of critical values for a modified version of the Studentized range distribution (called the generalized Studentized range distribution) were provided by Bryant and Paulson (1976) and Thigpen and Paulson (1974). About a decade later Hochberg and Varon-Salomon (1984) demonstrated that the Tukey–Kramer procedure maintains the familywise error rate at the nominal level even when the covariate is random. As a bonus the *T–K* method provides higher power and shorter simultaneous confidence intervals in most cases.

Although the *T–K* is classified by some statisticians as an approximate procedure in the case of ANCOVA (Westfall et al., 1999), there is no reason to avoid using it in constructing simultaneous confidence intervals. (An exact but computationally intensive method is available and has been implemented in *SAS*) (Westfall et al., 1999.) Unlike the Fisher–Hayter procedure, which is for hypothesis tests *only*, the *T–K* procedure can be employed for both hypothesis tests and simultaneous confidence intervals. The disadvantage of the *T–K* is that it has lower power than the Fisher–Hayter when it is used for hypothesis tests.

Computation Procedure for Tukey–Kramer Tests and Simultaneous Confidence Intervals

The *T–K* test statistic is

$$\frac{\bar{Y}_{i\,\text{adj}} - \bar{Y}_{j\,\text{adj}}}{\sqrt{\dfrac{\text{MS}_{\text{Res}_\text{w}}\left[\dfrac{1}{n_i} + \dfrac{1}{n_j} + \dfrac{(\bar{X}_i - \bar{X}_j)^2}{\text{SS}_{\text{w}_x}}\right]}{2}}} = q,$$

where $\text{MS}_{\text{Res}_\text{w}}$ is the mean square error from the ANCOVA summary table and SS_{wx} is the sum of squares within groups from the ANOVA on the covariate X. The obtained value of q is compared with the critical value of the Studentized range statistic $q_{J,\,N\,-\,J\,-\,1}$. This value is found in the Studentized range table at the intersection of the column headed with the value equal to J (the number of groups) and the row containing degrees of freedom $= N - J - 1$. The hypothesis that populations i and j have equal adjusted population means is rejected if the obtained q is equal to or greater than the critical q.

The 95% simultaneous confidence intervals around pairwise differences between adjusted means are computed by using

$$\left(\bar{Y}_{i\,\text{adj}} - \bar{Y}_{j\,\text{adj}}\right) \pm q_{(.05,J,N-J-1)}\sqrt{\frac{\text{MS}_{\text{Res}_w}\left[\frac{1}{n_i}+\frac{1}{n_j}+\frac{(\bar{X}_i-\bar{X}_j)^2}{\text{SS}_{w_x}}\right]}{2}} = (L,\ U),$$

where L and U are the lower and upper limits defining the interval.

Numerical Example of T–K Tests and Simultaneous Confidence Intervals
Suppose that we are interested in all pairwise comparisons in the experiment. Because the adjusted means are 28.48, 40.33, and 36.19 the sample differences are

$$28.48 - 40.33 = -11.85;$$
$$28.48 - 36.19 = -7.71;\ \text{and}$$
$$40.33 - 36.19 = 4.14.$$

The statistic for testing H_0: $\mu_{1\,\text{adj}} = \mu_{2\,\text{adj}}$ is

$$\frac{-11.85}{\sqrt{64.64\left[\frac{1}{10}+\frac{1}{10}+\frac{25}{5700}\right]}} = 4.61 = q$$

and the 5% critical value of q is 3.51. The hypothesis is rejected.

If the two remaining pairwise comparisons are computed, it will be discovered that they are not statistically significant. We conclude that the only significant difference is the difference between treatments 1 and 2. This demonstrates that the Fisher–Hayter approach is more powerful than the *T–K* approach; the *F–H* identified two significant differences rather than one.

Software
Minitab's GLM routine provides Tukey–Kramer and other multiple comparison tests and simultaneous confidence intervals. Suppose you are interested in computing ANCOVA, adjusted means, residuals of the model, Cooks distance, residual plots, *T–K* tests, and simultaneous confidence intervals. The following commands are used:

```
MTB > GLM 'Y' = TX;
SUBC>   Covariates 'X';
SUBC>   Brief 1;
SUBC>   Means TX;
SUBC>   Residuals 'RESI4';
```

```
SUBC>   CookD 'COOK4';
SUBC>   GFourpack;
SUBC>   RType 1;
SUBC>   Pairwise TX;
SUBC>     Tukey.
```

Output

```
General Linear Model: Y versus TX

Factor  Type   Levels  Values
TX      fixed       3  1, 2, 3

Analysis of Variance for Y, using Adjusted SS for Tests
Source  DF   Seq SS   Adj SS   Adj MS      F      P

X        1  1567.36  1855.35  1855.35  28.70  0.000
TX       2   707.99   707.99   354.00   5.48  0.010
Error   26  1680.65  1680.65    64.64
Total   29  3956.00


S = 8.03992   R-Sq = 57.52%   R-Sq(adj) = 52.61%

Means for Covariates

Covariate   Mean  StDev
X          49.33  14.17

Least Squares Means for Y

TX   Mean  SE Mean
1   28.48    2.558
2   40.33    2.555
3   36.19    2.543


Tukey 95.0% Simultaneous Confidence Intervals
Response Variable Y
All Pairwise Comparisons among Levels of TX

TX = 1  subtracted from:

TX   Lower  Center  Upper  ---+---------+--------+--------+---
2    2.831  11.853  20.87                  (--------*--------)
3   -1.248   7.712  16.67             (--------*--------)
                           ---+---------+--------+--------+---

                            -10         0       10       20
```

```
TX = 2  subtracted from:

TX   Lower  Center  Upper  ---+---------+--------+--------+---
3   -13.08  -4.141  4.799  (--------*--------)
                           ---+---------+--------+--------+---

                            -10         0       10       20
```

Tukey Simultaneous Tests

```
Response Variable Y
All Pairwise Comparisons among Levels of TX

TX = 1  subtracted from:

    Difference       SE of           Adjusted
TX    of Means  Difference  T-Value   P-Value
2       11.853       3.635    3.261    0.0084
3        7.712       3.610    2.136    0.1020

TX = 2  subtracted from:

    Difference       SE of           Adjusted
TX    of Means  Difference  T-Value   P-Value
3       -4.141       3.602   -1.150    0.4931

Residual Plots for Y
```

(Neither these plots nor the requested Cook's distance values are shown here.)

The output for the *T–K* tests (labeled as "Tukey simultaneous tests") shows "*T*-values" but these are not based on conventional two-sample *t*-tests; they provide the *p*-values that are associated with *q*. (It turns out that *q* can be easily transformed to a *t*-statistic using $\frac{q}{\sqrt{2}} = t$.) The *T–K* simultaneous confidence intervals are labeled as "Tukey 95% simultaneous confidence intervals," and adjusted means are labeled as "least squares" means.

9.5 PLANNED PAIRWISE AND COMPLEX COMPARISONS: BONFERRONI

Occasionally an experiment is designed in which the experimenter has interest in a limited number of pairwise and/or complex comparisons. Suppose that the three following comparisons were planned (before data collection) in a six-group experiment: (1 vs. 2), (3 vs. 4), and (5 vs. 6).

The Bonferroni procedure is appropriate for *planned* pairwise comparisons of this type (as well as complex comparisons) rather than for comparisons that are suggested by the outcome of the experiment.

The advantages of this procedure over the *T–K* procedure are: (1) there is greater power when the number of comparisons relative to the total number of possible pairwise comparisons is small, and (2) complex comparisons can be tested. If an investigator should decide to make a large number of planned pairwise comparisons (say, almost as many as the total number of pairwise comparisons), the power advantage of the Bonferroni procedure will be lost. It is only when the number of comparisons is relatively small that the Bonferroni is superior to Fisher–Hayter or Tukey–Kramer.

Computation Procedure for Bonferroni Tests and Simultaneous Confidence Intervals

A formula for carrying out a Bonferroni test is

$$\frac{c_1\left(\bar{Y}_{1\,\text{adj}}\right) + c_2\left(\bar{Y}_{2\,\text{adj}}\right) + \cdots + c_J\left(\bar{Y}_{J\,\text{adj}}\right)}{\sqrt{\text{MS}_{\text{Res}_\text{w}}\left[1 + \dfrac{\text{MS}_{\text{b}_x}}{\text{SS}_{\text{w}_x}}\right]\left[\dfrac{(c_1)^2}{n_1} + \dfrac{(c_2)^2}{n_2} + \cdots + \dfrac{(c_J)^2}{n_J}\right]}} = t_B,$$

where

$c_1, c_2, \ldots, c_J$ are the contrast coefficients for comparisons of interest;
$\bar{Y}_{1\,\text{adj}}, \ldots, \bar{Y}_{J\,\text{adj}}$ are the adjusted means;
$n_1, \ldots, n_J$ are the sample sizes associated with groups;
$\text{MS}_{\text{Res}_\text{w}}$ is the error term obtained from the ANCOVA summary table;
MS_{b_x} is the mean square between groups obtained from summary table for ANOVA on covariate; and
SS_{w_x} is the sum of squares within groups obtained from summary table for ANOVA on covariate.

The critical value against which the absolute obtained t_B is compared (i.e., $t_{\text{B}(\alpha,\, C',\, N-J-1)}$) is usually called the Bonferroni *t*-statistic, which can be found in the appendix. (Dunn (1961) was the first to tabulate the critical values of the Bonferroni *t* and for this reason the method is sometimes called the Dunn- or Dunn–Bonferroni procedure.) It is important to note that the second term in the subscript (i.e., C') is the number of planned comparisons—not the number of groups or the number of covariates. The third term is the degrees of freedom associated with the ANCOVA error term.

Each $100(1-\alpha)\%$ interval in the set of planned comparisons is constructed by using

$$c_1\left(\bar{Y}_{1\,\text{adj}}\right) + c_2\left(\bar{Y}_{2\,\text{adj}}\right) + \cdots + c_J\left(\bar{Y}_{J\,\text{adj}}\right) \mp t_{\text{B}(\alpha, C', N-J-1)} \times \sqrt{\text{MS}_{\text{Res}_\text{w}}\left[1 + \frac{\text{MS}_{\text{b}_x}}{\text{SS}_{\text{w}_x}}\right]\left[\frac{(c_1)^2}{n_1} + \frac{(c_2)^2}{n_2} + \cdots + \frac{(c_J)^2}{n_J}\right]}$$

Numerical Example of Bonferroni Tests and Simultaneous Confidence Intervals
If the investigator conducting the example study planned to compare the treatment 1 adjusted mean with the treatment 2 adjusted mean, and the treatment 2 adjusted mean

with the treatment 3 adjusted mean, two of three possible pairwise comparisons are involved. Because the ANCOVA F is significant, the Fisher–Hayter procedure is the best choice as a testing procedure; but if simultaneous confidence intervals are desired, the Fisher–Hayter does not qualify. *Minitab* output in the previous section shows that the *T–K* simultaneous confidence intervals associated with the two-planned contrasts are similar to the Bonferroni intervals. Computation is shown below for the planned Bonferroni tests and intervals.

Tests

1.
$$\frac{1(28.48) - 1(40.33) + 0(36.19)}{\sqrt{64.64[1 + (63.33/5700)][(1)^2/10 + (-1)^2/10 + (0)^2]}}$$
$$= \frac{-11.85}{\sqrt{(65.36)(0.2)}} = \frac{-11.28}{3.62} = -3.27 = t_{\mathrm{B_{obt}}}$$

2.
$$\frac{0(28.48) + 1(40.33) - 1(36.19)}{\sqrt{64.64[1 + (63.33/5700)][(0)^2/10 + (1)^2/10 + (-1)^2/10]}}$$
$$= \frac{4.14}{\sqrt{(65.36)(0.2)}} = \frac{4.14}{3.62} = 1.14 = t_{\mathrm{B_{obt}}}$$

The critical value against which the obtained t_{B} values are compared is $t_{\mathrm{B}(\alpha,\, 2,\, 26)}$ or 2.38 for $\alpha = .05$.

Simultaneous confidence intervals are as follows:

1. The confidence interval around $\left[1\left(\bar{Y}_{1\,\mathrm{adj}}\right) - 1\left(\bar{Y}_{2\,\mathrm{adj}}\right) + 0\left(\bar{Y}_{3\,\mathrm{adj}}\right)\right]$ on $\left[1\left(\mu_{1\,\mathrm{adj}}\right) - 1\left(\mu_{2\,\mathrm{adj}}\right) + 0\left(\mu_{3\,\mathrm{adj}}\right)\right]$ is $-11.85 \pm 2.38(3.62) = (-20.47, -3.23)$.
2. The confidence interval around $\left[0\left(\bar{Y}_{1\,\mathrm{adj}}\right) + 1\left(\bar{Y}_{2\,\mathrm{adj}}\right) - 1\left(\bar{Y}_{3\,\mathrm{adj}}\right)\right]$ on $\left[0\left(\mu_{1\,\mathrm{adj}}\right) + 1\left(\mu_{2\,\mathrm{adj}}\right) - 1\left(\mu_{3\,\mathrm{adj}}\right)\right]$ is $4.14 \pm 2.38(3.62) = (-4.48, 12.76)$.

Nonrandomized Studies

If the data have been obtained from a randomized experiment, the formulas presented above are appropriate. If the data have been obtained from a biased assignment or nonrandomized study, the following formula is more appropriate:

$$\frac{c_1\bar{Y}_{1\,\mathrm{adj}} + c_2\bar{Y}_{2\,\mathrm{adj}} + \cdots + c_J\bar{Y}_{J\,\mathrm{adj}}}{\sqrt{\mathrm{MS}_{\mathrm{Res_w}}\left[\frac{c_1^2}{n_1} + \frac{c_2^2}{n_2} + \cdots + \frac{c_J^2}{n_J}\left[\left(c_1\bar{X}_1 + c_2\bar{X}_2 + \cdots + c_J\bar{X}_J\right)/\mathrm{SS}_{\mathrm{w}_x}\right]\right]}} = t_{\mathrm{B}}.$$

Note that the error term is computed separately for each contrast when this formula is employed. The critical value is $t_{\mathrm{B}(\alpha,\, C',\, N-J-1)}$, which is the same value used with the randomized experiment formula.

9.6 ANY OR ALL COMPARISONS: SCHEFFÉ

The Scheffé procedure is the most general of the available procedures for making multiple comparisons. Although it can be used for any type or number of comparisons, it is recommended for only certain types of comparisons. The Scheffé approach is most useful when (1) the experimenter discovers comparisons he would like to test after examining the outcome of the experiment (i.e., data snooping) and (2) a very large mixture of pairwise and/or complex comparisons are planned. When planned complex comparisons are involved, the choice between the Scheffé and Bonferroni procedures can easily be made by selecting the one with the smaller critical value.

The Scheffé test can be used to test any possible contrast; if the ANCOVA F is significant there is some contrast [not necessarily an (i,j) contrast] that is significant using the Scheffé procedure. If the ANCOVA F is not significant, there is no contrast that will be declared significant using the Scheffé procedure.

Computation Procedure for Scheffé Tests

The test statistic is

$$\frac{c_1\left(\bar{Y}_{1\,\text{adj}}\right) + c_2\left(\bar{Y}_{2\,\text{adj}}\right) + \cdots + c_J\left(\bar{Y}_{J\,\text{adj}}\right)}{\sqrt{\text{MS}_{\text{Res}_\text{w}}\left[1 + \left(\text{MS}_{\text{b}_x}/\text{SS}_{\text{w}_x}\right)\right]\left\{\left[(c_1)^2/n_1\right] + \left[(c_2)^2/n_2\right] + \cdots + \left[(c_J)^2/n_J\right]\right\}}} = F',$$

where

$c_1, c_2, \ldots, c_J$ are the contrast coefficients for comparison of interest;
$\bar{Y}_{1\,\text{adj}}, \bar{Y}_{2\,\text{adj}}, \ldots, \bar{Y}_{J\,\text{adj}}$ are the adjusted means;
$n_1, n_2, \ldots, n_J$ are the sample sizes associated with groups $1, 2, \ldots, J$;
$\text{MS}_{\text{Res}_\text{w}}$ is the error term obtained from ANCOVA summary table;
MS_{b_x} is the mean square between groups obtained from summary table for ANOVA on covariate; and
SS_{w_x} is the sum of squares within groups obtained from summary table for ANOVA on covariate.

The critical value against which the obtained F' is compared is

$$\sqrt{(J-1)F_{(\alpha, J-1, N-J-1)}}.$$

Simultaneous Confidence Intervals Based on Scheffé Procedure

The simultaneous 95% confidence intervals are computed by using the following formula for each comparison:

$$c_1\left(\bar{Y}_{1\,\text{adj}}\right) + c_2\left(\bar{Y}_{2\,\text{adj}}\right) + \cdots + c_J\left(\bar{Y}_{J\,\text{adj}}\right) \pm \sqrt{(J-1)F_{(\alpha, J-1, N-J-1)}}$$

$$\times \sqrt{\text{MS}_{\text{Res}_\text{w}}\left[1 + \left(\text{MS}_{\text{b}_x}/\text{SS}_{\text{w}_x}\right)\right]\left[\frac{c_1^2}{n_1} + \frac{c_2^2}{n_2} + \cdots + \frac{c_J^2}{n_J}\right]}.$$

Numerical Example of Scheffé Tests and Simultaneous Confidence Intervals
Suppose that the six (i, j) comparisons in the example study are of interest. The contrast estimates are as follows:

$$1(28.48) - 1(40.33) + 0(36.19) = -11.85 = \hat{\psi}_1$$
$$1(28.48) + 0(40.33) - 1(36.19) = -7.71 = \hat{\psi}_2$$
$$0(28.48) + 1(40.33) - 1(36.19) = 4.14 = \hat{\psi}_3$$
$$1(28.48) - 0.5(40.33) - 0.5(36.19) = -9.78 = \hat{\psi}_4$$
$$-0.5(28.48) + 1(40.33) - 0.5(36.19) = 7.995 = \hat{\psi}_5$$
$$-0.5(28.48) - 0.5(40.33) + 1(36.19) = -1.785 = \hat{\psi}_6$$

The test statistics for the contrast estimates are as follows:

$$\frac{-11.85}{\sqrt{64.64[1 + (63.33/5700)]\left[(1)^2/10 + (-1)^2/10 + (0)^2/10\right]}} = \frac{-11.85}{3.62} = -3.28$$

$$\frac{-7.71}{\sqrt{64.64[1 + (63.33/5700)]\left[(1)^2/10 + (-1)^2/10 + (-1)^2/10\right]}} = \frac{-7.71}{3.62} = -2.13$$

$$\frac{-4.14}{\sqrt{64.64[1 + (63.33/5700)]\left[(0)^2/10 + (1)^2/10 + (-1)^2/10\right]}} = \frac{4.14}{3.62} = 1.14$$

$$\frac{-9.78}{\sqrt{64.64[1 + (63.33/5700)]\left[(1)^2/10 + (-0.5)^2/10 + (-0.5)^2/10\right]}} = \frac{-9.78}{3.13} = -3.12$$

$$\frac{-7.995}{\sqrt{64.64[1 + (63.33/5700)]\left[(-0.5)^2/10 + (1)^2/10 + (-0.5)^2/10\right]}} = \frac{7.995}{3.13} = 2.55$$

$$\frac{-1.785}{\sqrt{64.64[1 + (63.33/5700)]\left[(-0.5)^2/10 + (-0.5)^2/10 + (1)^2/10\right]}} = \frac{1.785}{3.13} = 0.57$$

The absolute value of each obtained F' is compared with the critical value of F', which is

$$\sqrt{2\left[F_{(.05,2,26)}\right]} = \sqrt{2[3.37]} = 2.60.$$

The hypotheses

$$\mu_{1\,\text{adj}} - \mu_{2\,\text{adj}} = 0.0 \quad \text{and}$$
$$\mu_{1\,\text{adj}} - \frac{\mu_{2\,\text{adj}} + \mu_{3\,\text{adj}}}{2} = 0.0$$

are rejected at the .05 level because the test statistics associated with $\hat{\psi}_1$ and $\hat{\psi}_4$ yield absolute values of F'_{obt} that exceed the critical value of F' (i.e., 3.28 and 3.12 are greater than 2.60).

The 95% simultaneous confidence intervals on Ψ_1, Ψ_2, Ψ_3, Ψ_4, Ψ_5, and Ψ_6 are

$\hat{\psi}$	Confidence interval
$-11.85 \pm (2.60)(3.62)$	$= (-21.26, -2.44)$
$-7.71 \pm (2.60)(3.62)$	$= (-17.12, 1.70)$
$4.14 \pm (2.60)(3.62)$	$= (-5.27, 13.55)$
$-9.78 \pm (2.60)(3.13)$	$= (-17.92, -1.64)$
$7.995 \pm (2.60)(3.13)$	$= (-0.14, 16.13)$
$1.785 \pm (2.60)(3.13)$	$= (-6.35, 9.92)$

The probability is at least .95 that the entire set of confidence intervals constructed according to the Scheffé procedure will cover all of the population contrasts they estimate. Alternatively, the probability is $\leq$.05 that one or more of the population contrasts will not be covered by such a set of intervals.

Nonrandomized Studies

The Scheffé formula for nonrandomized studies is

$$\frac{c_1\bar{Y}_{1\,\text{adj}} + c_2\bar{Y}_{2\,\text{adj}} + \cdots + c_J\bar{Y}_{J\,\text{adj}}}{\sqrt{\text{MS}_{\text{Res}_w}\left[\frac{c_1^2}{n_1} + \frac{c_2^2}{n_2} + \cdots + \frac{c_J^2}{n_J}\left[\left(c_1\bar{X}_1 + c_2\bar{X}_2 + \cdots + c_J\bar{X}_J\right)/\text{SS}_{w_x}\right]\right]}} = F'.$$

The critical value of F' is

$$\sqrt{(J-1)F_{(\alpha,J-1,N-J-1)}}.$$

9.7 IGNORE MULTIPLE COMPARISON PROCEDURES?

The publication requirements in many areas are quite rigid regarding the necessity of using multiple comparison procedures to handle the multiplicity problem. This is true in both psychology and medicine. But this is not a universal position. Some very bright statisticians and epidemiologists either do not worry about adjusting for them (e.g., Gelman and Hill, 2007) or are quite opposed to doing so (Rothman, 1990). I will attempt to summarize their positions in a couple of lines. The Gelman–Hill position is that (1) point null hypotheses are almost never true and (2) we already know that about 5% of 95% confidence intervals exclude the true values, so simply keep that in mind. The Rothman position is that (1) when the data are actual observations (rather than random numbers), it is more reasonable to entertain the premise that nature follows regular laws than to use chance as the first-order explanation and (2) the

increase in type II error that accompanies corrections for multiplicity is unwarranted. Many other statisticians are not in support of multiple comparison procedures (e.g., Cobb, 1998) and have provided reasoned counter arguments. It is, however, a rare manuscript submission in psychology or NIH grant proposal that will fly without acknowledgement of the issue.

9.8 SUMMARY

The ANCOVA F-test is often only a starting point in the analysis of studies with three or more groups. To analyze differences between particular pairs of adjusted means (pairwise comparisons) or differences based on complex linear combinations of adjusted means (complex comparisons), multiple comparison procedures are recommended. Four different multiple comparison procedures were described: the Fisher–Hayter procedure, the Tukey–Kramer procedure, the Bonferroni procedure, and the Scheffé method. The Fisher–Hayter was suggested for hypothesis tests on pairwise comparisons. The Tukey–Kramer was suggested for both hypothesis tests and simultaneous confidence intervals on pairwise comparisons. The Bonferroni procedure was recommended for hypothesis tests and simultaneous confidence intervals on a small number of planned comparisons of any type (simple or complex). The Scheffé method was suggested for both hypothesis tests and simultaneous confidence intervals on any comparisons that are suggested by the data or for a large number of planned pairwise and/or complex comparisons.

The correct interpretation of multiple comparison procedures is based on the concept of familywise error. The α associated with a multiple comparison procedure is the probability that one or more of the family of comparisons will be incorrectly declared significant. Simultaneous confidence intervals have a parallel interpretation. That is, the set of population differences between adjusted means will be included in the simultaneous confidence intervals on these differences at least 95% of the time.

CHAPTER 10

Multiple Covariance Analysis

10.1 INTRODUCTION

It is not unusual for a researcher to be interested in controlling more than one source of unwanted variability. Suppose that a state department of education is interested in comparing the effectiveness of a sex education instructional package that includes a large television component with a similar package that does not include the television aspect. If students from a large geographic area are randomly assigned to the two treatments, there is likely to be sizable sampling variability both between and within treatment groups with respect to variables such as race, parent's income, previous educational experiences, and religion. If measures are available on these variables, multiple covariance analysis can be employed as a control procedure by using all the measures as covariates. The issues discussed in previous chapters for the case of a single covariate generalize directly to the case of multiple covariates.

There are other procedures for controlling several sources of unwanted (nuisance) variability, such as direct experimental control, selection, and factorial designs. An investigator could control for race, parent's income, previous educational experiences, and religion by selecting subjects who belong to the same race and religion, have parents at one income level, and have the same educational experiences. This approach would require a considerable amount of effort to obtain a reasonable sample size; more importantly, it would require that the selected subjects in some way be isolated from other students for treatment purposes. It is not realistic to think that such subjects could be (or should be) isolated for this purpose. The use of each of these variables as a factor (in addition to the treatment factor) in a factorial ANOVA design is a more practical strategy. But there are weaknesses with this approach also. When a control variable is a continuous variable (e.g., income or amount of experience) information is lost by forming just a few levels of such factors; thus if income is broken down into low, medium, and high levels, there will be information lost concerning differences in income within any one of these crude categories. On the other hand, if a large

The Analysis of Covariance and Alternatives: Statistical Methods for Experiments, Quasi-Experiments, and Single-Case Studies, Second Edition. Bradley E. Huitema.

number of levels is included (say, 10 levels of income) there will be a correspondingly large loss of error degrees of freedom that will tend to decrease the power of the analysis.

Multiple ANCOVA circumvents the problems mentioned above. All available subjects can be included in the analysis, there is no problem of coarse grouping, and each covariate (regardless of the number of different values of X) has one degree of freedom. All covariates can be included simultaneously in one analysis of covariance. The adjusted treatment mean differences are then interpreted as being independent of all variables included as covariates. This interpretation is straightforward when the data are based on a randomized design.

Multiple ANCOVA is often employed with nonrandomized studies to eliminate bias in the comparison of several groups. But the problems here are essentially the same as were mentioned with one covariate. Multiple ANCOVA will *not* equate groups that differ on the covariates. It may *reduce* bias on Y that is predictable from the covariates, but it will generally not *eliminate* bias.

Suppose an evaluation of the effect of an advertising program is undertaken by comparing the sales in one city where the program was used with another city where the program was not used. It is reasonable in this case, to ask whether the cities are equivalent on all variables relevant to sales. Let us say that income level, average age, and political conservatism are believed to be the major dimensions that distinguish the two cities. If these variables are employed as multiple covariates and the adjusted mean difference on Y is statistically significant, this does *not* mean that the advertising program was effective. Two problems must be kept in mind in the interpretation of such an analysis. First, it is unlikely that differences between the groups can be *completely* explained by a few variables. If all variables that are required to explain differences between groups on a particular dependent variable are not included as covariates, the groups will not be equated. Second, even if the appropriate variables are included in the analysis, measurement error associated with the measurement of these variables will generally lead to an underadjustment of the means. Multiple ANCOVA will yield means on Y adjusted for the variability accounted for by the covariates *as measured*. It would be appropriate to state that the groups were equated on the measured or obtained scores but not on the true scores.

Fortunately, these problems do not yield ANCOVA and multiple ANCOVA useless with observational studies. As long as it is recognized that these analyses generally *reduce* rather than remove bias in the comparison of groups, there is no problem. Very often the best that can be done with these designs is to employ as many covariates as can reasonably be expected to explain group differences.

When covariates are chosen an attempt should be made to select variables that are not highly redundant. That is, the correlations among the covariates should be considered. If the correlations are too high, it should be recognized that (1) the estimates of the parameters may be unstable and (2) the computer programs employed to estimate the equations required in multiple ANCOVA may not perform properly. The latter problem is easy to identify because most computer programs will generally abort the computation and provide a message stating that the "inverse of the matrix can't be computed," that the matrix is "singular," or that "a linear dependence among

predictor variables exists." All these messages refer to the problem of extremely high correlation among variables in the model. These problems are referred to in the regression literature under the topic of multicollinearity.

It is possible to use up to $N - (J + 1)$ covariates, where N is the total number of subjects and J is the number of groups. But it is good practice to limit the number of covariates to the extent that the ratio

$$\frac{C + (J - 1)}{N},$$

(where C is the number of covariates) does not exceed 0.10. If this ratio is greater than 0.10, the ANCOVA F-test is valid but the estimates of the adjusted means are likely to be unstable. That is, if a study with a high $C + (J - 1)/N$ ratio is cross-validated, it can be expected that the equation that is used to estimate the adjusted means in the original study will yield very different estimates for another sample from the same population.

The hypothesis tested with multiple ANCOVA is the same as with simple ANCOVA, except that the adjusted means are adjusted with respect to multiple covariates rather than a single covariate. A graphic representation of simple and multiple ANCOVA is presented in Figure 10.1. It can be seen that the adjusted population means in simple ANCOVA are conditional on the single covariate X. Under multiple ANCOVA the adjusted population means are conditional on both X_1 and X_2. In general, if C covariates are employed, the adjusted means are conditional on C covariates and the multiple ANCOVA F-test is a test of the equality of the elevations of J hyperplanes.

The interpretation problems described in Chapters 6 and 7 for simple ANCOVA also apply to the multiple ANCOVA model, which is

$$Y_{ij} = \mu + \alpha_j + \beta_1(X_{1_{ij}} - \bar{X}_{1_{..}}) + \beta_1(X_{2_{ij}} - \bar{X}_{2_{..}}) + \cdots + \beta_C(X_{C_{ij}} - \bar{X}_{C_{..}}) + \varepsilon_{ij},$$

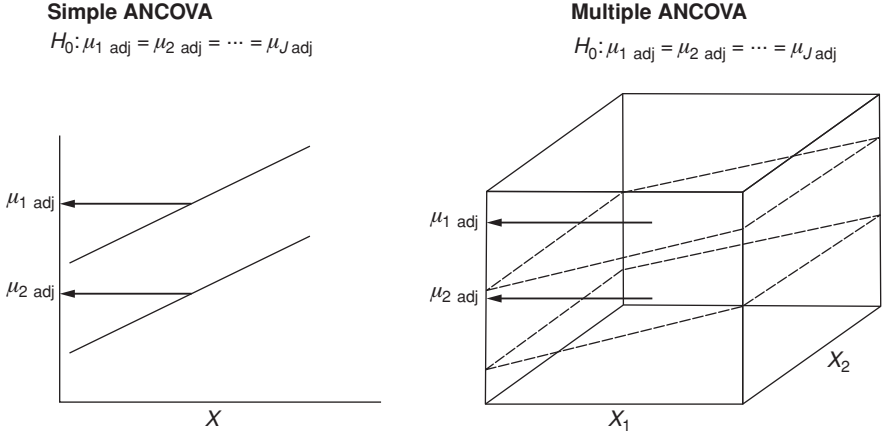


Figure 10.1 Comparison of simple and multiple covariance analysis.

where

μ is the grand population mean of all Y scores;
α_j is the effect of jth treatment;
$\beta_1, \beta_2 \ldots \beta_C$ are the partial regression coefficients associated with covariates 1 through C;
$X_{1_{ij}}, X_{2_{ij}} \ldots X_{C_{ij}}$ are the scores of covariates X_1 through X_C for the ith subject in jth group; and
ε_{ij} is the deviation of the observation from adjusted mean of group of which it is a member (i.e., $Y_{ij} - \mu_{j\,\text{adj}}$).

The assumptions associated with this model are also straightforward extensions of those associated with simple ANCOVA.

10.2 MULTIPLE ANCOVA THROUGH MULTIPLE REGRESSION

Multiple ANCOVA can easily be carried out using multiple regression. The rationale and computation are straightforward extensions of simple ANCOVA. Suppose that two covariates had been involved in the three-group study described in Chapter 6 rather than one. In addition to the aptitude test scores originally used as the covariate, we are going to use scores from an academic motivation test. The predictors for the complete multiple covariance analysis are presented in Table 10.1. Aptitude scores are in the X_1 column and academic motivation scores are in the X_2 column.

By regressing Y on the first four columns, we obtain the R^2, which provides us with the proportion of the variability in Y accounted for by both covariates and the dummy variables. We then regress Y on X_1 and X_2 to obtain the proportion (R^2) accounted for by only the covariates. The difference between the two R^2-values represents the proportion of the variability accounted for by the dummy variables (i.e., group membership or treatment effect) that is independent of variability accounted for by X_1 and X_2. The F-test of the adjusted treatment effects takes the general form tabulated as follows:

Source	SS	*df*	MS	*F*
Adjusted treatment	$\left(R^2_{yD,X} - R^2_{yX}\right) SST$	$J - 1$	$SS_{AT}/(J-1)$	MS_{AT}/MS_{res_w}
Multiple within residual	$\left(1 - R^2_{yD,X}\right) SST$	$N - J - C$	$SS_{res_w}/(N-J-C)$	
Multiple total residual	$\left(1 - R^2_{yX}\right) SST$	$N - C - 1$		

or

$$\frac{\left(R^2_{yD,X} - R^2_{yX}\right)/(J-1)}{\left(1 - R^2_{yD,X}\right)/(N-J-C)} = F$$

Table 10.1 Predictors for Multiple ANCOVA and Associated Tests

(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	
d_1	d_2	X_1	X_2	d_1X_1	d_2X_1	d_1X_2	d_2X_2	Y
1	0	29	3	29	0	3	0	15
1	0	49	3	49	0	3	0	19
1	0	48	2	48	0	2	0	21
1	0	35	5	35	0	5	0	27
1	0	53	5	53	0	5	0	35
1	0	47	9	47	0	9	0	39
1	0	46	3	46	0	3	0	23
1	0	74	7	74	0	7	0	38
1	0	72	6	72	0	6	0	33
1	0	67	8	67	0	8	0	50
0	1	22	3	0	22	0	3	20
0	1	24	2	0	24	0	2	34
0	1	49	4	0	49	0	4	28
0	1	46	4	0	46	0	4	35
0	1	52	5	0	52	0	5	42
0	1	43	4	0	43	0	4	44
0	1	64	8	0	64	0	8	46
0	1	61	7	0	61	0	7	47
0	1	55	6	0	55	0	6	40
0	1	54	5	0	54	0	5	54
0	0	33	2	0	0	0	0	14
0	0	45	1	0	0	0	0	20
0	0	35	5	0	0	0	0	30
0	0	39	4	0	0	0	0	32
0	0	36	3	0	0	0	0	34
0	0	48	8	0	0	0	0	42
0	0	63	8	0	0	0	0	40
0	0	57	4	0	0	0	0	38
0	0	56	9	0	0	0	0	54
0	0	78	7	0	0	0	0	56

where

$R^2_{yD,X}$ is the coefficient of multiple determination obtained by regressing dependent variable on group membership dummy variables and covariates;

R^2_{yX} is the coefficient of multiple determination obtained by regressing dependent variable on covariates;

N is the total number of subjects;

J is the number of groups; and

C is the number of covariates.

The obtained F is evaluated with $F_{(\alpha, J-1, N-J-C)}$.

For the data in Table 10.1, we can see that columns 1 and 2 = D and columns 3 and 4 = X. The R^2-values are

$$R^2_{yD,X} = R^2_{y1234} = 0.754186 \quad \text{and}$$
$$R^2_{yX} = R^2_{y34} = 0.596264$$

The difference or unique contribution of $D = 0.157922$.

Source	SS	*df*	MS	*F*
Adjusted treatment	(0.157922)3956 = 624.74	2	312.37	8.03
Multiple within residual	(0.245814)3956 = 972.44	25	38.90	
Multiple total residual	(0.403736)3956 = 1597.18	27		

or

$$\frac{0.157922/2}{0.245814/25} = \frac{0.78961}{0.009833} = 8.03$$

The critical values of F for the .05 and .01 significance levels are 3.38 and 5.57, respectively. Multiple covariance has resulted in a higher F-value in this case than either ANOVA or single-covariate ANCOVA. The ANOVA F is 1.60 (not significant), and the single-covariate ANCOVA F is 5.48 ($p < .05$).

The required sums of squares can be obtained more directly from the regression analysis summary tables. The reason for using the R^2 approach illustrated here is to simply convey the essential aspects of the analysis using proportions. I recommend using dedicated ANCOVA routines for actual data analysis.

10.3 TESTING HOMOGENEITY OF REGRESSION PLANES

Just as we assume homogeneous regression slopes in simple ANCOVA, we assume homogeneous regression planes or hyperplanes in multiple covariance analysis. A test of homogeneous regression planes should be carried out before proceeding with the interpretation of multiple ANCOVA. The general linear regression approach to this test is described in this section. The advantage of the approach presented here is that the conceptual framework and the computation can be seen to be just a slight extension of the main ANCOVA. The test statistic is as follows:

$$\frac{\left(R^2_{yD,X,DX} - R^2_{yD,X}\right)/(C(J-1))}{\left(1 - R^2_{yD,X,DX}\right)/(N - [J(C+1)])} = F$$

where

$R^2_{yD,X,DX}$ is the coefficient of multiple determination obtained by regressing dependent variable on group membership dummy variables, covariates, and products of dummy variables and covariates;

$R^2_{yD,X}$ is the coefficient of multiple determination obtained by regressing dependent variable on group membership dummy variables and covariates;

N is the total number of subjects;

J is the number of groups; and

C is the number of covariates.

The obtained F is evaluated with $F_{[\alpha,C(J-1),N-[J(C+1)]]}$. The right-hand R^2 in the numerator is obtained as a step in the computation of the main ANCOVA analysis; $R^2_{yD,X,DX}$, the left-hand R^2 in the formula, will exceed $R^2_{yD,X}$ if a separate multiple regression equation is fitted to each individual group; in this case smaller residuals will be obtained than when a single pooled within-group multiple regression equation is fitted to all groups. For the example data of Table 10.1 we see that columns 1 and $2 = D$, columns 3 and $4 = X$, and columns 5 through $8 = DX$. Hence,

$$R^2_{yD,X,DX} = R^2_{y12345678} = 0.78573 \quad \text{and}$$
$$R^2_{yD,X} = R^2_{y1234} = 0.754186$$

The difference or unique contribution of DX is equal to 0.031547. The test statistic is

$$\frac{0.031547/4}{0.214267/21} = 0.77 = F$$

The critical value of F using $\alpha = .05$ is 2.84. Equivalently, the form found in the following table can be used:

Source	SS	*df*	MS	*F*
Heterogeneity of planes	$\left(R^2_{yD,X,DX} - R^2_{yD,X}\right) SST$	$C(J-1)$	MS_{het}	MS_{het}/MS_{res_i}
Multiple residual$_i$	$\left(1 - R^2_{yD,X,DX}\right) SST$	$N - [J(C+1)]$	MS_{res_i}	
Multiple residual$_w$	$\left(1 - R^2_{yD,X}\right) SST$	$N - J - C$		

With the example data the summary is

Source	SS	*df*	MS	*F*
Heterogeneity of Planes	124.80	4	31.20	0.77
Multiple residual$_i$	847.64	21	40.36	
Multiple residual$_w$	972.44	25		

Because the obtained F is less than the critical value of F we have little reason to doubt the homogeneity of the regression planes. Hence, we need not be concerned with the problem of treatment-covariate interaction.

10.4 COMPUTATION OF ADJUSTED MEANS

Adjusted means in multiple ANCOVA may be computed in a manner similar to the procedure suggested for simple ANCOVA. After the regression equation associated with $R^2_{yD,X}$ and the grand covariate means are obtained, each adjusted mean is computed from the equation

$$\bar{Y}_{j\,\text{adj}} = b_0 + b_1(d_1) + \cdots + b(d_{J-1}) + b(\bar{X}_1..) + \cdots + b(\bar{X}_C..)$$

The regression equation associated with $R^2_{yD,X}$ for the example of this section (i.e., R^2_{y1234}) includes the following coefficients:

$$b_0 = 7.9890045$$
$$b_1 = -6.82978$$
$$b_2 = 4.40365$$
$$b_3 = 0.276592$$
$$b_4 = 2.83490$$

The grand means for the two covariates are

$$\bar{X}_1.. = 49.333 \quad \text{and}$$
$$\bar{X}_2.. = 5.000.$$

Adjusted means are then obtained from the equation

$$\bar{Y}_{j\,\text{adj}} = 7.9890045 - 6.82978(d_1) + 4.40365(d_2) + 0.276592(\bar{X}_1..) + 2.83490(\bar{X}_2..).$$

The adjusted means for the three groups are

$$\begin{aligned}\bar{Y}_{1\,\text{adj}} &= 7.9890045 - 6.82978(1) + 4.40365(0) + 0.276592(49.333) + 2.83490(5.0)\\ &= 7.9890045 - 6.82978 + 0 + 13.645113 + 14.1745\\ &= \underline{28.98}\end{aligned}$$

$$\begin{aligned}\bar{Y}_{2\,\text{adj}} &= 7.9890045 - 6.82978(0) + 4.40365(1) + 0.276592(49.333) + 2.83490(5.0)\\ &= 7.9890045 - 0 + 4.40365 + 13.645113 + 14.1745\\ &= \underline{40.21}\end{aligned}$$

$$\begin{aligned}\bar{Y}_{3\text{ adj}} &= 7.9890045 - 6.82978(0) + 4.40365(0) + 0.276592(49.333) + 2.83490(5.0)\\ &= 7.9890045 - 0 + 0 + 13.645113 + 14.1745\\ &= \underline{35.81}\end{aligned}$$

10.5 MULTIPLE COMPARISON PROCEDURES FOR MULTIPLE ANCOVA

The formulas presented in Chapter 9 for multiple comparison tests must be modified for multiple covariance analysis. Many readers will not be familiar with the matrix notation employed in these formulas. Most textbooks on applied multivariate analysis contain chapters on matrix algebra.

The terms in the formulas presented in Tables 10.2 and 10.3 are defined as follows.

MS multiple res_w is the error mean square employed in overall multiple ANCOVA; $\mathbf{W}_\text{x}^{-1}$ is the inverse of the pooled corrected within group sum of products matrix for the covariates; that is,

$$\mathbf{W}_\text{x} = \begin{bmatrix} \sum x_1^2 & \sum x_1 x_2 & \cdots & \sum x_1 x_C \\ \sum x_2 x_1 & \sum x_2^2 & \cdots & \sum x_2 x_C \\ \cdot & \cdot & \cdots & \cdot \\ \cdot & \cdot & \cdots & \cdot \\ \cdot & \cdot & \cdots & \cdot \\ \sum x_C x_1 & \sum x_C x_2 & \cdots & \sum x_C^2 \end{bmatrix}$$

$\mathbf{S}_\text{b}$ is the between-group variance–covariance matrix for the covariates; that is, $\mathbf{S}_{\text{b}_\text{x}} = \frac{1}{j-1}\mathbf{B}_\text{x}$ where $\mathbf{B}_\text{x} - \mathbf{T}_\text{x} - \mathbf{W}_\text{x}\ \mathbf{T}_\text{x}$ is the corrected total sum of products matrix for the covariates, $\text{tr}(\mathbf{W}_\text{x}^{-1}\mathbf{S}_{\text{b}_\text{x}})$ is the trace of the product $\mathbf{W}_\text{x}^{-1}\mathbf{S}_{\text{b}_\text{x}}$, and n_h is the harmonic mean of the sample sizes associated with treatments i and j.

The vector $\mathbf{d}$ in the formulas in Table 10.3 is the column vector of differences between the ith and jth group means on the covariates. For example, to test the difference between adjusted means 1 and 3 in a three-group experiment with C covariates, we have

$$\begin{array}{c} \\ \text{Covariate 1} \\ \text{Covariate 2} \\ \vdots \quad \vdots \\ \text{Covariate } C \end{array} \begin{array}{c} \begin{array}{ccc} \text{Group 1} & & \text{Group 3} \end{array} \\ \begin{bmatrix} \bar{X}_{1,1} & - & \bar{X}_{1,3} \\ \bar{X}_{2,1} & - & \bar{X}_{2,3} \\ \vdots & \vdots & \vdots \\ \bar{X}_{C,1} & - & \bar{X}_{C,3} \end{bmatrix} \end{array} \begin{array}{c} \mathbf{d} \\ \begin{bmatrix} d_1 \\ d_2 \\ \vdots \\ d_C \end{bmatrix} \end{array}$$

Table 10.2 Formulas for Multiple Comparisons with Multiple Covariates in Randomized Experiments

Procedure	Formula	Critical Value
Fisher–Hayter	$\dfrac{\bar{Y}_{i\,\mathrm{adj}} - \bar{Y}_{j\,\mathrm{adj}}}{\sqrt{\dfrac{\mathrm{MS\ multiple\ Res_w}[1 + \mathrm{tr}(\mathbf{W}_\mathrm{x}^{-1}\mathbf{S}_{\mathrm{b_x}})}{n_\mathrm{h}}}} = q$	Studentized range $q_{J-1,\,N-J-C}$
Tukey–Kramer	$\dfrac{\bar{Y}_{i\,\mathrm{adj}} - \bar{Y}_{j\,\mathrm{adj}}}{\sqrt{\dfrac{\mathrm{MS\ multiple\ Res_w}[1 + \mathrm{tr}(\mathbf{W}_\mathrm{x}^{-1}\mathbf{S}_{\mathrm{b_x}})}{n_\mathrm{h}}}} = q$	Studentized range $q_{J,\,N-J-\mathrm{C}}$
Bonferroni	$\dfrac{c_1\bar{Y}_{1\,\mathrm{adj}} + c_2\bar{Y}_{2\,\mathrm{adj}} + \cdots + c_J\bar{Y}_{J\,\mathrm{adj}}}{\sqrt{\mathrm{MS\ multiple\ Res_w}\left[\dfrac{c_1^2}{n_1} + \dfrac{c_2^2}{n_2} + \cdots + \dfrac{c_J^2}{n_J}\right][1 + tr(\mathbf{W}_\mathrm{x}^{-1}\mathbf{S}_{\mathrm{b_x}})]}} = t$	Bonferroni $t_{B,C',\,N-J-C}$
Scheffé	$\dfrac{c_1\bar{Y}_{1\,\mathrm{adj}} + c_2\bar{Y}_{2\,\mathrm{adj}} + \cdots + c_J\bar{Y}_{J\,\mathrm{adj}}}{\sqrt{\mathrm{MS\ multiple\ Res_w}\left[\dfrac{c_1^2}{n_1} + \dfrac{c_2^2}{n_2} + \cdots + \dfrac{c_J^2}{n_J}\right][1 + \mathrm{tr}(\mathbf{W}_\mathrm{x}^{-1}\mathbf{S}_{\mathrm{b_x}})]}} = F'$	$F' = \sqrt{(J-1)F_{\alpha,\,J-1,\,N-J-C}}$

Table 10.3 Formulas for Multiple Comparisons with Multiple Covariates in Nonrandomized Studies

Procedure	Formula	Critical Value
Fisher–Hayter	$\dfrac{\bar{Y}_{i\,\text{adj}} - \bar{Y}_{j\,\text{adj}}}{\sqrt{\dfrac{\text{MS multiple Res}_\text{w}\left[\dfrac{1}{n_i} + \dfrac{1}{n_j} + \mathbf{d}'\mathbf{W}_\text{x}^{-1}\mathbf{d}\right]}{2}}} = q$	Studentized range $q_{J-1,\,N-J-C}$
Tukey–Kramer	$\dfrac{\bar{Y}_{i\,\text{adj}} - \bar{Y}_{j\,\text{adj}}}{\sqrt{\dfrac{\text{MS multiple Res}_\text{w}\left[\dfrac{1}{n_i} + \dfrac{1}{n_j} + \mathbf{d}'\mathbf{W}_\text{x}^{-1}\mathbf{d}\right]}{2}}} = q$	Studentized range $q_{J,\,N-J-\text{C}}$
Bonferroni	$\dfrac{c_1\bar{Y}_{1\,\text{adj}} + c_2\bar{Y}_{2\,\text{adj}} + \cdots + c_J\bar{Y}_{J\,\text{adj}}}{\sqrt{\text{MS multiple Res}_\text{w}\left[\dfrac{c_1^2}{n_1} + \dfrac{c_2^2}{n_2} + \cdots + \dfrac{c_J^2}{n_J} + \mathbf{d}'\mathbf{W}_\text{x}^{-1}\mathbf{d}\right]}} = t$	Bonferroni $t_{B,C',\,N-J-C}$
Scheffé	$\dfrac{c_1\bar{Y}_{1\,\text{adj}} + c_2\bar{Y}_{2\,\text{adj}} + \cdots + c_J\bar{Y}_{J\,\text{adj}}}{\sqrt{\text{MS multiple Res}_\text{w}\left[\dfrac{c_1^2}{n_1} + \dfrac{c_2^2}{n_2} + \cdots + \dfrac{c_J^2}{n_J} + \mathbf{d}'\mathbf{W}_\text{x}^{-1}\mathbf{d}\right]}} = F'$	$F' = \sqrt{(J-1)F_{\alpha,J-1,N-J-C}}$

Using contrast coefficients the vector **d** is

	Group				
	1	2	$\cdots$	J	**d**
Covariate 1	$c_1\bar{X}_{1,1}$	$c_2\bar{X}_{1,2}$	$+\cdots+$	$c_J\bar{X}_{1,J}$	d_1
Covariate 2	$c_1\bar{X}_{2,1}$	$c_2\bar{X}_{2,2}$	$+\cdots+$	$c_J\bar{X}_{2,J}$	d_2
$\vdots$ $\vdots$	$\vdots$	$\vdots$	$\cdots$	$\vdots$	$\vdots$
Covariate C	$c_1\bar{X}_{C,1}$	$c_2\bar{X}_{C,2}$	$+\cdots+$	$c_J\bar{X}_{C,J}$	d_C

The transpose of **d** is the row vector $\mathbf{d}'$.

Example 10.1: Pairwise Comparisons Using Fisher–Hayter Multiple covariate analysis on the data presented in Table 10.1 yields the following adjusted means, covariate means and error mean square:

$$\bar{Y}_{1\text{ adj}} = 28.98 \quad \bar{X}_{1,1} = 52.00 \quad \bar{X}_{2,1} = 5.1$$
$$\bar{Y}_{2\text{ adj}} = 40.21 \quad \bar{X}_{1,2} = 47.00 \quad \bar{X}_{2,2} = 4.8$$
$$\bar{Y}_{2\text{ adj}} = 35.81 \quad \bar{X}_{1,3} = 49.00 \quad \bar{X}_{2,3} = 5.1$$

$$\text{MS}_{\text{resw}} = 38.90$$

Because the experiment was based on a randomized design the differences among covariate means are small for both covariates. The formula for the Fisher–Hayter test found in Table 10.2 is appropriate for all pairwise contrasts. This formula requires the computation of $\text{tr}(\mathbf{W}_x^{-1}\mathbf{S}_{b_x}$; this computation involves the following steps:

1. Compute $\mathbf{W}_x$.

$$\mathbf{W}_x = \begin{bmatrix} 5700.0 & 591.0 \\ 591.0 & 149.4 \end{bmatrix}$$

2. Compute the inverse of $\mathbf{W}_x$.

$$\mathbf{W}_x^{-1} = \begin{bmatrix} 0.000297 & -0.001177 \\ -0.001177 & 0.011348 \end{bmatrix}$$

3. Compute the corrected between-group sum of products matrix $\mathbf{B}_x$.

$$\mathbf{B}_x = \text{T}_x - \text{W}_x = \begin{bmatrix} 5826.67 & 598.00 \\ 598.00 & 150 \end{bmatrix} - \begin{bmatrix} 5700.00 & 591.00 \\ 591.00 & 149.40 \end{bmatrix}$$
$$= \begin{bmatrix} 126.67 & 7.00 \\ 7.00 & 0.60 \end{bmatrix}$$

4. Multiply $\mathbf{B}_x$ by the scalar $1/(J-1)$ to obtain the between-group variance–covarince matrix $\mathbf{S}_{b_x}$.

$$\mathbf{S}_{b_x} = \frac{1}{2}\begin{bmatrix} 126.67 & 7.00 \\ 7.00 & 0.60 \end{bmatrix} = \begin{bmatrix} 63.335 & 3.50 \\ 3.50 & 0.30 \end{bmatrix}$$

5. Compute the product $\mathbf{W}_x^{-1}\mathbf{S}_{b_x}$.

$$\mathbf{W}_x^{-1}\mathbf{S}_{b_x} = \begin{bmatrix} 0.000297 & -0.001177 \\ -0.0011777 & 0.011348 \end{bmatrix}\begin{bmatrix} 63.335 & 3.50 \\ 3.50 & 0.30 \end{bmatrix}$$
$$= \begin{bmatrix} 0.014691 & 0.000686 \\ -0.034827 & -0.000715 \end{bmatrix}$$

6. Compute the trace $\mathbf{W}_x^{-1}\mathbf{S}_{b_x}$.

$$\text{tr}\left(\mathbf{W}_x^{-1}\mathbf{S}_{b_x}\right) = \begin{bmatrix} 0.014691 & 0.000686 \\ -0.034827 & -0.000715 \end{bmatrix} = 0.0139759$$

The error term for each of the differences between adjusted means is

$$\sqrt{\frac{\text{MS multiple Res}_w\left[1 + \text{tr}\left(\mathbf{W}_x^{-1}\mathbf{S}_{b_x}\right)\right]}{n_h}} = \sqrt{\frac{38.90[1 + .0139759]}{10}} = 1.986$$

All pairwise contrasts are divided by this term to obtain q.

Group	Adjusted Mean Difference	Obtained q
1 vs. 2	$28.98 - 40.21 = -11.23$	-5.65
1 vs. 3	$28.98 - 35.81 = -6.83$	-3.44
2 vs. 3	$40.21 - 35.81 = 4.40$	2.22

The absolute value of each q_{obt} is evaluated by comparing it with the critical value of $q_{J-1,\,N-J-C}$ for a specified value of α. If we use $\alpha = .05$ the critical $q_{3-1,\,30-3-2)} = q_{2,25} = 2.913$. Two of the three pairwise contrasts are statistically significant.

If the study had not involved randomized groups, the differences among means on the covariates could have been large. In this case the Fisher–Hayter formula in Table 10.3 would be appropriate. An equivalent formula is used in *Minitab* and also in

other software packages that compute multiple comparisons for multiple ANCOVA. The formula is

$$\frac{\bar{Y}_{i\,\text{adj}} - \bar{Y}_{j\,\text{adj}}}{\sqrt{\dfrac{\text{MS multiple Res}_\text{w}\left[\dfrac{1}{n_i} + \dfrac{1}{n_j} + [\mathbf{d}'\mathbf{W}_\text{x}^{-1}\mathbf{d}]\right]}{2}}} = q$$

The adjusted means, the error term MS multiple res_w, and $\mathbf{W}^{-1}$ have already been computed. The column vector $\mathbf{d}$ is simply the vector of differences between the ith and jth group means on the covariate. Because there are three groups and three pairwise contrasts in this example, the $\mathbf{d}$ vectors associated with the three contrasts are

Group	Covariate Mean Differences	$= \mathbf{d}$
1 vs. 2	$\bar{X}_{1,1} - \bar{X}_{1,2} = 52 - 47$	$= \begin{bmatrix} 5.0 \\ 0.3 \end{bmatrix}$
	$\bar{X}_{2,1} - \bar{X}_{2,2} = 5.1 - 4.8$	
1 vs. 3	$\bar{X}_{1,1} - \bar{X}_{1,3} = 52 - 49$	$= \begin{bmatrix} 3.0 \\ 0.0 \end{bmatrix}$
	$\bar{X}_{2,1} - \bar{X}_{2,3} = 5.1 - 5.1$	
2 vs. 3	$\bar{X}_{1,2} - \bar{X}_{1,3} = 47 - 49$	$= \begin{bmatrix} -2.0 \\ -0.3 \end{bmatrix}$
	$\bar{X}_{2,2} - \bar{X}_{2,3} = 4.8 - 5.1$	

If we carry out the computation for the contrast of treatments one and two we find

$$\frac{-11.23}{\sqrt{\dfrac{38.90\left[\dfrac{1}{10} + \dfrac{1}{10} + [.004915]\right]}{2}}} = -5.625 = q.$$

This value differs only slightly from the value obtained with the formula in Table 10.2 for randomized designs. The critical value of q is, as before, 2.913. The qs for the other pairwise contrasts are also essentially the same as with the randomized-group formula. The similarity of the results using the two formulas is expected when the differences on the covariates are small.

Simultaneous Confidence Intervals

The radical terms associated with the Tukey–Kramer, Bonferroni, and Scheffé formulas in Tables 10.2 and 10.3 can be employed to construct simultaneous confidence intervals. A specified adjusted mean difference plus and minus the product of the radical term times the associated critical value yields the confidence interval.

Example of Simultaneous Confidence Intervals

The Tukey–Kramer 95% percent simultaneous confidence intervals require the same matrix manipulations shown for the Fisher–Hayter tests. The interval for the contrast of the *i*th and *j*th adjusted means is

$$(\bar{Y}_{i\,\text{adj}} - \bar{Y}_{j\,\text{adj}}) \mp \left[\sqrt{\frac{\text{MS multiple Res}_\text{w}\left[1 + \text{tr}(\mathbf{W}_\text{x}^{-1}\mathbf{S}_{\text{b}_\text{x}})\right]}{n_\text{h}}}\right] q_{J-1,N-J-C}$$

for the randomized-groups experiment, and

$$(\bar{Y}_{i\,\text{adj}} - \bar{Y}_{j\,\text{adj}}) \mp \left[\sqrt{\frac{\text{MS multiple Res}_\text{w}\left[\frac{1}{n_i} + \frac{1}{n_j} + \mathbf{d}'\mathbf{W}_\text{x}^{-1}\mathbf{d}\right]}{n_\text{h}}}\right] q_{J-1,N-J-C}$$

for any design. The output from Minitab (which uses the latter approach) is shown below. Note the signs; for the first comparison the routine subtracts the mean for group 1 from the mean for group 2 rather than the other way around.

The 95% simultaneous confidence intervals for the three example pairwise contrasts are

Groups	Adjusted Mean Difference	95% Simultaneous CIs
1 vs. 2	11.23	(4.21, 18.26)
1 vs. 3	6.83	(−.16, 13.82)
2 vs. 3	−4.40	(−11.36, 2.55)

10.6 SOFTWARE: MULTIPLE ANCOVA AND ASSOCIATED TUKEY–KRAMER MULTIPLE COMPARISON TESTS USING *MINITAB*

Input

The data are entered in the *Minitab* worksheet as follows:

```
Row   tx   x1   x2    Y
  1    1   29    3   15
  2    1   49    3   19
  3    1   48    2   21
  4    1   35    5   27
  5    1   53    5   35
  6    1   47    9   39
  7    1   46    3   23
  8    1   74    7   38
  9    1   72    6   33
 10    1   67    8   50
 11    2   22    3   20
```

```
 12   2  24   2  34
 13   2  49   4  28
 14   2  46   4  35
 15   2  52   5  42
 16   2  43   4  44
 17   2  64   8  46
 18   2  61   7  47
 19   2  55   6  40
 20   2  54   5  54
 21   3  33   2  14
 22   3  45   1  20
 23   3  35   5  30
 24   3  39   4  32
 25   3  36   3  34
 26   3  48   8  42
 27   3  63   8  40
 28   3  57   4  38
 29   3  56   9  54
 30   3  78   7  56
```

The menu input commands to compute multiple ANCOVA and Tukey–Kramer tests and simultaneous confidence intervals are

> Stat → ANOVA → General Linear Model → Responses: *Y* →
> → Model: *Tx* → Comparisons → PairwiseComparisons →
> Terms: *Tx* → Method: Tukey → Confidence Interval → Test →
> OK → OK

```
The corresponding command line editor commands are
MTB > GLM 'Y' = tx;
SUBC>  Covariates 'x1' 'x2';
SUBC>  Brief 2;
SUBC>  Pairwise tx;
SUBC>   Tukey;
SUBC>   NoGrouping.
```

Output:

```
General Linear Model: Y versus tx
Factor  Type   Levels  Values
tx      fixed       3  1, 2, 3
Analysis of Variance for Y, using Adjusted SS for Tests

Source  DF   Seq SS  Adj SS  Adj MS    F        P
x1       1  1567.36  257.21  257.21   6.61  0.016
x2       1   791.46  708.21  708.21  18.21  0.000
tx       2   624.74  624.74  312.37   8.03  0.002
Error   25   972.44  972.44   38.90
Total   29  3956.00
```

```
S = 6.23678   R-Sq = 75.42%   R-Sq(adj) = 71.49%

Means for Covariates

Covariate    Mean   StDev
x1         49.333  14.175
x2          5.000   2.274
```

Least Squares Means for Y

```
tx   Mean  SE Mean
1   28.98    1.988
2   40.21    1.982
3   35.81    1.974
```

Tukey 95.0% Simultaneous Confidence Intervals

```
Response Variable Y
All Pairwise Comparisons among Levels of tx
tx = 1  subtracted from:

tx   Lower  Center  Upper  -+---------+---------+---------+-----
2   4.2061  11.233  18.26                (-----*-----)
3  -0.1588   6.830  13.82           (-----*-----)
                           -+---------+---------+---------+-----
                         -10         0         10        20

tx = 2  subtracted from:
tx  Lower  Center  Upper  -+---------+---------+---------+-----
3  -11.36  -4.404  2.552  (-----*-----)
                           -+---------+---------+---------+-----
                         -10         0         10        20
```

Tukey Simultaneous Tests

```
Response Variable Y
All Pairwise Comparisons among Levels of tx
tx = 1 subtracted from:

    Difference       SE of           Adjusted
tx    of Means  Difference  T-Value   P-Value
2       11.233       2.823    3.979    0.0015
3        6.830       2.808    2.432    0.0566
tx = 2  subtracted from:
```

```
    Difference        SE of            Adjusted
tx    of Means  Difference  T-Value    P-Value
3       -4.404       2.795   -1.576     0.2745
```

10.7 SUMMARY

Multiple covariance analysis is a straightforward extension of ANCOVA with one covariate. Several covariates can be expected to increase the power of an experiment more than a single covariate if the covariates are well chosen. Also, the use of more than one covariate will generally remove more bias when means are compared. Still, even the use of many covariates will not generally *eliminate* bias in the comparison of groups from different populations.

The computation required for multiple ANCOVA is an extension of the general linear regression approach employed with simple ANCOVA. The dependent variable is first regressed on all group membership dummy variables and all covariates. Next, the dependent variable is regressed on all covariates. The difference between the R^2-values associated with these two regressions represents the proportion of the total variation on Y that is accounted for by group differences that are independent of the covariates. The ANCOVA F-test evaluates the difference between means that have been adjusted by all covariates. The multiple comparison tests covered in Chapter 6 must be modified to accommodate multiple covariates. Results based on these tests are conditional on all covariates included in the model.

PART IV

Alternatives for Assumption Departures

CHAPTER 11

Johnson–Neyman and Picked-Points Solutions for Heterogeneous Regression

11.1 INTRODUCTION

Recall that testing for homogeneous population regression slopes is a routine aspect of evaluating the adequacy of the ANCOVA model. When the slopes are heterogeneous an alternative to ANCOVA should be considered. I describe two related approaches to appropriately analyze the heterogeneous regression case. The first one is known as the Johnson–Neyman (J–N) technique (Johnson and Neyman, 1936). I label the second approach as picked-points analysis (PPA). The more popular term coined by Rogosa (1980) is "pick-a-point." (Because the recommended implementation of the procedure actually involves multiple points rather than a single point, picked-points analysis may be somewhat more descriptive.) Before describing these methods I briefly review some of the problems introduced by heterogeneous slopes.

When heterogeneous regression slopes are present this implies that the magnitude of the treatment effect is not the same at different levels of X. Three varieties of this problem are illustrated in Figure 11.1. Panel A illustrates a situation in which the adjusted means are equal even though group 1 is superior to group 2 at low values of X and inferior to group 2 at high levels of X. If the data are not visually examined or if the homogeneity of regression slopes test is not carried out to identify the problem, ANCOVA will lead to the conclusion that there is not a treatment effect because the adjusted means are equal. Panel B illustrates a case in which the adjusted means are different, but it appears that there are no treatment effects at low levels of X and large effects at high levels of X. Panel C illustrates the most common case where adjusted means are different and one group is consistently superior to the other group regardless of the level of X. The only descriptive problem with the adjusted means

The Analysis of Covariance and Alternatives: Statistical Methods for Experiments, Quasi-Experiments, and Single-Case Studies, Second Edition. Bradley E. Huitema.

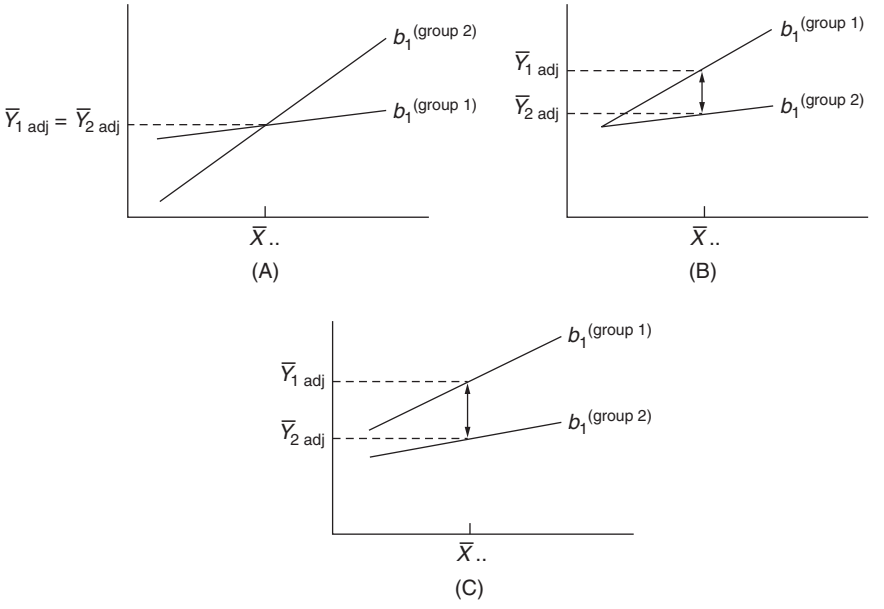


Figure 11.1 Three forms of heterogeneous regression slopes.

in this case is that the adjusted difference does not reveal the changing degree to which one group is superior to the other as a function of X. The difference between adjusted means does not adequately describe the outcome of the experiment because it overestimates the effects at some levels of X and underestimates the effects at other levels of X. This problem as well as other problems of inference and interpretation can be solved using J–N and PPA solutions.

Although the solutions described in this chapter are not well known, there are some common solutions to a similar problem in conventional two-factor analysis of variance. Hence, it is useful to revive memories of approaches used in the analysis of two-factor designs in order to track the general ideas associated with analyzing studies with heterogeneous regression slopes. The homogeneity of regression slopes test in the one-factor ANCOVA is analogous to the $A \times B$ interaction test in a two-factor ANOVA. Likewise, PPA tests applied to one-factor ANCOVA designs with heterogeneous slopes are analogous to simple main effects tests in two-factor ANOVA designs with interaction.

Recall that if interaction is present in a two-factor ANOVA design there is interest in identifying the levels of factor B at which simple effects of factor A are present (and/or levels of A at which B effects are present). That is, a separate test for simple effects of factor A is performed at each level of factor B. Similarly, the PPA approach provides a separate test for treatment effects at each X *point picked* by the researcher. Simple effects tests on the effects of factor A in a two-factor design are usually computed at each level of factor B; the number of levels of factor B is usually very small and no thought goes into deciding the number of tests to compute because it is generally dictated by the number of levels of factor B in the design. But, in the

case of one-factor ANCOVA, the covariate X is usually a continuous variable and the researcher must pick the points on X that are of interest. Three points (a very low point, the mean, and a very high point) are often sufficient to provide a useful description of the outcome.

The J–N approach differs from the PPA approach in that it does not require points on X to be picked in order to carry out the analysis. This is because the purpose of the J–N approach is to provide *regions* of nonsignificance and significance instead of a separate test for the difference on Y at individual *points picked* on X. The J–N and PPA computation procedures are presented in Section 11.2 for the frequently encountered case of two independent groups and one covariate. Corresponding procedures for more than two groups, multiple covariates, correlated samples, and two-factor designs are described in subsequent sections. Recently developed robust versions of these methods (Watcharotone et al., 2010) are briefly described.

11.2 J–N AND PPA METHODS FOR TWO GROUPS, ONE COVARIATE

Johnson–Neyman Technique

Suppose that the data listed in Table 11.1 and displayed in Figure 11.2 are based on an experiment in which two methods of therapy are the treatments and scores on a sociability scale are employed as the covariate. The dependent variable is an aggressiveness score based on a behavioral checklist.

Suppose the ANCOVA and homogeneity of regression slopes tests have been computed using the previously described regression approach (Chapter 7) based on the predictors included in Table 11.2 or by using a dedicated ANCOVA routine.

Table 11.1 Data from a Two-Group One-Covariate Experiment

Therapy 1		Therapy 2	
X	Y	X	Y
1	10	1	5
2	10	1.5	6
2	11	2.5	6
3	10	3.5	7
4	11	4.5	8
5	11	4.5	9
5	10	5	9
6	11	6	9
6	11.5	6	10.5
7	12	7	11
8	12	7	12.5
8	11	7.5	12.5
9	11	8	14
10	12.5	9	14.5
11	12	10	16

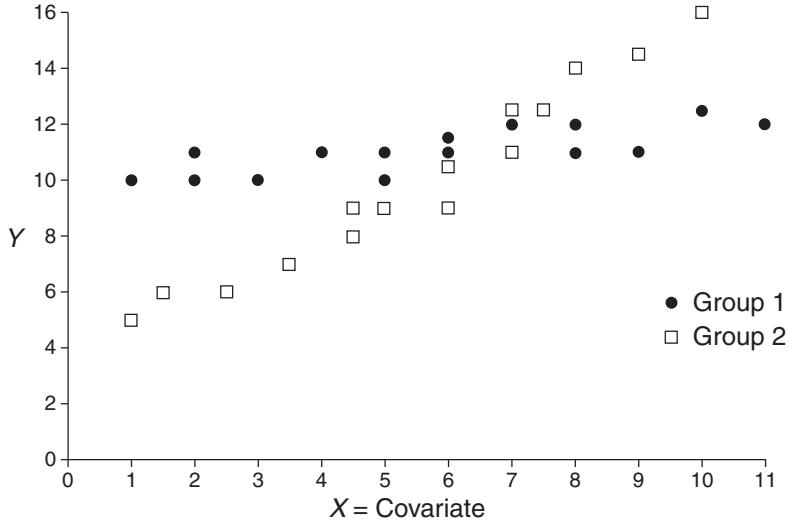


Figure 11.2 Plot of aggressiveness scores (Y) on sociability scores (X) for two treatment groups.

The ANCOVA $F_{\text{obt}} = 2.28$, which does not exceed the critical value for a 5% test. At first glance it appears that the investigator may reasonably conclude that statistically significant treatment effects are not present. But this conclusion should be questioned if the ANCOVA assumption of homogeneous slopes appears not to be approximately met.

The test on the assumption of homogeneous regression slopes yields $F_{\text{obt}} = 141.89$, which, when compared with the critical $F_{(.05,\,1,\,26)} = 4.23$, is obviously significant ($p < .001$). A plot of the data with the individual slopes superimposed is shown in Figure 11.3 (ignore the vertical lines for now). It clearly reveals why the homogeneity test yields such a large obtained F. The values of the sample slopes for groups 1 and 2 are .209 and 1.237, respectively.

This figure confirms that the ANCOVA model is quite inappropriate because the test on the difference between adjusted means does not attend to the obvious treatment differences at high and low levels of X. Two relevant questions at this stage are (1) "What are the values on X associated with nonsignificant treatment effects?" and (2) "What are the values on X associated with significant treatment effects?" These questions are answered using the J–N technique; it is computed below.

Nonsignificance Region

The limits of the region of nonsignificance on X are computed by using

$$X_{L1} = \frac{-B - \sqrt{B^2 - AC}}{A} \quad \text{and}$$

$$X_{L2} = \frac{-B + \sqrt{B^2 - AC}}{A}$$

Table 11.2 Required Columns for ANCOVA Through Regression Analysis (Data from Table 11.1)

Subject	*Y*	*X*	*D*	*DX*
1	10	1	1	1
2	10	2	1	2
3	11	2	1	2
4	10	3	1	3
5	11	4	1	4
6	11	5	1	5
7	10	5	1	5
8	11	6	1	6
9	11.5	6	1	6
10	12	7	1	7
11	12	8	1	8
12	11	8	1	8
13	11	9	1	9
14	12.5	10	1	10
15	12	11	1	11
16	5	1	0	0
17	6	1.5	0	0
18	6	2.5	0	0
19	7	3.5	0	0
20	8	4.5	0	0
21	9	4.5	0	0
22	9	5	0	0
23	9	6	0	0
24	10.5	6	0	0
25	11	7	0	0
26	12.5	7	0	0
27	12.5	7.5	0	0
28	14	8	0	0
29	14.5	9	0	0
30	16	10	0	0

where

X_{L1} and X_{L2} = limits of nonsignificance region,

$$A = \frac{-F_{(\alpha,1,N-4)}}{n-4}(\text{SSres}_\text{i})\left(\frac{1}{\sum x_1^2}+\frac{1}{\sum x_2^2}\right)+\left(b_1^{(\text{group 1})}-b_1^{(\text{group 2})}\right)^2,$$

$$B = \frac{-F_{(\alpha,1,N-4)}}{n-4}(\text{SSres}_\text{i})\left(\frac{\bar{X}_1}{\sum x_1^2}+\frac{\bar{X}_2}{\sum x_2^2}\right) + \left(b_0^{(\text{group 1})}-b_0^{(\text{group 2})}\right)\left(b_1^{(\text{group 1})}-b_1^{(\text{group 2})}\right),$$

$$C = \frac{-F_{(\alpha,1,N-4)}}{n-4}(\text{SSres}_\text{i})\left(\frac{N}{n_1 n_2}+\frac{\bar{X}_1^2}{\sum x_1^2}+\frac{\bar{X}_2^2}{\sum x_2^2}\right)+\left(b_0^{(\text{group 1})}-b_0^{(\text{group 2})}\right)^2,$$

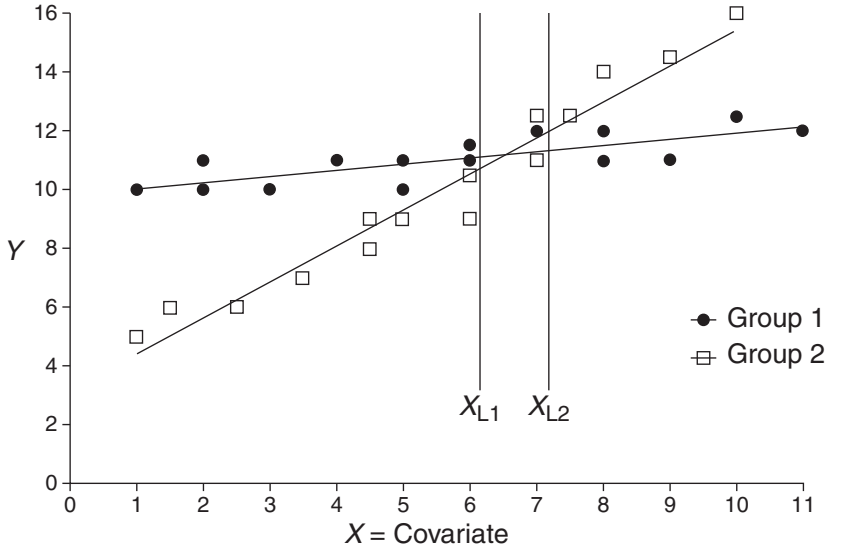


Figure 11.3 Plot of individual regression lines and limits for the J–N region of nonsignificance.

where

$F_{(\alpha,\,1,\,N-4)}$ is the critical value of F statistic for desired level of α and 1 and $N - 4$ degrees of freedom (where N is the total number of subjects; i.e., $n_1 + n_2$);
SSres_i is the individual residual sum of squares (described in Chapter 6);
$\bar{X}_1$ and $\bar{X}_2$ are covariate means for samples 1 and 2, respectively;
$\sum x_1^2$ and $\sum x_2^2$ are covariate sums of squares for samples 1 and 2, respectively;
$b_0^{(\text{group 1})}$ is the regression intercept for sample 1;
$b_0^{(\text{group 2})}$ is the regression intercept for sample 2;
$b_1^{(\text{group 1})}$ is the regression slope for sample 1; and
$b_1^{(\text{group 2})}$ is the regression slope for sample 2.

The summary statistics and critical value of F for the example data are

Group 1	Group 2
$n_1 = 15$	$n_2 = 15$
$\bar{X}_1 = 5.8$	$\bar{X}_2 = 5.533$
$\sum x_1^2 = 130.4$	$\sum x_2^2 = 99.23$
$b_0^{(\text{group 1})} = 9.857$	$b_0^{(\text{group 2})} = 3.155$
$b_1^{(\text{group 1})} = 0.209$	$b_1^{(\text{group 2})} = 1.237$
Critical $F_{(.05,\,1,\,26)} = 4.23$	

The values for A, B, C and the limits on the region are

$$\bar{X}_2 = 5.533,$$
$$A = 1.0253304$$
$$B = -6.7155607$$
$$C = 43.71518698$$
$$X_{L1} = \frac{6.7155607 - 0.525590}{1.0253304} = 6.04$$
$$X_{L2} = \frac{6.7155607 + 0.525590}{1.0253304} = 7.06$$

Hence, the region of nonsignificance is 6.04 through 7.06 using $\alpha = .05$. If we select a specific point on the sociability scale X that falls in this region we conclude that we have insufficient evidence to claim a treatment effect on Y. If instead we select a point below 6.04, we have strong evidence to claim that aggression is higher under treatment 1. If we select a point above 7.06, we conclude that we have strong evidence to claim that treatment 2 aggressiveness scores are higher.

Picked-Points Analysis

The PPA approach requires the researcher to pick points on the X dimension that are of interest. I recommend selecting the mean and two additional points. The additional points of most interest will usually approach the extremes of the observed X distribution. Consider the sociability scores in Figure 11.2. Note that data are available from both groups for the value 1.0, which is near the lower end of the observed X distribution; similarly, near the upper end of the distribution there are data from both groups at the value 10.0. It would be reasonable to perform PPA using 1.0, the mean (equal to 5.665), and 10.0 as the three points picked for analysis. A reasonable general approach is to pick values of X falling (1) one standard deviation below the mean, (2) at the mean, and (3) one standard deviation above the mean. In the case of the example data these points correspond to 2.85, 5.665, and 8.48. The standard deviation used in this case is based on the total (both samples combined) sample distribution of X. If X is highly skewed, I recommend using quartiles 1, 2, and 3.

Regardless of the method of selecting points, the overriding issue is that there must be data from both groups at or very near the chosen values. To pick points beyond the observed data is to engage in speculation unless convincing additional data outside the experiment demonstrate that the linear model holds beyond the observed range. For the example data, suppose we pick $X = 20$ as one of the points and use PPA to compare therapy effects at this point. Note in Figure 11.2 that there are no data near this value; extreme and unjustified extrapolation is involved in estimating effects in this case.

Two methods of performing PPA are included in this section. I first describe a practical approach that requires no knowledge of matrix algebra or notation, and no special software; it can be accomplished with a single regression analysis once the appropriate predictor variables are constructed. The test statistic is based on the *t*-distribution. The second approach, which provides an *F*-statistic, involves a matrix solution that is likely to be of interest only to the methodologically oriented reader. The reader interested only in the simplest method of computing an appropriate analysis should skip the subsection describing the second method.

Method I: Regression Approach

Method I involves the following steps:

1. Pick the first point of interest.
2. Once a point is picked, that point it is used to center the *X* variable. That is, the picked point is subtracted from each value in the *X* column; I suggest that the result be labeled as the "PP centered *X*" or simply "PPC*X*." (I hesitate to label this column simply as "centered *X*" because this term is almost universally interpreted as "deviation from the mean.")
3. After the PPC*X* variable is constructed it is multiplied by the 0–1 dummy variable *D* used to identify the two groups; the product is labeled D*PPC*X*. I recommend using "1" to identify subjects in the first group and "0" to identify subjects in the second group (especially if the second group is a control and the first group is treated).
4. A three-predictor multiple regression analysis is then carried out where *Y* is regressed on *D*, PPC*X*, and *D**PPC*X*. The coefficient for the *D* variable is the estimated treatment effect at the picked point. The *t*-value associated with this coefficient tests the statistical significance of the treatment effect estimate at the picked point. The estimated standard error for the coefficient can be used in the construction of a confidence interval for the effect at the point picked. Most recent regression software such as *Minitab* (version 16) and *SPSS* provide this confidence interval. This four-step procedure is repeated for each picked point; if three points are picked the whole routine is applied three times. Each analysis uses a different value to center the *X* variable. The complete PPA analysis consists of the whole set of results. In some studies, however, an additional aspect of the analysis may be useful; it is described next.

Test on Difference Between Effect Estimates

Occasionally, there may be interest in a formal test of the difference between the effect estimated at one point on *X* and the effect estimated at another point on *X*. A completely inadequate approach is to simply compare the decisions associated with the different tests. That is, if the treatment effect at one point on *X* is declared significant but the treatment effect at another point on *X* is declared nonsignificant,

this is not convincing evidence of differential effects. Instead, I recommend the following formal test statistic to evaluate the difference between effects:

$$\frac{b_{Di} - b_{Dj}}{\sqrt{s_{Di}^2 + s_{Dj}^2 - 2s_{Di\,Dj}}} = t,$$

where

b_{Di} and b_{Dj} are the treatment effect estimates at X points i and j;
s_{Di}^2 and s_{Dj}^2 are the error variance estimates for coefficients b_{Di} and b_{Dj}; and
s_{DiDj} is the covariance between the two treatment effect estimates.

The covariance can be estimated using $[(s_{Di})(s_{Dj})(r)]$, where r is the correlation between (1) the residuals from regressing the dummy variable D on the PPCX and D*PPCX variables used in estimating the effect at X point i and (2) the residuals from the regression of D on the PPCX and D*PPCX variables used to estimate the effect at X point j. The critical value of t is based on $N - 4$ degrees of freedom. This test can be viewed as an analog to an interaction contrast test in multiple factor ANOVA.

Computational Examples of PPA: Method I

Three computational examples of PPA are included in this section. Examples 11.1 and 11.2 are based on the same data used above to describe the J–N technique. Example 11.3 is based on data from a small experimental study that reveals a rather different but much more common type of heterogeneity than is shown in Figure 11.2.

Example 11.1 PPA for a Single Picked Point Return to the data displayed in Figure 11.2 and once again suppose there is interest in estimating the treatment effect for subjects who have a sociability score of 8. That is, the picked point on X is equal to 8 (which corresponds to Q3). Enter the Y, D, and X columns shown in Table 11.2 in the *Minitab* worksheet as shown below. Once the data are entered (where column 1 = Y, column 2 = D, and column 3 = X) two additional columns need to be constructed: a column that is centered at 8 and the product column of D times the centered variable. In this case "PPCX" = $X - 8$. The *Minitab* commands for constructing the PPCX and D*PPCX columns are

```
MTB > let  c4 = X-8
MTB > let  c5 = c2*c4
```

The following columns are now included in the worksheet:

Y	D	X	X-8	D*X-8
10	1	1	-7	-7
10	1	2	-6	-6
11	1	2	-6	-6
10	1	3	-5	-5

```
11      1       4      -4       -4
11      1       5      -3       -3
10      1       5      -3       -3
11      1       6      -2       -2
11.5    1       6      -2       -2
12      1       7      -1       -1
12      1       8       0        0
11      1       8       0        0
11      1       9       1        1
12.5    1      10       2        2
12      1      11       3        3
 5      0       1      -7        0
 6      0       1.5    -6.5      0
 6      0       2.5    -5.5      0
 7      0       3.5    -4.5      0
 8      0       4.5    -3.5      0
 9      0       4.5    -3.5      0
 9      0       5      -3        0
 9      0       6      -2        0
10.5    0       6      -2        0
11      0       7      -1        0
12.5    0       7      -1        0
12.5    0       7.5    -0.5      0
14      0       8       0        0
14.5    0       9       1        0
16      0      10       2        0
```

The required regression analysis is then computed using the following commands:

```
MTB > Regress 'Y' 3  'D' 'X-8' 'D*X-8';
SUBC> GFourpack;
SUBC> RType 1;
SUBC> Constant;
SUBC> Brief 2.
```

Output:

```
Regression Analysis: Y versus D, X-8, D*X-8

The regression equation is Y = 13.1 - 1.53 D + 1.24 X-8 -
1.03 D*X-8

Predictor      Coef          SE Coef          T         P
Constant       13.0512       0.2318        56.29     0.000
D              -1.5257       0.3120        -4.89     0.000
X-8             1.23698      0.06506       19.01     0.000
D*X-8          -1.02839      0.08633      -11.91     0.000
```

Note that the value of the coefficient associated with the indicator variable D is -1.5257 and the related p-value is $<.001$. The coefficient indicates that subjects with sociability scores of 8 have an estimated aggressiveness score under treatment 1 that is about one and a half points lower than it is under treatment 2. This descriptive result is consistent with visual inspection of the data in Figure 11.3. Also, the inferential result agrees with the J–N upper region of significance because significant treatment effects are associated with values of X that exceed 7.06.

The estimate of the effect can be interpreted as the difference between the predicted value on Y for subjects in the first group whose sociability scores equal 8 and the predicted value on Y for subjects in the second group whose sociability scores equal 8. The fitted regression equation (i.e., $\hat{Y} = 13.051 - 1.526(D) + 1.237(X - 8) - 1.028(D^*[X - 8])$) is the basis for both predictions. The application of this equation to the data for each group yields the following predicted values:

$$\text{Group 1: } \hat{Y} = 13.051 - 1.526(1) + 1.237(0) - 1.028(D^*[0]) = 11.525, \quad \text{and}$$
$$\text{Group 2: } \hat{Y} = 13.051 - 1.526(0) + 1.237(0) - 1.028(D^*[0]) = 13.051.$$

The difference between these predicted values is $(11.525 - 13.051) = -1.53$, which is the value of the coefficient associated with the indicator variable D.

Example 11.2: Three Picked Points The same initial data used above are analyzed here using three picked points. The three points are 1, the mean $\bar{X}..$ (5.6665), and 10. Hence, the three required PPCX variables are defined as $X - 1$, $X - 5.6665$, and $X - 10$. Next, the product of each centered column times D is constructed. Then three regression analyses are carried out. First, Y is regressed on D, $X - 1$, and $D^*(X - 1)$; second, Y is regressed on D, $X - 5.6665$, and $D^*(X - 5.6665)$; and last, Y is regressed on D, $X - 10$, and $D^*(X - 10)$. The crucial aspects of these analyses are shown below.

```
Regression Analysis: Y versus D, CenteredPP1, D*CenteredPP1

The regression equation is

Y = 4.39 + 5.67 D + 1.24 CenteredPP1 - 1.03 D*CenteredPP1

Predictor              Coef      SE Coef           T         P
Constant             4.3923       0.3391       12.95     0.000
D                    5.6731       0.4660       12.17     0.000
Centered PP1         1.23698      0.06506      19.01     0.000
D*Centered PP1      -1.02839      0.08633     -11.91     0.000


Regression Analysis: Y versus D, CenteredMean, D*CenteredMean
```

```
The regression equation is
Y = 10.2 + 0.874 D + 1.24 CenteredMean - 1.03 D*CenteredMean

Predictor              Coef      SE Coef           T         P
Constant            10.1647       0.1676       60.66     0.000
D                    0.8741       0.2369        3.69     0.001
CenteredMean         1.23698      0.06506      19.01     0.000
D*CenteredMean      -1.02839      0.08633     -11.91     0.000


Regression Analysis: Y versus D, CenteredPP10, D*CenteredPP10

The regression equation is

Y = 15.5 - 3.58 D + 1.24 CenteredPP10 - 1.03 D*CenteredPP10

Predictor              Coef      SE Coef           T         P
Constant            15.5252       0.3353       46.30     0.000
D                   -3.5825       0.4441       -8.07     0.000
CenteredPP10         1.23698      0.06506      19.01     0.000
D*CenteredPP10      -1.02839      0.08633     -11.91     0.000
```

The output for the *D* variable in each of the three analyses is summarized below:

```
Predictor      Coef        SE Coef            T         P
   D         5.6731         0.4660        12.17     0.000
   D         0.8741         0.2369         3.69     0.001
   D        -3.5825         0.4441        -8.07     0.000
```

Note that the dependent variable score estimate is higher for treatment 1 than for treatment 2 at X points 1 and $\bar{X}$.., but the reverse is true at point 10. Each of the three D estimates is statistically significant, but the sizes of the estimates are quite different. These discrepancies in the size of D are quite understandable if Figure 11.3 is scrutinized.

It is usually unnecessary to follow up with additional analysis of data such as these, but occasionally there may be strong interest in testing the difference between two values of D. Suppose the researcher wants to know if a strong argument can be made that there is a population difference between the size of the treatment effect at $X = 1$ and the size of the treatment effect at the mean on X. The application of the test on the difference between effect estimates (described earlier in this section) yields:

$$t = \frac{5.6731 - .8741}{\sqrt{(.4660)^2 + (.2369)^2 - 2(.0555)}} = 11.91.$$

This value far exceeds the critical value of t based on 26 degrees of freedom. It is clear that the difference between the two treatment effect estimates is not explainable on the basis of sampling error.

Example 11.3: Packard–Boardman Data Packard and Boardman (1988) present data from the area of ecological physiology that nicely demonstrate the differences among several relevant analyses and the advantage of fitting the best model. They performed a randomized-group experiment in which 16 snapping turtle eggs were randomly assigned to one of two environmental conditions ($n_1 = 8$ and $n_2 = 8$). The conditions were the level of water potential of the substrate on which the eggs rested. The dependent variable was the size of the turtle, measured as dry mass (g) of the carcass at hatching. The data are listed in Table 11.3.

The dry mass (Y) means for groups 1 and 2 are 1.480 and 1.413, respectively. A one-factor ANOVA testing the difference between these means is shown below.

```
One-way ANOVA: Dry mass versus Group

Source     DF          SS           MS        F          P
Group       1      0.0179       0.0179     0.88      0.364
Error      14      0.2843       0.0203
Total      15      0.3021
```

It can be seen that this analysis leads to the conclusion that there is no treatment effect because $p = .364$. The investigators were aware, however, that the size of hatchling turtles is usually positively related to the size of the eggs in which they develop. Their data, plotted in Figure 11.4, indicate that the relationship between these two variables is quite high.

This means that some of the within-group variation on dry mass (Y) is predictable from egg size. The implication of this is that within-group variation should not be conceptualized as only unexplainable random error (which is assumed under the ANOVA

Table 11.3 Turtle Data from Packard and Boardman (1988)

Turtle	Egg Mass	Dry Mass	Group
1	10.223	1.184	1
2	11.184	1.371	1
3	12.251	1.676	1
4	11.922	1.662	1
5	11.485	1.509	1
6	11.625	1.539	1
7	11.303	1.481	1
8	11.662	1.417	1
9	11.415	1.364	2
10	11.684	1.508	2
11	11.668	1.535	2
12	11.322	1.387	2
13	12.553	1.522	2
14	12.213	1.502	2
15	10.814	1.256	2
16	10.493	1.230	2

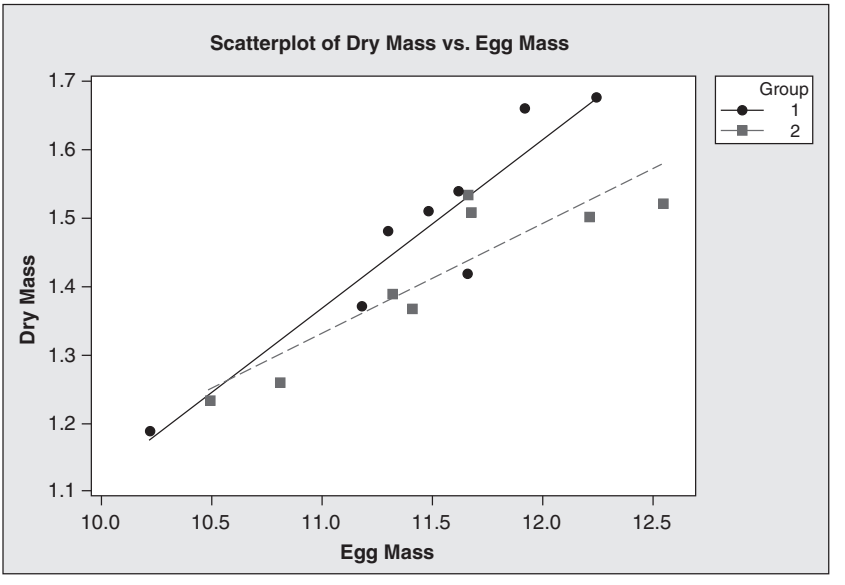


Figure 11.4 Packard–Boardman (1988) turtle data illustrating the regression of carcass dry mass on egg mass separately in two treatment groups.

model). Because measurements of egg size (g) were collected during the study, they are available for use as a covariate in an ANCOVA. The results of ANCOVA, using egg size as the covariate and dry mass as the dependent variable, are shown below.

```
ANCOVA: Y versus Group

Factor  Levels  Values
Group        2  1, 2

Analysis of Covariance for Y

Source          DF      Adj SS          MS       F           P
Covariates      1       0.23044     0.23044      55.66       0.000
Group           1       0.02526     0.02526       6.10       0.028
Error           13      0.05382     0.00414
Total           15      0.30214

                  Adjusted Means
Group       N           Y
1           8       1.4862
2           8       1.4066
```

The difference between the adjusted means (.080) analyzed in this ANCOVA is about the same size as the difference between the unadjusted means (.067) that was

tested using ANOVA, but the ANCOVA is statistically significant. Compare the error mean square in ANOVA with the error mean square in ANCOVA. Note that the latter is dramatically smaller (.004 rather than .020). This is the main reason the *p*-value is so much smaller for ANCOVA (.028) than for ANOVA (.364).

The next step in carrying out a complete ANCOVA is to test the assumption of homogeneous regression slopes. I recommend setting α at a very liberal level for this test (say, .20). The input and output for the test is shown below. Recall that the *Minitab* GLM output for this test appears in the line labeled "Group*X." Both the obtained *p*-value and the plot of the data in Figure 11.4 suggest that the slopes are not homogeneous.

```
MTB > glm c3 = c1 c2 c1*c2;
SUBC> covariate c2.

General Linear Model: Y versus Group

Factor  Type   Levels  Values
Group   fixed       2  1, 2

Analysis of Variance for Y, using Adjusted SS for Tests

Source     DF     Seq SS     Adj SS    Adj MS        F       P
Group       1   0.017889   0.009217  0.009217     2.58   0.134
X           1   0.230438   0.238991  0.238991    66.82   0.000
Group*X     1   0.010895   0.010895  0.010895     3.05   0.106
Error      12   0.042922   0.042922  0.003577
Total      15   0.302144
```

Because the slopes appear to be heterogeneous a method that accommodates this property should be considered. I carried out three PPA analyses after picking three points. The picked points were selected to equal one standard deviation below the mean, the mean, and one standard deviation above the mean; these correspond to 10.870, 11.489, and 12.108, respectively. The effect estimates and *p*-values associated with these points are .0256 ($p = .563$), .080 ($p = .020$), and .134 ($p = .009$). Note that the effect estimate from the ANCOVA (.080) is the same as the PPA estimate when the picked point is the mean on *X*, but the *p*-value for PPA (.020) is smaller than the ANCOVA *p*-value, which is .028. This is a typical finding; it occurs because heterogeneity of slopes introduces variability to the ANCOVA error sum of squares (i.e., $SSres_w$) that is not present in the error sum of squares associated with the PPA.

Method II: Matrix Approach

The general matrix approach described here applies whether there is one covariate or multiple covariates. After the method is described it is applied to the one-covariate aggression study analyzed earlier (using J–N) to show the correspondence of the two methods.

Formally, the null hypothesis associated with the PPA test can be written as

$$H_0 : \mathbf{x}_s' \boldsymbol{\Delta} = 0.0,$$

where

$\mathbf{x}_s$ is the $(C + 1) \times 1$ column vector of a specified set of covariate scores augmented by scalar 1 as first element;

$\boldsymbol{\Delta}$ is the difference vector $\boldsymbol{\beta}^{(\text{group 1})} - \boldsymbol{\beta}^{(\text{group 2})}$, where $\boldsymbol{\beta}^{(\text{group 1})}$ is the vector of least-squares regression parameters obtained by regressing Y on covariates 1 through C for population 1; in other words,

$$\boldsymbol{\beta}^{(\text{group 1})} = \begin{bmatrix} \beta_0^{(\text{group 1})} \\ \beta_1^{(\text{group 1})} \\ \vdots \\ \beta_C^{(\text{group 1})} \end{bmatrix},$$

and correspondingly,

$$\boldsymbol{\beta}^{(\text{group 2})} = \begin{bmatrix} \beta_0^{(\text{group 2})} \\ \beta_1^{(\text{group 2})} \\ \vdots \\ \beta_C^{(\text{group 2})} \end{bmatrix}.$$

The product $\mathbf{x}_s' \boldsymbol{\Delta}$ is the difference between the two population means on Y at point $\mathbf{x}_s$.

When analyzing sample data the point estimate of the effect parameter $\mathbf{x}_s' \boldsymbol{\Delta}$ is $\mathbf{x}_s' \mathbf{D}$. The F-ratio for testing the null hypothesis is

$$F = \frac{\mathbf{x}_s' \mathbf{D} \mathbf{D}' \mathbf{x}_s}{\mathbf{x}_s' \mathbf{V} \mathbf{x}_s},$$

where

$\mathbf{x}_s$ is a $(C + 1) \times 1$ column vector of a *s*pecified set of picked points augmented by scalar 1 as first element; that is,

$$\mathbf{x}_s = \begin{bmatrix} 1 \\ X_1 \\ X_2 \\ \vdots \\ X_C \end{bmatrix},$$

where

$X_1, X_2, \ldots X_C$ are picked points on covariates 1 through C,

D is the $(C+1) \times 1$ column vector of differences in least-squares estimates $\mathbf{b}^{(\text{group 1})}$ and $\mathbf{b}^{(\text{group 2})}$; that is,

$$\mathbf{D} = \begin{bmatrix} b_0^{(\text{group 2})} - b_0^{(\text{group 1})} \\ b_1^{(\text{group 2})} - b_1^{(\text{group 1})} \\ b_2^{(\text{group 2})} - b_2^{(\text{group 1})} \\ \vdots \qquad\qquad \vdots \\ b_C^{(\text{group 2})} - b_C^{(\text{group 1})} \end{bmatrix} = [\mathbf{b}^{(\text{group 2})} - \mathbf{b}^{(\text{group 1})}]$$

The variance of the difference in estimated regression lines, planes, or hyperplanes at $\mathbf{x}_s$ is

$$\mathrm{V} = \mathrm{MSres_i} \left\lfloor (\mathbf{X}_1'\mathbf{X}_1)^{-1} + (\mathbf{X}_2'\mathbf{X}_2)^{-1} \right\rfloor,$$

where

$\mathrm{MSres_i}$ is the error MS employed in the homogeneity of regression test described in Chapter 6; briefly,

$$\frac{\left(1 - R^2_{yD,X,DX}\right) \mathrm{SST}}{N - J(C+1)}$$

where

$\mathbf{X}_1$ is the $n \times (C+1)$ matrix of raw covariate scores for group 1 augmented by a column vector of ones as the first column of the matrix; that is,

$$\mathbf{X}_1 = \begin{bmatrix} 1 & X_{11} & X_{21} & \cdots & X_{C1} \\ 1 & X_{12} & X_{22} & \cdots & X_{C2} \\ \vdots & \vdots & \vdots & \vdots\vdots\vdots & \vdots \\ 1 & X_{1n_1} & X_{2n_1} & \cdots & X_{Cn_1} \end{bmatrix}$$

where

$\mathbf{X}_2$ is the corresponding matrix for group 2.

The obtained F is compared with $F_{[\alpha,\, 1,\, N - J(C+1)]}$.

Computational Example of Method II

Example 11.1 data listed in Table 11.1 are employed to illustrate the equivalence of the regression approach (described earlier as method I) and the F-test procedure described here. Once again, we are interested in testing for treatment effects in a

subpopulation of subjects who have *X* scores of 8. The steps involved in performing the matrix approach using *Minitab* matrix commands are shown below.

```
MTB > Read 15 2 m1.

DATA> 1 1
DATA> 1 2
DATA> 1 2
DATA> 1 3
DATA> 1 4
DATA> 1 5
DATA> 1 5
DATA> 1 6
DATA> 1 6
DATA> 1 7
DATA> 1 8
DATA> 1 8
DATA> 1 9
DATA> 1 10
DATA> 1 11
15 rows read.

MTB > Read 15 2 m2.

DATA> 1 1
DATA> 1 1.5
DATA> 1 2.5
DATA> 1 3.5
DATA> 1 4.5
DATA> 1 4.5
DATA> 1 5
DATA> 1 6
DATA> 1 6
DATA> 1 7
DATA> 1 7
DATA> 1 7.5
DATA> 1 8
DATA> 1 9
DATA> 1 10
15 rows read.

MTB > Transpose M1 m3.
MTB > Multiply M3 M1 m4.
MTB > Invert M4 m5.
MTB > Transpose M2 m6.
MTB > Multiply M6 M2 m7.
MTB > Invert M7 m8.
MTB > Add M5 M8 m9.
MTB > Read 2 2 m10.
DATA> .42 0
```

```
DATA> 0 .42
2 rows read.

MTB > Multiply M10 M9 m11.
MTB > Read 1 2 m12.
DATA> 1 8
1 rows read.

MTB > Read 2 1 m13.
DATA> 6.702
DATA> -1.028
2 rows read.

MTB > Transpose M12 m14.
MTB > Transpose M13 m15.
MTB > Multiply M12 M13 m16.

Answer = -1.5220

MTB > Multiply M16 M15 m17.
MTB > Multiply M17 M14 m18.

Answer = 2.3165

MTB > Multiply M12 M11 m19.
MTB > Multiply M19 M14 m20.

Answer = 0.0973
MTB > let k1= 2.3165/.0973

MTB > print k1
Data Display
K1   23.8078
```

The obtained F of 23.8078 far exceeds the critical value of F, which is 4.23 (based on 1 and 26 degrees of freedom). The corresponding p-value is .000046. It is concluded that aggression is lower under therapy 1 than it is under therapy 2 for a subpopulation of subjects having sociability scores $= 8$. The outcome of this F-test is consistent with both the previously computed J–N technique (which provides a region of nonsignificance of 6.04 through 7.06) and the PPA t-test computed using method I.

The correspondence among the t, F, and J–N outcomes in this example is not just coincidental; it is mathematical. The F-value from PPA method II is the square of the t-value from PPA method I. Similarly, it turns out that all values of $\mathbf{x}_s$ that satisfy the inequality

$$\mathbf{x}_s'\mathbf{D}\mathbf{D}'\mathbf{x}_s - F_{[\alpha,1,N-J(C+1)]}\mathbf{x}_s'\mathbf{V}\mathbf{x}_s \geq 0$$

define the J–N region of significance.

There is an easy way to demonstrate the correspondence of the PPA approaches and the J–N region of nonsignificance. First, compute the J–N region of nonsignificance; then use the upper limit of the computed region as the picked point in a PPA. If this is done with the example aggression data where the upper nonsignificance limit (X_{L2}) is 7.06, the p-value associated with the PPA is .051. Hence, it is obvious that an X value of 7.06 is right on the edge of the nonsignificance region. If a somewhat larger value of X is picked, the PPA will yield a p-value less than .05, as would be expected.

Confidence Intervals

When confidence intervals rather than significance tests are desired (a good idea), compute

$$\mathbf{x}_s'\mathbf{D} \pm \sqrt{F_{[\alpha,1,N-J(C+1)]}}\sqrt{\mathbf{x}_s'\mathbf{V}\mathbf{x}_s}.$$

For the example, the 95% confidence interval (using a covariate score of 8) is

$$[\,1 \quad 8\,]\begin{bmatrix} 6.705 \\ -1028 \end{bmatrix} \pm \sqrt{4.232}\sqrt{0.09731} = (-2.17, -0.88).$$

Given the individual point $\mathbf{x}_s$ the probability is .95 that the obtained confidence interval will cover the true population mean difference $\mathbf{x}_s'\boldsymbol{\Delta}$. The negative sign of $\mathbf{x}_s'\mathbf{D}$ in this example (i.e., $x_s'D = -1.52$) indicates that the expected value for group 1 is lower than the expected value for group 2; all values (in this example) of $\mathbf{x}_s'\mathbf{D}$ are negative for covariate scores above the point of intersection of the two slopes and positive for covariate values below the point of intersection. The point of intersection (on the X dimension) for the two-group one-covariate case can be computed using the following formula:

$$\frac{b_0^{(\text{group 1})} - b_0^{(\text{group 2})}}{b_1^{(\text{group 2})} - b_1^{(\text{group 1})}}.$$

In the case of multiple covariates, there will generally be many combinations of covariate scores that are associated with no difference on Y.

Standardized Effect Size

A standardized effect size can be computed using $\frac{\mathbf{x}_s'\mathbf{D}}{s_w} = g$, where s_w is the pooled within-group standard deviation. In this situation g applies to the subpopulation of subjects having covariate values included in vector $\mathbf{x}_s$.

11.3 A COMMON METHOD THAT SHOULD BE AVOIDED

Suppose that the data of Table 11.2 are viewed as a two-factor ANOVA problem. Factor A is type of therapy (group 1 vs. group 2), and factor B is sociability (classified as high or low). Subjects with sociability scores equal to or less than 5 are classified as "low"; scores above 5 are classified as "high." The data arrangement (Y scores) is as follows:

		Factor B: Sociability	
		Low	High
Factor A: Therapy	Group 1	10, 10, 11, 10, 11, 11, 10	11, 11.5, 12, 12, 11, 11, 12.5, 12
	Group 2	5, 6, 6, 7, 8, 9, 9	9, 10.5, 11, 12.5, 12.5, 14, 14.5, 16

The cell means are as follows:

	B_1	B_2
A_1	10.43	11.63
A_2	7.14	12.50

The ANOVA summary is as follows:

Source	SS	df	MS	F
A: Therapy	8.53	1	8.53	3.97
B: Sociability	80.17	1	80.17	37.26
$A \times B$	32.31	1	32.31	15.02
Within cell	55.95	26	2.15	
Total	176.97	29		

The $A \times B$ interaction is significant, which tells us that the simple effects of therapy for those classified as "low" on sociability differs from the simple effects of therapy for those classified as "high" on sociability. Simple main-effects tests are then carried out to determine whether there are simple effects of therapy for each category of sociability.

Source	SS	df	MS	F	p-value
A_1 vs. A_2 at B_1	37.79	1	37.79	17.57	.000283
A_1 vs. A_2 at B_2	3.06	1	3.06	1.42	.244168
Within cell	55.95	26	2.15		

These tests lead us to conclude that the 3.29 point difference between therapies is significant for subjects classified as "low" on sociability, but the −.87 point difference between therapies for subjects classified as "high" is not.

A comparison of these results with the outcome of J–N and PPA analyses reveals two advantages of the latter approaches. First, the J–N and PPA methods allow statements about treatment differences for *any* observed value of X. The classification of X into "low" and "high" levels for ANOVA has discarded much information on this dimension. The two simple main effects tests evaluate the hypotheses that (1) there are no treatment effects for subjects with an average sociability score of 3.18 (i.e., the average X score for all subjects classified as "low" on X) and (2) there are no treatment effects for subjects with an average sociability score of 7.84 (i.e., the average X score for all subjects classified as "high" on X). Variation on X within the "low" and "high" categories is ignored when these tests are used.

Second, the precision of the tests is much lower with the ANOVA approach. The difference between the two types of therapy is clearly significant for an X score of 7.84 according to the outcome of the J–N and PPA analyses. The t-value for PPA when the point picked on $X = 7.84$ is 4.49 ($p = .000129$). The ANOVA simple main-effects test on a similar difference (i.e., A_1 vs. A_2 at B_2) is clearly *not* significant. What is the main reason the effect estimate is statistically significant with J–N and PPA but not with the ANOVA simple main-effects test? Note the size of the error mean square (MS) for the different methods:

$$\text{MSres}_\text{i} = 0.42 \text{ (used in J–N and PPA)}, \quad \text{and}$$
$$\text{MS within cell} = 2.15 \text{ (used in ANOVA)}.$$

In this example the pooled within-treatment residual mean square (MSres_i) is much smaller than the within-cell mean square, primarily because a high linear relationship exists between X and Y. (A lower relationship between X and Y would yield a smaller power advantage for the J–N technique.) An additional advantage of J–N and PPA is that they consume fewer error degrees of freedom when compared with a two-factor ANOVA having more than two levels of the classification factor. Hence, they are evaluated against a lower criterion.

In summary, J–N and PPA methods are preferable to two-factor ANOVA and simple main effects tests because (1) they are more powerful and (2) they allow tests of treatment effects for any and all observed levels of the covariate.

11.4 ASSUMPTIONS

The assumptions underlying the appropriate use of the J–N and PPA approaches are essentially the same as those associated with ANCOVA—with one exception. It is not, of course, assumed that the regressions are homogeneous. The general equivalence of the assumptions of these two procedures is not surprising because the purpose and computation are similar for both.

The basic difference between the two procedures is that treatment effects are estimated at the grand covariate mean with ANCOVA, but with J–N and PPA the

treatment effects are estimated as a function of the covariate score. A complete description of all the assumptions is not presented here because they are equivalent to ANCOVA assumptions (described in Chapter 8). I do, however, summarize the findings of studies on the consequences of violations of several of the assumptions. The major assumptions are that

1. The residuals of the individual within-group regressions of Y on X are independent.
2. The residuals are normally distributed.
3. The residuals have homogeneous variance for each value of X (the homoscedasticity assumption).
4. The residuals have homogeneous variance across treatment groups (i.e., the conditional variances for all J groups are equal).
5. The regression of Y on X is linear.
6. The levels of the covariate are fixed.
7. The covariate is measured without error.

It is known that the independence assumption is critical and that violations of this assumption will generally lead to an increase in the probability of type I error. The consequences of violating the normality assumption appear to be minor unless extreme departures (or outliers) are present (Mendro, 1975). Violation of the assumption of homogeneity of conditional variances across treatments has relatively little effect. Shields (1978) found that when this assumption was violated, the effects were essentially the same as is found with ANOVA and ANCOVA. That is, if sample sizes are equal, the probability of type I error is little affected. But this is not the case with unequal sample sizes. When the larger variance is combined with the larger group, the probability of type I error is less than the nominal α. When large variances are combined with small groups, the probability of type I error is greater than the nominal α. These results are consistent with those of several other older studies. Erlander and Gustavsson (1965) evaluated the effects of heterogeneous conditional variances where one population variance was twice the size of the other. This degree of heterogeneity had little effect on either the confidence band for the treatment difference or on the region of significance. Borich and Wunderlich (1973) came to a similar conclusion. Mendro (1975) investigated the effects of severe heterogeneity of conditional variances. When the larger:smaller conditional variance ratio was 4, the probability of type I error was found to follow the same pattern reported by Shields; that is, positive bias in F when the smaller samples are associated with the larger variance and negative bias when the smaller samples are associated with the smaller variance.

With respect to homoscedasticity, Shields found the J–N technique to be sensitive to violations with both equal and unequal sample sizes. She found that as the variance for a fixed value of the covariate increased, the probability of including that X-value in the region of significance increased.

The effects of violating the assumption of fixed error-free covariates have been investigated by Rogosa (1977). If the covariates are random error-free variables rather than fixed, the type I error rate is essentially unchanged. The power of the test,

however, is lower in this case of random variables. Rogosa (1977) also investigated the effects of measurement error.

Three major consequences of measurement error in the covariate are:

1. An inevitable shrinking of the J–N region of significance.
2. Reduced probability that the homogeneity of regression slopes test will identify heterogeneous slopes (this means that the J–N technique will be used less frequently when measurement error is present because ANCOVA is used when the slopes are evaluated as being homogeneous).
3. Possible increase in type I error.

Although it may seem that points 1 and 3 are contradictory, they are not. When observed scores (scores that contain measurement error) are used, values of X that lie in the region of significance may lie in the region of nonsignificance if true scores are used. Details on how this can occur can be found in the excellent work of Rogosa (1977).

11.5 TWO GROUPS, MULTIPLE COVARIATES

Generalized J–N Methods

Generalizations of the J–N technique are available to accommodate multiple covariates. If two covariates are involved, the J–N technique can be used to identify the region of significant differences in the plane of the covariates. An illustration of heterogeneous regression planes can be seen in Figure 11.5. If three or more covariates are employed, a region of significant differences in the hyperspace of the covariates can be obtained. Because four-dimensional representations are beyond my psychomotor skills, an illustration of heterogeneous regression hyperplanes is not provided here.

But others have not been daunted by the task. Valiant attempts have been made to apply sophisticated symbolic and graphical software to provide four-dimensional graphical contour plots to simplify interpretation in the three-covariate case (Hunka,

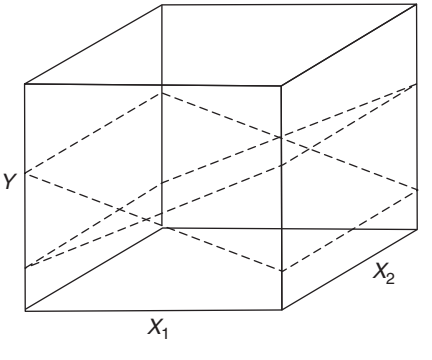


Figure 11.5 Illustration of heterogeneous regression planes in two-covariate case.

1994, 1995; Hunka and Leighton, 1997). The outcome of this work is interesting; software to produce plots of this type is available at

http//www.wolfram.com.cgi-bin/MathSource/Enhancements/Statistics/0208-066T

and

http//www.wolfram.com.cgi-bin/MathSource/Enhancements/Statistics/0207-953

The approach provides a quasi four-dimensional graphical representation that involves three-dimensional contour plotting where one of the covariates is set to a constant value to reduce the plot of the region of significance to two dimensions.

Alas, I do not recommend the generalized J–N approach for more than two covariates. My impression is that most audiences exposed to the graphics designed to communicate the multidimensional results of such analyses are baffled. The problem is not the software; rather it is the inherent complexity of the sample space.

Generalized PPA

It was pointed out earlier in the initial description of the matrix approach to PPA that it applies to both a single covariate and to multiple covariates. An example of applying it to a three-covariate problem is described here.

Computational Example: Two Groups, Three Covariates

Suppose that the following data were collected in a two-group three-covariate experiment:

Group 1				Group 2			
X_1	X_2	X_3	Y	X_1	X_2	X_3	Y
1	9	3	10	1	8	2	5
2	7	5	10	1.5	9	4	6
2	9	4	11	2.5	9	4	6
3	9	3	10	3.5	8	5	7
4	9	3	11	4.5	8	4	8
5	8	4	11	4.5	8	4	9
5	7	4	10	5.5	7	3	9
6	9	4	11	6	8	3	9
6	6	3	11.5	6	6	4	10.5
7	8	5	12	7	7	3	11
8	7	4	12	7	5	5	12.5
8	7	5	11	7.5	8	5	12.5
9	6	6	11	8	6	6	14
10	6	6	12.5	9	7	5	14.5
11	6	6	12	10	6	6	16

The predictors used for ANCOVA and the homogeneity of regression hyperplanes tests are as follows:

D	X_1	X_2	X_3	DX_1	DX_2	DX_3	Y
1	1	9	3	1	9	3	10
1	2	7	5	2	7	5	10
1	2	9	4	2	9	4	11
1	3	9	3	3	9	3	10
1	4	9	3	4	9	3	11
1	5	8	4	5	8	4	11
1	5	7	4	5	7	4	10
1	6	9	4	6	9	4	11
1	6	6	3	6	6	3	11.5
1	7	8	5	7	8	5	12
1	8	7	4	8	7	4	12
1	8	7	5	8	7	5	11
1	9	6	6	9	6	6	11
1	10	6	6	10	6	6	12.5
1	11	6	6	11	6	6	12
0	1	8	2	0	0	0	5
0	1.5	9	4	0	0	0	6
0	2.5	9	4	0	0	0	6
0	3.5	8	5	0	0	0	7
0	4.5	8	4	0	0	0	8
0	4.5	8	4	0	0	0	9
0	5.5	7	3	0	0	0	9
0	6	8	3	0	0	0	9
0	6	6	4	0	0	0	10.5
0	7	7	3	0	0	0	11
0	7	5	5	0	0	0	12.5
0	7.5	8	5	0	0	0	12.5
0	8	6	6	0	0	0	14
0	9	7	5	0	0	0	14.5
0	10	6	6	0	0	0	16

The ANCOVA F-value (2.14) is not statistically significant, but the homogeneity of regression F-value (59.55) certainly is. Because the regression hyperplanes are clearly heterogeneous, we ignore the outcome of ANCOVA and employ PPA instead.

Suppose that we want to determine whether the two treatments differ for a population of subjects having the following covariate scores on covariates 1, 2, and 3, respectively: $X_1 = 4$, $X_2 = 9$, and $X_3 = 3$. The following steps are carried out:

1. Center the first covariate by subtracting 4 from each value in X_1.

2. Center the second covariate by subtracting 9 from each value in X_2.

3. Center the third covariate by subtracting 3 from each value in X_3.

4. Multiply *D* times the first centered covariate.
5. Multiply *D* times the second centered covariate.
6. Multiply *D* times the third centered covariate.
7. Regress *Y* on *D* and the six variables constructed using steps 1 to 6.

These seven steps are shown below as *Minitab* input commands; the first seven lines correspond to steps 1 to 7.

Input:

```
MTB > let c6=X1-4
MTB > let c7=X2-9
MTB > let c8=X3-3
MTB > let c9=c2*c6
MTB > let c10=c2*c7
MTB > let c11=c2*c8
MTB > Regress 'Y' 7 'D' 'X1-4'-'D*X3-3';
SUBC>   Constant;
SUBC>   Brief 2.
```

The follow output shows (1) the data in the worksheet (including the original covariates that are not used in the ultimate analysis) and (2) relevant portions of the output from the regression of *Y* on predictor variables *D* and the six variables generated using steps 1 to 6.

Output:

(1) Worksheet

```
MTB > print c1-c11

                        Data Display

Row     Y  D    X1  X2  X3  X1-4  X2-9  X3-3  D*X1-4  D*X2-9  D*X3-3
  1  10.0  1   1.0   9   3  -3.0     0     0      -3       0       0
  2  10.0  1   2.0   7   5  -2.0    -2     2      -2      -2       2
  3  11.0  1   2.0   9   4  -2.0     0     1      -2       0       1
  4  10.0  1   3.0   9   3  -1.0     0     0      -1       0       0
  5  11.0  1   4.0   9   3   0.0     0     0       0       0       0
  6  11.0  1   5.0   8   4   1.0    -1     1       1      -1       1
  7  10.0  1   5.0   7   4   1.0    -2     1       1      -2       1
  8  11.0  1   6.0   9   4   2.0     0     1       2       0       1
  9  11.5  1   6.0   6   3   2.0    -3     0       2      -3       0
 10  12.0  1   7.0   8   5   3.0    -1     2       3      -1       2
 11  12.0  1   8.0   7   4   4.0    -2     1       4      -2       1
 12  11.0  1   8.0   7   5   4.0    -2     2       4      -2       2
 13  11.0  1   9.0   6   6   5.0    -3     3       5      -3       3
```

14	12.5	1	10.0	6	6	6.0	-3	3	6	-3	3
15	12.0	1	11.0	6	6	7.0	-3	3	7	-3	3
16	5.0	0	1.0	8	2	-3.0	-1	-1	0	0	0
17	6.0	0	1.5	9	4	-2.5	0	1	0	0	0
18	6.0	0	2.5	9	4	-1.5	0	1	0	0	0
19	7.0	0	3.5	8	5	-0.5	-1	2	0	0	0
20	8.0	0	4.5	8	4	0.5	-1	1	0	0	0
21	9.0	0	4.5	8	4	0.5	-1	1	0	0	0
22	9.0	0	5.0	7	3	1.0	-2	0	0	0	0
23	9.0	0	6.0	8	3	2.0	-1	0	0	0	0
24	10.5	0	6.0	6	4	2.0	-3	1	0	0	0
25	11.0	0	7.0	7	3	3.0	-2	0	0	0	0
26	12.5	0	7.0	5	5	3.0	-4	2	0	0	0
27	12.5	0	7.5	8	5	3.5	-1	2	0	0	0
28	14.0	0	8.0	6	6	4.0	-3	3	0	0	0
29	14.5	0	9.0	7	5	5.0	-2	2	0	0	0
30	16.0	0	10.0	6	6	6.0	-3	3	0	0	0

(2) Regression of Y on D, all centered covariates, and the products of D times each centered covariate.

```
Regression Analysis: Y versus D, X1-4, ...

The regression equation is

Y = 7.38 + 3.48 D + 1.02 X1-4 - 0.318 X2-9 + 0.430 X3-3 -
0.761 D*X1-4 + 0.442 D*X2-9 - 0.501 D*X3-3

Predictor       Coef   SE Coef       T       P
Constant      7.3833    0.3182   23.20   0.000
D             3.4847    0.4236    8.23   0.000
X1-4         1.02427   0.09527   10.75   0.000
X2-9         -0.3180    0.1874   -1.70   0.104
X3-3          0.4302    0.1744    2.47   0.022
D*X1-4       -0.7609    0.1270   -5.99   0.000
D*X2-9        0.4416    0.2668    1.66   0.112
D*X3-3       -0.5007    0.2736   -1.83   0.081

S = 0.588280   R-Sq = 95.7%   R-Sq(adj) = 94.3%
```

It can be seen by inspecting the coefficient for variable D that the estimated outcome difference between the expected value of treatment 1 and the expected value of treatment 2 is 3.4847 points on the aggression scale for a subpopulation of subjects having $X_1 = 4$, $X_2 = 9$, and $X_3 = 3$. The p-value associated with the test on this difference is less than .001. The same result is obtained when the computation is carried out using the matrix approach described in Section 11.2. Although it can be seen that the computation required for three or more covariates is straightforward, I encourage restraint when tempted to use more than two.

11.6 MULTIPLE GROUPS, ONE COVARIATE

The original presentation of the J–N technique (Johnson and Neyman, 1936) considered the case of two groups. If three or more groups are involved, several alternatives, two of which are described in Potthoff (1964), are available. A simple but conservative procedure is to employ the two-group analysis to each pair of groups in the experiment, with one slight modification. Substitute the Bonferroni F-statistic in place of the conventional F-statistic where F is called for in the computational formulas.

If four groups are involved one might be interested in as many as

$$\frac{J(J-1)}{2} = \frac{4(4-1)}{2} = 6$$

comparisons of pair of groups using the J–N procedure. The Bonferroni F critical value is determined using $df_1 = 1$, $df_2 = N - J(C+1)$, and α set at the desired familywise level divided by the number of planned individual J–N analyses $= C' = 6$. For example, if alpha is set at .05 for the whole collection of comparisons the critical value of F used in each individual analysis is based on $\alpha/C' = .05/6 = .0083$. If the investigator has interest in and plans (before data are collected) to analyze only four of the six possible comparisons, the critical value of F is determined using $df_1 = 1$, $df_2 = N - J(C+1)$, and $C' = 4$. Similarly, when the PPA approach is applied instead of the J–N technique, the critical value is based on the Bonferroni t-statistic.

Alternative Multiple-Group Technique

An alternative to the multiple-group approach just described involves the following procedure. First, run the homogeneity of regression test. If this test yields a significant F-value, the overall homogeneity of regression hypothesis is rejected, but it does not lead to the conclusion that all slopes differ from each other. Follow up the overall test with additional homogeneity of regression tests on all combinations of pairs of slopes. Frequently, the slopes are homogeneous for most of the groups. Apply ANCOVA to the collection of groups that have homogeneous slopes; apply either the J–N procedure or PPA for comparisons involving groups with heterogeneous slopes. For example, if four groups are employed in an experiment and the overall homogeneity of regression test is significant, run follow-up homogeneity of regression tests for the six comparisons. Suppose the following pattern of homogeneity of regression test results is obtained:

Groups Involved in Slope Comparison	Statistically Significant?
1,2	No
1,3	No
1,4	Yes
2,3	No
2,4	Yes
3,4	Yes

Because comparisons of all groups suggest that the slopes of groups 1, 2, and 3 are homogeneous, a single ANCOVA is appropriate for these groups. The slope of group 4 differs from the slopes of the others; separate J–N or PPA comparisons between group 4 and each of the other groups would then be carried out. Use Bonferroni critical values based on $C' = 4$ for the one ANCOVA and the three J–N analyses if there is interest in controlling the error rate familywise.

11.7 ANY NUMBER OF GROUPS, ANY NUMBER OF COVARIATES

Situations in which more than two groups are involved can be analyzed pairwise using J–N and PPA regardless of the number of covariates. The Bonferroni F-statistic is substituted for the conventional F-statistic in all computations if the experimenter desires to maintain the error rate at α for the whole collection of pairwise contrasts.

Suppose that three groups had been involved in the example described in Section 11.2 rather than two. If the experimenter selects a point $\mathbf{x}_s$ and computes $\mathbf{x}_s'\mathbf{D}$ for each pair of groups, he or she will have three $\mathbf{x}_s'\mathbf{D}$ values. Each value may be evaluated using PPA. Because three tests are being carried out, the critical value is based on the Bonferroni F (or Bonferroni t if the regression approach is used) where $C' = 3$, rather than the conventional statistic. The probability of making a type I error in the collection of three tests is equal to or less than α. (A type I error is committed when the difference $\mathbf{x}_s'\mathbf{D}$ is declared significant when the true difference $\mathbf{x}_s'\Delta$ is zero.) If simultaneous confidence intervals rather than tests are of interest, Bonferroni critical values are used in the PPA confidence interval formula (see Table 11.4).

11.8 TWO-FACTOR DESIGNS

The homogeneity of regression slopes assumption applies to the case of two-factor designs as well as to one-factor designs. The conventional homogeneity of regression

Table 11.4 Formulas for PPA Individual and Simultaneous Tests and Confidence Intervals

Conventional PPA Formulas	
Individual Significance Tests	Individual Confidence Intervals
$F = (\mathbf{x}_s'\mathbf{D}\mathbf{D}'\mathbf{x}_s)/(\mathbf{x}_s'\mathbf{V}\mathbf{x}_s)$	$\mathbf{x}_s'\mathbf{D} \pm \sqrt{F_{[\alpha,1,N-J(C+1)]}}\sqrt{\mathbf{x}_s'\mathbf{V}\mathbf{x}_s}$
Critical value $= F_{[\alpha,1,N-J(C+1)]}$	
Modified PPA Formulas for Simultaneous Inference	
Simultaneous Significance Tests	Simultaneous Confidence Intervals
Tests are computed by using conventional PPA tests, but critical value is:	$\mathbf{x}_s'\mathbf{D} \pm \sqrt{(C+1)F_{[\alpha,C+1,N-J(C+1)]}}\sqrt{\mathbf{x}_s'\mathbf{V}\mathbf{x}_s}$
$(C+1)F_{[\alpha,C+1,N-J(C+1)]}$	

test associated with the independent sample two-factor design is a test of the equality of the cell slopes. This test is directly relevant to the interpretation of the $A \times B$ interaction and any simple main-effects tests that may be of interest. If the cell slopes are homogeneous, main-effects tests, the interaction test, and simple main-effects tests are all appropriate. Whereas heterogeneity of cell slopes invalidates the $A \times B$ interaction test, the main-effects tests may be appropriate in this case. Additional testing may help decide whether the main effects should be interpreted.

It is assumed that the slopes within the levels of factor A are homogeneous. This assumption is tested by using the conventional homogeneity of regression test on the various levels of factor A (not on the cells). If this test reveals heterogeneity of slopes for the different levels of factor A, follow-up homogeneity of regression tests on the pairs of levels of factor A should be run in the case that the number of levels is greater than two. Use ANCOVA to test the adjusted effects on factor A for those levels having homogeneous slopes. Use the J–N procedure or PPA to analyze comparisons involving levels of A with heterogeneous slopes. The same procedure is recommended for factor B. The general rule to keep in mind is that homogeneity of slopes should be present for the levels or cells being compared with ANCOVA. If this is not the case for any comparison of interest, use J–N or PPA for those comparisons that involve heterogeneous regression slopes.

11.9 INTERPRETATION PROBLEMS

Two-group one-covariate J–N and PPA procedures yield results no more difficult to interpret than are ANCOVA results. There are three issues however, that should be kept in mind when interpreting the outcome of these analyses: (1) the extrapolation problem, (2) the possibility of having heterogeneous slopes but no subjects for which differences are significant, and (3) simultaneous inference problems in the situations where two or more regions are computed. These points are discussed in the remainder of this section.

The Extrapolation Problem

It is important to restrict the generalization of results to the range of values of the covariate(s) observed in the samples. Consider the data illustrated in Figure 11.2. The J–N analysis indicates that X scores below 6.04 and above 7.06 are associated with significant differences between the two treatments. The range of X included in the study is 1 through 11. Suppose someone wants to know if the effects of the two treatments are different at $X = 20$. It is possible to test this hypothesis by observing that an X value of 20 falls in the upper J–N significance region, but no inferential statements should be made concerning treatment effects at levels of X far outside the range of scores included in the observed data. If there are no data on subjects with X scores near 20 we have no justification for making statements concerning treatment effects at such extreme X scores. It might not be unreasonable to extrapolate results to X scores of 12 or 13, but extrapolation to more extreme scores would be poor practice. If the effects of the treatments for subjects with X scores of 20 are important,

the investigator should make sure to include subjects with this score in the experiment. Extrapolation is useful for hypothesis generation, but it can be dangerous if it is not recognized as a form of speculation.

Inconsistencies in Outcomes of Homogeneity of Regression and Johnson–Neyman Tests

A large homogeneity of regression slopes F-ratio is usually a sign that the J–N procedure will reveal regions of significance and nonsignificance. It is possible, however, to obtain a significant homogeneity of regression F-ratio, but no region of significance *within the range of the sample data*. This will generally occur only when the heterogeneity of slopes is slight and the associated F is just barely significant. The regions of significance may lie outside the range of covariate scores included in the sample. As has already been pointed out, a researcher should be very careful about making statements of significant differences for ranges of the covariate(s) for which there are no data.

Simultaneous Inference

The probability of making a type I error in a two-group problem using J–N or PPA is the probability of rejecting H_0 when the true population difference between means at a specified point on X (in the one-covariate case) is zero. Hence, if several points on X are selected and a statement concerning the population difference at each of these points is made, the probability of making a type I error is α for each statement. Suppose, however, that we want to hold the probability of making a type I error at α for all statements simultaneously. In other words, if we state that differences are significant for many different values of X, we want α to be the probability that one or more of the whole family of statements is incorrect.

Similarly, if a 95% confidence interval is computed for each of many points on X using the conventional PPA procedure, we can state for each of the many confidence intervals that the probability is .95 that the obtained confidence interval will span the population mean difference. The probability that the whole collection of confidence intervals simultaneously includes all population differences is *not* .95. The exact probability is lower than .95 in this case. It is possible, however, to employ a slightly modified PPA procedure to obtain simultaneous tests and confidence intervals. The conventional and modified formulas for the two-group case are shown in Table 11.4.

If more than two groups are involved the Bonferroni F based on $df_1 = C + 1$, $df_2 = N - J(C + 1)$, and $C' = [J(J - 1)]/2$ is substituted for the conventional F in the formulas for simultaneous inference shown in Table 11.4. When this situation is encountered, we can be at least $100(1 - \alpha)\%$ confident that all confidence intervals across all $[J(J - 1)]/2$ group comparisons contain the population mean differences.

The modified formulas presented here are essentially the same as those presented by Potthoff (1964). Potthoff presented another modified formula for simultaneous confidence intervals based on the same rationale as Scheffé's multiple comparison

procedure. The procedure described in this section is generally more powerful than either Potthoff's Scheffé-type procedure (unless sample sizes are *very* small) or Erlander and Gustavsson's (1965) procedure. All of these procedures have been compared by Cahen and Linn (1971).

In summary, the conventional J–N and PPA procedures are appropriate if the experimenter wants to maintain the probability of making a type I error at α for each individual point on X. If, on the other hand, the objective is to make a probability statement about a whole set of differences, the simultaneous method may be employed. Using this approach the probability of one or more false rejections of H_0 in the whole collection or set of statements is less than or equal to α.

Consider again Example 11.1 analyzed using the matrix approach in Section 11.2. The PPA test yields $F_{\text{obt}} = 23.81$ at the covariate score 8. The critical value of 4.23 is based on the conventional F-statistic with 1 and $N - J(C + 1)$ degrees of freedom. It is appropriate to use this critical value because the experimenter wants to hold the probability of making a type I error at .05 for this particular comparison. Suppose, however, that the experimenter wants to make a statement for each value of X included in the experiment. Further, he or she wants the probability of making one or more type I errors in the whole collection of tests to be less than .05. The modified formula for the critical value yields $2(3.37) = 6.74$. Each test is evaluated using this critical value.

It can be seen that the critical value for the simultaneous approach is much larger than the critical value for the individual approach. Corresponding to the difference in critical values is a large difference in power. The power reduction is large enough that the researcher should carefully consider whether simultaneous inference is necessary; usually it is not. Issues of this type have been debated in the statistical literature for half a century under the general heading of multiplicity. The opinions are still diverse. My recommendation in the context of PPA is to plan to make three tests before performing the experiment and to use individual tests. For example, pick points on X falling one standard deviation below the mean, the mean, and one standard deviation above the mean. Make it clear that the error rate is per-comparison at each point when reporting the results.

11.10 MULTIPLE DEPENDENT VARIABLES

If multiple dependent variables are employed, the experimenter has a choice of error rates to consider. The experimenter may decide to treat each dependent variable as a separate experiment. In this case the conventional J–N procedure is applied to each dependent variable, and a separate statement (or set of statements) is made for each dependent variable. On the other hand, if the experimenter decides to treat the whole family of dependent variables as one experiment, he or she may want the probability of making a type I error to be equal to or less than α for the whole family of dependent variables. A conservative approach is to increase the critical value by substituting the Bonferroni F based on $df_1 = 1$, $df_2 = N - J(C + 1)$, and C' (the number of dependent variables) in place of the conventional F critical value.

Suppose that five dependent variables rather than one had been involved in the two-group one-covariate problem presented in Section 11.2. The conventional J–N or PPA is performed on each dependent variable; the critical value for each of the five tests or confidence intervals is

$$F_{\mathrm{B}[(\alpha,C,1,N-J(C+1)]} = F_{\mathrm{B}[.05,5,1,26]} = 7.72.$$

If we want the probability of type I error to be less than say .05 for all dependent variables *and* all levels of X simultaneously, the critical value is

$$(C+1)F_{\mathrm{B}[\alpha,C',C+1,N-J(C+1)]}.$$

For example, if there are 30 subjects, two groups, one covariate, five dependent variables, and $\alpha = .05$, the critical value for each test is $(2)F_{\mathrm{B}[.05,\,5,\,2,\,26]} = 11.06$ for each test.

11.11 NONLINEAR JOHNSON-NEYMAN ANALYSIS

The J–N technique, like the analysis of covariance, is based on the assumption that the relationship between the covariate and the dependent variable is linear. It is possible, however, to modify the analysis to deal with the situation in which the XY relationship is nonlinear. This modification of the J–N technique and a computer program for handling the rather complex computations involved have been presented in Wunderlich and Borich (1974) and Borich et al. (1976).

Before resorting to the Wunderlich-Borich approach, the nature of the nonlinearity should be identified. If the nonlinearity is of monotonic form, a transformation of the original data may yield linearity. The transformed data can then be substituted for the original data in a conventional J–N analysis. In the unlikely situation where the nonlinear relationship is not monotonic, the Wunderlich–Borich modifications should be considered.

11.12 CORRELATED SAMPLES

The conventional J–N technique is appropriate for two independent samples. Where one sample is observed under two conditions, or where matched pairs are randomly assigned to two conditions, the independent sample formula is inappropriate. A J–N alternative for this rare case can be found in the first edition of this book (Huitema, 1980, Chapter 13).

11.13 ROBUST METHODS

As the number of variables in the analysis increases so does the chance of encountering outliers that can distort or invalidate the results. If each variable is carefully diagnosed before performing a one- or two-covariate J–N or PPA the conventional

approaches are likely to be satisfactory. There are, however, situations in which outliers (especially multivariate outliers) are not identified. Robust methods are a safe way to avoid potential problems associated with outliers, whether they are identified or not. Effective robust versions of J–N and PPA methods are available (Watcharotone, 2010; Watcharotone et al., 2010). These methods provide much higher power when outliers are present; the more extreme the outliers the greater the gain in power over the conventional OLS methods.

11.14 SOFTWARE

Software for several conventional J–N models is described in Hayes and Matthes (2009), Hunka (1994, 1995), Hunka and Leighton (1997), Karpman (1980, 1983, 1986), Lautenschlager (1987), and Preacher et al. (2006). Routines for robust estimation of these models are described in Watcharotone et al. (2010) and are available at joseph.mckean@wmich.edu.

11.15 SUMMARY

The assumption of homogeneity of regression is important for the correct interpretation of ANCOVA. When this assumption is violated J–N and picked-points solutions are appropriate alternative analyses. Rather than evaluating the treatment effects at the covariate mean, as is the case with ANCOVA, these procedures evaluate the treatment effects as a function of the level of the covariate. The J–N procedure identifies regions of the X dimension that are associated with significant differences between the treatment groups (if such differences exist). Likewise, a region of nonsignificance is identified; it is the range of X scores associated with nonsignificant differences. The points-picked analysis provides a test for treatment effects at all points on the X dimension that are picked by the researcher. Extensions of both approaches are available for several groups, several covariates, two-factor designs, multiple dependent variables, and correlated samples. Robust versions are also available.

CHAPTER 12

Nonlinear ANCOVA

12.1 INTRODUCTION

The relationship between the covariate and the dependent variable scores is not always linear. Because an assumption underlying the ANCOVA model is that the within-group relationship between X and Y is linear, researchers should be aware of the problem of nonlinearity. If ANCOVA is employed when the data are nonlinear, the power of the F-test is decreased and the adjusted means may be poor representations of the treatment effects.

Two reasons for nonlinear relationships between X and Y are inherent nonlinearity of characteristics and scaling error. It is quite possible that the basic characteristics being measured are not linearly related. For example, the relationship between extroversion (X) and industrial sales performance (Y) could be predicted to be nonlinear. Those salespeople with very low extroversion scores may have poor sales performance because they have difficulty interacting with clients. Those with very high extroversion scores may be viewed as overly social and not serious about their work. Hence, very low or very high extroversion scores may be associated with low sales performance, whereas intermediate extroversion scores may be associated with high sales performance.

Another example of expected nonlinearity might be found between certain measures of motivation (X) and performance (Y). Psychologists working in the area of motivation sometimes hypothesize that there is an optimal level of motivation or arousal for an individual working on a specific task. At very low or very high levels of arousal, performance is lower than at the optimal level of arousal. In both examples, the relationship between X and Y scores is expected to be nonlinear because the relationship between the basic characteristic underlying the observed (measured) scores is expected to be nonlinear. This distinction between the measured and underlying or basic scores is important. It is quite possible that the relationship between observed X and Y scores is nonlinear when the relationship between the basic X and Y characteristics is linear. When this occurs, the problems of scaling error are involved.

The Analysis of Covariance and Alternatives: Statistical Methods for Experiments, Quasi-Experiments, and Single-Case Studies, Second Edition. Bradley E. Huitema.

There are several types of scaling errors that can produce nonlinearity, but probably the most frequently encountered type results in either "ceiling" or "floor" effects. In either case the problem is that the instrumentation or scale used in the measurement of either the X or the Y variable (or both) may not be adequate to reflect real differences in the characteristics being measured. For example, if most of the subjects employed in a study obtain nearly the highest possible score on a measure, there are likely to be unmeasured differences among those who get the same high score. The measurement procedure simply does not have sufficient "ceiling" to reflect differences among the subjects on the characteristics being measured. Suppose most subjects get a score of 50 on a 50-point pretest that is employed as a covariate; the test is much too easy for the subjects included in the experiment. If the scores on this measure are plotted against scores on a posttest that is of the appropriate difficulty level, nonlinearity will be observable. Here the inherent relationship between the X and Y characteristics is linear, but the obtained relationship between the observed measures is not linear. Hence, one reason for nonlinearity in the XY relationship is scaling error or inappropriate measurement. Regardless of the reason for nonlinearity, the linear ANCOVA model is inappropriate if the degree of nonlinearity is severe.

12.2 DEALING WITH NONLINEARITY

A routine aspect of any data analysis is to plot the data. This preliminary step involves plotting the Y scores against the X scores for each group. Severe nonlinearity will generally be obvious in both the trend observed in the scatter plot and in the shape of the marginal distributions. More sensitive approaches for identifying nonlinearity include visual inspection of the residuals of the ANCOVA model and fitting various alternative models to the data. Once it has been decided that nonlinearity is problematic, the next step is to either (1) seek a transformation of the original X and/or Y scores that will result in a linear relationship for the transformed data or to (2) fit an appropriate polynomial ANCOVA model to the original data.

Data Transformations

If the relationship between X and Y is nonlinear but monotonic (i.e., Y increases when X increases but the function is not linear), a transformation of X should be attempted. Logarithmic, square root, and reciprocal transformations are most commonly used because they usually yield the desired linearity. Advanced treatments of regression analysis should be consulted for details on these and other types of transformation (e.g., Cohen et al., 2003).

Once a transformation has been selected, ANCOVA is carried out in the usual way on the transformed data. For example, if there is reason to believe that the relationship between $\log_e X$ and Y is linear, ANCOVA is carried out using $\log_e X$ as the covariate. It must be pointed out in the interpretation of the analysis, however, that $\log_e X$ rather than X was the covariate.

A method of determining whether a transformation has improved the fit of the model to the data is to plot the scores and compute ANCOVA for both untransformed and transformed data. A comparison of the plots and ANCOVAs will reveal the effect of the transformation.

Polynomial ANCOVA Models

If the relationship between X and Y is not monotonic, a simple transformation will not result in linearity. In the nonlinear-monotonic situation, the values of Y increase as value of X increases. In the nonlinear-nonmonotonic situation, Y increases as X increases only up to a point, and then Y decreases as X increases. If we transform X to $\log_e X$ for the nonmonotonic situation, the $\log_e X$ values increase as X increases and nonlinearity is still present when $\log_e X$ and Y are plotted. The simplest alternative in this case is to fit a second-degree polynomial (quadratic) ANCOVA model. This model is written as

$$\bar{Y}_{ij} = \mu + \alpha_j + \beta_1 \left(X_{ij} - \bar{X}..\right) + \beta_2 \left(X_{ij}^2 - \left(\overline{X^2..}\right)\right) + \varepsilon_{ij},$$

where

Y_{ij} is the dependent variable score of ith individual in jth group;
μ is thepopulation mean on Y;
α_j is the effect of treatment j;
β_1 is the linear effect regression coefficient;
X_{ij} is the covariate score for ith individual in jth group;
$\bar{X}..$ is the mean of all observations on covariate;
β_2 is the curvature effect coefficient;
X_{ij}^2 is the squared covariate score for ith individual in jth group;
$\left(\overline{X^2..}\right)$ is the mean of squared observations on covariate (i.e., $\sum_{i=1}^{N} X_{ij}^2/N$); and
ε_{ij} is the error component associated with ith individual in jth group.

This model differs from the linear model in that it contains the curvature effect term $\beta_2(X_{ij}^2 - (\overline{X^2..}))$. If the dependent variable scores are a quadratic rather than a linear function of the covariate, this model will provide a better fit and will generally yield greater power with respect to tests on adjusted means.

The quadratic ANCOVA is computed by using X and X^2 as if they were two covariates in a multiple covariate analysis. The main ANCOVA test, the homogeneity of regression test, the computation of adjusted means, and multiple comparison tests are all carried out as with an ordinary two-covariate ANCOVA. If the relationship between X and Y is more complex than a quadratic function, a higher degree polynomial may be useful. The third-degree polynomial (cubic) ANCOVA model is written as

$$\bar{Y}_{ij} = \mu + \alpha_j + \beta_1 \left(X_{ij} - \bar{X}..\right) + \beta_2 \left(X_{ij}^2 - \left(\overline{X^2..}\right)\right) + \beta_3 \left(X_{ij}^3 - \left(\overline{X^3..}\right)\right) + \varepsilon_{ij}.$$

This model will provide a good fit if the relationship between the covariate and the dependent variable is a cubic function. Cubic ANCOVA is carried out by employing X, X^2, and X^3 as covariates in a multiple covariance analysis. Higher degree polynomials can be employed for more complex functions, but it is very unusual to encounter such situations.

Higher degree polynomial models virtually always fit sample data better than do simpler polynomial models, but this does not mean that the more complex models are preferable to the simpler ones. Care must be taken not to employ a more complex model than is required; there are essentially two reasons to keep the model as simple as possible. First, a degree of freedom is lost from the ANCOVA error mean square (i.e., MS_{Res_w}) for each additional term in the ANCOVA model. If the number of subjects is not large, the loss of degrees of freedom can easily offset the sum-of-squares advantage of a better fit afforded by the more complex model. Even though the sum-of-squares residual is smaller with more complex models, the mean-square error can be considerably larger with complex models. The consequences of the larger error term are less precise estimates of the adjusted means, and, correspondingly, less precise tests on the difference between adjusted means. This problem is illustrated in Section 12.3. The second reason for not employing a more complex model than is required is the law of parsimony. If a linear model fits the data almost as well as a quadratic model, the simpler model should usually be chosen because the interpretation and generalization of results is more straightforward.

Two additional points on the use of polynomial regression models are relevant to the polynomial ANCOVA described here. First, it is not necessary that the covariate be a fixed variable. This point was made earlier in the discussion of assumptions for ANCOVA but is reiterated here for nonlinear ANCOVA because, as Cramer and Appelbaum (1978) observed, it is sometimes mistakenly believed that polynomial regression is appropriate only with X fixed. Second, the parameters of the polynomial regression are sometimes difficult to estimate with certain multiple regression computer programs because these programs will not, with certain data sets, yield the inverse of the required matrix. This problem develops because X, X^2, X^3, and so on are all highly correlated. These computational difficulties can generally be reduced by transforming the raw X scores to deviation scores (i.e., centered scores) before the regression analysis is carried out. That is, in quadratic ANCOVA, for example, $(X - \bar{X})$ and $(X - \bar{X})^2$ rather than X and X^2 should be used as the covariates. Additional details on this problem in the context of conventional regression analysis can be found in Bradley and Srivastava (1979) and Budescu (1980).

12.3 COMPUTATION AND EXAMPLE OF FITTING POLYNOMIAL MODELS

The computation and rationale for quadratic ANCOVA are essentially the same as for multiple ANCOVA. Consider the following data:

(1)		(2)	
Experimental Group		Control Group	
X	Y	X	Y
13	18	11	13
7	14	2	1
17	7	19	2
14	14	15	9
3	8	8	10
12	19	11	15

Previous research or theoretical considerations may suggest that the relationship between X and Y is best described as a quadratic function. A scatter plot of these data appears to support a quadratic model. Hence, the experimenter has a reasonable basis for deciding to employ the quadratic ANCOVA model. The computation of the complete quadratic ANCOVA through the general linear regression procedure is based on the following variables:

(1)	(2)	(3)	(4)	(5)	(6)
D	X	X^2	DX	DX^2	Y
1	13	169	13	169	18
1	7	49	7	49	14
1	17	289	17	289	7
1	14	196	14	196	14
1	3	9	3	9	8
1	12	144	12	144	19
0	11	121	0	0	13
0	2	4	0	0	1
0	19	361	0	0	2
0	15	225	0	0	9
0	8	64	0	0	10
0	11	121	0	0	15

We now proceed as if we were performing a multiple covariance analysis using X and X^2 as the covariates. As before, the main test on adjusted treatment effects is based on the coefficients of multiple determination R^2_{yX} and $R^2_{yD,X}$.

The term R^2_{yX} represents the proportion of the total variability explained by the quadratic regression (i.e., the regression of Y on X and X^2), whereas $R^2_{yD,X}$ represents the proportion of the total variability explained by the quadratic regression and the treatments. Hence, the difference between the two coefficients represents the proportion of the variability accounted for by the treatments that is independent of that accounted for by quadratic regression. The proportion of unexplained variability is, of course, $1 - R^2_{yD,X}$.

Column 1 in this example is the only dummy variable (because there are only $J - 1$ dummy variables), columns 2 and 3 are the covariate columns, columns 4 and 5 are the interaction columns (not used in the main analysis), and column 6 contains the dependent variable scores. The regression analyses yield the following:

$$R^2_{yD,X} = R^2_{y123} = 0.918903 \quad \text{and}$$
$$R^2_{yX} = R^2_{y23} = 0.799091.$$

Difference or unique contribution of dummy variable beyond quadratic regression = 0.119812.

Total sum of squares = 361.67. The general form of the quadratic ANCOVA summary is as follows:

Source	SS	*df*	MS	*F*
Adjusted treatment	$\left(R^2_{yD,X} - R^2_{yX}\right)$ SST	$J - 1$	$SS_{AT}/(J - 1)$	MS_{AT}/MS_{Res_w}
Quadratic residual$_w$	$\left(1 - R^2_{yD,X}\right)$ SST	$N - J - 2$	$SS_{Res_w}/(N - J - 2)$	
Quadratic residual$_t$	$\left(1 - R^2_{yX}\right)$ SST	$N - 1 - 2$		

The quadratic ANCOVA summary for the example data is as follows:

Source	SS	*df*	MS	*F*
Adjusted treatment	$(0.119812)361.67 = 43.33$	1	43.33	11.82 ($p = .009$)
Quadratic residual$_w$	$(1 - 0.918903)361.67 = 29.33$	8	3.67	
Quadratic residual$_t$	$(1 - 0.799091)361.67 = 72.66$	9		

Adjusted means and multiple comparison procedures are also dealt with as they are under the multiple ANCOVA model. The adjusted means for the example data are obtained through the regression equation associated with $R^2_{y_{123}}$. The intercept and regression weights are

$$b_0 = -5.847359$$
$$b_1 = 3.83111$$
$$b_2 = 3.66943$$
$$b_3 = -0.17533$$

The group 1 dummy score, the grand mean covariate score, and the grand mean of the squared covariate scores are 1, 11, and 146, respectively. Hence, $\bar{Y}_{1\,\text{adj}} = -5.847359 + 3.83111(1) + 3.66943(11) - 0.17533(146) = 12.75$. The group 2 dummy score, the grand mean covariate score, and the grand mean of the squared covariate scores are 0, 11, and 146, respectively. Hence, $\bar{Y}_{2\,\text{adj}} = -5.847359 + 3.83111(0) + 3.66943(11) - 0.175333(146) = 8.92$.

Just as the test of the homogeneity of regression planes is an important adjunct to the main F test in multiple ANCOVA, the test of the homogeneity of the quadratic regressions for the separate groups should be carried out in quadratic ANCOVA. This test is computed in the same manner as the test of the homogeneity of regression planes.

The form of the summary is as follows:

Source	SS	*df*	MS	*F*
Heterogeneity of quadratic regression	$\left(R^2_{yD,X,DX} - R^2_{yD,X}\right)$ SST	$2(J-1)$	MS_{het}	$\dfrac{\text{MS}_{\text{het}}}{\text{MS}_{\text{Res}_i}}$
Quadratic residual$_i$	$\left(1 - R^2_{yD,X,DX}\right)$ SST	$N-(J-3)$	MS_{Res_i}	
Quadratic residual$_w$	$\left(1 - R^2_{yD,X}\right)$ SST	$N-J-2$		

A more general form, appropriate for testing the homogeneity of any degree (denoted as C) polynomial regression, is as follows:

Source	SS	*df*	MS	*F*
Heterogeneity of polynomial regression	$\left(R^2_{yD,X,DX} - R^2_{yD,X}\right)$ SST	$C(J-1)$	MS_{het}	$\dfrac{\text{MS}_{\text{het}}}{\text{MS}_{\text{Res}_i}}$
Polynomial residual$_i$	$\left(1 - R^2_{yD,X,DX}\right)$ SST	$N-J(C+1)$	MS_{Res_i}	
Polynomial residual$_w$	$\left(1 - R^2_{yD,X}\right)$ SST	$N-J-C$		

For the example data, the necessary quantities are

$$R^2_{yD,X,DX} = R^2_{y12345} = 0.944817 \quad \text{and}$$
$$R^2_{yD,X} = R^2_{y123} = 0.918903.$$

Difference or heterogeneity of regression $= 0.025914$.
Total sum of squares $= 361.67$.

The summary is as follows:

Source	SS	*df*	MS	*F*
Heterogeneity of polynomial regression	$(0.025914)361.67 = 9.37$	2	4.68	1.41 ($p = .32$)
Polynomial residual$_i$	$(1 - 0.944817)361.67 = 19.96$	6	3.33	
Polynomial residual$_w$	$(1 - 0.918903)361.67 = 29.33$	8		

The obtained F-value is clearly not significant; we conclude that there is little evidence to argue that the population quadratic regressions for the experimental

and control groups are different. The quadratic ANCOVA model is accepted as a reasonable representation of the data.

Comparison of Quadratic ANCOVA with Other Models

It was mentioned earlier that the complexity of the model employed should be sufficient to adequately describe the data but that it should not be more complex than is required. The results of applying four different models to the data of the example problem are tabulated as follows:

Model	Obtained F	Degrees of Freedom	p-value
ANOVA	2.62	1,10	.137
Linear ANCOVA	2.38	1,9	.157
Quadratic ANCOVA	11.82	1,8	.009
Cubic ANCOVA	9.96	1,7	.016

The F of the simplest model, ANOVA, when compared with the linear ANCOVA F, illustrates the fact that ANOVA can be more powerful than ANCOVA when the correlation between the covariate and the dependent variable is low. The F of the most complex of the four models, cubic ANCOVA, when compared with the quadratic F, illustrates the fact that more complex models do not necessarily lead to greater precision. The greatest precision is obtained with the model that is neither too simple nor more complex than is necessary for an adequate fit.

Minitab Input and Output

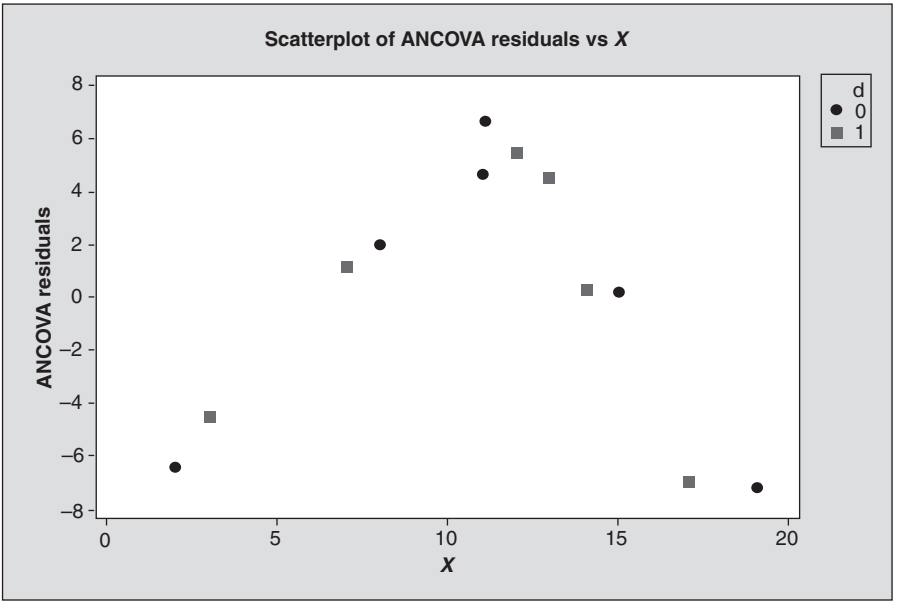
Input for estimating the linear ANCOVA model:

```
MTB > ancova Y=d;
SUBC> covariate X;
SUBC> means d;
SUBC> residuals c7.
```

Output for linear ANCOVA:

```
ANCOVA: Y versus d
Factor  Levels  Values
d            2  0, 1

Analysis of Covariance for Y
Source      DF  Adj SS     MS     F      P
Covariates   1    3.41   3.41  0.11  0.749
d            1   75.00  75.00  2.38  0.157
Error        9  283.25  31.47
Total       11  361.67
```

```
S = 5.61004   R-Sq = 21.68%   R-Sq(adj) = 4.28%

Covariate     Coef  SE Coef       T      P
X           0.1067    0.324  0.3293  0.749

Adjusted Means
d  N       Y
0  6   8.333
1  6  13.333

MTB > Plot 'ANCOVA Residuals'*'X';
SUBC>   Symbol 'd'.

Scatterplot of ANCOVA Residuals vs X
```

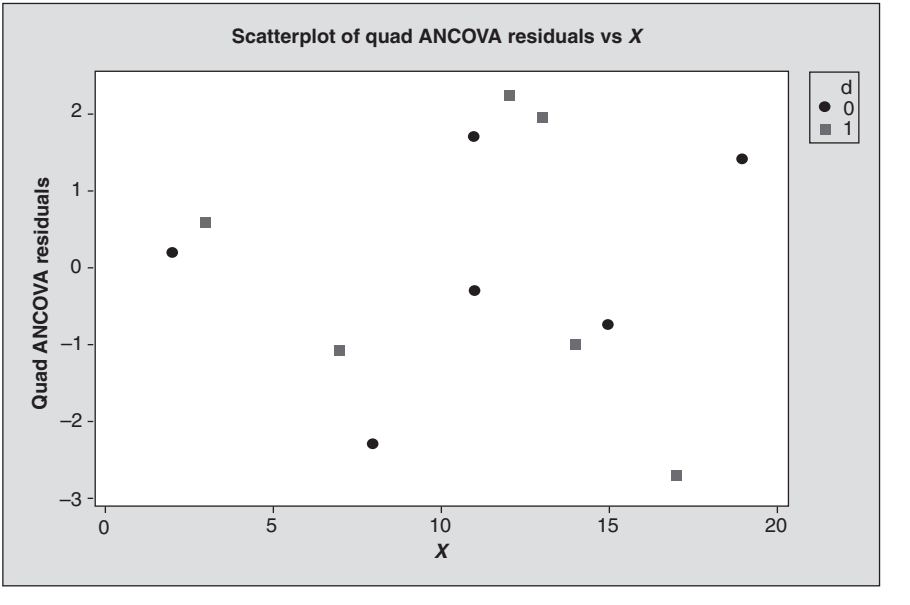


It is obvious from inspecting the plot of the residuals of the linear ANCOVA model shown above that this model is inappropriate. A quadratic model appears to be a good contender so it is estimated next.

Input to compute quadratic ANCOVA. The variable d is a (1, 0) dummy variable indicating group membership, c2 = the covariate X, and c3 = X^2.

```
MTB > ancova Y=d;
SUBC> covariates c2 c3;
SUBC> means d;
SUBC> residuals c8.
```

```
ANCOVA: Y versus d
Factor  Levels  Values
d            2  0, 1

Analysis of Covariance for Y
Source      DF  Adj SS      MS      F      P
Covariates   2  257.34  128.67  35.10  0.000
d            1   43.33   43.33  11.82  0.009
Error        8   29.33    3.67
Total       11  361.67

S = 1.91472   R-Sq = 91.89%   R-Sq(adj) = 88.85%

Covariate     Coef  SE Coef       T      P
X           3.6694   0.4421   8.299  0.000
X*X        -0.1753   0.0211  -8.322  0.000

Adjusted Means
d  N       Y
0  6   8.918
1  6  12.749

MTB > Plot 'Quad ANCOVA Residuals'*'X';
SUBC>   Symbol 'd'.

Scatterplot of Quad ANCOVA Residuals vs X
```



Note that the residuals of the quadratic ANCOVA model indicate no additional forms of nonlinearity or other departures from assumptions. This is confirmed by estimating the cubic ANCOVA model. Note in the output that the *p*-value on the cubic coefficient is .77.

Input for estimating the cubic ANCOVA model:

```
MTB > Let c9 = X*X*X
MTB > ancova Y=d;
SUBC> covariates c2 c3 c9;
SUBC> means d;
SUBC> residuals c10.
```

Output for cubic ANCOVA model:

```
ANCOVA: Y versus d
Factor  Levels  Values
d            2  0, 1

Analysis of Covariance for Y
Source      DF   Adj SS      MS      F      P
Covariates   3  257.720  85.907  20.77  0.001
d            1   41.179  41.179   9.96  0.016
Error        7   28.947   4.135
Total       11  361.667

S = 2.03354   R-Sq = 92.00%   R-Sq(adj) = 87.42%

Covariate     Coef  SE Coef        T      P
X           3.2073   1.5909   2.0161  0.084
X*X        -0.1231   0.1733  -0.7104  0.500
X*X*X      -0.0016   0.0054  -0.3040  0.770

Adjusted Means
d  N       Y
0  6   8.945
1  6  12.722
```

12.4 SUMMARY

The assumption of the conventional ANCOVA model that the covariate and the dependent variable are linearly related will not always be met. Severe nonlinearity generally can be easily identified by inspecting the *XY* scatter plot within groups. If the relationship is nonlinear but monotonic, it is likely that a simple transformation (generally of the *X* variable) can be found that will yield a linear relationship

between transformed X and Y. Analysis of covariance is then applied by using the transformed variable as the covariate. If the relationship is not monotonic, the simple transformation approach will not be satisfactory, and the more complex approach of employing some polynomial of X should be attempted. Generally, a quadratic or cubic ANCOVA model will fit the data. Complex polynomial models should be employed only if simpler ones are obviously inadequate. Simpler models are preferred because results based on complex models are (1) more difficult to interpret and generalize and (2) less stable. When polynomial ANCOVA models are clearly called for, the computation involves a straightforward extension of multiple ANCOVA.

CHAPTER 13

Quasi-ANCOVA: When Treatments Affect Covariates

13.1 INTRODUCTION

It is pointed out in Chapter 8 that the covariate(s) should be measured before treatments are applied. If the covariate is affected by the treatments, an undesirable consequence is that ANCOVA will remove a portion of the treatment effect from the dependent variable. This effect is easy to demonstrate.

The first example of ANCOVA described in Chapter 6 involves a randomized-group experiment using an aptitude measure as the covariate X; the design was appropriately structured so that the covariate was measured before the treatments were applied. Suppose that a slightly modified version of the experiment is carried out, but now the covariate is measured *after* the treatments are applied. Whereas minor chance differences between group means were observed on the covariate in the original experiment, this is not true in the modified experiment. Here nonchance treatment effects are part of the explanation for differences between groups on the covariate. The effects of these covariate differences on the outcome of ANCOVA are a concern.

The original covariate means are 52, 47, and 49, for treatment groups 1, 2, and 3, respectively. The modified experiment (covariate measured after treatments) yields covariate means of 54, 69, and 54, for treatments 1, 2, and 3, respectively. All other aspects of the two experiments are the same; that is, the variances within groups on X, the correlation within groups between X and Y, and the dependent variable scores are identical. *Minitab* ANCOVA summaries for both the original and the modified data are shown below.

The Analysis of Covariance and Alternatives: Statistical Methods for Experiments, Quasi-Experiments, and Single-Case Studies, Second Edition. Bradley E. Huitema.

ANCOVA on Original Data

```
Analysis of Covariance for Y

Source       DF   Adj SS       MS       F       P
Covariates    1  1855.35  1855.35  28.70  0.000
Tx            2   707.99   354.00   5.48  0.010
Error        26  1680.65    64.64
Total        29  3956.00

Adjusted Means
Tx   N       Y
1   10  28.479
2   10  40.331
3   10  36.190
```

ANCOVA on Modified Data

```
Analysis of Covariance for Y

Source       DF   Adj SS       MS       F       P
Covariates    1  1855.35  1855.35  28.70  0.000
Tx            2   214.53   107.27   1.66  0.210
Error        26  1680.65    64.64
Total        29  3956.00

Adjusted Means
Tx   N       Y
1   10  32.853
2   10  33.295
3   10  38.853
```

Note that the sum of squares for the adjusted treatment effect has dropped from 707.99 to 214.35. Correspondingly, the *p*-value for the adjusted treatment effect has changed from .01 to .21. The consequence of using ANCOVA in this situation is obvious; the covariate has removed over two-thirds of the treatment effect sum of squares but the error sum of squares is unchanged. A clearly statistically significant result has been adjusted away and the adjusted means have become misleading. This is a situation where ANCOVA should be avoided. But what analysis will avoid these problems?

By far the most frequently used alternative analysis in situations such as this is a conventional ANOVA. Because this approach yields unbiased estimates in the long run it is often adequate. But I do not recommend this almost universal practice. It is wasteful because it ignores useful covariate information. There is a better approach.

13.2 QUASI-ANCOVA MODEL

I recommend an approach that has major advantages over both conventional ANOVA and ANCOVA when the treatments have affected the covariate. I refer to this as a

quasi-ANCOVA analysis. It involves a modification of the conventional ANCOVA model that completely removes the influence of covariate mean differences from the treatment-effect estimates on the dependent variable. A comparison of the conventional ANCOVA and quasi-ANCOVA models is shown below.

Conventional ANCOVA Model

$$Y_{ij} = \mu + \alpha_j + \beta_1 \left(X_{ij} - \bar{X}..\right) + \varepsilon_{ij},$$

where

Y_{ij} is the dependent variable score of ith individual in jth group;
μ is the overall population mean (on dependent variable);
α_j is the effect of treatment j;
β_1 is the linear regression coefficient of Y on X;
X_{ij} is the covariate score for ith individual in jth group;
$\bar{X}..$ is the grand covariate mean; and
ε_{ij} is the error component associated with ith individual in jth group.

Quasi-ANCOVA Model

$$Y_{ij} = \mu + \alpha_j + \beta_1 \left(X_{ij} - \bar{X}_j\right) + \varepsilon_{ij}$$

where

Y_{ij} is the dependent variable score of ith individual in jth group;
μ is the overall population mean (on dependent variable);
α_j is the effect of treatment j;
β_1 is the linear regression coefficient of Y on X;
X_{ij} is the covariate score for ith individual in jth group;
$\bar{X}_j$ is the covariate mean for treatment j; and
ε_{ij} is the error component associated with ith individual in jth group.

The two models look the same except for what appears to be a trivial difference in one subscript. Note that the grand covariate mean $\bar{X}..$ in the conventional model is replaced with the treatment-group covariate mean $\bar{X}_j$ in the quasi-ANCOVA model. This minor change has a profound effect on the properties of the analysis. The estimation of the quasi-ANCOVA model is carried out just like conventional ANCOVA except that the covariate scores (X) are replaced by the residuals of the ANOVA model applied to the covariate.

The residuals in ANOVA are obtained by subtracting the individual treatment-group mean from each observation within the treatment group. The sum of these residuals within each treatment group is equal to zero. For example, if the scores in one treatment group are 3, 4, and 8, the mean is 5 and the corresponding residuals are $(3 - 5) = -2$, $(4 - 5) = -1$, and $(8 - 5) = -3$. The sum of these three residuals is zero and, of course, the mean of the residuals is also zero. The fact that the mean of the ANOVA residuals on the X variable is zero for each treatment group is the key to eliminating bias in the ANCOVA adjusted means that is introduced when treatments affect X.

Recall that the treatment effect in the ANOVA model is defined as $(\mu_j - \mu) = \alpha_j$, where μ_j is the population mean for treatment j and μ is the mean of all the

population means (i.e., $\frac{\mu_1+\mu_2+\cdots+\mu_J}{J} = \mu$). The adjusted treatment effect in ANCOVA is $\alpha_j - \beta_1\left(\bar{X}_j - \bar{X}..\right)$, so the difference between the unadjusted treatment effect and the adjusted treatment effect is $\beta_1\left(\bar{X}_j - \bar{X}..\right)$. Correspondingly, the unadjusted sample mean is $\bar{Y}_j$ and the adjusted sample mean is $\bar{Y}_j - b_w\left(\bar{X}_j - \bar{X}..\right)$. It can be seen that no adjustment takes place if the within-group slope is zero and/or the covariate means are equal.

The rationale for the adjustment is that the within-group regression slope is predictive of the amount of between-group regression toward the mean on Y that can be expected to occur in a randomized-group experiment. Differences between groups on the covariate are the result of chance and the adjustment provides a precise estimate of the mean on Y that would have occurred if the treatment covariate mean had been equal to the covariate grand mean. This works because, in the case of randomized two-group experiments, the expectation for the mean difference on the covariate is $E\left(\bar{X}_1 - \bar{X}_2\right) = 0$. But when the treatments affect the covariate the covariate mean differences are no longer explainable as resulting only from chance. In this case, the sample covariate mean differences are partly a function of the treatments and, therefore, $E\left(\bar{X}_1 - \bar{X}_2\right) \neq 0$. This means that the adjustment process can no longer be expected to provide unbiased estimates of the treatment effects. A large *systematic* difference between covariate means will result in a large bias in the adjustment if the slope is large.

As can be seen in the expression for the adjustment, the larger the difference between the covariate means and the larger the slope, the larger the amount of inappropriate adjustment. If there is absolutely no effect of the treatment on Y the expectation for the adjusted effect will not be zero; this is because the expectation for the adjustment is not zero when the treatment affects the covariate. That is, $E\left[b_w\left(\bar{X}_j - \bar{X}..\right)\right] \neq 0$ when the treatment affects the covariate and the slope is nonzero. Because the bias on the outcome Y stems from systematic differences between covariate means induced by the treatments, the solution is to remove differences between covariate means. This is exactly what occurs when the original covariate scores (X) are replaced by the within-group residuals from an ANOVA performed on X. The mean of the X residuals within each group is exactly zero. One consequence of this is that the "adjusted" means associated with the quasi-ANCOVA are identical to the unadjusted means because the adjustment is zero. This might lead one to conclude that the results of applying quasi-ANCOVA are the same as simply ignoring the covariate and just performing ANOVA on Y. Although the treatment parameter estimates are identical, the inferential results are quite different. This is demonstrated next.

13.3 COMPUTATIONAL EXAMPLE OF QUASI-ANCOVA

The computation of quasi-ANCOVA consists of two steps. First, compute the residuals of the ANOVA model applied to the covariate. Second, compute ANCOVA using the residuals from the first step as the covariate. These steps are carried out on the modified data mentioned in Section 13.1 (where treatments affected the covariate) using the *Minitab* commands shown below.

ANOVA on X to Get Residuals

The *X* variable that had been affected by the treatments (labeled as "New *X*1") was entered in column c5 and it was decided that column c7 would be used for the residuals of ANOVA applied to column c5. It can be seen below that the first command line requests a one-way ANOVA on column c5, the subcommand in the second line requests that the residuals be computed and assigned to column c7, and the third line requests the treatment-group means on the data in column c5.

Commands:

```
MTB > anova c5=Tx;
SUBC> resid c7;
SUBC> means Tx.
```

Output:

```
ANOVA: New X1 versus Tx

Factor  Type   Levels  Values
Tx      fixed       3  1, 2, 3

Analysis of Variance for New X1

Source  DF      SS     MS     F      P
Tx       2  1500.0  750.0  3.55  0.043
Error   27  5700.0  211.1
Total   29  7200.0

S = 14.5297   R-Sq = 20.83%   R-Sq(adj) = 14.97%

Means
Tx   N  New X1
1   10  54.000
2   10  69.000
3   10  54.000
```

Quasi-ANCOVA

Quasi-ANCOVA is a conventional ANCOVA in which the covariate consists of the residuals of ANOVA applied to the covariate *X*. Note below that the covariate is specified as column c7, which contains the *X* residuals.

Commands:

```
MTB > ancova Y=Tx;
SUBC> covariate c7;
SUBC> means Tx.
```

Output:

```
ANCOVA: Y versus Tx

Factor  Levels  Values
Tx           3  1, 2, 3

Analysis of Covariance for Y

Source      DF   Adj SS       MS      F      P
Covariates   1  1855.35  1855.35  28.70  0.000
Tx           2   420.00   210.00   3.25  0.055
Error       26  1680.65    64.64
Total       29  3956.00

S = 8.03992   R-Sq = 57.52%   R-Sq(adj) = 52.61%

Covariate          Coef  SE Coef      T      P
Resid ANOVA X1   0.5705    0.106  5.357  0.000

Adjusted Means
Tx   N       Y
1   10  30.000
2   10  39.000
3   10  36.000
```

Comparison of Quasi-ANCOVA with ANOVA

The "adjusted means" in the output shown above are actually unadjusted means because no mean adjustment takes place when the covariate means are equal, as they must be when the covariate consists of within-group residuals on X. It might seem, therefore, that there is no reason to use quasi-ANCOVA instead of ANOVA because both analyses are used to evaluate differences among exactly the same outcome means. But recall that conventional ANCOVA has two advantages over ANOVA: (1) adjustment of means to correct for chance differences between groups on X before the experiment begins and (2) increased precision resulting from a reduction in error variation. Although quasi-ANCOVA has no mean adjustment (because it is the wrong thing to do if the treatments affect the covariate), it retains the second advantage. This is the reason for the label "quasi."

Table 13.1 summarizes the results of quasi-ANCOVA, ANOVA, and other analyses applied to the same outcome data. Note the difference between the results of quasi-ANCOVA and ANOVA that are presented in the last two rows of the table. Although the means and the sums of squares for the treatment effects (listed under SS_{AT}) are identical, the error sum of squares for quasi-ANCOVA is less than half that found using ANOVA. This difference in estimated error variation greatly affects the resulting p-values: .05 for quasi-ANCOVA and .22 for ANOVA. The second row contains the

Table 13.1 Comparison of Results of ANCOVA Applied to Experiment Where Treatments Have No Effect on *X* (Line One) with ANOVA, ANCOVA, and Quasi-ANCOVA Where Treatments Have a Large Effect on *X*

Treatments Affect Covariate?	Analysis Method	SS_{AT}	SS_{error}	Mean Estimates	*p*-value
No	ANCOVA	707.99	1680.65	28.48 40.33 36.19	.010
Yes	ANCOVA	214.53	1680.65	32.85 33.30 38.85	.210
Yes	Quasi-ANCOVA	420.00	1680.65	30.00 39.00 36.00	.055
Yes	ANOVA	420.00	3536.00	30.00 39.00 36.00	.220

results of conventional ANCOVA applied to exactly the same data as were used for quasi-ANCOVA and ANOVA. It can be seen that the SS_{AT} for conventional ANCOVA is about half the size of the treatment sum of squares for quasi-ANCOVA, but the error SS is the same for these analyses. Because the conventional ANCOVA treatment-effect adjustment has removed much of the effect, the *p*-value is much larger than for quasi-ANCOVA.

The first row summarizes the results of ANCOVA applied to the same *Y* scores as in all of the other analyses, but the covariate was not affected by the treatments. Note that the sum of squares for the treatment effects is larger for this analysis than it is using quasi-ANCOVA applied to data where the covariate was affected by treatments. The reason for the difference is that the chance differences on the covariate result in somewhat larger differences among the adjusted means than exist among the unadjusted means. By chance the covariate mean for the first treatment group is larger than for the other groups so the outcome mean is adjusted downward (from 30.00 to 28.48); similarly, the second group covariate mean is smaller than the others so the mean is adjusted upward (from 39.00 to 40.33). The third group covariate mean is only slightly below the grand covariate mean so the adjustment is minor (from 36.00 to 36.19). This shows that the treatment effects estimated by ANCOVA in a randomized experiment where the treatment does not affect the covariate may differ from the treatment-effect estimate provided by quasi-ANCOVA (when applied to the same data). The reason is that quasi-ANCOVA does not adjust the means at all, even though adjustment is justified when the covariate is not affected by the treatments.

This leads to the conclusion that conventional ANCOVA is the method of choice when the treatments do not affect the covariate.

13.4 MULTIPLE QUASI-ANCOVA

If data have been collected on multiple covariates and treatments have affected all of them, a multiple quasi-ANCOVA is appropriate. Each covariate affected by treatments can bias conventional multiple ANCOVA. No new principles are involved in this case. The first step is to compute the residuals of fitting the ANOVA model to each covariate. Second, compute multiple ANCOVA using the residuals computed in step one as covariates. In Section 13.5, this approach is applied to multiple covariate data (originally presented in Chapter 10) that have been modified to demonstrate the effects of using multiple covariates affected by treatments. A listing of the original and modified data can be found at the end of this chapter.

If data have been collected on multiple covariates and some of them have been affected by treatments but others have not, it is appropriate to include them all in the analysis. The covariates that have been affected should be transformed to ANOVA residuals and those unaffected should be left in the original form.

13.5 COMPUTATIONAL EXAMPLE OF MULTIPLE QUASI-ANCOVA

Compute ANOVA Residuals on All Covariates

Two covariates are involved in this example. The first step is to compute ANOVA residuals on both covariates. This was illustrated above for the first covariate; the residuals for this covariate were assigned to *Minitab* worksheet column c7. The commands for computing the ANOVA residuals for the second covariate are shown below. It can be seen that they are assigned to column c8.

Commands:

```
MTB > Name c8 "RESI1"
MTB > Oneway 'X2' C1;
SUBC>   Residuals 'RESI1'.
```

Output:

```
One-way ANOVA: X2 versus Tx

Source  DF      SS    MS     F      P
Tx       2    0.60  0.30  0.05  0.947
Error   27  149.40  5.53
Total   29  150.00

S = 2.352   R-Sq = 0.40%   R-Sq(adj) = 0.00%
```

```
                              Individual 95% CIs For Mean Based on
                              Pooled StDev

Level   N   Mean  StDev  ------+--------+--------+--------+--
1      10  5.100  2.378     (--------------*--------------)
2      10  4.800  1.814  (--------------*--------------)
3      10  5.100  2.767     (--------------*--------------)
                          ------+--------+--------+--------+--
                               4.0      5.0      6.0      7.0

Pooled StDev = 2.352
```

Three analyses are shown below. The first one provides multiple ANCOVA results for the situation where the covariates are not affected by treatments; the second and third analyses provide multiple ANCOVA and quasi-ANCOVA results for the situation where both covariates are affected by treatments. A summary of the results of these analyses is provided in Table 13.2.

The worksheet is arranged with original covariates X_1 and X_2 in columns c2 and c3, the treatment-affected covariates are in columns c5 and c6, and the X_1 and X_2 ANOVA residuals are in columns c7 and c8.

Table 13.2 Comparison of Results of Multiple ANCOVA Applied Where Treatments Affect Neither X_1 Nor X_2 (Line One) with ANOVA, Multiple ANCOVA and Multiple Quasi-ANCOVA Where Treatments Have a Large Effect on both X_1 and X_2

Treatments Affect Covariates?	Analysis Method	SS_{AT}	SS_{error}	Mean Estimates	*p*-value
No	Multiple ANCOVA	624.74	972.44	28.98 40.21 35.81	.002
Yes	Multiple ANCOVA	44.48	972.44	33.93 34.25 36.82	.572
Yes	Multiple quasi-ANCOVA	420.00	972.44	30.00 39.00 36.00	.011
Yes	ANOVA	420.00	3536.00	30.00 39.00 36.00	.220

Multiple ANCOVA: Covariates Not Affected by Treatments

Commands:

```
MTB > ancova Y=Tx;
SUBC> covariates c2 c3;
SUBC> means Tx.
```

Output:

```
ANCOVA: Y versus Tx

Factor  Levels  Values
Tx           3  1, 2, 3

Analysis of Covariance for Y

Source      DF   Adj SS       MS      F      P
Covariates   2  2563.56  1281.78  32.95  0.000
Tx           2   624.74   312.37   8.03  0.002
Error       25   972.44    38.90
Total       29  3956.00

S = 6.23678   R-Sq = 75.42%   R-Sq(adj) = 71.49%

Covariate    Coef  SE Coef      T      P
X1         0.2766    0.108  2.571  0.016
X2         2.8349    0.664  4.267  0.000

Adjusted Means
Tx   N       Y
1   10  28.979
2   10  40.212
3   10  35.809
```

Multiple ANCOVA: Covariates Affected by Treatments

Commands:

```
MTB > ancova Y=Tx;
SUBC> covariates c5 c6;
SUBC> means Tx.
```

Output:

```
ANCOVA: Y versus Tx

Factor  Levels  Values
Tx           3  1, 2, 3
```

```
Analysis of Covariance for Y

Source      DF   Adj SS       MS      F      P
Covariates   2  2563.56  1281.78  32.95  0.000
Tx           2    44.48    22.24   0.57  0.572
Error       25   972.44    38.90
Total       29  3956.00

S = 6.23678   R-Sq = 75.42%   R-Sq(adj) = 71.49%

Covariate    Coef  SE Coef      T      P
New X1     0.2766    0.108  2.571  0.016
New X2     2.8349    0.664  4.267  0.000

Adjusted Means
Tx   N       Y
1   10  33.934
2   10  34.250
3   10  36.816
```

Multiple Quasi-ANCOVA: Covariates Affected by Treatments

Commands:

```
MTB > ancova Y=Tx;
SUBC> covariates c7 c8;
SUBC> means Tx.
```

Output:

```
ANCOVA: Y versus Tx

Factor  Levels  Values
Tx           3  1, 2, 3

Analysis of Covariance for Y

Source      DF   Adj SS       MS      F      P
Covariates   2  2563.56  1281.78  32.95  0.000
Tx           2   420.00   210.00   5.40  0.011
Error       25   972.44    38.90
Total       29  3956.00

S = 6.23678   R-Sq = 75.42%   R-Sq(adj) = 71.49%

Covariate         Coef  SE Coef      T      P
Resid ANOVA X1  0.2766    0.108  2.571  0.016
Resid ANOVA X2  2.8349    0.664  4.267  0.000
```

```
Adjusted Means
Tx    N        Y
1    10   30.000
2    10   39.000
3    10   36.000
```

The first two analysis of Table 13.2 illustrate the difference between multiple ANCOVA applied to the same outcome data when treatments do and do not affect the covariates. It can be seen in row one that when treatments do not affect the covariates the differences among adjusted means are large and the p-value is very small. In contrast, the second analysis shows the results when the treatments affect the covariates. Note that there is relatively little difference among the adjusted means and that the p-value is very large.

Results displayed below the first analysis provide direct comparisons of three competing analyses (multiple ANCOVA, multiple quasi-ANCOVA, and ANOVA) where multiple covariates are affected by treatments. The inferential results of neither multiple ANCOVA nor ANOVA suggest that there is a treatment effect, whereas quasi-ANCOVA provides strong inferential support for a treatment effect ($p = .01$). The error sum of squares for the most typical analysis of this type of data (ANOVA) is about four times the size of the error sum of squares associated with quasi-ANCOVA.

The overall pattern of results from the different analyses shown in Table 13.2 is the same as is shown in Table 13.1 but the additional covariate has increased the adjustments to both between-group and within-group variation. For example, the error sums of squares for ANOVA, quasi-ANCOVA, and multiple quasi-ANCOVA are 3536, 1680.65, and 974.44, respectively. There is a message here for the analysis of designs where the covariates are affected by treatments: To adopt ANOVA when good covariates are available (for quasi-ANCOVA) is to throw away valuable information that can have a dramatic effect on the inferential outcome of the study.

13.6 SUMMARY

Designs in which the covariate(s) are affected by the treatments should not be analyzed using ANCOVA. The covariate will usually remove a portion of the treatment effect if ANCOVA is used in this situation. But this does not mean that ANOVA (which ignores the covariates) is the appropriate alternative analysis. Predictive covariates can dramatically reduce the error sum of squares and for this reason they are useful in the appropriate analysis. Quasi-ANCOVA is the recommended alternative. It involves a simple modification of conventional ANCOVA that eliminates the between-groups adjustment that causes bias (when treatments affect covariates) but retains the within-group adjustment that increases precision. This analysis can be performed using conventional ANCOVA routines by substituting residuals of ANOVA performed on the covariates for the original covariates.

Data Listing

The variables used in the analyses described in this chapter are listed below:

Row	Tx	X1	X2	Y	New X1	New X2	Resid ANOVA X1	Resid ANOVA X2
1	1	29	3	15	31	3.2	-23	-2.1
2	1	49	3	19	51	3.2	-3	-2.1
3	1	48	2	21	50	2.2	-4	-3.1
4	1	35	5	27	37	5.2	-17	-0.1
5	1	53	5	35	55	5.2	1	-0.1
6	1	47	9	39	49	9.2	-5	3.9
7	1	46	3	23	48	3.2	-6	-2.1
8	1	74	7	38	76	7.2	22	1.9
9	1	72	6	33	74	6.2	20	0.9
10	1	67	8	50	69	8.2	15	2.9
11	2	22	3	20	44	5.1	-25	-1.8
12	2	24	2	34	46	4.1	-23	-2.8
13	2	49	4	28	71	6.1	2	-0.8
14	2	46	4	35	68	6.1	-1	-0.8
15	2	52	5	42	74	7.1	5	0.2
16	2	43	4	44	65	6.1	-4	-0.8
17	2	64	8	46	86	10.1	17	3.2
18	2	61	7	47	83	9.1	14	2.2
19	2	55	6	40	77	8.1	8	1.2
20	2	54	5	54	76	7.1	7	0.2
21	3	33	2	14	38	3.3	-16	-3.1
22	3	45	1	20	50	2.3	-4	-4.1
23	3	35	5	30	40	6.3	-14	-0.1
24	3	39	4	32	44	5.3	-10	-1.1
25	3	36	3	34	41	4.3	-13	-2.1
26	3	48	8	42	53	9.3	-1	2.9
27	3	63	8	40	68	9.3	14	2.9
28	3	57	4	38	62	5.3	8	-1.1
29	3	56	9	54	61	10.3	7	3.9
30	3	78	7	56	83	8.3	29	1.9

CHAPTER 14

Robust ANCOVA/Robust Picked Points

14.1 INTRODUCTION

Occasionally data are encountered that contain one or more unexplained outliers and/or severe departures from normality. Although ANCOVA is affected little by many types of nonnormality, there are situations where this is not true. Unequal sample sizes combined with treatment distributions that differ greatly in terms of skewness and variance seem to be particularly problematic. Even if the response distributions are essentially identical, a single outlier can completely invalidate an analysis in the sense that both parameter estimates and inference are greatly distorted. Examples of these problems are described in this chapter along with brief descriptions of two methods for contending with them.

14.2 RANK ANCOVA

In education and behavioral sciences, the most frequently recommended alternative to conventional parametric ANCOVA is a method proposed by several methodologists years ago but usually attributed to Conover and Inman (1982). It involves (1) combining groups; (2) transforming X to the rank on X, transforming Y to the rank on Y; and (3) applying a conventional ANCOVA using ranked X as the covariate and ranked Y as the dependent variable. The main advantages of this approach are that it is easy to carry out, easy to explain, and it is very effective when the original data contain one or more outliers. A demonstration of these properties is shown next.

The data from the first ANCOVA example presented in Chapter 6 are repeated below along with the *Minitab* commands to transform the data, and the results of the rank ANCOVA:

The Analysis of Covariance and Alternatives: Statistical Methods for Experiments, Quasi-Experiments, and Single-Case Studies, Second Edition. Bradley E. Huitema.

Commands used to transform the data to ranks:

```
MTB > Let 'Rank X' = RANK(X)
MTB > Let 'Rank Y' = RANK(Y)

MTB >  print TX X Y 'Rank X' 'Rank Y'
```

Data Display

Row	TX	X	Y	Rank X	Rank Y
1	1	29	15	3.0	2.0
2	1	49	19	16.5	3.0
3	1	48	21	14.5	6.0
4	1	35	27	5.5	8.0
5	1	53	35	19.0	15.5
6	1	47	39	13.0	19.0
7	1	46	23	11.5	7.0
8	1	74	38	29.0	17.5
9	1	72	33	28.0	12.0
10	1	67	50	27.0	27.0
11	2	22	20	1.0	4.5
12	2	24	34	2.0	13.5
13	2	49	28	16.5	9.0
14	2	46	35	11.5	15.5
15	2	52	42	18.0	22.5
16	2	43	44	9.0	24.0
17	2	64	46	26.0	25.0
18	2	61	47	24.0	26.0
19	2	55	40	21.0	20.5
20	2	54	54	20.0	28.5
21	3	33	14	4.0	1.0
22	3	45	20	10.0	4.5
23	3	35	30	5.5	10.0
24	3	39	32	8.0	11.0
25	3	36	34	7.0	13.5
26	3	48	42	14.5	22.5
27	3	63	40	25.0	20.5
28	3	57	38	23.0	17.5
29	3	56	54	22.0	28.5
30	3	78	56	30.0	30.0

Commands used to perform the rank ANCOVA:

```
MTB > ancova 'Rank Y'=TX;
SUBC> covariate 'rank X';
SUBC> means TX.
```

Output for Rank ANCOVA:

```
ANCOVA: Rank Y versus TX
Factor  Levels  Values
TX         3    1, 2, 3

Analysis of Covariance for Rank Y
Source        DF     Adj SS      MS       F       p
Covariates     1    1080.87   1080.87   31.17   0.000
TX             2     364.37    182.18    5.25   0.012
Error         26     901.53     34.67
Total         29    2244.00

Adjusted Means
TX     N   Rank Y
1     10   10.863
2     10   19.318
3     10   16.318
```

The *p*-value from this analysis is very close to the value found using parametric ANCOVA but the "adjusted means" are very different because these are actually adjusted mean ranks. Some researchers consider this change in interpretation a major disadvantage of the method, especially if the original response measure was chosen because it is widely understood.

The potential strength of the rank ANCOVA method over conventional ANCOVA is demonstrated next. Let's return to the data analyzed above and change the last observation in the last group from 56 to 1156 so that we now have an obvious outlier. If we now rerun both the conventional parametric ANCOVA and the rank ANCOVA on the data that now contain the outlier, the *p*-values are .28 and .01, respectively. The conventional ANCOVA adjusted means associated with the data set containing the outlier are all greatly affected; they are: 13.55, 53.39, and 148.06 rather than 28.48, 40.33, and 36.19 found with the original data. Note that all three adjusted means are affected even though the outlier is only in the third group. The distortion in the first and third groups is massive. There is no individual score in the first group that is as small as the adjusted mean for this group and there is no individual score in the third group that is anywhere near the adjusted mean for this group. Even though the difference between the largest and smallest adjusted means is about 10 times as large as the maximum difference in the analysis without the outlier, the *p*-value is much larger. This occurs because the mean square error with the outlier included is 47,569 rather than 354 found in the analysis without the outlier. Fortunately, even if the researcher does not note the outlier before performing the analysis, the conventional ANCOVA diagnostic procedures shout in unison. The normal probability plot has essentially no points on the expected line, the unusual residual stands out in the residual plot, and Cook's distance is huge (2.13).

14.3 ROBUST GENERAL LINEAR MODEL

Because rank ANCOVA is not affected by the outlier problem illustrated in the previous section, one might guess that I strongly recommend it. Although it has performed admirably in the example, I prefer a robust linear model approach (based on so-called Wilcoxon estimation) instead. Unlike the simple rank transformation approach, the robust linear model approach is a unified methodology that includes important model diagnosis procedures that are similar to those used with OLS linear models. Details of this specific robust approach are not covered here, but they can be found in the work of Hettmansperger and McKean (1998), McKean (2004), McKean et al. (1999), McKean et al. (1993), and McKean and Vidmar (1994). Software for performing the analysis is called *RGLM*. It was developed by Terpstra and McKean (2004) and is available at http://www.stat.wmich.edu/mckean/.

Two concepts that are used to explain why the robust linear model approach performs better than conventional OLS approaches in the presence of outliers and distribution departures from normality are the influence function and the breakdown point (see, e.g., McKean, 2004). The influence function is a measure of the change in an estimator (such as a mean or a regression coefficient) that occurs when an outlier is added to the data. Least-squares estimators are seriously affected by outliers. In fact the influence of an outlier (on either a predictor or the dependent variable) is said to be unbounded because the regression coefficient is increasingly influenced by an outlier as it becomes more and more extreme. In contrast, the influence of an outlier on the robust estimator is bounded.

The breakdown point refers to the proportion of bad data the estimator can tolerate before becoming meaningless. Consider the sample mean. For a sample of size n, if one observation is moved to plus infinity, the breakdown point of $1/n$ converges to zero as n increases to infinity. The mean is meaningless with a single outlier in this case because it moves to infinity. But in the case of the median, the breakdown point is .5 because half the data must be moved to infinity before the median moves to infinity and becomes meaningless. Much work in robust analysis is directed toward the development of estimators with high breakdown points. Two robust ANCOVA methods are discussed here. It will be shown that there are important differences between them.

Applications of the robust linear model ANCOVA to two examples of sample data will clarify some of the advantages of this procedure relative to rank ANCOVA. First, the example data just analyzed using rank ANCOVA are used to point out one advantage. Then I use data mentioned by McKean and Vidmar (1994) in a second example that illuminates a largely unrecognized weakness of the rank ANCOVA method.

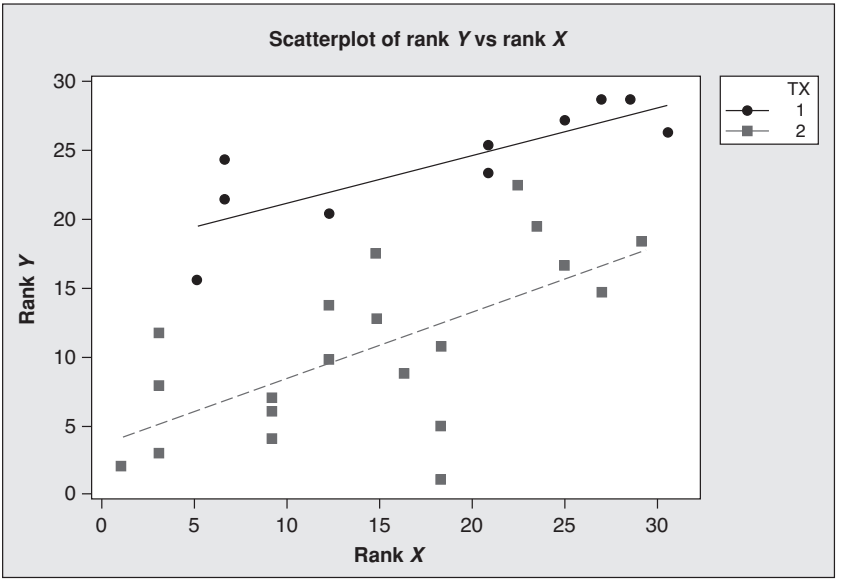
Example 14.1 The first example involves the data with the large outlier (analyzed above using traditional ANCOVA and rank ANCOVA). Briefly, it produces a p-value of .03; so like rank ANCOVA it results in the rejection of the null hypothesis whereas traditional ANCOVA does not. But unlike rank ANCOVA, the robust regression approach provides descriptive results that are in the metric of the original dependent variable. Whereas rank ANCOVA provides adjusted mean ranks, the *RGLM* estimate refers to adjusted means, which are almost always more informative. This facilitates

interpretation, allows reporting results using confidence intervals and tests relevant to the original metric, and leads to multiple comparison and picked-points generalizations that are straightforward. These generalizations are not possible using simple rank transformations.

Example 14.2 The second example is based on data from Shirley (1981) who carried out a three-group experiment to evaluate the effects of antidotes on the time it takes rats to enter an experimental chamber. I have combined data from the second and third groups (because they are essentially identical) to produce the two-group (control vs. antidote) data set shown below.

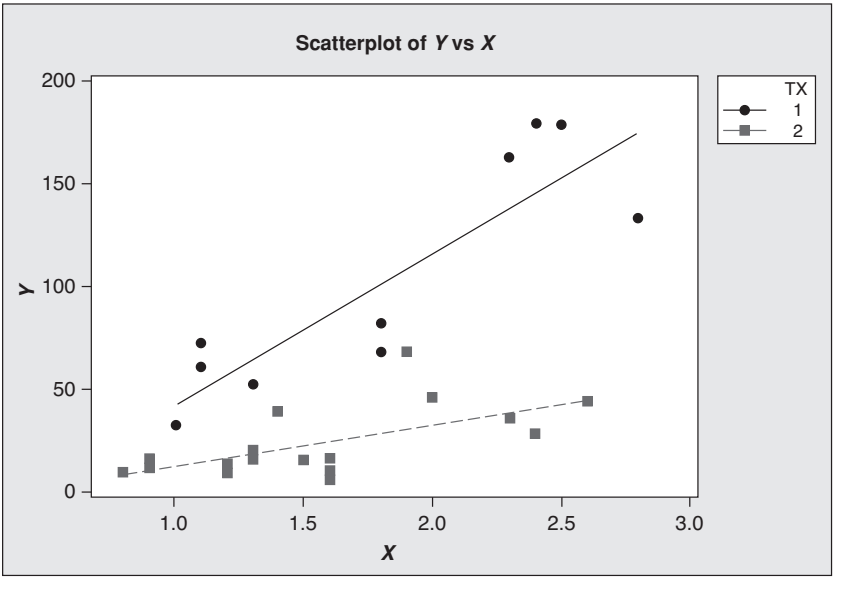
Row	*TX*	*X*	*Y*
1	1	1.8	79.1
2	1	1.3	47.6
3	1	1.8	64.4
4	1	1.1	68.7
5	1	2.5	180.0
6	1	1.0	27.3
7	1	1.1	56.4
8	1	2.3	163.3
9	1	2.4	180.0
10	1	2.8	132.4
11	2	1.6	10.2
12	2	0.9	3.4
13	2	1.5	9.9
14	2	1.6	3.7
15	2	2.6	39.3
16	2	1.4	34.0
17	2	2.0	40.7
18	2	0.9	10.5
19	2	1.6	0.8
20	2	1.2	4.9
21	2	1.3	14.8
22	2	2.3	30.7
23	2	0.9	7.7
24	2	1.9	63.9
25	2	1.2	3.5
26	2	1.3	10.0
27	2	1.2	6.9
28	2	2.4	22.5
29	2	1.4	11.4
30	2	0.8	3.3

The data were transformed to ranks and the following plot illustrates those transformed values.



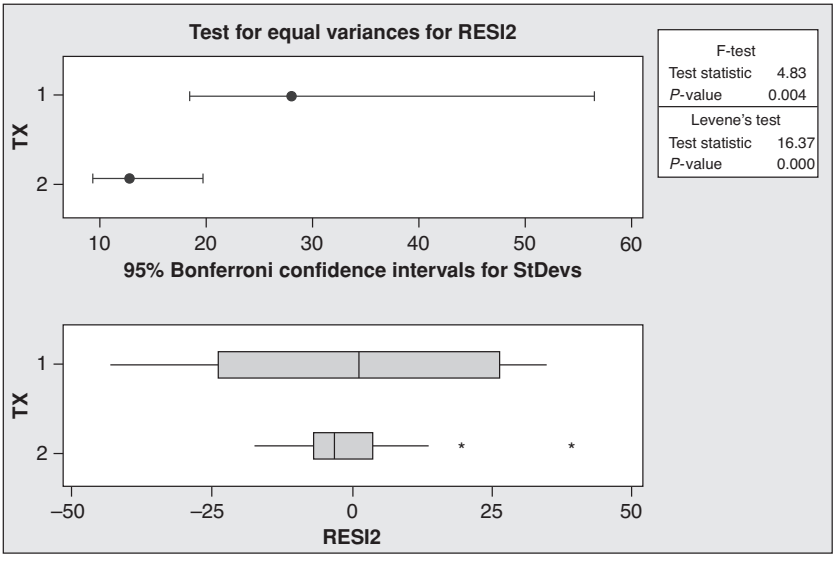
Conventional ANCOVA and the test for homogeneity of regression slopes were then carried out on these ranks. The p-value for the latter test is .49, which supports the adequacy of rank ANCOVA and implies that the effect of the treatments is not a function of the covariate. In this situation it appears that rank ANCOVA ($p < .001$) is quite satisfactory.

But a crucial point is missed if we settle on this analysis. *Ranks obscure the nature of the original data*. A plot of the original X and Y scores is shown below.



A comparison of the two plots makes it apparent that the striking heterogeneity of slopes in the original data is completely obliterated by ranking. The ranked data clearly do not convey the critical fact that the treatment effect depends on the level of the covariate. When the test for homogeneous slopes is applied to the original data, the resulting $F = 17.70$ ($p \leq .001$). The rank ANCOVA completely misses this point. This does not mean, however, that the conventional alternatives for accommodating heterogeneous slopes should be adopted.

A complete analysis of the residuals from fitting a separate slope to each group reveals two departures from ANCOVA assumptions (in addition to heterogeneity of slopes). First, the conditional variances are heterogeneous, as can be seen in the following plot:



Second, there is convincing evidence of an outlier in the second group. A formal outlier test involves comparing the *t*-residual statistics (provided as an option when running regression using *Minitab*) for observations in the second group with the Bonferroni *t* critical value based on the number of observations in the group ($n_2 = 20$) and $n_2 - 2$ *df*. The largest *t*-residual is 4.41 and the critical Bonferroni $t = 4.23$ for $\alpha = .01$. This is strong evidence that this observation is an outlier. The second largest *t*-residual is only 1.59, so it appears that there is only one outlier. Because (1) the slopes are heterogeneous, (2) the variances are heterogeneous (along with unequal sample size), and (3) an outlier is present, it is reasonable to estimate a picked-points model using robust regression.

Robust Picked-Point Analysis

It can be seen in Table 14.1 that three points on *X* were picked and that *RGLM* provides results from both OLS and robust estimation. The predictor variables were

Table 14.1 Output from Program *RGLM* Illustrating Robust Wilcoxon and Ordinary Least-Squares Results for Picked-Points Analysis of Shirley Data. Points Were Picked to Estimate the Treatment Effect at (1) One SD Below the *X* Mean (1.026), (2) the *X* Mean (1.603), and (3) One SD Above the *X* Mean (2.18).

Picked Point = 1.026
Wilcoxon

Column of *X*	Beta	Std. Errors	*T*-Ratio
1	.533469E+01	.501223E+01	.106433E+01
2	.278221E+02	.102273E+02	.272039E+01
3	.184502E+02	.770648E+01	.239411E+01
4	.666191E+02	.116106E+02	.573779E+01

The final full model dispersion is .530641E+03.
The estimate of scale is .173010E+02.
Scale estimate tauhat-star is .142725E+02.
LS

Column of *X*	Beta	Std. Error	*T*-Ratio
1	.663031E+01	.607386E+01	.109161E+01
2	.330728E+02	.116750E+02	.283279E+01
3	.210437E+02	.879737E+01	.239204E+01
4	.557636E+02	.132541E+02	.420727E+01

The final full model dispersion is .101417E+05.
The estimate of scale is .197501E+02.
R-squared .876418E+00.
Picked Point = 1.603 = $\bar{X}$
Wilcoxon

Column of *X*	Beta	Std. Error	*T*-Ratio
1	.160415E+02	.349464E+01	.459031E+01
2	.662237E+02	.698355E+01	.948281E+01
3	.185715E+02	.770725E+01	.240961E+01
4	.665164E+02	.116117E+02	.572838E+01

The final full model dispersion is .530635E+03.
The estimate of scale is .173027E+02.
Scale estimate tauhat-star is .140633E+02.
LS

Column of *X*	Beta	Std. Error	*T*-Ratio
1	.187725E+02	.450825E+01	.416403E+01
2	.652484E+02	.797131E+01	.818540E+01
3	.210437E+02	.879737E+01	.239204E+01
4	.557636E+02	.132541E+02	.420727E+01

Table 14.1 (*Continued*)

The final full model dispersion is .101417E+05.
The estimate of scale is: .197500E+02.
R-squared .876418E+00.
Picked Point = 2.18
Wilcoxon

Column of *X*	Beta	Std. Error	*T*-Ratio
1	.267753E+02	.619289E+01	.432355E+01
2	.104600E+03	.899959E+01	.116227E+02
3	.185882E+02	.762720E+01	.243710E+01
4	.664237E+02	.114911E+02	.578043E+01

The final full model dispersion is .530638E+03.
The estimate of scale is .171230E+02.
LS

Column of *X*	Beta	Std. Error	*T*-Ratio
1	.309147E+02	.743573E+01	.415758E+01
2	.974240E+02	.103803E+02	.938546E+01
3	.210436E+02	.879737E+01	.239204E+01
4	.557636E+02	.132541E+02	.420727E+01

The final full model dispersion is .101417E+05.
The estimate of scale is .197500E+02.

constructed exactly as shown in Chapter 11 for conventional picked-points analysis. The second estimated "beta" shown in the output is the treatment effect estimate.

The estimates of the treatment effect at the three picked points on X are 27.82, 66.22, and 104.60. The corresponding t-values are 2.72, 9.48, and 11.62. Consequently, it can be stated that there is strong evidence of a treatment effect at each of the picked values of X. Because standard error estimates are provided confidence intervals can be easily computed.

Efficiency

A measure of the efficiency of the robust method relative to OLS is provided by the ratio of squares of "scale" estimates shown in the output. This is essentially the ratio of the error mean squares for the two types of estimation: OLS scale/Wilcoxon scale $= \frac{19.75^2}{17.30^2} = 1.30$. This indicates that the robust method is considerably more efficient in the sense that OLS has a mean square error estimate that is 30% larger than the robust error estimate.

Multiple Comparison Tests for Robust ANCOVA

In the case of a two-group robust ANCOVA, the regression model is estimated using only a dummy variable column and a covariate column, so there are three coefficients

including the intercept; the one associated with the dummy variable is the robust adjusted mean difference. When three or more groups are involved the regression approach to ANCOVA is carried out using the same general approach described in Chapter 7, but using robust regression.

There may be interest in performing multiple comparison tests following a robust linear model ANCOVA with three or more groups. Recall that multiple comparison tests require: (1) the adjusted means, (2) the adjusted mean square error (MS_{Res_w}), (3) the difference between the comparison group covariate means, and (4) the sum of squares within groups on the covariate. A robust analog of each of the four is available in a version of *RGLM* under development. But output from the version illustrated in Table 14.1 provides enough information for an adequate test.

The robust adjusted means can be computed using the approach described for least-squares ANCOVA in Chapter 7 (Section 7.4), except that the coefficients of the equation are based on the fitted robust linear model. The estimate of scale from the same model used to estimate the adjusted means (i.e., the regression of Y on dummy variables and the covariate) is substituted for MS_{Res_w}. The covariate means and the within-group sum of squares on the covariate can be computed in the usual way, unless there are covariate outliers. In this case the covariate means and within-group sum of squares based on robust estimates are more appropriate.

14.4 SUMMARY

Robust methods have advantages over conventional ANCOVA when severe nonnormality and/or outlier problems are a concern. Two robust ANCOVA methods that are very efficient in the presence of outliers are described. Rank ANCOVA is simple to compute using conventional ANCOVA or regression software, but it does not provide results that adequately describe the nature of the original data. The robust linear model approach provides many advantages in addition to efficiency. It produces descriptive results that capture the nature of the original data, provides confidence intervals in the original metric, can be used with minor adjustments to compute multiple comparison tests, and is easily generalized to picked-points analysis. Major gains in the adequacy of both descriptive statistics and power relative conventional ANCOVA are likely when outliers exist.

CHAPTER 15

ANCOVA for Dichotomous Dependent Variables

15.1 INTRODUCTION

It is common to encounter dependent variables that are not approximately continuous. Examples can be found in all research areas. A behavioral science researcher may want to know if social support systems have an effect on whether students graduate, a pharmaceutical researcher may want to know if two drugs have differential effects on patient survival, a manager may want to investigate the effects of a new production process on having an accident, and a medical market researcher may want to know if potential customers exposed to various types of information buy a product. Each of these examples has a dependent variable that is dichotomous (i.e., graduate vs. do not graduate, survive vs. do not survive, accident vs. no accident, and purchase vs. no purchase). Variables of this type lead to descriptive and inferential problems when analyzed using standard ANCOVA or corresponding regression methods.

The two categories of the dichotomous dependent variable are usually assigned the values zero and one. If conventional regression analysis is used to estimate the ANCOVA model when Y is a 0–1 outcome and the predictors are dummy variables (used to identify treatment conditions) and covariates, three problems will be encountered. They are

1. The values predicted from the equation may be outside the possible range for probability values (i.e., zero through one).
2. The homoscedasticity assumption will be violated because the variance on Y will differ greatly for various combinations of treatments and covariate scores.
3. The error distributions will not be normal.

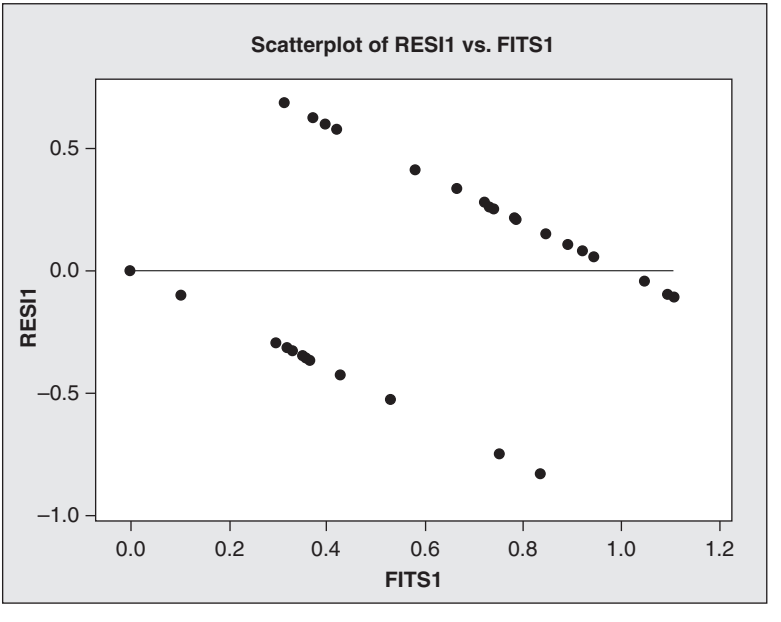
These problems are easily demonstrated.

The Analysis of Covariance and Alternatives: Statistical Methods for Experiments, Quasi-Experiments, and Single-Case Studies, Second Edition. Bradley E. Huitema.

The achievement outcome data presented in Example 6.1 (Chapter 6) are approximately continuous, but they can be modified (for purposes of this chapter) to become dichotomous. This was accomplished by changing each achievement score to 0 if the original value was 33 or less, and to 1 if the original value was equal to or greater than 34. Hence, each subject was classified as either unsuccessful (0) or successful (1) with respect to achievement. This 0–1 dependent variable was then regressed (using OLS) on the group-membership dummy variables and the covariate in order to estimate the parameters of the ANCOVA model.

After the model was fitted the equation was used to predict the probability of academic success for each of the 30 subjects. That is, dummy variable and covariate scores for each subject were entered in the equation and $\hat{Y}$ was computed for each subject. It turned out that one of the predicted values was negative and one was greater than one. This is an undesirable property for a procedure that is intended to provide a probability value. But this is not the only problem with the analysis.

Recall that the conventional approach for identifying departures from ANCOVA assumptions involves inspecting the residuals of the fitted model. The residuals shown below are from Example 6.1.



It is obvious that the residuals cannot be characterized as random. It can be seen that the variation of the residuals around the mean of zero is small at the lower and upper ends of the "Fits" distribution and very large in the middle. This pattern can be expected when a conventional OLS model is fitted to a dichotomous outcome. Problems other than heterogeneity of variance are also apparent (e.g., departures from normality). Because these problems can be anticipated when OLS models are applied to studies using binary (dichotomous) dependent variables some alternative is usually sought.

15.2 LOGISTIC REGRESSION

Although several alternatives are available to analyze dichotomously scaled outcomes, by far the most frequently used approach is to apply logistic regression modeling. Logistic models have much in common with ordinary least squares models but the correspondence is clouded by statistics, terminology, and estimation procedures that are unfamiliar. The unfamiliarity begins with the nature of the quantity that is modeled.

The focus of research that uses a dichotomous outcome variable is on the probability that some event (such as graduating from high school or having a heart attack) will occur. The population probability of the event is often denoted as π. Although the value of this parameter is what we want to know, there are problems if we attempt to model it directly. It might seem that we should be able to easily estimate the following regression model:

$$\pi_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \cdots + \beta_m X_{mi}$$

It would be convenient if we could do so; unfortunately this is impossible. If it were possible it would be misspecified. The reason it is impossible is that we do not have a column of πs. That is, we cannot regress π on the predictors because the outcome score we have for each subject is not π; rather, the outcome score available is a zero or a one. Even if we had the π, we would discover that it is not a linear function of the parameters; instead it is a nonlinear function of the β parameters. This implies that neither conventional multiple regression nor polynomial regression qualify. Rather, a model of π that is nonlinear in the parameters is needed.

Fortunately, we need not give up on the conceptual convenience of the linear model. There is a simple way of transforming π to a value that *is* a linear function of the β parameters. The transformation is often called the logit link function, which is often denoted by $g(\pi)$ or logit (π). It is unfortunate that the term "logit" has caught on, because there is a more meaningful term that actually describes what it is; the alternative is "log-odds." The latter term makes sense if the term "odds" (or, to be more specific, "odds ratio") is understood.

Odds Ratio

The odds of an event occurring is defined as $\frac{\pi}{1-\pi}$. This is simply the population proportion (or probability) of a "1" response (the event occurred) divided by the proportion of "0" responses (the event did not occur). For example, if the proportion of subjects who pass a test is .75 the odds ratio is $\frac{.75}{.25} = 3$. The odds ratio for the occurrence of an event has properties that are very different than the properties of the proportion. One problem with proportions is that the substantive implications of a given difference in proportions that fall near .50 are often very different than that of the same difference falling near zero or one.

Suppose a study comparing two treatments finds that a difference in the proportion of patients who survive is $(.55 - .45) = .10$, and a second study finds a difference

of (.15 − .05) = .10. Although the treatment effect in each study is an *absolute* difference in proportions of .10, the *relative* improvement in the two studies is quite different. The first treatment in the first study resulted in a relative increase .10/.45 = 22%, whereas the first treatment in the second study resulted in a relative increase of .10/.05 = 200%.

Whereas the proportion is confined to values zero through one, the odds ratio ranges from $-\infty$ to $+\infty$. This important property has implications for justifying the assumptions of the logistic model, which models the logit.

Logit

The population logit is defined as $\log_e(\frac{\pi}{1-\pi})$. It can be seen that this is simply the log (using base e) of the odds ratio; the less popular term "log-odds" is certainly more descriptive than logit, but I stick with the more popular term in the remainder of the chapter. The sample estimate of the logit computed for subject i is denoted as

$$\log_e\left(\frac{\hat{\pi}_i}{1-\hat{\pi}_i}\right)$$

15.3 LOGISTIC MODEL

The logistic regression model can be written as

$$\log_e\left(\frac{\pi}{1-\pi}\right) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_m X_m,$$

where $X_1, X_2, \ldots, X_m$ are the predictor variables.

Note that $\log_e(\frac{\pi}{1-\pi})$ is modeled as a linear function of the predictors. This model is one of a family of models known as generalized linear models. It is not possible to estimate the parameters of this model using OLS procedures. Instead, the estimation is carried out using the method of maximum likelihood. Software for maximum likelihood estimation of this model is widely available. The *Minitab* binary logistic regression routine is used in subsequent examples.

Probability Estimation

After the parameters are estimated there is often interest in computing the probability of an "event" for certain subjects or groups. Suppose a study uses a dichotomous variable where the event is experiencing a heart attack. If this event is scored as 1 and not having a heart attack is scored as 0, the probability of being a 1 given certain scores on the predictors is likely to be of interest. The probability estimates

are computed using

$$\hat{\pi}_i = \frac{e^{b_0+b_1X_{1i}+b_2X_{2i}+\cdots+b_mX_{mi}}}{1+e^{b_0+b_1X_{1i}+b_2X_{2i}+\cdots+b_mX_{mi}}}$$

This is sometimes called the logistic function.

15.4 DICHOTOMOUS ANCOVA THROUGH LOGISTIC REGRESSION

I now return to the example described at the end of Section 15.1. Dichotomous ANCOVA can be computed on 0–1 outcome data via logistic regression; the approach is very similar to the method used to compute conventional ANCOVA through OLS regression. The software and the details are different but the general ideas are the same.

The first step is to regress the dichotomous dependent variable on the dummy variables and the covariate using logistic regression. Second, regress the dichotomous dependent variable on only the covariate. The output from both steps is shown below for the three-group example described in Section 15.1.

```
Binary Logistic Regression: Dichotomous Y versus D1, D2, X1
Link Function: Logit
Response Information
Variable       Value  Count
Dichotomous Y  1      18    (Event)
               0      12
               Total  30
Logistic Regression Table
                                              Odds    95% CI
Predictor     Coef    SE Coef     Z      P Ratio Lower  Upper
Constant  -4.32031   1.92710 -2.24 0.025
D1        -1.49755   1.14341 -1.31 0.190  0.22  0.02   2.10
D2         1.42571   1.23264  1.16 0.247  4.16  0.37  46.60
X1        0.101401 0.0400053  2.53 0.011  1.11  1.02   1.20

Log-Likelihood = −13.966
Test that all slopes are zero: G = 12.449, DF = 3, P-Value
= 0.006


Binary Logistic Regression: Dichotomous Y versus X1
Link Function: Logit
Response Information
Variable       Value  Count
Dichotomous Y  1      18     (Event)
               0      12
               Total  30
```

```
Logistic Regression Table
                                                    Odds    95% CI
Predictor       Coef    SE Coef      Z      P  Ratio  Lower  Upper
Constant   -3.38749    1.75255  -1.93  0.053
X1        0.0790462  0.0362325   2.18  0.029   1.08   1.01   1.16

Log-Likelihood = −16.993
Test that all slopes are zero: G = 6.394, DF = 1, P-Value
= 0.011
```

Note that like OLS regression output, there is a column heading for predictors, coefficients, and standard errors for the predictors. Unlike OLS output, there is neither a column of t-values nor an ANOVAR summary table with the F-test for the multiple regression. Instead we find z, p, the odds ratio, the 95% confidence interval on the odds ratio, the log-likelihood, and a G-statistic along with the related degrees of freedom and p-value. The z- and p-values are direct analogs to the t- and p-values in OLS, and the G-statistic is the analog to the ANOVAR F. The G-value can be interpreted as a chi-square statistic. The log-likelihood is related to the notion of residual variation but it will not be pursued in this brief introduction. Additional detail on logistic regression is available in the excellent work of Hosmer and Lemeshow (2000).

A very convenient property is associated with the two G-values shown above. Denote the first one as $G_{(D_1,D_2,X)}$; it is associated with three predictors (D_1, D_2, and the covariate X) in this example. Denote the second one as $G_{(X)}$; it is associated with only one predictor (the covariate X).

The difference $(G_{(D_1,D_2,X)} - G_{(X)}) = X^2_{AT}$. This chi-square statistic is used to test for adjusted treatment effects. The null hypothesis can be written as: $\mathrm{H}_0 : \pi_{1\,\mathrm{adj}} = \pi_{2\,\mathrm{adj}} = \cdots = \pi_{j\,\mathrm{adj}}$, where the $\pi_{j\,\mathrm{adj}}$ are the adjusted population probabilities of a "1" response. This hypothesis is the analog to the hypothesis tested using conventional ANCOVA, and the chi-square statistic is the analog to the conventional ANCOVA F-test. The G-values and the associated degrees of freedom described in the output shown above are summarized in Table 15.1.

The p-value associated with an obtained chi-square of 6.100 with two degrees of freedom is .047. This is implies that at least one of the three adjusted group probabilities of academic success differs from the others.

The approach shown in this example generalizes to any number of treatment groups. If there are two treatments, the first regression includes one dummy variable and one covariate; the second regression includes only the covariate. Similarly, if

Table 15.1 Omnibus Test for Adjusted Treatment Effects in an Experiment with Three Treatment Groups, One Covariate, and a Dichotomous Outcome Variable

Predictors in the Model	G-Statistic	Degrees of Freedom
D_1, D_2, X	$G_{(D_1,D_2,X)} = 12.449$	3
X	$G_{(X)} = 6.349$	1
	Difference $= 6.100 = \chi^2_{\mathrm{AT}}$	Difference $= 2$

there are four groups, the first regression includes three dummy variables and the covariate; the second regression includes only the covariate.

Adjusted Probabilities (or Proportions)

As mentioned previously, the chi-square test shown above is a test on differences among groups with respect to adjusted probabilities of the event "1" occurring. The adjusted probability for group j is computed by entering (a) the group membership dummy variable scores associated with group j and (b) the grand means of covariates X_1 through X_C in the fitted equation shown below:

$$\hat{\pi}_{j\,\text{adj}} = \frac{e^{b_0+b_1d_1+\cdots+b_{J-1}d_{J-1}+b\bar{X}_{1..}+\cdots+b\bar{X}_{C..}}}{1+e^{b_0+b_1d_1+\cdots+b_{J-1}d_{J-1}+b\bar{X}_{1..}+\cdots+b\bar{X}_{C..}}}$$

Minitab will compute these adjusted probabilities. When entering menu commands for logistic regression, click on "Predictions" and enter the appropriate dummy variable scores and grand covariate means in the window below "Predicted event probabilities for new observations." The corresponding command line editor commands used to compute the adjusted mean probability for the first treatment group in the example study are listed below.

```
MTB > Blogistic 'Dichotomous Y' = D1 D2 X1;
SUBC> Logit;
SUBC> Eprobability 'EPRO4';
SUBC> Brief 1;
SUBC> Predict 1 0 49.333;
SUBC> PEProbability 'PEProb4'.
```

The complete logistic regression analysis output (not shown here) appears; the last portion of the results contains the adjusted mean probability for group 1 (labeled as a predicted event probability). It can be seen below that $\hat{\pi}_{1\,\text{adj}} = .3067$. The last line of output confirms that the values entered for dummy variables and the grand covariate mean are 1, 0, and 49.333.

Output

```
Predicted Event Probabilities for New Observations
New Obs      Prob   SE Prob         95% CI
      1  0.306740 0.170453 (0.0842125, 0.680404)

Values of Predictors for New Observations
New Obs  D1  D2      X1
      1   1   0  49.333
```

The same approach is used to compute the adjusted probabilities for groups 2 and 3. They are: $\hat{\pi}_{2\,\text{adj}} = .8917$ and $\hat{\pi}_{3\,\text{adj}} = .6642$. The chi-square statistic is the omnibus test for differences among these three values; they are the essential descriptive results that will be reported. An alternative is to convert the probabilities to the corresponding odds ratios (i.e., .44, 8.23, and 1.98).

Table 15.2 Test for Homogeneous Logistic Regressions in an Experiment with Three Treatment Groups, One Covariate, and a Dichotomous Outcome Variable

Predictors in the Model	G-Statistic	Degrees of Freedom
D_1, D_2, X, D_1X, D_2X	$G_{(\mathbf{D,X,DX})} = 13.716$	5
D_1, D_2, X	$G_{(\mathbf{D,X})} = 12.449$	3
	Difference $= 1.267 = \chi^2_{\mathrm{AT}}$	Difference $= 2$

15.5 HOMOGENEITY OF WITHIN-GROUP LOGISTIC REGRESSION

Recall from Chapter 7 that the assumption of homogeneous within-group regression slopes can be tested using a model comparison F-test. A chi-square analog to this test is described in this section for logistic regression.

Two logistic regression models are estimated. The first model is based on all dummy variables required to identify groups, the covariate, and the products of each dummy variable times the covariate. The second model includes only the dummy variables and the covariate. The G-statistic associated with the second fitted model is subtracted from the G-statistic associated with the first fitted model to provide the chi-square statistic used to test the homogeneity of the logistic regression slopes. The summary of the application of this procedure to the example data of the previous section is shown in Table 15.2.

The p-value associated with the obtained chi-square is .53; it is concluded that there are insufficient data to reject the homogeneous regression assumption.

15.6 MULTIPLE COVARIATES

The approach described in the case of one covariate generalizes directly to multiple covariates. The first step involves the logistic regression of the 0–1 outcome variable on all required dummy variables and all covariates. The second step involves the regression of the 0–1 outcome variable on all covariates. The corresponding G-statistics are $G_{(\mathbf{D,X})}$ and $G_{(\mathbf{X})}$.

The example shown below has three groups and two covariates. The 0–1 outcome scores are identical to those used in the previous section; the two covariates are the same as shown in the example of multiple covariance analysis presented in Chapter 10 (Table 10.1). As can be seen in the output shown below, the first regression has two dummy variables and two covariates. The second regression has two covariates.

```
Binary Logistic Regression: Dichotomous Y versus D1,
D2, X1, X2
Link Function: Logit
Response Information
Variable       Value  Count
Dichotomous Y  1      18     (Event)
               0      12
               Total  30
```

```
Logistic Regression Table
                                                  Odds   95% CI
Predictor      Coef    SE Coef      Z      P Ratio Lower Upper
Constant   -4.78770    2.10304 -2.28 0.023
D1         -1.82844    1.44725 -1.26 0.206  0.16  0.01  2.74
D2          1.46749    1.33841  1.10 0.273  4.34  0.31 59.78
X1        0.0410009 0.0496310  0.83 0.409  1.04  0.95  1.15
X2         0.735960  0.391550  1.88 0.060  2.09  0.97  4.50
Log-Likelihood = -11.093
Test that all slopes are zero: G = 18.196, DF = 4, P-Value
= 0.001
```

```
Binary Logistic Regression: Dichotomous Y versus X1, X2
Link Function: Logit
Response Information
Variable       Value  Count
Dichotomous Y  1      18     (Event)
               0      12
               Total  30
Logistic Regression Table
                                                  Odds   95% CI
Predictor      Coef    SE Coef      Z      P Ratio Lower  Upper
Constant   -3.49999    1.80934 -1.93 0.053
X1        0.0144772 0.0472354  0.31 0.759  1.01  0.92   1.11
X2         0.698334  0.363340  1.92 0.055  2.01  0.99   4.10

Log-Likelihood = −14.177
Test that all slopes are zero: G = 12.027, DF = 2, P-Value
= 0.002
```

The difference between the two G-statistics is 6.169 and the difference between the degrees of freedom associated with the G-statistics is 2. The *p*-value associated with an obtained chi-square value of 6.169 is .0457. Hence, it is concluded that there are differences among treatments with respect to the probability of academic success. The estimated probability of success for each treatment and the associated predictor scores used to compute it can be seen in the output listed below.

Group 1

```
Predicted Event Probabilities for New Observations
New Obs      Prob   SE Prob        95% CI
      1  0.286258 0.199094 (0.0560666, 0.730323)
Values of Predictors for New Observations
New Obs  D1  D2     X1    X2
      1   1   0  49.3333   5
```

Group 2

```
Predicted Event Probabilities for New Observations
New Obs       Prob    SE Prob        95% CI
      1  0.915468 0.0856622 (0.552986, 0.989563)
Values of Predictors for New Observations
New Obs  D1  D2     X1    X2
      1   0   1  49.3333   5
```

Group 3

```
Predicted Event Probabilities for New Observations
New Obs       Prob    SE Prob        95% CI
      1  0.713984 0.198870 (0.270144, 0.943934
Values of Predictors for New Observations
New Obs  D1  D2     X1    X2
      1   0   0  49.3333   5
```

15.7 MULTIPLE COMPARISON TESTS

Multiple comparisons among the mean adjusted event probabilities estimated for the various treatment groups may be of interest. If so, approximate analogs to Fisher–Hayter and Tukey–Kramer approaches are shown in Table 15.3. The standard errors SE_i and SE_j shown in this table are included in the previously described *Minitab* output associated with the option for "Predicted Event Probabilities for New Observations." The critical values are based on infinite degrees of freedom.

Example 15.1 Consider the three-group two-covariate example in the previous section. The relevant output is shown before the beginning of this section. Note that the three adjusted event probabilities of success are .286, .915, and .714, for treatments 1, 2, and 3, respectively. The standard errors for these three probabilities are .199094, .0856622, and .198870, respectively. Tests on all three pairwise comparisons may

Table 15.3 Multiple Comparison Formulas for Mean Adjusted Event Probabilities

Type of Test	Formula	Critical Value
Fisher–Hayter	$\dfrac{\hat{\pi}_i - \hat{\pi}_j}{\sqrt{\dfrac{(SE_i)^2 + (SE_j)^2}{2}}} = q$	$q_{J-1,\infty}$
Tukey–Kramer	$\dfrac{\hat{\pi}_i - \hat{\pi}_j}{\sqrt{\dfrac{(SE_i)^2 + (SE_j)^2}{2}}} = q$	$q_{J,\infty}$

Table 15.4 Summary of Parallel Analyses Applied to Continuous and Dichotomous Outcome Data

ANOVA	Logistic ANOVA	ANCOVA	Logistic ANCOVA	Multiple ANCOVA	Multiple Logistic ANCOVA
$p = .220$	$p = .178$	$p = .010$	$p = .047$	$p = .002$	$p = .046$
$\bar{Y}_1 = 30$	$\hat{\pi}_1 = .40$	$\bar{Y}_{1\,\text{adj}} = 28.48$	$\hat{\pi}_{1\,\text{adj}} = .31$	$\bar{Y}_{1\,\text{adj}} = 28.98$	$\hat{\pi}_{1\,\text{adj}} = .29$
$\bar{Y}_2 = 39$	$\hat{\pi}_2 = .80$	$\bar{Y}_{2\,\text{adj}} = 40.33$	$\hat{\pi}_{2\,\text{adj}} = .89$	$\bar{Y}_{2\,\text{adj}} = 40.21$	$\hat{\pi}_{2\,\text{adj}} = .92$
$\bar{Y}_3 = 36$	$\hat{\pi}_3 = .60$	$\bar{Y}_{3\,\text{adj}} = 36.19$	$\hat{\pi}_{3\,\text{adj}} = .66$	$\bar{Y}_{3\,\text{adj}} = 35.81$	$\hat{\pi}_{3\,\text{adj}} = .71$

be of interest. For example, the FH-type test for treatments 1 and 2 is computed as follows:

$$\frac{.286 - .915}{\sqrt{\dfrac{(.199094)^2 + (.0856622)^2}{2}}} = 4.10 = q.$$

The critical value of q (found in the column headed with $J - 1 = 2$) for a 5% test and infinite degrees of freedom is 2.77; it is concluded that $\pi_{1\,\text{adj}} \neq \pi_{2\,\text{adj}}$. Corresponding tests for comparing treatments 1 versus 3 and 2 versus 3 yield p-values of .14 and .25, respectively.

15.8 CONTINUOUS VERSUS FORCED DICHOTOMY RESULTS

The example of dichotomous outcome data analyzed in this chapter was obtained by simply forcing quantitative data into a dichotomy. It is of interest to compare results of conventional ANCOVA on the original data (previously presented in Chapters 6, 7 and 10) with those using logistic analysis on the transformed data. Whenever information is thrown away by forcing continuous data into a dichotomy (almost always a bad practice), there is usually a loss of sensitivity unless there are outliers in the original data. So larger p-values are expected using dichotomized data. Table 15.4 summarizes results from parallel analyses.

The pattern of outcomes is completely consistent for all analyses regardless of the descriptive measure. That is, treatment 1 yields the poorest performance and treatment 2 yields the highest performance with respect to means, adjusted means, proportions, and adjusted proportions. Inferentially, however, it can be seen that the p-values for conventional ANCOVA and multiple ANCOVA are considerably smaller than for the logistic counterparts. This confirms common wisdom regarding the effects of forcing a continuous variable into a dichotomy. Of course, when the dependent variable is a true dichotomy (e.g., alive versus dead) the methods of this chapter are recommended.

15.9 SUMMARY

Logistic regression is recommended for estimating the parameters of a modified ANCOVA model that is designed for dichotomous outcome variables. Dichotomous ANCOVA can be carried out using two regressions. First, the dichotomous outcome is regressed on all group membership dummy variables and all covariates using logistic regression. Second, the dichotomous outcome is regressed on covariates. The G-statistic from the second regression is subtracted from the G-statistic from the first regression to compute a chi-square statistic. This chi-square is used to test the hypothesis that the adjusted event probabilities are equal for all treatments. The methods shown in this chapter generalize to other analyses related to ANCOVA such as tests for homogeneous regression, picked points analysis, and quasi-ANCOVA.

CHAPTER 16

Designs with Ordered Treatments and No Covariates

16.1 INTRODUCTION

It has been pointed out in previous chapters that a common methodological error is to apply ANOVA to experiments that have quantitatively scaled independent variables. For example, a randomized group experiment with five levels of the independent variable, where the treatments are doses of a drug (say, 10, 20, 30, 40, and 50 mg) might be analyzed using ANOVA. Because a conventional one-factor ANOVA treats dosage levels as if they differ in type rather than amount, this is an example of a mismatch between the inherent nature of the independent variable and the scaling that is assumed when applying ANOVA. A major consequence of applying ANOVA rather than simple linear regression analysis in this situation is low power and excessive additional analyses in the form of often unnecessary multiple comparison tests. This chapter makes a similar point in the context of experiments with independent variables that are conceptualized as neither purely qualitative nor quantitative.

16.2 QUALITATIVE, QUANTITATIVE, AND ORDERED TREATMENT LEVELS

Fante et al. (submitted) investigated the effects of three types of training method on fluency. The three types of training were (A) traditional, (B) traditional plus study objectives, and (C) fluency building. The dependent variable was a quantitative measure of product knowledge. There was no basis for treating this as a simple regression problem because there was no logic for attaching a quantitative scale value of treatment potency to each treatment. More specifically, it could not be argued that adding study objectives (treatment B) to the traditional treatment (treatment A)

The Analysis of Covariance and Alternatives: Statistical Methods for Experiments, Quasi-Experiments, and Single-Case Studies, Second Edition. Bradley E. Huitema.

should improve the adequacy of training by the same amount as using fluency training (treatment C) instead of study objectives (treatment B). No meaningful quantitative value of independent variable "amount" was provided for each treatment in the sense that it is in a drug dosage study where the amount of a drug would be quantified in milligrams or some related metric. But this does not mean that the independent variable should be treated as purely qualitative (i.e., nominal). Even though it is common to classify independent variables into just two categories (qualitative and quantitative) this distinction does not exhaust the possibilities regarding the nature of the independent variable. Many experiments are carried out using treatments that are different condition types, but the types can be ordered.

Initially the independent variable in the fluency experiment may appear to be purely qualitative because the treatments are labeled as different types (rather than amounts) of training. But there was an a priori basis for ordering the three treatment types. Theoretical arguments presented before carrying out the experiment suggested that the traditional approach should be the least effective, the addition of study objectives to the traditional approach should be somewhat more effective, and the fluency building approach should be the most effective. The rank order regarding the relative potency of the three treatments led to the following prediction regarding the outcome on the dependent variable: $\mu_A \leq \mu_B \leq \mu_C$. This prediction can be viewed as a formal statement of the alternative hypothesis; that is, the alternative to the null hypothesis of equal population means. Ordered treatments of this type have implications for the method of analysis. Although several different appropriate methods are available for analyzing data of this type, these methods are almost invisible in applied statistics and research design textbooks. I refer to the class of tests that acknowledge treatment order as monotone tests. This term is used because these tests are relevant to the hypothesis that increases on the rank of the treatments that are accompanied by increases (decreases) on the outcome measure Y.

If an experiment has ordered treatment levels but is analyzed as if the scaling of the treatments is qualitative, there are two problems. First, a conventional parametric ANOVA will have lower power (often far lower) than will an appropriate analysis that acknowledges the order. The reason for this is that the ANOVA F-test is uniformly powerful against differences among population means that exist in all possible directions. The power is the same for detecting all alternatives to the null hypothesis that yield the same noncentrality parameter (a parameter measuring the degree of departure from the null hypothesis). This is an advantage if the treatments are purely qualitative, but it is a disadvantage when the treatments are ordered, because the order implies interest in fewer than all possible alternatives.

Suppose, for example, that the null hypothesis is false in a four-group experiment and the population means are 3, 7, 10, and 17 for populations one through four, respectively. The F-test will have the same power in this case as it would be if the means were 17, 3, 7, and 10 for populations one through four, respectively. The noncentrality parameter for ANOVA is exactly the same in these two situations because ANOVA is not sensitive to the order of the means. But the power of monotone tests is a function of the correlation between the predicted treatment order and the

actual order of the population outcome means. Because monotone tests are designed to utilize the information regarding treatment order, they generally have higher power than the omnibus ANOVA test. This is true even if there are minor discrepancies between the predicted and actual population orders.

The second problem with applying ANOVA to designs with ordered treatments is that researchers routinely compute and report results of multiple comparison tests for all pairwise comparisons when the F-test is significant. Although this practice often provides a welcome level of completeness in an experiment with qualitative treatments, it is generally unnecessary when the treatments are ordered. If the purpose of the experiment is to provide information on whether there is a monotonic relationship between the order of the treatments and the outcome measure, a single test that evaluates a single monotonic contrast is all that is required. The essential issue of interest is the pattern of the whole collection of outcome means rather than pairwise mean differences.

This is similar to the type of interest one has in a correlational study. A single correlation coefficient is sufficient to describe the linear relationship between two continuous variables (if the linear model is apt). Although the correlational researcher could compute t-tests on the difference between means on Y for each pair of points on X, this would bloat the analysis and miss the point. It is essentially the same with ordered treatments.

Basis for Ordering Treatment Levels

There are several situations in which one is likely to encounter ordered treatment levels. Among them is one that must be avoided. If the experiment has already been completed it may be tempting to first inspect the descriptive outcome and then order the treatments according to the value of the obtained means. If monotone analysis is then applied to the ordered means we have a form of *post hoc* analysis at its worst; it will generate type I error rates that are far above the nominal level established for the monotone test. This problem does not occur with appropriate *a priori* justifications for treatment order. Several of these justifications are described next.

Theory

Treatments are often derived from theory regarding the nature of the phenomena being investigated. Indeed, much research is carried out for theory testing purposes. A useful theory is based on constructs that can be operationalized as independent and dependent variables that are subject to empirical investigation. The operations that define treatment levels often can be ordered more easily than they can be quantitatively scaled. For example, in learning theory the salience of different social stimuli may be more easily ordered than assigned quantitative scale values. In medical research, theory may predict that the severity of disease is related to degree of change on certain biomarkers, but severity may not be easily quantified. It is common to use coarse disease staging categories to label disease severity rather than to attempt more precise quantitative measurement.

Previous Research

It was pointed out that monotone testing involves predicting the order of treatment effectiveness before the experiment is carried out, but it is not appropriate to order the treatments *post hoc*. If the order is established on the basis of outcome means observed after the experiment is completed it cannot be claimed that the order was based on prediction. But if the outcome of a *previous* experiment is used as the basis for ordering the treatments in the design of a current experiment there is a basis for the prediction and therefore a sound justification for a monotone test. This type of justification is natural in programmatic research where each experiment in a sequence informs the next.

Decomposition of Complex Treatments

Suppose you are interested in evaluating the relative importance of specific aspects of a multicomponent treatment that is designed to reduce LDL-cholesterol in patients with heart disease. The treatment consists of three components: (1) a low dosage (10 mg) statin, (2) an aerobic exercise routine, and (3) a Mediterranean diet. Perhaps a previous clinical trial has shown the complete treatment program (consisting of these three components) to be effective relative to a control condition, but you are interested in the relative importance of the different components.

It is widely believed that the two lifestyle components (diet and exercise) are moderately effective, but still they are far less effective than the heavy artillery of a specific statin. It is also widely believed that diet is more important than exercise in reducing LDL-cholesterol (although exercise is better at increasing HDL-cholesterol). Hence, the predicted order of effectiveness of the three components (beginning with the least effective) is 1 (Exercise), 2 (Diet), and 3 (Statin). A randomized three-group design using these three conditions could be carried out and analyzed using a monotone test. But such an experiment does not exhaust the ordered treatment design possibilities.

The predicted order regarding the relative effectiveness of three components leads naturally to hypotheses regarding the effectiveness of combinations of these components. For example, it could be argued that the following order is consistent with the originally stated beliefs regarding the effectiveness of exercise, diet, and statin: 1 = Exercise, 2 = Diet, 3 = Diet plus Exercise, 4 = Statin, 5 = Statin plus Exercise, 6 = Statin plus Diet, 7 = Statin plus Exercise plus Diet. This order assumes that diet plus exercise is less effective than statin alone. But if there is no empirical or theoretical support for this order it might be just as reasonable to predict that the combination of exercise and diet are of the same effectiveness as statin alone. In this case it would be reasonable to assign the same (tied) rank to both conditions; that is, assign rank 3.5 to combined diet and exercise as well as to statin alone. The use of tied ranks is acceptable for more than two groups, but usually this is not necessary.

Justifications other than theory, previous research, and treatment decomposition may be used for ordering treatments. The key issues are that (a) there is interest in evaluating whether the predicted order is correct and (b) the order is not established *post hoc*.

Three monotone methods for the analysis of one-factor randomized-group designs are described in the remainder of this chapter. The first is a parametric approach, the

second is a new nonparametric approach that is available in two versions, and the third is based on an atypical application of ordinal logistic regression. Extensions of these analyses to designs that include one or more covariates are described in Chapter 17.

16.3 PARAMETRIC MONOTONE ANALYSIS

Hypotheses

Both ANOVA and parametric monotone analysis test the following omnibus null hypothesis: H_0: $\mu_1 = \mu_2 = \cdots = \mu_J$. This is a formal statement regarding the equality of all population means; it is equivalent to the statement that all population means are equal to the grand population mean μ. If this null hypothesis is rejected using the ANOVA F-test, it is claimed that sufficient evidence exists for the alternative hypothesis. That is, a statistically significant ANOVA F implies that at least one population mean differs from the grand population mean; formally, H_A: $\mu_j \neq \mu$ for at least one j.

Although both ANOVA and monotone analysis test the same null hypothesis, the alternative hypothesis is much more specific for the latter. The monotone alternative states that the population dependent variable means are monotonically ordered. For example, it may be stated that the order of population means is lowest for the first treatment and highest for group J. That is, H_A: $\mu_1 \leq \mu_2 \leq \cdots \leq \mu_J$ (with at least one strict inequality). This is the alternative hypothesis when the test is set up in directional form, as is usually the case. If a nondirectional test is desired, there are two monotonic alternative hypotheses (viz., H_A: $\mu_1 \leq \mu_2 \leq \cdots \leq \mu_J$ and H_A: $\mu_1 \geq \mu_2 \geq \cdots \geq \mu_J$). The nondirectional form would be used in the rare situation where the researcher has a basis for ordering the treatments but no basis for predicting the direction of the treatment effects. The details of carrying out the parametric version of monotone analysis are as follows.

Abelson–Tukey Test

Although several tests have been proposed for the ordered treatment case, I prefer the one proposed by Abelson and Tukey (1963). The key to implementing this approach is a set of contrast coefficients (presented in Table 16.1) that satisfy what Abelson and Tukey describe as the *maximin* criterion. These coefficients provide a nearly ideal way of combining means for purposes of detecting a monotonic relationship between the independent and dependent variables. Some of the ideas behind of the maximin criterion are described next.

Suppose it is known that a set of five treatment conditions has been ordered in terms of a known key property that distinguishes one treatment from another. Let's say this property is drug dosage and the doses are 10, 20, 30, 40, and 50 mg for groups one through five, respectively. Denote the dosage variable as X and the dose for the jth group as X_j. We would like to construct a set of contrast coefficients c_j that will correlate perfectly with the X_j. This will be easy in the situation described here

because we know the exact value of X_j for each treatment group and we can see that the increase from one treatment group to the next is linear. This means that any set of contrast coefficients that vary linearly should correlate perfectly with the X_j. For example, consider the set of $c_j = [-2, -1, 0, 1, 2]$. These coefficients follow a linear progression and the correlation between c_j and X_j is 1.0.

Now let's change the situation a little. Still assume that the amount of some key characteristic distinguishes one treatment from another, but now X_j is unknown. When I state that X_j is unknown, understand that it is the specific values that are unknown but there is partial knowledge under the assumption that X_j is ordered. That is, we have limited knowledge that allows us to order the treatments in terms of the key characteristic without knowing the exact amount of the characteristic. In this case we would like to develop a set of contrast coefficients c_j that will maximize the correlation with the largely unknown X_j. Suppose we use the same set of contrast coefficients we used previously, i.e., $[-2, -1, 0, 1, 2]$, but now the unknown X_j are actually 3, 4, 7, 21, and 67. In this case, the correlation between c_j and X_j is .85. This correlation is less than perfect because the progression of these contrast coefficients is linear, whereas the progression of the X_j is not. As a third example, suppose the values of unknown X_j are 7, 8, 9, 10, and 97. The correlation of these values with the linear coefficients is only .73.

Note in these three examples that the correlation between the linear contrast coefficients and the X_j is perfect only in the case when the progression of X_j is exactly linear. But also note that the lowest correlation is still reasonably high (.73). These correlations are of interest because the power of the monotone test is a function of the correlation between the contrast coefficients and the unknown X_j; the higher the absolute correlation the higher the power.

The goal in developing a satisfactory analysis of monotone data is to find contrast coefficients that will be, in general, highly correlated with the unknown X_j regardless of their specific values. The maximin coefficients developed by Abelson and Tukey accomplish this goal. If we apply maximin coefficients to the two examples in the previous paragraph the correlations are .87 and .79; these are somewhat higher than corresponding correlations that were found using the linear contrast coefficients (i.e., .85 and .73). Hence, even though the linear coefficients are quite good, the maximin coefficients are better.

The reason Abelson and Tukey use the term "maximin" to describe their contrasts is because the c_j were developed in an attempt to *maxi*mize the *min*imum squared correlation between c_j and the unknown X_j considering all possibilities under the monotonicity assumption. The formula for computing the *j*th maximin coefficient is as follows:

$$c_j = \sqrt{(j-1)[1-((j-1)/J)]} - \sqrt{j(1-j/J)},$$

where J is the number of ordered treatment groups. Although these coefficients are provided in Table 16.1 for $J = 2 - 20$, the formula is required for the unlikely case that J exceeds 20.

Table 16.1 Maximin Contrast Coefficients (c_j) for Parametric Monotone Analysis

(a) $J = 2$–10

j	$J=2$	$J=3$	$J=4$	$J=5$	$J=6$	$J=7$	$J=8$	$J=9$	$J=10$
1	−.707	−.816	−.866	−.894	−.913	−.926	−.935	−.943	−.949
2	.707	.000	−.134	−.201	−.242	−.269	−.289	−.305	−.316
3		.816	.134	.000	−.070	−.114	−.144	−.167	−.184
4			.866	.201	.070	.000	−.045	−.076	−.100
5				.894	.242	.114	.045	.000	−.032
6					.913	.269	.144	.076	.032
7						.926	.289	.167	.100
8							.935	.305	.184
9								.943	.316
10									.949

(b) $J = 11$–20

j	$J=11$	$J=12$	$J=13$	$J=14$	$J=15$	$J=16$	$J=17$	$J=18$	$J=19$	$J=20$
1	−.953	−.957	−.961	−.964	−.966	−.968	−.970	−.972	−.973	−.975
2	−.326	−.333	−.340	−.346	−.351	−.354	−.358	−.362	−.364	−.366
3	−.198	−.209	−.218	−.226	−.233	−.238	−.243	−.248	−.252	−.255
4	−.118	−.133	−.145	−.155	−.163	−.170	−.178	−.183	−.188	−.192
5	−.056	−.075	−.090	−.102	−.113	−.122	−.130	−.136	−.142	−.147
6	.000	−.024	−.043	−.059	−.071	−.082	−.092	−.100	−.107	−.113
7	.056	.024	.000	−.019	−.035	−.047	−.059	−.068	−.077	−.084
8	.118	.075	.043	.019	.000	−.015	−.029	−.040	−.050	−.058
9	.198	.133	.090	.059	.035	.015	.000	−.014	−.024	−.034
10	.326	.209	.145	.102	.071	.047	.029	.014	.000	−.011
11	.953	.333	.218	.155	.113	.082	.059	.040	.024	.011
12		.957	.340	.226	.163	.122	.092	.068	.050	.034
13			.961	.346	.233	.170	.130	.100	.077	.058
14				.964	.351	.238	.178	.136	.107	.084
15					.966	.354	.243	.183	.142	.113
16						.968	.358	.248	.188	.147
17							.970	.362	.252	.192
18								.972	.364	.255
19									.973	.366
20										.975

Source: This table is a modified and expanded (by the author) version of portions of Table 1 in Abelson and Tukey (1963).

Abelson and Tukey describe two additional contrasts (other than linear and maximin) that have desirable properties; they label these as linear-2 and linear-2–4 contrasts. These are essentially simple approximations to the maximin contrast. These approaches had a place at the time the Abelson–Tukey article was published (1963), but the major justification for using them (computational simplicity) is not an issue in the current age. I generally recommend the maximin approach rather than the approximations; software entry of minimax coefficients is no more difficult than

entry of the approximations. There are, however, two exceptions to this general recommendation.

If there are only two groups I recommend that you use -1 and 1 as the two coefficients. If there are three groups, use -1, 0, and 1 as the coefficients. These coefficients will yield the same test results as will the maximin coefficients, but the contrast estimates are more transparent. In the two-group case, the contrast estimate is simply the difference between the two means. In the three-group case, the contrast estimate is simply the difference between the first and third means.

Abelson–Tukey Test Statistic

The test statistic is defined as

$$\frac{\sum_{j=1}^{J} c_j \bar{Y}_j}{\sqrt{\mathrm{MS_W}\left[\sum_{j=1}^{J} \frac{c_j^2}{n_j}\right]}} = t,$$

where

c_j is the Abelson–Tukey maximin contrast coefficient associated with treatment j;
$\bar{Y}_j$ is the jth sample mean;
$\mathrm{MS_W}$ is the within-group mean square from the ANOVA on Y;
n_j is the sample size associated with the jth treatment; and
t is the test statistic.

The directional form of the test is carried out by comparing the obtained value of t with the critical value of t (for specified α_1) based on $N - J$ degrees of freedom. The nondirectional form of the test is carried out by comparing the absolute obtained value of t with the critical value of t (for specified α_2) based on $N - J$ degrees of freedom.

Assumptions

The assumptions associated with the Abelson–Tukey test are essentially the same as those underlying conventional ANOVA (viz., errors are independent and population distributions are normally distributed with homogeneous variances), but the treatment levels are assumed to be ordered. This means that conventional diagnostic methods used for ANOVA are also relevant with the Abelson–Tukey test.

Standardized Effect Size and η^2 for the Maximin Contrast

A standardized effect size measure for the maximin contrast can be computed using

$$\frac{\sum_{j=1}^{J} c_j \bar{Y}_j}{\sqrt{\mathrm{MS_W}}} = g_{\text{maximin}}.$$

The proportion of the total variation on the dependent variable explained by the maximin contrast can be estimated using

$$\frac{\left[\dfrac{n\left(\sum_{j=1}^{J} c_j \bar{Y}_j\right)^2}{\sum_{j=1}^{J} c_j^2}\right]}{\mathrm{SS}_{\text{total}}} = \hat{\eta}^2_{\text{maximin}}.$$

If the sample size (n) is not the same for all groups substitute the maximin contrast *w*eighted *h*armonic mean (n_{mwh}) of the sample sizes in the numerator of this formula. This mean is computed using $\dfrac{\sum_{j=1}^{J} c_j^2}{\sum_{j=1}^{J}\left(\frac{c_j^2}{n_j}\right)} = n_{\text{mwh}}$.

Example 16.1 Consider the data in Table 16.2 to be from a randomized-group design where the levels are five different methods of treating depression. All treatments were applied for nine months; a measure of well-being was then obtained from each patient. The five treatment conditions were originally labeled A, B, C, D, and E. These letters were attached arbitrarily to the different types of treatment; that is, the independent variable was viewed initially as qualitative and the letters were simply used to name the treatment conditions. Then, *before* the experiment was carried out, it was recognized that there was a sound theoretical justification for ordering the treatments in terms of a treatment characteristic that was believed to be likely to affect the outcome. Treatment E was predicted to be least effective, treatment C was viewed as likely to be somewhat more effective, treatments A and B were viewed as still more effective, and finally treatment D was predicted to be most effective. Because the researcher was interested in the ordered hypothesis and was able to provide a predicted rank order of treatment effectiveness, a monotone analysis was justified.

Table 16.2 ANOVA and Monotone Analysis Applied to a Five Group Experiment: Treatments A, B, C, D, and E (Rank 1 = Lowest Predicted Effectiveness)

Rank 1	Rank 2	Rank 3	Rank 4	Rank 5
Treatment E (Rank 1)	Treatment C (Rank 2)	Treatment A (Rank 3)	Treatment B (Rank 4)	Treatment D (Rank 5)
8	10	2	12	6
4	4	2	6	3
1	2	7	6	10
4	2	5	2	8
1	5	9	5	11
$\bar{Y}_E = 3.6$	$\bar{Y}_C = 4.6$	$\bar{Y}_A = 5.0$	$\bar{Y}_B = 6.2$	$\bar{Y}_D = 7.6$

ANOVA

Source		SS	*df*	MS	*F*	*p*
Between groups	47.60	4	11.90	1.142	.37	.37
Within groups	208.40	20	10.42			
Total	256.00	24				

Contrast coefficients (from Table 16.1): [−.894, −.201, 0, .201, .894]

Monotone t

$$\frac{\sum_{j=1}^{J} c_j \bar{Y}_j}{\sqrt{\mathrm{MS_W}\left[\sum_{j=1}^{J} \frac{c_j^2}{n_j}\right]}} = \frac{-.894(3.6) - .201(4.6) + 0(5.0) + .201(6.2) + .894(7.6)}{\sqrt{10.42\left[\frac{-.894^2}{5} + \frac{-.201^2}{5} + \frac{0^2}{5} + \frac{.201^2}{5} + \frac{.894^2}{5}\right]}}$$

$$= \frac{3.90}{1.87} = 2.09 = t_{\mathrm{obt}}$$

Critical value of t (for $\alpha_1 = .05$) based on $df = 20$ is 1.725.
Critical value of t (for $\alpha_2 = .05$) based on $df = 20$ is 2.086.
p-value (directional test) = .025.
p-value (nondirectional test) = .050.

$$g_{\mathrm{maximin}} = \frac{\sum_{j=1}^{J} c_j \bar{Y}_j}{\sqrt{MS_{\mathrm{W}}}} = \frac{3.90}{\sqrt{10.42}} = 1.21.$$

$$\hat{\eta}^2_{\mathrm{maximin}} = \frac{\left[\frac{5(3.9)^2}{1.67847}\right]}{256} = .18.$$

The results are shown in Table 16.2. Note that the pattern of the means strongly suggests that there is a monotonic increasing relationship between the rank order of the treatments and the order of the dependent variable means. The results of ANOVA and the monotone analysis are quite different. ANOVA leads to the conclusion that there are insufficient data to reject the null hypothesis whereas the monotone analysis clearly rejects it. Note that the p-value for ANOVA is .37 whereas it is only .025 (directional) for the monotone analysis. Hence, the null hypothesis is rejected and it is concluded that the outcome is consistent with the monotonic alternative $\mu_1 \leq \mu_2 \leq \mu_3 \leq \mu_4 \leq \mu_5$.

Interpretation Considerations

A major advantage of monotone testing relative to ANOVA and associated multiple comparison tests is that only one test is needed regardless of the number of treatments. Because we simply want to know whether Y is a monotonic function of the order of the treatments, multiple comparisons of the individual means are not necessary.

The issue of ordered treatments is acknowledged during three stages of an investigation: research design, statistical analysis, and interpretation. During the research design stage the substantive question is formulated in monotonic terms. Specifically, do the means on the dependent variable tend to increase with increases in the rank order of the treatment conditions? Next, during the analysis stage, the order of the treatment conditions must be known when specifying the contrast. Last, during the interpretation stage, the results are described as supporting or not supporting the monotonic alternative to the null hypothesis. Clarity of purpose at the research design stage usually removes ambiguity regarding an appropriate method of analysis and leads naturally to straightforward statements regarding the results.

When the average group scores on Y increase with the order of the treatment levels the relationship is described as *monotonic increasing*. Alternatively, if the average Y scores decrease with the order of the treatment levels, the relationship is described as *monotonic decreasing*. Appropriate graphics describing the pattern of the outcome means should always accompany the verbal description of the results.

What If ANOVA F Is Significant and the Monotone t Is Not?

It is possible for the null hypothesis to be retained using the monotone test but rejected using ANOVA (although it is usually the other way around). This is bothersome to some researchers; it should not be. Remember that the monotone test and ANOVA answer different questions. When the null hypothesis is retained using the monotone test one simply concludes that there are insufficient data to support the argument that a monotonic relationship exists between the independent and dependent variables; this conclusion does not speak to the issue of other forms of relationship. This is neither a weakness nor a quirk of monotone analysis.

The same issue occurs with many more traditional types of analysis. For example, consider the case of applying simple linear regression analysis to data that are best described using a nonlinear function. A nonsignificant linear relationship does not necessarily mean that there is no systematic relationship whatsoever. If a regression slope is not statistically significant one should entertain several reasons for this outcome; among them: (a) there is no linear relationship between X and Y, (b) there is a linear relationship in the population but the test had insufficient power to detect it, and (c) the linear model does not adequately represent the data and therefore some other functional form should be considered. Essentially the same issues should be considered when a monotone test is not statistically significant.

Comparison of ANOVA, Monotone Analysis, and Simple Regression

It may be helpful to point out the differences among ANOVA, monotone analysis, and simple regression analysis with respect to the question answered. There is a clear

difference in the level of specificity regarding what one can conclude as we move from the ANOVA F-test, through the monotone test, and finally to simple linear regression analysis. The ANOVA F-test is the most diffuse because we can conclude only that at least one population mean differs from the grand population mean (or that at least two population means differ). The monotone test is much more specific because we can conclude that an increase on the independent variable is associated with an increase on the dependent variable (when the relationship is monotonic increasing). Simple regression analysis is the most specific. Not only can we state that an increase on the independent variable is associated with an increase on the dependent variable (when the relationship is positive), but we can also state how much Y increases with a one unit increase on X. But remember that all of these analyses are based on different assumptions regarding the nature of the independent variable. A qualitative independent variable is assumed when using ANOVA, ordered levels of the independent variable are assumed when using monotone analysis, and a quantitative independent variable is assumed under the simple linear regression model.

If the independent variable is quantitative it is possible to use any of these three analyses, but it is generally a mistake to use conventional ANOVA or monotone analysis if the simple regression model fits the data. Because neither ANOVA nor monotone analyses utilize all of the quantitative information, they have lower power and provide answers to less specific questions. The power lost by using monotone analysis rather than regression is usually small relative to the loss associated with using ANOVA. Monotone analysis can actually be more powerful than regression if the data exhibit moderate departures from linearity.

Trend Analysis Versus Regression Analysis

A fourth analytic approach, not previously discussed, is known as trend analysis. This is a popular method in psychological research with quantitative independent variables. It usually consists of the following routine:

1. Test for differences among the means using ANOVA F. If F is not significant, terminate the analysis; if F is significant, proceed to step 2.
2. Test for linear trend of the means.
3. Perform an F-test for lack of fit of the linear model; if lack of fit is identified, follow up with a test for quadratic trend. Continue this process with higher order polynomials until no significant results are obtained for lack of fit and higher order trend tests.

The purported logic of this routine, as described in many textbooks, is essentially as follows. The preliminary first step (ANOVA) will tell you whether it makes sense to proceed to subsequent steps. If, for example, the ANOVA F is not significant the researcher concludes that there are no systematic differences among means and therefore it is useless to look for trend. But if the ANOVA F is significant it makes sense to evaluate the possibility of trend. If linear trend is identified, compute a lack

of fit test; it will tell you whether to bother fitting trend components more complex than the linear function.

Unfortunately, this line of thought is not justified. It reflects a lack of understanding of a retained null hypothesis. This is a problem with both the preliminary ANOVA and lack of fit tests. In addition, there is a different problem with the tests for trend. Briefly, these problems are described below.

First, the preliminary F-test is unnecessary and is potentially misleading. It has been pointed out previously that ANOVA has low power (relative to competing procedures) in the context of quantitative independent variables. It often will lead the researcher to incorrectly conclude that there is no trend just because mean differences have not been detected using F. Second, the lack of fit test has low power against the type of nonlinearity most likely to be of interest (viz., a quadratic function). Third, the F-tests for the various trend components have fewer degrees of freedom than the corresponding regression tests and are less powerful under realistic situations. I need not perseverate on the details of these problems, because there is a desirable alternative method of analysis that should be considered. An example may help describe it.

Suppose the data shown in Table 16.2 came from an experiment in which the independent variable is quantitative (rather than just ordered as is stated in the table). Let the quantitative levels of the independent variable be 10, 20, 30, 40, and 50 for treatments one through five, respectively. Fit a simple linear regression model to the data. This analysis yields intercept and slope estimates of 2.520 and .096, respectively. These coefficients are essential aspects of the descriptive analysis.

These results imply that the dependent variable scores increase approximately one point for every 10-point increase on the independent variable. A quick glance at the treatment means immediately confirms this interpretation. The inferential test on the slope yields an F of 5.05 (p-value = .035). This is strong evidence that a linear function describes some of the variation among the means. Indeed, a comparison of the regression sum of squares (46.08) with the between group sum of squares (47.60) reveals that almost all of the variation of between-group means can be modeled as a linear function. The issue of possible minor curvature can be investigated by fitting higher order polynomial models; generally it is sufficient to fit only quadratic and cubic models.

The curvature coefficient estimate in the quadratic model yields a nondirectional p-value of .11. The cubic component in the cubic model yields a p-value of .22. Hence, the overall conclusion of these analyses is that there is strong evidence of positive linear trend, but not curvature. This approach completely avoids all of the tests in the typical trend analysis.

Because I did not go through the details of conventional trend analysis one might suspect that the conventional tests for linear, quadratic, and cubic functions using trend analysis are the same as the regression model tests for these same functions, but they are not. The error terms for these two approaches differ. In the case of the data in the current example one would not compute the trend tests according to the conventional procedure because the preliminary ANOVA F has a p-value of .37. But if the trend test is computed in spite of the preliminary test result, the obtained F for linear trend

is 4.42. Note that this is smaller than the obtained F of 5.05 for linear trend using the linear regression approach. The difference between these two Fs is a reflection of fewer degrees of freedom for the error MS using the trend analysis approach.

My recommendation is to avoid conventional trend analysis and use polynomial regression when the independent variable is quantitative. If the Abelson–Tukey procedure is used in this case rather than regression, some information is ignored and consequently some power is lost. The power difference between the two is small, however, relative to the power difference between ANOVA and the Abelson–Tukey test. Note that in the case of the example data the p-values (all nondirectional) for ANOVA, Abelson–Tukey, and linear regression are .37, .05, and .03, respectively. Because directional tests are usually justified in the case of ordered treatments, the power advantage of Abelson–Tukey and regression tests over ANOVA are even larger than is suggested by the comparison of these nondirectional p-values.

16.4 NONPARAMETRIC MONOTONE ANALYSIS

Several methods of nonparametric monotone analysis are available in the statistical literature, although they are not usually referred to under this rubric. The most well-known method, proposed independently by Terpstra (1952) and Jonckherre (1954), is usually called either the Terpstra–Jonckherre test or simply the Jonckherre test. This test is known among statisticians but it, like the Abelson–Tukey procedure, has not been implemented in most statistical software packages.

A more recent nonparametric monotone procedure has been proposed by McKean et al. (2001). This approach has several advantages over the Terpstra–Jonckherre test. First, it provides descriptive information in the form of a measure of association (a form of effect size) between the independent and dependent variables. Second, it provides a confidence interval on the measure of association. Third, the inferential aspects are based on bootstrap methodology that is believed to provide more exact results in the small sample case. Because the test incorporates both the well-known Spearman rank order correlation coefficient and bootstrap methods I have labeled it as the *S*pearman *b*ased *b*ootstrap (SBB) monotone method. The fourth advantage is that the general approach is easily extended to handle covariates.

The use of the Spearman correlation in the test for ordered treatments is an application rather different than the typical situation in which the Spearman is applied. Whereas X and Y are both random variables under the correlational design (where the Spearman coefficient is usually applied), the X variable is fixed in the ordered treatment design considered here. Although the application of the Spearman coefficient to the ordered treatment randomized group experiment appears to be new, the theory supporting this type of approach is well established (see Lehmann and D'Abrera, 1975, p. 291; Tryon and Hettmansperger, 1973).

Although the recommended SBB methodology is computer intensive, an alternative method of computing the monotonic test is also available; the latter method can easily be carried out by hand. The issues of tied ranks, small sample size, and information lost through the rank transformation on Y may lead to minor inaccuracy in

the p-values when the alternative method is used. The recommended SBB approach is not affected by these issues. An example illustrating the application of both the simple approach and the SBB approach is describe below.

Hypotheses and Notation

The null and alternative hypotheses are written as

H_0: $\theta_1 = \theta_2 = \cdots = \theta_J$ and
H_A: $\theta_1 \leq \theta_2 \leq \cdots \leq \theta_J$ (with at least one strict inequality),

respectively, where J is the number of populations and the parameter θ_j is the dependent variable median associated with the jth population. Let X_{ij} denote the rank order of the treatment condition for the ith observation associated with the jth group, for $i = 1, \ldots, n_j$, and $j = 1, \ldots, J$. Correspondingly, let Y_{ij} denote the ith observation on the dependent variable associated with the jth group, for $i = 1, \ldots, n_j$, and $j = 1, \ldots, J$.

Simple Approach

The test of the null hypothesis is carried out by computing a test of significance of the Spearman correlation (r_s) between the ordered levels of the independent variable (X) and the dependent variable (Y). If the null hypothesis presented above is true, then it is also true that the population Spearman correlation $\rho_s = 0$. Correspondingly, if the monotone alternative hypothesis presented above is true, then $\rho_s \neq 0$. Hence, the null hypothesis can be evaluated by testing the statistical significance of r_s.

The conventional t-test for the Spearman coefficient involves computing:

$$\frac{r_s}{\sqrt{\frac{1 - r_s^2}{N - 2}}} = t.$$

The obtained value of t is compared with the critical value of t based on $N - 2$ degrees of freedom.

The application of this approach is described below using the data presented in Table 16.2. These data have been rearranged in Table 16.3 in a form that may make the computation easier to follow.

Sample medians = 4.0, 4.0, 5.0, 6.0, and 8.0 for groups 1 through 5, respectively.

Correlation of X and Y ranks = Spearman correlation = .452.

Squared Spearman correlation = .204.

Conventional test on Spearman correlation:

$$\frac{.452}{\sqrt{\frac{1 - .204}{25 - 2}}} = 2.43 = t_{obt}.$$

Table 16.3 Computational Example of Spearman Based Monotone Method

X (Ordered Tx)	Y	Rank of X	Rank of Y	Mean Rank on Y
1	8	3	19.5	
1	4	3	10	
1	1	3	1.5	Gp. 1 = 8.5
1	4	3	10	
1	1	3	1.5	
2	10	8	22.5	
2	4	8	10	
2	2	8	5	Gp. 2 = 11.1
2	2	8	5	
2	5	8	13	
3	2	13	5	
3	2	13	5	
3	7	13	18	Gp. 3 = 12.4
3	5	13	13	
3	9	13	21	
4	12	18	25	
4	6	18	16	
4	6	18	16	Gp. 4 = 15.0
4	2	18	5	
4	5	18	13	
5	6	23	16	
5	3	23	8	
5	10	23	22.5	Gp. 5 = 18.0
5	8	23	19.5	
5	11	23	24	

Critical $t = 2.069$ (for $\alpha_2 = .05$) or 1.714 (for $\alpha_1 = .05$).

Directional test p-value (from *Minitab*) = .012.

Decision: Reject H_0: $\theta_1 = \theta_2 = \cdots = \theta_5$.

The X and Y data are listed in the first two columns. It can be seen in column one that each subject in the first treatment is identified as a member of group one; therefore, there are five ones. The same pattern holds for the other groups. The dependent variable scores appear in the second column.

The 25 scores in the first column are ranked and entered in the third column; rank 3 is assigned to all subjects in the first group because 3 is the average of the five ranks (i.e., 1, 2, 3, 4, and 5) associated with this group; the same pattern holds for all groups. (Note that there will always be a very large number of tied values on the X variable in the case of an experiment because every member within each treatment group has the same tied X rank.) Next, the Y scores are transformed to ranks and entered in column four. The rank transformations are easily accomplished

with the aid of any standard statistics software package. A conventional Pearson correlation coefficient is then computed between columns three and four. The Pearson correlation between the ranks on X and the ranks on Y is called the Spearman rank order correlation. It can be seen in the table that the Spearman rank order correlation is .452 and that this coefficient is statistically significant. It is concluded that there is a monotonic relationship between the order of the treatments and the associated population medians.

SBB Method

The SBB method differs from the method just described in that the hypothesis test itself is based on a computer intensive approach that differs from the conventional t on the Spearman coefficient. Specifically, the inferential test is based on bootstrap methodology that is described in McKean et al. (2001). The classic reference on the use of bootstrapping as the foundation for inference is Efron and Tibshirani (1993). This methodology has been implemented for the monotone test in Web based software that is available at http://www.stat.wmich.edu/.

The output from this program is shown below for the example data:

Value of Spearman's $\rho = .452160 \times 10^{0}$
p-value for bootstrap test $= .150000 \times 10^{-1}$
Lower confidence interval value, based on the bootstrap $= .168127 \times 10^{0}$
Upper confidence interval value, based on the bootstrap $= .720404 \times 10^{0}$
Value of Spearman's $\rho^2 = .204448 \times 10^{0}$
Lower confidence interval value for $\rho^2 = .282668 \times 10^{-1}$
Upper confidence interval value for $\rho^2 = .518982 \times 10^{0}$

Note that the Spearman coefficient is labeled as "Spearman's ρ" rather than r_s. Also, note the use of scientific notation. The results rewritten in conventional format are as follows:

$$r_s = .45$$
$$p\text{-value (directional)} = .015$$
$$90\% \text{ Confidence interval on } \rho = (.17, .72)$$
$$r_s^2 = .20$$
$$90\% \text{ Confidence interval on } \rho^2 = (.03, .52)$$

The squared Spearman coefficient may be of interest as a descriptive statistic. It describes the proportion of the sample variation in rank scores on the dependent

variable that can be explained given knowledge of treatment order. In the case of the example data we see that this proportion is approximately .20. The 90% confidence interval on the population proportion is (.03, .52).

As with the simpler test, it can be stated that there are sufficient data to convincingly support the hypothesis that population medians are a monotonic function of the order of the treatments. The conventional test on the correlation and the SBB approach lead to the same conclusion but the p-value obtained with the latter is slightly larger (.015 rather than .012); this is consistent with statistical theory. The conventional test is slightly liberal in the case of small N (say, ≤ 30).

16.5 REVERSED ORDINAL LOGISTIC REGRESSION

An additional method for the analysis of experiments with ordered treatment levels involves the use of a procedure that was designed for a completely different purpose. This procedure is known as ordinal logistic regression analysis. The conventional context for applying ordinal logistic regression involves continuous predictor variables and a dependent variable that consists of ordered categories. Hence, the typical application of ordinal regression appears to have no relevance to an ordered treatment study because the latter involves a dependent variable that is assumed to be approximately continuous.

But note that both the conventional ordinal regression application and the ordered treatment experiment are similar in that one variable is ordered. It just happens that we call this variable the independent variable in the case of the ordered treatment experiment whereas we call it the dependent variable in the case of conventional ordinal regression applications. In both cases we are interested in discovering whether there is convincing evidence of a relationship between a continuous variable and a variable with ordered categories. Because the overall purpose of the analysis is the same in both cases, we can reverse the role of the independent and dependent variables and apply ordinal logistic regression to data from an experiment with ordered treatment levels (Huitema, 2009). That is, when conventional ordinal logistic regression software commands are used, the actual ordered treatment categories become (for purposes of performing the analysis) the dependent variable and the actual dependent variable becomes the predictor. Hence, I call this approach the reversed ordinal logistic regression analysis.

Example 16.2 Consider the data in Table 16.2. When these data are rearranged we have the first two columns shown in Table 16.3. Note that the first column indicates the rank order of the five treatments and the second column contains the dependent variable scores; label these columns "Gp Rank" and Y, respectively. If we enter these two columns in any standard software package that performs ordinal logistic regression the actual roles of these two variables must be reversed. For example, with *Minitab* we enter "Gp Rank" in the "Response" window and "Y" in the "Model"

window. Of course, the command line editor can be used instead of the menu approach. The command line editor input that was used for the example data is shown below along with a small portion of the output.

Input Commands

```
MTB > set c1
DATA> 5(1) 5(2) 5(3) 5(4) 5(5)
DATA> end
MTB > print c1 c2
```

Output

Data Display

Row	Gp	Rank	Y
1		1	8
2		1	4
3		1	1
4		1	4
5		1	1
6		2	10
7		2	4
8		2	2
9		2	2
10		2	5
11		3	2
12		3	2
13		3	7
14		3	5
15		3	9
16		4	12
17		4	6
18		4	6
19		4	2
20		4	5
21		5	6
22		5	3
23		5	10
24		5	8
25		5	11

Input

```
MTB > OLogistic 'Gp Rank' = Y;
SUBC>   Logit;
SUBC>   Brief 2.
```

Output

```
Ordinal Logistic Regression: Gp Rank versus Y
Link Function: Logit
Log-Likelihood = -37.817
Test that all slopes are zero: G = 4.837, DF = 1,
P-Value = 0.028
```

The inferential portion of the output to which we attend is the *G*-test statistic and the associated *p*-value. The *G*-value is interpreted as a chi-square statistic; the *p*-value shown for *G* is based on the chi-square distribution with one degree of freedom.This *p*-value is not directional. If the directional *p*-value is desired simply divide the nondirectional *p*-value by two. This approach works in the present unusual application of ordinal logistic regression because the test statistic (i.e., *G*) is based on one degree of freedom; when two or more degrees of freedom are involved (as there may be in other ordinal regression applications) directional tests do not apply.

Although much additional output usually accompanies ordinal regression analysis, only a small portion is shown here because much of it is tangential to the present application. The overriding issue is whether there is convincing evidence of a monotonic relationship between the ordered treatments and the dependent variable; the *G* test statistic attends to this question. The output of most logistic regression programs includes several measures of association, but I do not recommend any of them. The squared correlation between the group variable and dependent variable is generally an adequate descriptive measure; it is far more transparent than the special purpose measures developed for logistic regression. An excellent reference on conventional applications of ordinal logistic regression is Hosmer and Lemeshow (2000).

Comparison of Methods for Analyzing Experiments with Ordered Treatments

Several methods have been mentioned for the analysis of experiments with ordered treatment levels. The application of these methods to the example data set suggests that there is high agreement among them with respect to the decision reached. The p_1-values obtained using the Jonckherre, Abelson–Tukey, Spearman-Based Bootstrap, and Reverse Ordinal Logistic Regression procedures are .02, .025, .015, and .014, respectively. These results are in stark contrast to those obtained using methods that do not acknowledge the order of the treatments. Parametric and nonparametric (i.e., Kruskal–Wallis) ANOVA tests yield *p*-values of .37 and .29, respectively. Although the *p*-values for the various monotonic methods are similar in the case of this particular example, there are several issues that one should keep in mind when choosing among these analyses.

First, a thorough comparative investigation of these methods has not yet been performed. Research comparing these procedures with respect to type I error, power, and robustness under a wide variety of sample sizes and distributions is needed. In the absence of such evaluative studies, arguments for choosing one method oven another must be based on known general properties of each method.

For example, there are large differences in robustness from method to method; this can be predicted from theory. Because the Abelson–Tukey method is based on parametric assumptions it is affected by outliers more than are the nonparametric methods. (This is easy to demonstrate by adding an outlier to the example data set and comparing methods.) Other issues such as transparency and software availability are also relevant in making a choice.

The transparency of an analysis should always be a high priority. Although all of these analyses are less familiar to research workers than are standard methods, there are large differences from one method to another regarding how the analysis works. It is easy to see that the Abelson–Tukey method, for example, is simply a test on a contrast between two weighted means, and that the weights are easily conceptualized. The Spearman-Based Bootstrap is similarly easy to conceptualize because the essence of the method is simply a test of the well-known Spearman correlation between the ordered independent and dependent variables. Both of these methods are easily generalized to include covariates (as shown in Chapter 17) and both have the additional advantage of easy power estimation. The property of transparency holds for neither the Jonckherre test nor the Reverse Ordinal Logistic Regression method. A practical problem with implementing the Reverse Ordinal Logistic Regression approach is that researchers often find the data entry procedure to be counterintuitive and confusing as a result of the reversed labels for the independent and dependent variables. Also, researchers who are unfamiliar with ordinal logistic regression software are often confused by the output, which differs substantially from output associated with OLS and binary logistic regression. The best choice among the monotone analysis methods will often be based on the extent to which the researcher has an intuitive understanding of the logic for the analysis.

16.6 SUMMARY

Independent variables are frequently classified as either quantitative or qualitative when an analytic procedure is selected. Those that are classified as quantitative are usually analyzed using regression whereas those that are classified as qualitative are usually analyzed using ANOVA. This dichotomous approach for classifying types of independent variables completely ignores the fact that many experiments have ordered treatment levels that are neither quantitative nor purely qualitative. If the order is not acknowledged in the analysis (as in the case of ANOVA) there are two disadvantages. First, the power of the test is lower than with methods that incorporate information regarding treatment order. Second, the conventional analysis is cumbersome and cluttered because it frequently includes unnecessary multiple comparison tests.

Some form of monotone analysis is recommended as the appropriate alternative to both simple regression analysis and ANOVA when the treatments are ordered. Monotone methods are generally far more powerful than ANOVA; they are also more parsimonious because they are focused on a single monotonic contrast. Both parametric and nonparametric methods of monotone analysis are available. The Abelson–Tukey method is the recommended parametric approach and the Spearman

Based Bootstrap is the recommended nonparametric method. A third alternative, the Reversed Ordinal Logistic Regression method is based on an unusual application of conventional ordinal logistic regression. Whereas the Abelson–Tukey test is based on a pre-specified weighted combination of treatment means, the general idea behind the Reversed Ordinal Logistic Regression approach is to determine whether the log odds of being above a specific treatment level category is a linear function the dependent variable. Evidence that treatment level is predictable from knowledge of dependent variable scores implies a monotonic relationship between the independent and dependent variables and that a treatment effect exists.

CHAPTER 17

ANCOVA for Ordered Treatments Designs

17.1 INTRODUCTION

Chapter 16 introduced monotone methods for the analysis of one-factor independent sample designs that have ordered treatment levels. In this chapter I introduce generalizations of monotone methods that are appropriate for designs having both ordered treatment levels and one or more covariates. The advantages of including covariates in monotone analysis are the same as in the case of ANCOVA applied to experiments with qualitative independent variables. That is, power is increased and effect estimates are adjusted to provide less ambiguous comparisons. Both parametric and nonparametric monotone methods are described; the considerations for choosing one type over the other are essentially the same as with designs not having ordered treatments.

17.2 GENERALIZATION OF THE ABELSON–TUKEY METHOD TO INCLUDE ONE COVARIATE

I begin with an example that was introduced in Section 16.1. Fante et al. (submitted) were interested in the effects of three types of training on verbal fluency. The training types were ordered on the basis of theory and previous research. Sixty subjects participated; each was randomly assigned to one of the three types of training. Before the treatments were carried out a typing test was administered. It was believed that chance differences between groups on typing skill could cloud the interpretation of differences observed on the dependent variable. This was a reasonable concern because the verbal fluency outcome was measured using timed typed responses. Hence, there was interest in using pretreatment typing speed (words per minute) as a covariate. This allowed an examination of the effects of the treatments on verbal fluency that was essentially uncontaminated by pretreatment differences in typing speed.

The Analysis of Covariance and Alternatives: Statistical Methods for Experiments, Quasi-Experiments, and Single-Case Studies, Second Edition. Bradley E. Huitema.

The analysis of this completely randomized one factor design was not a conventional ANOVA for two reasons. First, the treatments were ordered and there was interest in whether there was a monotonic relationship between the predicted order and the average dependent variable scores for the three groups. Second, potentially useful covariate information was available before the experiment was carried out. Both the order of the treatments and the covariate information are acknowledged in the analysis described next. I describe this analysis as an ANCOVA generalization of the Abelson–Tukey monotone method.

Abelson–Tukey Monotone Test for the Single Covariate Case

Certain aspects of three analyses with which you are already familiar are required in order to compute the monotone test statistic. These analyses are identified in the following preliminary steps: (1) compute ANOVA on the covariate X, (2) compute ANCOVA on Y, and (3) obtain the maximin coefficients from Table 16.1. The general form of the test statistic is as follows:

$$\frac{c_1(\bar{Y}_{1\text{ adj}}) + c_2(\bar{Y}_{2\text{ adj}}) + \cdots + c_J(\bar{Y}_{J\text{adj}})}{\sqrt{\text{MS}_{\text{Resw}}\left[\frac{c_1^2}{n_1} + \frac{c_2^2}{n_2} + \cdots + \frac{c_J^2}{n_J} + \frac{\left(c_1(\bar{X}_1) + c_2(\bar{X}_2) + \cdots + c_J(\bar{X}_J)\right)^2}{\text{SS}_{\text{W}_\text{X}}}\right]}} = t,$$

where

$c_1, c_2, \ldots c_J$ are the maximin contrast coefficients,
$\bar{Y}_{1\,\text{adj}}, \bar{Y}_{2\,\text{adj}}, \ldots \bar{Y}_{J\,\text{adj}}$ are the adjusted means,
$\bar{X}_1, \bar{X}_2, \ldots \bar{X}_J$ are the covariate means,
MS_{Resw} is the mean square residual within groups, and
SS_{Wx} is the within groups sum of squares on the covariate.

The obtained value of t is compared with the critical value of t based on $N - J - 1$ degrees of freedom. If $t_{\text{obt}} \geq t_{\text{critical}}$ then reject H_0: $\mu_{1\text{ adj}} = \mu_{2\text{ adj}} = \cdots = \mu_{J\text{ adj}}$; the alternative hypothesis is H_A: $\mu_{1\text{ adj}} \leq \mu_{2\text{ adj}} \leq \cdots \leq \mu_{J\text{ adj}}$.

A measure of association for this monotone analysis is:

$$\hat{\eta}^2_{\hat{\Psi}\text{ Maximin}} = \frac{\text{SS}_{\hat{\Psi}\text{ Maximin}}}{\text{SS}_{\text{Res}_\text{T}}},$$

where

$$\text{SS}_{\hat{\Psi}\text{ Maximin}} \approx \frac{\hat{\Psi}^2_{\text{Maximin}}}{\sum_{j=1}^{J} \frac{c_j^2}{n_j}},$$

(This method of estimating the sum of squares for the contrast is an approximation, but it is adequate for the current application),

$$\hat{\Psi}_{\text{Maximin}} = c_1(\bar{Y}_{1\text{ adj}}) + c_2(\bar{Y}_{2\text{ adj}}) + \cdots + c_J(\bar{Y}_{J\text{ adj}}), \quad \text{and}$$

Table 17.1 Example of Abelson–Tukey Test Applied to a Small Experiment with One Covariate

Ordered Treatment	Covariate Score (WPM)	DV Score (Fluency)
1	21.6	.8
1	23.8	.6
1	27.2	.6
2	31.4	1.4
2	32.8	4.0
2	34.2	2.6
3	33.4	5.0
3	26.4	2.2
3	22.0	3.2

SS_{Res_T} is the total residual sum of squares from the total regression of the dependent variable on the covariate (found in the ANCOVA summary table as the sum of the adjusted treatment sum of squares plus the within-group residual sum of squares).

The $\hat{\eta}^2_{\text{Maximin}}$ coefficient can be interpreted as the approximate proportion of the observed residual variation on the dependent variable (i.e., that variation not explained by the covariate) that is explained by the estimated maximin contrast.

A very small randomly selected subset of the complete ($N = 60$) fluency data (from Fante et al., submitted) is presented in Table 17.1 to illustrate the computations.

ANOVA on covariate X (WPM):

Source	SS	*df*	MS	*F*	*p*
Between	113.98	2	56.99	3.98	.08
Within	85.95	6	14.32		
Total	199.93	8			

Covariate means:

$$\bar{X}_1 = 24.2$$
$$\bar{X}_2 = 32.8$$
$$\bar{X}_3 = 27.267$$

ANCOVA on Y (Fluency):

Source	SS	*df*	MS	*F*	*p*
AT	8.54	2	4.27	3.91	.09
Res_W	5.45	5	1.09		
Res_T	13.99	7			

Adjusted means:

$$\bar{Y}_{1\text{ adj}} = 1.258$$
$$\bar{Y}_{2\text{ adj}} = 1.950$$
$$\bar{Y}_{3\text{ adj}} = 3.592$$

Maximin contrast coefficients (from Table 16.3):

$$c_1 = -.816$$
$$c_2 = .000$$
$$c_3 = .816$$

Test statistic:

$$\frac{-.816(1.258) + .000(1.950) + .816(3.592)}{\sqrt{1.09\left[\frac{-.816^2}{3} + \frac{.000^2}{3} + \frac{.816^2}{3} + \frac{[-.816(24.2) + .000(32.8) + .816(27.267)]^2}{85.95}\right]}} = 2.54 = t_{\text{obt}}$$

Critical value of $t = 2.015$ for $\alpha_1 = .05$ and $df = 5$.

p_1-value (from *Minitab*) $= .026$.

Decision: Reject H_0 and conclude that there is a monotonic increasing relationship between the order of the treatments and the mean population fluency scores.

The sum of squares for the contrast and the η-squared statistic are:

$$\text{SS}_{\hat{\Psi}\text{ Maximin}} = \frac{1.905^2}{.444} = 8.18 \quad \text{and}$$

$$\hat{\eta}^2_{\hat{\Psi}\text{ Maximin}} = \frac{\text{SS}_{\hat{\Psi}\text{ Maximin}}}{\text{SS}_{\text{Res}_\text{T}}} = \frac{8.18}{13.99} = .58$$

Approximately 58% of the observed variation in fluency scores that is not explained by the covariate is explained by the maximin contrast between adjusted treatment means.

Although Table 17.1 illustrates the analysis using maximin coefficients, the same t-value is obtained if the more intuitive three group coefficients (described in Chapter 16) are used. The contrast estimate using the latter approach is simply the difference between the first and third adjusted means (2.33 points in the example).

17.3 ABELSON–TUKEY: MULTIPLE COVARIATES

Just as it is possible to use more than one covariate in the analysis of experiments with qualitative treatment variables, it is possible to generalize the Abelson–Tukey

procedure to handle multiple covariates. The test statistic for the multiple covariate generalization of the Abelson–Tukey procedure is presented below.

$$\frac{c_1(\bar{Y}_{1\text{ adj}}) + c_2(\bar{Y}_{2\text{ adj}}) + \cdots + c_J(\bar{Y}_{J\text{ adj}})}{\sqrt{\text{MS}_{\text{Resw}}\left[\frac{c_1^2}{n_1} + \frac{c_2^2}{n_2} + \cdots + \frac{c_J^2}{n_J} + \mathbf{d}^{\text{T}}\mathbf{W}_{\text{X}}^{-1}\mathbf{d}\right]}} = t$$

where

$c_1, c_2, \ldots c_J$ are the maximin contrast coefficients,
$\bar{Y}_{1\text{ adj}}, \bar{Y}_{2\text{ adj}}, \ldots \bar{Y}_{J\text{ adj}}$ are the adjusted means,
MS_{Resw} is the mean square residual within groups from the multiple ANCOVA,

$$\mathbf{d} = \begin{bmatrix} c_1\bar{X}_{1,1} + c_2\bar{X}_{1,2} + \cdots + c_J\bar{X}_{1,J} \\ c_1\bar{X}_{2,1} + c_1\bar{X}_{2,2} + \cdots + c_1\bar{X}_{2,J} \\ \cdot \quad \cdot \quad \cdots \quad \cdot \\ \cdot \quad \cdot \quad \cdots \quad \cdot \\ \cdot \quad \cdot \quad \cdots \quad \cdot \\ c_1\bar{X}_{C,1} + c_1\bar{X}_{C,2} + \cdots + c_1\bar{X}_{C,J} \end{bmatrix} = \begin{bmatrix} d_1 \\ d_2 \\ \cdot \\ \cdot \\ \cdot \\ d_C \end{bmatrix},$$

(where the first and second subscripts on the sample covariate means denote the covariate and the group, respectively),

where

$\mathbf{d}^{\text{T}}$ = transpose of vector $\mathbf{d}$,
$\mathbf{W}_{\mathbf{X}}^{-1}$ = inverse of matrix $\mathbf{W}_{\text{X}}$, where

$$\mathbf{W}_{\mathbf{x}}\begin{bmatrix} \sum x_1^2 & \sum x_1x_2 & \ldots & \sum x_1x_C \\ \sum x_2x_1 & \sum x_2^2 & \ldots & \sum x_2x_C \\ \cdot & \cdot & \cdots & \cdot \\ \cdot & \cdot & \cdots & \cdot \\ \cdot & \cdot & \cdots & \cdot \\ \sum x_Cx_1 & \sum x_Cx_2 & \ldots & \sum x_C^2 \end{bmatrix}.$$

(This is the deviation (i.e., centered) score *within* group sum of products matrix for the covariates.)

17.4 RANK-BASED ANCOVA MONOTONE METHOD

The method described in this section is one of three robust approaches that can be used when covariates are included in the design. It is based on a multiple regression

approach applied to ranked data. The major advantage of this method relative to the other robust approaches is that it is relatively easy to understand and carry out; it can be computed with standard statistical software. The disadvantages are that (1) it provides hypothesis tests only (no confidence intervals are provided), and (2) it has less theoretical and diagnostic support than does the robust method described in Section 17.7. I label the approach of this section as the rank-based ANCOVA monotone method and the method of Section 17.7 as the robust ANCOVA *R*-estimate monotone method. A thorough head-to-head comparison of the statistical properties of these two methods has not yet been carried out.

The rank-based method is performed using the following steps:

1. Assign the score 1 to each subject in the treatment group predicted to have the lowest dependent variable score, assign the score 2 to each subject in the group predicted to have the second to the lowest dependent variable score, and so on; assign the value equal to J to each subject in the group predicted to have the highest dependent variable score.
2. Rank the group scores that are described in step 1 (this is easily accomplished using *Minitab* or *SPSS*); denote the resulting variable as G_{Rank}.
3. Rank the covariate scores; denote the resulting variable as X_{Rank}.
4. Rank the dependent variable scores; denote the resulting variable as Y_{Rank}.
5. Fit the following regression model:

$$Y_{\text{Rank } i} = \beta_0 + \beta_1(G_{\text{Rank } i}) + \beta_2(X_{\text{Rank } i}) + \varepsilon_{\text{Rank } I},$$

   where $Y_{\text{Rank } i}$, $G_{\text{Rank } i}$, and $X_{\text{Rank } i}$ are the rank scores on Y, G, and X for subject i.
6. Interpret the conventional test of significance for the first partial regression coefficient b_1 (the estimate of β_1) as an approximate test of H_0: $\theta_{1\text{ adj}} = \theta_{2\text{ adj}} = \cdots = \theta_{J\text{ adj}}$ (where $\theta_{j\text{ adj}}$ denotes the rank-covariate adjusted median for the jth population). Equivalently, the null hypothesis may be written as H_0: $\beta_1 = 0$. If the null hypothesis is rejected this is evidence that there is a monotonic relationship between the ordered treatments variable and the dependent variable holding constant the covariate rank. If the b_1 coefficient is positive the alternative hypothesis can be written as H_A: $\theta_{1\text{ adj}} \leq \theta_{2\text{ adj}} \leq \cdots \leq \theta_{J\text{ adj}}$ and, equivalently, as H_A: $\beta_1 > 0$.

Tests of significance regarding all regression coefficients are provided by virtually all conventional regression software. Although these tests are based on normal distribution theory that assumes continuous normally distributed observations, there is evidence that the properties of these tests are usually satisfactory when applied to ranked data (Conover and Iman, 1981).

The application of these steps to the fluency data presented in Table 17.1 yields the ranks shown in Table 17.2 and the results that follow.

The regression of Y_{Rank} on G_{Rank} and X_{Rank} produces the following coefficients, test statistics, and p-values:

Coefficient	t-value	p-value
$b_0 = -.299$	$-.200$	.85
$b_1 = .746$	3.189	$<.02$
$b_2 = .313$	1.412	.21

The first partial regression coefficient (b_1) and the associated p-value are the most relevant portions of these regression results. Note that the sign of the coefficient is positive and the p-value (nondirectional) is less than .02. Hence, H_0: $\beta_1 = 0$ is rejected (also, H_0: $\theta_1 = \theta_2 = \theta_3$ is rejected) and it can be concluded that when the rank on the covariate (WPM) is held constant the evidence for a monotonic increasing relationship between the independent variable (treatment order) and the population medians on the dependent variable (fluency) is convincing.

An equivalent approach is to compute the partial correlation between Y_{Rank} and G_{Rank} holding constant X_{Rank}. The significance test on this partial correlation coefficient provides the same p-value as the test on the first partial regression coefficient described above using regression analysis. An advantage of the correlation approach is that the squared partial correlation coefficient provides a measure of monotonic effect size. In the case of the fluency example this squared value is .63. It can be interpreted as the proportion of the variation on the dependent variable ranks that is explained by the ordered independent variable, holding constant the covariate ranks. In the case of the fluency data the value of this squared partial correlation coefficient is similar to the value of $\hat{\eta}^2_{\text{Maximin}}$ (i.e., .58) that was obtained using the parametric maximin contrast approach.

Table 17.2 Ranks Based on Fluency Data in Table 17.1

G_{Rank}	X_{Rank}	Y_{Rank}
2	1	3
2	3	1.5
2	5	1.5
5	6	4
5	7	8
5	9	6
8	8	9
8	4	5
8	2	7

17.5 RANK-BASED MONOTONE METHOD WITH MULTIPLE COVARIATES

The procedure described in the previous section is easily extended to the case of multiple covariates. The independent variable, the dependent variable, and all covariates are transformed to ranks; the ranked dependent variable is regressed on the ranked independent variable and all ranked covariates. As in the single covariate case, the focus is on the resulting estimate of the first partial regression coefficient and its p_1-value.

17.6 REVERSED ORDINAL LOGISTIC REGRESSION WITH ONE OR MORE COVARIATES

Chapter 16 describes the Reversed Ordinal Logistic Regression approach for analyzing ordered treatment experiments having no covariates. This approach is easily extended to handle one or multiple covariates.

When no covariates are involved, the general idea is to determine whether the log odds of being above a specific ordered treatment category is a linear function of the dependent variable. This is determined by reversing the roles of the independent and dependent variables and carrying out an ordinal logistic regression analysis. That is, the analysis is carried out by specifying the treatment group ranks as the dependent variable (for purposes of performing the analysis) and the actual dependent variable is specified as the predictor variable.

Both the purpose of the analysis and the procedure are similar if a covariate is involved. In this case we want to determine if the log odds of being above a specific treatment category is a linear function of the dependent variable, holding constant the covariate. This is accomplished by regressing the ranks of the treatment groups on both the actual dependent variable and the covariate using ordinal logistic regression. The test on the coefficient associated with the actual dependent variable is the test of interest. I recommend that the test on this coefficient be carried out using a likelihood-ratio model comparison approach rather than the conventional z statistic or Wald χ^2 (one degree of freedom) statistic produced by typical ordinal regression software. The Wald test is less trustworthy in this application than is the likelihood-ratio model comparison approach recommended here. The recommended approach is described below.

Example 17.1 Consider the data presented in Table 17.1. The first column contains the treatment order associated with each subject. This column should be specified in the "dependent variable" field of relevant ordinal logistic regression software (such as *Minitab* or *SPSS*). Next, enter columns two and three (i.e., the covariate "words per minute" and the actual dependent variable "fluency") as predictors. In the case of *SPSS*, there is no field-labeled "predictors." Instead there are two fields; one is labeled "factor(s)" and the other is labeled "covariate(s)." Both the actual covariate and the

actual dependent variable are entered in the "covariate(s)" field. That is, for purposes of this analysis both "words per minute" and "fluency" are classified as covariates.

The likelihood-ratio test on the coefficient for the actual dependent variable (fluency) can be carried out using the following steps:

1. Regress the ordered treatment variable on both the actual dependent variable (fluency) and the covariate (WPM); the ordinal logistic regression output from fitting this "full" model will include a χ^2 test statistic for the overall "full model" regression. *Minitab* labels this statistic as the "*G* test" of the hypothesis that all slopes are zero; *SPSS* labels this statistic as the "final χ^2." The obtained value of this statistic provided by both programs is 8.414.
2. Regress the actual dependent variable on the covariate alone; the ordinal logistic regression output from this regression will include a G or χ^2 test statistic for this "restricted model." The value of this statistic as provided by both programs is 0.221.
3. Subtract the obtained restricted-model χ^2 from the obtained full-model χ^2 to compute the likelihood-ratio χ^2 test statistic on the adjusted monotonic treatment effect. That is, $(\chi^2_{\text{Full}} - \chi^2_{\text{Restricted}}) = \chi^2_{\text{Adj Tx}}$. The value of this test statistic is (8.414 – .221) = 8.193; this is used to evaluate whether there is a monotonic effect of the treatments on the dependent variable, holding the covariate constant. One degree of freedom is associated with this χ^2 test because the full-model $df = 2$ and the restricted-model $df = 1$; hence, the df for the model comparison = 2 – 1 = 1. The one-tailed p-value associated with this test is <.01.

A comparison of the results from this method with those from the Abelson–Tukey and rank-based monotone ANCOVA methods reveals identical decisions for all procedures (assuming $\alpha = .05$) and similar but not identical p-values. Directional p-values for the Abelson–Tukey, rank-based, and reversed ordinal logistic methods are .03, < .01, and <.01, respectively.

17.7 ROBUST *R*-ESTIMATE ANCOVA MONOTONE METHOD

The method of this section is essentially a combination of the Abelson–Tukey procedure and a robust general linear model described by Hettmansperger and McKean (1998) and McKean (2004). Specifically, it incorporates maximin contrast coefficients and robust estimates of location and error variance. The advantage of this combined approach over the rank methods described in Sections 17.3 and 17.4 is that the robust general linear model brings with it the whole armamentarium of the general linear model framework, including theory and diagnostics for evaluating the aptness of the model. A disadvantage is that the computation requires routines that are not currently available in the most popular statistics software packages. A version running in the *R* computing environment (Terpstra and McKean, 2004) is currently available. Another

potential disadvantage is slightly lower power than competing procedures when the assumptions for the parametric test are met.

The steps involved in computing the analysis are listed below:

1. Compute the adjusted measures of location using a robust *R*-estimation routine.
2. Compute the maximin contrast by applying the maximin contrast coefficients to the robust adjusted measures of location.
3. Compute t using:

$$\frac{c_1(\bar{Y}'_{1\text{ adj}}) + c_2(\bar{Y}'_{2\text{ adj}}) + \cdots + c_J(\bar{Y}'_{J\text{ adj}})}{\sqrt{\hat{\tau}^2\left[+\dfrac{\left(c_1(\bar{X}_1) + c_2(\bar{X}_2) + \cdots + c_J(\bar{X}_J)\right)^2}{\text{SS}_{\text{W}_\text{X}}}\right]}} = t,$$

where $\hat{\tau}^2$ is the estimate of scale associated with the estimated robust linear model and $\bar{Y}'_{j\text{ adj}}$ is a robust adjusted mean. (These estimators are provided in the output of the *RGLM* software used to fit the ANCOVA model; see Chapter 14 for an example from this routine.) The degrees of freedom $= N - J - 1$, where N is the total number of observations and J is the number of groups.

17.8 SUMMARY

The advantages of ordered treatments analysis described in Chapter 16 are combined with the advantages of ANCOVA in this chapter. Four different methods of combining the two are described. The first method is essentially an application of a modified version the parametric Abelson–Tukey method to adjusted means from ANCOVA, but with an adjusted error term. The second method is a simple rank-based approach that can be computed using ordinary multiple regression software. The third approach involves the use of a reversed ordinal logistic regression model. The fourth approach utilizes recently developed robust linear model estimation. Each method can be expected to have power advantages over methods that do not incorporate a covariate; each method can incorporate more than one covariate.

PART V

Single-Case Designs

CHAPTER 18

Simple Interrupted Time-Series Designs

18.1 INTRODUCTION

The focus of this chapter is on two forms of the simple interrupted time-series design. The first form is a single-case design that involves obtaining measurements from a single subject at many time points before and after the introduction of an intervention. The second form also involves the collection of observations at many time points before and after an intervention, but the observations are obtained from a compound unit that consists of more than one organism. The recommended analysis for both forms can be viewed as a variant of ANCOVA in which time is the covariate.

Single-Case Designs

The intensive and systematic experimental study of individual subjects has a long history in both medicine and psychology. Early examples in medicine can be found in Claude Bernard's *Introduction to the Study of Experimental Medicine* (1865). At about the same time when Bernard's book was published, the foundations of experimental psychology (in areas such as psychophysics, memory, and physiological psychology) were developed using the intensive study of one or very few subjects. But, by the beginning of the twentieth century, the individual differences movement and the subsequent development of randomized-group designs and associated methods of statistical inference led to a dramatic fall in the popularity of methods using single subjects. By the 1950s, the intensive study of single subjects was viewed as rather odd in many mainstream medical and psychology journals. Indeed, most of these journals systematically rejected manuscripts presenting results based on these designs.

The Analysis of Covariance and Alternatives: Statistical Methods for Experiments, Quasi-Experiments, and Single-Case Studies, Second Edition. Bradley E. Huitema.

One area within psychology, however, resisted the focus on group experiments and associated statistical analyses. Sophisticated single-case research designs (which are very different than "case studies") were developed, refined, and promoted by behavioral psychologists (e.g., Sidman, 1960; Skinner, 1956). The most popular variants of these designs include the reversal design, the multiple baseline design, and the alternating treatments design. The essential elements of these designs are now often combined to yield many flexible hybrid versions. Journals that encourage the publication of data based on these designs have been active for over 40 years (e.g., the *Journal of Applied Behavior Analysis* and the *Journal of the Experimental Analysis of Behavior*). In medicine another design variant, sometimes described as an *n*-of-1 randomized trial is occasionally used, but it does not have the popularity of the single-subject designs used in psychology.

Although single-subject designs have been largely invisible in the major journals in psychology and the medical sciences, it appears that they are gaining acceptance. Over a dozen behavioral science textbooks that provide thorough coverage of these designs (e.g., Barlow et al., 2009; Cooper et al., 2006; Johnston and Pennypacker, 2008) are now available. In addition, a steady stream of expository articles that describe properties of these designs can be seen in both behavioral science and medical science journals (e.g., Guyatt et al., 1990; Lundervold and Belwood, 2000; Madsen and Bytzer, 2002; Morgan and Morgan, 2001; Senn, 1998; Stricker and Trierweiler, 1995). Still, the acceptance of these designs by funding agencies and mainstream journals does not equal that of conventional group designs. Although many general research methods courses in the behavioral sciences include brief mention of single-case research, the major experimental design textbooks in psychology and medicine contain essentially no coverage of these designs. Reasons for this neglect appear to include convention, lack of awareness, failure to distinguish between case studies and structured intervention studies, and an absence of adequate and widely accepted statistical methods.

Statistical Methods for Single-Case Designs

Many statistical analyses have been proposed for single-case research, but none of them is popular in the disciplines that most frequently use these designs. One reason this is true is that visual analysis of the graphed data (instead of visual analysis supplemented with formal statistical analysis) has been and continues to be the approach recommended in single-case design textbooks (e.g., Cooper et al., 2006; Johnston and Pennypacker, 2008); it is the approach encouraged in the major behavioral journals. Although arguments for the statistical analysis of this type of research have been made for many years (e.g., Huitema, 1986b), the proportion of single-case experiments analyzed using formal statistical methods is small.

Three justifications for statistical analysis of single-subject designs include: (1) appropriate analyses tend to have high credibility throughout the scientific community, (2) many granting agencies and journals often demand formal statistical results, and (3) single-case studies that report no effect metric are likely to be ignored in

meta-analytic studies. Unfortunately, several single-case statistical methods that have been recommended in social science textbooks and journals contain fatal flaws and are unacceptable, even in the context of simple two-phase designs. Critiques of some of these procedures can be found elsewhere (e.g., Huitema, 2004; Huitema and McKean, 2000b; Huitema et al., 2008).

An inspection of the few existing textbooks that mention statistical analyses for single-case designs reveals a common weakness with nearly all of them. The analyses are presented and illustrated for only the elementary two-phase (AB) design. This variant is considered anathema by behavioral researchers who usually use more sophisticated and convincing single-case experiments.

Although the two-phase single-case design is not considered credible by some researchers (primarily because it is highly vulnerable to the internal validity threat of history), it is an important design for at least three reasons. First, it is simple to implement. Second, it is the essential building block for more convincing reversal and multiple baseline designs. Third, it is similar to the interrupted time-series quasi-experiment that has a long history of acceptance in the literature on quasi-experimentation.

Single-Case AB Designs Versus Interrupted Time-Series Quasi-Experiments

Two-phase single-case designs and interrupted time-series quasi-experiments have much in common. Both involve the systematic collection of data at many time points before and after the introduction of some intervention. The key distinction between the two is the nature of the unit that provides the response measures to be analyzed. Single-case designs (also called single-subject, single-participant, or operant designs) usually refer to the situation where time-series data are obtained from one person or organism. The interrupted time-series quasi-experiment usually refers to a two-phase design in which the pre- and postintervention time-series data are provided by some compound unit such as a classroom, a city, or a county.

Suppose there is interest in changing a single person's frequency of unsafe job behavior. A trained observer collects data for 15 days before and 14 days after a planned intervention is introduced. The sequence of 29 observations constitutes a time-series and some method of data analysis is required to evaluate the size and nature of possible change after the intervention. Note that each of the 29 observations is based on a single subject. Contrast this with a study where there is interest in evaluating the effects of some public health intervention on the annual cancer death rate reported for an entire county. Although the county (a compound unit) may include thousands of individuals, the researcher simply records a single number (i.e., the number of residents who died of cancer) each year before and after the intervention. The resulting collection of numbers constitutes the interrupted time-series to be analyzed.

The composition of a compound unit may be the same during all measurement occasions; it may change partially, or it may change completely (as when studying the annual number of car accidents reported for 16-year-old drivers in a specific state).

But regardless of the size or composition of the unit of interest, the purpose of the analysis is to measure and evaluate change associated with an intervention to that unit. Internal and external validity issues, however, are not the same when the unit is a single-subject as it is when a compound unit is involved. Knowledge regarding the type of subjects studied and the between-subject variation in outcome are among the key issues in establishing the generality of results of an intervention. This variation may be investigated using several different experiments, each with a single subject, or by using a single experiment containing multiple subjects.

The present chapter describes methods for the analysis of two-phase designs where the unit of analysis is either an individual subject or some compound unit. Chapter 19 provides detailed examples of the application of two-phase analyses. Chapter 20 describes analyses for the reversal design and Chapter 21 introduces analyses for three versions of the multiple baseline design. Familiarity with the essentials of conducting research using single-case designs is presumed throughout these chapters. Useful references on design aspects (rather than analysis) include Barlow et al. (2009) and Cooper et al. (2006).

18.2 LOGIC OF THE TWO-PHASE DESIGN

An understanding of the essential descriptive parameters associated with the analysis of two-phase (AB) studies rests on the logic of the design. Data in the baseline phase provide measures of initial level, variability, trend, and serial dependency. These measures provide a basis for projections of what is expected to occur during the second phase in the absence of an intervention; such projections are sometimes called counterfactuals. The researcher's interest lies in the difference between counterfactual measures and measures that are based on what actually occurs after the intervention is introduced. Although change with respect to any of the measures mentioned above may be of interest, the major focus is usually on two change measures. The first is known as *level change* and the second is known as *slope change*. Although the interpretation of both of these measures is straightforward, level change is frequently misconceptualized, mislabeled (as intercept change), and incorrectly computed (Huitema, 2004; Huitema and McKean, 2000a).

Level Change

One possible measure of level change indicates the amount by which an intervention changes the expected value of the response at the beginning of the intervention phase relative to the expected value predicted from the data in the first (baseline) phase. If there are n_1 observations in the first phase and n_2 observations in the second phase, the first observation in the intervention phase occurs at time $n_1 + 1$. The level change reasonably can be defined (under the assumption that an adequate model describes the data for each phase) as the difference between (a) the predicted (counterfactual) value of Y at time $n_1 + 1$ based on a model of the first phase data and (b) the expected value of Y at time $n_1 + 1$ based on a model of the second-phase data. It is

crucial that both of these estimates be associated with exactly the same time point (viz., $n_1 + 1$). Although various time-series intervention models may use different procedures to compute the two level estimates, all acceptable procedures estimate level change at a common time point. Surprisingly, several time-series methods have been promoted that do not follow this rule; critiques of these methods are available elsewhere (Huitema, 2004; Huitema et al., 2008).

It is important to be aware that the concept of level change does not necessarily refer to the difference between the means of the two phases. Level change refers to a shift in elevation at time $n_1 + 1$ that is unexplained by possible within-phase trends. When the slope is zero within each phase the difference between phase means is equal to the level change, but if nonzero slope is present in either or both phases then level change is not the same as the mean difference. In the latter case, the mean difference is misleading as a measure of the effect of the intervention because (1) it does not convey the fact that the size of the effect (if any) is a function of the within-phase time period, (2) it may be large even when there is no effect whatsoever, and (3) it may not reveal an effect when one is present.

Slope (Trend) Change

Slope change provides the second major way of characterizing the effect of an intervention. Here the term slope has its traditional meaning. It simply refers to the average change on Y given a one-unit change on X, where, in the time-series context, the X variable is time. If the intervention has an effect, it may produce a change in level, a change in slope, or both. Because a reasonable representation of intervention effects often requires measures of both level change and slope change, an adequate descriptive analysis will usually include accurate estimates of both of them. Although interventions can also interrupt the structure of time-series data by changing the variance or in other more subtle ways (e.g., Stoline et al., 1980), level change and slope change provide the two most basic effect measures.

Illustrations of level change and slope change are presented in Figure 18.1. The elevation of the tip of the arrow at the right end of the first phase indicates the level of that phase. The elevation of the tip of the arrow at the left end of the second phase indicates the level for phase two. Hence, the difference in elevation of the tips of the two arrows indicates level change. Panel A illustrates a step-function level change (i.e., the full change takes place immediately), but there is no trend in either phase so there is no slope change. Panel B illustrates a large level change (even though there is no change in the phase means), but no slope change. Panel C illustrates a clear change in slope, but the level does not change. Panel D illustrates large change in both level and slope.

18.3 ANALYSIS OF THE TWO-PHASE (AB) DESIGN

Many statistical methods have been proposed for the analysis of change in two-phase studies. Among them are ARIMA models and time-series regression models.

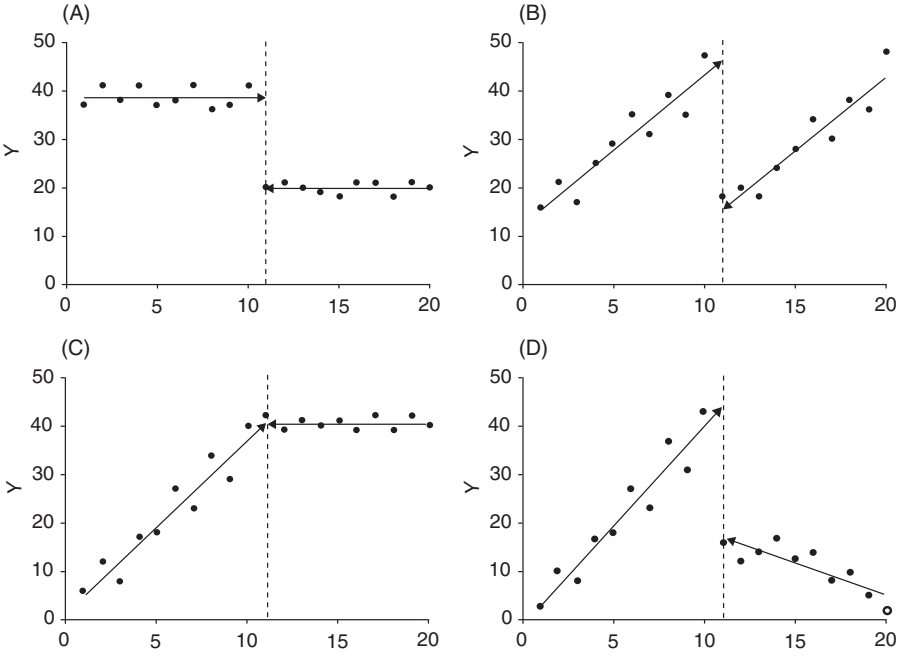


Figure 18.1 Two-phase designs illustrating four outcome patterns: (A) large level change but no change in slope, (B) large level change, steep slope in both phases but no change in slope, (C) no level change, large change in slope, and (D) large level change and large slope change.

ARIMA Intervention Models

A popular approach for the analysis of time-series data of various forms is widely known in the statistical community as ARIMA (*a*uto*r*egressive *i*ntegrated *m*oving *a*verage) modeling. Versions of ARIMA modeling that were developed for the interrupted time-series design (Box and Tiao, 1965, 1975) are sometimes called "intervention analysis." These methods have long been recommended for the analysis of single-case designs (e.g., Glass et al., 1975).

This approach requires one to build (i.e., identify, fit, and diagnose) a model that explains the response variable as a function of autoregressive, moving average, differencing, and intervention parameters as discussed in Box et al. (2008). Although the ARIMA model building approach is dominant in the statistical analysis of general time-series data, it has three disadvantages in the context of typical two-phase experiments. First, the recommended minimum number of observations in the time-series for building a model is 50–100. Few single-case studies approach this minimum. While it appears that this disadvantage has sometimes been overstated in the single-subject metholological literature, it is true that a small number of observations lead to the dual problems of (a) ambiguity regarding the identification of the most appropriate model and (b) minor bias in the estimation of time-series parameters. However,

because the effect estimates based on various plausible ARIMA models are usually quite similar, the identification problem need not disqualify ARIMA modeling. The second disadvantage is the complexity of the method. Although the computational aspects are easily handled with appropriate software, a sound understanding of the parameter estimates associated with these analyses requires substantial training. Even seasoned researchers (and a few authors of expository articles describing these methods) occasionally misinterpret the somewhat opaque parameters. The third disadvantage is that the focus of the analysis is on valid statistical inference rather than clear description of the observed data. The first purpose of a statistical analysis is to provide a transparent view of the outcome of the study. The parameters of a conventional ARIMA interrupted time-series model do not necessarily provide this transparency.

The transparency problem is most easily observed in the case of data that contain deterministic trend in either or both of the phases. This is a situation where conventional ARIMA models are less than ideal; they are better suited to the case of stochastic trend (although it is possible to specify an ARIMA model with deterministic trends). Unfortunately, deciding whether trend is deterministic or stochastic is challenging, even using the most recent and sophisticated diagnostic tests available for this purpose. This is especially true in the case of short series.

One of the first steps in ARIMA modeling of data that may have trend is to apply a low-order differencing transformation to remove the trend; the remainder of the estimation aspects of the analysis (e.g., fitting coefficients that describe intervention, autoregressive, and/or moving average parameters) is performed on the transformed data. The view of trend under ARIMA modeling is that it is a nuisance to be eliminated before the remainder of the analysis can be completed. (Section 19.4 presents an example and additional detail on this issue.) I prefer to view description as primary and inference as secondary; an approach consistent with this view is described next.

Time-Series Regression Models

The approaches described in the remainder of this chapter explicitly describe the observed data (including trend) using appropriate terms in a regression model. When the errors of such a model violate the independence assumption an expanded regression model that accommodates the dependency is applied. Models of this type are relatively simple to implement, interpret, and extend to complex designs. The intervention effect parameter estimates provided by this approach are consistent with the way single-case researchers usually conceptualize change. A benefit of this consistency is acceptance of these methods by researchers.

Four time-series regression models for the analysis of two-phase designs are described in this section. They differ in terms of (a) the number of parameters used to describe the intervention effects and (b) the assumed nature of the errors. Model I contains four parameters; the focus of this model is on both level change and slope change. Model II contains two parameters; the focus of this model is on level change alone. Independent errors are assumed for models I and II, but extensions of these models are available when the errors appear to be dependent. Model III is similar

Table 18.1 Four Intervention Models Characterized by Type of Intervention Effect Parameters and Assumption Regarding Relationship Among Errors

	Type of Intervention Parameters	
Assumed Relationship among Errors	Level Change and Slope Change	Level Change Only
Independent	Model I: H–M four-parameter	Model II: two-parameter
Autocorrelated	Model III: H–M four-parameter	Model IV: two-parameter

to model I in that the focus is on both level change and slope change, but the errors are assumed to follow an autoregressive process. Model IV is similar to model II in that the focus is on level change alone, but the errors are assumed to follow an autoregressive process. These characteristics are summarized in Table 18.1.

Many two-phase studies can be adequately characterized using one of these intervention models. Two strategies for selecting one of them are described in the following section.

18.4 TWO STRATEGIES FOR TIME-SERIES REGRESSION INTERVENTION ANALYSIS

Strategy I: Simply Fit Model III

One strategy for selecting an analysis of two-phase data is to routinely fit model III. This simple strategy is often satisfactory, especially if the sample size is relatively large in each phase. Because model III includes the parameters necessary to describe level change, slope change, and autocorrelation among the errors (whether autocorrelation is present or not), it can accommodate many types of two-phase data. Once the data are plotted and carefully inspected, the essentials of the analysis are complete in one pass through an appropriate computer routine that fits regression models with autoregressive errors. Two routines of this type are described in Chapter 19.

Strategy II: Two-Stage Model Selection

Although strategy I is easy to apply and often provides a satisfactory solution, my general recommendation is to use strategy II instead. The two-stage approach is more cumbersome to implement, but it has three advantages over strategy I that make it worth the additional trouble. First, the interpretation of the results may be simplified if the chosen model contains fewer parameters than are included in model III. Second, the power of the tests for intervention effects is likely to be much higher when fewer parameters are required. Third, the whole analysis often can be carried out using conventional OLS regression routine rather than specialized software required to apply strategy I.

The first stage (A) of strategy II involves fitting models I and II using OLS, and then performing a model comparison test to select one of the two. The second stage (B) involves an evaluation of the errors of the selected model by applying conventional regression diagnostics to the residuals to determine whether they are reasonably consistent with the assumptions of linearity, homoscedasticity, and normality. If serious violations of any of these assumptions are identified, conventional regression remediation procedures are applied. An additional evaluation of the residuals is then carried out to evaluate conformity with the assumption of independent errors. If the errors of the chosen model (i.e., I or II) appear not to be dependent, that model is adopted. If dependency among errors is identified, the corresponding two- or four-parameter model that includes an autoregressive component (viz., model III or model IV) is adopted instead. Details regarding the two stages (A and B) of strategy II are described in the following section.

18.5 DETAILS OF STRATEGY II

Stage A: Preliminary Modeling

The first stage involves plotting the raw data, fitting models I and II, and computing a model comparison test to identify the more appropriate of the two models.

Model I

Model I is written as follows:

$$Y_t = \beta_0 + \beta_1 T_t + \beta_2 D_t + \beta_3 SC_t + \varepsilon_t,$$

where

Y_t is the dependent variable score at time t;
β_0 is the process intercept;
β_1, β_2, and β_3 are the process partial regression coefficients;
T_t is the value of the time variable T at time t;
D_t is the value of the level-change dummy variable D (0 for the first phase and 1 for the second phase) at time t;
SC_t is the value of the slope-change variable SC defined as $[T_t - (n_1 + 1)]D$
n_1 is the number of observations in the first phase; and
ε_t is the error of the process at time t.

The errors are assumed to be independent, normally distributed, have a mean of zero at each time point, and have the same variance at all time points.

It can be seen that this model has much in common with conventional multiple regression. The most obvious difference is that this is a model of responses measured at N time points (n_1 in the first phase and n_2 in the second) on a single unit rather than a model of the response of many subjects measured at one time point. For this reason the subscript t is used to denote the time point of measurement in a sample of

N time points whereas conventional regression notation uses the subscript i to denote a subject in a sample of N different subjects.

Because this four-parameter model measures both level change and slope (trend) change, it accommodates all of the outcome patterns illustrated in Figure 18.1. Any multiple linear regression routine can be used to fit the model and perform inferential aspects of the analysis. The key issue in correctly implementing this model is the specification of the design matrix.

Because there are major misunderstandings in the literature (in both the behavioral sciences and medicine) regarding the appropriate specification of the predictor variables for four-parameter intervention models, a correct specification (mentioned above) is reiterated and described with additional detail below; examples are provided in the next chapter. Detailed discussions of the problems associated with popular alternative methods used to define the predictor variables are available elsewhere (Huitema, 2004; Huitema and McKean, 2000a, 2000b; Huitema et al., 2008, submitted).

The parameters of model I are estimated by regressing Y on the set of three predictor variables (T, D, and SC). The predictor variables are specified as follows:

Time variable (T): The time value sequence $1, 2, \ldots, N$ (where $n_1 + n_2 = N$) is the first predictor.

0–1 Treatment dummy variable (D): The value zero is entered for each observation in the baseline phase and the value one is entered for each observation in the intervention phase.

Slope change variable (SC): Each value in the slope change column is computed using $SC_t = [T_t - (n_1 + 1)]D_t$.

When the predictor variables are defined as shown here, the four coefficients of the model are interpreted as described below.

The intercept β_0 is the expected elevation on Y at time period zero (the time period that occurs immediately before the first measurement is obtained). The first partial regression coefficient β_1 is the slope in the baseline phase. The second partial regression coefficient β_2 is the level change that is measured at time $n_1 + 1$. This can be interpreted as the difference between (a) the value of Y that is predicted to occur at time $n_1 + 1$ on the basis of baseline data and (b) the value of Y that is predicted to occur at time $n_1 + 1$ on the basis of intervention phase data. If the process present during the baseline phase continues unchanged during the second phase (i.e., there is no interruption of the process), the two predictions will be the same; if the intervention shifts the level, the two predictions will be different. The third partial regression coefficient β_3 is the change in slope from the baseline phase to the intervention phase. The value of the slope in the intervention phase can be computed by adding the slope change coefficient to the value of the slope in the baseline phase; that is, $\beta_1 + \beta_3 =$ second (i.e., intervention phase) slope. After model I is estimated the same data are then analyzed using model II.

Model II

Model II is a reduced form of model I. It is written as follows:

$$Y_t = \beta_0 + \beta_1 D_t + \varepsilon_t,$$

where

Y_t is the dependent variable score at time t;
β_0 is the process intercept;
β_1 is the process slope;
D_t is the value of the level-change dummy variable D (zero for the first phase and one for the second phase) at time t; and
ε_t is the process error of the model at time t.

The errors in the process are assumed to be independent, normally distributed, have a mean of zero at each time point, and have constant variance from one time point to another.

Interpretation of the Coefficients of Model II

The two parameters of this model are easily interpreted. The intercept β_0 is equal to the mean of the first phase of the process. The slope β_1 is the level change, and $\beta_0 + \beta_1$ is equal to the mean of the second phase of the process. When Model II is valid the level change coefficient is equivalent to the difference between the two phase means; this is not generally true under Model I.

Deciding Between Models I and II

If there is no trend in either phase, there is no need for all of the parameters in model I. Parameters β_1 and β_3 are superfluous in this case because they measure the first phase slope and change in slope, both of which are assumed to be zero in model II. The application of model I in this case presents no problem in terms of bias in the level change estimate, but the power of the test on level change is negatively affected. That is, the power of the test on level change is higher using an analysis based on model II than it is using an analysis based on model I. In contrast, a model II analysis provides a biased estimate of level change if slope is present. Because the absence of bias usually should have higher priority than power when selecting an analysis, it is generally better to use model I when there is little justification for claiming that trend is not present. A formal method to help decide whether the simpler model is justified can be helpful.

Model Comparison Test

The following statistic can be used to test the joint null hypothesis that parameters β_1 and β_3 in model I are equal to zero (i.e., H_0: $\beta_1 = \beta_3 = 0$). If this hypothesis is rejected, model I is preferred over model II because the first phase slope, slope change, or both are not equal to zero. The two-parameter model is inadequate in this situation because it does not contain enough parameters to adequately describe the

data. The model comparison test statistic is defined as follows:

$$\frac{\left(\mathrm{SS}_{\mathrm{Reg}_{\mathrm{Model\,I}}} - \mathrm{SS}_{\mathrm{Reg}_{\mathrm{Model\,II}}}\right)/2}{\mathrm{MS}_{\mathrm{Res}_{\mathrm{Model\,I}}}} = F,$$

where

$\mathrm{SS}_{\mathrm{Reg}_{\mathrm{Model\,I}}}$ is the regression sum of squares based on model I;
$\mathrm{SS}_{\mathrm{Reg}_{\mathrm{Model\,II}}}$ is the regression sum of squares based on model II; and
$\mathrm{MS}_{\mathrm{Res}_{\mathrm{Model\,I}}}$ is the residual mean square based on model I.

The obtained value of F is compared with the critical value of F based on $df = 2, N - 4$. A liberal alpha level (say, $\alpha = .10$) should be used for this test.

If it is known before the experiment is conducted that the process under investigation has deterministic slope, there is no need for the model comparison test. In this case model I should be adopted even if the model comparison test does not lead to a rejection of the null hypothesis. But in most cases the investigator will not know the underlying process and the model comparison test will be relevant. Once a decision has been made regarding the choice between models I and II, conformity with assumptions regarding the chosen model should be investigated.

Stage B: Evaluate the Properties of the Errors of the Selected OLS Model

The adequacy of both the descriptive and inferential aspects of a regression analysis depends upon whether the assumptions of the model are approximately met. The assumptions regarding the errors of models I and II are essentially the same as in the case of conventional regression models (viz., the errors are independent and normally distributed, with mean zero and constant variance). Hence, the methods of evaluating conformity with the assumptions are essentially the same as with conventional regression models, but there is one exception. Because many observations are obtained from one unit that is measured continuously or repeatedly during a sequence of N time periods (rather than one observation from N independent subjects) the assumption of independent errors should be questioned. If, in general, the error measured at time t systematically provides information regarding the value of the error measured at one or more subsequent time points in the series, the errors are not independent. In this case the errors are said to be dependent or autocorrelated. Tests are available to determine whether the errors are autocorrelated; there are several reasons why it is useful to know.

The most frequently discussed motivation for identifying possible dependence of errors is that it has an effect on the standard errors associated with the coefficients of the regression model. If the errors are positively autocorrelated, the standard errors are underestimated and the t-values are too large; if the autocorrelation is negative the standard errors are overestimated and the t-values are too small. Consequently, p-values and confidence intervals associated with tests on level-change and slope-change coefficients are distorted by dependency of errors. Although these inferential problems are important, note that nothing has been stated regarding bias in the parameter estimates. Autocorrelation among the errors does *not* introduce bias to the

level-change and slope-change coefficients. But the presence of autocorrelated errors may mean that the model is wrong.

Errors are autocorrelated for a reason; often the reason is not known. If the model does not contain parameters that adequately describe the data, the model is said to be misspecified. Misspecified models can be expected to have autocorrelated errors. For example, if model II (a two-parameter model) is inappropriately applied to data that require four parameters (e.g., model I) for adequate description, the errors will be autocorrelated. This implies that a test for autocorrelated errors is a general test for model misspecification. Actually there are several reasons for autocorrelated errors. Among them: (1) the model is misspecified because it needs additional intervention predictors, (2) the functional form describing the data within phases is misspecified, or (3) the errors follow some time-series process (even though the deterministic portion of the model is reasonable). Examples of the effects of several different types of misspecification on autocorrelation can be found elsewhere (e.g., Huitema and McKean, 1998).

Some writers believe that the errors of single-case time-series designs *must* be autocorrelated (e.g., Jones et al., 1977); they interpret a lack of autocorrelation as evidence that behavior is not systematic. This is one of many misunderstandings regarding autocorrelation that have been discussed at length in the literature (see Huitema, 1986a, 1988; Huitema and McKean, 1998). Adequate empirical investigations of this issue conclude that autocorrelation is not a problem in the majority of applied single-case behavioral experiments *if the model is reasonably specified* (e.g., Huitema, 1985, 1986a; Huitema and McKean, 1998; Methot, 1995; Sideridis and Greenwood, 1997). This conclusion holds for data of the type frequently encountered in behavioral and medical research. Daily sampling of behavior that can change very rapidly is typical in research of this type. Outcomes that change slowly are much more likely to be autocorrelated, but they may yield little residual autocorrelation if the interval between observations is sufficiently long. This can occur, for example, in epidemiological studies where disease rates are reported on an annual basis. Published research in many content areas other than psychology and medicine suggests that independent errors are quite common (e.g., Van Den Noortgate and Onghena, 2003, p. 8), but they are far from universal.

Models I and II adequately describe the data from many two-phase experiments. Although fitting both of them and then applying a model comparison test to identify the more appropriate one is an important initial stage of the analysis, the better of the two models still may be inadequate. A test for autocorrelation among the errors of the model initially identified should be carried out. If the errors of the chosen model appear not to be autocorrelated, that model is adopted for the final analysis. But if the errors appear to be autocorrelated, then either model III or IV becomes relevant. Details regarding autocorrelation and two tests for autocorrelated errors are presented next.

Autocorrelation and Tests for Autocorrelated Errors

Autocorrelation among the errors of the model refers to the general notion that errors measured at time t are somewhat predictable from errors measured at an earlier

point in time (say, time $t - 1$). If this is true, the errors are not independent and the inferential statistics regarding intervention effects are distorted. A formal test for autocorrelated errors is by far the most frequently used method to facilitate deciding whether the errors of time-series regression models are independent. Because the errors are unknown they are estimated using the residuals of the fitted equation. Hence, tests for autocorrelation are carried out on the residuals and the inference is with respect to the errors of the model. (Technically, there is always minor autocorrelation among the residuals as a result of the OLS estimation procedure, but this is not a practical issue of concern.) Two tests that are appropriate for this purpose are the classic Durbin–Watson (D–W; Durbin and Watson, 1950, 1951) test and the simple Huitema–McKean (H–M; Huitema and McKean, 2000a) test. When one of these tests suggests that the errors are autocorrelated, the inferential results of the OLS analysis are questionable and models that accommodate autocorrelated errors (models III and IV) should be considered. Details regarding the D–W and H–M tests are presented next.

The D–W Test

The D–W test statistic d is provided as part of the output of almost all comprehensive OLS regression routines; it is described in most textbooks on regression analysis. The null hypothesis associated with this test states that the value of the lag-1 autocorrelation among the process errors is equal to zero; that is, $\rho_1 = 0$. A rejection of this null hypothesis using the traditional D–W test leads to the conclusion that the errors are positively autocorrelated.

The obtained test statistic d can be evaluated in two ways. The standard method is to compare it with two critical values (d_L and d_U) provided in tables of the critical values for the D–W test (see Appendix 18.1). Unlike conventional tests based on t and F distributions that have a single critical value (given α, df, and N), the traditional way of carrying out the D–W test involves a bounded approach. The test is performed by comparing the obtained value of d with both the lower and upper bounds (d_L and d_U) associated with the number of parameters in the model (P) and the total sample size N. The null hypothesis is rejected if the obtained d is *less* than d_L and it is retained if d is *larger* than d_U; if d falls between d_L and d_U, the test is declared inconclusive. (A test for negative autocorrelation can be performed by substituting $(4 - d)$ in place of d in the routine described above.)

The inconclusive region is present because the exact critical value depends on both the general form of the design matrix and the specific values it contains. Because these values differ from one data set to another, the exact critical value also differs from one data set to another. Hence, d_L and d_U are the bounds on the exact critical value determined by Durbin and Watson. Although the testing routine is confusing to many researchers, it is still recommended in many contemporary textbooks; it can be avoided in two ways. The best alternative is to use special purpose software that provides the D–W p-value for the specific data set. Such routines are not available in current versions of *Minitab* or SPSS; specialized software that has this capability is often expensive and/or not user friendly. A solution to these problems is contained in Appendix 18.2 in the form of a *Minitab* macro that conveniently provides a p-value

(based on the beta distribution) for the D–W test. A second way to avoid the problems with the traditional D–W bounds procedure is to use a completely different test.

The H–M Test

A second method of testing H_0: $\rho_1 = 0$ is the simple H–M test (Huitema and McKean, 2000a). This test is easy to compute; the only aspect of the test that requires software is the lag-1 autocorrelation coefficient (r_1) that is provided by both *Minitab* and SPSS. Because this test is not bounded, only one critical value is associated with a given level of α. Type I error and power for this test are essentially the same as for the D–W. The disadvantage is that it is less general than the D–W. It was developed specifically for design matrices of the form discussed here for interrupted time-series experiments; it does not apply to general regression problems where the design matrices deviate from the form used for the intervention models described here.

The test statistic $z_{\text{H–M}}$ is defined as follows:

$$\frac{r_1 + \dfrac{P}{N}}{\sqrt{\dfrac{(N-2)^2}{(N-1)N^2}}} = z_{\text{H–M}},$$

where

r_1 is the sample lag-1 autocorrelation coefficient, which is usually defined as

$$r_1 = \frac{\displaystyle\sum_{t=2}^{N} (e_t)\,(e_{t-1})}{\displaystyle\sum_{t=1}^{N} e_t^2},$$

where

e_t is the residual associated with time t;
P is the number of parameters in the model;
N is the total number of observations in the experiment; and
$z_{\text{H–M}}$ is the test statistic that is interpreted as a normal deviate.

The obtained value of $z_{\text{H–M}}$ is compared with the critical value of z associated with the chosen level of α. The critical values for a directional test are 1.282 ($\alpha = .10$), 1.645 ($\alpha = .05$), and 2.326 ($\alpha = .01$). Alternatively, p-values may be computed. This is easily accomplished because, under the null hypothesis, the test statistic is distributed approximately as a standard normal deviate. This means that the p-value associated with any obtained $z_{\text{H–M}}$ can be obtained by referring to a table of the standard normal distribution or by applying a computer subroutine for the normal distribution.

Other formal tests for autocorrelation exist. Most major software packages that provide the correlogram (a plot of autocorrelation coefficients computed for various

lags) include standard tests or confidence intervals that are used to evaluate autocorrelation among the errors. These approaches work well only when the number of observations is large; they lack power otherwise. For this reason the D–W and H–M tests recommended in this section are more appropriate than the tests on the autocorrelation function (provided in most software packages) when the number of observations is not large. This recommendation holds even though it is often pointed out that tests for lag-1 autocorrelation (e.g., D–W and H–M) ignore dependency that may be revealed at higher order lags. Although this is true, both old and new tests that evaluate autocorrelation for a large portion of the entire autocorrelation function generally have much lower power than do lag-1 tests applied to series with fewer than 100 observations (Huitema and McKean, 2007a, 2007b). Alas, when N is very small, no available test is satisfactory.

Small N *Alternatives to Traditional Autocorrelation Tests*

Because no traditional autocorrelation testing procedure is sufficiently powerful with short series (say, $N < 50$) three alternatives may be considered in this case. First, α can be set at a level that is more liberal than the conventional level of .05. I recommend using $\alpha = .10$ (directional) or .20 (nondirectional). The critical value for the H–M test that is consistent with this recommendation is 1.282. The second alternative procedure is to simply specify an absolute value for the sample lag-1 autocorrelation coefficient (say, .30) beyond which the independent errors model will be considered invalid.

The third alternative is to bypass all procedures for determining whether autocorrelated errors exist; instead, always use an intervention model that accommodates autoregressive errors. This is exactly what is done when strategy I is used. The advantage of this approach is that it simplifies the analysis because the autocorrelation testing step is completely eliminated. But two potential disadvantages are associated with always adopting an autoregressive model for the errors. First, it has been shown that fitting autoregressive models to correct for error rate problems introduced by dependency among errors can actually make matters worse in the case of small N (Huitema et al., 1994). Second, if an autoregressive model (e.g., model III or IV) is applied when no dependency exists, the power to detect intervention effects is generally lower than it is when using a model that assumes independence (e.g., model I or II). Hence, I recommend the use of models that accommodate autoregressive errors only when there is reasonable evidence to justify the additional complexity in the model. A formal test for autocorrelation using $\alpha = .10$ (directional) provides such evidence.

Recall that the purpose of evaluating the independence assumption is to determine whether the conventional models (models I and II) appear to be adequate for describing the data. If no autocorrelation is identified among the errors of the model that was selected in stage A, that model is adopted. But if autocorrelation is identified in the initially selected model, the final model should acknowledge it. This involves fitting model III or IV, depending on the number of parameters required. Models III and IV are described below.

Model III—Four Parameters and Autoregressive Errors

Model III should be considered if (a) model I is chosen over model II on the basis of the model comparison test, and (b) the errors of model I are determined to be autocorrelated. Model III is identical to model I except that the errors ε_t are not independent; rather, they follow an autoregressive process. This means that the error at time t is a linear combination of the errors present at preceding time points. In most experiments where the errors are not independent the process can be adequately described as first-order autoregressive (i.e., the autoregressive order $p = 1$). In this case model III can be written as:

$$Y_t = \beta_0 + \beta_1 T_t + \beta_2 D_t + \beta_3 SC_t + \phi_1 \varepsilon_{t-1} + u_t,$$

where

Y_t is the dependent variable score at time t;
β_0 is the process intercept;
β_1, β_2, and β_3 are the process partial regression coefficients;
T_t is the value of the time variable T at time t;
D_t is the value of the level-change dummy variable D (0 for the first phase and 1 for the second phase) at time t;
SC_t is the value of the slope change variable SC [defined as $[T - (n_1 + 1)]D$ where n_1 is the number of observations in the first phase];
$\varepsilon_t = Y_t - [\beta_0 + \beta_1 T_t + \beta_2 D_t + \beta_3 SC_t]$; it is assumed that the $\varepsilon_t = \phi_1 \varepsilon_{t-1} + u_t$, where ϕ_1 is the lag-1 autoregressive coefficient; and
u_t (the disturbance at time t) $Y_t - (\beta_0 + \beta_1 T_t + \beta_2 D_t + \beta_3 SC_t + \phi_1 \varepsilon_{t-1})$.

The disturbances are assumed independent, normally distributed with mean zero, and constant variance from one time period to another.

Model III includes a coefficient that measures autocorrelation among the errors; this coefficient is not found in model I because it assumes independent errors. The parameters of model III can be estimated using the double bootstrap procedure (McKnight et al., 2000) that is implemented in software illustrated in Chapter 19. I recommend this procedure when $N < 50$. If $N \geq 50$, use either the double bootstrap procedure or a routine that provides maximum-likelihood estimates of regression models with first-order autoregressive errors. Major software packages such as *SAS* and *SPSS* (*Trends* module) provide this capability.

A generalized version of model III that is appropriate for situations where the errors follow a more complex AR(p) autoregressive structure can be written as follows:

$$Y_t = \beta_0 + \beta_1 T_t + \beta_2 D_t + \beta_3 SC_t + \phi_1 \varepsilon_{t-1} + \phi_2 \varepsilon_{t-2} + \cdots + \phi_p \varepsilon_{t-p} + u_t.$$

The predictors in this model are the same as those shown earlier for the basic version of model III; the $\phi_1, \phi_2, \ldots, \phi_p$ are the lag-1 through lag-p autoregressive coefficients. These coefficients are necessary when errors measured at time t are predictable from errors measured at time periods $t - 1, t - 2, \ldots, t - p$. The order p is determined by the how far back in the error series there is predictability of ε_t.

Computation of these coefficients is not described here, but it can be found in many textbooks on time-series analysis (e.g., Enders, 2010).

Most single-case studies have little or no residual autocorrelation; when autocorrelation is present, it usually can be adequately modeled using a first-order (lag-1) autoregressive coefficient, so neither the generalized version of model III nor a more complex model (that includes moving average or combined autoregressive-moving average coefficients) is likely to be needed. For this reason, subsequent discussion of model III refers to the basic (*AR*-1) version.

Model IV—Two Parameters and Autoregressive Errors

Model IV should be considered if (a) model II was chosen over model I on the basis of the model comparison test, and (b) the errors of model II were determined to be autocorrelated. The only difference between models II and IV is the nature of the errors; it is assumed that the errors are independent under model II, but autocorrelated under model IV. The most basic version of model IV can be written as follows:

$$Y_t = \beta_0 + \beta_1 D_t + \phi_1 \varepsilon_{t-1} + u_t,$$

where

Y_t is the dependent variable score at time t;
β_0 is the process intercept (equal to the mean of the first phase in the process);
β_1 is the slope coefficient (equal to the level change in this model);
D_t is the value of the level change dummy variable D (0 for the first phase and 1 for the second phase) at time t,

$$\varepsilon_t = Y_t - [\beta_0 + \beta_1 D_t],$$

where

ϕ_1 is the lag-1 autoregressive coefficient; and
u_t is the disturbance $(Y_t - (\beta_0 + \beta_1 D_t + \phi_1 \varepsilon_{t-1}))$ at time t.
It is assumed that u_t is independent and normally distributed with mean zero and a common variance σ^2 at all time points.

This model can be generalized to handle errors having a more complex form of dependency that is described by a pth order autoregressive process, where p is the order of the lag. The generalized version of model IV can be written as follows:

$$Y_t = \beta_0 + \beta_1 D_t + \phi_1 \varepsilon_{t-1} + \phi_2 \varepsilon_{t-2} + \cdots + \phi_p \varepsilon_{t-p} + u_t.$$

It can be seen that the only difference between the basic version and the generalized version is in the autoregressive structure of the errors. Because the generalized version is not often required it will not be discussed further. Subsequent reference to models III and IV refer to the basic (*AR*-1) versions.

As with model III, the recommended estimation procedure for model IV is the double bootstrap routine illustrated in Chapter 19. This recommendation is for the

case of small N. In the case of $N \geq 50$, both the double bootstrap and maximum-likelihood estimation are appropriate.

In summary, models III and IV are simply extensions of models I and II, respectively, that accommodate dependency among the errors. The level-change and slope-change intervention effect parameters associated with these models are interpreted in essentially the same manner as they are with models I and II.

18.6 EFFECT SIZES

In most two-phase studies where the dependent variable is well understood the level-change and slope-change statistics provide excellent descriptions of the outcome of the intervention. Some journals, however, require authors to include a measure of standardized effect size or some measure of the degree of association between the independent and dependent variables as part of the analysis of most experimental research, whether the design is a conventional group comparison or a single-case intervention study. Although measures of this type are commonly reported in group-based experimental and quasi-experimental applications, they are not a part of the single-case research culture. If conventional measures of standardized effect size (such as d and g) and measures of association (such as R^2 and $\hat{\eta}^2$) are applied to single-case studies they are easily misinterpreted. This occurs because the nature of the unit of analysis is often ignored.

Population Versus Process

An understanding of appropriate measures of standardized effect size in single-case designs requires a clear distinction between two key concepts: population and process. Recall that a conventional between-subject randomized-group design involves the selection of N subjects from a population of potential interest. When the total sample has been randomly selected from a population of interest (and all subjects complete the experiment), the researcher is entitled to generalize the experimental results back to the population from which the subjects were drawn. The sample means are viewed as unbiased estimates of the corresponding population parameters, μ_1 and μ_2. Similarly, the sample standardized effect size (using Hedges' g) is computed using

$$\frac{\bar{Y}_1 - \bar{Y}_2}{\sqrt{\dfrac{\sum_{i=1}^{n_1} y_i^2 + \sum_{i=1}^{n_2} y_i^2}{n_1 + n_2 - 2}}} = \frac{\bar{Y}_1 - \bar{Y}_2}{s_w} = g.$$

This is an estimate of the population standardized effect size $\frac{\mu_1 - \mu_2}{\sigma_w} = \delta$.

Recall that the standard deviation in a group-based experiment is a measure of variation among subjects in an entire distribution of different subjects. Suppose

that the treatment-group mean exceeds the control-group mean and the standardized effect size is 0.7. This measure provides information regarding the performance of treated subjects relative to subjects who are in the control condition. For example, it can be determined (from the unit normal distribution) that approximately 76% of the treatment-group subjects have scores that exceed the control distribution mean whereas only 50% of the control scores exceed this mean. It can be seen in the formula for *g* that the within-group standard deviation is the basis for both the standardization and the interpretation.

Because a single-case design involves *N* observations that have been sampled from a *process* generated by a single unit treated and measured repeatedly over time whereas a traditional group design involves *N* different units that are sampled from a *population* of different units treated and measured once, the population parameters in the two-group design are not the same as the process parameters in the single-subject design. Although group experiments and single-case experiments are both carried out to evaluate whether treatment effects exist, these designs answer different questions. The change in process level in the single-subject design should not be interpreted as the differential effect of the treatment and control conditions in a population of different subjects. There is usually no justification to claim that the change in process level (for example) for one subject measured repeatedly yields an unbiased estimate of the population average causal effect $\mu_1 - \mu_2$.

Because there is a difference in the interpretation of the effects in these two types of design, there is also a difference in the interpretation of standardized effect sizes and measures of association applied to these designs. Once again, the distinction between the distribution of responses from various subjects and the distribution of responses from a single subject must be made if standardized effect sizes based on the two types of design are to be interpreted reasonably.

Effect Measures for the Two-Phase Single-Case Design

R^2

A measure of the degree of association between the level-change predictor (dummy) variable and the dependent variable is provided by the coefficient of determination. This measure describes the proportion of the variation in the observed single-case behavior that is explained by level change; it is computed as the ratio of the sum of squares regression for the level-change dummy variable over the total sum of squares. This value is appropriate for both models II and IV; it is automatically produced by typical multiple regression software when model II is estimated. If model I or III is appropriate, the ratio $(\mathrm{SS}_{\mathrm{Reg}\,T,D,SC} - \mathrm{SS}_{\mathrm{Reg}\,T,SC})/\mathrm{SS}_{\mathrm{total}}$ estimates the proportion of the total variation that is explained by level change independent of trend in either phase.

The SLC Statistic

A method of estimating the *s*tandardized *l*evel *c*hange (*SLC*) may be defined as:

$$\mathrm{SLC} = \frac{b_{\mathrm{LC}}}{\sqrt{\mathrm{MS}_{\mathrm{Res}}}},$$

where

b_{LC} is the estimate of level change for the final model (i.e., b_2 for models I and III, and b_1 for models II and IV); and

MS_{Res} is the mean square residual from fitting model I (when the final model is either model I or III) or the mean square residual from model II (when the final model is either model II or IV).

The denominator is the estimate of the standard deviation of the prediction errors; recall that in conventional regression models this is often called the estimated standard error of estimate. Under models II and IV the estimated standard error of estimate is the same as the pooled within-phase standard deviation. Hence, in this case, *SLC* is the number of standard deviation units that the level has changed. It is important to remember, however, that this standard deviation is based on the individual subject's behavior across time; it is not the standard deviation for a sample of different subjects.

Suppose a single-subject study is carried out to evaluate the effect of a method of reducing depression and that model II fits the data. The level of depression before intervention is 23; after the intervention is applied the level is 16. Hence, the level-change coefficient is -7. If $\sqrt{MS_{Res}} = 2$, then $SLC = -3.5$. In this case the level observed after the intervention must be interpreted as an extreme departure from the level observed before the intervention. But this finding tells us nothing about the severity of this individual's depression (either before or after intervention) relative to that found in a defined population of depressed patients. This information can be conveyed using a different measure.

The CPR Statistic

Suppose there is interest in evaluating a single subject's reduction in depression in a normative sense. I describe a measure that provides this information as the *c*hange in *p*ercentile *r*ank (*CPR)* associated with a relevant normative group. *CPR* is a hybrid measure that uses both single-case results and normative-group data (if they are available) to convert the level change for a single subject to the corresponding percentile change in a relevant norm group.

Return to the depression example described above where the reduction in the level of depression for a single patient was seven points. Suppose that normative data on the depression measure are available for a defined population of depressed patients. Further, suppose that the mean and standard deviation in the depressed population are 30 and 7, respectively. If the individual patient's level-change coefficient is -7, this represents a reduction of one standard deviation in the population. But this change is not interpreted in the same manner as is the standardized effect size *g* in the case of a two-group design.

Recall that a *g* statistic of -1 means that the treatment-group mean is one standard deviation unit below the control-group mean (assuming that the treated group has lower depresession than the control group). In this case 50% of the treated population has depression scores above the treated mean whereas 84% of the control population

has depression scores above the treated population mean. Correspondingly, it can be stated that one-half of the control population depression scores fall above the control mean whereas only 16% of the treated population has depression scores that high. Clearly the *g* statistic is based on *mean* differences; but this is not true of the *CPR* statistic. Rather, the latter refers to the change between the population percentile rank associated with the individual patient's baseline level and the population percentile rank associated with the same patient's level in the intervention phase.

The computation of the *CPR* statistic is straightforward. In the example, the treatment appears to have changed the subject's depression score from a value that is equivalent to the score of a patient falling one standard deviation below the depressed population mean to a score that is equivalent to the score of a patient falling two standard deviations below the mean. The proportion of the standard normal distribution falling below $z = -1$ (i.e., the standard score associated with the baseline level of 23) is .1587 and the proportion of the standard normal distribution falling below $z = -2$ (i.e., the standard score associated with the baseline level of 16) is .0228. Hence, approximately 16% of the depressed population is less depressed than this patient was at baseline whereas only 2% of the depressed population is less depressed than this subject was after intervention. The change in the percentile rank (*CPR*) (percentile rank at baseline minus percentile rank after intervention) is $(16\% - 2\%) = 14\%$. Alternatively, approximately 84% of the norm group had a higher depression score than this patient did at baseline; approximately 98% of the norm group had a higher depression score than this patient did after intervention. These values are illustrated in Figure 18.2.

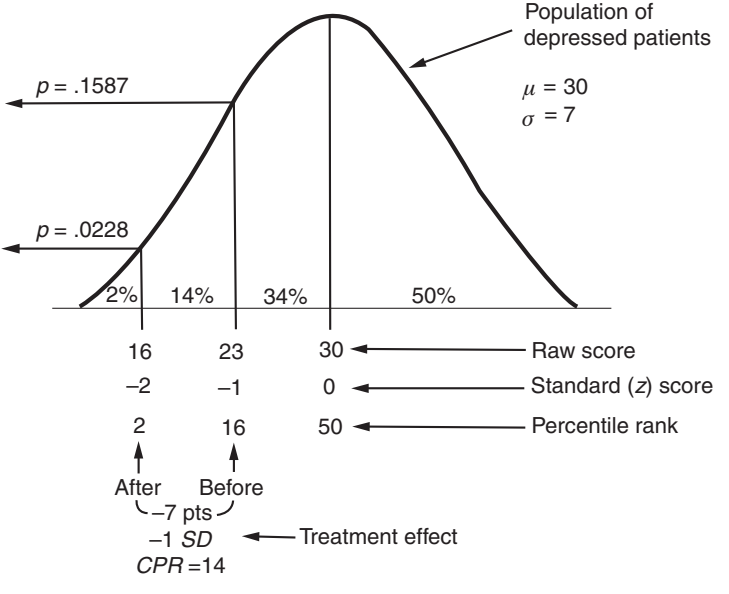


Figure 18.2 Areas of the standard normal distribution that are relevant to the computation of the CPR statistic.

Effect Measures for Two-Phase Intervention Designs Using a Compound Unit

When the analysis is carried out on a compound unit (instead of a single subject) the interpretation of the parameter estimates should acknowledge the nature of the unit. For example, suppose an academic department consists of 15 faculty and the total number of federal grant applications written by the faculty as a whole is available for many years before and after large financial rewards are provided for proposal submissions. The data for individual faculty members are not available; rather, the total number submitted by the whole department for each year is available. This is neither a traditional repeated-measures group design (where data on each individual subject are available) nor a conventional single-subject design, but the recommended regression methods are the same as in the case of the single-subject design. The generalization of results, however, is different. The estimates of level change, standardized level change, and measures of association generalize to a time-series process generated by a population of faculty members (like those studied) rather than to a single-subject process.

18.7 SAMPLE SIZE RECOMMENDATIONS

Recommendations regarding the minimum number of time-series observations required per phase for an adequate two-phase study are sought frequently. In some areas of research where data tend to be quite stable from one observation period to another I recommend $n_1 \geq 6$ and $N \geq 12$. In many other situations these minimum values are ridiculously inadequate. Hence, I feel compelled to be circumspect regarding this issue. Among the considerations that are relevant to such a recommendation are the following:

1. The type of process being sampled
2. The total number of observations in both phases combined
3. Whether the process contains cycles within the duration of the planned experiment
4. The sampling interval
5. The degree to which the errors are autocorrelated
6. The number of parameters to be estimated
7. The specific type of parameters to be estimated
8. Previous information regarding the process
9. The reason the intervention is introduced
10. The size and type of effect that is present

Unfortunately, the researcher will not have information regarding all of these considerations at the time the experiment is designed.

The issue most researchers have in mind when they ask about recommended sample size is the power of the analysis. An adequate study is not the same thing as a study with adequate power, although power is a relevant concern. Power can be estimated relatively easily and then translated into a sample size recommendation. But the dominant concern should be the adequacy of the comparison. One usually

presumes the adequacy of the comparison in the case of a randomized two-group experiment, regardless of the sample size. More skepticism is called for with two-phase time-series designs; one reason for this is that much is riding on the adequacy of projections made using the baseline data.

Short baseline phases can easily lead to self-deception regarding the adequacy of a study. Two-phase designs with short baseline phases are very vulnerable to the dual problems of regression toward the mean and overfitting. In the former, a short series of unusually high or low baseline values is often the very reason an intervention is initiated. In this case it should be anticipated that subsequent observations will be less extreme than are the initial observations (i.e., regression toward the mean will occur); a baseline of reasonable length will clarify whether the observations that initially seem unusual are simply a characteristic property of behavior that occurs now and then in the baseline process. If this is not done and the intervention is introduced immediately after the discrepant observations are observed, it will be difficult to say whether to change after intervention should be attributed to it. Consequently, it is essential to obtain adequate baseline data in order to eventually determine whether subsequent observations actually deviate from the baseline pattern.

Clearly, the justification for drawing causal conclusions from the two-phase design rests heavily on projections from the baseline phase. Suppose the underlying baseline process has no trend but it does have substantial variation. Further, suppose only a few baseline observations (say, three) are sampled from this process and the slope is computed on these meager data. The value of the sample slope is unlikely to be equal to zero even though the process slope is exactly zero. The sample regression line will essentially chase the error manifest in the sample. Projections from this slope will therefore be heavily based on random error. It would not be unusual to encounter an extremely steep slope in this situation—a slope that would provide a very misleading counterfactual value. In a sense, OLS works too well in these situations because the fitted line follows the sample idiosyncrasies; this is an example of overfitting. An example of data that present these problems may be helpful.

Data from one of three subjects included in a study reported by Scherrer and Wilder (2008) is illustrated in Figure 18.3. Note that there are only three observations in the baseline phase, but a reasonable number ($n_2 = 21$) in the intervention phase.

If the recommended strategy (i.e., strategy II) for the analysis of AB designs is carried out, model I is identified as the appropriate model because (1) the model comparison test strongly suggests that model II is inadequate ($F = 27.03$ and $p < .001$) and (2) the errors do not appear to be dependent ($r_1 = -.105$, $z_{\text{H–M}} = .32$, and $p = .75$). The following parameter estimates and associated inferential results are obtained using OLS (*Minitab*) to fit model I:

Four-Parameter (Model I) Estimates:

Predictor	Coef	SE Coef	T	P
Constant	83.000	6.395	12.98	0.000
Time	-21.000	2.960	-7.09	0.000
D	93.974	6.634	14.17	0.000
SC	21.288	2.964	7.18	0.000

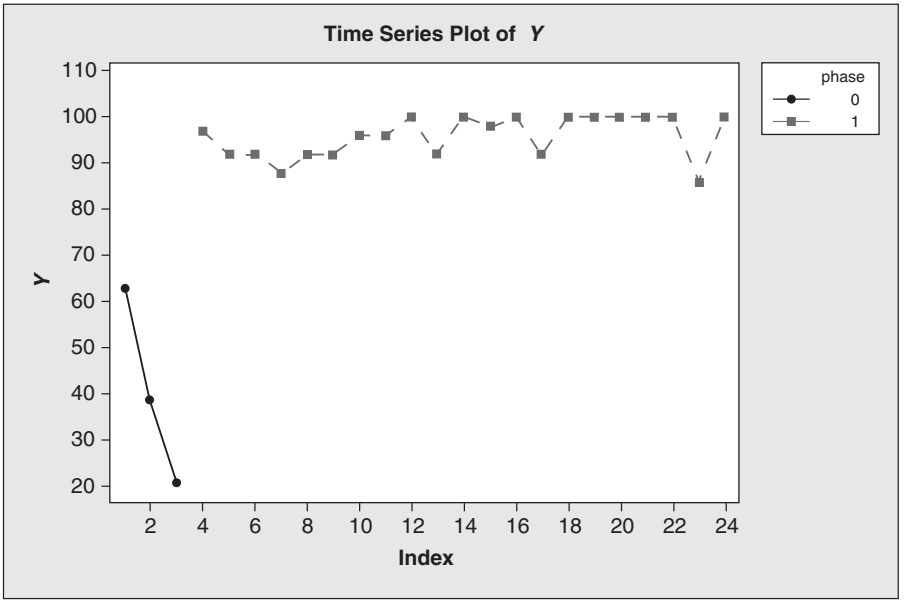


Figure 18.3 Data from a single subject where $n_1 = 3$ and $n_2 = 21$. (*Data Source*: Scherrer and Wilder, 2008.)

The level for the baseline phase is estimated using: $b_0 + b_1(n_1 + 1) = 83 - 21(4) = -1$. A negative value is not actually possible in the experiment because the dependent variable is the number of times that certain behaviors occur within an interval. Because a negative baseline level is substantively impossible, the level-change estimate (i.e., 93.974) is substantively meaningless. (Recall that the level-change estimate is defined as the difference between the level that is estimated from the baseline data at time $n_1 + 1$ and the level that is estimated from the intervention phase data at time $n_1 + 1$.) An inspection of the baseline data immediately indicates that it is not credible to project a continuation of the trend established by the trajectory of the three observations in this phase. In other words, the trend in this baseline realization can't possibly reflect the trend in the true baseline process. Instead, the regression line simply follows the idiosyncracies of this three data-point realization. Virtually any study with random variation is subject to this problem if the baseline sample is very short. This is one reason why short baselines should be avoided.

This is a situation in which the model should not include a slope parameter for the first phase because there are so few observations. Indeed, if model I (the four-parameter model that includes both level-change and slope-change parameters) is fitted it is likely to provide a misleading estimate of level change. Hence, neither model I nor the previously recommended model comparison test (that contrasts models I and II) are recommended in the case of inadequate baseline data. The best that can be done in this situation is to simply presume that there is no slope in the

unknown baseline process and to fit a three-parameter model that contains no slope parameter for the baseline phase. This model is

$$Y_t = \beta_0 + \beta_1 D + \beta_2 S_2 + \varepsilon_t,$$

where

β_0 is the baseline level;
β_1 is the level change parameter;
D is the 0–1 dummy variable indicating baseline and intervention phases;
β_2 is the slope for the intervention phase;
S_2 is the second phase time variable; and
ε_t is the error at time t.

The predictors used to estimate this three-parameter model are the same as those used to estimate model I, except for the eliminated first variable (time). That is, the predictors in this model include the level-change dummy variable and the slope-change variable.

The results of fitting this model using the data illustrated in Figure 18.3 are shown below. Because the predictor S_2 defined for the three-predictor model is constructed in the same way as is the predictor *SC* used in model I, the *SC* label is used on the output. But keep in mind that the coefficient associated with it is the phase 2 slope rather than slope change (alternatively, view it as the change from the *assumed* baseline slope of zero).

Level-change estimates based on this model are likely to be more satisfactory than those based on the four-parameter model (model I) even if moderate slope exists in the first phase process. If, after fitting this three-parameter model, the t-test on b_2 is not statistically significant, model II (i.e., the model with only level and level change coefficients) should be adopted.

Three-Parameter Model Estimates:

```
Predictor      Coef    SE Coef       T       P
Constant     41.000      4.423    9.27   0.000
D            51.974      5.476    9.49   0.000
SC           0.2883     0.2761    1.04   0.308
```

A comparison of the level-change estimate from this analysis with the level-change estimate in the previous analysis (model I) reveals a dramatic difference (i.e., 93.97 vs. 51.97). Also, note that it does not appear to be necessary to include the parameter that measures slope in the intervention phase because the p-value on the estimate is .31. Consequently, it is reasonable to eliminate that parameter from the model. If this is done the result is model II. Recall that the only parameters in this model are β_0 and β_1 and that they measure the baseline level and level change, respectively. To estimate these parameters we simply regress Y on the level-change indicator variable. The results of fitting this two-parameter model are as follows:

Two-Parameter (Model II) Estimates:

Predictor	Coef	SE Coef	T	P
Constant	41.000	4.432	9.25	0.000
D	54.857	4.738	11.58	0.000

It can be seen that the value of the level-change estimate is similar to the value obtained using the three-parameter model. (A heterogenous variance version of the model II analysis yields an identical effect estimate and a slightly different *p*-value). Because the two- and three-parameter models assume that the slope is zero in the first phase, they are not as vulnerable to over-fitting as is the four-parameter model (model I). Consequently the level-change estimates based on these models are more trustworthy than those based on model I when the baseline phase is very short.

Additional results from the example study (described in Chapter 21) support the recommendation to fit a model with no baseline slope parameter. Data were obtained from two additional subjects who were exposed to the same intervention, but they had relatively long baseline phases. The level-change estimates were essentially the same for these subjects under models I and II. The model II level-change estimates for all three subjects were similar: 54.86, 52.8, and 46.00. These results lead to the conclusion that fitting model I in the case of unstable and tiny baseline phases is the wrong approach. The negative consequences of fitting model I in this situation are far more serious than when fitting model II where moderate nonzero slope is characteristic of the underlying process.

In summary, if the process is known from previous evidence to be very stable across time, half-dozen observations in each phase may be sufficient for estimation purposes. I do not, however, recommend fitting the preliminary four-parameter model (model I) if the baseline is shorter than this. On the other hand, if there is no previous evidence regarding the behavior of the process in the absence of an intervention, the baseline phase should be long enough to convincingly establish the baseline variation and trajectory. In some cases this may require over 50 observations. Certainly there should be enough baseline observations to capture the nature of the process of interest. Fortunately, it is possible to both reduce the number of observations required in each phase and to increase internal validity at the same time. More sophisticated single-case designs are required to accomplish these dual goals; two designs and associated analyses that do so are pursued in Chapters 20 and 21.

18.8 WHEN THE MODEL IS TOO SIMPLE

Although the analyses described in this chapter are satisfactory for many two-phase studies, there are situations in which the nature of the data or the questions of interest require alternative methods. For example, in the area of pilot safety, Rantz (2007) studied the effects of a behavioral intervention on pilot checklist behavior. This preliminary work indicated that the nature of the intervention effect was not adequately captured using a step-function model such as model II. Rather, it was shown that there was a three-increment pattern that consisted of (1) a major immediate increase in behavior during the first day after the intervention was initiated, (2) a

smaller increase on the second day after intervention, and (3) an additional increase that characterized the data during the third and all subsequent intervention sessions. This preliminary evidence was then used to develop an intervention model for the analysis of a subsequent larger study (Huitema, 2008; Rantz, 2009).

The model that was developed includes parameters to measure all three increments. The dynamic nature of the intervention effect (i.e., the pattern of change during the intervention phase) is captured using these parameters. A similar approach was used in a study regarding driver behavior and pedestrian safety (Shurbutt et al., 2009). Alternative (nonregression) approaches for intervention analysis that may be useful in the case of a large number of observations are described in Box et al. (2008).

18.9 SUMMARY

The focus of this chapter is on two forms of the simple two-phase interrupted time-series design. The first form is a single-case design that involves obtaining measurements from a single subject during many time points before and after the introduction of an intervention. The second form involves observations obtained from a compound unit that consists of more than one organism. The recommended analysis for both forms may involve a variant of ANCOVA in which time is the covariate and the errors are either independent or autocorrelated.

Two strategies are proposed for modeling the data. The first strategy is to simply assume a four-parameter time-series regression model with autocorrelated errors, and to fit a model of this type. The standard results of such an analysis include measures of intercept, first phase slope, level change, slope change, and lag-1 dependency among residuals. The second strategy involves performing two stages in order to identify the best model. The first stage includes a test to determine whether a two-parameter model (that measures intercept and level change) is preferable to a four-parameter model. The second stage involves evaluations of the residuals of the model chosen in the first stage. If it is determined that the errors of the initially chosen model are not independent, a term for autocorrelated errors is included in the ultimate model.

Two types of standardized effect size are described. One measures level change relative to the within-phase variation associated with the subject or compound unit. The second measure frames the individual level change in terms of change in percentile rank in a relevant population.

Not all two-phase applications are adequately analyzed using the methods described in this chapter. More complex models should be considered when step-function or linear-trend functions do not reasonably approximate the nature of the intervention effects.

APPENDIX 18.1 CRITICAL VALUES FOR THE DURBIN–WATSON TEST FOR AUTOCORRELATED ERRORS

Durbin–Watson statistic: 5% significance points of d_L and d_U.

This table is a slightly modified version of the D–W tables presented in: Savin and White (1977). The Durbin–Watson test for serial correlation with extreme sample sizes or many regressors. *Econometrica*, *45*, 1989–1996.

	$m = 1$		$m = 2$		$m = 3$		$m = 4$		$m = 5$		$m = 6$		$m = 7$		$m = 8$		$m = 9$		$m = 10$	
n	d_L	d_U	d_L	d_U	d_L	d_U	d_L	d_U	d_L	d_U	d_L	d_U	d_L	d_U	d_L	d_U	d_L	d_U	d_L	d_U
6	0.610	1.400	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—
7	0.700	1.356	0.467	1.896	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—
8	0.763	1.332	0.559	1.777	0.367	2.287	—	—	—	—	—	—	—	—	—	—	—	—	—	—
9	0.824	1.320	0.629	1.699	0.455	2.128	0.296	2.588	—	—	—	—	—	—	—	—	—	—	—	—
10	0.879	1.320	0.697	1.641	0.525	2.016	0.376	2.414	0.243	2.822	—	—	—	—	—	—	—	—	—	—
11	0.927	1.324	0.758	1.604	0.595	1.928	0.444	2.283	0.315	2.645	0.203	3.004	—	—	—	—	—	—	—	—
12	0.971	1.331	0.812	1.579	0.658	1.864	0.512	2.177	0.380	2.506	0.268	2.832	0.171	3.149	—	—	—	—	—	—
13	1.010	1.340	0.861	1.562	0.715	1.816	0.574	2.094	0.444	2.390	0.328	2.692	0.230	2.985	0.147	3.266	—	—	—	—
14	1.045	1.350	0.905	1.551	0.767	1.779	0.632	2.030	0.505	2.296	0.389	2.572	0.286	2.848	0.200	3.111	0.127	3.360	—	—
15	1.077	1.361	0.946	1.543	0.814	1.750	0.685	1.977	0.562	2.220	0.447	2.471	0.343	2.727	0.251	2.979	0.175	3.216	0.111	3.438
16	1.106	1.371	0.982	1.539	0.857	1.728	0.734	1.935	0.615	2.157	0.502	2.388	0.398	2.624	0.304	2.860	0.222	3.090	0.155	3.304
17	1.133	1.381	1.015	1.536	0.897	1.710	0.779	1.900	0.664	2.104	0.554	2.318	0.451	2.537	0.356	2.757	0.272	2.975	0.198	3.184
18	1.158	1.391	1.046	1.535	0.933	1.696	0.820	1.872	0.710	2.060	0.603	2.258	0.502	2.461	0.407	2.668	0.321	2.873	0.244	3.073
19	1.180	1.401	1.074	1.536	0.967	1.685	0.859	1.848	0.752	2.023	0.649	2.206	0.549	2.396	0.456	2.589	0.369	2.783	0.290	2.974
20	1.201	1.411	1.100	1.537	0.998	1.676	0.894	1.828	0.792	1.991	0.691	2.162	0.595	2.339	0.502	2.521	0.416	2.704	0.336	2.885
21	1.221	1.420	1.125	1.538	1.026	1.669	0.927	1.812	0.829	1.964	0.731	2.124	0.637	2.290	0.546	2.461	0.461	2.633	0.380	2.806
22	1.239	1.429	1.147	1.541	1.053	1.664	0.958	1.797	0.863	1.940	0.769	2.090	0.677	2.246	0.588	2.407	0.504	2.571	0.424	2.735
23	1.257	1.437	1.168	1.543	1.078	1.660	0.986	1.785	0.895	1.920	0.804	2.061	0.715	2.208	0.628	2.360	0.545	2.514	0.465	2.670
24	1.273	1.446	1.188	1.546	1.101	1.656	1.013	1.775	0.925	1.902	0.837	2.035	0.750	2.174	0.666	2.318	0.584	2.464	0.506	2.613
25	1.288	1.454	1.206	1.550	1.123	1.654	1.038	1.767	0.953	1.886	0.868	2.013	0.784	2.144	0.702	2.280	0.621	2.419	0.544	2.560
26	1.302	1.461	1.224	1.553	1.143	1.652	1.062	1.759	0.979	1.873	0.897	1.992	0.816	2.117	0.735	2.246	0.657	2.379	0.581	2.513
27	1.316	1.469	1.240	1.556	1.162	1.651	1.084	1.753	1.004	1.861	0.925	1.974	0.845	2.093	0.767	2.216	0.691	2.342	0.616	2.470
28	1.328	1.476	1.255	1.560	1.181	1.650	1.104	1.747	1.028	1.850	0.951	1.959	0.874	2.071	0.798	2.188	0.723	2.309	0.649	2.431
29	1.341	1.483	1.270	1.563	1.198	1.650	1.124	1.743	1.050	1.841	0.975	1.944	0.900	2.052	0.826	2.164	0.753	2.278	0.681	2.396
30	1.352	1.489	1.284	1.567	1.214	1.650	1.143	1.739	1.071	1.833	0.998	1.931	0.926	2.034	0.854	2.141	0.782	2.251	0.712	2.363
31	1.363	1.496	1.297	1.570	1.229	1.650	1.160	1.735	1.090	1.825	1.020	1.920	0.950	2.018	0.879	2.120	0.810	2.226	0.741	2.333

(*Continued*)

	$m=1$		$m=2$		$m=3$		$m=4$		$m=5$		$m=6$		$m=7$		$m=8$		$m=9$		$m=10$	
n	d_L	d_U	d_L	d_U	d_L	d_U	d_L	d_U	d_L	d_U	d_L	d_U	d_L	d_U	d_L	d_U	d_L	d_U	d_L	d_U
32	1.373	1.502	1.309	1.574	1.244	1.650	1.177	1.732	1.109	1.819	1.041	1.909	0.972	2.004	0.904	2.102	0.836	2.203	0.769	2.306
33	1.383	1.508	1.321	1.577	1.258	1.651	1.193	1.730	1.127	1.813	1.061	1.900	0.994	1.991	0.927	2.085	0.861	2.181	0.796	2.281
34	1.393	1.514	1.333	1.580	1.271	1.652	1.208	1.728	1.144	1.808	1.079	1.891	1.015	1.978	0.950	2.069	0.885	2.162	0.821	2.257
35	1.402	1.519	1.343	1.584	1.283	1.653	1.222	1.726	1.160	1.803	1.097	1.884	1.034	1.967	0.971	2.054	0.908	2.144	0.845	2.236
36	1.411	1.525	1.354	1.587	1.295	1.654	1.236	1.724	1.175	1.799	1.114	1.876	1.053	1.957	0.991	2.041	0.930	2.127	0.868	2.216
37	1.419	1.530	1.364	1.590	1.307	1.655	1.249	1.723	1.190	1.795	1.131	1.870	1.071	1.948	1.011	2.029	0.951	2.112	0.891	2.197
38	1.427	1.535	1.373	1.594	1.318	1.656	1.261	1.722	1.204	1.792	1.146	1.864	1.088	1.939	1.029	2.017	0.970	2.098	0.912	2.180
39	1.435	1.540	1.382	1.597	1.328	1.658	1.273	1.722	1.218	1.789	1.161	1.859	1.104	1.932	1.047	2.007	0.990	2.085	0.932	2.164
40	1.442	1.544	1.391	1.600	1.338	1.659	1.285	1.721	1.230	1.786	1.175	1.854	1.120	1.924	1.064	1.997	1.008	2.072	0.952	2.149
45	1.475	1.566	1.430	1.615	1.383	1.666	1.336	1.720	1.287	1.776	1.238	1.835	1.189	1.895	1.139	1.958	1.089	2.022	1.038	2.088
50	1.503	1.585	1.462	1.628	1.421	1.674	1.378	1.721	1.335	1.771	1.291	1.822	1.246	1.875	1.201	1.930	1.156	1.986	1.110	2.044
55	1.528	1.601	1.490	1.641	1.452	1.681	1.414	1.724	1.374	1.768	1.334	1.814	1.294	1.861	1.253	1.909	1.212	1.959	1.170	2.010
60	1.549	1.616	1.514	1.652	1.480	1.689	1.444	1.727	1.408	1.767	1.372	1.808	1.335	1.850	1.298	1.894	1.260	1.939	1.222	1.984
65	1.549	1.616	1.514	1.652	1.480	1.689	1.444	1.727	1.408	1.767	1.372	1.808	1.335	1.850	1.298	1.894	1.260	1.939	1.222	1.984
70	1.583	1.641	1.554	1.672	1.525	1.703	1.494	1.735	1.464	1.768	1.433	1.802	1.401	1.838	1.369	1.874	1.337	1.910	1.305	1.948
75	1.598	1.652	1.571	1.680	1.543	1.709	1.515	1.739	1.487	1.770	1.458	1.801	1.428	1.834	1.399	1.867	1.369	1.901	1.339	1.935
80	1.611	1.662	1.586	1.688	1.560	1.715	1.534	1.743	1.507	1.772	1.480	1.801	1.453	1.831	1.425	1.861	1.397	1.893	1.369	1.925
85	1.624	1.671	1.600	1.696	1.575	1.721	1.550	1.747	1.525	1.774	1.500	1.801	1.474	1.829	1.448	1.857	1.422	1.886	1.396	1.916
90	1.635	1.679	1.612	1.703	1.589	1.726	1.566	1.751	1.542	1.776	1.518	1.801	1.494	1.827	1.469	1.854	1.445	1.881	1.420	1.909
95	1.645	1.687	1.623	1.709	1.602	1.732	1.579	1.755	1.557	1.778	1.535	1.802	1.512	1.827	1.489	1.852	1.465	1.877	1.442	1.903
100	1.654	1.694	1.634	1.715	1.613	1.736	1.592	1.758	1.571	1.780	1.550	1.803	1.528	1.826	1.506	1.850	1.484	1.874	1.462	1.898
150	1.720	1.746	1.706	1.760	1.693	1.774	1.679	1.788	1.665	1.802	1.651	1.817	1.637	1.832	1.622	1.847	1.608	1.862	1.594	1.877
200	1.758	1.778	1.748	1.789	1.738	1.799	1.728	1.810	1.718	1.820	1.707	1.831	1.697	1.841	1.686	1.852	1.675	1.863	1.665	1.874

	$m = 11$		$m = 12$		$m = 13$		$m = 14$		$m = 15$		$m = 16$		$m = 17$		$m = 18$		$m = 19$		$m = 20$	
n	d_L	d_U	d_L	d_U	d_L	d_U	d_L	d_U	d_L	d_U	d_L	d_U	d_L	d_U	d_L	d_U	d_L	d_U	d_L	d_U
16	0.098	3.503	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—
17	0.138	3.378	0.087	3.557	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—
18	0.177	3.265	0.123	3.441	0.078	3.603	—	—	—	—	—	—	—	—	—	—	—	—	—	—
19	0.220	3.159	0.160	3.335	0.111	3.496	0.070	3.642	—	—	—	—	—	—	—	—	—	—	—	—
20	0.263	3.063	0.200	3.234	0.145	3.395	0.100	3.542	0.063	3.676	—	—	—	—	—	—	—	—	—	—
21	0.307	2.976	0.240	3.141	0.182	3.300	0.132	3.448	0.091	3.583	0.058	3.705	—	—	—	—	—	—	—	—
22	0.349	2.897	0.281	3.057	0.220	3.211	0.166	3.358	0.120	3.495	0.083	3.619	0.052	3.731	—	—	—	—	—	—
23	0.391	2.826	0.322	2.979	0.259	3.128	0.202	3.272	0.153	3.409	0.110	3.535	0.076	3.650	0.048	3.753	—	—	—	—
24	0.431	2.761	0.362	2.908	0.297	3.053	0.239	3.193	0.186	3.327	0.141	3.454	0.101	3.572	0.070	3.678	0.044	3.773	—	—
25	0.470	2.702	0.400	2.844	0.335	2.983	0.275	3.119	0.221	3.251	0.172	3.376	0.130	3.494	0.094	3.604	0.065	3.702	0.041	3.790
26	0.508	2.649	0.438	2.784	0.373	2.919	0.312	3.051	0.256	3.179	0.205	3.303	0.160	3.420	0.120	3.531	0.087	3.632	0.060	3.724
27	0.544	2.600	0.475	2.730	0.409	2.859	0.348	2.987	0.291	3.112	0.238	3.233	0.191	3.349	0.149	3.460	0.112	3.563	0.081	3.658
28	0.578	2.555	0.510	2.680	0.445	2.805	0.383	2.928	0.325	3.050	0.271	3.168	0.222	3.283	0.178	3.392	0.138	3.495	0.104	3.592
29	0.612	2.515	0.544	2.634	0.479	2.755	0.418	2.874	0.359	2.992	0.305	3.107	0.254	3.219	0.208	3.327	0.166	3.431	0.129	3.528
30	0.643	2.477	0.577	2.592	0.512	2.708	0.451	2.823	0.392	2.937	0.337	3.050	0.286	3.160	0.238	3.266	0.195	3.368	0.156	3.465
31	0.674	2.443	0.608	2.553	0.545	2.665	0.484	2.776	0.425	2.887	0.370	2.996	0.317	3.103	0.269	3.208	0.224	3.309	0.183	3.406
32	0.703	2.411	0.638	2.517	0.576	2.625	0.515	2.733	0.457	2.840	0.401	2.946	0.349	3.050	0.299	3.153	0.253	3.252	0.211	3.348
33	0.731	2.382	0.668	2.484	0.606	2.588	0.546	2.692	0.488	2.796	0.432	2.899	0.379	3.000	0.329	3.100	0.283	3.198	0.239	3.293
34	0.758	2.355	0.695	2.454	0.634	2.554	0.575	2.654	0.518	2.754	0.462	2.854	0.409	2.954	0.359	3.051	0.312	3.147	0.267	3.240
35	0.783	2.330	0.722	2.425	0.662	2.521	0.604	2.619	0.547	2.716	0.492	2.813	0.439	2.910	0.388	3.005	0.340	3.099	0.295	3.190
36	0.808	2.306	0.748	2.398	0.689	2.492	0.631	2.586	0.575	2.680	0.520	2.774	0.467	2.868	0.417	2.961	0.369	3.053	0.323	3.142
37	0.831	2.285	0.772	2.374	0.714	2.464	0.657	2.555	0.602	2.646	0.548	2.738	0.495	2.829	0.445	2.920	0.397	3.009	0.351	3.097

(*Continued*)

	$m = 11$		$m = 12$		$m = 13$		$m = 14$		$m = 15$		$m = 16$		$m = 17$		$m = 18$		$m = 19$		$m = 20$	
n	d_L	d_U	d_L	d_U	d_L	d_U	d_L	d_U	d_L	d_U	d_L	d_U	d_L	d_U	d_L	d_U	d_L	d_U	d_L	d_U
38	0.854	2.265	0.796	2.351	0.739	2.438	0.683	2.526	0.628	2.614	0.575	2.703	0.522	2.792	0.472	2.880	0.424	2.968	0.378	3.054
39	0.875	2.246	0.819	2.329	0.763	2.413	0.707	2.499	0.653	2.585	0.600	2.671	0.549	2.757	0.499	2.843	0.451	2.929	0.404	3.013
40	0.896	2.228	0.840	2.309	0.785	2.391	0.731	2.473	0.678	2.557	0.626	2.641	0.575	2.724	0.525	2.808	0.477	2.829	0.430	2.974
45	0.988	2.156	0.938	2.225	0.887	2.296	0.838	2.367	0.788	2.439	0.740	2.512	0.692	2.586	0.644	2.659	0.598	2.733	0.553	2.807
50	1.064	2.103	1.019	2.163	0.973	2.225	0.927	2.287	0.882	2.350	0.836	2.414	0.792	2.479	0.747	2.544	0.703	2.610	0.660	2.675
55	1.129	2.062	1.087	2.116	1.045	2.170	1.003	2.225	0.961	2.281	0.919	2.338	0.877	2.396	0.836	2.454	0.795	2.512	0.754	2.571
60	1.184	2.031	1.145	2.079	1.106	2.127	1.068	2.177	1.029	2.227	0.990	2.278	0.951	2.330	0.913	2.382	0.874	2.434	0.836	2.487
65	1.231	2.006	1.195	2.049	1.160	2.093	1.124	2.138	1.088	2.183	1.052	2.229	1.016	2.276	0.980	2.323	0.944	2.371	0.908	2.419
70	1.272	1.987	1.239	2.026	1.206	2.066	1.172	2.106	1.139	2.148	1.105	2.189	1.072	2.232	1.038	2.275	1.005	2.318	0.971	2.362
75	1.308	1.970	1.277	2.006	1.247	2.043	1.215	2.080	1.184	2.118	1.153	2.156	1.121	2.195	1.090	2.235	1.058	2.275	1.027	2.315
80	1.340	1.957	1.311	1.991	1.283	2.024	1.253	2.059	1.224	2.093	1.195	2.129	1.165	2.165	1.136	2.201	1.106	2.238	1.076	2.275
85	1.369	1.946	1.342	1.977	1.315	2.009	1.287	2.040	1.260	2.073	1.232	2.105	1.205	2.139	1.177	2.172	1.149	2.206	1.121	2.241
90	1.395	1.937	1.369	1.966	1.344	1.995	1.318	2.025	1.292	2.055	1.266	2.085	1.240	2.116	1.213	2.148	1.187	2.179	1.160	2.211
95	1.418	1.930	1.394	1.956	1.370	1.984	1.345	2.012	1.321	2.040	1.296	2.068	1.271	2.097	1.247	2.126	1.222	2.156	1.197	2.186
100	1.439	1.923	1.416	1.948	1.393	1.974	1.371	2.000	1.347	2.026	1.324	2.053	1.301	2.080	1.277	2.108	1.253	2.135	1.229	2.164
150	1.579	1.892	1.564	1.908	1.550	1.924	1.535	1.940	1.519	1.956	1.504	1.972	1.489	1.989	1.474	2.006	1.458	2.023	1.443	2.040
200	1.654	1.885	1.643	1.896	1.632	1.908	1.621	1.919	1.610	1.931	1.599	1.943	1.588	1.955	1.576	1.967	1.565	1.979	1.554	1.991

Note: m is the number of predictors in the model not counting the intercept. For example, the four-parameter model discussed in Chapters 18 and 19 has an intercept and three predictors so the appropriate column is headed $m = 3$.

APPENDIX 18.2 *MINITAB* MACROS* FOR THE COMPUTATION OF (1) THE D–W TEST STATISTIC AND (2) THE *p*-VALUE FOR THE D–W TEST STATISTIC

(1) D-W Test Statistic

```
MACRO
DW RESIDS STAT
MCONSTANT N NM1 STAT
MCOLUMN RESIDS TEMP1 TEMP2
NOECHO
LET N = COUNT(RESIDS)
COPY RESIDS TEMP1;
USE 2:N.
LET NM1 = N − 1
COPY RESIDS TEMP2;
USE 1:NM1.
LET STAT = SUM((TEMP1 - TEMP2)**2)/SUM(RESIDS**2)
PRINT STAT
ECHO
ENDMACRO
```

(2) p-Value for the D-W Test Statistic

```
MACRO
PBETA X N P DIN ALT
MCONSTANT N P CAPP CAPQ ED VD V1 V2 D PVALUE ALT K I J DIN
MCONSTANT PVALUE1 PVALUE2 DTEMP
MCOLUMN P2 Q2 Q3 C.1-C.N CTEMP
MMATRIX A X TX MX R PM INVX O
# 'ALT VALUE: -1 rho<0; 0 rho<>0; 1 rho>0'
NOECHO
LET D = DIN
DEFINE 0 N N A
LET K=-1/N
DEFINE K N N O
COPY O C.1-C.N
DO I = 1 : N
 LET C.I(I)=1+C.I(I)
ENDDO
COPY C.1-C.N O
COPY A C.1-C.N
LET K=N-1
```

*These *Minitab* macros were written by Professor Joseph W. McKean, Department of Statistics, Western Michigan University. They can be downloaded from his Website at: www.stat.wmich.edu/mckean/mckean.html

```
LET C.1(1)=1
LET C.2(1)=-1
LET C.K(N)=-1
LET C.N(N)=1
DO I = 2 : K
 LET C.I(I)=2.0
 LET J=I+1
 LET C.J(I)=-1.0
 LET J=I-1
 LET C.J(I)=-1.0
ENDDO
COPY C.1-C.N A
MULTIPLY O X X
TRANSPOSE X TX
MULTIPLY TX A R
MULTIPLY R X R
MULTIPLY TX X MX
INVERT MX INVX
MULTIPLY R INVX PM
DIAGONAL PM P2
MULTIPLY PM PM PM
DIAGNOAL PM Q3
MULTIPLY A A A
MULTIPLY TX A R
MULTIPLY R X R
MULTIPLY R INVX PM
DIAGONAL PM Q2
LET CAPP=2*(N-1)-SUM(P2)
LET CAPQ=2*(3*N-4)-2*SUM(Q2)+SUM(Q3)
LET ED=CAPP/(N-P-1)
LET VD=2*(CAPQ-CAPP*ED)/((N-P-1)*(N-P+1))
LET V1=(ED*(4.0-ED)/VD-1.0)*ED/4.0
LET V2=(4.0/ED-1.0)*V1
IF ALT=1
 LET D=D/4
 CDF D PVALUE;
 BETA V1 V2.
ELSE IF ALT=-1
 LET D=D/4
 CDF D PVALUE;
 BETA V1 V2.
 LET PVALUE = 1 - PVALUE
ELSE
 LET D=D/4
 CDF D PVALUE1;
 BETA V1 V2.
 LET DTEMP = 1 - D
 IF PVALUE1 < .5
  CDF DTEMP PVALUE2;
```

```
  BETA V2 V1.
  LET PVALUE2 = 1 - PVALUE2
 ELSE
  LET PVALUE1 = 1 - PVALUE1
  CDF DTEMP PVALUE2;
  BETA V2 V1.
 ENDIF
 LET PVALUE = PVALUE1 + PVALUE2
ENDIF
PRINT PVALUE
PRINT V1 V2
ECHO
ENDMACRO
```

(3)

```
MACRO
QBETA X N P
MCONSTANT N P CAPP CAPQ ED VD V1 V2 K I J
MCOLUMN P2 Q2 Q3 C.1-C.N PVALUE DLOW DUP
MMATRIX A X TX MX R PM INVX O
NOECHO
DEFINE 0 N N A
LET K=-1/N
DEFINE K N N O
COPY O C.1-C.N
DO I = 1 : N
 LET C.I(I)=1+C.I(I)
ENDDO
COPY C.1-C.N O
COPY A C.1-C.N
LET K=N-1
LET C.1(1)=1
LET C.2(1)=-1
LET C.K(N)=-1
LET C.N(N)=1
DO I = 2 : K
 LET C.I(I)=2.0
 LET J=I+1
 LET C.J(I)=-1.0
 LET J=I-1
 LET C.J(I)=-1.0
ENDDO
COPY C.1-C.N A
MULTIPLY O X X
TRANSPOSE X TX
MULTIPLY TX A R
MULTIPLY R X R
```

```
MULTIPLY TX X MX
INVERT MX INVX
MULTIPLY R INVX PM
DIAGONAL PM P2
MULTIPLY PM PM PM
DIAGNOAL PM Q3
MULTIPLY A A A
MULTIPLY TX A R
MULTIPLY R X R
MULTIPLY R INVX PM
DIAGONAL PM Q2
LET CAPP=2*(N-1)-SUM(P2)
LET CAPQ=2*(3*N-4)-2*SUM(Q2)+SUM(Q3)
LET ED=CAPP/(N-P-1)
LET VD=2*(CAPQ-CAPP*ED)/((N-P-1)*(N-P+1))
LET V1=(ED*(4.0-ED)/VD-1.0)*ED/4.0
LET V2=(4.0/ED-1.0)*V1
SET PVALUE
.99 0.975 .95 .90
END
INVCDF PVALUE DUP;
BETA V1 V2.
LET DUP=4*DUP
SET PVALUE
.01 0.025 .05 .10
END
INVCDF PVALUE DLOW;
BETA V1 V2.
LET DLOW=4*DLOW
PRINT PVALUE DLOW DUP
ECHO
ENDMACRO
```

CHAPTER 19

Examples of Single-Case AB Analysis

19.1 INTRODUCTION

Four example data sets from two-phase studies are analyzed in this chapter. The studies illustrate key aspects regarding analytic choice and interpretation. One example was chosen to illustrate why the regression methods recommended in Chapter 18 may be preferred to ARIMA modeling. Computational aspects are described in considerable detail for each example.

19.2 EXAMPLE I: CANCER DEATH RATES IN THE UNITED KINGDOM

A study reported by Burk (1980) was designed to evaluate the effects of the introduction of fluoridation to drinking water on cancer death rates (CDR) in the United Kingdom. The data are shown in Figure 19.1. Each data point indicates the annual death rate in Birmingham for a specific year. It can be seen that annual CDR data were collected for 19 years; six of these years occurred before the introduction of fluoridation and 12 years occurred during a period when fluoridation was present. This example of a two-phase design was chosen for three reasons. First, the number of observations is small enough that the reader can easily repeat the analysis to confirm the results. Second, a comparison of the results based on the recommended analysis with those based on a frequently encountered alternative should clarify the importance of using the recommended design matrix. Third, the example illustrates why designs with only two phases are often considered untrustworthy.

The Analysis of Covariance and Alternatives: Statistical Methods for Experiments, Quasi-Experiments, and Single-Case Studies, Second Edition. Bradley E. Huitema.

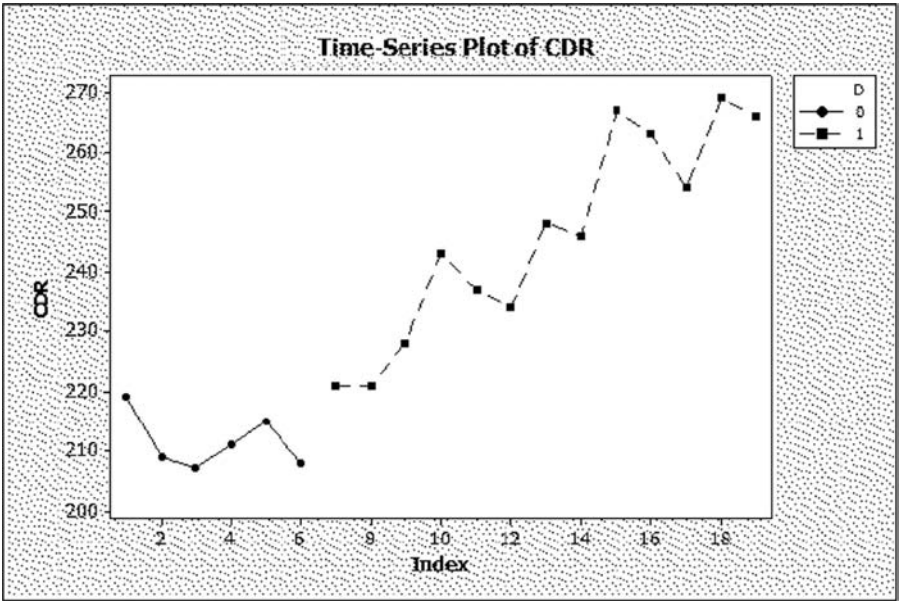


Figure 19.1 Annual cancer death rate in Birmingham, UK, from 1959 through 1977.

Strategy II Analysis

Stage A

The *Minitab* input commands (using the command line editor) for creating *T* (time), *D* (dummy), and *SC* (slope change) predictor variable columns are as follows:

```
MTB > set c1
DATA> 1:19
DATA> set c2
DATA> 6(0) 13(1)
DATA> set c3
DATA> 7(0) 1:12
DATA> end
```

Once these columns (1–3) are created, the dependent variable scores (CDR) are entered in column 4. Models I and II are then estimated using conventional regression commands. Model I is estimated by regressing CDR on all three predictors, and Model II is estimated by regressing CDR on only the dummy variable. The predictor variables and CDR data are listed in Table 19.1. The graphic representation of the CDR data is presented in Figure 19.1.

Note that the time index (1–19) rather than the actual year of measurement is plotted on the abscissa. Year 1959 corresponds to the first time point and year 1977

Table 19.1 Input for Model I Analysis of UK Cancer Death Rates (CDR): Predictors Are Time (*T*), Level Change Dummy (*D*), and Slope Change (*SC*)

T	*D*	*SC*	*CDR*
1	0	0	219
2	0	0	209
3	0	0	207
4	0	0	211
5	0	0	215
6	0	0	208
7	1	0	221
8	1	1	221
9	1	2	228
10	1	3	243
11	1	4	237
12	1	5	234
13	1	6	248
14	1	7	246
15	1	8	267
16	1	9	263
17	1	10	254
18	1	11	269
19	1	12	266

corresponds to time point 19. It is important that the actual years *not* be used in the analysis as the time variable (see Huitema and McKean, 2000b).

Model I Output

```
Regression Analysis: CDR versus Time, D, SC

The regression equation is CDR = 215 - 0.94 Time + 13.1 D +

5.04 SC


Predictor     Coef  SE Coef      T      P
Constant   214.800    5.734  37.46  0.000
Time        -0.943    1.472  -0.64  0.532
D           13.130    6.581   2.00  0.065
SC           5.042    1.542   3.27  0.005

S = 6.15947   R-Sq = 93.3%   R-Sq(adj) = 92.0%
```

```
Analysis of Variance

Source          DF      SS      MS      F      P
Regression       3  7937.9  2646.0  69.74  0.000
Residual Error  15   569.1    37.9
Total           18  8506.9

Durbin-Watson statistic = 2.27220
```

Model II Output

```
Regression Analysis: CDR versus D


The regression equation is

CDR = 212 + 34.4 D


Predictor     Coef  SE Coef      T      P
Constant   211.500    5.976  35.39  0.000
D           34.423    7.224   4.76  0.000

S = 14.6376   R-Sq = 57.2%   R-Sq(adj) = 54.7%


Analysis of Variance

Source          DF      SS      MS      F      P
Regression       1  4864.5  4864.5  22.70  0.000
Residual Error  17  3642.4   214.3
Total           18  8506.9

Durbin-Watson statistic = 0.531226
```

The predictor variables (Time, D, and SC), the dependent variable (CDR), and the residual columns RESI1 and RESI2 (the residuals from fitting Models I and II, respectively), are printed as follows:

```
MTB > print c1-c4 c6 c7
Data Display

Row  Time  D  SC  CDR    RESI1     RESI2
1     1    0   0  219   5.1429    7.5000
2     2    0   0  209  -3.9143   -2.5000
3     3    0   0  207  -4.9714   -4.5000
4     4    0   0  211  -0.0286   -0.5000
```

5	5	0	0	215	4.9143	3.5000
6	6	0	0	208	-1.1429	-3.5000
7	7	1	0	221	-0.3297	-24.9231
8	8	1	1	221	-4.4286	-24.9231
9	9	1	2	228	-1.5275	-17.9231
10	10	1	3	243	9.3736	-2.9231
11	11	1	4	237	-0.7253	-8.9231
12	12	1	5	234	-7.8242	-11.9231
13	13	1	6	248	2.0769	2.0769
14	14	1	7	246	-4.0220	0.0769
15	15	1	8	267	12.8791	21.0769
16	16	1	9	263	4.7802	17.0769
17	17	1	10	254	-8.3187	8.0769
18	18	1	11	269	2.5824	23.0769
19	19	1	12	266	-4.5165	20.0769

Comparison of Models I and II

Model comparison test statistic:

$$\frac{\left(\mathrm{SS}_{\mathrm{Reg}_{\text{Model I}}} - \mathrm{SS}_{\mathrm{Reg}_{\text{Model II}}}\right)/2}{\mathrm{MS}_{\mathrm{Res}_{\text{ModelI}}}} = F \text{ and}$$

$$\frac{(7937.86 - 4864.52)/2}{37.94} = 40.50 = F_{\text{obt}}.$$

The degrees of freedom for the model comparison test $= 2$ and $N - 4$; the critical value (using $\alpha = .10$) is $F_{2,\,15} = 2.695$. The evidence is very strong for the conclusion that model I is preferable to model II. Clearly, model II does not adequately describe the data, but as can be seen in Figure 19.2, the regression lines associated with model I capture the essential time progression of CDR in both phases of the series.

An inspection of the residuals of model I (see Figure 19.3) reveals no outliers, but there is a hint that minor heterogeneity of the pre- and postintervention variances may be present.

Formal tests of homogeneity of the pre- and postintervention error variances (see F-test and Levene's test results in Figure 19.4) both have p-values above .40; no further attention to this issue is necessary.

Normality of the errors is evaluated using the probability plot shown in Figure 19.5. It can be seen that all residuals are within the bounds indicated by the curved lines; hence, it does not appear that there are extreme departures from normality.

The next property to be scrutinized is the independence of the errors of model I. The Durbin–Watson (D–W) test statistic (requested as one of the options when models I and II were estimated) is part of the output shown above for the two models. Note that the obtained D–W values are 2.27 and 0.53 for models I and II, respectively. A comparison of these values with the critical values indicates that there is essentially no evidence for autocorrelation among the errors of model I whereas there is exceedingly strong evidence of autocorrelation among the errors of model II.

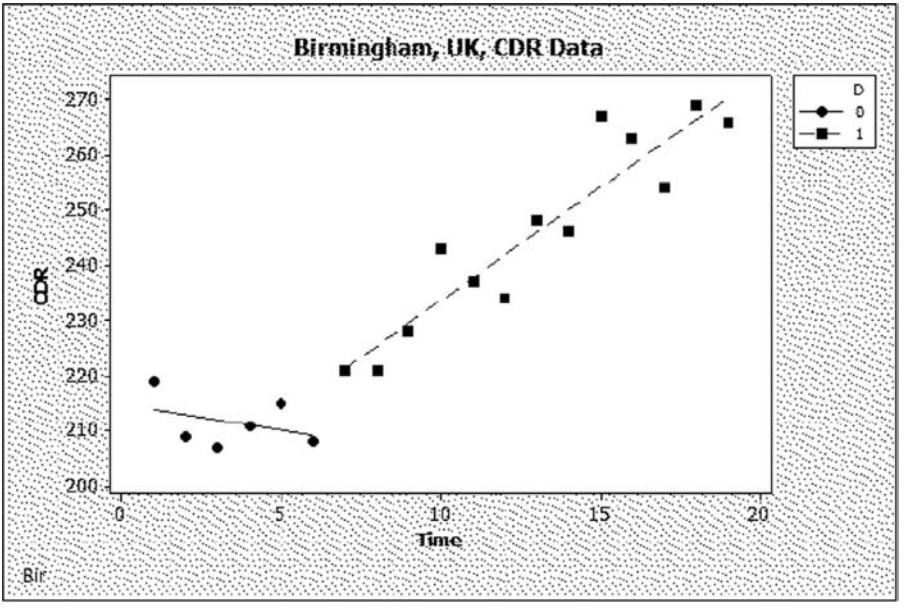


Figure 19.2 Within phase regression lines associated with fitting model I to Birmingham, UK, CDR data.

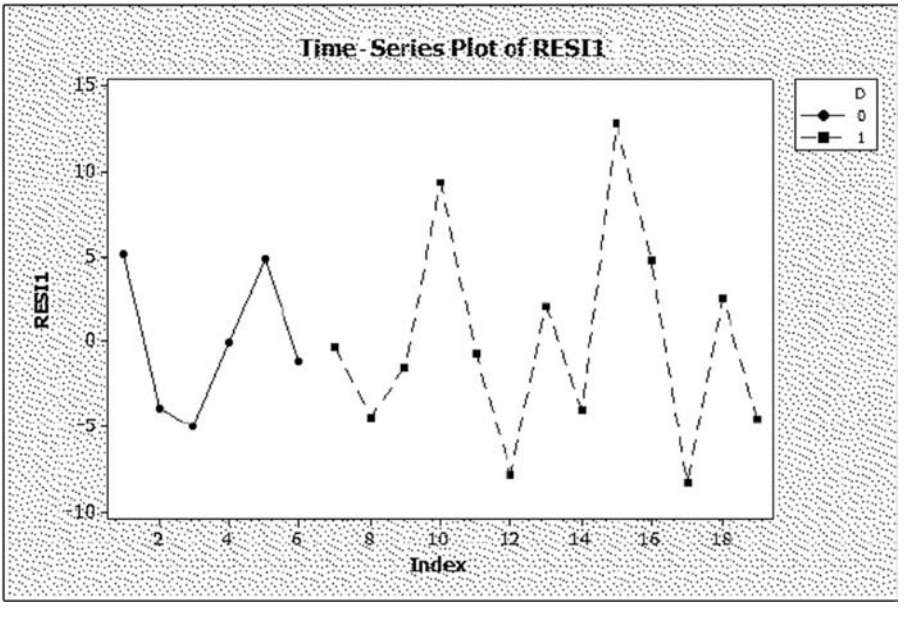


Figure 19.3 Model I residuals based on Birmingham, UK, annual cancer death rate data.

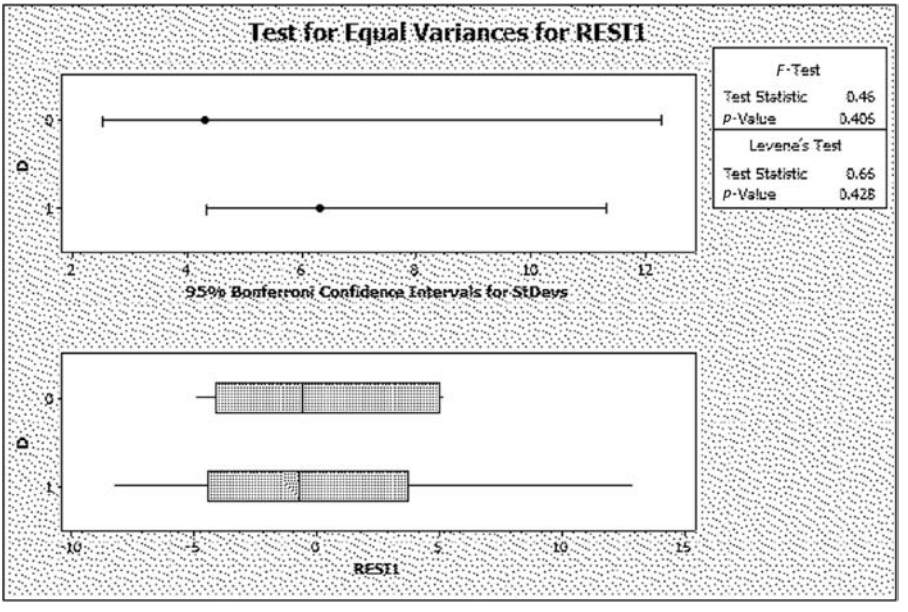


Figure 19.4 Tests for homogeneity of phase A and phase B model I error variances for Birmingham, UK, CDR data.

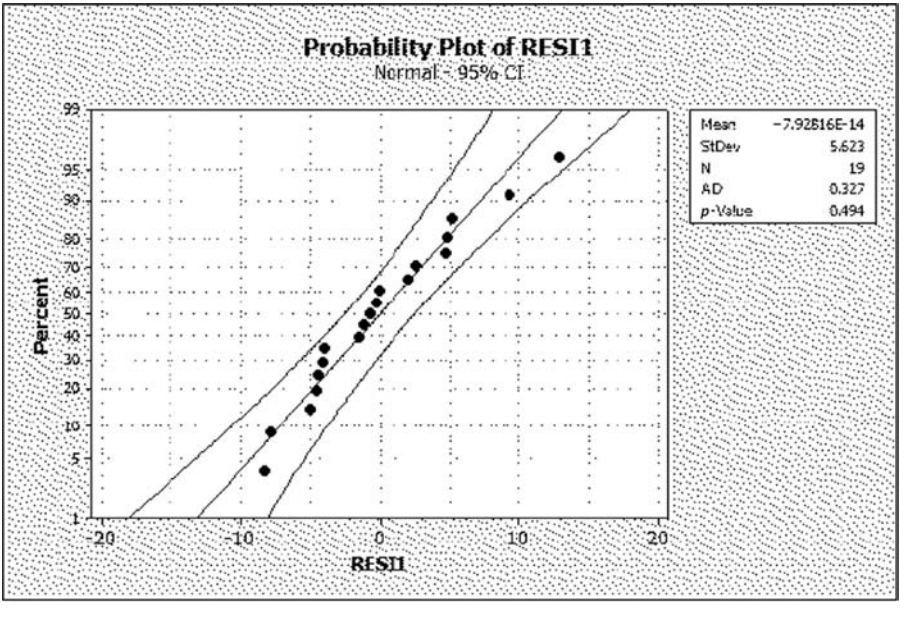


Figure 19.5 Probability plot of the residuals from fitting model I to Birmingham, UK, CDR data.

If the H–M test is applied instead of the D–W, the same conclusions are reached. The H–M nondirectional p-values for models I and II are .88 and .0002, respectively.

Note that the autocorrelation tests support the conclusion of the model comparison test in that both types of test identify strong evidence of misspecification when using model II. These two types of test focus on different aspects of the data. The model comparison test implies that much more variation is explained using model I than when using model II (93% vs. 57%, respectively), whereas the tests for autocorrelation lead to the conclusion that model II is inadequate to describe the data because the errors are dependent. One reason for autocorrelation is a failure to include important predictor variables in the model. The lack of autocorrelation among the errors of model I implies that there is no obvious need for additional parameters in the model. Because the errors appear not to be autocorrelated and the other diagnostics for model departure imply that the model is well specified, it is reasonable to accept model I and to proceed with the interpretation of the parameter estimates and the associated inferential results.

It can be seen that the value of the level change coefficient is 13.13 ($p = .06$). The meaning of this value can be seen by inspecting the regression lines in Figure 19.2. The year after the baseline ends is the first year (1965) that fluoridation could have had an effect. One can project forward from the baseline data using the first phase regression to year 1965, which is time period seven. The projected value is $b_0 + b_1(7) = 215 - .94(7) = 208.42$, a value that is consistent with a visual projection in the graph. This is the value that is predicted to occur in 1965 in the absence of an intervention. The elevation of the right-hand regression line at time period seven is equal to $b_0 + b_1(T) + b_2(D) + b_3(SC) = 215 - .94(7) + 13.13(1) + 5.04(0) = 221.55$. This is the CDR value predicted (from the second phase regression) to occur in 1965 under the condition that the fluoridation intervention has been introduced. Hence, two predictions are made regarding CDR in 1965. The difference between these predictions defines the level change coefficient. That is, $(221.55 - 208.42) = (13.13)$. But, because the p-value associated with this level change coefficient is larger than the value specified for alpha ($\alpha = .05$), it is concluded that the data are insufficient to state that there is a level change in the process.

The importance of correctly specifying the design matrix cannot be overstated. The CDR data described above were previously analyzed (Manly, 1993) using an alternative (but common) design matrix; the reported level change estimate was -9893.93 rather than the correct value (13.13). The reported value is conceptually impossible in terms of the substance of the study (as is obvious from a visual inspection of the data in either Figure 19.1 or Figure 19.2), but is the correct mathematical outcome of the inappropriate design matrix specification.

The slope change coefficient of 5.04 describes the change between the baseline slope and the intervention phase slope. The slope in the baseline phase (-0.94) plus the slope change (5.04) yields the slope in the second phase (4.10). A visual inspection of the data in Figure 19.2 for the intervention years (1965–1977) confirms that there was an average increase in CDR of about 4 units per year after fluoridation was introduced. Because the p-value associated with the slope change is only .005, the evidence for change appears to be persuasive.

What cannot be gleaned from this analysis, however, is the reason for the change. It is tempting to attribute the observed increase in CDR to the introduction of fluoridation, but this design is very vulnerable to the internal validity threats of "history" and "mortality" because it was carried out over a long period. Unlike the typical single-case design that involves data collected daily for several weeks or months, the fluoridation study used annual measurements collected over years (1959 through 1977) from a compound unit. As the duration of a study of this type increases, so does the plausibility of the argument that postbaseline condition changes and/or unit composition changes are responsible for change on the response measure. Indeed, additional data supplied by Cook-Mozaffare et al. (1981) and Cook-Mozaffare and Doll (1981) convincingly demonstrate that demographic changes in the population of Birmingham occurred during the same time interval as the fluoridation intervention; these changes (not the fluoridation intervention) best explain the increase in CDR. Problems of this type occur frequently with two-phase designs.

19.3 EXAMPLE II: FUNCTIONAL ACTIVITY

Dyer et al. (1984) studied the effects of an intervention (a supervision program) on a measure of functional activity in severely handicapped students. Linearly transformed data collected from one residential house are displayed in Figure 19.6 and are listed below:

Phase 1 ($n_1 = 16$): 5, 3, 10, 0, 6, 4, 6, 6, 5, 5, 8, 5, 4, 11, 7, 4.

Phase 2 ($n_2 = 31$): 12, 17, 16, 16, 12, 12, 15, 10, 17, 18, 12, 17, 16, 10, 14, 13, 17, 13, 16, 15, 18, 17, 14, 12, 13, 16, 14, 13, 17, 15, 17.

Three predictors (time, level change dummy variable, and slope change variable) were entered in the *Minitab* worksheet in columns C1 through C3; the dependent variable scores were entered in C4. Commands for estimating model I and portions of the associated output follow:

```
MTB > regr c4 on 3, c1-c3;
SUBC> resid c5.


Regression Analysis
The regression equation is

Y = 4.40 + 0.137 Time + 7.39 D - 0.101 SC

Predictor       Coef       Stdev    t-ratio        p
Constant       4.400       1.278       3.44    0.001
Time          0.1368      0.1322       1.03    0.307
D              7.388       1.538       4.80    0.000
SC            -0.1013      0.1410      -0.72    0.476

s = 2.438       R-sq = 77.5%     R-sq(adj) = 75.9%
```

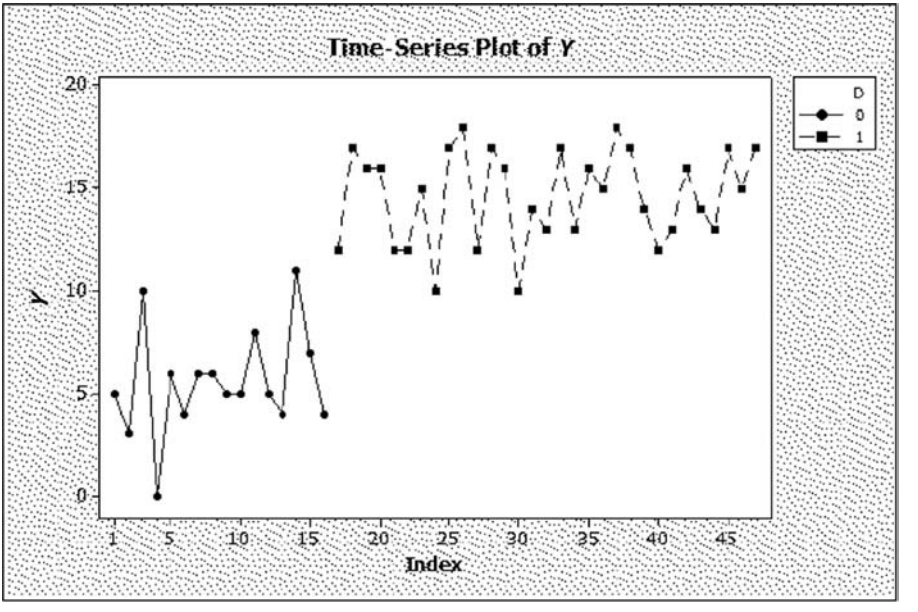


Figure 19.6 Functional activity data from a group of handicapped students living in a residential house. (Based on Dyer et al., 1984.)

```
Analysis of Variance

SOURCE        DF         SS         MS          F         p
Regression     3     880.06     293.35      49.36     0.000
Error         43     255.55       5.94
Total         46    1135.62
```

Next, model II was estimated and the model comparison test was carried out.

```
MTB > Name c6 = 'RESI2'
MTB > Regress 'Y' 1 'D';
SUBC>   Residuals 'RESI2'.


Regression Analysis

The regression equation is
Y = 5.56 + 9.08 D

Predictor       Coef       Stdev    t-ratio         p
Constant      5.5625      0.6067       9.17     0.000
D             9.0827      0.7471      12.16     0.000

s = 2.427       R-sq = 76.7%     R-sq(adj) = 76.1%
```

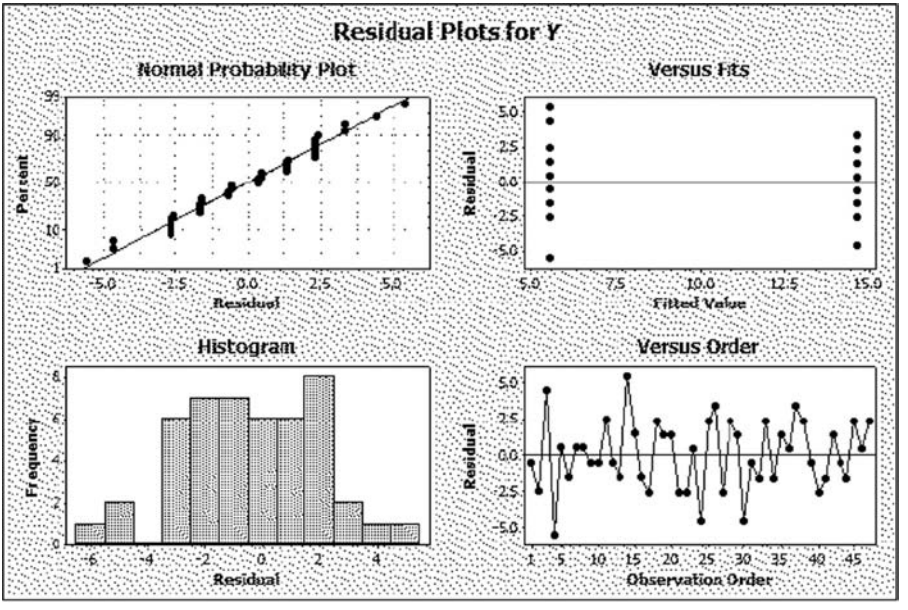


Figure 19.7 Four-in-one plots of the residuals from fitting model II to the functional activity data.

```
Analysis of Variance

SOURCE        DF         SS         MS          F         p
Regression     1     870.58     870.58     147.82     0.000
Error         45     265.03       5.89
Total         46    1135.62
```

$$\text{Model comparison test: } \frac{(880.06 - 870.58)/2}{5.94} = 0.80 = F.$$

The p-value associated with the obtained F (based on 2 and 43 degrees of freedom) is .46. It is concluded that a model with only the intercept and the level change dummy variable (i.e., model II) is satisfactory. The residuals of model II are then evaluated for conformity with assumptions of the model. The *Minitab* "four in one" plots of the residuals are shown in Figure 19.7.

These plots reveal no outliers and no obvious departures from linearity, homoscedasticity, or normality assumptions. The next issue is whether the errors of the model appear to be independent. The residuals from model II are in column C6; command *acf* provides a lag-1 autocorrelation estimate of $-.21$; the H–M test on this coefficient yields a p-value $= .23$. If the D–W test is used instead, it yields the

same conclusion. Hence, the assumption of independent errors appears to be satisfied; therefore, it is reasonable to proceed with the interpretion of the results of fitting model II. The level change coefficient = 9.08; this change is statistically significant ($p < .001$). The 95% confidence interval on the level change is (7.58, 10.59).

The unit of analysis in this example is the residential house; the postintervention functional activity level in this house is considerably more than three standard deviation units higher than it was during the baseline phase. Specifically, the standardized level change for the process = SLC = $\frac{9.08}{\sqrt{5.89}} = 3.74$. Although this is an extremely large improvement relative to the baseline for this individual house, there may be interest in interpreting the level increase (9.08 points) relative to a normative distribution of functional activity scores in an established population of interest. Suppose the mean in such a population is 20 and the standard deviation is 7. When the baseline level (5.5625) and the intervention level (14.6452) are transformed to population standard scores, the meaning of the level change can be stated in the population sense. The standard scores are

$$\frac{(5.5625 - 20)}{7} = -2.06 \text{ and}$$
$$\frac{(14.6452 - 20)}{7} = -0.76.$$

The areas of the standard normal distribution below −2.06 and −0.76, are .02 and .22, respectively. It appears that the intervention has changed the functional activity level for the average treated person in the residential house from the 2nd percentile to the 22nd percentile of the norm group of interest (i.e., the CPR is 20).

19.4 EXAMPLE III: CEREAL SALES

The data illustrated in Figure 19.8 were presented in a recent time-series analysis textbook (Montgomery et al., 2008) as representing weekly ($N = 104$) cereal sales for a specific brand. Another company introduced a competing product at the 88th time point; there was interest in determining whether the competing product had an effect on the sales of the original product. This is a nice data set for illustrating the difference between ARIMA modeling and regression modeling with autoregressive errors. I begin with a brief overview of some key aspects of ARIMA models and how they differ from the regression approach recommended in Chapter 18. Details of applying the regression approach to the sales data are also illustrated.

Montgomery et al. (2008) identified and estimated an ARIMA (0, 1, 1) intervention model (using standard procedures that are not described here). An ARIMA model is usually characterized in terms of (1) the autoregressive order, (2) the order of differencing, and (3) the moving average order. Hence, the ARIMA (0, 1, 1) notation tells us that the model used to analyze the sales data had no autoregressive parameters; one differencing parameter, and one moving average parameter. A

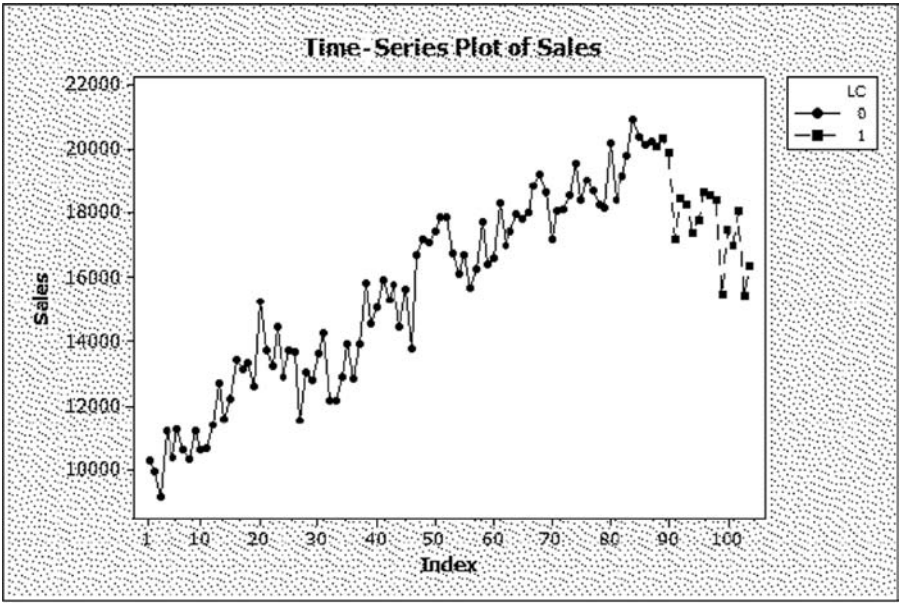


Figure 19.8 Cereal sales data before and after a competing cereal is introduced. (Data from Montgomery et al., 2008.)

thorough description of these parameters can be found in most time-series analysis books (e.g., Box et al., 2008; Montgomery et al., 2008); the focus here is on the differencing parameter because this is the first issue of concern in the modeling process.

It can be seen in Figure 19.8 that the data increase during the baseline phase. Acceptable estimates of autoregressive and moving average parameters are distorted when this type of nonstationarity is present. Hence, nonstationarity is removed by subjecting the data to one or more differencing transformations. In this example the differencing was specified as first order; hence, the transformation is simply $(Y_t - Y_{t-1}) = D_t$. First differencing removes trend that appears to be linear; second order differencing is required to remove nonlinear trend. A plot of the first differences can be seen in Figure 19.9.

Note that first differencing has removed the trend that is obvious in the original data shown in Figure 19.8. However, differencing has also introduced significant lag-1 autocorrelation ($r_1 = -.44$); this can be seen in the form of a choppy pattern in the differenced data. High differences tend to be followed immediately by low differences, and low differences tend to be followed immediately by high differences. The pattern of the autocorrelation function (not shown here) clearly indicates that the differences should be modeled using what is known as a first order moving average process. Hence, the (0, 1, 1) ARIMA model identified and estimated by Montgomery

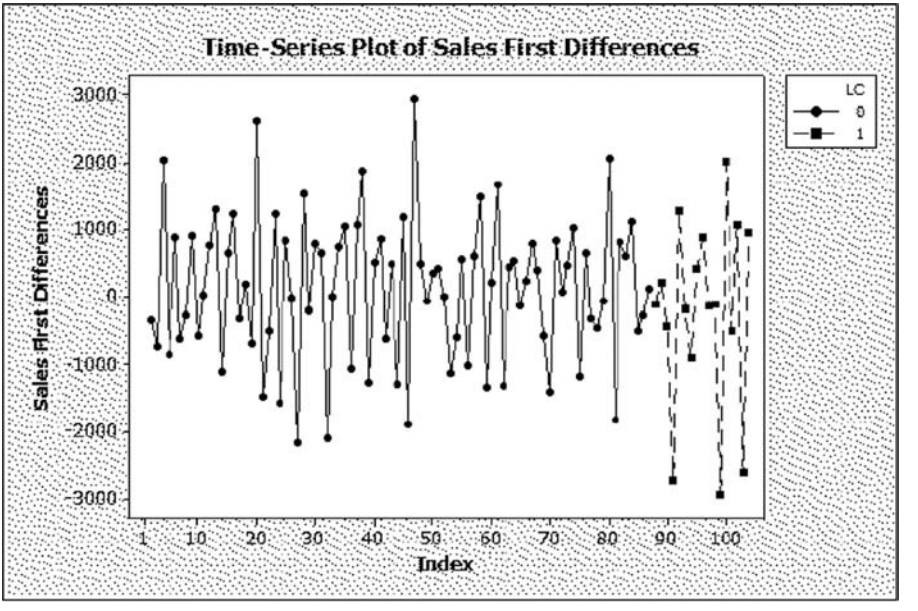


Figure 19.9 Plot of the first differences of the cereal sales data.

et al. (2008) is certainly correct. They present the following SAS output that describes the intervention effect:

The ARIMA Procedure

Conditional Least Squares Estimation

Parameter	Estimate	Standard Error	t value	Approx Pr > |t|	Lag	Variable	Shift
MA1,1	0.75886	0.06555	11.58	<.0001	1	Sales	0
NUM1	-2071.5	637.43397	-3.25	0.0016	0	Step	0

The first parameter estimate (0.75886) is the moving average coefficient (usually denoted as $\hat{\theta}$) used to model dependency in the differenced data. (Thorough descriptions of this measure can be found in Box et al., 2008; it will not be covered here.) The second parameter estimate (2071.5) is a measure of the intervention effect; I denote this estimate as $\hat{\delta}_w$. Because the p-value associated with this estimate is .0016, it is concluded that there is an intervention effect. The curious researcher should ask two questions about this output.

First, exactly what does the effect estimate -2071.5 actually mean in this study? One can neither point to a specific aspect of the graphed representation of the original data to see what -2071.5 means, nor find a detailed description of this estimator in the typical time-series textbook description of this model. The $\hat{\delta}_w$ coefficient is sometimes called a measure of level change, but it is not the same as the easily

computed and conceptualized level-change estimate based on the regression models discussed in Chapter 18.

The expression shown below (based on early mathematical-statistics work) describes the level change estimator for the (0, 1, 1) ARIMA intervention analysis:

$$\hat{\delta}_w = \frac{1-\hat{\theta}}{1-\hat{\theta}^{2n_2}}\left[\sum_{i=1}^{n_2}\hat{\theta}^{i-1}Y_{n_1+i} + \hat{\theta}^{n_2}\sum_{i=1}^{n_2}\hat{\theta}^{n_2-i}Y_{n_1+i}\right] - \frac{1-\hat{\theta}}{1-\hat{\theta}^{2n_1}}\left[\sum_{i=1}^{n_1}\hat{\theta}^{n_1-i}Y_i + \hat{\theta}^{i-1}Y_i\right],$$

where $\hat{\theta}$ is the estimate of the moving average parameter; n_1 and n_2 are the number of observations in the pre-and postintervention phases, respectively; and Y_i is the dependent variable score at the ith time period within a phase. It is not difficult to reach the conclusion that the meaning of this level change coefficient is anything but transparent.

Not surprisingly, the interpretation of $\hat{\delta}_w$ has tripped up many researchers and methodologists. It is widely believed that the test on $\hat{\delta}_w$ reduces to the independent sample t-test on the difference between the mean of the preintervention data and mean of the postintervention data when $\theta = 0$. This is far from true. It turns out that $\hat{\delta}_w$ refers to the difference between two exponentially weighted moving averages. This, of course, is not obvious to the researcher inspecting output of typical software providing results for this model. The observations immediately before and after the intervention are weighted very heavily; those near the beginning of the preintervention series and those near the end of the postintervention series receive very little weight. The weights change dramatically with the value of the moving average parameter. This is essential information that can be presented graphically (e.g., Huitema, 1986b), but it does not appear in the output of the major software packages. This is unfortunate because misinterpretations of $\hat{\delta}_w$ are inevitable without knowledge of how the data are weighted. For example, when the value of θ (the moving average coefficient) is zero, the effect estimate is based completely on only two observations! Specifically, the two observations include the one immediately before the intervention, and the one immediately after the intervention. All the other observations, regardless of the length of the series, are ignored because they are assigned weight zero.

In some studies, the observations immediately after the intervention are of least interest and the observations at the end of the intervention phase are of most interest. Or, in the more typical case, there is equal interest in all observations throughout the intervention phase. These are situations where the ARIMA (0, 1, 1) model is clearly not consistent with the purpose of the investigation. Regression models are almost always more reasonable in cases like this unless the process is known to contain a unit root, (which implies that the autoregressive parameter $a_1 = 1.0$).

The second question the curious researcher is likely to ask is, "Why is there only one effect measure in this example? Where is the slope change estimate?" After all, it seems obvious that there is a trend in the baseline and a change in trend after

intervention; it is natural to be interested in estimates of baseline slope and slope change. But measures of these characteristics are not a part of the ARIMA (0, 1, 1) intervention model and therefore they do not appear in the associated output. The reason they are not included is because under this model, trend is conceptualized as stochastic (i.e., random variation across time) rather than deterministic (i.e, systematic change across time). Hence, because trend is viewed as nuisance variation it is transformed away using differencing. In a sense the ARIMA model is saying, "Don't bother with the trend, it is just noise in the process." Hence, the absence of a description of trend or change in trend in the intervention analysis is consistent with the assumptions of the model. This is the right thing to do as long as within-phase trend is a generated by what is known as a unit root process. It is a mistake to estimate a regression equation on data generated by a unit root process because to do so is to simply chase random variation in a stochastic process.

In contrast, if the data are generated by what is known as a trend-stationary process, they are a deterministic function of time. It is a mistake to difference a trend-stationary process. Differencing such a process introduces autocorrelation. More importantly, a differenced series does not describe the dynamics of change in the original series; instead, it removes information regarding such change. This is why the output for the ARIMA (0, 1, 1) intervention model does not include estimates of slope and slope change.

Hence, slope and slope change are both viewed as irrelevant from the point of view of the ARIMA (0, 1, 1) model because these deterministic characteristics are assumed not to exist in an interrupted unit root (or difference-stationary) process. In contrast, these are necessary parameters in modeling an interrupted trend-stationary process that has deterministic slope and slope change; of course, regression models I and III described in Chapter 18 contain these parameters. The latter models use the baseline trend to estimate the counterfactual value of the observation that follows the intervention. If the data are trending in a deterministic fashion, this is the right thing to do. In contrast, the forecast function for the ARIMA (0, 1, 1) model is flat; this is why level change is the only effect estimate associated with this model. This makes sense only if the observed data trajectory (e.g., see Figure 19.8) is generated by a stochastic process rather than by a deterministic process.

The difference in the way trend is conceptualized under difference-stationary and trend-stationary processes brings up the obvious question, "How do you know which process generated the apparent trend in the observed data?" Curiously, I have never seen this topic mentioned in the area of intervention analysis. I know of no published intervention research (either ARIMA or regression-based) in either psychology or medicine that mentions the issue.

Occasionally, economists working in areas that do not involve intervention analysis apply one or more tests to determine whether the data-generating process is difference-stationary or trend-stationary. Such tests may be useful if a reasonable sample size is available. Among the more popular methods for this purpose are the Dickey–Fuller tests. When the two relevant variants of these tests are applied to the example data, they overwhelmingly reject ($p < .001$) the hypothesis that the cereal sales process contains a unit root. This means that the obvious increase in the baseline data should

not be viewed as stochastic; instead, it is argued that there is strong evidence for a deterministic trend and therefore regression methods should be considered. Most (but not all) two-phase time-series data that I have encountered in psychological and medical research appear to be adequately modeled using the regression approaches described in Chapter 18.

Regression Analysis of Sales Data

Application of the regression approach to the cereal sales data is illustrated next, beginning with *Minitab* commands for constructing the columns that can be used in carrying out analyses using *Minitab*, SPSS, and the time-series double bootstrap routine (TSDB).

```
MTB set c1
DATA> 104(1)
DATA> set c2
DATA> 1:104
DATA> set c3
DATA> 87(0) 17(1)
DATA> set c4
DATA> 88(0) 1:16
DATA> end
```

The resulting columns, along with the dependent variable column, are shown in Appendix 19.1. The complete collection of columns defines the **X:Y** matrix that is required to fit model III using TSDB. Column c1 need not be entered when using *Minitab* to fit models I and II. (The third column is labeled LC rather than D in this section to prevent confusion with the difference D that is used in ARIMA modeling.)

Model I Minitab Output

```
Regression Analysis: Sales versus Time, LC, SC

The regression equation is
Sales = 10259 + 115 Time - 812 LC - 326 SC

Predictor       Coef    SE Coef        T        P
Constant     10258.5      200.2    51.25    0.000
Time         115.318      3.951    29.19    0.000
LC            -811.8      474.1    -1.71    0.090
SC           -325.68      45.99    -7.08    0.000

S = 925.503   R-Sq = 90.8%   R-Sq(adj) = 90.5%
```

```
Analysis of Variance
Source            DF          SS           MS        F      P
Regression         3   842304783    280768261   327.79  0.000
Residual Error   100    85655655       856557
Total            103   927960438

Durbin-Watson statistic = 1.49738

Autocorrelation Function: RESI1

Lag            AC       T     LBQ
  1      0.251116    2.56    6.75
  2      0.187050    1.80   10.53
  3      0.171616    1.60   13.75
  4      0.088314    0.80   14.61
  5     -0.002339   -0.02   14.61
  6     -0.071761   -0.65   15.18
  7     -0.072002   -0.65   15.77
  8     -0.037140   -0.33   15.93
  9     -0.097115   -0.87   17.03
 10     -0.039035   -0.35   17.21
```

Model II Output

```
Regression Analysis: Sales versus LC

The regression equation is
Sales = 15332 + 2579 LC


Predictor     Coef   SE Coef       T        P
Constant   15332.5     306.4   50.03    0.000
LC          2579.2     758.0    3.40    0.001

S = 2858.35   R-Sq = 10.2%   R-Sq(adj) = 9.3%

Analysis of Variance

Source            DF          SS         MS        F       P
Regression         1    94604689   94604689    11.58   0.001
Residual Error   102   833355749    8170154
Total            103   927960438

Durbin-Watson statistic = 0.164745


MTB > ACF 'Sales Resid 2-P'.
```

```
Autocorrelation Function: Sales Resid 2-P

Lag         ACF       T      LBQ
  1    0.900646    9.18    86.82
  2    0.866171    5.45   167.90
  3    0.832520    4.18   243.55
  4    0.787455    3.42   311.91
  5    0.745219    2.93   373.75
  6    0.705982    2.57   429.82
  7    0.674750    2.31   481.56
  8    0.639105    2.09   528.47
  9    0.609375    1.91   571.56
 10    0.584337    1.77   611.60
```

Model Comparison and Autocorrelation Tests

The obtained F-value associated with the model comparison test = 436.46. Obviously there is a convincing justification to prefer model I to model II. Similarly, the D–W (and H–M) test leaves no doubt regarding the inadequacy of model II. The residuals of model II are unusually highly autocorrelated ($r_1 > .90$), which indicates a very poorly specified model. The residuals of model I are also significantly autocorrelated, but far less so ($r_1 = .25$). The model comparison test selects model I as the better of the two models, but the residuals of this model are autocorrelated. Therefore, model III is chosen for the final analysis. It can be seen in the lower right-hand panel of Figure 19.10 that there is some evidence of the classical meandering pattern that is typical of positively autocorrelated residuals.

Because the sample size is reasonably large ($N = 104$) the model can be estimated adequately using either the double bootstrap software (TSDB) or a maximum-likelihood routine for regression models with autoregressive errors (e.g., *SPSS Autoreg*). Portions of the output from both approaches are illustrated below. TSDB output is shown first.

TSDB Output

```
                TSDB Timeseries Results on DV = Y/10000

Parameter Estimates and Test that parameter is zero

Parameter    Estimate  t-ratio   p-value
Beta 1       1.027069   36.676   2.00252e-59
Beta 2        .011483   21.132   1.80194e-38
Beta 3       -.073061   -1.216   0.226876
Beta 4       -.032890   -5.193   1.1075e-06

(Note: The four coefficients are numbered as
1-4 rather than 0-3)
```

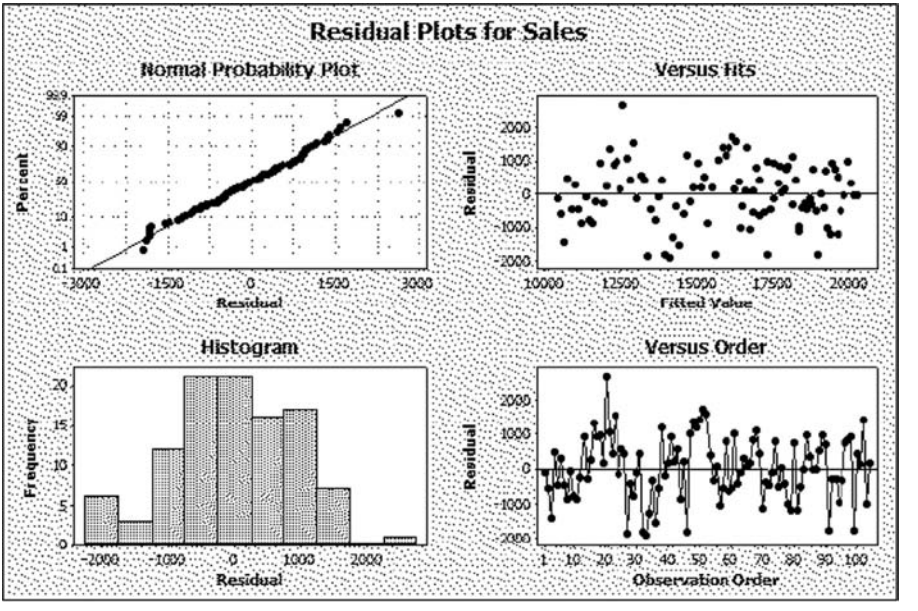


Figure 19.10 Residuals of model I for cereal sales data.

```
    Variance      Covariance  Matrix of Parameter     Estimates
 0.000784206  -1.21593e-05           0.000115795   3.13468e-05
-1.21593e-05   2.95303e-07          -8.80331e-06  -5.80965e-07
 0.000115795  -8.80331e-06            0.00360783  -0.000282953
 3.13468e-05  -5.80965e-07          -0.000282953    4.0113e-05

 AR 1 = .311
```

Because the original sales metric (see Figure 19.8) yields very large values (some over 20,000), it causes an estimation problem when using the TSDB program. An AR1 coefficient of .99 (found in the present example) is evidence that the regression coefficients and the associated *t*- and *p*-values are incorrect; this problem is easily solved. Simply divide the original outcome scores by a value (most conveniently 10, 100, 1000, or 10,000) that transforms them to values not exceeding the number of observations in the study. Each original value in the sales example was divided by 10,000, as indicated on the top of the output for TSDB shown on the previous page. (However, 1000 could have been used instead; both values result in transformed scores far less than *N*.) It follows that the parameter estimates shown in the output (but not the AR1 estimate) must be multiplied by 10,000 in order to interpret them in terms of the original sales metric. Therefore, the level change estimate is 10,000 $(-.073061) = -730.61$, and the slope change estimate is $10{,}000(-.032890) = -328.90$. The *t*-values and *p*-values shown in the output apply to both the original and the transformed data.

SPSS Autoreg Output

```
FINAL PARAMETERS:

Number of residuals  104
Standard error       900.30613
Log likelihood       -852.58457
AIC                  1715.1691
SBC                  1728.3911


Variables in the Model:

                    B         SEB          T-RATIO  APPROX. PROB.
AR1               .24928      .09715      2.565924      .01178891
VAR00002     115.17489     5.05569     22.781260      .00000000
VAR00003    -755.28595   576.05532     -1.311134      .19284535
VAR00004    -329.41841    56.04880     -5.877350      .00000006
CONSTANT   10261.67558   256.82005     39.956676      .00000000
```

The level-change and slope-change estimates and tests based on TSDB and *Autoreg* procedures are similar but not identical. One reason for the difference is that the bootstrap approach provides less biased estimates of the autoregressive parameter (note that the estimates are .31 and .25 for TSDB and *SPSS Autoreg*, respectively). Because autoregressive coefficients affect the regression coefficient estimates, the results of the two procedures are expected to be slightly different. Similarly, because the standard errors for the two procedures are based on different foundations (i.e., empirical for TSDB and theoretical for *SPSS Autoreg*), the inferential aspects are also expected to be somewhat different.

The change estimates and inferential results based on OLS, *Autoreg*, and TSDB are summarized in Table 19.2 along with the level-change estimate reported in Montgomery et al. (2008) for the (0, 1, 1) ARIMA intervention model.

It can be seen that the level-change estimates produced by OLS, TSDB, and *Autoreg* are similar, and that the ARIMA estimate is very different. The slope change estimates associated with the three regression methods are almost identical. The slope

Table 19.2 Summary of Change Statistics Associated with OLS, TSDB, *SPSS Autoreg*, and *SAS* ARIMA (0, 1, 1) Applied to Cereal Sales Data

Method	Level Change	LC *t*-Value	LC *p*-Value	Slope Change	SC *t*-Value	SC *p*-Value
OLS	−812	−1.71	.09	−326	−7.08	<.001
TSDB	−731	−1.22	.23	−329	−5.19	<.001
SPSS Autoreg	−755	−1.31	.19	−329	−5.88	<.001
SAS ARIMA (0, 1, 1)	−2072	−3.25	.0016	–	–	–

estimate for the baseline phase (not shown in the table) is essentially the same value (115) for all three regression methods. The intervention appears to have changed the slope from 115 during baseline to about −214 after intervention (because the baseline slope plus the slope change = 115 − 329 = −214). This essential descriptive information is absent from the ARIMA analysis.

19.5 EXAMPLE IV: PARACETAMOL POISONING

Morgan et al. (2007) were interested in the effects of legislation that was introduced in 1998 to limit the pack size of paracetamol (known as acetaminophen in the United States) sold in shops in the United Kingdom. This legislation was introduced as a possible way to reduce paracetamol poisoning, which is the leading cause of liver failure in both the United States and Great Britain. The researchers initially focused on age-standardized mortality rates for paracetamol poisoning in England and Wales for the years 1993 through 2004. Because the effects of the legislation were expected to first appear in 1999, the years 1993 through 1998 were considered the baseline phase and years 1999 through 2004 were considered the intervention phase. The data are illustrated in Figure 19.11. This is an exceedingly short series, but it represents only a portion of a study that contains other data.

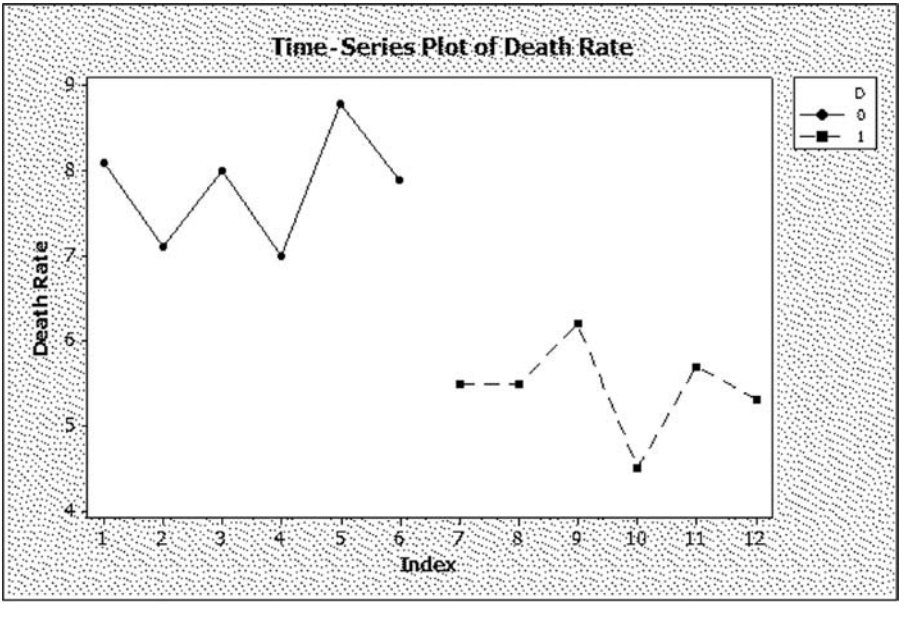


Figure 19.11 Age-standardized death rate per million associated with poisoning involving paracetamol for years 1993 (index value 1) through 2004 (index value 12). (Adapted from Morgan et al., 2007.)

Minitab input and output used to fit models I and II is shown below:

```
MTB > set c1
DATA> 12(1)
DATA> set c2
DATA> 1:12
DATA> set c3
DATA> 6(0) 6(1)
DATA> set c4
DATA> 7(0) 1:5
DATA> end
MTB > Print c1-c5

 Data Display

Row  Unity  Time    D  SC  PDR
  1      1     1    0   0  8.1
  2      1     2    0   0  7.1
  3      1     3    0   0  8.0
  4      1     4    0   0  7.0
  5      1     5    0   0  8.8
  6      1     6    0   0  7.9
  7      1     7    1   0  5.5
  8      1     8    1   1  5.5
  9      1     9    1   2  6.2
 10      1    10    1   3  4.5
 11      1    11    1   4  5.7
 12      1    12    1   5  5.3
```

The column of ones is constructed in case the TSDB program is necessary for the analysis; of course, this column is not necessary for estimating the OLS models using *Minitab*.

Model I

Input Commands:

```
MTB > Name c6 "RESI1"
MTB > Regress 'PDR' 3 'Time'-'SC';
SUBC>   Residuals 'RESI1';
SUBC>   Constant;
SUBC>   DW;
SUBC>   Brief 1.
```

Output:

```
 Regression Analysis: PDR versus Time, D, SC

The regression equation is
```

```
PDR = 7.51 + 0.089 Time - 2.53 D - 0.149 SC


Predictor      Coef  SE Coef      T      P
Constant     7.5067   0.6265  11.98  0.000
Time         0.0886   0.1609   0.55  0.597
D           -2.5267   0.7935  -3.18  0.013
SC          -0.1486   0.2275  -0.65  0.532


S = 0.672964   R-Sq = 82.4%   R-Sq(adj) = 75.8%


Analysis of Variance

Source           DF       SS      MS      F      P
Regression        3  17.0036  5.6679  12.52  0.002
Residual Error    8   3.6230  0.4529
Total            11  20.6267

Durbin-Watson statistic = 3.28732
```

Model II

Input Commands:

```
MTB > Name c7 "RESI2"
MTB > Regress 'PDR' 1 'D';
SUBC>   Residuals 'RESI2';
SUBC>   GFourpack;
SUBC>   RType 1;
SUBC>   Constant;
SUBC>   DW;
SUBC>   Brief 1.
```

Output:

```
Regression Analysis: PDR versus D

The regression equation is
PDR = 7.82 - 2.37 D

Predictor      Coef  SE Coef     T      P
Constant     7.8167   0.2524  30.97  0.000
D           -2.3667   0.3570  -6.63  0.000

S = 0.618331   R-Sq = 81.5%   R-Sq(adj) = 79.6%
```

```
Analysis of Variance

Source          DF      SS      MS      F      P
Regression       1  16.803  16.803  43.95  0.000
Residual Error  10   3.823   0.382
Total           11  20.627

Durbin-Watson statistic = 3.09706

MTB > ACF 'RESI2'.

Autocorrelation Function: RESI2

Lag         ACF      T    LBQ
  1   -0.561973  -1.95  4.82
  2    0.293665   0.80  6.27
  3   -0.260898  -0.67  7.54
```

Model Comparison Test

$$\frac{(17.004 - 16.803)/2}{.453} = 0.222 = F$$

The p-value associated with the obtained F is .81; therefore, model II is selected as the more appropriate of the two. The residuals from fitting this model are illustrated in Figure 19.12.

The variation in the two phases is similar and no outliers are present. Because the number of observations is exceedingly small it is not expected that normality will be well approximated in the histogram, but the normal probability plot does not reveal any unusual departures. The plot in the lower right suggests negative autocorrelation because the pattern is one in which a small residual tends to be followed by a large residual and a large residual tends to be followed by a small residual. This impression is confirmed by the lag-1 autocorrelation coefficient of $-.56$. The H–M test on this coefficient is statistically significant when α is set at the recommended level of .20 (nondirectional). Hence, inferential tests and confidence intervals from OLS are conservative (i.e., t-values are too small in absolute value and confidence intervals are too wide). Model IV is selected because the data are adequately described using two parameters (to describe level and level change) plus an autoregressive coefficient to capture the dependency in the errors.

The TSDB program for fitting model IV requires a combined **X:Y** matrix as input. Because all the required columns for this matrix were included among those produced using *Minitab* input commands (shown above), *Minitab* can be used to easily provide the combined **X:Y** matrix. The design matrix required for model IV consists of the

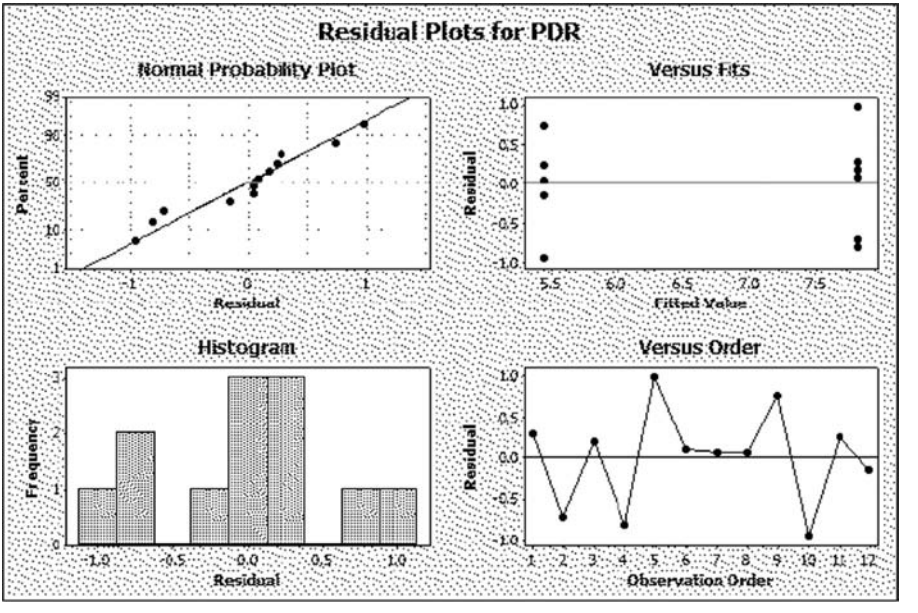


Figure 19.12 Residual plots for paracetamol death rate data.

unity column and the dummy variable column; these are columns c1 and c3 in the original (i.e., model I) matrix. The dependent variable vector **Y** is column c5 of the model I matrix. Hence, only columns c1, c3, and c5 are needed. But, if we use the *Minitab* "print" command to print these three columns, we will discover that the "row" column is always printed; this column should not be included in the **X:Y** matrix. The way around this problem is to copy the variables of interest into a matrix. The following commands copy variables c1, c3, and c5 into a matrix labeled as m1; this is the required **X:Y** matrix for fitting model IV using program TSDB.

```
MTB > copy c1 c3 c5 m1
MTB > print m1

Data Display
Matrix M1

1   0   8.1
1   0   7.1
1   0   8.0
1   0   7.0
1   0   8.8
1   0   7.9
1   1   5.5
1   1   5.5
```

```
 1   1   6.2
 1   1   4.5
 1   1   5.7
 1   1   5.3
```

Once this matrix is printed, go to http://fisher.stat.wmich.edu/joe/TSDB/Timeseries.html

Cut and paste the matrix into the "Data Input" box that appears on the first page of TSDB; then click the "Submit Data" button. The output appears on the second page as shown below.

```
Timeseries Results

Parameter Estimates and Test that parameter is zero

Parameter       Estimate    t-ratio        p-value
Beta 1          7.790834     37.246    3.59472e-11
Beta 2         -2.315002     -7.248    4.82665e-05

Variance Covariance Matrix of Parameter Estimates

 0.0437542          -0.0504428
-0.0504428           0.102015

Bootstrap Residual MSE = .365827

Bootstrap Estimates of AR Parameters

Parameter          Estimate
     AR 1             -.235

Variance Covariance Matrix of AR Estimates

           .235096
```

Note that the level-change estimate is -2.32, which is similar to the OLS estimate of -2.37. The corresponding t-values on these estimates are -7.25 and -6.63 for TSDB and OLS, respectively.

At this stage it is tempting to conclude that the intervention has reduced the death rate by 2.32 deaths per million. But the cautious researcher should ask two questions. First, are sufficient data available in the meager baseline ($n = 6$) phase to know the nature of the process? Second, are there jumps of the size observed here in the total series (i.e., combined baseline to intervention phases) also observed in a long baseline series in the absence of intervention? Because we do not have data from a long baseline series we simply do not know the answer to this crucial question; therefore, the results should be viewed as suggestive but not convincing even though the p-value is less than .00005. The second question is, "Are there plausible explanations other

than the intervention that can explain the decrease in paracetamol poisoning deaths?" Fortunately, the investigators collected additional data that sheds light on this issue.

Time-series data regarding fatal poisoning involving aspirin, paracetamol compounds, antidepressants, and nondrug poisoning suicide were collected in addition to data on paracetamol poisoning. Interestingly, these series also move downward after the introduction of the intervention. Although these results do not point to a specific confounding variable that affects all types of fatal poisoning, they suggest that the drop in paracetamol poisoning was a part of a general downward national trend in poisoning deaths due to unspecified causes. These data cast serious doubt on the preliminary conclusion that the legislation intervention on packet size was the cause of the drop in paracetamol poisoning. A formal method for comparing evidence for the drop in paracetamol poisoning relative to evidence for the drop in poisoning from other drugs is presented in Chapter 21.

19.6 SUMMARY

Four examples of the analysis of two-phase designs are illustrated in this chapter. Two of the examples describe major problems in attempting to draw causal conclusions from two-phase designs that involve long data collection periods and few observations. An example of a long series clarifies the differences between the recommended regression approaches and the more popular ARIMA approach. The nature of the models required to adequately describe data collected in very different application areas is demonstrated often to be quite simple. The need for both additional data and more sophisticated designs is also noted.

APPENDIX 19.1 X:Y INPUT MATRIX FOR TSDB PROGRAM (CEREAL SALES DATA FROM MONTGOMERY ET AL., 2008)

1	1	0	0	10,245
1	2	0	0	9,893
1	3	0	0	9,155
1	4	0	0	11,194
1	5	0	0	10,338
1	6	0	0	11,212
1	7	0	0	10,578
1	8	0	0	10,300
1	9	0	0	11,192
1	10	0	0	10,617
1	11	0	0	10,635
1	12	0	0	11,392
1	13	0	0	12,686
1	14	0	0	11,568
1	15	0	0	12,204
1	16	0	0	13,435

APPENDIX 19.1 (*Continued*)

1	17	0	0	13,120
1	18	0	0	13,299
1	19	0	0	12,602
1	20	0	0	15,222
1	21	0	0	13,735
1	22	0	0	13,224
1	23	0	0	14,455
1	24	0	0	12,873
1	25	0	0	13,704
1	26	0	0	13,683
1	27	0	0	11,498
1	28	0	0	13,025
1	29	0	0	12,807
1	30	0	0	13,597
1	31	0	0	14,237
1	32	0	0	12,130
1	33	0	0	12,138
1	34	0	0	12,879
1	35	0	0	13,929
1	36	0	0	12,853
1	37	0	0	13,926
1	38	0	0	15,796
1	39	0	0	14,531
1	40	0	0	15,034
1	41	0	0	15,898
1	42	0	0	15,269
1	43	0	0	15,744
1	44	0	0	14,450
1	45	0	0	15,634
1	46	0	0	13,744
1	47	0	0	16,675
1	48	0	0	17,164
1	49	0	0	17,083
1	50	0	0	17,425
1	51	0	0	17,848
1	52	0	0	17,856
1	53	0	0	16,717
1	54	0	0	16,120
1	55	0	0	16,671
1	56	0	0	15,643
1	57	0	0	16,244
1	58	0	0	17,726
1	59	0	0	16,392
1	60	0	0	16,604
1	61	0	0	18,279
1	62	0	0	16,951
1	63	0	0	17,394

(*Continued*)

APPENDIX 19.1 (*Continued*)

1	64	0	0	17,935
1	65	0	0	17,798
1	66	0	0	18,018
1	67	0	0	18,807
1	68	0	0	19,193
1	69	0	0	18,607
1	70	0	0	17,186
1	71	0	0	18,024
1	72	0	0	18,091
1	73	0	0	18,542
1	74	0	0	19,547
1	75	0	0	18,368
1	76	0	0	19,020
1	77	0	0	18,697
1	78	0	0	18,233
1	79	0	0	18,156
1	80	0	0	20,213
1	81	0	0	18,374
1	82	0	0	19,188
1	83	0	0	19,795
1	84	0	0	20,904
1	85	0	0	20,399
1	86	0	0	20,122
1	87	0	0	20,237
1	88	1	0	20,110
1	89	1	1	20,321
1	90	1	2	19,877
1	91	1	3	17,157
1	92	1	4	18,432
1	93	1	5	18,246
1	94	1	6	17,343
1	95	1	7	17,768
1	96	1	8	18,646
1	97	1	9	18,514
1	98	1	10	18,397
1	99	1	11	15,463
1	100	1	12	17,472
1	101	1	13	16,958
1	102	1	14	18,031
1	103	1	15	15,408
1	104	1	16	16,356

CHAPTER 20

Analysis of Single-Case Reversal Designs

20.1 INTRODUCTION

The major weakness of the AB two-phase design is potential confounding of the intervention effect with the effect of other events that occur at the same time. This type of confounding is known in the quasi-experimental design literature as the internal validity threat of history. Several more sophisticated time-series single-case designs are available to contend with this issue. The two most frequently used variants are known as the reversal (or withdrawal) design and the multiple baseline design. Both of them dramatically reduce confounding effects of this type and provide additional advantages as well. These designs are described thoroughly in the major single-case design references (e.g., Barlow et al., 2009; Cooper et al., 2006; Johnston and Pennypacker, 2008). The focus of the present chapter is on the reversal design.

The simplest version of the reversal design begins as an AB design, but when the intervention condition is withdrawn at the end of the second phase, the conditions present during the baseline are reintroduced during a third phase. This arrangement is described as the ABA reversal design. The third phase provides additional evidence regarding the effect of the intervention. If the dependent variable scores change in the desired direction when the intervention is introduced in the B phase and then change back to the original baseline level when the intervention is withdrawn, the evidence for an intervention effect is stronger than in the case of an AB design. These designs may effectively rule out the interpretation that some event other than the intervention is responsible for change on the dependent variable.

Frequently a fourth phase is added, during which the intervention is introduced a second time. The resulting arrangement is called an ABAB design. This design provides very strong evidence for causal effects when (1) the behavior during the two baseline phases is similar, (2) the behavior during the two intervention phases

The Analysis of Covariance and Alternatives: Statistical Methods for Experiments, Quasi-Experiments, and Single-Case Studies, Second Edition. Bradley E. Huitema.

is similar, (3) a rapid change occurs with the introduction of the intervention each time, and (4) a rapid return to the baseline level occurs with the withdrawal of the intervention each time. Many researchers favor this design because the effect of introducing the intervention is demonstrated twice. Further, the effect of withdrawing the intervention is demonstrated twice and the experiment ends in the intervention condition. The condition present at the end of the experiment is often important in clinical work where it may be unethical to remove a condition (such as a treatment for high blood pressure) that has been demonstrated to have desirable effects.

Although reversal designs are usually quite persuasive in the evaluation of many types of behavioral and medical intervention, the situations in which they can be used are somewhat restricted. There are several logical, ethical, and practical problems that may rule out the use of these designs. Reversal designs are often impractical because many interventions are one-shot treatments with irreversible effects such as surgical ablation or teaching a skill (e.g., how to ride a bicycle). Once these treatments are applied it may be illogical to expect the behavior of the subject to revert to a preintervention level. Hence, reversal designs are appropriate only if the dependent variable scores can be expected to return to the level observed during the baseline phase when the intervention is withdrawn. Occasionally, behavior expected to revert to baseline level sometimes does not do so, often because conditions other than the intervention maintain it. In this case the evidence for an intervention effect is less persuasive than when the behavior reverses.

In other situations, ethical arguments are incompatible with design requirements. For example, if an intervention changes a behavior that is harmful to the subject it may be argued that it should not be withdrawn. Problems of this type reduce the number of situations in which reversal designs are appropriate. Most of these problems can be solved using some version of the multiple baseline design. This design is considered in Chapter 21.

20.2 STATISTICAL ANALYSIS OF REVERSAL DESIGNS

The traditional behavior-analytic approach for the analysis of reversal designs is visual inspection of the graphed data (e.g., Kazdin, 1982; Parsonson and Baer, 1986; Poling et al., 1995). Occasionally, however, a statistical method is used to analyze the data from these designs. Sometimes ABA and ABAB designs are analyzed by applying a simple two-phase method to each pair of adjacent phases. There are four potential problems with this approach. First, serious flaws are associated with many analyses that have been proposed for two-phase designs (see Huitema, 2004; Huitema and McKean, 2000b; Huitema et al., 2008); such problems are multiplied if these methods are applied to reversal and multiple baseline designs. Second, even in the case where the two-phase method is adequate for a single comparison, the piecemeal approach of performing a separate analysis on each pair of adjacent phases in complex reversal designs results in power that is lower than it would be if all data in the experiment were considered simultaneously. Third, multiple comparison error rate problems are not acknowledged. Fourth, and most important, the focus on each individual two-phase comparison rather than on the overall pattern of all changes in

the experiment ignores the logic of the design and is inconsistent with the approach used in an adequate visual analysis.

The logic of the reversal design should be reflected in the statistical model. Consider the typical ABAB design. An intervention is viewed as effective if a specific pattern of change from phase to phase is observed. Suppose the responses are expected to be higher during intervention conditions than during baseline conditions. A predicted pattern of sequential change from one phase to the next is easily specified in this case. That is, it is predicted that there will be (1) an increase after the first phase, (2) a decrease after the second phase, and (3) an increase after the third phase. If interventions act as hypothesized before the experiment is carried out, the predicted pattern of change and the observed pattern of change will be consistent. Hence, an adequate statistical analysis should formally incorporate the predicted change pattern in evaluating the strength of the evidence for intervention effects. Methods of this type are introduced next.

The recommended analysis provides measures of level change and slope change for each pair of adjacent phases, overall measures of change, significance tests on all change measures, and two types of standardized effect size. The following six steps are involved in performing the analysis:

Step 1: Predict the direction of change for each pair of adjacent phases. Although change can be described with respect to both level and slope, it is adequate to simply predict the direction of change (i.e., increase or decrease) without specifying additional detail regarding the form of change. This should be done during the design stage before data have been collected. It is unacceptable to collect the data first and then specify the direction of change based on the outcome of the study. The predictions should be based on theory and/or previous research results; the analysis does not apply to studies having no basis for predictions.

The predicted direction of change is recorded using negative and positive signs. A negative sign is recorded if a decrease is expected to occur and a positive sign is recorded if an increase is expected to occur. The collection of these signs defines the anticipated pattern of change from one phase to another. The number of condition changes in a reversal design with J phases is $J - 1$. Hence, an ABA design includes two changes; an ABAB design has three changes, and so on.

Step 2: Fit the full (level-change and slope-change) model. In the case of the ABAB reversal design, the full regression model is written as follows:

$$Y_t = \beta_0 + \beta_1 T_t + \beta_2 LC_{1t} + \beta_3 SC_{1t} + \beta_4 LC_{2t} + \beta_5 SC_{2t} + \beta_6 LC_{3t} + \beta_7 SC_{3t} + \varepsilon_t,$$

where

Y_t is the dependent variable score at time t;
β_0 is the intercept for the first phase;
β_1 is the slope for the first phase;
β_2 is the first level change (phase1 vs. phase 2);
β_3 is the first slope change (phase1 vs. phase 2);
β_4 is the second level change (phase 2 vs. phase 3);
β_5 is the second slope change (phase 2 vs. phase 3);

β_6 is the third level change (phase 3 vs. phase 4);
β_7 is the third slope change (phase 3 vs. phase 4);
T_t is the time variable T measured at time t;
LC_{1t}, LC_{2t}, and LC_{3t} are the level-change dummy variables measured at time t;
$SC_{1t,}$ SC_{2t}, and SC_{3t} are the slope-change variables measured at time t; and
ε_t is the error measured at time t.

The design matrix required for fitting this model is shown in Table 20.1.

When conventional software for ordinary least-squares (OLS) regression is used as the estimation method, it is not necessary to enter the first column of the design matrix on the worksheet; most regression routines (e.g., *Minitab*, *SAS*, *and SPSS*) automatically produce this column by default unless the user specifically requests that the model be fitted with no intercept.

Although this design matrix applies to experiments with four phases, it is easily modified for designs with fewer phases. For example, if the design has only three phases (e.g., an ABA design), simply remove the last two columns and the bottom rows (associated with the last phase) from this matrix. In the case of the AB design,

Table 20.1 Design Matrix X for Fitting the Full Model to ABAB Reversal Data

$$
\mathbf{X} = \left[\begin{array}{cccccccc}
1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\
1 & 2 & 0 & 0 & 0 & 0 & 0 & 0 \\
. & . & . & . & . & . & . & . \\
. & . & . & . & . & . & . & . \\
. & . & . & . & . & . & . & . \\
1 & n_1 & 0 & 0 & 0 & 0 & 0 & 0 \\
\hline
1 & n_1+1 & 1 & 0 & 0 & 0 & 0 & 0 \\
1 & n_1+2 & 1 & 1 & 0 & 0 & 0 & 0 \\
. & . & . & 2 & . & . & . & . \\
. & . & . & . & . & . & . & . \\
. & . & . & . & . & . & . & . \\
1 & n_1+n_2 & 1 & n_2-1 & . & . & . & . \\
\hline
1 & n_1+n_2+1 & 1 & n_2 & 1 & 0 & 0 & 0 \\
1 & n_1+n_2+2 & 1 & n_2+1 & 1 & 1 & 0 & 0 \\
. & . & . & . & . & . & . & . \\
. & . & . & . & . & . & . & . \\
. & . & . & . & . & . & . & . \\
1 & n_1+n_2+n_3 & 1 & n_2+n_3-1 & 1 & n_3-1 & 0 & 0 \\
\hline
1 & n_1+n_2+n_3+1 & 1 & n_2+n_3 & 1 & n_3 & 1 & 0 \\
1 & n_1+n_2+n_3+2 & 1 & n_2+n_3+1 & 1 & n_3+1 & 1 & 1 \\
. & . & . & . & . & . & . & 2 \\
. & . & . & . & . & . & . & . \\
. & . & . & . & . & . & . & . \\
1 & n_1+n_2+n_3+n_4 & 1 & n_2+n_3+n_4-1 & 1 & n_3+n_4-1 & 1 & n_4-1
\end{array}\right]
$$

n_1, n_2, n_3, and n_4 refer to the number of observations in phases one through four, respectively.

the last four columns and all rows associated with the last two phases are removed; the result is the four-parameter matrix described in Chapter 19.

Similarly, the matrix can be expanded if more than four phases are involved. Two additional columns that follow the same pattern shown here are required for each additional phase; rows to accommodate the additional phase(s) are entered at the bottom. For example, when there are five phases the two added columns consist of a 0–1 column (ones indicating fifth phase observations and zeroes otherwise) and a column with zeroes for the first four phases followed with zero through $n_5 - 1$ for the fifth phase.

Step 3: Fit the reduced (level-change) model. The reduced model includes an intercept and level-change parameters, but excludes the time and slope-change parameters contained in the full model. Hence, only columns one, three, five, and seven of the original full model **X** matrix for the ABAB design (shown in Table 20.1) are used. The reduced model is written as follows:

$$Y_t = \beta_0 + \beta_1 LC_{1t} + \beta_2 LC_{2t} + \beta_3 LC_{3t} + \varepsilon_t.$$

The level-change parameters in this model are equivalent to differences between adjacent phase means. These differences do not include all pairwise contrasts. For example, there is no direct comparison between phases one and four. Although it is possible to develop a separate test for comparisons of this type, I recommend against it. Whereas the comparison of adjacent phases involves relatively little extrapolation, the comparison of phases separated by many time points stretches the credibility of the model. Although interest in the comparison of nonadjacent phases is natural, the estimates based on such comparisons are often untrustworthy; hence, they should be avoided.

Step 4: Compare the full and reduced models. A test comparing the full and reduced models determines whether the first slope and the set of slope-change parameters included in the full model are necessary. If the model comparison F-test is statistically significant, it is convincing evidence that the full model is more satisfactory than the reduced model. If the test is nonsignificant, the need to retain the slope and slope-change parameters is not established and the reduced model is selected unless there is a compelling argument to include them.

The form of the test is shown below:

$$\frac{[\mathrm{SS}_{\mathrm{Reg(Full)}} - \mathrm{SS}_{\mathrm{Reg(Reduced)}}]/(df_{\mathrm{Reg(Full)}} - df_{\mathrm{Reg(Reduced)}})}{\mathrm{MS}_{\mathrm{Res(Full)}}} = F,$$

where

$\mathrm{SS}_{\mathrm{Reg\ Full}}$ is the regression sum of squares for the full model;
$\mathrm{SS}_{\mathrm{Reg\ Reduced}}$ is the regression sum of squares for the reduced model;
df_{Full} is the regression degrees of freedom for the full model;
df_{Reduced} is the regression degrees of freedom for the reduced model; and
$\mathrm{MS}_{\mathrm{Res\ (Full)}}$ is the residual mean square for the full model.

The obtained F is compared with the critical value of the F-distribution based on $(df_{\mathrm{Full}} - df_{\mathrm{Reduced}})$ numerator degrees of freedom, and $N - m - 1$ denominator degrees of freedom (where m is the number of predictors in the full model).

Step 5: Evaluate the errors of the selected model. Conventional regression diagnostic methods are applied to evaluate the errors of the selected model. If no problems are identified, move on to the next step. If it appears that the errors do not reasonably approximate the assumptions of the regression model, remedial action should be considered (see Chapter 5).

For example, if autocorrelation among the errors is identified, the *TSDB* routine should be considered; this routine is especially recommended when the number of observations is small ($N < 50$). *TSDB* requires entry of the single compound matrix **X:Y**, which is the combined complete design matrix **X** (including the unity column) plus the dependent variable vector **Y**. The *SPSS AUTOREG* routine is an alternative approach for fitting a regression model with autoregressive errors when the number of observations in the total series is large.

Regardless of the method required to remediate the problems in the original analysis, the model should be re-estimated. The parameter estimates from the revised analysis are used in the next step.

Step 6: Compute overall effect estimates and tests. A researcher experienced in the visual analysis of reversal designs focuses primarily on the overall pattern of change rather than each individual change. Correspondingly, each quantitative measure of change in the sequence of phase changes provides additional evidence regarding the effect. In most cases it is appropriate to cumulate the quantitative evidence from all phase changes; that is the purpose of the methods presented next.

Overall Level Change

Each level-change coefficient in the intervention model is associated with a test statistic. A quantitative measure of the overall level change is defined below, along with an approximate test of significance for this statistic.

$$\text{Overall level change statistic:} \frac{\sum_{c=1}^{C} \frac{1}{\hat{\sigma}_c^2} (PS_c)(OS_c)\left|b_{\text{LC}_c}\right|}{\sum_{c=1}^{C} \frac{1}{\hat{\sigma}_c^2}} = \text{LC}_{\text{Overall}},$$

where

PS_c is the sign predicted for change c;
OS_c is the sign observed for level-change coefficient b_{LC_c};
$\hat{\sigma}_c^2$ is the error variance estimate for level change c (i.e., the square of the standard error of coefficient b_{LC_c} provided in regression output); and
C is the number of level-change coefficients [i.e., $C = (J - 1)$, where J is the number of phases in the design].

$$\text{Test statistic for overall level change:} \frac{\sum_{c=1}^{C} (PS_c)(OS_c)\,|z_c|}{\sqrt{C}} = z_{\text{Overall LC}},$$

where $z_c = z$-statistic associated with observed level-change coefficient b_{LC_c}.

Overall Slope Change

The corresponding statistics for overall slope change are

$$\text{Overall slope change statistic:} \frac{\sum_{c=1}^{C} \frac{1}{\hat{\sigma}^2_{\text{SC}_c}} (PS_c)\,(OS_c)\,\left|b_{\text{SC}_c}\right|}{\sum_{c=1}^{C} \frac{1}{\hat{\sigma}^2_{\text{SC}_c}}} = \text{SC}_{\text{Overall}} \quad \text{and}$$

$$\text{Test statistic for overall slope change:} \frac{\sum_{c=1}^{C} (PS_c)(OS_c)\,\left|z_{\text{SC}_c}\right|}{\sqrt{C}} = z_{\text{Overall SC}},$$

where

PS_c is the sign predicted for change c;
OS_c is the sign observed for slope-change coefficient b_{SC_c};
z_{SC_c} is the z-statistic associated with observed slope-change coefficient b_{SC_c}; and
C is the number of slope-change coefficients.

The coefficients and tests associated with the model selected earlier (in Steps 4 and 5) provide the ingredients for the overall statistics. Note that the expressions for the overall test statistics include individual z-statistics. Conventional regression routines do not provide these values; instead, they provide t-values and associated p-values. The p-values on the individual coefficients, however, can be transformed to z-statistics using the Stouffer method. This is accomplished by (1) dividing the p-value by two (to obtain the one-tailed p-value), and (2) finding the z-value that is associated with the one-tailed p-value.

For example, suppose the t-value associated with a level-change coefficient is 2.72 and the associated p-value is .008; the one-tailed p is .004. The corresponding value of z can be obtained from a table of the normal distribution, or by using a computer routine. For example, if *Minitab* is used, one way to compute z is to follow this path:

```
Menu Bar → Calc → Probability Distributions →
Normal → Inverse Cumulative Probability → Input
Constant → (Enter .004) → OK
```

The following output appears:

```
Inverse Cumulative Distribution Function
Normal with mean = 0 and standard deviation = 1
P( X < = x )      x
      0.004      -2.65207
```

The value −2.65207 is the z-score below which the proportion of the normal distribution is .004. It is important to assign the same sign to z as is associated with

t. So in this example, the *t*-value of 2.72 is transformed to a *z*-value of 2.65. This transformation allows the expression for the overall test to be very simple.

A problem sometimes occurs in implementing this procedure; the *p*-value reported in software output for one (or more) of the partial regression coefficients may be zero throughout the three places that are usually provided. Although this is not a problem in conventional applications, it is a problem when the *p*-value is needed for determination of *z*. The solution depends upon the software. *SPSS* provides the *p*-value throughout many places when the normally reported *p*-value is double clicked. *Minitab* provides *p*-values for many places by using the relevant probability distribution function. For example, if the *t*-value is 2.72 and the associated degrees of freedom = 84, the *Minitab* path is as follows:

```
Menu Bar → Calc → Probability Distributions → t →
Cumulative Probability → Noncentrality parameter: 0.0 →
Degrees of Freedom 84 → Input Constant: -2.72 → OK
```

The output is as follows:

```
Cumulative Distribution Function
Student's t distribution with 84 DF
  x        P( X < = x )
-2.72       0.0039660
```

Note that the one-tailed *p*-value reported here goes out seven places; the *p*-value that accompanies the *Minitab* regression output goes out only three. Also, note in the input that the *t*-value entered as the "input constant" has a negative sign even though the obtained value of *t* is positive. If the positive value had been entered, the output would have shown the area below 2.72 and it would have been necessary to subtract that value from 1.0 in order to get the area above 2.72, which is what we want. One step is saved when the negative value of *t* is entered because, as is shown above, the output provides the area below it, which is equal to the one-tailed *p*-value we want.

Standardized Effect Sizes

It was pointed out in Chapter 19 that the choice of an adequate standardized effect size requires a clear distinction between different ways of conceptualizing the unit of analysis. Recall that a "single-case" design may involve either a single organism (e.g., a person) or a single group (or some other compound unit) that provides a single measure at each time point (e.g., annual death rates for a city) rather than outcome measures for each individual in the compound unit. This issue is also relevant to reversal designs and for the same reasons. That is, the distinction between the two types of case has implications for the generality of the results. When the case is a single subject, the results generalize to the time-series process generated by that subject. When the case is a group, the results generalize to a group time-series process; if the group is selected to be representative of a specified population then the results may be

generalized to that population. Regardless of the unit of analysis, the interpretation should make it clear that the effects were estimated from a unit that was treated in a specific sequence and that different results may be found if the sequence is changed.

The standardized effect size for each pair of adjacent phases in a reversal design may be computed using

$$\frac{b_{\mathrm{LC}_c}}{\sqrt{\mathrm{MS}_{\mathrm{Res}}}} = g_c,$$

where b_{LC_c} is the level-change coefficient for phase comparison c and $\mathrm{MS}_{\mathrm{Res}}$ is the mean square residual for the model that provides the LC coefficients. Because the data are collected in a specific order, each effect size should be identified by order.

The same general approach may be applied using the overall level-change statistic. The standardized overall level-change statistic is computed as follows:

$$\frac{\mathrm{LC}_{\mathrm{Overall}}}{\sqrt{\mathrm{MS}_{\mathrm{Res}}}} = g_{\mathrm{Overall}}.$$

Change in Percentile Rank (CPR)

A modification of the *CPR* statistic for the AB single-subject design presented in Chapter 19 is pertinent to reversal designs. Recall that the mean and standard deviation of a relevant norm group must be available in order to compute *CPR*. The first step involves computing the norm group percentile rank that corresponds to the single-subject's baseline level (estimated using b_0 in the case of the reduced model, and $b_0 + b_1$ $(n_1 + 1)$ in the case of the full model). Next, compute the percentile rank associated with (baseline level + $\mathrm{LC}_{\mathrm{Overall}}$) if the overall intervention effect is to increase behavior. If the effect of the treatment is to decrease behavior, compute the percentile rank associated with (baseline level − $\mathrm{LC}_{\mathrm{Overall}}$). The difference between the two percentile ranks describes the intervention effect in terms of the shift in normative standing.

Measure of Association: R^2

The R^2 statistic was recommended in Chapter 19 as a measure of the proportion of the variation on the dependent variable that is explained by the independent variable in a two-phase design. The same statistic is relevant in the case of experiments based on multiple-phase reversal designs. The only difference is the number of predictors in the model.

20.3 COMPUTATIONAL EXAMPLE: PHARMACY WAIT TIME

Slowiak et al. (2008) studied the effects of two interventions on the amount of time that customers wait for prescriptions to be filled in a hospital pharmacy. Customer satisfaction with wait time was also measured. These measures were obtained throughout

Table 20.2 Signs Associated with Predicted Direction of Change for Adjacent Phases

Comparison Phases	Sign for Predicted Direction of Change
A, B	−
B, C	−
C, B′	+
B′, A	+

each day of the 89 work-day experiment. A five-phase interrupted time-series reversal design was used; the phases and conditions were described as follows: A = baseline, B = feedback, C = feedback plus goal setting, B′ = feedback, and A′ = follow-up baseline. Note that this design differs from the previously described reversal designs in that there is more than one type of intervention and there are five phases.

Step 1: Predict the change pattern. Theory and previous research regarding feedback and goal-setting treatment conditions allowed four predictions to be made regarding the direction of change from phase to phase. First, the researchers predicted that the feedback condition (B) would decrease wait time relative to the baseline condition (A); therefore, the sign of the first predicted level-change coefficient (shown in Table 20.2) is negative. Second, the feedback plus goal setting condition (C) was predicted to decrease wait time to a level below that associated with feedback alone (B), so the second predicted sign is negative. Third, it was predicted that the change from condition C to the second feedback condition (B′) would increase wait time, so the third predicted sign is positive. Fourth, the change from condition B′ to the second baseline condition (A′) was predicted to increase wait time, so the fourth predicted sign is positive. These phase comparisons and signs are required in the last step of the analysis.

Step 2: Fit the full intervention model. Because a multiple-phase experiment may require a measure of level and slope to describe the responses in each phase, the number of parameters in the full model is twice the number of phases. Hence, 10 parameters are required in this five-phase design. This means that nine predictor variable columns are required (in addition to the unity column that is automatically added by OLS software to estimate the intercept). The form of the columns illustrated in Table 20.1 was followed in specifying the predictors. The *Minitab* commands used to construct these columns are described next.

The sample sizes associated with phases one through five are $n_1 = 24$, $n_2 = 15$, $n_3 = 29$, $n_4 = 15$, and $n_5 = 6$. The *Minitab* command line editor was opened and the following lines were entered:

```
MTB > set c1
DATA> 1:89
DATA> set c2
DATA> 24(0) 65(1)
DATA> set c3
```

```
DATA> 25(0) 1:64
DATA> set c4
DATA> 39(0) 50(1)
DATA> set c5
DATA> 40(0) 1:49
DATA> set c6
DATA> 68(0) 21(1)
DATA> set c7
DATA> 69(0) 1:20
DATA> set c8
DATA> 83(0) 6(1)
DATA> set c9
DATA> 84(0) 1:5
DATA> end
```

It can be seen that the "set c1" command creates the time variable that contains the values 1–89, "set c2" creates the first level-change column with 24 zeros and 65 ones, "set c3" creates the first slope-change variable with 25 zeroes followed by the sequence 1–64, and so on. Each of the 10 columns is associated with one parameter estimate. The parameter estimates are defined as follows:

b_0 is the intercept for the first phase;
b_1 is the slope for the first phase;
b_2 is the first level change (phase1 vs. phase 2);
b_3 is the first slope change (phase1 vs. phase 2);
b_4 is the second level change (phase 2 vs. phase 3);
b_5 is the second slope change (phase 2 vs. phase 3);
b_6 is the third level change (phase 3 vs. phase 4);
b_7 is the third slope change (phase 3 vs. phase 4);
b_8 is the fourth level change (phase 4 vs. phase 5); and
b_9 is the fourth slope change (phase 4 vs. phase 5).

The meaning of the level-change and slope-change coefficients discussed in Chapter 19 for two-phase designs generalizes to the coefficients in this design and to other designs with more than two phases.

The dependent variable scores (wait time in minutes) were entered in column 10. The data are listed below ($N = 89$; read from left to right, with the first data point in each phase in bold):

18	22	18	18	19	18	20	29	28	20	29	25	15	14	31
21	22	22	17	27	17	19	17	18	**23**	23	20	19	19	15
20	17	15	21	17	17	19	13	20	**17**	21	16	15	14	23
18	18	18	15	24	18	19	19	22	17	17	12	19	19	19
16	20	17	15	18	20	14	16	**24**	24	23	20	20	27	18
19	21	25	23	15	14	19	22	**20**	17	27	22	22	27	

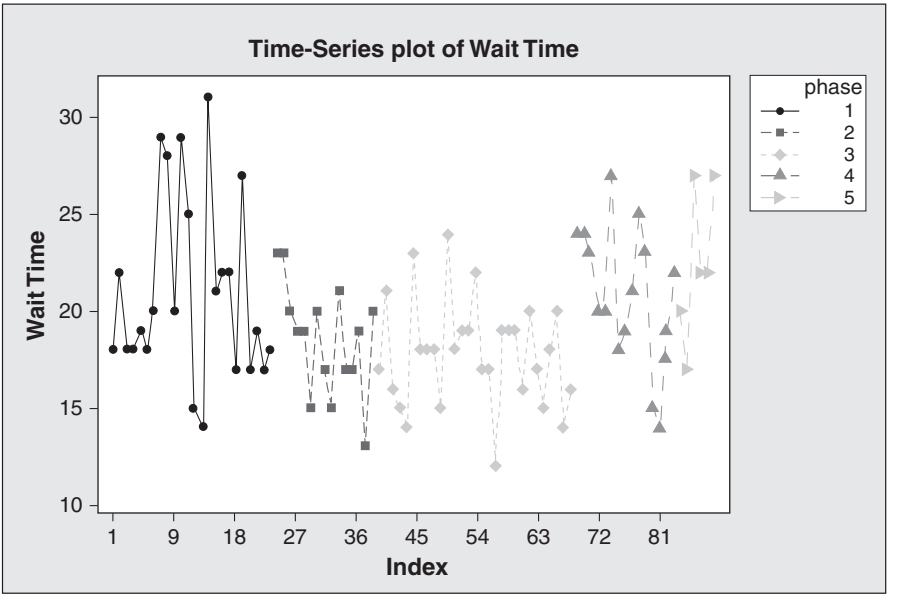


Figure 20.1 Pharmacy wait time data collected throughout five design phases.

These data are illustrated in Figure 20.1.

The dependent variable (wait time) was regressed on predictor variables c1–c9. The *Minitab* input commands for this regression and the related options of interest are shown below:

```
MTB > Regress 'Wait Time' 9 'Time'-'SC4';
SUBC>   Constant;
SUBC>   DW;
SUBC>   Brief 2.
```

Selected portions of the output associated with fitting the full model are shown below:

```
The regression equation is
Wait Time = 21.5 - 0.037 Time + 0.33 LC1 - 0.295 SC1
            + 2.50 LC2 + 0.291 SC2 + 6.29 LC3 - 0.319 SC3
            + 1.24 LC4 + 1.65 SC4

Predictor        Coef          SE Coef          T            P
Constant       21.467          1.502        14.29        0.000
Time           -0.0374         0.1051       -0.36        0.723
LC1             0.326          2.308         0.14        0.888
SC1            -0.2948         0.2375       -1.24        0.218
```

```
LC2              2.496        2.327          1.07       0.287
SC2              0.2908       0.2272         1.28       0.204
LC3              6.286        2.217          2.84       0.006
SC3             -0.3193       0.2272        -1.41       0.164
LC4              1.238        3.226          0.38       0.702
SC4              1.6464       0.8782         1.87       0.065

Analysis of Variance

Source            DF          SS         MS         F         P
Regression         9      337.17      37.46      2.95     0.005
Residual Error    79     1003.59      12.70
Total             88     1340.76

Durbin-Watson statistic = 2.15169
```

After the full model was estimated, a reduced model containing only the intercept and four level-change coefficients was estimated. Input and output for the reduced model is shown below:

Input for fitting the reduced model and related statistics:

```
MTB > Regress'Wait Time '4 'LC1' 'LC2' 'LC3' 'LC4;
SUBC>   Residuals 'RESI1';
SUBC>   Cookd 'COOK1';
SUBC>   GFourpack;
SUBC>   RType 1;
SUBC>   Constant;
SUBC>   DW;
SUBC>   Brief 1.
```

Relevant portions of the reduced model output:

```
Regression Analysis: Wait Time versus LC1, LC2, LC3, LC4

The regression equation is

Wait Time = 21.0 - 2.47 LC1 - 0.74 LC2 + 3.14 LC3
            + 1.57 LC4

Predictor      Coef  SE Coef        T      P
Constant    21.0000   0.7403    28.37  0.000
LC1         -2.467    1.194     -2.07  0.042
LC2         -0.740    1.153     -0.64  0.523
LC3          3.140    1.153      2.72  0.008
LC4          1.567    1.752      0.89  0.374

S = 3.62683   R-Sq = 17.6%   R-Sq(adj) = 13.7%
```

```
Analysis of Variance

Source           DF       SS     MS     F      P
Regression        4   235.84  58.96  4.48  0.002
Residual Error   84  1104.93  13.15
Total            88  1340.76

Durbin-Watson statistic = 1.99669

MTB > ACF 'RESI1'.

Autocorrelation Function: RESI1

Lag       ACF      T    LBQ
 1  -0.011582  -0.11   0.01
 2  -0.162318  -1.53   2.47
 3   0.012755   0.12   2.48
 4   0.012931   0.12   2.50
 5  -0.022827  -0.21   2.55
 6  -0.111621  -1.03   3.76
 7  -0.108287  -0.98   4.92
 8   0.028281   0.25   5.00
 9   0.106578   0.96   6.15
10  -0.077272  -0.69   6.76
11  -0.046702  -0.41   6.99
12  -0.103226  -0.91   8.11
13  -0.067263  -0.59   8.59
14  -0.017703  -0.15   8.63
15  -0.049116  -0.43   8.89
16   0.064785   0.56   9.36
17   0.065622   0.57   9.84
18  -0.024535  -0.21   9.91
19   0.120583   1.04  11.59
20   0.081795   0.70  12.38
21  -0.049236  -0.42  12.66
22  -0.160878  -1.36  15.79
```

The two models were then compared to determine whether the first phase slope and the slope-change parameters were needed to model the data. The statistics required for the model comparison test are shown below:

$SS_{Reg\ (Full)} = 337.17$	$df_{Reg\ (Full)} = 9$
$SS_{Reg\ (Reduced)} = 235.84$	$df_{Reg\ (Reduced)} = 4$
$SS_{Res\ (Full)} = 1003.59$	$df_{Res\ (Full)} = 79$

Test statistic:

$$\frac{[SS_{Reg(Full)} - SS_{Reg(Reduced)}]/(df_{Reg(Full)} - df_{Reg(Reduced)})}{MS_{Res(Full)}} = F \quad \text{and}$$

$$\frac{[337.17 - 235.84]/(9 - 4)}{12.70} = 1.60.$$

The critical value of F is based on numerator degrees of freedom $= (df_{\text{Reg(Full)}} - df_{\text{Reg(Reduced)}}) = 5$, and denominator degrees of freedom $= N - m - 1 = df_{\text{Resid(Full)}} = 89 - 9 - 1 = 79$. The critical value of $F = 2.33$ for $\alpha = .05$; therefore, it was concluded that the justification for including the slope-change parameters in the model is not strong. (The same conclusion is reached if α is set at .10.) Consequently, the simpler (reduced) model that contains only level-change coefficients and the intercept was tentatively adopted for these data. Diagnostic methods were then applied to evaluate the adequacy of the model.

Diagnostics

The residual plots shown in Figure 20.2 reveal no major departures from linearity and homoscedasticity, tests for autocorrelated errors reveal no hint of a dependency problem (note in the output that $DW = 2.0$ and $r_1 = -.01$); however, the residual distribution has a minor positive skew. Square root and log transformations were carried out to determine if they improve the analysis. Although both transformations resulted in less skew, they had virtually no effect on the inferential results and, if used, they complicate the interpretation of the results. Consequently, the reduced model based on the untransformed wait times was selected as the final model.

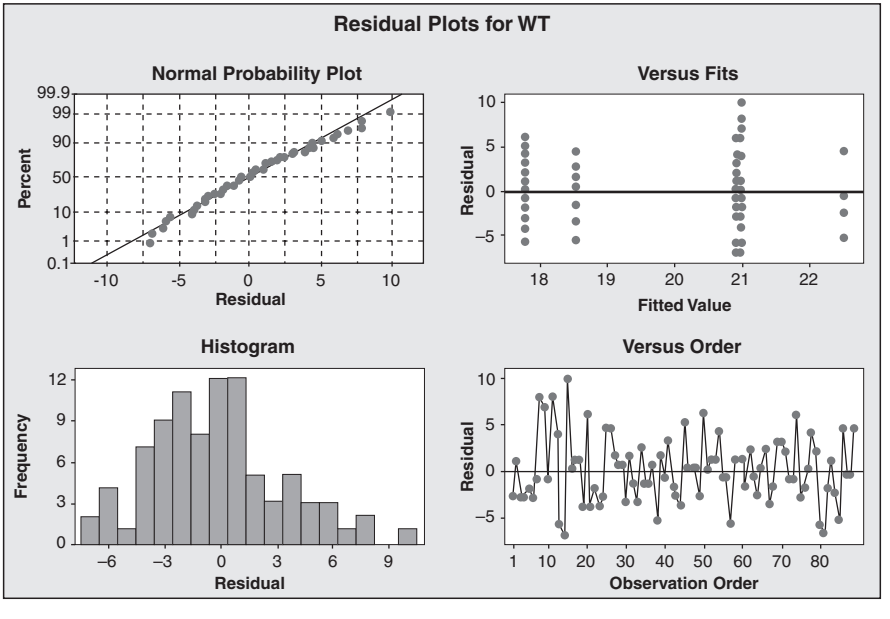


Figure 20.2 *Minitab* four-in-one residual plots for pharmacy wait time experiment.

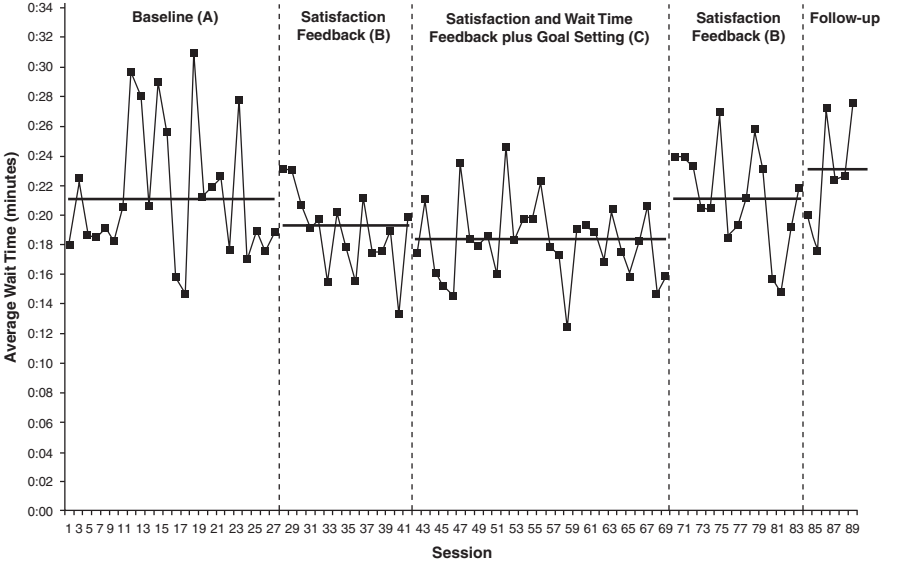


Figure 20.3 Wait time level for each phase estimated using the reduced model. (From Slowiak et al., 2008.)

Table 20.3 Summary Statistics Required for Computing $LC_{Overall}$ and $z_{Overall}$

Phases	b_{LC_c}	$\hat{\sigma}_c^2$	t_c	p_1	z_c	PS_c	OS_c	$(PS_c)(OS_c)\lvert z_c\rvert$
1,2	−2.47	1.43	−2.07	.021	−2.03	–	–	2.03
2,3	−.74	1.33	−.64	.262	−.64	–	–	.64
3,4	3.14	1.33	2.72	.004	2.65	+	+	2.65
4,5	1.57	3.07	.89	.187	.89	+	+	.89

$$\sum_{c=1}^{C} (PS_c)\,(OS_c)\,|z_c| = 6.21$$

The reduced model level estimates are illustrated using a horizontal line in each phase in Figure 20.3. It can be seen that each level change from one phase to the next is consistent with the direction predicted before the experiment was begun. The specific level-change values (i.e., the b_{LC_c}) are listed in Table 20.3.

Overall Level Change and Test for Overall Level Change

The overall level change statistic ($LC_{Overall}$) and the test statistic for overall level change ($z_{Overall}$) are shown below. The values required to compute these statistics are summarized in Table 20.2. Note that it contains the estimated level-change coefficients, error variances for these coefficients, associated t, z and one-tailed

p-values, predicted signs, observed signs, and products of predicted and observed signs times the absolute *z*-values.

Overall Level-Change Statistic

The computation of the overall level-change statistic requires the individual level-change coefficients, the associated variance estimates, and the signs for the observed and predicted effects summarized in Table 20.2. The final computation is shown below:

Phases	$\frac{1}{\hat{\sigma}_c^2}$	$\frac{1}{\hat{\sigma}_c^2}(PS_c)(OS_c)\lvert b_{\mathrm{LC}_c}\rvert$
1,2	.699	1.727
2,3	.752	.556
3,4	.752	2.361
4,5	.326	.512
	$\sum_{c=1}^{C}\frac{1}{\hat{\sigma}_c^2} = 2.529$	$\sum_{c=1}^{C}\frac{1}{\hat{\sigma}_c^2}(PS_c)(OS_c)\lvert b_{\mathrm{LC}_c}\rvert = 5.156$

$$\frac{\sum_{c=1}^{C}\frac{1}{\hat{\sigma}_c^2}(PS_c)(OS_c)\lvert b_{\mathrm{LC}_c}\rvert}{\sum_{c=1}^{C}\frac{1}{\hat{\sigma}_{\mathrm{LC}_c}^2}} = \frac{5.156}{2.529} = 2.03 = \mathrm{LC}_{\mathrm{Overall}}$$

This measure indicates that the overall effect of introducing or withdrawing the conditions is to change the absolute wait time by about 2 min.

A positive value of the $\mathrm{LC}_{\mathrm{Overall}}$ statistic means that the overall change is consistent with the direction predicted. But this does not mean that each individual change is an increase, because both increases and decreases are predicted for an effective intervention, depending on the phase comparison. A negative value of $\mathrm{LC}_{\mathrm{Overall}}$ occurs only when the predicted and observed signs do not agree. But it is possible to have a positive value of $\mathrm{LC}_{\mathrm{Overall}}$ even when the predicted and observed signs do not have perfect agreement; this occurs when the size of the change coefficients associated with agreeing signs is larger than the change coefficients associated with the disagreeing signs.

Test for Overall Level Change

The test for overall level change requires the individual level-change estimates, the associated one-tailed *p*-values, and the signs associated with the set of predicted and

observed level changes; these are summarized in Table 20.2.

$$\frac{\sum_{c=1}^{C}(PS_c)(OS_c)\,|z_c|}{\sqrt{C}} = \frac{6.21}{\sqrt{4}} = 3.105 = z_{\text{Overall LC}}$$

The one-tailed p-value (in this case the area above $z = 3.105$) is $< .001$. Hence, it is concluded that the cumulative evidence for level change associated with the interventions in this study is very strong.

The observed direction of level change associated with each of the four condition changes is completely consistent with the pattern of predicted level change made before the experiment was carried out. It can be seen in Table 20.2 that the correlation between the PS column and the *OS* column is perfect. But, the test statistic is based on more than the agreement between these two columns. That is, the test is a function of both the agreement between *PS* and *OS* and the amount of evidence associated with each level change.

Optional Change Measures and Tests for Each Type of Intervention

Unlike the conventional ABAB reversal design that includes only two types of condition (either baseline or intervention), the example design (ABCB′A′) includes three conditions (baseline, intervention B, and intervention C). Because there are two types of intervention in this example, it is relevant to present separately the average change that involves the A and B phases and the average change that involves the C and B phases. An inspection of the individual level-change coefficients that contribute to the overall level-change statistic indicates that the average absolute change from condition A (baseline) to condition B (feedback) and B to A is about 2 min [i.e., $(2.47 + 1.57)/2 = 2.02$]; similarly, the average absolute change from B to C (feedback plus goal setting) and C to B is about 2 min [i.e., $(.74 + 3.14)/2 = 1.94$]. Label these averages (2.02 and 1.94) as $\text{LC}_{\text{Feedback}}$ and $\text{LC}_{\text{Feedback + Goal Setting}}$, respectively.

The separate tests for the two interventions follow the same general form as the test for overall level change, but the first one is limited to the AB and B′A′ phase changes and the second one is limited to the BC and CB′ phase changes. The required values required for the computations are shown in Table 20.2.

Separate Test for Feedback Effects

The first and fourth rows in the table are associated with the AB and B′A′ phase changes. The test statistic z is the sum of the $(PS_c)(OS_c)|z_c|$ values found in the last column of the table for rows one and four, divided by the square root of 2.

$$z = \frac{2.03 + .89}{\sqrt{2}} = 2.06$$

The p_1-value associated with the obtained z is .02.

Separate Test for Feedback Plus Goal Setting Effects

The second and third rows in the table are associated with the BC and CB′ phase changes. The test statistic z is the sum of the $(PS_c)(OS_c)|z_c|$ values found in the last column of the table for rows two and three, divided by the square root of 2.

$$z = \frac{.64 + 2.65}{\sqrt{2}} = 2.33$$

The p_1-value associated with this obtained z is .01. These two separate tests can be viewed as a way of partitioning the information regarding the overall level-change effects into that which is supplied by each separate type of intervention. Because the overall test cumulates evidence from all phase changes in this example, it has a larger test statistic than either one of the separate tests.

Standardized Level Changes

The standardized overall level change is

$$\frac{\mathrm{LC}_{\mathrm{Overall}}}{\sqrt{\mathrm{MS}_{\mathrm{Res}}}} = 2.03/3.627 = .56,$$

and the separate standardized level changes for feedback and feedback plus goal setting interventions are

$$\frac{\mathrm{LC}_{\mathrm{Feedback}}}{\sqrt{\mathrm{MS}_{\mathrm{Res}}}} = 2.02/3.627 = .56 \text{ and}$$

$$\frac{\mathrm{LC}_{\mathrm{Feedback\ +\ Goal\ Setting}}}{\sqrt{\mathrm{MS}_{\mathrm{Res}}}} = 1.94/3.627 = .53.$$

The mean square residual used in the computation of these three standardized level changes is based on the reduced model. Had this been a conventional reversal design (e.g., ABAB or ABABA), only the overall effect size would have been computed.

R^2

The obtained value of R^2 from the reduced model is .176; this is the proportion of the total variation in wait time that is explained by the interventions. This R^2 is classified as a medium effect size using conventional criteria for measures of this type, however, a more reasoned approach for evaluating the importance of the result described next.

Correlation of Wait Time with Satisfaction

Although the focus of the study was on wait time as the dependent variable, data on customer satisfaction with wait time were also collected. The analytic approach

described for wait time also applies to the satisfaction data. It turns out that the reduced model is a good fit for the satisfaction data as well. The overall test on satisfaction is statistically significant ($p < .01$) and $R^2 = .19$. More importantly, the correlation between wait time and satisfaction is $-.57$ ($p < .001$). As wait time goes down, satisfaction goes up. The size of the relationship between these two variables is rather striking when one considers the moderate size of the change in wait time. It appears that about a third (i.e., $.57^2 = .32$) of the variation on satisfaction is explained by wait time. A reasonable conclusion is that the 2-min reduction in wait time produced by the interventions is important to customers.

20.4 SUMMARY

The conventional reversal design has three or more phases during which the intervention condition is introduced and withdrawn. Each additional phase provides additional information regarding the effects of the intervention. It is appropriate that the analysis acknowledge the cumulative evidence in overall descriptive measures and associated test statistics.

The first step in the analysis requires the researcher to thoughtfully consider the change pattern that is expected to accompany introductions and withdrawals of the intervention. This step is performed before data are collected. The second step involves fitting a regression model that includes level-change and slope-change parameters. Third, a reduced model that includes only level-change parameters is fitted. Fourth, a model comparison test is performed to identify the better of the two models. Fifth, the residuals of the selected model are diagnosed for departures from the assumptions of the regression model. When serious violations of linearity, homoscedasticity, or normality are encountered at step five, conventional remedial procedures are applied and the model is refitted. If autocorrelated errors are identified in the initially selected model, the data are reanalyzed using a regression model that accommodates autoregressive errors. The results of the remedial analysis are then used in step six.

Step six involves the use of the parameter estimates and the associated inferential measures from the selected model to construct cumulative descriptive and inferential measures of the overall effects of the intervention. The cumulative measures include overall measures of level change and slope change, test statistics for overall level change and slope change, standardized effect sizes, and a measure of association between the independent and dependent variables. Additional measures of overall effects are included when the design has more than one type of intervention.

CHAPTER 21

Analysis of Multiple-Baseline Designs

21.1 INTRODUCTION

Multiple baseline designs are widely recognized in many areas of research (especially applied behavior analysis) as easily implemented, highly sensitive, and internally valid. Many areas of research in which randomized-group designs and reversal single-case designs are disqualified by practical or ethical considerations are easily investigated using at least one of the variants of the multiple-baseline design.

Variants of the Multiple-Baseline Design

There are essentially three major variants of this design. They may be labeled as: (1) multiple-baseline across subjects, (2) multiple-baseline across settings, and (3) multiple-baseline across dependent variables (or *behaviors*). A brief overview of these variants is presented in the remainder of this section. More thorough descriptions are available in Barlow et al. (2009), Cooper et al. (2006), and Johnston and Pennypacker (2008). The distinctions among these variants have implications for the choice of statistical analysis. Methods of analysis for the variants are presented in subsequent sections.

Multiple-Baseline Across Subjects

The multiple-baseline design across subjects is not really a "single-case" design because it contains at least two subjects or units. Hence, it is more appropriately labeled as a "very small sample" design. This variant is essentially a collection of AB designs, but there is a unique feature regarding the timing of condition changes that leads to high internal validity. Instead of introducing the intervention to all subjects at the same time, the intervention is introduced at a different time to each subject according to a planned staggered sequence. The staggered initiation of the intervention makes it implausible that that an event unrelated to the intervention is the

The Analysis of Covariance and Alternatives: Statistical Methods for Experiments, Quasi-Experiments, and Single-Case Studies, Second Edition. Bradley E. Huitema.

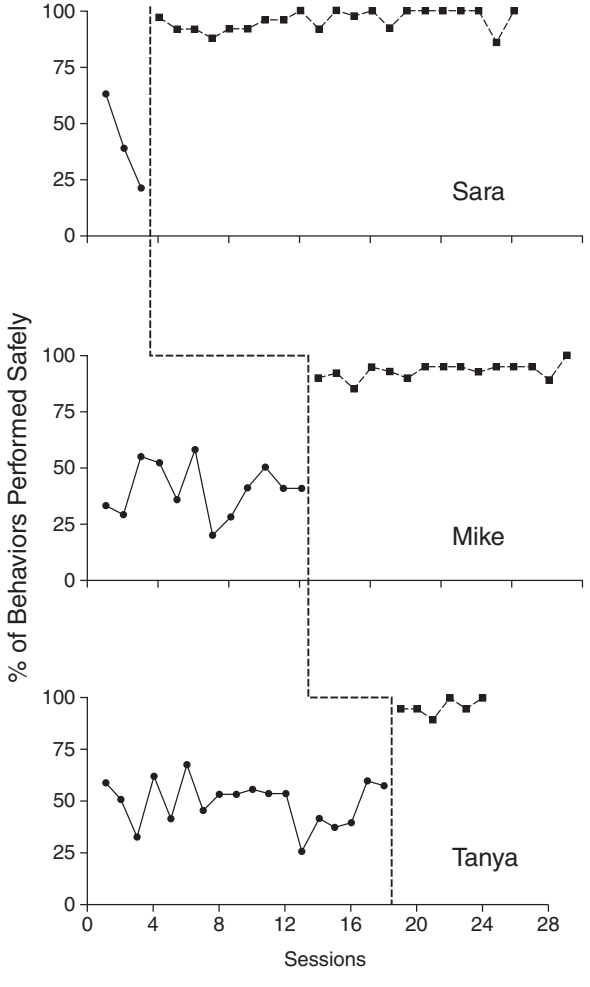


Figure 21.1 Example of multiple-baseline design across subjects. (*Data source*: Scherrer and Wilder, 2008)

cause of the apparent effect on each subject. An example from Scherrer and Wilder (2008) is illustrated in Figure 21.1.

Here the subjects were three cocktail servers and the intervention was training to reduce the risk of developing musculoskeletal disorders from tray carrying. Data regarding tray carrying were obtained by direct observation using trained observers (with high interrater reliability). Subjects were unaware that their performance was being observed, and the observers did not know when each participant received training. The subjects were instructed to not share information regarding training, and they were isolated from each other in the sense that they worked on different days. Note in the figure that the intervention was never introduced to more than one subject at a time and that the number of baseline observations is different for each subject. The internal validity threat of history is quite implausible in this experiment.

Multiple-Baseline Across Settings

The second variant is the multiple-baseline design across settings. It involves a single subject who is measured under baseline and intervention conditions in several different settings. The dependent variable is the same in all settings. For example, consider a study that evaluated the effects of an intervention on the frequency of a child's inappropriate behavior. The child was observed in three different settings (home, school, and bus) each day of the study and the method of counting instances of specific inappropriate behaviors was the same in all three settings.

Multiple-Baseline Across Dependent Variables (or Behaviors)

The third design variant is the multiple-baseline design across dependent variables (or behaviors). It involves the situation in which data are collected on two or more response measures from one subject (or other unit) in one setting. Designs like this are common in applied behavioral studies where data on several different aspects of behavior (such as rate of responding, accuracy, and latency) or different behaviors (such as safe lifting, wearing safety glasses, and wearing helmets in a workplace) are recorded during each observation session.

Design Variant Implications for Analysis

The analysis of the multiple-baseline design is more complex than is the analysis of reversal designs because two types of possible dependency need to be acknowledged. The first type of dependency is within-series autocorrelation of errors (previously discussed in the context of simple AB and reversal designs); this type of dependency is possible with all variants of the multiple-baseline design. The second type of dependency is correlation between errors of the multiple series; this type of correlation (sometimes called cross-correlation) is likely in some multiple-baseline studies but not in others.

Three analytic approaches are described in subsequent sections. They differ with respect to the presumed error structure within and between the multiple series. *Case I* presumes independence of errors both within series and between series. *Case II* presumes dependency of errors within series but independence of errors between series. *Case III* presumes independent errors within series and dependent errors between series. There is also a case where errors are dependent both within and between series, but the analysis of this case is not pursued here; it is currently under development.

21.2 CASE I ANALYSIS: INDEPENDENCE OF ERRORS WITHIN AND BETWEEN SERIES

I mentioned earlier that the multiple-baseline design consists of a small collection of AB designs that have been structured to have the intervention introduced at different time points. A common multiple baseline across subjects design has 2 to 4 subjects (or some other unit). It is typical that these subjects are essentially isolated from each other (as in the example illustrated in Figure 21.1). When the subjects are independent,

it is common to have neither between-series correlation nor autocorrelation within series; this situation describes Case I.

Several statistics are useful in evaluating the outcome of the multiple baseline design. I recommend the following: (1) a measure describing overall level change (across the different series), (2) a test for overall level change, and (3) measures of overall effect size. These measures are described in the remainder of this section for Case I.

LC_{Overall}

The following three steps are involved in computing the overall level change statistic LC_{Overall}:

Step 1. Apply the regression methodology described in Chapter 18 (for the AB design) to the data based on each unit.

Step 2. List the parameter estimates and associated inferential results from the separate regression analyses in the tabular form shown in Table 21.1. The results shown in this table are based on the application of OLS regression to contrived academic outcome data presented in Koehler and Levin (2000). The units were three classrooms and the intervention was designed to improve performance on weekly academic assessments. Each classroom was assessed weekly for 10 weeks (see Figure 21.2).

The values summarized in Table 21.1 are used in computing both the weighted overall level change measure and the test on this statistic.

Table 21.1 Summary of Values Required for Computation of the Weighted Overall Level Change

Unit	b_{LC_j}	$\hat{\sigma}_j$	t_j	p_1	z_j	$\hat{\sigma}_j^2$	$\frac{1}{\hat{\sigma}_j^2}$
1	0.600	0.5788	1.037	.170	0.95	0.335	2.98
2	2.214	0.7083	3.126	.007	2.45	0.502	1.99
3	2.500	1.1487	2.176	.031	1.87	1.319	0.76
					$\sum_1^J z_j = 5.28$	$\sum_{j=1}^J \left(\frac{1}{\hat{\sigma}_j^2}\right) = 5.73$	

where

J is the number of units (the number of classrooms in this example);
b_{LC_j} is the level change coefficient estimated for the jth unit;
$\hat{\sigma}_j$ is the estimated standard error for the jth level change coefficient;
t_j is the t-value associated with the jth level change coefficient (estimated for the jth unit);
p_1 is the one-tailed p-value associated with t_j;
z_j is the normal deviate associated with p_1 for unit j;
$\hat{\sigma}_j^2$ is the variance estimate associated with the jth LC coefficient; and
$\frac{1}{\hat{\sigma}_j^2}$ is the reciprocal of $\hat{\sigma}_j^2$.

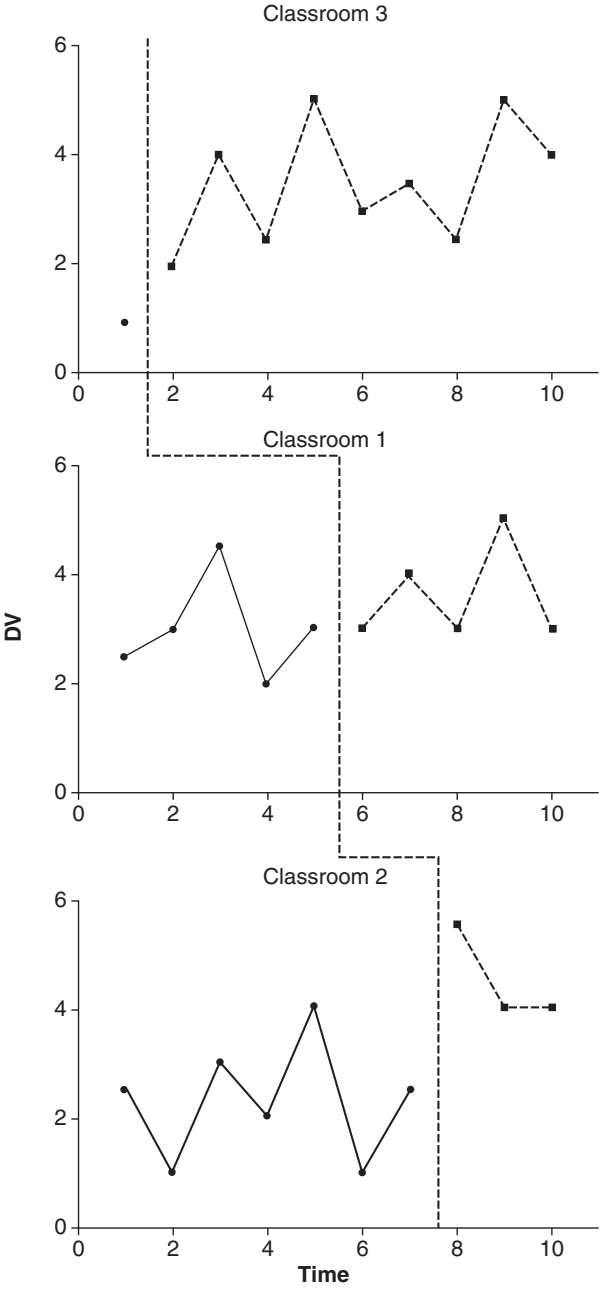


Figure 21.2 Example of multiple-baseline data from Koehler and Levin (2000).

Step 3. Compute the weighted overall level change measure, the test on the overall level change measure, the overall standardized effect size, and, if normative data are available, the CPR measure.

$$\frac{\sum_{j=1}^{J} \frac{1}{\hat{\sigma}_j^2} b_{\mathrm{LC}_j}}{\sum_{j=1}^{J} \frac{1}{\hat{\sigma}_j^2}} = LC_{\mathrm{Overall}}.$$

The LC_{overall} value for the example data is:

$$\frac{2.98\,(.60) + 1.99\,(2.214) + .76\,(2.50)}{5.73} = \frac{8.09}{5.73} = 1.41.$$

It can be seen that LC_{Overall} is simply a weighted average of the J level-change coefficients, where the weights are the reciprocals of the error variance estimates for the individual level change coefficients. The justification for these weights rests on the notion that an *LC* coefficient with small variance is more precise than an *LC* coefficient with large variance; therefore, coefficients with high variance should have lower weight than those with low variance. In the example, the first *LC* coefficient is weighted more heavily than are the second and third *LC* coefficients because the variance estimate associated with the first coefficient is the smallest of the three.

LC_{Overall} Test Statistic

Compute the overall level change test statistic using: $\frac{\sum_{j=1}^{J} z_j}{\sqrt{J}} = z_{\mathrm{Overall}}$. The example data yield the following value for this test statistic: $\frac{5.28}{\sqrt{3}} = 3.05$. The z_{Overall}-statistic is distributed approximately as a standard normal deviate; the associated one-tailed p-value is approximately .001.

Overall Standardized Effect Size

Two very different types of effect size (other than the unstandardized overall level change) are relevant in a multiple-baseline experiment. The first type involves standardization based on within-unit variation. This standardization is carried out using:

$$\frac{LC_{\mathrm{Overall}}}{\sqrt{\frac{\sum_{j=1}^{J} \mathrm{SS}_{\mathrm{Residual}_j}}{\sum_{j=1}^{J} \left(N_j - P\right)}}} = \text{Overall standardized effect size}$$

where

$SS_{Residual_j}$ is the sum of squares residual from fitting the two phase model to unit j;

$\sum_{j=1}^{J} SS_{Residual_j}$ is the pooled within unit residual sum of squares;

N_j is the number of observations in the series associated with unit j;

P is the number of parameters in the intervention model applied to unit j;

$(N_j - P)$ is the residual degrees of freedom for unit j; and

$\sum_{j=1}^{J} (N_j - P)$ is the pooled within unit residual degrees of freedom.

Because each unit (i.e., classroom) provided 10 assessments (one each week) the common $N_j = 10$. The residual SS for the example data are 6.70, 8.43, and 9.50 for units one through three, respectively, so the pooled residual SS = 24.63. The pooled within-unit residual degrees of freedom is 24, because the data from each individual unit has $N_j - 2$ degrees of freedom.

$$\frac{LC_{\text{Overall}}}{\sqrt{\dfrac{\sum_{j=1}^{J} SS_{Residual_j}}{\sum_{j=1}^{J} (N_j - P)}}} = \frac{1.41}{\sqrt{\dfrac{24.63}{24}}} = 1.41/1.013 = 1.39.$$

This statistic describes the overall level change that has been standardized by the within-unit standard deviation, thus yielding a standardized effect-size measure.

CPR

An additional type of overall level-change measure is the change in percentile rank (*CPR*). This measure describes the change between the percentile rank associated with the overall baseline level and the percentile rank associated with the overall postintervention level. The computation requires an estimate of the overall intercept (i.e., the weighted overall baseline level), the overall level change, and the mean and standard deviation of a norm group.

Suppose the summary statistics in Table 21.2 are from three patients sampled from a population of cardiac cases who were told to get at least 30 min of aerobic exercise daily (the standard recommendation for all cardiac patients in this population). Then a 10-week multiple baseline experiment was carried out to determine whether compliance with the recommended exercise treatment could be improved (e.g., using incentives). The baseline behavior recorded weekly for each patient was the number of days per week that the patient got at least 30 min of aerobic exercise. An intervention was introduced to patients 1, 2, and 3, after baseline data were collected over a different interval for each patient. A separate regression model of the type described in Chapter 18 was fit to the data from each patient, and the intercepts were recorded.

Table 21.2 Summary of Values Required for Computation of the Weighted Overall Baseline Level

Patient	b_o	$\hat{\sigma}^2_{b_o}$	$\frac{1}{\hat{\sigma}^2_{b_o}}$	$\frac{1}{\hat{\sigma}^2_{b_o}} b_o$
1	3.00	0.17	5.95	17.85
2	2.29	0.15	6.64	15.21
3	1.00	1.19	0.84	0.84
			$\sum \frac{1}{\hat{\sigma}^2_{b_0}} = 13.43$	$\sum \frac{1}{\hat{\sigma}^2_{b_0}} b_o = 33.90$

Because each intercept measures the baseline level for an individual patient there is interest in describing the overall level of these baselines before the intervention was introduced. The weighted overall baseline level is computed using:

$$\frac{\sum \frac{1}{\hat{\sigma}^2_{b_o}} b_o}{\sum \frac{1}{\hat{\sigma}^2_{b_0}}} = \bar{b}_o = \text{Weighted overall baseline level.}$$

The estimated intercepts, error variances, reciprocals of the error variance estimates (i.e., the weights), and the products of the weights times the intercepts are summarized in Table 21.2 for the three patients.

The weighted overall baseline level $\bar{b}_o = 33.90/13.43 = 2.52$. Suppose that the overall level change $= 1.41$; it can be seen that the intervention changed the overall level from 2.52 during baseline to 3.93 (1.41 + 2.52) during intervention. Hence, the overall compliance was about two and a half days per week before the intervention and almost four days per week after the intervention was introduced. Although this change is likely to be viewed as being of clinical importance, a more formal quantitative description of the meaning of the overall level change may also be of interest.

Suppose normative data are available (or can be collected) regarding the distribution of compliance (i.e., the number of days per week that cardiac patients perform the recommended exercise in the absence of special interventions). Let us say the distribution is approximately normal and has a mean and standard deviation of 2.75 and 0.75, respectively. This means that the overall baseline level obtained in the experiment falls at percentile rank 38 ($z = -0.31$) and the overall intervention level falls at percentile rank 94 ($z = 1.57$). Hence, the change in percentile rank (CPR) $= 56$.

When Slope Change Is Present

Although intervention studies that do not require a slope-change parameter are typical, it is not particularly unusual to encounter data that do. That is, model I (described

in Chapter 18) is sometimes required to model the data from one or more individual subjects in a multiple-baseline design. When this occurs, an additional summary table is completed that is of the same general form as Table 21.1, but slope-change coefficients and the associated statistics are substituted for the level-change statistics in the table. The method of cumulating the evidence on slope change is parallel to the method shown for level change.

21.3 CASE II ANALYSIS: AUTOCORRELATED ERRORS WITHIN SERIES, INDEPENDENCE BETWEEN SERIES

The methods described in the previous section apply to the case of independent errors both within and between units. Occasionally the errors within units may be autocorrelated with any of the three variants of the design. If this occurs a minor modification of the Case I analysis is called for. The Case II analysis requires that the parameter estimates and *p*-values associated with each individual series be based on the *TSDB* results rather than the OLS results illustrated above for Case I. Hence, the Case II analysis is carried out as follows:

Step 1. Apply the methodology described in Chapter 18 (for the AB design) to the data from each unit.

Step 2. Enter the parameter estimates and associated inferential results from these analyses (each based the *TSDB* double bootstrap routine) using the form shown in Table 21.1. If some units have independent errors, use OLS results in the summary table for those units and use results from *TSDB* for the autocorrelated units.

Step 3. Compute the weighted overall level change measure, the test on the overall level change measure, the overall standardized effect size, and, if normative data are available, the CPR measures, as described in Section 21.2.

21.4 CASE III ANALYSIS: INDEPENDENT ERRORS WITHIN SERIES, CROSS-CORRELATION BETWEEN SERIES

Correlation between error series is possible with all design variants, but it is most likely to be found using the multiple-baseline design across dependent variables (or behaviors). This is especially likely when seemingly different dependent variables actually measure a similar construct. The analytic problem introduced by between-series correlation is that each outcome series provides less unique information than is provided when the series are independent. This is a concern because the overall test statistic that applies to the case of independent error series is inflated when the series are positively correlated. The inflated test statistic results in type I error that exceeds the nominal value (e.g., .05). Consequently, the analysis recommended in this section incorporates information regarding dependency between the errors of the various series included in the design.

Both the within-series error structure and the between-series error structure can be described using appropriate measures of correlation. Autocorrelations are computed at various lags to describe the within-series structure and cross-series correlations are computed at various lags to describe the between-series structure. Although it is possible for these structures to be quite complex, this is usually not true. Autocorrelation (if present) is usually adequately described as a lag-1 process. Cross-correlations between dependent variables are usually highest at lag-zero; that is, the errors of one dependent variable do not usually lead or lag the errors of other dependent variables in multiple-baseline experiments. Hence, the methods proposed here apply when: (1) the within-series process errors are either independent or are described by a lag-1 autoregressive process, and (2) the between-series cross-correlations are either zero or, if nonzero, are highest at lag-zero.

Step 1. Compute the overall level change statistic using the same approach described under Case I.

Step 2. Compute the test for overall level change using the following statistic:

$$\frac{\sum_{j=1}^{J} t_j}{\sqrt{\mathbf{J}^{\mathrm{T}}\mathbf{R}\,\mathbf{J}}} = t_{\mathrm{Overall}},$$

where

$\boldsymbol{J}$ is the column vector of J ones;
$\boldsymbol{J}^{\mathrm{T}}$ is the transpose of $\boldsymbol{J}$;
t_j is the t-value associated with the jth level change coefficient (estimated for the jth dependent variable);
$\mathbf{R} = J \times J =$ intercorrelation matrix of the residuals of the J fitted intervention models; and
t_{Overall} is the test statistic for overall level change.

The approximate degrees of freedom $= 0.5(N - 2)(1 + 1/J^2)$ where $N = n_1 + n_2 =$ the common total number of observations on each dependent variable. This *df* approximation usually results in values that are not integers, but the *Minitab* routine for the t distribution accommodates such values. The test statistic and the associated degrees of freedom approximation are based on theoretical work (O'Brien, 1984; Tamhane and Logan, 2004) directed toward applications that are quite removed from the one described here. It has been shown, however, that the test performs very well using the time-series designs discussed here (Awosoga, 2009).

It is useful to compare this test statistic with the corresponding test statistic that was previously described for Case I (independent error series). The statistic that applies to Case I is:

$$\frac{\sum_{j=1}^{J} z_j}{\sqrt{J}} = z_{\mathrm{Overall}}.$$

The denominator for this test statistic is simply the square root of the number of series. Note that the denominator of the test statistic for the case of cross-correlated error series is $\sqrt{\mathbf{J}^{\mathrm{T}}\mathbf{R}\mathbf{J}}$ rather than $\sqrt{J}$ and that the test statistic is t rather than z.

Why the Intercorrelations of the Series Must Be Included in the Analysis

The reason the correlations between series must be acknowledged in the analysis is pursued here. Suppose that three dependent variables are involved and that the corresponding error series are known to be independent of each other. In this case the correlation matrix computed on the errors is of the following form:

$$\mathrm{R} = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}.$$

Note that the value of each correlation on the main diagonal is one and each off-diagonal correlation is equal to zero. Hence, when the error series are independent of each other, the sum of the elements in the correlation matrix is equal to the number of series. Suppose that each level change z-value is -1.51 and that α has been set at .05. In this case each one-tailed p-value is .066; a traditional (i.e., one dependent variable at a time) testing approach leads to the conclusion that an effect has not been demonstrated. The recommended cumulative approach, however, leads to the following overall level change test statistic:

$$\frac{\sum_{j=1}^{J} z_j}{\sqrt{3}} = \frac{-4.53}{1.73} = -2.615.$$

The corresponding one-tailed p-value is .004. Thus, we can conclude that there is very strong support for a change in level when all the evidence is acknowledged.

Now, the example is modified so that the correlations between series are no longer zero. Instead, each series is perfectly correlated with the other series. That is, the value of the correlation between each series and any other series is perfect (i.e., 1.0). In this case the form of the correlation matrix is:

$$\mathrm{R} = \begin{bmatrix} 1 & 1 & 1 \\ 1 & 1 & 1 \\ 1 & 1 & 1 \end{bmatrix}.$$

This implies that a standard score at a specific time point within one series is the same on all of the other series. Suppose, once again, that each level change z-value is -1.51. Unlike the Case I example where each of the three series provides completely unique information (because the series are known to be independent), the present example involves three series that are completely redundant. Indeed, it can be argued

that only one dimension has been measured (even though scores are available on three scales). This redundancy is obvious in the correlation matrix and it will be incorporated in the error term in the overall analysis. The Case III test statistic that acknowledges this complete redundancy is shown below:

$$\frac{\sum_{j=1}^{J} z_j}{\sqrt{\mathbf{J}^{\mathrm{T}}\mathbf{R}\mathbf{J}}} = \frac{-4.53}{\sqrt{9}} = \frac{-4.53}{3} = -1.51 = z_{\mathrm{Overall}}.$$

Note that z_{Overall} is equal to the z for each individual dependent variable. After all, if each dependent variable measures exactly the same dimension then the overall test should be equivalent to a test on a single variable. If the correlation among the variables is not acknowledged and the Case I analysis is used, the value of the test statistic is -2.615, instead of the correct value -1.51. This means that the incorrect use of the Case I analysis will lead to the conclusion that the evidence is very strong ($p_1 = .004$) when the correct conclusion is that the inferential evidence is not convincing ($p_1 = .066$).

Although this example illustrates why it is important to acknowledge correlation among the series, note that the test statistic shown here (z_{Overall}) is not the same as the test statistic t_{Overall} recommended earlier for Case I. The reason is that z_{Overall} is appropriate when the process cross-correlations are known whereas t_{Overall} is appropriate when the process cross-correlations are estimated from sample data. Estimating the process correlations from sample data introduces a source of variation in the test statistic (due to sampling error in the correlations) that is not present when the process cross-correlations are known. Hence, the t-distribution better approximates the distribution of the test statistic under the null hypothesis than does the unit normal distribution. Because it is virtually always necessary to estimate the correlations from sample data, t_{Overall} is the more realistic statistic. The exception to this is Case I where the various series can be assumed to be independent (and therefore uncorrelated) by design.

Example 21.1: Case III Analysis

Case I data from Koehler and Levin (2000) analyzed earlier (see Table 21.1) are reanalyzed in this section to illustrate the required computations for a Case III analysis. Here it is not assumed that the three error series are independent; therefore, the between-series error correlations are estimated from the residuals by computing Pearson correlation coefficients between each column of residuals. Listed below are the original data (also used in the Case I analysis described earlier), the associated predictors (D1, D2, and D3) for the level change model, the residuals for each unit, the correlations among the residuals, the matrix commands for the matrices required to compute the test statistic, and the commands necessary to compute the test statistic and its one-tailed p-value.

```
MTB > print c1-c6
Data Display
Row      Unit 1     Unit 2   Unit 3    D1      D2     D3
 1        2.5        2.5      1.0      0       0      0
 2        3.0        1.0      2.0      0       0      1
 3        4.5        3.0      4.0      0       0      1
 4        2.0        2.0      2.5      0       0      1
 5        3.0        4.0      5.0      0       0      1
 6        3.0        1.0      3.0      1       0      1
 7        4.0        2.5      3.5      1       0      1
 8        3.0        5.5      2.5      1       1      1
 9        5.0        4.0      5.0      1       1      1
10        3.0        4.0      4.0      1       1      1

MTB > print c7-c9
Data Display
Row          RESI1         RESI2          RESI3
 1           -0.5         0.21429         -0.0
 2           -0.0        -1.28571         -1.5
 3            1.5         0.71429          0.5
 4           -1.0        -0.28571         -1.0
 5           -0.0         1.71429          1.5
 6           -0.6        -1.28571         -0.5
 7            0.4         0.21429         -0.0
 8           -0.6         1.00000         -1.0
 9            1.4        -0.50000          1.5
10           -0.6        -0.50000          0.5

Correlations among residuals of units 1, 2, and 3:
                    RESI1              RESI2
                    RESI2  0.147
                    RESI3  0.558       0.423
               Cell Contents: Pearson correlation

MTB > Read 3 1 m1

DATA> 1
DATA> 1
DATA> 1

3 rows read.
MTB > Read 3 3 m2
DATA> 1   .15 .56
DATA> .15  1  .42
DATA> .56 .42  1

3 rows read.
MTB >  Transpose m1 m3.
MTB > Multiply m3 m2 m4.
MTB > Multiply m4 m1 m5.
```

```
Answer = 5.2600
MTB > print m1-m5

Data Display
Matrix M1
  1
  1
  1

Matrix M2

1.00  0.15  0.56
0.15  1.00  0.42
0.56  0.42  1.00

Matrix M3
1  1  1

Matrix M4
1.71  1.57  1.98

Matrix M5
5.26

MTB > let c1=6.339/sqrt(5.26)
MTB > print c1

Data Display
C1
2.76394

MTB > cdf 2.76394;
SUBC> t 4.44.

Cumulative Distribution Function
Student's t distribution with 4.44 DF

  x          P( X <= x )
2.76394        0.977437
MTB >  let c2=1-.974678
MTB > print c2

Data Display
C2
0.025322

MTB > let c3=1-.977437
MTB > print c3

Data Display
C3
0.022563
```

It can be seen that matrix M1 = $\mathbf{J}$, matrix M2 = $\mathbf{R}$, matrix M3 = $\mathbf{J}^{\mathrm{T}}$, matrix M4 = $\mathbf{J}^{\mathrm{T}}\mathbf{R}$, and Matrix M5 = $\mathbf{J}^{\mathrm{T}}\mathbf{R}\mathbf{J}$.

The sum of the individual level change t-values shown in Table 21.1 is 6.339. Hence, the overall level change test statistic $\frac{\sum_{j=1}^{J} t_j}{\sqrt{\mathbf{J}^{\mathrm{T}}\mathbf{R}\mathbf{J}}} = \frac{6.339}{\sqrt{5.26}} = 2.76394$. The one-tailed p-value associated with $.5\,(N-2)\left(1+\frac{1}{J^2}\right) = .5\,(10-2)\left(1+\frac{1}{3^2}\right) = 4.44$ degrees of freedom is .023. (Recall that directional (one-tailed) tests are justified when the anticipated direction of the intervention change is predicted before data collection; otherwise non-directional tests should be used.) It can be concluded that the evidence for an overall level change is persuasive. Estimates of overall level change and overall standardized effect size are computed as shown for Case I.

21.5 INTERVENTION VERSUS CONTROL SERIES DESIGN

An alternative to both the reversal design and the multiple baseline design is described in this section. This design is similar to a multiple baseline design in that time-series data are collected from two or more units across time, but it differs in that an intervention is introduced to only one unit; one or more additional units serve as controls. As with multiple baseline designs, a separate model is estimated for each unit. But rather than cumulating evidence from the different units to estimate the overall change, a comparison is made between the evidence for change in the experimental unit and the evidence for change in the control unit(s). An example is illustrated in Figure 21.3. These data are from the paracetamol study described in Chapter 19 (Section 19.5).

Recall that the intervention was the introduction of legislation to limit the pack size of paracetamol in an effort to reduce poisoning, and the dependent variable was the annual age-standardized paracetamol death rate (PDR). Part of reasoning behind the legislation was that many suicides are not planned days in advance and therefore small pack sizes may make it less likely that impulsive suicide attempts (or accidental overdoses) will be fatal. The upper part of the figure illustrates the PDR before (1993–1998) and after (1999–2004) the introduction of the legislation. The data shown in the lower portion of the figure illustrate the aspirin death rate for the same years. The aspirin death rate is a meaningful comparison because the legislation targeted only paracetamol. It can be seen that both PDR and aspirin death rates decrease after the intervention, but the effect appears to be larger on PDR.

A tempting but inappropriate method of analyzing data such as these is to perform an intervention analysis on each independent series. If there is a significant effect on the experimental series and a nonsignificant effect on the control series it might seem that the evidence for change on the experimental series is significantly greater than evidence for change on the control series. This is not true. The comparison of conclusions (i.e., "significant" vs. "nonsignificant") is not a test for differential intervention effects in the two series. Suppose α is set at the conventional 5% level and the p-values for level change turn out to be .0499 and .0501 for the experimental and control series, respectively. One result is declared statistically significant and

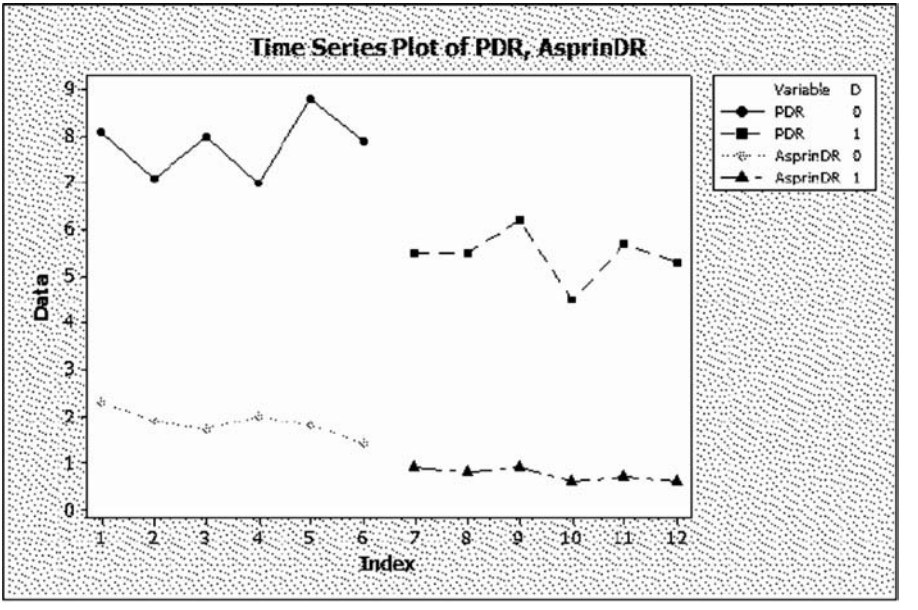


Figure 21.3 Annual death rates associated with paracetamol poisoning (upper) and aspirin poisoning (lower) before and after intervention.

the other is not, but the *p*-values are essentially identical. What is needed is not a comparison of conclusions; rather, a formal comparison of *p*-values is called for. An approach for doing this is presented next.

Testing the Difference between *p*-Values Obtained in Experimental and Control Series

Paracetamol Analysis

The paracetamol data analysis presented in Section 19.5 is based on model IV. A portion of the *TSDB* output for these data is repeated below:

```
                    Timeseries Results
  Parameter Estimates and Test that parameter is zero
  Parameter        Estimate         t-ratio        p-value
  Beta 1           7.790834         37.246         3.59472e-11
  Beta 2          -2.315002         -7.248         4.82665e-05

  Variance Covariance Matrix of Parameter Estimates

  0.0437542     -0.0504428
 -0.0504428      0.102015

  Bootstrap Residual MSE = .365827
```

Aspirin Analysis

The aspirin death rate data are

Baseline: 2.3, 1.9, 1.7, 2.0, 1.8, 1.4, and
Intervention: 0.9, 0.8, 0.9, 0.6, 0.7, 0.6

The model comparison test (see Chapter 18) comparing models I and II yields $F = 7.13$ ($p = .017$) and the p-value for the H–M test on the residuals of this model is .68.

Consequently, model I is chosen because the fit is much better using four parameters rather than two, and it appears that the errors of this model are not autocorrelated. The *Minitab* results from fitting this model using OLS are as follows:

Regression Analysis: Asprin DR versus Time, D, SC

```
The regression equation is AsprinDR = 2.30 - 0.129 Time -
                    0.500 D + 0.0686 SC

       Predictor       Coef       SE Coef       T       P
       Constant       2.3000      0.1464     15.72   0.000
       Time          -0.12857     0.03758    -3.42   0.009
       D             -0.5000      0.1854     -2.70   0.027
       SC             0.06857     0.05315     1.29   0.233

       S = 0.157208   R-Sq = 95.3%   R-Sq(adj) = 93.5%

     Analysis of Variance
     Source          DF      SS       MS        F         P
     Regression       3    3.9823   1.3274    53.71     0.000
     Residual Error   8    0.1977   0.0247
     Total           11    4.1800
```

If we focus on level change we see that the level change coefficients are -2.315 and -0.50 for paracetamol and aspirin, respectively. The corresponding p-values are reported in the output as 4.82665e−05 and .027. These are two-tailed p-values; the required one-tailed p values are:

Paracetamol series: $p_1 = .000024133$, and
Aspirin series: $p_1 = .013537$

Test Statistic for Comparing Two Independent *p*-Values

Note that these p-values result in the conclusion that there is a statistically significant level change in both series, but this is not the current question. We are now questioning whether the p-value associated with level change on the paracetamol series

is significantly different from the p-value associated with level change on the aspirin series. The test statistic relevant to this question is

$$\frac{z_E - z_C}{\sqrt{2}} = z$$

where z_E is the standard normal deviate associated with the experimental series level change p_1, z_C is the standard normal deviate associated with the control series level change p_1, and z is the test statistic that is evaluated as a standard normal deviate.

Note that each p_1-value must be transformed to a corresponding normal deviate. This approach is the same as the one described in Chapter 20 for another application. The *Minitab* menu commands for obtaining the standard normal deviates associated with the example p_1-values (i.e., .000024133 and .013537) are as follows:

```
Menu Bar → Calc → Probability Distributions→
Normal → Inverse Cumulative Probability → Input Constant →
(Enter .000024133) → OK
```

Inverse Cumulative Distribution Function

Normal with mean = 0 and standard deviation = 1

P(X < = x)	x
0.0000241	-4.06387

The value of z below which the area is .0000241 is $-4.06387 = z_E$. The same approach applied to the aspirin p_1-value of .013537 yields $z_C = -2.21$. The test statistic is computed as follows:

$$\frac{-4.064 - (-2.210)}{\sqrt{2}} = -1.31 = z.$$

The area of the standard normal distribution below $z = -1.31$ is .095. The two-tailed area is .19; hence, it is concluded that there is not a statistically significant difference between the paracetamol and aspirin p-values. That is, it has not been demonstrated that the evidence for level change on the paracetamol series is significantly more persuasive than the evidence for level change on the control (aspirin) series.

Useful descriptive values to accompany the inferential result are likely to be of interest. The first is the standardized level change and the second is the percentage reduction in death rate level relative to the baseline level. The standardized level change statistics are:

$$\frac{-2.135}{\sqrt{.3658}} = -3.53 \quad \text{and} \quad \frac{-.50}{.157208} - 3.18,$$

where -2.135 and -0.50 are the level change estimates for paracetamol and aspirin, respectively, 0.3658 is the bootstrap residual mean square from the *TSDB* output

for the paracetamol series, and 0.157208 is the estimated standard error of estimate (denoted as "s" in *Minitab* output) for the aspirin series.

The computation of the percentage reduction in death rate level after intervention requires estimates of the baseline level and the level change coefficient. The baseline level is estimated by the intercept in the case of model II; this is the model used for the paracetamol series; it can be seen in the *TSDB* output shown above that the intercept is 7.79. In the case of model I (used for the aspirin series) the baseline level is estimated using $b_0 + b_1(n_1 + 1)$. Hence, the baseline level for the aspirin series is: $2.3 - .12857(7) = 1.4$. The reduction in level after intervention is $\frac{-2.32}{7.79} = -.30$ or 30% for paracetamol and $\frac{-.50}{1.40} = -.36$ or 36% for aspirin.

The analyses described in this section are quite different than the ones reported by Morgan et al. (2007), but the overall conclusion is the same. Those authors also collected data on death rates associated with other drugs and on nondrug suicides. Interestingly, there was an unexplained general downward trend on all of these "control" series after the legislation was introduced. This study illustrates, once again, the problem of drawing causal inferences from two-phase (AB) studies. If control data had not been available in this example the statistically significant level change identified on the paracetamol series would almost certainly have been attributed to the intervention. The importance of having at least one control series is difficult to overstate in research of this type.

21.6 SUMMARY

There are essentially three major variants of the multiple-baseline design: (1) multiple baseline across subjects, (2) multiple-baseline across settings, and (3) multiple-baseline across dependent variables. All three variants are practical and have high in internal validity. They sometimes differ in terms of the most appropriate method of statistical analysis. All of the proposed analytic approaches are extensions of the methods recommended for the analysis of simple two-phase AB designs. The extensions acknowledge the logic of the design and cumulate information provided by each of the multiple series to provide overall measures of level and slope change, overall measures of effect size, and tests on these statistics. Several variants of these analyses are presented that differ with respect to the type of error dependency (if any) that is present within series and between series.

A design that has multiple series but is not classified as a multiple-baseline design is also described. This design involves an experimental series to which the intervention is applied and one (or more) additional series that plays the role of a control that has not been subject to the intervention. A test that contrasts the evidence for change in the experimental series against the evidence for change in the control series is presented. This design and analysis approach may provide protection against the internal validity threats of history and maturation to which most conventional AB designs are vulnerable.

PART VI

ANCOVA Extensions

CHAPTER 22

Power Estimation

22.1 INTRODUCTION

Power estimation is an important step in the design of experiments, regardless of the method of analysis. It can reduce the number of experiments that are doomed to failure before they begin. Recall from Chapter 1 that a researcher who has a good understanding of research results in a content area of interest can often provide an informed opinion regarding differences between means that fall into three categories: (1) trivial or unimportant, (2) of questionable importance, and (3) of definite practical or theoretical importance. When this can be done it is the third category that is of most interest in the determination of power. A meaningful power analysis requires an answer to the following question: What is the smallest difference between means that would be considered to be of practical or theoretical importance? The answer to this question is the starting point for a power analysis.

The experiment should be designed so that a difference judged as important will be detected by the statistical analysis. This is especially important in grant applications. Funding agencies are only interested in supporting experiments that have a high probability of rejecting the null hypothesis when it is false. This will not occur unless the analysis has adequate power. This brief chapter first describes how to estimate power for ANOVA by using the *Minitab* routine for this purpose. The same routine is appropriate for ANCOVA power estimation.

22.2 POWER ESTIMATION FOR ONE-FACTOR ANOVA

In the case of a simple two-group randomized experiment, the power of the test (either t or F) depends on the true (population) difference(s) between the means, the number of subjects in each group (i.e., the sample size n), the size of α chosen for the test, and the standard deviation within groups. Hence, in order to estimate power one needs to provide (1) the difference between population means that is large

The Analysis of Covariance and Alternatives: Statistical Methods for Experiments, Quasi-Experiments, and Single-Case Studies, Second Edition. Bradley E. Huitema.

enough to be of interest, (2) the sample size, (3) the chosen level of α, and (4) an estimate of the within-group standard deviation. An estimate of the within-group standard deviation may be obtained from a previous study, if one can be found; the dependent variable should be the same as in the planned study and the subjects should be similar. If previous research is not available to provide this estimate, it may be necessary to run a pilot study. In either case the estimate is provided by the square root of the within-group MS obtained from the analysis of variance from the previous experiment or pilot study. Once the four ingredients listed above are provided, power can be estimated by computing a coefficient know as phi (not the phi coefficient referred to in correlation analysis) and finding it in power tables (available in most older experimental design textbooks) to determine power. The contemporary approach is to use widely available software routines for this purpose. The *Minitab* power routine is very convenient to use. An example is shown below.

Suppose an experimenter plans a study similar to the experiment described as Example 3.1 (Chapter 3). The dependent variable will be the same, the subjects will be drawn from the same population, and three treatments (somewhat different than the original treatments) will be included. A difference of 12 points is judged as the minimum difference between population means that is of definite practical importance. Therefore, there is an interest in determining the power for detecting a difference between population means that differ by 12 points, given the planned sample size of 10 subjects per group in a three-group design, alpha set at .05, and a within-group standard deviation that is the same as in the previous experiment (i.e., Experiment 1 in Chapter 3). The analysis of variance summary table from the previous experiment shows that $MS_w = 131$; therefore, the within-group standard deviation is 11.44. The *Minitab* menu commands for power are as follows:

Stat → *Power and Sample Size* → *One-Way ANOVA* → *NUMBER OF LEVELS*: 3 → *Sample sizes*: 10 → *Value of the maximum difference between means*: 12 → *Standard deviation*: 11.44 → *OK*

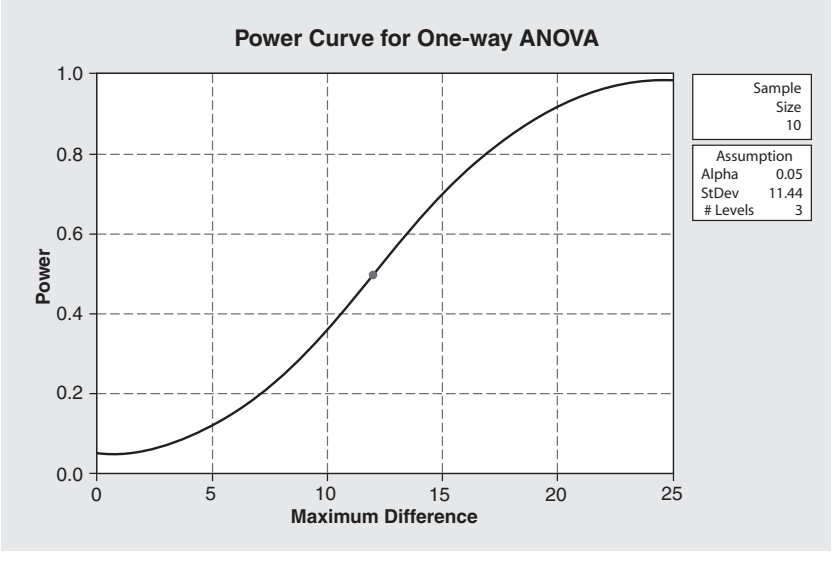
The command line editor commands are

```
MTB > Power;
SUBC>   OneWay 3;
SUBC>     Sample 10;
SUBC>     MaxDifference 12;
SUBC>     Sigma 11.44.
```

The output is

```
Power and Sample Size
One-way ANOVA
Alpha = 0.05  Assumed standard deviation = 11.44
Factors: 1  Number of levels: 3
   Maximum  Sample
Difference    Size     Power
       12      10     0.496900
The sample size is for each level
```

The default for alpha is .05, but this can be set to any value. After opening the "Power and Sample Size for One-Way ANOVA" window, an "Options" button and a "Graph" button can be seen. If you click on options, another window opens and you will see where to type the "significance level" (i.e., alpha) you want. If you click the graph button and the "Display power curve" box, the power curve will appear at the same time as the session output. This is a very useful plot; I recommend requesting it routinely because it can immediately provide answers to many questions regarding the effects of changing the size of the effect to be detected. The following power curve is produced for the example data:

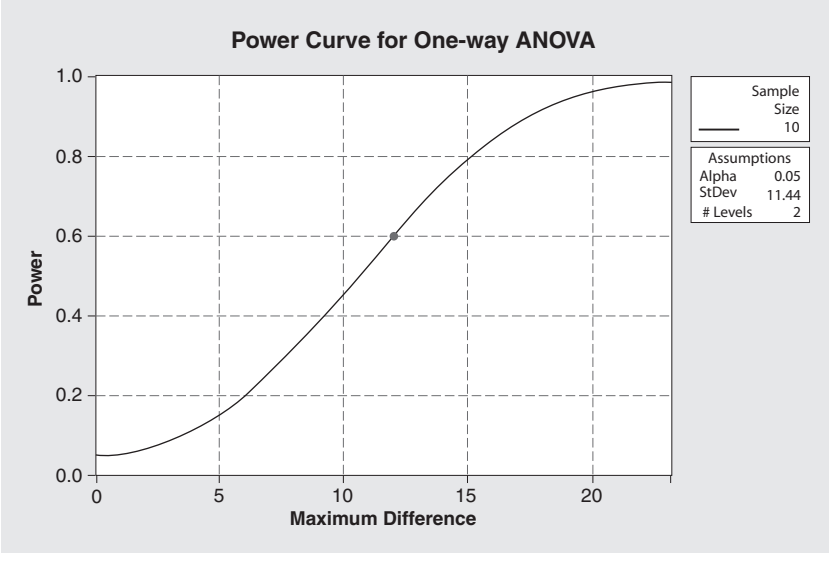


It can be seen from the elevation of the dot that the power estimate is about .50; this is unsatisfactory. Although conventions for acceptable power are not quite as firmly established as those for α, it is a rare funding agency that would look favorably upon a power estimate that is as low as .50. After all, this is saying that there is only a 50-50 chance of success (i.e., detecting a difference of 12 points when it exists). A power value of .80 is widely considered acceptable, but I prefer to use .90.

Note in the figure that power increases rapidly as the difference between means increases. If you had chosen a difference of over 17 points as the definition for an important finding then the power is above .80. But if the true difference is five points, note that the power is just a little above .10. Clearly, the value chosen as a difference of importance is crucial; power is a function of the choice.

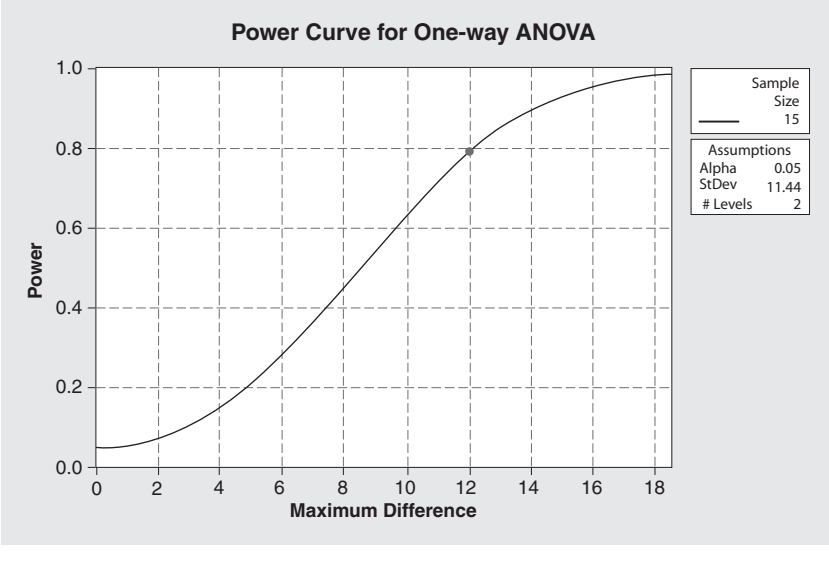
Three other aspects of the experiment were mentioned above for two-group experiments: α, n, and the within-group standard deviation σ_w. But this list is not exhaustive in the case of ANOVA and ANCOVA. The number of groups in the design is largely unrecognized as having a substantial effect on power. Suppose that the experiment described above can be simplified so that it contains only two groups rather than three. If everything about the original experiment is kept the same (i.e., α, n, and σ_w), but now the number of groups is two, power will go up even though *fewer* subjects

are included in the experiment (i.e., 20 rather than 30). This is demonstrated in the power curve shown below.



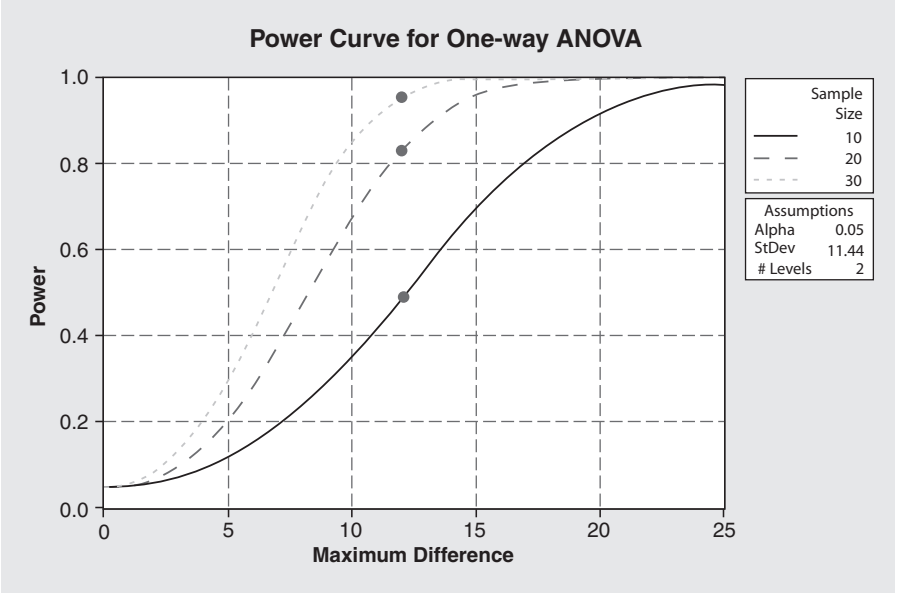
Note that power is slightly over .60 (rather than the previously computed value of .50 in the three-group design) even though the other properties of the design are identical.

Now redesign the experiment so that the total number of subjects is the same as in the original design (i.e., $N = 30$) but use only two groups so that the sample size is $n = 15$. Now compute power for this design. The power curve is shown below:



Note that power is now almost .80 using the same N as in the original three-group experiment that has power = .50. The message here is that careful consideration should be given to the need for each treatment group.

The next issue is sample size. It is relevant to consider what happens to power as the sample size is increased. This is conveniently illustrated for the original example data by simply entering 10, 20, 30 rather than 10 in the window for "sample sizes" in *Minitab.* (This prompt for sample size is not asking for a separate sample size for each group; it always assumes equal sample size for each group; this is why the number 10 was entered only once following sample size in the original menu input shown above.) The power curves for three sample sizes are shown next.

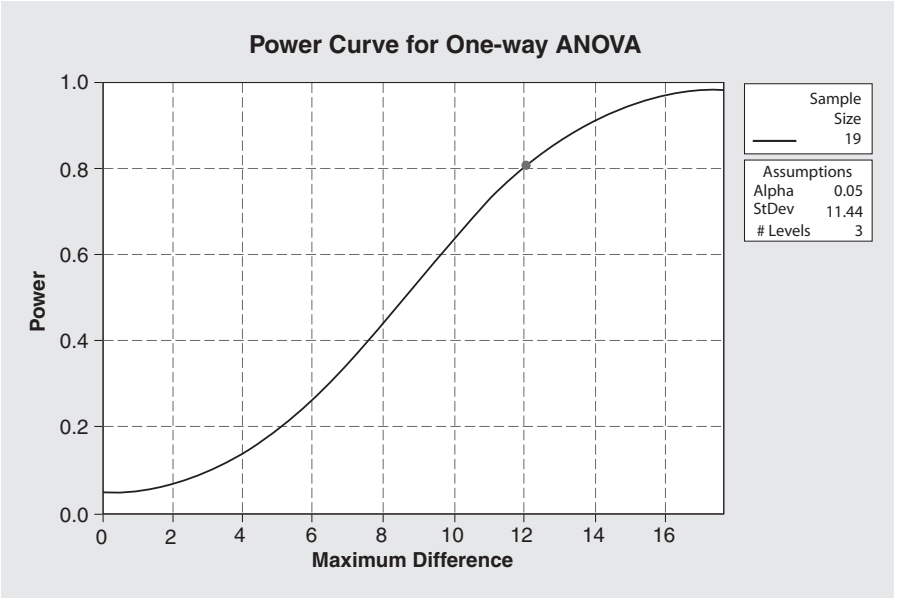


The power values are .50, .83, and .96, for $n = 10$, 20, and 30, respectively. The curves clearly show the advantage of increasing sample size.

Determination of Sample Size for Specified Power

Because sample size is easily manipulated in the design of an experiment the issue of power is often dealt with by first deciding on the desired power level and then determining the sample size that is required to achieve it. Suppose power is set at .80 and the other aspects of the design are the same as shown above for the original power analysis. The appropriate *Minitab* menu commands are

Stat → *Power and Sample Size* → *One-Way ANOVA* → *Number of Levels*: 3 → *Value of the maximum difference between means*: 12 → *Power values* .80 → *Standard deviation*: 11.44 → *OK*



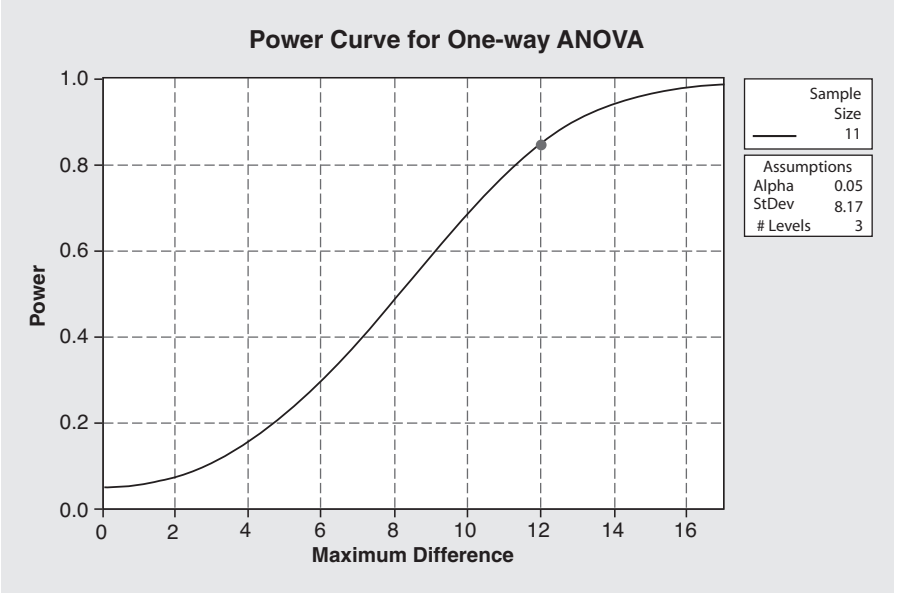
The required number of subjects per group is 19 to achieve power = .80.

22.3 POWER ESTIMATION FOR ANCOVA

Return to the example mentioned above. Once again suppose that the experimenter is interested in designing an experiment that will use the same dependent variable and subject pool as in the original experiment, but the treatments will be somewhat different. Again, the difference between population means that is considered to be of practical importance is 12 points. But now the experimenter has discovered that she has access to aptitude score data on the subjects who will be included in the experiment. Further, she has identified published research results indicating that the within-group correlation between aptitude scores and achievement scores of the type she plans to use is approximately .70. The square of this value provides an estimate of the proportion of the within-group variation on achievement that is predictable from aptitude scores.

This implies that the within-group standard deviation she used when computing power for ANOVA is an overestimate. If she uses aptitude scores as a covariate in an ANCOVA the error mean square should be smaller than the error mean square in ANOVA. The relationship between the ANOVA error mean square and the ANCOVA error mean square is ANCOVA MS error = ANOVA MS error times

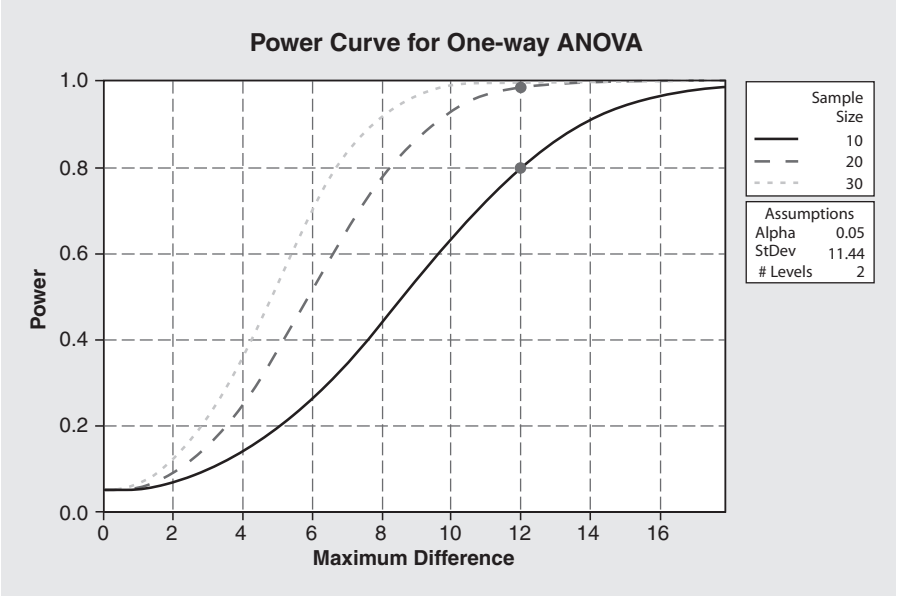
$(1 - r^2)$. MS_w from a previous study was 131. Recall that the square root of this value (i.e., 11.44) was used in the original power analysis and in determining the number of subjects needed per group. Because the estimated mean square for ANCOVA is $131(1 - .70^2) = 66.81$, the standard deviation to be used to estimate power for ANCOVA is $\sqrt{66.81} = 8.17$. We simply substitute this value for the conventional within-group standard deviation in the *Minitab* power routine. The outcome of this is shown below.



The output is not appropriately labeled as an ANCOVA power curve because the ANOVA routine has been used; the only difference in computation procedures for ANOVA and ANCOVA is the value entered as the standard deviation. Note the effect of including the covariate on the sample size (shown in the box in the upper right) that is required to achieve power = .80. Whereas a sample size of 19 is required for ANOVA, the required ANCOVA sample size is only 11. Hence, the experiment requires a total $N = 57$ if ANOVA is used whereas the required $N = 33$ if ANCOVA is used instead.

The ANCOVA power curve shown below reveals an important feature of power analysis that has not yet been mentioned.

Note that the power estimates are .80, .99, and .99, for sample sizes 10, 20, and 30, respectively. This implies that requiring the experimenter to use $n > 20$ is wasteful, because $n = 20$ is large enough to virtually guarantee detection of the desired difference. Of course, the situation is quite different if it is decided that a difference



of eight points is of practical importance. In this case the power values are .44, .78, and .93, for $n = 10$, 20, and 30, respectively.

22.4 POWER ESTIMATION FOR STANDARDIZED EFFECT SIZES

Some researchers prefer to estimate ANOVA power for standardized effect sizes rather than for effects measured in the original metric. I prefer the original metric, but it is simple to adapt the *Minitab* routine to the standardized metric. For example, suppose you want to determine the sample sizes necessary for a two-group ANOVA to detect conventionally defined small, medium, and large standardized effects (with power = .80). Simply enter .20, .50, and .80 as the "values of the maximum difference between means" and enter 1.00 as the value of the standard deviation. In the case of ANCOVA use $1 - r^2$ as the value of the standard deviation.

22.5 SUMMARY

The estimation of power may facilitate the design of experiments in two ways. First, it may prevent the initiation of an experiment that has a poor chance of detecting an important difference. Second, it may eliminate the use of more subjects than are required to answer the question of interest. Power is affected by the size of the difference between population means, the level of significance (α) set for the test, the sample size, the within-group standard deviation, the number of groups in the experiment, and the degree of sample size balance. There is little difference between the method used to estimate power for ANCOVA and the method that applies to ANOVA.

CHAPTER 23

ANCOVA for Randomized-Block Designs

23.1 INTRODUCTION

ANCOVA and blocking are widely viewed as competing methods for controlling nuisance variation. It is usually argued that ANCOVA is superior when the relationship between the covariate and the dependent variable is strong, and that blocking is better when the relationship is not linear (because the blocking factor is function free). Although there is truth in this conventional wisdom, there are better ways of dealing with most types of nonlinearity. Quadratic ANCOVA is usually preferable. It has two advantages: (1) Only one degree of freedom is consumed by the curvature component, whereas each block consumes an additional degree of freedom in the RB analysis. (2) Quadratic ANCOVA provides adjusted means, RB ANOVA does not.

The typical way of characterizing the comparison of "ANCOVA" with "blocking" confounds two issues that should be kept separate: design and analysis. ANCOVA is a method of analysis; it is not a design. Blocking is a method of design; it is not an analysis. Hence, a literal comparison of ANCOVA with blocking makes no sense because this implies a comparison of a method of analysis with a type of design. The actual comparison that should be articulated is the analysis of the randomized-group design using ANCOVA versus the analysis of the randomized-block design using RB ANOVA. The failure to distinguish the design from the analysis has led most researchers believe that ANCOVA does not apply to the randomized-block design. The purpose of this brief chapter is to demonstrate that it is desirable to analyze the typical randomized-block design using either (1) conventional ANCOVA or (2) a more complex analysis that includes a covariate.

Blocking is almost always a good idea. Recall from Chapter 3 that it involves forming subgroups on the basis of some measurement (such as a pretest) that is substantially correlated with the dependent variable. Then, subjects within blocks are randomly assigned to treatments. The traditional version has as many subjects within

The Analysis of Covariance and Alternatives: Statistical Methods for Experiments, Quasi-Experiments, and Single-Case Studies, Second Edition. Bradley E. Huitema.

each block as there are treatment conditions; this implies that there is one subject at the intersection of each treatment and block. The details of setting up a design of this type are described next.

23.2 CONVENTIONAL DESIGN AND ANALYSIS EXAMPLE

A randomized-block experiment was designed to have three treatments and five blocks. Pretest (blocking variable) data were obtained on 15 subjects. The ordered blocking variable scores appear below.

0, 10, 13, 21, 24, 35, 36, 38, 60, 70, 72, 91, 97, 99, 99

The subjects who obtained the scores of 0, 10, and 13 constitute block 1. These subjects were randomly assigned to treatments 1, 2, and 3. The second block consists of subjects who obtained the scores of 21, 24, and 35; they were randomly assigned to the three treatments. The process was repeated for the remaining blocks. The treatments were applied and dependent variable scores were obtained. Four analyses were applied to the data. The first three are summarized in Table 23.1.

Table 23.1 Randomized-Group ANOVA, Randomized-Block ANOVA, and Randomized-Group ANCOVA Applied to Data from a Randomized-Block Design

One-factor ANOVA for randomized group design:

Source	DF	SS	MS	F	P
Gp	2	2221	1110	0.19	0.826
Error	12	68603	5717		
Total	14	70824			

One-factor Randomized Block ANOVA:

Source	DF	SS	MS	F	P
Bk	4	66213.1	16553.3	55.41	0.000
Gp	2	2220.9	1110.5	3.72	0.072
Error	8	2389.7	298.7		
Total	14	70823.7			

One-factor ANCOVA:

Source	DF	Adj SS	MS	F	P
Gp	2	431	216	14.33	0.001
Error	11	166	15		

Adjusted Means		
Gp	N	Y
1	5	105.03
2	5	117.58
3	5	107.79

Analytic Approaches

The first analysis in the table is a conventional one-factor randomized-group (RG) ANOVA that ignores both the blocks and the covariate. The second analysis is the conventional randomized-block (RB) ANOVA that includes the blocking factor, but the blocking *scores* (X) do not appear in this analysis; indeed, the X scores may be discarded as soon as the blocks are formed. The third analysis is a conventional RG ANCOVA that uses X as the covariate. This analysis ignores blocks; it is carried out as if the data came from a randomized-group design. A fourth analysis (not included in the table) is based on a model that includes treatments, blocks, and the covariate. Because this is not a standard model it is described subsequently in detail.

The treatment means associated with the first two analyses are, of course, identical. They are 95.40, 125.20, and 109.80 for treatments 1, 2, and 3, respectively. Although there is a dramatic difference in the size of the randomized-group and randomized-block ANOVA p-values (i.e., .83 vs. .07), they both fail to detect an effect. In contrast, the ANCOVA p-value is only .001 even though the differences among the adjusted means are smaller than the differences among the unadjusted means of the RG and RB analyses.

The Role of X

Recall that X scores have no direct role in the conventional RB ANOVA of the RB design. That is, the X scores are used only in the *design* of the experiment (to form blocks), not for any other purpose. In contrast, there is a second role for X when ANCOVA is involved. Not only is X used at the design stage to form blocks, but it also is used as the covariate in the *analysis*. It is widely believed, however, that there is no need to do so and that if ANCOVA is used with this design it offers no advantage over conventional RB ANOVA. One purpose of this chapter is to demonstrate that this is not true.

Design Variants

The application of ANCOVA to the example data illustrates the use of X in both design and analysis stages of a typical version of the RB experiment. A second variant involves using one variable that may be categorical (e.g., litters or hospitals) to form blocks and a second variable (perhaps a quantitative variable discovered after the experiment is completed) as a covariate. A third design variant involves the same general structure as the first, but the number of subjects within each block is a multiple (other than one) of the number of treatments. I recommend using conventional ANCOVA for the first variant and a more complex method (described in the next section) for the second variant. The third variant is not a true randomized-block design; rather it is a two-factor design with one classification factor (blocks) and one treatment factor (the manipulated factor); it can be analyzed as described in Chapter 24.

23.3 COMBINED ANALYSIS (ANCOVA AND BLOCKING FACTOR)

The Model

The model for the combined analysis can be written as follows:

$$Y_{ij} = \mu + \alpha_j + \beta_1 \left(X_{ij} - \bar{X}..\right) + \gamma_i + \varepsilon_{ij},$$

where the terms are the same as in the conventional ANCOVA model except for the gamma, which is included to measure block effects.

The test for adjusted treatment effects is carried out using a regression model-comparison approach that is similar to the one described in Chapters 7 and 10 for conventional ANCOVA. The only difference is the required predictor variables.

Required Predictor Variables

The required predictor variables for the combined analysis are shown in Table 23.2 for the same example outcome data analyzed in Table 23.1. The first column simply identifies 15 subjects and is not used in the analysis; subjects 1 to 5 were exposed treatment 1, subjects 6 to 10 were exposed to treatment 2, and subjects 11 to 15 were exposed to the third treatment. Columns d_1 and d_2 are the $J - 1$ treatment indicator variables (using 1, 0, −1 coding). The value 1 in column d_1 indicates that a subject belongs to treatment 1, −1 indicates that a subject is in the last treatment, and 0 indicates that a subject is in neither the first nor the last condition. The value 1 in column d_2 indicates that a subject belongs to treatment 2, −1 indicates that a subject is in the last treatment, and 0 indicates that a subject is in neither the second nor

Table 23.2 Example Input Data Required for Combined Analysis

ID	d1	d2	d3	d4	d5	d6	X	Y
1	1	0	1	0	0	0	0	9
2	1	0	0	1	0	0	24	48
3	1	0	0	0	1	0	38	76
4	1	0	0	0	0	1	70	144
5	1	0	−1	−1	−1	−1	99	200
6	0	1	1	0	0	0	13	40
7	0	1	0	1	0	0	35	89
8	0	1	0	0	1	0	36	89
9	0	1	0	0	0	1	91	192
10	0	1	−1	−1	−1	−1	99	216
11	−1	−1	1	0	0	0	10	20
12	−1	−1	0	1	0	0	21	48
13	−1	−1	0	0	1	0	60	128
14	−1	−1	0	0	0	1	72	149
15	−1	−1	−1	−1	−1	−1	97	204

the last condition. Similarly, columns d_3 through d_6 are the $K-1$ block indicator variables (where K is the number of blocks). Column X contains the covariate scores and column Y the dependent variable scores.

Test for Adjusted Treatment Effects

The model comparison of interest is the full combined model

$$Y_{ij} = \mu + \alpha_j + \beta_1 \left(X_{ij} - \bar{X}..\right) + \gamma_i + \varepsilon_{ij}$$

versus the following reduced form of the model

$$Y_{ij} = \mu + \beta_1 \left(X_{ij} - \bar{X}..\right) + \gamma_i + \varepsilon_{ij}.$$

A difference in the predictive effectiveness of these models implies that the hypothesis of no adjusted treatment effect is false. The model comparison F-test of this hypothesis involves four steps.

First, regress Y on the indicator variables for treatments, the indicator variables for blocks, and the covariate. Second, regress Y on the indicator variables for blocks and the covariate. Third, subtract the SS Regression for the second regression from the SS Regression for the first regression to obtain the Adjusted Treatment SS. Last, form the following ratio:

$$\frac{\left(\text{SS Regr}_{\text{Full}} - \text{SS Regr}_{\text{Reduced}}\right) \big/ \left(df\ \text{Regr}_{\text{Full}} - df\ \text{Regr}_{\text{Reduced}}\right)}{\text{MSres}_{\text{Full}}} = F$$

The degrees of freedom are $J-1$ and $(J-1)(K-1)-1$.

The application of these steps to the example data yields:

$$\frac{\left(\text{SS}_{\text{TX, BK, X}} - \text{SS}_{\text{BK, X}}\right) \big/ \left(df_{\text{TX, BK, X}} - df_{\text{BK, X}}\right)}{\text{MSres}_{\text{Full}}} = \frac{(70687 - 70280)/(7-5)}{19.429}$$

$$= \frac{203.500}{19.429} = 10.474$$

The degrees of freedom are 2 and 7 and the p-value is .008.

Adjusted Means

The adjusted means to which the combined test applies are based on the full model regression coefficients. That is,

$$\bar{Y}_{j\,\text{adj}} = b_0 + b_1 d_1 + b_2 d_2 + \cdots + b_{(J-1)} d_{(J-1)} + b_J 0 + b_{J+1} 0 + \cdots + b_{(J+K-2)} 0 + b_{J+K-1} \bar{X}..$$

For the example, the first adjusted mean $= b_0 + b_1(1) + b_7(\bar{X}..) = 14.056 + (-5.691) + 1.884(51) = 104.44$; the adjusted means for treatments 2 and 3 are 118.04 and 107.92, respectively.

A comparison of the results of this analysis with the results of the conventional ANCOVA shown at the bottom of Table 23.1 reveals a larger F for the conventional ANCOVA. The main reason for this is that ANCOVA has more degrees of freedom for the error term (i.e., 11 for ANCOVA versus 7 for the combined analysis). The sum of squares for error is actually smaller for the combined analysis. The differences among adjusted means are slightly larger for the combined analysis but, because the blocks introduce additional dependency among the adjusted means, the adjusted treatment sum of squares is slightly smaller.

Because the example design is based on very small samples and uses the same X variable for both block formation and as the covariate in the analysis, the conventional ANCOVA provides a smaller error term than other analyses. The combined analysis is recommended for large samples where one variable (that may be qualitative) is used for block formation and a second variable is used as the covariate. In the latter case a well-chosen covariate will substantially reduce error variation that is not explained by the blocking factor.

23.4 SUMMARY

Conventional ANCOVA is a highly efficient method for the analysis of conventional RB designs where the covariate is used for block formation. It usually has higher power than the typical RB ANOVA because variation on X not captured by blocks remains in the error term. A second approach, described as a combined ANCOVA and block analysis is recommended for the case where a one variable is used for block formation and a second variable is used as a covariate.

CHAPTER 24

Two-Factor Designs

24.1 INTRODUCTION

The ANOVA associated with randomized-block and repeated-measurement designs described in Chapter 3 is employed to test the same hypothesis as is tested with completely randomized ANOVA and ANCOVA designs. That is, $H_0 : \mu_1 = \mu_2 = \cdots = \mu_J$. With all these designs there is just one factor that is of experimental interest. A comparison of the randomized-group ANOVA partitioning with (1) the randomized-group ANCOVA, (2) the randomized-block ANOVA, and (3) the repeated-measurement ANOVA partitioning reveals a common pattern. It can be seen in Figure 24.1 that in the case of the randomized-group analysis the total sum of squares is partitioned into only two components: treatment and within treatment. All other analyses involve breaking down the within-treatment sum of squares into two separate components.

Variation accounted for by X in ANCOVA, by blocks in the randomized-block ANOVA, and by subjects in the repeated-measurement analysis has the same type of effect for all these analyses. These sources of variation are removed from the within-treatment sum of squares; consequently, the error term in the F ratio is reduced. The effect of this is to increase the power of the analysis. These strategies for reducing the size of the error variation by employing covariates, blocks, or subjects can be generalized to designs in which the experimenter is interested in evaluating the effects of two independent variables.

The application of covariance analysis to two-factor independent sample designs is described in this chapter. If the reader has not been previously exposed to two-factor designs, I recommend that another source be consulted before completing this chapter. Excellent descriptions are presented in Keppel and Wickens (2004), Kirk (1995), Howell (2010), and Maxwell and Delaney (2004).

The Analysis of Covariance and Alternatives: Statistical Methods for Experiments, Quasi-Experiments, and Single-Case Studies, Second Edition. Bradley E. Huitema.

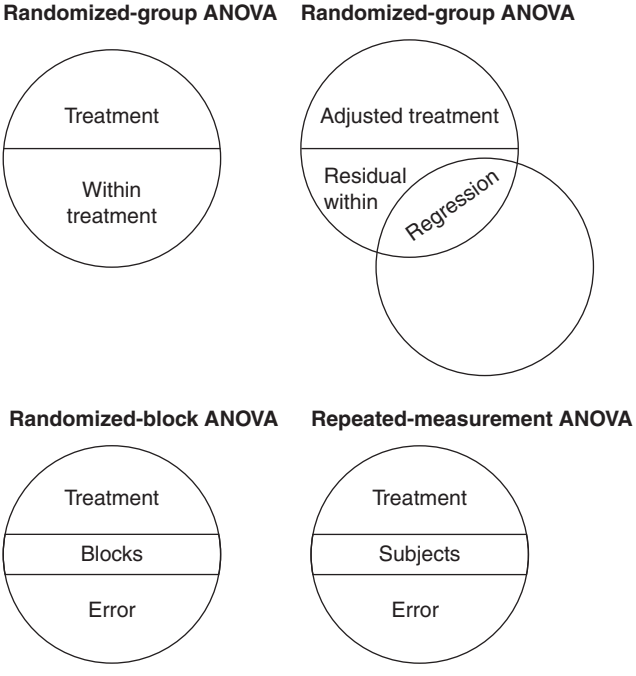


Figure 24.1 Partitioning for analyses of four one-factor designs.

Purpose of Independent Sample Two-Factor Designs

It is typical for experimenters to be interested in evaluating the effects of more than one independent variable in a single experiment. For example, an experimenter might be interested in studying the effects of two different types of reinforcement (monetary and social) and two different creative thinking programs (Purdue and Khatena) on the frequency of novel responses produced under controlled conditions. The two factors in this design are (1) type of reinforcement (factor A) and (2) type of program (factor B). Each factor has two levels. An advantage of designing the experiment as one two-factor design rather than as two one-factor designs is that possible interaction effects can be studied. Separate studies of the effects of type of reinforcement and type of program will provide no information on the effects of both factors applied simultaneously. The three major hypotheses that are tested with a two-factor design are

$$H_0 : \mu_{A_1} = \mu_{A_2} = \cdots = \mu_{A_I}$$

$$H_0 : \mu_{B_1} = \mu_{B_2} = \cdots = \mu_{B_J}$$

H_0 : Simple effects of factor A are consistent across all levels of factor B,

where μ_{A_1} through μ_{A_I} are the population marginal means associated with treatment levels A_1 through A_I, and μ_{B_1} through μ_{B_J} are the population marginal means associated with treatment levels B_1 through B_J. These three hypotheses are equivalent to stating that the A, B, and $A \times B$ effects in the two-factor ANOVA model are equal to zero. The model is

$$Y_{ijk} = \mu + \alpha_i + \gamma_j + \alpha\gamma_{ij} + \varepsilon_{ijk},$$

where

μ is the population mean common to all observations;
α_i is the effect of ith level of factor A;
γ_j is the effect of jth level of factor B;
$\alpha\gamma_{ij}$ is the interaction effect of ith level of factor A and jth level of factor B; and
ε_{ijk} is the error component associated with kth observation in ith level of factor A and jth level of factor B.

Hence, if population data from a two-factor experiment that has two levels of factor A (i.e., $I = 2$) and two levels of factor B (i.e., $J = 2$) are available, the marginal means reflect the overall (main) effects of the two factors. It can be seen in Table 24.1 that each factor A marginal mean is averaged across both levels of factor B and that each factor B marginal mean is averaged across both levels of factor A. The *cell means* are employed in describing possible interaction between the two factors. The difference between cell means $\mu_{A_1B_1}$ and $\mu_{A_2B_1}$ (the simple effect of A at B_1) is not the same as the difference between $\mu_{A_1B_2}$ and $\mu_{A_2B_2}$ (the simple effect of A at B_2). Because these two differences are not equal, it is said that the simple effects of factor A are not consistent at the two levels of factor B and, therefore, that factors A and B interact.

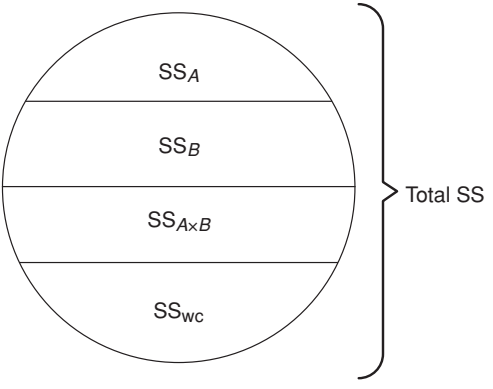
When a two-factor experiment is carried out, the analysis of variance is employed to test the hypotheses that (1) the I factor A marginal means are equal, (2) the J factor B marginal means are equal, and (3) the difference between the factor A cell means is the same at all levels of factor B. The summary table for the independent sample two-factor analysis of variance is shown in Table 24.2, along with a diagrammatic representation of the partitioning of the total sum of squares. It can be seen in the summary table that three F ratios are obtained, one for each of the null hypotheses

Table 24.1 Hypothetical Population Means Associated with a Two-Factor Design

		Factor B		
		Level B_1	Level B_2	Marginal Means for Factor $A \downarrow$
Factor A	Level A_1	$\mu_{A_2B_1} = 10$	$\mu_{A_1B_2} = 30$	$\mu_{A_1} = 20$
	Level A_2	$\mu_{A_2B_1} = 5$	$\mu_{A_2B_2} = 15$	$\mu_{A_2} = 10$
	Marginal Means $\rightarrow$ for Factor B	$\mu_{B_1} = 7.5$	$\mu_{B_1} = 22.5$	

Table 24.2 ANOVA Summary and Partitioning for Two-Factor Independent Cells Design

Source	SS	df	MS	F
Factor A	SS_A	$I-1$	$SS_A/(I-1)$	MS_A/MS_{wc}
Factor B	SS_B	$J-1$	$SS_B/(J-1)$	MS_B/MS_{wc}
$A \times B$ Interaction	$SS_{A\times B}$	$(I-1)(J-1)$	$SS_{A\times B}/[(I-1)(J-1)]$	$MS_{A\times B}/MS_{wc}$
Within cell	SS_{wc}	$N-IJ$	$SS_{wc}/(N-IJ)$	MS_A/MS_{wc}
Total		$N-1$		



described earlier. Each obtained F value is compared with the corresponding critical value of F that is based on the degrees of freedom indicated below.

Obtained F	Compared with	Critical F
F_A	Compared with	$F_{(\alpha,I-1,N-IJ)}$
F_B	Compared with	$F_{(\alpha,J-1,N-IJ)}$
$F_{A\times B}$	Compared with	$F_{[\alpha,(I-1)(J-1),N-IJ]}$

If the obtained value equals or exceeds the critical value, the associated null hypothesis is rejected. Correspondingly, standard software for this design provides a p-value along with each of these F ratios.

It is pointed out in most introductions to the two-factor ANOVA that the sample sizes associated with the various cells of the design should be either equal or proportional. An example of unequal but proportional sample sizes is as follows:

	B_1	B_2
A_1	$n=15$	$n=30$
A_2	$n=10$	$n=20$
A_3	$n=20$	$n=40$

Note that the number of subjects in each B_2 cell is twice the number in the B_1 cell of the same row. There are, of course, many ways of achieving proportionality of the cell frequencies. The reasons for designing the experiment so that equality or proportionality of cell sizes exists are computational and conceptual. If the sample sizes are disproportional, the design is said to be nonorthogonal. Essentially, this means that the various sources of variability in the design are not independent of each other. When the sample sizes are proportional, the design is said to be orthogonal because various sum of squares (i.e., SS_A, SS_B, $SS_{A\times B}$, and SS_{wc}) are all independent of each other. That is, the sum of squares for any one source of variability in the experiment has nothing to do with the sum of squares for any other source. Suppose that the following sums of squares have been computed for a given set of data:

$$SS_A = 400$$
$$SS_B = 200$$
$$SS_{A\times B} = 225$$
$$SS_{WC} = 500$$

If a constant value of 50, for example, is added to the score of each subject who falls in one of the levels of factor A, the sum of squares for factor A will change. But the other sum of squares will *not* change if the design is proportional. If it is disproportional, the B sum of squares can be affected by the factor A effect. Hence, the sources of variability are not independent if the design involves disproportionality. This problem occurs if the conventional procedure for the computation of the sum of squares is followed. An interpretation issue develops if one factor affects the sum of squares for another factor. If the sum of squares are not independent, it is not possible to unambiguously describe the effects of a particular factor because the factors are confounded.

Many procedures have been developed to cope with this problem. Before modern computational routines were developed, some statisticians recommended randomly discarding observations from cells to achieve proportionality; others suggested that values be made up (estimated) and added to certain cells to achieve proportionality. An argument can be made, however, that to make up data is unethical and that to throw it away is immoral! There are other alternatives.

Although the controversy concerning how to analyze the nonorthogonal design has been running for decades, there is still a lack of agreement. However, the number of credible approaches has been boiled down to essentially two or three. These approaches are usually labeled by the manner in which the sums of squares for the factors are computed. The terms "type I sum of squares," "type II sum of squares," and "type III sum of squares" appear as options for the analysis of two-factor designs in the better software packages. The first and third options are the most frequently recommended and the choice between them creates the major remaining controversy.

Those who favor the type I approach argue that ANOVA is no different than any other modeling problem and that one should allow the data to determine the model. The analysis begins by testing for interaction. If the interaction is not statistically significant, the terms used to estimate it are removed from the model. Then the sum of squares for factor A is estimated; next, the sum of squares for factor B conditional on factor A (i.e., SS $B|A$) is estimated. But, because the choice of the label A or B for a given factor is arbitrary, one also runs the analysis by estimating the SS for factor B first, followed with the estimation of SS $A|B$. The preferred sequence is sometimes based on theory regarding the nature of the variables under study. Tests are then carried out on the main effects terms for the preferred sequence. The main advantage of this approach is that power for main effect tests is increased by the removal of the interaction term when it is not necessary. But the criterion for removal is statistical significance; this is an issue because the power for interaction is relatively low. This implies that small samples are likely to lead to models with no interaction term even if interaction is present.

I prefer the type III approach for experimental research. Consider the properties of the tests for A, B, and interaction in an orthogonal design. Recall that the sums of squares for all factors are orthogonal in this case. This means that there is no concern that the effects of one factor are contaminated by any other factor. That is, SS_A, SS_B, and $SS_{A\times B}$ are completely independent of each other. This is as it should be because the questions of interest in an experiment are separate questions. The desire is to obtain an answer for each question that is unconfounded by the other terms in the model *whether the design is orthogonal or not*. (The specific question answered using the type I approach changes when the design is nonorthogonal.)

This implies that the quest should be to find an analysis that provides answers to the independent questions of interest to the researcher. This leads to the type III solution because it provides tests of the same hypotheses as are tested using an orthogonal design even with large departures from orthogonality. There is a penalty in terms of lower power on the main effect tests if interaction is not present, but if imbalance is slight (as is usually the case in experimental research) this problem is minor. Unbiased effect estimates should be a higher priority than power.

There is a role for other methods in the case of nonexperimental research when there is interest in estimating parameters of existing finite populations that differ in size (as in sample survey work). But experimental research typically does not involve interest in generalizing results to populations of different sizes. Details of computing the type III two-factor ANCOVA is presented next.

24.2 ANCOVA MODEL AND COMPUTATION FOR TWO-FACTOR DESIGNS

The application of covariance analysis to the two-factor independent sample design serves the same purpose as in the case of one-factor designs. That is, power is increased and means are adjusted for chance pretreatment differences measured by

the covariate. The hypotheses associated with the two-factor ANCOVA are

$$H_0 : \mu_{A_1\,\text{adj}} = \mu_{A_2\,\text{adj}} = \cdots = \mu_{A_I\,\text{adj}}$$

$$H_0 : \mu_{B_1\,\text{adj}} = \mu_{B_2\,\text{adj}} = \cdots = \mu_{B_J\,\text{adj}}$$

H_0 : Adjusted Simple effects of factor A are consistent across all levels of factor B.

As is the case with two-factor ANOVA, these hypotheses are equivalent to stating that A, B, and $A \times B$ effects in the two-factor ANCOVA model are equal to zero.

The Model

The two-factor ANCOVA model is

$$Y_{ijk} = \mu + \alpha_i + \gamma_j + \alpha\gamma_{ij} + \beta_1 \left(X_{ijk} - \bar{X}..\right) + \varepsilon_{ijk},$$

where

μ = overall population mean
α_i = effect of ith level of factor A
γ_j = effect of jth level of factor B
$\alpha\gamma_{ij}$ = interaction effect of ith level of factor A and jth level of factor B
$\beta_1(X_{ijk} - \bar{X}..)$ = effect of covaritate
ε_{ijk} = error component associate with kth observation in ith level of factor A and jth level of factor B

In the case of a completely randomized two-factor design, the adjusted population means and the unadjusted population means are equal when there are no treatment effects, and the null hypotheses associated with ANCOVA and ANOVA are equivalent.

Suppose in the previously described two-factor experiment where the two factors were type of reinforcement and type of creativity training program, that the experimenter decides it would be useful to control for verbal fluency. If the dependent variable is the number of novel responses produced in a specific period of time, it could be argued that much of the variability in the production of such responses should be predictable from a measure of verbal fluency. Differences among subjects in verbal fluency will contribute to the within-cell error variance. If a measure of verbal fluency can be obtained before the treatments are applied, the power of the tests on all effects will be increased because within-group variability on the number of novel responses produced (Y) will be decreased by information contained in the verbal fluency measure (X).

Computation of Type III Analysis for Orthogonal and Nonorthogonal Designs

The procedure described in this section is appropriate for both orthogonal and nonorthogonal designs; it applies to both ANOVA and ANCOVA. Table 24.3 contains

Table 24.3 Data for a 2 × 2 Independent Sample ANCOVA

		Factor B			
		B_1		B_2	
Factor A	A_1	X	Y	X	Y
		2	1	1	4
		5	8	7	7
		3	2	3	5
		9	9	6	8
	A_2	X	Y	X	Y
		3	2	9	6
		4	2	1	3
		7	6	3	4
		8	5	5	4

data from an independent sample experiment that has two levels of factor A and two levels of factor B and a covariate X. This is described as a 2 × 2 independent sample ANCOVA design. The predictor variables required for the appropriate regression analysis of these data is provided in Table 24.4. It can be seen that the first three columns contain dummy variables. For two-factor designs, we employ the procedure described next to construct the dummy variables.

Table 24.4 Predictor Variables for Computation of Two-Factor ANCOVA

	(1) d_{a_1}	(2) d_{b_1}	(3) $d_{a_1}d_{b_1}$	(4) X	(5) $d_{a_1}X$	(6) $d_{b_1}X$	(7) $d_{a_1}d_{b_1}X$	Y
Cell A_1B_1	1	1	1	2	2	2	2	1
	1	1	1	5	5	5	5	8
	1	1	1	3	3	3	3	2
	1	1	1	9	9	9	9	9
Cell A_1B_2	1	−1	−1	1	1	−1	−1	4
	1	−1	−1	7	7	−7	−7	7
	1	−1	−1	3	3	−3	−3	5
	1	−1	−1	6	6	−6	−6	8
Cell A_2B_1	−1	1	−1	3	−3	3	−3	2
	−1	1	−1	4	−4	4	−4	2
	−1	1	−1	7	−7	7	−7	6
	−1	1	−1	8	−8	8	−8	5
Cell A_2B_2	−1	−1	1	9	−9	−9	9	6
	−1	−1	1	1	−1	−1	1	3
	−1	−1	1	3	−3	−3	3	4
	−1	−1	1	5	−5	−5	5	4

First, the total number of dummy variables required in a two-factor design is equal to the number of cells minus one. The number of cells is equal to the number of levels on factor A times the number of levels on factor B. Hence, a 2×2 design contains four cells and requires three dummy variables; a 2×3 design requires five dummy variables, and so on. The sum of the number of dummy variables associated with factor A, factor B, and the $A \times B$ interaction is equal to the total number required. If there are I levels of factor A, there must be $I - 1$ dummy variables to identify the levels of factor A. If there are J levels of factor B, there must be $J - 1$ dummy variables to identify the levels of factor B. The $A \times B$ interaction requires $(I - 1)(J - 1)$ dummy variables. It can be seen, then, that $I - 1$, $J - 1$, and $(I - 1)(J - 1)$ add up to $(IJ) - 1$ dummy variables. The form of the dummy variables required for the analysis is not the same as was employed in Chapter 7 for the one-factor design. The dummy variables differ in that the possible values are 1, -1, and 0 rather than 1 and 0. The rules for assigning these values are quite straightforward.

In each dummy-variable column each member of the cell being identified is assigned the value 1; those subjects who fall in the last cell are assigned the value -1 and all others, the value 0. If there are only two levels of a factor, the subjects in the first level are assigned the value 1 and those in the second level the value -1; there is no "zero" in the dummy variable column because in the case of two levels the subjects are members of either the first or the last level. A few examples should clarify the procedure.

Consider the data in Table 24.4. Because this is a 2×2 design the total number of dummy variables required is $(2 \times 2) - 1$, or 3. The first dummy variable is labeled d_{a_1} because this variable will identify which subjects are members of level A_1 and which subjects are members of level A_2. The subjects in level A_1 are assigned 1s, and those in level A_2 are assigned -1s. This dummy variable can be seen in the first column of Table 24.4. Note that the first eight subjects are members of level A_1, and the last eight subjects are members of level A_2. The second dummy variable column (d_{b_1}) is used to identify members of levels B_1 and B_2 using 1s and -1s, respectively. The third dummy variable column $(d_{a_1}d_{b_1})$ is the $A \times B$ interaction column. The values in this column are obtained by computing the products of the values in columns d_{a_1} and d_{b_1}.

Covariate values are entered in the fourth column (X), and the next three columns are simply the products of the dummy variables and the covariate. That is, the values in d_{a_1} times the values in X yield the values in $d_{a_1}X$. The corresponding situation holds for $d_{b_1}X$ and $d_{a_1}d_{b_1}X$. The last column contains scores on the dependent variable.

Examples of dummy variables for 2×3 and 3×4 designs can be seen in Table 24.5. (The columns for the products of the dummy variables and covariate have been omitted from the 3×4 matrix to save space.) The notation employed with the dummy variables is as follows:

$D_a = all\ I - 1$ dummy variables required to identify group membership of levels of factor A

$D_b = all\ J - 1$ dummy variables required to identify group membership in levels of factor B

$D_{a\times b} = all\ (I - 1)(J - 1)$ product dummy variables

Table 24.5 Predictor Variables for 2 × 3 and 3 × 4 ANCOVA Designs

	2 × 3 Design Matrix											
Cell	d_{a_1}	d_{b_1}	d_{b_2}	$d_{a_1}d_{b_1}$	$d_{a_1}d_{b_2}$	X	$d_{a_1}X$	$d_{b_1}X$	$d_{b_2}X$	$d_{a_1}d_{b_1}X$	$d_{a_1}d_{b_2}X$	Y
A_1B_1	1	1	0	1	0							
A_1B_2	1	0	1	0	1							
A_1B_3	1	−1	−1	−1	−1							
A_2B_1	−1	1	0	−1	0							
A_2B_2	−1	0	1	0	−1							
A_2B_3	−1	1	−1	1	1							

	3 × 4 Design Matrix												
Cell	d_{a_1}	d_{a_2}	d_{b_1}	d_{b_2}	d_{b_3}	$d_{a_1}d_{b_1}$	$d_{a_1}d_{b_2}$	$d_{a_1}d_{b_3}$	$d_{a_2}d_{b_1}$	$d_{a_2}d_{b_2}$	$d_{a_2}d_{b_3}$	X	Y
A_1B_1	1	0	1	0	0	1	1	0	0	0	0		
A_1B_2	1	0	0	1	0	0	0	0	0	0	0		
A_1B_3	1	0	0	0	1	0	1	1	0	0	0		
A_1B_4	1	0	−1	−1	−1	−1	−1	−1	0	0	0		
A_2B_1	0	1	1	0	0	0	0	0	1	0	0		
A_2B_2	0	1	0	1	0	0	0	0	0	1	0		
A_2B_3	0	1	0	0	1	0	0	0	0	0	1		
A_2B_4	0	1	−1	−1	−1	0	0	0	−1	−1	−1		
A_4B_1	−1	−1	1	0	0	−1	0	0	−1	0	0		
A_4B_2	−1	−1	0	1	0	0	−1	0	0	−1	0		
A_4B_3	−1	−1	0	0	1	0	0	−1	0	0	1		
A_4B_4	−1	−1	−1	−1	−1	1	1	1	1	1	−1		

d_{a_1} = label attached to first dummy variable associated with factor A (if there are more than two levels of factor A, the labels attached to the $I-1$ dummy variables are $d_{a_1}, d_{a_2}, d_{a_3}, \ldots, d_{a_{I-1}}$)
d_{b_1} = label attached to first dummy variable associated with factor B (if there are more than two levels of factor B, labels attached to remaining dummy variables are $d_{b_2}, d_{b_3}, \ldots, d_{b_{J-1}}$)
$d_{a_1}d_{b_1}$ = label attached to product of d_{a_1} and d_{b_1}, similar labels are attached to the additional product dummy variables required when there are more than two levels either factor

Once the predictor columns have been constructed, the two-factor ANOVA, ANCOVA, and homogeneity of regression tests can easily be carried out with the aid of a multiple linear regression computer program. The procedures for carrying out each of these analyses are described next.

Required Regressions for ANOVA

For two-factor ANOVA, the required R^2 values are as follows:

1. $R^2_{yD_a,D_b,D_{a\times b}}$. This is the squared multiple correlation coefficient associated with the regression of the dependent variable (Y) on all dummy variables used to identify the levels of factor A (D_a), all dummy variables used to identify the levels of factor B (D_b), and all interaction dummy variables ($D_{a\times b}$). This

coefficient yields the proportion of the total sum of squares that is accounted for by factor A, factor B, and the $A \times B$ interaction.

2. $R^2_{yD_b,D_{a\times b}}$. This is the squared multiple correlation coefficient associated with the regression of the dependent variable (Y) on all dummy variables used to identify the levels of factor B (D_b) and all interaction dummy variables ($D_{a\times b}$). This coefficient yields the proportion of the total sum of squares that is accounted for by factor B and the $A \times B$ interaction.
3. $R^2_{yD_a,D_{a\times b}}$. This is the squared multiple correlation coefficient associated with the regression of the dependent variable (Y) on all dummy variables used to identify the levels of factor A (D_a) and all interaction dummy variables ($D_{a\times b}$). This coefficient yields the proportion of the total sum of squares that is accounted for by factor A and the $A \times B$ interaction.
4. $R^2_{yD_a,D_b}$. This is the squared multiple correlation coefficient associated with the regression of the dependent variable (Y) on all dummy variables used to identify the levels of factor A (D_a) and all dummy variables used to identify the levels of factor B (D_b). This coefficient yields the proportion of the total sum of squares that is accounted for by factors A and B.
5. These R^2 values are then employed to compute the required sum of squares for the type III analysis. Table 24.6 provides the format for the summary of the analysis. The first line of the summary is for factor A. Note that the R^2 based on all A, B, and $A \times B$ dummy variables is the first coefficient and that the R^2 based on B and $A \times B$ dummy variables is subtracted from the former coefficient. The only reason for a difference between these two R^2 values can be seen in the subscripts. If the R^2 based on A, B, and $A \times B$ is larger than the R^2 based on B and $A \times B$, the difference is explained by factor A. Hence, the difference between the two R^2 values yields the proportion of Y that is accounted for by factor A independent of factor B and the $A \times B$ interaction. If this difference is multiplied by the total sum of squares, the sum of squares for factor A is obtained.

Table 24.6 Summary Table for Two-factor Independent Sample ANOVA for Orthogonal or Nonorthogonal Designs

Source	SS	*df*	MS	*F*
Factor A	SST $(R^2_{yD_a,D_b,D_{a\times b}} - R^2_{yD_b,D_{a\times b}})$	$I - 1$	$\mathrm{SS}_A/(I-1)$	$\mathrm{MS}_A/\mathrm{MS}_{\mathrm{WC}}$
Factor B	SST $(R^2_{yD_a,D_b,D_{a\times b}} - R^2_{yD_a,D_{a\times b}})$	$J - 1$	$SS_B/(J-1)$	$\mathrm{MS}_B/\mathrm{MS}_{\mathrm{WC}}$
$A \times B$ Interaction	SST $(R^2_{yD_a,D_b,D_{a\times b}} - R^2_{yD_a,D_b})$	$(I-1)(J-1)$	$\mathrm{SS}_{A\times B}/[(I-1)(J-1)]$	$\mathrm{MS}_{A\times B}/\mathrm{MS}_{\mathrm{WC}}$
Within cell	SST $(1 - R^2_{yD_a,D_b,D_{a\times b}})$	$N - IJ$	$\mathrm{SS}_{\mathrm{WC}}/(N-IJ)$	
Total	(SST)	$N - 1$		

Cell means are based on the following regression equation:
$\bar{Y}_{ij} = b_0 + b_1(d_1) + b_2(d_2) + \cdots + b_{IJ-1}(d_{IJ-1})$

The second line of the summary table follows the same pattern as the first. The left-hand R^2 is based on A, B, and $A \times B$ dummy variables. The right-hand R^2 is based on A and $A \times B$ dummy variables. The difference between the two coefficients describes the proportion of Y that is accounted for by factor B independent of factor A and the $A \times B$ interaction. This difference times the total sum of squares yields the sum of squares for factor B.

The third line of the summary table follows the same pattern as the first. Once again, the left-hand R^2 is based on A, B, and $A \times B$ dummy variables. The right-hand R^2 is based on A and B dummy variables. The difference between the two R^2 values can be explained by the difference between the sets of dummy variables associated with each value. Because the right-hand R^2 is based on A and B whereas the left-hand R^2 is based on A, B, and $A \times B$, the $A \times B$ interaction must account for the difference. This difference is the proportion of the total sum of squares that is accounted for by the $A \times B$ interaction independent of the A and B main effects. The product of the total sum of squares times the difference between these two R^2 values yields the $A \times B$ interaction sum of squares.

The labels for first three lines in the table are sometimes embellished to clarify the nature of the sums of squares. That is, the first line may be labeled as $A|B$, $A \times B$ using more complete notation to make it clear that the SS for factor A is conditional on B and $A \times B$. Similarly, the second line (for factor B) is sometimes labeled as $B|A$, $A \times B$, and the third line for (the interaction) may be denoted as $A \times B|A$, B. This more cumbersome notation is not used in Table 24.6, but it should be understood that the sum of squares for each effect is conditional on the other factors in the design.

The next line involves the subtraction of the R^2 based on A, B, and $A \times B$ dummy variables from one. Because the R^2 value describes the proportion of the total sum of squares that is accounted for by all the systematic sources in this design (i.e., factors A, B, and $A \times B$ interaction), the difference between this value and one yields the proportion that is not systematic (i.e., error). The product of this difference times the total sum of squares yields the within cell sum of squares.

If the design is orthogonal, the sum for the first four lines will be equal to the total sum of squares. If the design is nonorthogonal, the first four lines will not add up to the total sum of squares. In either case the F-tests are appropriate for the purpose of testing the following hypotheses:

$$H_0 : \mu_{A_1} = \mu_{A_2} = \cdots = \mu_{A_I}$$

$$H_0 : \mu_{B_1} = \mu_{B_2} = \cdots = \mu_{B_I}$$

H_0 : Simple effects of A are consistent at all levels of B

The regression equation associated with $R^2_{yD_a,D_b,D_{a\times b}}$ is the regression of Y on all dummy variables associated with A, B, and $A \times B$; this equation is shown at the bottom of Table 24.6. Note that it can be used to compute the cell means. When the dummy values for cell i,j are entered into the equation, the mean of cell i,j is the value predicted. It can be shown that the intercept b_0 is the grand (unweighted) mean of the cells and that each partial regression coefficient b_i is the effect (deviation of

the cell mean from the grand unweighted mean) of membership in the cell associated with the dummy variable. For example, the effect of membership in the first level of factor A is equal to the first partial regression coefficient. If there are only two levels of factor A, the effect of membership in the second level is *minus* the value of the first partial regression coefficient. If there are three levels of A, the first two partial regression coefficients are equal to the effects of levels A_1 and A_2, respectively. The effect of membership in A_3 (in the case of three levels of A) is *minus* the sum of the first two partial regression coefficients. This method of interpretation of these coefficients holds for all levels and factors.

ANOVA Example

The design matrix provided in Table 24.4 was employed to compute ANOVA on the Y variable. The required R^2 values are

$$R^2_{yD_a,D_b,D_{a\times b}} = R^2_{y1,2,3} = 0.12921$$
$$R^2_{yD_b,D_{a\times b}} = R^2_{y2,3} = 0.02809$$
$$R^2_{yD_a,D_{a\times b}} = R^2_{y1,3} = 0.10393$$
$$R^2_{yD_a,D_b} = R^2_{y1,2} = 0.12640$$

The total sum of squares is 89.00. The analysis is summarized in Table 24.7. The regression equation associated with $R^2_{y1,2,3} = R^2_{yD_a,D_b,D_{a\times b}}$ is

$$\hat{\mathrm{Y}} = 4.75 + 0.75(d_1) - 0.375(d_3) - 0.125(d_3).$$

The cell means are

$$\bar{Y}\,A_1B_1 = 4.75 + 0.75(1) - 0.375(1) - 0.125(1) = 5$$
$$\bar{Y}\,A_1B_2 = 4.75 + 0.75(1) - 0.375(-1) - 0.125(-1) = 6$$
$$\bar{Y}\,A_2B_1 = 4.75 + 0.75(-1) - 0.375(1) - 0.125(-1) = 3.75$$
$$\bar{Y}\,A_2B_2 = 4.75 + 0.75(-1) - 0.375(-1) - 0.125(1) = 4.25$$

Table 24.7 ANOVA (on Y) Summary for Data of Tables 24.3 and 24.4

Source	SS	*df*	MS	*F*
A	$89(0.12921 - 0.02809) = 9.00$	1	9.00	1.39
B	$89(0.12921 - 0.10393) = 2.25$	1	2.25	.35
$A \times B$	$89(0.12921 - 0.12640) = 0.25$	1	0.25	.05
Within cell	$89(1 - 0.12921) = 77.50$	12	6.46	
Total	89	15		

Table 24.8 Multiple Comparison Error Term Formulas for Comparisons of Marginal Means in Two-Factor Independent Sample ANOVA[a]

Procedure	Error Term	Critical Value
Fisher–Hayter	$\sqrt{\frac{MS_{\text{wc}}}{2}\left[\frac{1}{n_{m_i}}+\frac{1}{n_{m_j}}\right]}$	Studentized range $q_{(\alpha,L-1,N-IJ)}$
Tukey–Kramer	$\sqrt{\frac{MS_{\text{wc}}}{2}\left[\frac{1}{n_{m_i}}+\frac{1}{n_{m_j}}\right]}$	Studentized range $q_{(\alpha,L,N-IJ)}$
Bonferroni	$\sqrt{MS_{\text{WC}}\left[\frac{C_1^2}{n_{m_1}}+\frac{C_2^2}{n_{m_2}}+\cdots+\frac{C_1^2}{n_{m_L}}\right]}$	Bonferroni $t_B(\alpha, C', N-IJ)$
Scheffé	(Same formula as Bonferroni)	$\sqrt{(L-1)F_{(\alpha,L-1,N-IJ)}}$

[a] *Notation*:

$C' =$ Number of planned comparisons
$n_{m_i}, n_{m_j} =$ Number of subjects associated with ith and jth marginal means involved in comparison
$L =$ Number of levels associated with factor involved in comparison; it will be equal to either I (number of levels of factor A) or J (number of levels of factor B)
All α values refer to nondirectional tests.

The procedure used for this ANOVA is appropriate for both orthogonal and nonorthogonal designs and for any number of levels of the two factors. It should be mentioned, however, that in the case of two levels of each factor (i.e., a 2×2 design), it is not necessary to compute all the regressions employed in this general procedure. If the computer program provides a test of the significance of each partial regression coefficient, the analysis is complete in one pass. That is, when Y is regressed on d_{a_1}, d_{b_1}, and $d_{a_1 \times b_1}$, the tests of significance of the three associated partial regression coefficients are equivalent to the three F-tests in the ANOVA summary table. Tests on partial regression coefficients are usually t values; it is the squared t values that are equal to the ANOVA F values.

Intuitively, it is reasonable that these tests should be equivalent. Partial regression coefficients are measures of the independent effects of the variables in the equation. Likewise, the two-factor ANOVA tests the independent effects of factors A, B, and $A \times B$ interaction. Because the dummy variables are independent of each other across factors, and since each one is an indicator of group membership, the equivalence should, at least, not seem bizarre. When more than two levels are associated with factors A or B, there will probably be interest in multiple comparison tests or simultaneous confidence intervals for the factor or factors that have three or more marginal means. The formulas in Table 24.8 provide the appropriate error terms for these tests or intervals.

Computation Procedure for Two-factor ANCOVA

The two-factor ANCOVA hypotheses of equal adjusted marginal means for factors A and B, no interaction of $A \times B$, and the homogeneity of regression can be tested

using the computation described in the following paragraphs. As before with other designs, the procedure shown relies on the use of a multiple regression computer program.

The first step involves the construction of the design matrix using $1, 0, -1$ dummy variables. This step has been described earlier in this section. The second step involves the computation of the following squared multiple correlation coefficients:

1. $R^2_{yD_a,D_b,D_{a\times b},X,D_aX,D_bX,D_{a\times b}X}$. This is the coefficient associated with the regression of Y on all dummy variables used to identify levels of factor A, all dummy variables used to identify levels of factor B, all dummy variables used to identify the $A \times B$ interaction, the covariate(s), and the product of each of these dummy variables with the covariate. This R^2 value is interpreted as the proportion of the total sum of squares that is accounted for by factor A, factor B, the $A \times B$ interaction, the covariate(s), and the interaction of the covariate and the treatments.
2. $R^2_{yD_a,D_b,D_{a\times b},X}$. This is the coefficient associated with the regression of Y on all dummy variables used to identify levels of factor A, all dummy variables used to identify levels of factor B, the $A \times B$ interaction dummy variables, and the covariate(s). This R^2 value is interpreted as the proportion of the total sum of squares that is accounted for by factor A, factor B, the $A \times B$ interaction, and the covariate(s).
3. $R^2_{yD_b,D_{a\times b},X}$. This is the coefficient associated with the regression of Y on all dummy variables used to identify levels of factor B, the $A \times B$ interaction dummy variables, and the covariate(s). This R^2 value is interpreted as the proportion of the total sum of squares that is accounted for by factor B, the $A \times B$ interaction, and the covariate(s).
4. $R^2_{yD_a,D_{a\times b},X}$. This is the coefficient associated with the regression of Y on all dummy variables used to identify levels of factor A, the $A \times B$ interaction dummy variables, and the covariate. This R^2 value is interpreted as the proportion of the total sum of squares that is accounted for by factor A, the $A \times B$ interaction, and the covariate(s).
5. $R^2_{yD_a,D_b,X}$. This is the coefficient associated with the regression of Y on all dummy variables used to identify levels of factor A, levels of factor B, and the covariate(s). This R^2 value is interpreted as the proportion of the total sum of squares that is accounted for by factor A, factor B, and the covariate(s).

These five R^2 values are employed in the summary tables for the two-factor ANCOVA and the homogeneity of regression slopes test. Both summaries can be seen in Table 24.9.

The rationale for employing the R^2 values shown in Table 24.9 for the ANCOVA and homogeneity of regression tests is as follows. The adjusted effects of factor A are the effects of membership in the different levels of factor A that are independent of factor B, the $A \times B$ interaction, and the covariate(s). The difference between the R^2 value that is based on all variables that represent A, B, $A \times B$, and X and the R^2 value

Table 24.9 Summary Tables for Independent Sample Two-Factor ANCOVA and Homogeneity of Within-Cell Regression[a]

ANCOVA				
Source	SS	*df*	MS	*F*
Adjusted A	$\text{SST}\,(R^2_{yD_a,D_b,D_{a\times b},X} - R^2_{yD_b,D_{a\times b},X})$	$I-1$	$SS_{A\,\text{adj}}/(I-1)$	$\text{MS}_{A\,\text{adj}}/\text{MS}_{\text{Res}_{\text{wc}}}$
Adjusted B	$\text{SST}\,(R^2_{yD_a,D_b,D_{a\times b},X} - R^2_{yD_b,D_{a\times b},X})$	$J-1$	$\text{SS}_{B\,\text{adj}}/(J-1)$	$\text{MS}_{B\,\text{adj}}/\text{MS}_{\text{Res}_{\text{wc}}}$
Adjusted $A \times B$	$\text{SST}\,(R^2_{yD_a,D_b,D_{a\times b},X} - R^2_{yD_a,D_b,X})$	$(I-1)(J-1)$	$\text{SS}_{A\times B_{\text{adj}}}/[(I-1)(J-1)]$	$\text{MS}_{A\times B\,\text{adj}}/\text{MS}_{\text{Res}_{\text{wc}}}$
Res_{WC}	$\text{SST}\,(1 - R^2_{yD_a,D_b,D_{a\times b},X})$	$N-IJ-C$	$\text{SS}_{\text{Res}_{\text{WC}}}/(N-I)(J-C)$	
Res_t	$\text{SST}\,(1 - R^2_{yX})$	$N-C-1$		
Homogeneity of Within-Cell Regression Test				
Hetero. reg.	$\text{SST}\,(R^2_{yD_a,D_b,D_{a\times b},X,D_aX,D_bX,D_{a\times b}X} - R^2_{yD_a,D_b,D_{a\times b},X})$	$C(IJ-1)$	$\text{SS}_{\text{HR}}/[C(IJ-1)]$	$\text{MS}_{\text{HR}}/\text{MS}_{\text{Res}_i}$
Res_i	$\text{SST}\,(1 - R^2_{yD_a,D_b,D_{a\times b},X,D_aX,D_bX,D_{a\times b}X})$	$N-[IJ(C+1)]$	$\text{SS}_{\text{Res}_i}/N-[IJ(C-1)]$	
Res_{WC}	$\text{SST}\,(1 - R^2_{yD_a,D_b,D_{a\times b},X})$	$N-IJ-C$		

[a] Adjusted cell means are based on the regression equation associated with $R^2_{yD_a,D_b,D_{a\times b},X}$

$\bar{Y}_{ij\,\text{adj}} = b_0 + b_1(d_1) + b_2(d_2) + \cdots + b_{IJ-1}(d_{IJ-1}) + b_{X_1}(\bar{X}_1..) + b_{X_2}(\bar{X}_2..) + \cdots + b_{X_C}(\bar{X}_C..)$

where

$d_1 - d_{IJ-1} = IJ - 1$ Dummy variables.

$b_1 - b_{IJ-1} =$ Partial regression coefficients associated with the $IJ-1$ dummy variables.

$b_{X_1} - b_{X_C} =$ Partial regression coefficients associated with covariates 1 through C.

based on all these variables except those representing levels of factor A, is explained by the elimination of factor A variables from the equation. The value $R^2_{yD_a,D_b,D_{a\times b},X}$ will exceed $R^2_{yD_b,D_{a\times b},X}$ only if there are some differences among levels of A that are independent of the effects of B, $A \times B$, and X. Hence, the first line of the ANCOVA summary includes the difference between these two R^2 values. Because this difference is the proportion of the total sum of squares that is independently accounted for by factor A, the product of the total sum of squares times this difference yields the adjusted factor A sum of squares.

The same approach is taken in the second line of the summary. The left-hand R^2 is based on A, B, $A \times B$, and X, but the right-hand R^2 is based on A, $A \times B$, and X. The total sum of squares times the difference between these two R^2 values yields the sum of squares accounted for by factor B that is independent of the effects of A, $A \times B$, and X.

The adjusted $A \times B$ sum of squares is based on the same approach. Note in the third line of the summary that, once again, the left-hand R^2 is based on $A, B, A \times B$, and X. The right-hand R^2 is based on A, B, and X. The difference between the two R^2 values is the proportion of the total sum of squares that is accounted for by the $A \times B$ interaction independently of the effects of factor A, factor B, and the covariate(s). The product of the total sum of squares times this difference between R^2 values yields the adjusted $A \times B$ interaction sum of squares.

Under the two-factor ANCOVA model all variability not explained by $A, B, A \times B$, and X is termed *error* or *residual variability*. Since $R^2_{yD_a,D_b,D_{a\times b},X}$ is the proportion of the total variability accounted for by factor A, factor B, the $A \times B$ interaction, and X, one minus this value must be the proportion of the total sum of squares that is not explained by these sources of variability. The product of the total sum of squares times $(1 - R^2_{yD_a,D_b,D_{a\times b},X})$ yields the residual within-cell sum of squares that is entered on the fourth line of the summary. The last line of the summary contains the total residual sum of squares that is all variability not accounted for by X. The total sum of squares times $(1 - R^2_{yX})$ yields this sum of squares.

Homogeneity of Regression Test

A diagrammatic representation of the partitioning just described can be seen in the upper part of Figure 24.2. The lower part of this figure illustrates the manner in which the sum of squares residual within cells is further partitioned into heterogeneity of regression sum of squares and individual residual sum of squares. This additional partitioning is required for the homogeneity of within-cell regression test that is shown in the lower part of Table 24.9. The rationale for the partitioning is the same as in the case of a one-factor design.

It is assumed under the two-factor ANCOVA model that the regression slope (or plane or hyperplane in the case of multiple covariates) of Y on X is the same within each level of A for tests on A, within each level of B for tests on B, and within each cell for tests on $A \times B$ interaction. If these assumptions are not met, the most likely consequences are that the F-tests for adjusted A, B, and $A \times B$ effects

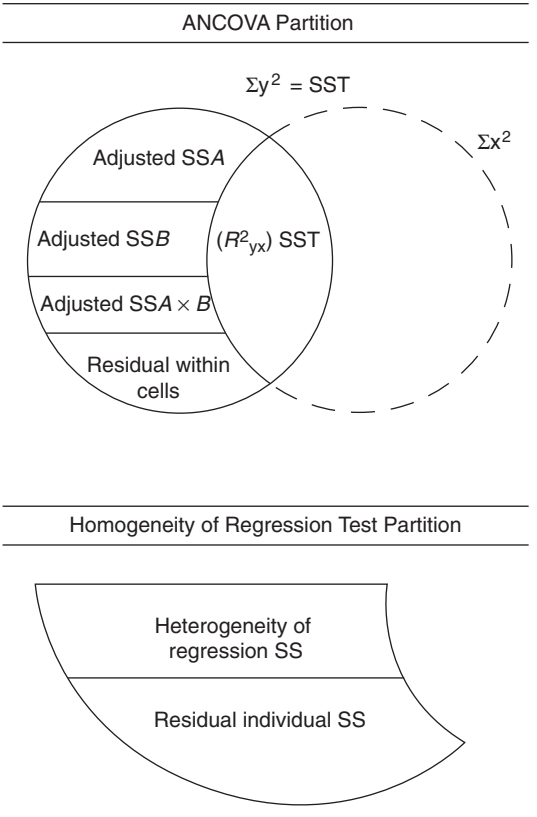


Figure 24.2 ANCOVA partitioning associated with independent cell sample two-factor designs.

will be conservatively biased and the adjusted means will be difficult to interpret. Typically, the approach to evaluating these assumptions is to test the homogeneity of the within-cell regressions. The method of performing this test involves first computing the within-cell residual sum of squares using the pooled within-cell slope. As can be seen in Table 24.9, this sum of squares is the error or within-cell residual sum of squares used in the two-factor ANCOVA. The second step is to compute the residual sum of squares within cells based on individual within-cell regressions rather than on the pooled within-cell regression. This can be accomplished by multiplying the total sum of squares times $(1 - R^2_{yD_a,D_b,D_{a\times b},X,D_aX,D_bX,D_{a\times b}X})$. The difference between the pooled within-cell residuals and the individual within-cell residuals reflects heterogeneity of the regression slopes. A somewhat more direct method of computing the heterogeneity of regression sum of squares is to compute the product of the total sum of squares times the difference between the two R^2 values shown in last two lines of Table 24.9. The mean-square heterogeneity of regression over the mean-square individual within-cell residual provides the F-test for the homogeneity of within-cell regression. If this F-test yields a nonsignificant F value, the assumption

of homogeneous within-cell regression slopes is retained, and the adjusted-cell and marginal means and the ANCOVA F-tests can be interpreted in a straightforward manner. If the homogeneity of regression slopes test yields a significant F value, the methods of Chapter 11 should be considered.

The computational formula for the adjusted-cell means is shown below the homogeneity of within-cell regression test in Table 24.9. Note that the equation is the one associated with the regression of Y on all dummy variables and all covariates (i.e., the equation associated with $R^2_{yD_a,D_b,D_{a\times b},X}$). An example follows.

Example 24.1 The predictors listed in Table 24.4 were employed to compute ANCOVA and the homogeneity of regression tests that are summarized in Table 24.10. The R^2 values that are required for this analysis are as follows (both general notation and specific design matrix column labels used with this data set are shown):

$$R^2_{yD_a,D_b,D_{a\times b},X,D_aX,D_bX,D_{a\times b}X} = R^2_{y1,2,3,4,5,6,7} = 0.84842$$
$$R^2_{yD_a,D_b,D_{a\times b},X} = R^2_{y1,2,3,4} = 0.72369$$
$$R^2_{yD_b,D_{a\times b},X} = R^2_{y2,3,4} = 0.57010$$
$$R^2_{yD_a,D_{a\times b},X} = R^2_{y1,3,4} = 0.65092$$
$$R^2_{yD_a,D_b,X} = R^2_{y1,2,4} = 0.72346$$

Table 24.10 Two-Factor Independent Sample ANCOVA and Homogeneity of Within-Group Regression Test for Data of Table 24.4

Source	SS	*df*	MS	*F*
	ANCOVA			
Adjusted A	$89(0.72369 - 0.57010) = 13.67$	1	13.67	6.11
Adjusted B	$89(0.72369 - 0.65092) = 6.48$	1	6.48	2.90
Adjusted $A \times B$	$89(0.72369 - 0.72346) = 0.02$	1	0.02	0.01
Res_{WC}	$89(1 - 0.72369) = 24.59$	11	2.235	
Res_t	$89(1 - 0.49995) = 44.50$[a]			
Critical value $= F_{(.05,1,11)} = 4.84$				
	Homogeneity of Within-Cell Regression Test			
Hetero. reg.	$89(0.84842 - 0.72369) = 11.10$	3	3.70	2.19
Res_i	$89(1 - 0.84842) = 13.49$	8	1.69	
Res_{WC}	$89(1 - 0.72369) = 24.59$	11		
Critical value $= F_{(.05,3,8)} = 4.07$				

[a] This value deviates slightly from the sum of the four values above it as a result of rounding error.

The total sum of squares (SST) is 89.00 and R^2_{yX}, which is not essential in the analysis, is 0.49995.

Regression equation associated with $R^2_{yD_a,D_b,D_{a\times b},X}$

$$\hat{Y} = 1.3539 + 0.9287\left(d_{a_1}\right) - 0.6431\left(d_{b_1}\right) - 0.0356\left(d_{a_1}d_{b_1}\right) + 0.7150(X)$$

Grand unweighted mean (i.e., mean of cell means) on $X = 4.75$; adjusted cell means:

$$\bar{Y}_{A_1B_1} = 1.3539 + 0.9287(1) - 0.6431(1) - 0.0356(1) + 0.7150(4.75) = 5.00$$
$$\bar{Y}_{A_1B_2} = 1.3539 + 0.9287(1) - 0.6431(-1) - 0.0356(-1) + 0.7150(4.75) = 6.36$$
$$\bar{Y}_{A_2B_1} = 1.3539 + 0.9287(-1) - 0.6431(1) - 0.0356(-1) + 0.7150(4.75) = 3.21$$
$$\bar{Y}_{A_2B_2} = 1.3539 + 0.9287(-1) - 0.6431(-1) - 0.0356(1) + 0.7150(4.75) = 4.43$$

		Factor B		Adjusted Marginal A Means
		B_1	B_2	
Factor A	A_1	5.00	6.36	$\bar{Y}_{A_1\,\text{abj}} = 5.68$
	A_2	3.21	4.43	$\bar{Y}_{A_2\,\text{abj}} = 3.82$
Adjusted Marginal B Means		$\bar{Y}_{B_1\,\text{abj}} = 4.11$	$\bar{Y}_{B_2\,\text{adj}} = 5.39$	

The adjusted marginal means in Table 24.10 were computed by averaging the adjusted-cell means associated with each factor level. Alternatively, they can be computed directly from the regression equation used to compute the cell means. When this approach is used, all terms in the equation are ignored except the intercept, the partial regression coefficient associated with the dummy variable(s) associated with the factor for which the adjusted marginal mean is being computed, and the partial regression coefficient associated with the covariate(s). The grand unweighted mean on Y is equal to the intercept plus the product of partial regression coefficient for X times the grand unweighted mean on X. The deviation from the Y unweighted mean due to the effect of level A_1 is equal to the partial regression coefficient associated with level A_1. The effect of level A_2 is minus the partial regression coefficient associated with A_1. The same approach applies to the main effects on factor B.

For the example data, the regression equation has the following coefficients:

$$b_0 = 1.3539$$
$$b_1 = 0.9287$$
$$b_2 = -0.6431$$
$$b_3 = -0.0356$$
$$b_4 = 0.7150$$

The grand unweighted mean on X is 4.75. The grand unweighted mean on Y is $b_0 + b_4\,(4.75) = 4.75$. The main effect of level A_1 is b_1. The main effect of level A_2 is $-b_1$; hence, the adjusted marginal means for factor A are

$$\bar{Y}_{A_{1\,\text{abj}}} = 4.75 + 0.93 = 5.68.$$
$$\bar{Y}_{A_{2\,\text{abj}}} = 4.75 - 0.93 = 3.82$$

Likewise, the adjusted marginal means for factor B are

$$\bar{Y}_{B_{1\,\text{abj}}} = 4.75 + (-0.64) = 4.11.$$
$$\bar{Y}_{B_{2\,\text{abj}}} = 4.75 - (-0.64) = 5.39$$

At this point it may be helpful to compare the results of ANOVA and ANCOVA on the example data. Both analyses are summarized in Table 24.11. Note that type of reinforcement, type of program, and the interaction of the two are nonsignificant in the case of ANOVA. With ANCOVA, the type of reinforcement has a statistically significant effect. The type of program and the interaction of reinforcement type and program type are nonsignificant.

The reasons ANOVA and ANCOVA reach different conclusions on the effect of type of reinforcement (factor A) can be seen by comparing the size of the error terms associated with the two models and by comparing the size of the effects being tested. The error mean square for ANOVA (the within-cell MS) is 6.46, whereas the error MS for ANCOVA (the within-cell MS residual) is only 2.24. Also, the difference between the marginal means is smaller than the difference between adjusted marginal means. The larger difference between adjusted marginal means is explained by the fact that the level with the lower covariate mean (A_1) had the higher dependent variable mean. Recall that the adjustment process works in such a way that when there is a positive relationship between the covariate and the dependent variable, the Y mean is adjusted upward for groups that fall below the X mean and downward for groups that fall above the X mean. In the example, the unadjusted mean difference is 1.5 points and the adjusted difference is 1.86 points. The overall result of the much smaller error term and the larger mean difference associated with ANCOVA is a much larger F value.

Computational Simplification for Two-Factor Independent Sample Designs with Two Levels of Each Factor: Any Number of Covariates

When a two-factor design contains only two levels of each factor, the general computation procedure described in Table 24.9 is appropriate, but there is a very efficient shortcut that can be employed. This shortcut is applicable to orthogonal and nonorthogonal designs with any number of covariates. If the design matrix is set up using the approach illustrated in Table 24.4, the regression of Y on the dummy variables associated with A, B, $A \times B$, and the covariate(s), is all that is required for the tests on the adjusted A, B, and $A \times B$ effects. The test on the first partial regression coefficient is equivalent to the test for adjusted A effects described in Table 24.9.

Table 24.11 Comparison of Two-factor Independent Sample ANOVA and ANCOVA for Data in Table 24.3

ANOVA		ANCOVA	
Source	F	Source	F
Factor A (type of reinforcement)	1.39	Adjusted A	6.11
Factor B (type of program)	0.35	Adjusted B	2.90
$A \times B$ Interaction (error MS = 6.46)	0.05	Adjusted $A \times B$ (error MS = 2.24)	0.01
Critical value of F (for $\alpha = .05$) = 4.75		Critical value of F (for $\alpha = .05$) = 4.84	

	B_1	B_2	
A_1	$\bar{Y} = 5.00$	$\bar{Y} = 6.00$	$\bar{Y}_{A_1} = 5.50$
A_2	$\bar{Y} = 3.75$	$\bar{Y} = 4.25$	$\bar{Y}_{A_2} = 4.00$ } ANOVA
	$\bar{Y}_{B_1} = 4.38$	$\bar{Y}_{B_2} = 5.12$	

	B_1	B_2	
A_1	$\bar{Y}_{\text{adj}} = 5.00$	$\bar{Y}_{\text{adj}} = 6.36$	$\bar{Y}_{A_1\,\text{adj}} = 5.68$
A_2	$\bar{Y}_{\text{adj}} = 3.21$	$\bar{Y}_{\text{adj}} = 4.43$	$\bar{Y}_{A_2\,\text{adj}} = 3.82$ } ANCOVA
	$\bar{Y}_{B_1\,\text{adj}} = 4.11$	$\bar{Y}_{B_2\,\text{adj}} = 5.39$	$\bar{X}_{A_1} = 4.50$
			$\bar{X}_{A_2} = 5.00$ }

	B_1	B_2
A_1	$\bar{X} = 4.75$	$\bar{X} = 4.25$
A_2	$\bar{X} = 5.50$	$\bar{X} = 4.50$
	$\bar{X}_{B_1} = 5.12$	$\bar{X}_{B_2} = 4.38$

The tests of significance of the second and third partial regression coefficients are equivalent to the tests of adjusted B and $A \times B$ effects, respectively.

Confidence Intervals for Two-Factor Designs with Two Levels on One Factor

In previous chapters I expressed a preference for confidence intervals over significance tests. That preference holds for two-factor designs. When a factor consists of only two levels, the following formula is appropriate:

$$\left(\bar{Y}_{\mathrm{m}_1\,\mathrm{adj}} - \bar{Y}_{\mathrm{m}_2\,\mathrm{adj}}\right) \pm \sqrt{\mathrm{MS}_{\mathrm{Res}_{\mathrm{wc}}}\left[\frac{1}{n_{\mathrm{m}_1}} + \frac{1}{n_{\mathrm{m}_2}} + \frac{\left(\bar{X}_{\mathrm{m}_1} - \bar{X}_{\mathrm{m}_2}\right)^2}{\mathrm{SS}_{\mathrm{WC_X}}}\right]}(t_{(\alpha, N-IJ-1)}),$$

where

$\bar{Y}_{\mathrm{m}_1\,\mathrm{adj}}, \bar{Y}_{\mathrm{m}_2\,\mathrm{adj}}$ = adjusted marginal means

$\bar{X}_{\mathrm{m}_1}, \bar{X}_{\mathrm{m}_2}$ = marginal means on covariate associated with same factor levels as $\bar{Y}_{\mathrm{m}_1\,\mathrm{adj}}$ and $\bar{Y}_{\mathrm{m}_2\,\mathrm{adj}}$

$n_{\mathrm{m}_1}, n_{\mathrm{m}_2}$ = number of subjects associated with marginal means 1 and 2

$\mathrm{MS}_{\mathrm{Res}_{\mathrm{wc}}}$ = mean square residual within cell associated with two-factor ANCOVA

$\mathrm{SS}_{\mathrm{wc_X}}$ = sum of squares within cell associated with a two-factor ANOVA on X [which can be computed by using $\mathrm{SST_X}(1 - R^2_{xD_aD_bD_{a\times b}})$ where $\mathrm{SST_X}$ is total sum of squares on covariate and $R^2_{xD_aD_bD_{a\times b}}$ is squared multiple correlation coefficeint associated with regression of on X dummy variables indicating levels of A, B and $A \times B$]

$t_{(\alpha, N-IJ-1)}$ = critical value of Student's t based on N (total number of subjects) minus IJ (total number of cells) minus one degree of freedom

This formula can be applied to the example data described previously in this section. The required terms can be found in Tables 24.10 and 24.11 except for $\mathrm{SS}_{\mathrm{wc_X}}$. This value is equal to

$$\mathrm{SST_X}\left(1 - R_{xD_a,D_b,D_{a\times b}}\right) = 107(1 - 0.03271) = 103.50.$$

The 95% confidence interval associated with the difference between the two levels of the factor A adjusted marginal means is

$$\begin{aligned}\left(\bar{Y}_{A_1\,\mathrm{abj}} - \bar{Y}_{A_2\,\mathrm{abj}}\right) &\pm \sqrt{2.235\left[\frac{1}{8} + \frac{1}{8} + \frac{(4.50 - 5.00)^2}{103.50}\right]}(2.201) \\ &= (5.68 - 3.82) \pm .751(2.201) = (.21, 3.51)\end{aligned}$$

The 95% confidence interval associated with the difference between the two levels of the factor B adjusted marginal means is

$$\left(\bar{Y}_{B_{1\,\text{abj}}} - \bar{Y}_{B_{2\,\text{abj}}}\right) \pm \sqrt{2.235\left[\frac{1}{8} + \frac{1}{8} + \frac{(5.12 - 4.38)^2}{103.50}\right]}(2.201)$$

$$= (4.11 - 5.39) \pm .755(2.201) = (-2.94, .38)$$

24.3 MULTIPLE COMPARISON TESTS FOR ADJUSTED MARGINAL MEANS

When more than two levels exist for a factor, multiple comparison procedures may be of interest for the analysis of specific contrasts among the marginal means. The formulas presented in Table 24.12 provide the appropriate error terms for multiple comparison hypothesis tests and simultaneous confidence intervals for the case of one covariate. If more than one covariate is involved, the formulas in Table 24.13 are appropriate. The terms in the formulas contained in Tables 24.12 and 24.13 are defined as follows:

$\text{MS}_{\text{Res}_{\text{wc}}}$ = within-cell residual MS that is the error term in two-factor ANCOVA

$n_{\text{m}_i}, n_{\text{m}_i}$ = sample sizes associated with ith and jth marginal means that are being compared

$n_{\text{m}_1}, n_{\text{m}_L}$ = sample sizes associated with each level of L

L = number of levels of factor being analyzed (equal to either I or J)

$\bar{X}_{\text{m}_i}, \bar{X}_{\text{m}_i}$ = marginal means on covariate for ith and jth levels of factor being analyzed

$\text{SS}_{\text{wc}_\text{X}}$ = within-cell sum of squares from two-factor ANOVA on covariate

$c_1 - c_L$ = contrast coefficients associated with levels 1 through L of factor being analyzed

N = total number of subjects

I, J = number of levels of factors A and B, respectively

C' = number of planned comparisons

C = number of covariates

$\mathbf{d}$ = column vector of differences between ith and jth marginal means on covariates

$\mathbf{d}'$ = transpose of $\mathbf{d}$

$W_{\text{wc}_\text{x}}^{-1}$ = inverse of within-cell sum of products matrix for covariates

Table 24.12 Multiple Comparison Formulas for Two-Factor Independent Sample Designs with One Covariate

Procedure	Error Term	Critical Value
Fisher–Hayter	$\sqrt{\dfrac{\mathrm{MS}_{\mathrm{Res_{wc}}}\left[\dfrac{1}{n_{m_1}}+\dfrac{1}{n_{m_j}}+\dfrac{(\bar{X}_{m_i}-\bar{X}_{m_j})^2}{\mathrm{SS}_{xy}}\right]}{2}}$	Studentized range $q_{(\alpha,L-1,N-IJ-1)}$
Tukey–Kramer	$\sqrt{\dfrac{\mathrm{MS}_{\mathrm{Res_{wc}}}\left[\dfrac{1}{n_{m_i}}+\dfrac{1}{n_{m_j}}+\dfrac{(\bar{X}_{m_i}-\bar{X}_{m_j})^2}{\mathrm{SS}_{\mathrm{wc}}}\right]}{2}}$	Studentized range $q_{(\alpha,L,N-IJ-1)}$
Bonferroni	$\sqrt{\mathrm{MS}_{\mathrm{Res_{wc}}}\left[\dfrac{c_1^2}{n_{\mathrm{m}_1}}+\dfrac{c_2^2}{n_{\mathrm{m}_2}}+\cdots+\dfrac{c_L^2}{n_{\mathrm{m}_L}}+\dfrac{(c_1\bar{X}_{\mathrm{m}_1}+c_2\bar{X}_{\mathrm{m}_2}+\cdots+c_L\bar{X}_{\mathrm{m}_L})^2}{SS_{\mathrm{wc_x}}}\right]}$	Bonferroni $t_{(\alpha,C',N-IJ-1)}$
Scheffé	(Same as Bonferroni)	$\sqrt{(L-1)F_{(\alpha,L-1,N-IJ-1)}}$

Table 24.13 Multiple Comparison Formulas for Two-Factor Independent Sample Designs with Any Number of Covariates

Procedure	Error Term	Critical Value
Fisher–Hayter	$\sqrt{\dfrac{\mathrm{MS}_{\mathrm{Res}_{\mathrm{wc}}}\left[\dfrac{1}{n_{\mathrm{cm}_i}}+\dfrac{1}{n_{\mathrm{cm}_j}}+\mathbf{d}'\mathbf{W}_{\mathrm{wc_x}}^{-1}\mathbf{d}\right]}{2}}$	Studentized range $q_{(\alpha,L-1,N-IJ-C)}$
Tukey–Kramer	$\sqrt{\dfrac{\mathrm{MS}_{\mathrm{Res}_{\mathrm{wc}}}\left[\dfrac{1}{n_{\mathrm{cm}_i}}+\dfrac{1}{n_{\mathrm{cm}_j}}+\mathbf{d}'\mathbf{W}_{\mathrm{wc_x}}^{-1}\mathbf{d}\right]}{2}}$	Studentized range $q_{(\alpha,L,N-IJ-C)}$
Bonferroni	$\sqrt{\mathrm{MSres}_{\mathrm{wc}}\left[\dfrac{c_1^2}{n_{\mathrm{m}_1}}+\dfrac{c_2^2}{n_{\mathrm{m}_2}}+\cdots+\dfrac{c_L^2}{n_{\mathrm{m}_L}}+\mathbf{d}'\mathbf{W}_{\mathrm{wc_x}}^{-1}\mathbf{d}\right]}$	Bonferroni $t_{(\alpha,C',N-IJ-C)}$
Scheffé	(Same formula as Bonferroni)	$\sqrt{(L-1)F_{(\alpha,L-1,N-IJ-1)}}$

An example of the term **d** (column vector) is as follows. If an investigator is interested in comparing marginal means 1 and 2 on factor A, the **d** vector is

$$\begin{array}{l} \\ \text{Covariate 1} \\ \text{Covariate 2} \\ \quad . \quad . \\ \quad . \quad . \\ \quad . \quad . \\ \text{Covariate } C \end{array} \begin{array}{c} \text{Level } A_1 \\ \left[\begin{array}{c} \bar{X}_{\mathrm{m}_{1,1}} \\ \bar{X}_{\mathrm{m}_{2,1}} \\ . \\ . \\ . \\ \bar{X}_{\mathrm{m}_{C,1}} \end{array}\right. \end{array} \begin{array}{c} - \\ \\ . \\ . \\ . \\ - \end{array} \begin{array}{c} \text{Level } A_2 \\ \left.\begin{array}{c} \bar{X}_{\mathrm{m}_{1,2}} \\ \bar{X}_{\mathrm{m}_{2,2}} \\ . \\ . \\ . \\ \bar{X}_{\mathrm{m}_{C,2}} \end{array}\right] \end{array} = \begin{array}{c} \mathbf{d} \\ \left[\begin{array}{c} d_1 \\ d_2 \\ . \\ . \\ . \\ d_C \end{array}\right] \end{array},$$

where

$\bar{X}_{\mathrm{m}_{1,1}} = A_1$ marginal mean on covariate 1
$\bar{X}_{\mathrm{m}_{1,2}} = A_2$ marginal mean on covariate 1
$d_1 = (\bar{X}_{\mathrm{m}_{1,1}} - \bar{X}_{\mathrm{m}_{1,2}})$, and so on

Simple Main Effects

The main effects associated with the two-factor design are no more difficult to interpret than are the adjusted means in a one-factor design as long as the $A \times B$ interaction is small or nonsignificant. If the interaction is significant, the interpretation of the main effects becomes ambiguous because the effects of A are not consistent across the different levels of B. In this case the main effects of A (or B) are still representative of the overall differences due to factor A (or B) but "overall" differences, which are reflected in the marginal means, are generally of little interest if the interaction is large. In this case the conventional two-factor ANCOVA should be supplemented

with "simple main effects" tests (or simple main effect confidence intervals). "Simple main effects" refer to cell means, whereas main effects refer to marginal means.

The simple main effects tests associated with the 2 × 2 ANCOVA design refer to tests on the following:

1. Differences between adjusted cell means A_1 and A_2 at level B_1
2. Differences between adjusted cell means A_1 and A_2 at level B_2
3. Differences between adjusted cell means B_1 and B_2 at level A_1
4. Differences between adjusted cell means B_1 and B_2 at level A_2

A convenient way of carrying out these tests is to employ the following formula for simple main effects comparisons associated with a given factor:

$$\frac{\bar{Y}_{\text{cm}_i\,\text{adj}} - \bar{Y}_{\text{cm}_j\,\text{adj}}}{\sqrt{\text{MS}_{\text{Res}_{\text{wc}}}\left[\frac{1}{n_{\text{cm}_i}} + \frac{1}{n_{\text{cm}_j}} + \frac{\left(\bar{X}_{\text{cm}_i} - \bar{X}_{\text{cm}_j}\right)^2}{\text{SS}_{\text{wc}_x}}\right]}} = t,$$

where the subscript "cm" denotes cell mean. The obtained t value is compared with the critical value of the t or the critical value of the Bonferroni t, depending upon whether the per-comparison or the family error rate is preferred. If there is interest in controlling the family error rate at α, the Bonferroni statistic $t_{B(\alpha,C',N-IJ-1)}$, where α is the same level as is employed in the overall main effects tests and C' is the number of simple main effect tests associated with the factor being analyzed. The denominator of this t formula can be employed in the construction of simultaneous confidence intervals for the simple main effect comparisons. That is, if we denote the error term of the t described above as $S_{\bar{Y}_{\text{cm}_i\,\text{adj}} - \bar{Y}_{\text{cm}_j\,\text{adj}}}$, the simultaneous confidence intervals are constructed by using

$$\left(\bar{Y}_{\text{cm}_i\,\text{adj}} - \bar{Y}_{\text{cm}_j\,\text{adj}}\right) \pm S_{\bar{Y}_{\text{cm}_i\,\text{adj}} - \bar{Y}_{\text{cm}_j\,\text{adj}}}\left[t_{B(\alpha,C,N-IJ-1)}\right].$$

This procedure is employed below with the example data originally presented in Table 24.3 to illustrate the computations. Note, however, that simple main effect procedures are not required for these data because the interaction is not significant. Only the simple main effect comparisons associated with factor A are carried out in the paragraphs that follow.

The simple main effect comparisons for factor A involve the adjusted cell means difference between A_1 and A_2 at B_1 and at B_2. It can be seen in Table 24.11 that the difference at B_1 is $(5.00 - 3.21) = 1.79$ and at B_2 $(6.36 - 4.43) = 1.93$. The confidence intervals associated with these two simple main effects comparisons are

$$(1.79) \pm \sqrt{2.235\left[\frac{1}{4} + \frac{1}{4} + \frac{(4.75 - 5.50)^2}{103.50}\right]}\left[t_{\text{B}_{(\alpha,C',N-IJ-1)}}\right]$$
$$= (1.79) \pm 1.063(2.593) = (-.97, 4.55)$$

$$(1.93) \pm \sqrt{2.235\left[\frac{1}{4}+\frac{1}{4}+\frac{(4.25-4.50)^2}{103.50}\right]}\left[t_{\mathrm{DB}_{(a,C',N-IJ-1)}}\right]$$
$$= (1.93) \pm 1.085(2.593) = (-.81, 4.67).$$

The probability is at least .95 so that the differences $(\mu_{A_{1\,\mathrm{adj}}} - \mu_{A_{2\,\mathrm{adj}}})$ at B_1 and $(\mu_{A_{1\,\mathrm{adj}}} - \mu_{A_{2\,\mathrm{adj}}})$ at B_2 are included in these intervals. You may note that both of these intervals are wider than the interval that was constructed for the adjusted marginal mean difference between A_1 and A_2. There are three reasons for the width of the simple effect intervals to differ from the width of the overall (marginal) interval. First, there is more sampling error associated with the adjusted cell means than with the adjusted marginal means because the cell values are based on fewer observations. Second, the differences between the covariate cell means (which are involved in the error term for simple main effects) are not the same as the difference between the marginal covariate means. Third, two statements are made in the case of the simple effects; only one is made in the case of the overall main effects. The Bonferroni approach controls α for the whole collection (two in this case) of simple main effect comparisons. The more comparisons there are, the wider the confidence intervals will be. The probability is at least $1 - \alpha$ that the whole collection of intervals will span the true differences between the adjusted cell means included in the simple main effect comparisons.

If multiple covariates are employed, the following formula for simple main effect comparisons is appropriate:

$$\frac{\bar{Y}_{\mathrm{cm}_{i\,\mathrm{adj}}} - \bar{Y}_{\mathrm{cm}_{j\,\mathrm{adj}}}}{\sqrt{\mathrm{MS}_{\mathrm{Res}_{\mathrm{wc}}}\left[\frac{1}{n_{\mathrm{cm}_i}}+\frac{1}{n_{\mathrm{cm}_j}}+\mathbf{d}'\mathbf{W}_{\mathrm{wc_X}}^{-1}\mathbf{d}\right]}} = t,$$

where the terms are defined as before and the obtained t value is compared with $t_{B(\alpha,C',N-IJ-1)}$.

Minitab Input and Output for Analysis of Data in Table 24.3

The ANCOVA described above can be computed using just a few commands in *Minitab*. The worksheet for the two-factor analysis requires just four columns: a column to identify levels of factor A, a column to identify levels of factor B, the covariate column, and the dependent variable column. The program automatically constructs the required dummy variable columns that are described earlier in this section; these remain invisible when using the routine illustrated below. The columns for analyzing the data in Table 24.3 are shown below (ignore the "row" column):

```
Row  A  B  X  Y
 1   1  1  2  1
 2   1  1  5  8
 3   1  1  3  2
 4   1  1  9  9
```

```
 5   1  2  1  4
 6   1  2  7  7
 7   1  2  3  5
 8   1  2  6  8
 9   2  1  3  2
10   2  1  4  2
11   2  1  7  6
12   2  1  8  5
13   2  2  9  6
14   2  2  1  3
15   2  2  3  4
16   2  2  5  4
```

The commands for the two-factor ANCOVA are

```
MTB > GLM  'Y' = A B A*B;
SUBC>   Covariates 'X';
SUBC>   Brief 1 ;
SUBC>   Means A B A*B.
```

Output:

The two-factor ANCOVA and the adjusted marginal and cell means are shown in boldface below.

```
General Linear Model: Y versus A, B
Factor  Type   Levels  Values
A       fixed       2  1, 2
B       fixed       2  1, 2
Analysis of Variance for Y, using Adjusted SS for Tests
Source  DF  Seq SS  Adj SS  Adj MS      F      P
X        1  44.495  52.908  52.908  23.67  0.000
A        1  13.410  13.669  13.669   6.11  0.031
B        1   6.482   6.477   6.477   2.90  0.117
A*B      1   0.020   0.020   0.020   0.01  0.926
Error   11  24.592  24.592   2.236
Total   15  89.000


S = 1.49520   R-Sq = 72.37%   R-Sq(adj) = 62.32%
Means for Covariates
Covariate   Mean  StDev
X          4.750  2.671


Least Squares Means for Y
A     Mean  SE Mean
1    5.679   0.5299
2    3.821   0.5299
```

```
B
1     4.107    0.5315
2     5.393    0.5315
A*B
1 1   5.000    0.7476
1 2   6.357    0.7512
2 1   3.214    0.7557
2 2   4.429    0.7485
```

Homogeneity of Regression Slopes Test

Recall that the computation of the homogeneity of regression slopes test requires fitting both the ANCOVA model and a model that fits separate slopes within each cell. The latter is fitted using the following commands:

```
MTB > GLM  'Y' = A B A*B A*X B*X A*B*X;

SUBC>   Covariates 'X';

SUBC>   Brief 1.
```

The output is as follows:

```
General Linear Model: Y versus A, B
Factor  Type   Levels  Values
A       fixed       2  1, 2
B       fixed       2  1, 2
Analysis of Variance for Y, using Adjusted SS for Tests
Source  DF  Seq SS  Adj SS  Adj MS      F      P
X        1  44.495  52.356  52.356  31.05  0.001
A        1  13.410   0.112   0.112   0.07  0.804
B        1   6.482  10.684  10.684   6.34  0.036
A*B      1   0.020   0.111   0.111   0.07  0.804
A*X      1   4.830   2.518   2.518   1.49  0.257
B*X      1   6.169   6.070   6.070   3.60  0.094
A*B*X    1   0.102   0.102   0.102   0.06  0.812
Error    8  13.491  13.491   1.686
Total   15  89.000
S = 1.29861   R-Sq = 84.84%   R-Sq(adj) = 71.58%
```

The error sum of squares from this model is the individual residual sum of squares (SS_{Res_i}). The error sum of squares from the two-factor ANCOVA is SS_{Res_w} and the difference between the two is the heterogeneity of regression slopes sum of squares. Hence, $(24.592 - 13.491) = 11.101$. The test statistic is $\frac{SS_{HR}/df_{HR}}{MS_{Res_i}} = \frac{11.101/3}{1.686} = 2.19 = F$. The p-value (based on $C(IJ - 1) = 3$ and $N - [IJ(C + 1)] = 8$ degrees of freedom is .17. It is concluded that heterogeneity of regression has not been demonstrated to be a concern in this experiment.

24.4 TWO-FACTOR ANOVA AND ANCOVA FOR REPEATED-MEASUREMENT DESIGNS

The two-factor design of the previous section involves the use of independent subjects in each cell. The design of this section involves the use of independent subjects in the various levels of factor A, but the same subjects are used in the various levels of factor B. This design is sometimes called a *split-plot* design or a *two-factor design with one repeated-measurement factor*. This design involves aspects of both the one-factor randomized-group design and the repeated-measurement one-factor design. The randomized-group aspect is involved in the random assignment of S subjects to the different treatment levels of factor A. In some varieties of this design, however, the assignment to the levels of A may not be random. Rather, subjects may be selected from different existing populations. For example, A_1 may be male subjects and A_2 female subjects. Hence, the levels of factor A may be under experimental control, or they may simply be classification levels. Factor B, the repeated-measurement factor, may involve levels of several types. If the levels of factor B are different treatments, the order in which a subject is exposed to the various treatments is randomly assigned for each subject independently. If the levels of factor B are different trials or observations of the same behavior at J different time points, there is no randomization for this factor.

An advantage of this design over the independent sample two-factor design is that the number of subjects required is smaller. Also, because each subject is observed under all levels of factor B, the size of the error term for testing the effects of this factor is generally smaller than with independent sample designs. It must be kept in mind, however, that the results with respect to factor B are generalized to a population of subjects who have been exposed to repeated-measurement conditions. Additional details on repeated measurement designs can be found in Keppel and Wickens (2004), Kirk (1995), Howell (2010), and Maxwell and Delaney (2004).

Suppose that a researcher has interest in evaluating the effects of two drugs and three conditions under which the drugs are administered, on anxiety. Hence, the two factors are A, drug type; and B, administration situation. The dependent variable is a scale of situational anxiety. Ten subjects are randomly assigned to the two drug levels A_1 and A_2. Then the order of administration of the three levels of factor B is randomly assigned for each subject independently. Before the experiment is initiated, the investigator notes apparently large differences in general (trait) anxiety among the subjects. His or her concern is that individual differences in anxiety level (that can be measured by using a general anxiety scale) will have an effect on the way subjects respond on the dependent variable. Hence, a measure of general anxiety is obtained from each subject before the experiment is started. This measure is then used as a covariate in the analysis after the experiment is completed. Contrived data appear in Table 24.14. Note that there are five subjects under each drug type. The five subjects associated with A_1 are not the same subjects associated with level A_2. The covariate score associated with each subject is shown in the X column, the three scores on the dependent variable (Y) under the three levels of B are in the next three columns, and the sum of the three Y scores for each subject is entered in the T column. The predictor variables required for the analysis are shown in Table 24.15. The

Table 24.14 Example Data for a Two-Factor Design with One Repeated-Measurement Factor and One Covariate

Factor *A*		Factor B, Administration Situation				
			B_1	B_2	B_3	
Drug Type		X	Y	Y	Y	T
	S_1	6	9	6	7	22
	S_2	5	6	7	7	20
A_1	S_3	4	8	5	6	19
	S_4	7	9	8	8	25
	S_5	4	7	5	6	18
	S_1	9	8	7	6	21
	S_2	6	4	3	4	11
A_2	S_3	3	3	3	2	8
	S_4	5	4	4	3	11
	S_5	6	3	4	4	11

$\{1, 0, -1\}$ dummy variables are constructed in exactly the same way as was described in the previous section for the two-factor independent sample design. The "ones" in the first column (d_{a_1}) identify observations associated with level A_1; the "minus ones" in this column identify observations associated with level A_2. Because there are two levels of factor A only one dummy variable is required for this factor (i.e., the number of dummy variables required for factor $A = I - 1 = 2 - 1 = 1$). Factor B has three levels; two dummy variables are required for this factor because the required number is the number of levels of B minus one (i.e., $J - 1 = 3 - 1 = 2$). A "one" is entered in column 2 (d_{b_1}) for each observation that is associated with level B_1. A "minus one" is entered for each observation associated with the last level (i.e., B_3). "Zero" is entered for observations that are associated with neither B_1 nor B_3 (i.e., B_2). The third column (d_{b_2}) contains "ones" to identify observations associated with level B_2, "minus ones" to identify those associated with B_3, and "zeros" to identify observations associated with neither B_2 nor B_3. The next two columns ($d_{a_1}d_{b_1}$ and $d_{a_1}d_{b_2}$) are product columns. Column 6 is the covariate column, and column 7 is the product of the d_{a_1} and X columns. Column 8 is the T column; each entry in this column is the total of the three Y scores across the three levels of B. Note that each covariate score and each T score appears three times for each subject. This occurs because each subject contributes three scores to the Y column; one score is obtained under each level of factor B. Because X and T scores for a given subject are the same regardless of the level of B, each X and T score must appear as many times as there are levels of B. Dependent variable scores are entered in the last column.

Unlike the predictor set associated with the two-factor independent sample design, the predictor set for this design does not contain the products of the dummy variables used to identify levels of B and X. Note that there are no columns labeled $d_{b_1}X$ or $d_{b_2}X$. This is because only factor A is adjusted by the covariate with this design. The

Table 24.15 Predictor Variables for 2 × 3 Two-Factor Design with One Repeated-Measurement Factor and One Covariate

Cell	(1) d_{a_1}	(2) d_{b_1}	(3) d_{b_2}	(4) $d_{a_1}d_{b_1}$	(5) $d_{a_1}d_{b_2}$	(6) X	(7) $d_{a_1}X$	(8) T	Y
	1	1	0	1	0	6	6	22	9
	1	1	0	1	0	5	5	20	6
A_1B_1	1	1	0	1	0	4	4	19	8
	1	1	0	1	0	7	7	25	9
	1	1	0	1	0	4	4	18	7
	1	0	1	0	1	6	6	22	6
	1	0	1	0	1	5	5	20	7
A_1B_2	1	0	1	0	1	4	4	19	5
	1	0	1	0	1	7	7	25	8
	1	0	1	0	1	4	4	18	5
	1	−1	−1	−1	−1	6	6	22	7
	1	−1	−1	−1	−1	5	5	20	7
A_1B_3	1	−1	−1	−1	−1	4	4	19	6
	1	−1	−1	−1	−1	7	7	25	8
	1	−1	−1	−1	−1	4	4	18	6
	−1	1	0	−1	0	9	−9	21	8
	−1	1	0	−1	0	6	−6	11	5
A_2B_1	−1	1	0	−1	0	3	−3	8	3
	−1	1	0	−1	0	5	−5	11	4
	−1	1	0	−1	0	6	−6	11	3
	−1	0	1	0	−1	9	−9	21	7
	−1	0	1	0	−1	6	−6	11	3
A_2B_2	−1	0	1	0	−1	3	−3	8	3
	−1	0	1	0	−1	5	−5	11	4
	−1	0	1	0	−1	6	−6	11	4
	−1	−1	−1	1	1	9	−9	21	6
	−1	−1	−1	1	1	6	−6	11	4
A_2B_3	−1	−1	−1	1	1	3	−3	8	2
	−1	−1	−1	1	1	5	−5	11	3
	−1	−1	−1	1	1	6	−6	11	4

covariate accounts for neither between-level nor within-level variability on factor B because the covariate score is constant for a given subject at all levels of this factor. The results of this are that (1) the covariate does not adjust the differences among the marginal factor B means, (2) the $A \times B$ interaction is not related to the covariate, and (3) the error term used to test the B and $A \times B$ effects is not related to the covariate.

A comparison of the predictors of this section with those of the previous section reveals two basic differences. Whereas there are no B and X product columns, it does contain the T column that is not found in the independent sample version. This difference is necessary because the partitioning of the total sum of squares is different with these two varieties of two-factor design. These differences can be seen in Figure 24.3.

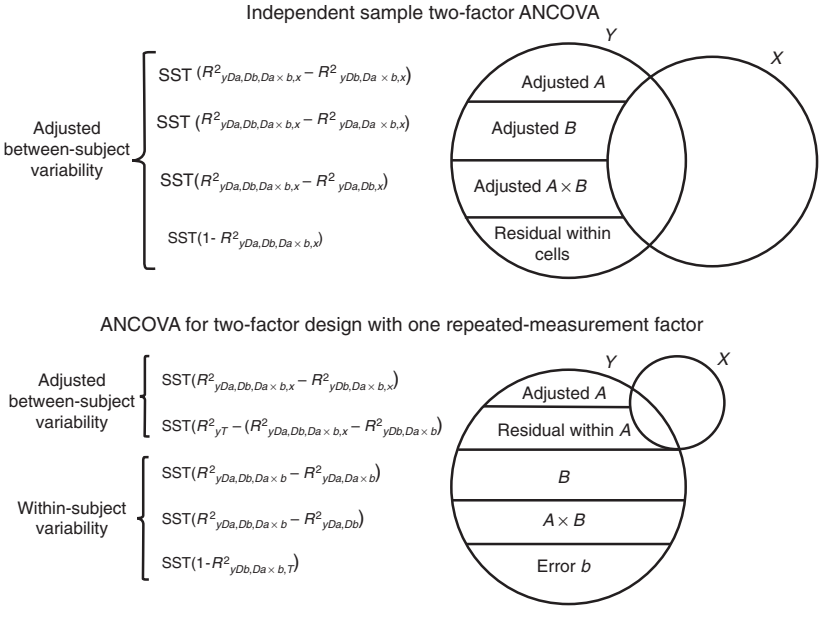


Figure 24.3 Diagrammatic representations of partitioning associated with independent samples and split-plot two-factor designs.

Note that the variability in the independent sample design is all between-subject variability. The repeated-measurement design contains variability that is partly attributed to differences between subjects and partly attributed to differences within subjects. The computation of ANOVA, ANCOVA, and homogeneity of regression tests for the repeated-measurement design requires the following R^2 values for the example data in Table 24.14.

$$
\begin{aligned}
R^2_{yD_a,D_b,D_{a\times b},X} &= R^2_{y1,2,3,4,5,6} = 0.87961 \\
R^2_{yD_b,D_{a\times b},X} &= R^2_{y2,3,4,5,6} = 0.24823 \\
R^2_{yT} &= R^2_{y8} = 0.85491 \\
R^2_{yD_a,D_b,D_{a\times b}} &= R^2_{y1,2,3,4,5} = 0.55469 \\
R^2_{yD_a,D_{a\times b}} &= R^2_{y1,4,5} = 0.51395 \\
R^2_{yD_a,D_b} &= R^2_{y1,2,3} = 0.53292 \\
R^2_{yD_b,D_{a\times b},T} &= R^2_{y2,3,4,5,8} = 0.91741 \\
R^2_{yD_a,X,D_aX} &= R^2_{y1,6,7} = 0.81716 \\
R^2_{yD_b,D_{a\times b}} &= R^2_{y2,3,4,5} = 0.06250 \\
R^2_{yD_a,X} &= R^2_{y1,6} = 0.81711
\end{aligned}
$$

The general form for the ANCOVA summary is presented in Table 24.16.

The appropriate R^2 values for the example data are entered in the ANCOVA summary (Table 24.17). The difference between the two R^2 values in the first line of the table is the proportion of the total sum of squares accounted for by factor A independently of B, $A \times B$, and X. The first R^2 in the second line provides the proportion of the total sum of squares that is accounted for by between-subject differences (i.e., 0.85491). The value subtracted from 0.85491 is 0.81711, which is the proportion of the total sum of squares accounted for by A and X independent of B and $A \times B$ (i.e., 0.87961 − 0.06250). Hence, the first two lines of the analysis (i.e., adjusted A effects and the residual within-levels of A) contain between-subject sum of squares that are independent of B, $A \times B$, and X.

The difference between the two R^2 values in line 3 is the proportion of the total sum of squares accounted for by B independent of A and $A \times B$; line 4 provides the R^2 values required to obtain the sum of squares for the $A \times B$ interaction.

The partitioning associated with B and $A \times B$ does not involve the use of the covariate in the computation of the R^2 values, but these sources of variability are still independent of all other sources of variability in the design including the covariate. The error term for the B and $A \times B$ effects (error b) is based on the sum of squares that is independent of B, $A \times B$, and all between-subject variability.

The ANOVA of this design differs from the ANCOVA only in the first two lines. The unadjusted A sum of squares is computed by using

$$\mathrm{SST}\left(R^2_{yD_a,D_b,D_{a\times b}} - R^2_{yD_b,D_{a\times b}}\right).$$

The factor A error sum of squares (error a) is computed by using $\mathrm{SST}[R^2_{y\mathrm{T}}]$ minus the unadjusted A sum of squares.

Comparison of ANOVA and ANCOVA

The difference between the general forms of ANOVA and ANCOVA for the two-factor repeated-measurement design can be seen by comparing Tables 24.16 and 24.18. Note that the marginal means for factor A are different under the two analyses but the factor B marginal means are identical. As was mentioned previously, this outcome is to be expected because the covariate adjusts the factor A means but not the factor B means.

Confidence Intervals for Adjusted Factor *A* Marginal Means

If there are only two levels of factor A (the nonrepeated factor), the confidence interval for the adjusted marginal means is computed as shown below. This formula is not appropriate if there are three or more levels of factor A and simultaneous confidence intervals are desired.

Table 24.16 General Form of ANCOVA Summary for Two-Factor Design with One Repeated-Measurement Factor[a]

Source	SS	*df*	MS	*F*
Adjusted A	SST $(R^2_{yD_a,D_b,D_{a\times b},X} - R^2_{yD_b,D_{a\times b},X})$	$I-1$	$\mathrm{SS}_{A_{\mathrm{adj}}}/(I-1)$	$\mathrm{MS}_{A\,\mathrm{adj}}/\mathrm{MS}_{\mathrm{Res_{wa}}}$
$\mathrm{Res_{wa}}$	SST $[R^2_{yT} - (R^2_{yD_a,D_b,D_{a\times b},X} - R^2_{yD_b,D_{a\times b},})]$	$S-I-1$	$\mathrm{SS}_{\mathrm{Res_{wa}}}/(S-I-1)$	$\mathrm{MS}_B/\mathrm{MS}_{\mathrm{E_b}}$
B	SST $(R^2_{yD_a,D_b,D_{a\times b}} - R^2_{yD_a,D_{a\times b}})$	$J-1$	$\mathrm{SS}_B/(J-1)$	$\mathrm{MS}_{A\times B}/\mathrm{MS}_{\mathrm{E_b}}$
$A \times B$	SST $(R^2_{yD_a,D_b,D_{a\times b}} - R^2_{yD_a,D_b})$	$(I-1)(J-1)$	$SS_{A\times B}/[(I-1)(J-1)]$	
Error b	SST $(1 - R^2_{yD_b,D_{a\times b},T})$	$\sum_{i=1}^{I}(J-1)(n_i-1)$	$\mathrm{SS}_{\mathrm{E}_b} \Big/ \sum_{i=1}^{I}(J-1)(n_i-1)$	
Res_t	SST $(1 - R^2_{yX})$	$N-1-C$		
Homogeneity of Regression Summary				
Heterogeneity	$\mathrm{SS}_{\mathrm{Res_{WA}}} - \mathrm{SS}_{\mathrm{Res}_i}$	$C(I-1)$	$\mathrm{SS}_{\mathrm{het}}/[C(I-1)]$	$\mathrm{MS}_{\mathrm{het}}/\mathrm{MS}_{\mathrm{Res}_i}$
Res_i	SST $(R^2_{yT} - R^2_{yD_a,X,D_aX})$	$S-[I(C+1)]$	$\mathrm{SS}_{\mathrm{Res}_i}/[S-[I(C+1)]]$	
Computation of Cell Means with Any Number of Levels of Factors A and B and Any Number of Covariates $\bar{Y}_{ij} = b_0 + b_1(d_1) + b_2(d_2) + \cdots + b_{IJ-1}(d_{IJ-1}) + b_{X_1}(\bar{X}..{}_1) + b_{X_2}(\bar{X}..{}_2) + \cdots + b_{X_C}(\bar{X}..{}_C)$				

[a]Notation:

$\mathrm{Res_{wa}}$ = residual within levels of A
Res_t = total residual
I = number of levels of factor A
J = number of levels of factor B
S = total number of subjects in experiment
n_i = number of subjects in ith level of factor A
N= total number of observations in experiment
C = number of covariates

Table 24.17 ANCOVA Summary for Data of Table 24.14

Source	SS	*df*	MS	*F*
Adjusted A	$119.47(0.87961 - 0.24823) = 75.43$	1	75.43	117
Res_{wa}	$119.47(0.85491 - 0.81711) = 4.52$	7	0.646	
B	$119.47(0.55469 - 0.51395) = 4.87$	2	2.44	3.95
$A \times B$	$119.47(0.55496 - 0.53292) = 2.63$	2	1.32	2.14
Error b	$119.47(1 - 0.91741) = 9.87$	16	0.617	
Res_t	$119.47(1 - 0.18573) = 97.3$	28		
Homogeneity of Regression Test				
Hetero. Regr.	$(\text{Res}_{\text{wa}} - \text{Res}_i) = 0.01$	1	0.01	0.01
Res_i	$119.47(0.85491 - 0.81716) = 4.51$	6	0.752	
Cell Means				

Regression equation associated with $R^2_{yD_a,D_b,D_{a\times b},X} = R^2_{y1,2,3,4,5,6}$:

$$\hat{Y} = 1.623 + 1.613(d_1) + 0.567(d_2) - 0.333(d_3) + 0.300(d_4) - 0.400(d_5) + 0.711(X)$$

$$\bar{Y}_{1,1\text{ adj}} = 1.623 + 1.613(1) + 0.567(1) - 0.333(0) + 0.300(1) - 0.400(0) + 0.711(5.5) = 8.01$$

$$\bar{Y}_{1,2\text{ adj}} = 1.623 + 1.613(1) + 0.567(0) - 0.333(1) + 0.300(0) - 0.400(1) + 0.711(5.5) = 6.41$$

$$\bar{Y}_{1,3\text{ adj}} = 1.623 + 1.613(1) + 0.567(-1) - 0.333(-1) + 0.300(-1) - 0.400(-1) + 0.711(5.5) = 7.01$$

$$\bar{Y}_{2,1\text{ adj}} = 1.623 + 1.613(-1) + 0.567(1) - 0.333(0) + 0.300(-1) - 0.400(0) + 0.711(5.5) = 4.19$$

$$\bar{Y}_{2,2\text{ adj}} = 1.623 + 1.613(-1) + 0.567(0) - 0.333(1) + 0.300(0) - 0.400(-1) + 0.711(5.5) = 3.99$$

$$\bar{Y}_{2,3\text{ adj}} = 1.623 + 1.613(-1) + 0.567(-1) - 0.333(-1) + 0.300(1) - 0.400(1) + 0.711(5.5) = 3.59$$

	B_1	B_2	B_3	Factor A Marginal Means
A_1	8.01	6.41	7.01	7.14
A_2	4.19	3.99	3.59	3.92
Factor B Marginal Means →	6.10	5.20	5.30	

Table 24.18 General Form of ANOVA on Factor A and Summary for Data of Table 24.14

Source	SS	*df*	MS	*F*
a. General Form—ANOVA				
A	SST $(R^2_{yD_a,D_b,D_{a\times b},} - R^2_{yD_b,D_{a\times b},})$	$I-1$	$SS_A/(I-I-1)$	$MS_A/MS_{E_{wa}}$
$Error_{WA}$	SST $[R^2_{yT} - (R^2_{yD_a,D_b,D_{a\times b},} - R^2_{yD_b,D_{a\times b},})]$	$S-1$	$SS_{Ea}/(S-1)$	
B $A \times B$ Error *b*	(Same as shown in Table 24.16 for ANCOVA)			
b. ANOVA Summary—Data of Table 24.14				
A	$119.47(0.55496 - 0.06250) = 58.80$	1	58.80	10.85
Error *A*	$119.47(0.85491 - 0.49219) = 43.33$	8	5.42	
B $A \times B$ Error *b*	(Same as shown in Table 24.17 for ANCOVA)			
Total	119.47	29		

[a]Computation of cell means is based on the regression equation associated with $R^2_{yD_a,D_b,D_{a\times b}} = R^2_{y1,2,3,4,5} : R^2_{yD_a,D_b,D_{a\times b}}$

$\hat{Y}_{ij} = b_0 + b_1(d_1) + b_2(d_2) + \cdots + b_{IJ-1}(d_{IJ-1}) = 5.533 + 1.400 + 0.567 - 0.333 + 0.300 - 0.400$

$\bar{Y}_{1,1} = 5.533 + 1.400(1) + 0.567(1) - 0.333(1) + 0.300(1) - 0.400(0) = 7.80$

$\bar{Y}_{1,2} = 5.533 + 1.400(1) + 0.567(0) - 0.333(1) + 0.300(0) - 0.400(0) = 6.20$

$\bar{Y}_{1,3} = 5.533 + 1.400(1) + 0.567(-1) - 0.333(-1) + 0.300(-1) - 0.400(-1) = 6.80$

$\bar{Y}_{2,1} = 5.533 + 1.400(-1) + 0.567(1) - 0.333(0) + 0.300(-1) - 0.400(0) = 4.40$

$\bar{Y}_{2,2} = 5.533 + 1.400(-1) + 0.567(0) - 0.333(1) + 0.300(0) - 0.400(-1) = 4.20$

$\bar{Y}_{2,3} = 5.533 + 1.400(-1) + 0.567(-1) - 0.333(-1) + 0.300(1) - 0.400(1) = 3.80$

	B_1	B_2	B_3	Factor A Marginal Means
A_1	7.80	6.2	6.8	6.93
A_2	4.4	4.2	3.8	4.13
Factor *B* Marginal Means →	6.1	5.2	5.3	

Table 24.19 Multiple Comparison Formulas for Two-factor Designs with One Repeated-Measurement Factor and One Covariate

Procedure	Error Term	Critical Value
Formulas for Factor *A*—Nonrepeated (Between Subject) Factor		
Fisher–Hayter	$\sqrt{\dfrac{\mathrm{MS}_{\mathrm{Res_{wc}}}\left[\dfrac{1}{Jn_i}+\dfrac{1}{Jn_j}+\dfrac{(\bar{X}_{m_i}-\bar{X}_{m_j})^2}{\mathrm{SS}_{\mathrm{wc_X}}}\right]}{2}}$	Studentized range $q_{(\alpha,I-1,S-I-1)}$
Tukey–Kramer	$\sqrt{\dfrac{\mathrm{MS}_{\mathrm{Res_{wc}}}\left[\dfrac{1}{Jn_i}+\dfrac{1}{Jn_j}+\dfrac{(\bar{X}_{m_i}-\bar{X}_{m_j})^2}{\mathrm{SS}_{\mathrm{wc_X}}}\right]}{2}}$	Studentized range $q_{(\alpha,I,S-I-1)}$
Bonferroni	$\sqrt{\mathrm{MS}_{\mathrm{Res_{wa}}}\left[\dfrac{c_1^2}{Jn_1}+\dfrac{c_2^2}{Jn_2}+\cdots+\dfrac{c_J^2}{Jn_1}\dfrac{(c_1\bar{X}_{\mathrm{m}_1}+c_2\bar{X}_{\mathrm{m}_2}+\cdots c_1\bar{X}_{\mathrm{m}_1})^2}{\mathrm{SS}_{\mathrm{wa_x}}}\right]}$	$t_{B(\alpha,C',S-I-1)}$
Scheffé	(Same as Bonferroni)	$F'=(I-1)\sqrt{F_{(\alpha,I-1,S-I-1)}}$
Formulas for Factor *B*—Repeated-measurement (Within-subject) Factor		
Fisher–Hayter	$\sqrt{\dfrac{\mathrm{MS}_{E_b}}{S}}$	Studentized range $q_{(\alpha,J-1,\Sigma(J-1)(n_j-1))}$
Tukey	$\sqrt{\dfrac{\mathrm{MS}_{E_b}}{S}}$	Studentized range $q_{(\alpha,J,\Sigma(J-1)(n_i-1))}$
Bonferroni	$\sqrt{\mathrm{MS}_{\mathrm{E}_b}\left[\dfrac{c_1^2}{S}+\dfrac{c_2^2}{S}+\cdots+\dfrac{c_J^2}{S}\right]}$	Bonferroni $t_{N(\alpha,C)\Sigma(J-1)(n_1-1))}$
Scheffé	(Same as Bonferroni)	$F'=(J-1)\sqrt{F_{(\alpha,J-1,\Sigma(J-1)(n_1-1))}}$

The application of the two-level formula to the example data yields the following 95% confidence interval:

$$\left(\bar{Y}_{\mathrm{m}_1\,\mathrm{adj}} - \bar{Y}_{\mathrm{m}_1\,\mathrm{adj}}\right) \pm \sqrt{\mathrm{MS}_{\mathrm{Res}_{\mathrm{wa}}}\left[\frac{1}{Jn_1} + \frac{1}{Jn_2} + \frac{\left(\bar{X}_{\mathrm{m}_1} - \bar{X}_{\mathrm{m}_2}\right)^2}{\mathrm{SS}_{\mathrm{wa}_x}}\right]}(t_{(\alpha, S-I-1)})$$

$$(6.95 - 3.72) \pm \sqrt{0.646\frac{1}{15} + \frac{1}{15} + \frac{(5.2 - 5.8)^2}{25.6}}(2.365)$$

$$(3.23) \pm 0.73 = (2.50, 3.96)$$

When three or more levels of factor A are involved, one of the other error term formulas in Table 24.19 should be used. The considerations in selecting among the different formulas are the same as those discussed in Chapter 5 for one-factor designs.

Confidence Intervals for Factor *B* Marginal Means

The simultaneous 95% confidence intervals for pairwise comparisons using the Tukey procedure are shown as follows for the example data:

$$(\bar{Y}_{m_i} - \bar{Y}_{m_j}) \pm \sqrt{\mathrm{MSE}_b/S}(q_{(\alpha, J, \Sigma(J-1)(n_1-1))})$$

$$(6.1 - 5.2) \pm \sqrt{0.617/10}(q_{.05,3,16)}) = (0.9) \pm 0.91 = (-.01, 1.81)$$

$$(6.1 - 5.2) \pm \sqrt{0.617/10}(q_{(.05,3,16)}) = (0.8) \pm 0.91 = (-0.11, 1.71)$$

$$(5.2 - 5.3) \pm \sqrt{0.617/10}(q_{(.05,3,16)}) = (-0.1) \pm 0.91 = (-1.01, 0.81)$$

Simple Main Effects—Factor A

If the $A \times B$ interaction is significant, comparisons between cell means may provide more relevant information than comparisons of marginal means. Simple main effect comparisons between levels of A at each level of B are based on the following error term and Bonferroni critical value:

Error Term	Critical Value
$\sqrt{\left[\frac{\mathrm{SS}_{\mathrm{Res_{wa}}+\mathrm{SSE}_b}}{(S-I-1)+[\Sigma(J-1)(n_i-1)]}\right]\left[\frac{1}{n_i} + \frac{1}{n_i} + \frac{(\bar{X}_{A_i} - \bar{X}_{A_i})^2}{\mathrm{SS}_{\mathrm{Wa}_x}}\right]}$	$t_{\mathrm{B}_{(\alpha, C', S-I-1+\Sigma(J-1)(n_i-1))}}$

The number of comparisons (C') associated with all simple main effect tests on factor A is computed by using $J[I(I-l)/2]$.

If multiple covariates are used, the error-term formula and the critical value are

$$\sqrt{\left[\frac{\mathrm{SS}_{\mathrm{Res}_{\mathrm{wa}}} + \mathrm{SSE}_b}{(S - I - C) + [\Sigma(J-1)(n_i - 1)]}\right]\left[\frac{1}{n_i} + \frac{1}{n_j} + \mathbf{d}'\mathbf{W}_{\mathrm{a_x}}^{-1}\mathbf{d}\right]}$$

and Bonferroni $t_{\mathrm{B}[\alpha, C', S-I-C+(\Sigma(J-1)(n_i-1))]}$

Simple Main Effects—Factor B

Simple main effect comparisons between the ith and jth levels of factor B at each level of A involve the following formulas for the error term and the critical value:

Error Term	Critical Value
$\sqrt{\mathrm{MS}_{\mathrm{E}_b}\left[\frac{1}{n_{\mathrm{cm}_i}} + \frac{1}{n_{\mathrm{cm}_j}}\right]}$	$t_{\mathrm{B}(\alpha, C', \sum(J-1)(n_i-1))}$

The number of comparisons $C' = I[J(J-1)/2]$ and n_{cm_i} and n_{cm_j} are the sample sizes associated with the ith and jth cell means being compared.

Independence of Errors Assumption

It was pointed out in Chapter 3 that the one-factor repeated-measurement design independence of error assumption should be questioned if the treatment order on the repeated factor is not determined randomly for each subject. The same issue is of concern for the repeated-measurement factor in two-factor designs. The methods of testing factor B and $A \times B$ effects described in this section will lead to positively biased F values if the independence of errors assumption is violated. Tests for departures from this assumption exist but they have very low power. Software for this design (such as *SPSS*) includes tests that accommodate violations if they occur. This is unlikely if the treatment order is determined randomly for the repeated factor.

Other Complex Designs

Covariance analysis can be applied to most of the more complex designs frequently described in the experimental design literature. Three-factor independent sample designs, for example, require no new principles. Additional dummy variables are required for the third factor and its interaction with the other factors, but the basic procedure involved in the construction of dummy variables and testing effects is unchanged. Milliken and Johnson (2002) should be consulted for the analysis of complex experimental designs not covered here.

24.5 SUMMARY

Analysis of covariance is not the only procedure for attempting to remove the effects of unwanted variability from an experiment. Other approaches include blocking and the use of subjects as their own control in repeated-measurement designs. The basic advantages of ANCOVA relative to these other approaches are that (1) the response measure is adjusted for chance differences that exist before treatments are applied, and (2) fewer steps are sometimes required in the design of the experiment because homogeneous blocks do not have to be formed.

Two two-factor designs include the independent-sample and repeated-measurement varieties. Both designs provide information on the overall (main) effects of two different independent variables or factors as well as information on the interaction of the two factors. The application of covariance analysis to the independent sample two-factor design generally results in greater power for tests on main effects and interaction. Effects are also adjusted for chance differences on the covariate that are related to Y.

The two-factor ANOVA applied to designs with one repeated-measurement factor involves between-subject comparisons on factor A and within-subject comparisons on factor B. This analysis generally has very high power on comparisons among levels of factor B but relatively low power on comparisons among levels of factor A. When covariance analysis is applied to this design, the power of factor A comparisons can be markedly increased if the covariate is well chosen. Hence, the covariance analysis of this design can result in very high power on factor A (if the covariate is highly correlated with the dependent variable); tests on B and the $A \times B$ interaction also have high power, but this is a result of the repeated measurement aspect of the design. Covariance analysis can be applied to other complex experimental designs.

CHAPTER 25

Randomized Pretest–Posttest Designs

25.1 INTRODUCTION

There is considerable confusion among researchers regarding the most reasonable way to analyze data from randomized pretest–posttest designs. The two most frequently encountered analyses are (1) one-factor ANOVA performed on the differences between pre- and postscores and (2) two-factor ANOVA with one between-subjects factor and one repeated-measures factor where the levels of the repeated factor are pre and post. Occasionally, one encounters a research report in which both of these analyses have been applied to the same data. Also, these designs have been analyzed using one-factor ANOVA on the posttest. The purpose of this chapter is to point out relationships among these methods, to explain why the two-factor ANOVA approach is often misleading, and to explain why ANCOVA is usually better than all the ANOVA alternatives.

25.2 COMPARISON OF THREE ANOVA METHODS

Consider the design shown in Table 25.1. Three different analyses using ANOVA can be applied to these data. First, it can be treated as a one-factor randomized design and analyzed using a *one-factor ANOVA on the posttest scores*. Second, it can be considered to be a one-factor randomized design and analyzed using a *one-factor ANOVA on the pre–post differences* shown in the last column. Third, it can be considered to be a two-factor repeated measures layout and analyzed using a *two-factor ANOVA with one between-subjects factor and one within-subjects factor*. Note that there are three groups, each with three subjects. The groups constitute the three levels of the between-subjects factor. The second factor (trials) is the repeated-measures (within subjects) factor where "pretest" is level one and "posttest" is level two. The results of these analyses are shown in Table 25.2.

The Analysis of Covariance and Alternatives: Statistical Methods for Experiments, Quasi-Experiments, and Single-Case Studies, Second Edition. Bradley E. Huitema.

Table 25.1 Example of Pretest–Posttest Randomized Design

Treatment Group	Trials		
	Pretest	Posttest	Difference
1	2	4	2
1	4	9	5
1	3	5	2
2	3	7	4
2	4	8	4
2	4	8	4
3	3	9	6
3	3	8	5
3	2	6	4

None of these analyses reveals a statistically significant effect. There is, however, a substantial difference between the size of the F ratio for ANOVA on posttest scores and ANOVA on difference scores (.86 vs. 2.25). These two analyses are straightforward to interpret. In one case the analysis evaluates the differences among the posttest means and in the other case the analysis evaluates the differences among

Table 25.2 Three Analyses of Variance for the Data Shown in Table 25.1

One-Factor ANOVA on Posttest Scores					
Source	SS	*df*	MS	*F*	*p*
Between	5.556	2	2.778	0.86	.47
Within	19.333	6	3.222		
Total	24.889	8			
One-Factor ANOVA on Difference Scores					
Source	SS	*df*	MS	*F*	*p*
Between	6	2	3.000	2.25	.19
Within	8	6	1.333		
Total	14	8			
Two-Factor Repeated-Measurement (Split-Plot) ANOVA					
Source	SS	*df*	MS	*F*	*p*
Between subjects	22.778	8	2.847		
(A) Tx	4.111	2	2.056	0.66	.55
Subjects within Tx	18.667	6	3.111		
Within subjects	79.000	9	8.778		
(B) Trials	72.000	1	72.000	108.00	.00
(AxB) Tx × trials	3.000	2	1.5	2.25	.19
Trials × sub w/Tx	4.000	6	0.667		
Total	101.778	17			

the three mean pretest–posttest differences. The third analysis, however, is very frequently misunderstood. Even the basic descriptive statistics associated with this two-factor approach are often misinterpreted.

Recall that two-factor designs provide marginal means on both factors. The first factor in the design is factor *A*, which is the treatment factor. It is commonly believed that the treatments factor and the associated means provide the main results of interest from this analysis. After all, the treatments factor is indeed the factor of interest in the one-factor ANOVA on the posttest scores because we want to know if the treatments have an effect on the posttest. It is common sense to believe that one should also be interested in the results of the treatments factor in the two-factor analysis. Alas, common sense fails in this case.

The nature of marginal means is often forgotten when analyzing this design. Consider the factor *A* marginal means. Recall that a marginal mean is computed in a balanced design by averaging all of the scores for a specific level of one factor over all levels of the other factor. But consider the nature of the scores being averaged. If you are interested in describing the effects of the treatments are you really interested in means that are computed by pooling together the pretest and posttest scores? Posttest means are of interest and the difference between pretest and posttest means are of interest, but means based on combined pretest and posttest data are not of interest. The differences among the marginal "treatment" means and the associated *F*-test are misleading and useless in the case of the pretest–posttest design; they should be ignored if this analysis is used. A comparison of the treatment sum of squares from the one-factor ANOVA on posttest scores with the treatment sum of squares from the two-factor ANOVA reveals the typical attenuating effect of the pretest data on the differences among the "treatment" means. The treatment SS values are 6 and 4.11, for the one- and two-factor analyses, respectively. The difference between these values is simply a reflection of the fact that differences among posttest means are larger than differences among the "treatment" marginal means in the two-factor analysis.

The aspect of the two-factor analysis that actually provides a result of interest is the interaction test. If you ponder this it makes sense because, in the context of this design, the interaction is saying that simple effects of factor *B* (trials) depend on the treatment level. But this is making something very simple somewhat complex; to state that there is an interaction is equivalent to stating that there are differences between groups on the amount gained. It turns out that the one-factor ANOVA *F*-test on the difference scores is algebraically equivalent to the interaction *F*-test in the two-factor analysis. Note in Table 25.2 that both *F*- and *p*-values are identical in these two analyses.

But do not interpret the equivalence of the two *F*-tests to mean that the choice between the two methods of analysis is arbitrary. There are two reasons to have a strong preference for the one-factor on difference scores approach. First, and most importantly, the results are more transparent. Second, the computation and interpretation of multiple comparison tests is straightforward. If the two-factor approach is followed instead, nonstandard interaction contrast tests will be needed to answer the questions that are answered simply in the one-factor analysis. My recommendation is to avoid the two-factor analysis of this design even though it is the most frequently recommended method. I also recommend avoiding the one-factor ANOVA on posttest

scores; it ignores pretest data and the typical result is substantially lower power than is provided by methods that acknowledge the pretest. But the recommendation to use the one-factor ANOVA on difference scores rather than the other two analyses is simply the best choice among the ANOVA approaches. ANCOVA is likely to be a better choice than any of the ANOVA methods.

25.3 ANCOVA FOR PRETEST–POSTTEST DESIGNS

ANCOVA is the preferred method of analysis for the randomized pretest–postest design. This design differs only slightly from the randomized designs used in previous examples of ANCOVA. Most of the ANCOVA examples in earlier chapters illustrate the use of a covariate that measures a characteristic that differs from the dependent variable. In this chapter the covariate is an early measurement on the variable to be used as the dependent variable rather than a separate variable measured in a different way. The measurement on the pretest may be planned, or, if the experiment involves the collection of data over a long treatment period (as in some learning experiments) the decision to employ preliminary (pre treatment) trials as the covariate may be made after the experiment is begun.

The difference-score ANOVA approach described in the previous section is most efficient when the slope of the pooled within-group regression of the posttest on the pretest is equal to 1.0. When the slope (which is not computed as a part of the ANOVA on differences) is in reality far from 1.0, the analysis will be less powerful than the ANCOVA approach described in this section. The reason is that ANCOVA fits the slope b_w to the data rather than assuming that it is equal to 1.0. Although pretest–posttest correlations are almost always high, they usually differ substantially from 1.0, if for no reason other than measurement error. This means that the ANCOVA fitted slope will provide a smaller estimate of error SS than will ANOVA on differences. In general, unless it is known at the time the experiment is designed that the slope is very close to 1.0, I recommend ANCOVA for the pretest–posttest randomized design. It is carried out using the pretest as the covariate and the posttest as the dependent variable using conventional ANCOVA software. The only instance in which ANCOVA should be avoided with this design is when it is known that the within-groups pretest–posttest correlation is very low (say $<$ .20). ANOVA on the posttest scores can have higher power than either ANOVA on differences or ANCOVA in this case, but very low pre–post correlations are unlikely in most areas of application.

A variant of the ANCOVA approach just described is to use the pretest as the covariate and the difference scores as the dependent variable, but there is no advantage. It turns out that this approach is algebraically equivalent to ANCOVA on posttest scores.

The advantage of ANCOVA over all three ANOVA approaches can be seen by comparing the ANCOVA results in Table 25.3 with the three ANOVA results in Table 25.2.

It can be seen that ANCOVA identifies a statistically significant effect ($p = .04$) whereas none of the three ANOVA approaches does so ($p = .47$, .19, and .19). Note

Table 25.3 Comparison of ANCOVA on Posttest with ANCOVA on Difference Scores Applied to the Data in Table 25.1

ANCOVA on Posttest Scores					
Adjusted treatment	8.959	2	4.480	7.00	.04
Res_w	3.2	5	0.640		
Res_t	12.159	7			
ANCOVA on Difference Scores					
Adjusted treatment	8.959	2	4.480	7.00	.04
Res_w	3.2	5	0.640		
Res_t	12.159	7			

that, as mentioned earlier, ANOVA on differences and repeated-measures ANOVA are equivalent, ANCOVA on posttest scores and ANCOVA on differences are equivalent, and ANOVA on posttest scores yields a far lower F value (and, of course, a higher p-value) than does any other procedure. The data for these analyses were contrived. An empirical example is presented next.

A reanalysis of some data collected by Fredericks (1969) is presented in Table 25.4 to illustrate some of the advantages of ANCOVA over ANOVA. Fredericks was interested in evaluating the effectiveness of two methods of improving the motor coordination of children having Down's syndrome (a chromosomal disorder caused by the presence of all or part of an extra 21st chromosome). The first method is the Doman–Delacato patterning treatment, and the second is a behavior-modification treatment that utilizes principles of shaping, reverse chaining, and social reinforcement. Seventy-two Down's syndrome children were randomly assigned to six groups. Only the first three groups are considered here.

The first group (group A) was treated with the Doman–Delacato patterning procedure four times a day, 15 minutes at a time for a period of 9 weeks. The second group (group B) was treated with the behavior-modification procedure for the same period of time. The third group (group C) was a control group that was not treated but was tested in accordance with the testing schedule employed with the first two groups. Subjects in all three groups were measured on a modified version of the Lincoln–Oseretsky motor development scale and on the Doman–Delacato profile before treatment began, every 2 weeks during treatment, and at the conclusion of the treatment program. Only the pretest and posttest data are included in our reanalysis.

These data have been chosen to show that (1) annoyingly large differences among pretest means in randomized designs are not atypical, and (2) the analysis of covariance can remove the basic interpretation difficulty associated with this problem. First, observe the pretest differences among the three groups on the two response measures. Fredericks (1969, p. 74), in commenting on his analysis (t tests) of the posttest scores on the Lincoln–Oseretsky scale, expresses concern for the adequacy of this analysis, "The fact that the behavior modification group mean is approximately 13 score points higher than either the Doman–Delacato group or the control group at the start of the experimental program somewhat clouds the interpretability of the obtained trend

Table 25.4 Down's Syndrome Data and Analysis: A Randomized Pretest–Posttest Design

	Lincoln–Oseretsky Scale						Doman–Delacato Profile					
	A		*B*		*C*		*A*		*B*		*C*	
Treatment[a]	Pretest	Posttest	Pretest	Posttest	Pretest	Posttest	Pretest	Posttest	Pretest	Posttest	Pretest	Posttest
	31	65	49	62	23	30	35	39.5	52	60	35	39.5
	10	16	0	0	41	68	32	35	12	12	48	54
	9	10	89	117	46	60	15	18	48	56	44	52
	65	76	44	68	0	4	46	54	48	50	18	18
	33	61	5	8	19	38	38	42.5	13	15	33.5	36.5
	0	5	52	105	6	25	6	10.5	39.5	42	23	23
	5	0	0	4	24	45	38	38	17	17	29	33
	0	0	2	5	0	0	16	17	38	39.5	9	9
	6	16	19	85	21	48	29	32	40	42	32	33
	41	52	67	122	26	49	32	35	50	60	37	41
			33	54	0	14			29	33	32	33

Data Analysis

Lincoln–Oseretsky Scale						Doman–Delacato Profile					
ANOVA on pretest $\bar{X}_A = 20.00$ $\bar{X}_B = 32.37$ $\bar{X}_C = 18.73$						ANOVA on pretest $\bar{X}_A = 28.70$ $\bar{X}_B = 35.14$ $\bar{X}_C = 30.95$					
Source	SS	*df*	MS	*F*	*p*	Source	SS	*df*	MS	*F*	*p*
Among	1,303.51	2	651.8	1.19	.32	Among	226.0	2	113.0	0.67	.52
Within	15,904.37	29	548.4			Within	4,916.37	29	169.5		
Total	17,207.88	31				Total	5,142.37	31			

ANOVA on posttest	ANOVA on posttest
$\bar{Y}_A = 30.10$ $\bar{Y}_B = 56.36$ $\bar{Y}_C = 34.64$	$\bar{Y}_A = 32.15$ $\bar{Y}_B = 38.77$ $\bar{Y}_C = 33.82$

Data Analysis

Lincoln–Oseretsky Scale						Doman–Delacato Profile					
Source	SS	*df*	MS	*F*	*p*	Source	SS	*df*	MS	*F*	*p*
Among	4,226.88	2	2,113	1.80	.18	Among	253.16	2	126.6	0.56	.58
Within	33,967.10	29	1,171			Within	6,536.34	29	225.4		
Total	38,194.88	31				Total	6,789.50	31			
ANCOVA						ANCOVA					
$\bar{Y}_{A\text{adj}} = 35.4$ $\bar{Y}_{B\text{adj}} = 44.6$ $\bar{Y}_{C\text{adj}} = 41.6$						$\bar{Y}_{A\text{adj}} = 35.6$ $\bar{Y}_{B\text{adj}} = 34.8$ $\bar{Y}_{C\text{adj}} = 34.7$					
Source	SS	*df*	MS	*F*	*p*	Source	SS	*df*	MS	*F*	*p*
Adjusted tr.	441.38	2	220.7	1.15	.33	Adjusted tr.	4.64	2	2.32	0.47	.63
Res_w	5,372.17	28	191.9			Res_w	137.25	28	4.90		
Res_t	5,813.55	30				Res_t	141.89	30			

(*Continued*)

Table 25.4 Down's Syndrome Data and Analysis: A Randomized Pretest–Posttest Design (*Continued*)

Summary of All Analyses					
Lincoln–Oseretsky Scale					
Treatment	*A*	*B*	*C*	*F*	*p*
$\bar{X}$	20.00	32.73	18.73	1.19(ANOVA)	.32
$\bar{Y}$	30.10	56.36	34.64	1.80(ANOVA)	.18
$\bar{Y}_{adj}$	35.4	44.6	41.6	1.15(ANCOVA)	.33
Doman–Delacato Profile					
Treatment	*A*	*B*	*C*	*F*	*p*
$\bar{X}$	28.70	35.14	30.95	0.67(ANOVA)	.52
$\bar{Y}$	32.15	38.77	33.82	0.56(ANOVA)	.58
$\bar{Y}_{adj}$	35.6	34.8	34.7	0.47(ANCOVA)	.63

[a]Treatments: *A*, Doman–Delacato; *B*, behavior modification; *C*, control.
Source: Data from Fredricks (1969).

differences and the much larger posttest difference favoring the behavior modification treatment."

This problem is attended to with the analysis of covariance. Note in Table 25.4 that the adjusted means associated with ANCOVA differ from the unadjusted values by several points. The adjustments are such that there is less difference among group means. This adjustment effect is even more dramatic on the Doman–Delacato profile. Note that there is almost a seven-point difference between the highest and the lowest unadjusted means but that the adjusted means are almost identical. Note that all these analyses yield nonsignificant differences. Descriptively, however, the results of ANCOVA are less ambiguous than those of ANOVA because the adjusted differences are the posttest differences that would be expected if the pretest scores for the three groups were equal.

Before concluding the discussion of this design, it should be mentioned that there is a large body of literature devoted to the complexities of the problems associated with analyses of pretest–posttest designs. It turns out, however, that most of this work deals with either the one-group pretest–posttest design or the nonequivalent group design; these designs are discussed in subsequent chapters. Most of the issues associated with these two designs are not relevant to the randomized pretest–posttest design discussed here.

25.4 SUMMARY

ANCOVA is the recommended procedure for the randomized pretest–posttest design. The analysis is carried out by using the pretest as the covariate and the posttest as the dependent variable. Frequently encountered analytic procedures include (1) one-factor ANOVA on the posttest, (2) one-factor ANOVA on difference scores, and (3) two-factor ANOVA (using treatments as the between subjects factor and the two measurement occasions as levels of the repeated measurement factor). These methods are less satisfactory than ANCOVA because they do not adjust for chance differences on the pretest and they generally have lower power.

CHAPTER 26

Multiple Dependent Variables

26.1 INTRODUCTION

It is not unusual for more than one dependent variable to be employed in many types of experimental and nonexperimental research. Indeed, in many clinical areas the use of a single dependent variable is the exception; the study of multiple measures is the rule. A behavioral science investigation of the effects of treatments for anxiety, for example, may include multiple standardized scales, physiological measures, and checklists based on behavioral observations. An educational researcher studying the effects of different types of instruction may use multiple measures of achievement, multiple measures of satisfaction with instruction, and measures of behavior, such as amount of content-relevant time spent outside the classroom. A clinical trial studying medical interventions for cardiovascular disease may include measures of cardiovascular death, death due to other causes, myocardial infarction, stroke, serum cholesterol indices, and quality of life measures.

Issues in the Analysis of Multiple Dependent Variables

Various procedures for analyzing studies with multiple dependent variables have been developed. They can be distinguished with respect to (1) how they control for multiplicity, (2) whether the test statistic acknowledges correlation among dependent variables, and (3) whether planning is required in the testing procedure.

Multiplicity

When an experiment involves an individual dependent variable the probability of making a Type I error using ANCOVA F may be labeled as α_i, where the subscript indicates that an *i*ndividual dependent variable is involved. In the case of an experiment with multiple dependent variables the probability that one or more of the set or family of F values will result in type I error is not in general equal to α_i. Instead,

The Analysis of Covariance and Alternatives: Statistical Methods for Experiments, Quasi-Experiments, and Single-Case Studies, Second Edition. Bradley E. Huitema.

the probability of type I error is greater than α_i and this probability increases as the number of dependent variables increases. The family error rate is labeled as α_f (where the subscript f indicates the whole *f*amily of tests). The only situation in which the use of conventional ANCOVA procedures on each dependent variable will yield a family error rate that is equal to α_i is when all measures are perfectly correlated (an unlikely situation).

Obviously, if several dependent variables are perfectly correlated, there is no reason to employ more than one of them because the others contribute no unique information. The reason for mentioning this extreme situation is simply to point out that when perfectly correlated dependent variables are analyzed the results of ANCOVA on any individual variable are the same on any other variable. All tests beyond the first are redundant. Likewise, the error rate for a single dependent variable in this situation must be the same as the error rate for the whole set; this is the only case in which $\alpha_i = \alpha_f$.

If the dependent variables are completely uncorrelated (another unlikely situation), the probability of making a type I error in the set or family is quite different than the error rate for a single dependent variable and $\alpha_f > \alpha_i$. Specifically, $\alpha_f = 1 - (1 - \alpha_i)^p$ where p is the number of dependent variables. Suppose that a study employing 10 dependent variables is analyzed by computing a univariate ANCOVA F-test on each dependent variable. If each test is run at the .05 level (i.e., $\alpha_i = .05$), the family error rate is $1 - (1 - .05)^{10}$ or .40. This is the probability that one or more of the 10 F-tests will be significant when the null hypothesis is true for each variable. Because this family error rate is quite high the researcher may want to employ a procedure that will keep the family error rate at .05 rather employ the conventional univariate procedure, which keeps the error rate for each individual F at .05.

Correlation among Dependent Variables

The use of more than one dependent variable in an experiment provides a source of information that is ignored when using separate univariate analyses on the set of dependent variables. If two or more dependent variables are employed, the researcher may gain insights concerning the nature of the data by analyzing the treatment effects with the use of procedures that utilize the relationships among the dependent variables.

Planned Tests

Although each dependent variable is acknowledged in some way in all of the methods described here, some methods require that the role of each variable be planned in the sense that some variables may be identified (before the experiment is performed) as more important than others. These approaches may lead to higher power than those that do not involve a planned ordering of importance.

Research Tradition

Although experiments using multiple response measures are ubiquitous throughout the behavioral and medical sciences, methods of analyzing these experiments

differ greatly from one area to another. The differences are revealed in both training and textbooks. I have before me the four most current, popular, and sophisticated behavioral science experimental design textbooks on the market. The average length is about 850 pages. The total amount of space devoted to the topic of multiple dependent variables is five lines! The same books include an average of 75 pages on multiple comparison procedures for the one-factor design alone. Hence, there is massive emphasis in behavioral science experiments on multiplicity in the form of multiple comparisons among groups on a single variable, but almost no acknowledgement of multiplicity with respect to multiple dependent variables. In contrast, researchers in epidemiology and clinical trials routinely discuss the issue multiplicity with respect to multiple response variables.

Why this difference between behavioral science and medical science? I suspect that if one were to ask the typical experimentally oriented behavioral science researcher why the issue of multiple response measures is not discussed in experimental design courses the response would be either (1) it is not important or (2) it is covered in multivariate analysis courses. In contrast, medical researchers, especially those in pharmaceutical research, are likely to point to governmental (e.g., FDA) scrutiny of multiplicity in evaluating claims for drug effectiveness (e.g., O'Neill, 2006). Regardless of the reasons for differences in research traditions, the fact is that there is no consensus on the most appropriate method to analyze studies with multiple response measures. Several common approaches are described in this chapter.

26.2 UNCORRECTED UNIVARIATE ANCOVA

The most common practice in many areas is to do nothing in the analysis to formally acknowledge multiplicity of response measures. That is, one simply computes a univariate ANCOVA on each dependent variable. It may be argued that readers of research reports can easily see how many response measures are used; therefore they can evaluate for themselves the approximate probability that a type I error is likely. For example, if results on 20 different response measures are reported and only one is statistically significant it may be reasonable to conclude that the outcome of the experiment is not convincing because one would expect to encounter a significant test result this often when the null is true.

Some skeptics who promote the use of other methods argue that authors frequently succumb to the temptation to simply not report the results on dependent variables that do not produce statistically significant results. Of course this practice makes it impossible for a reader to informally evaluate the probability of type I error in the experiment. I agree that this is a problem, but the argument that alternative methods are the solution strikes me as a non sequitur. Surely someone intent on hiding a variable delivering a nonsignificant univariate result will not be inclined to include that variable in a more elaborate procedure.

26.3 BONFERRONI METHOD

It has been pointed out that the family error rate for an experiment depends on the degree of relationship among the p dependent variables. When the correlation among them is perfect, the individual error rate is also the family error rate. If there is no correlation among the dependent variables, the family error rate is $1 - (1 - \alpha_i)^p$. Hence the probability that one or more of a set of F-tests will be declared significant when the null hypothesis is true for each variable lies between α_i and $1 - (1 - \alpha_i)^p$. If we are interested in keeping the probability of making a type I error at less than α_f it is reasonable to simply divide the desired α_f by the number of dependent variables (p) to obtain the α_i that should be employed for each test. That is, if a family α of .05 is desired and 10 dependent variables are employed, we find $.05/10 = .005 = \alpha_i$. If each test is run at the .005 level, the sum of the 10 individual α values can be no more than .05.

Hence,

$$\sum_{i=1}^{p} \alpha_i \leq \alpha_f$$

where the individual α_i are the individual type I error rates associated with dependent variables 1 through p. You may have noticed that $p(\alpha_i)$ or 10(.005) was used here as the maximum possible family error rate rather than the previously defined expression $1 - (1 - \alpha_i)^p$. It turns out that $p(\alpha_i)$, which is an approximate formula, yields values very close to those obtained with $1 - (1 - \alpha_i)^p$ (the exact formula) for conventional α values. For the present example, we find $1 - (1 - .005)^{10} = .04889$, whereas $10(.005) = .05$.

Computation of Bonferroni Adjusted *p*-values

Bonferroni corrected p-values (not to be confused with the number of dependent variables p) can be computed very simply. First, compute the conventional ANCOVA on each dependent variable and note the resulting p-value. Second, multiply each p-value by the number of dependent variables. If the product is less than α_f the result is declared statistically significant on that dependent variable. If the product is greater than 1.0, simply set it at 1.0.

26.4 MULTIVARIATE ANALYSIS OF COVARIANCE (MANCOVA)

A procedure known as *multivariate analysis of covariance* (MANCOVA) is an alternative to the Bonferroni F procedure. Like the Bonferroni F procedure, MANCOVA may be employed to control the family error rate. Unlike the Bonferroni F procedure, MANCOVA utilizes the relationships among the dependent variables and provides

an overall test on treatment differences. If the overall test is significant, a univariate test on each dependent variable or some other more complex multivariate procedure may be employed.

Only the basic notions of MANCOVA are described in this section. A thorough understanding of the more complex aspects of the procedure is based on a prerequisite familiarity with multivariate analysis of variance, canonical correlation, and discriminant analysis. Texts on multivariate analysis (e.g., Stevens, 2009) cover these topics in detail.

The similarity of univariate and multivariate analysis of covariance can be seen in Table 26.1 for the reader familiar with matrix algebra. Note that the null hypothesis associated with MANCOVA is similar to the hypothesis associated with univariate ANCOVA, but adjusted population means ($\mu_{j\,\mathrm{adj}}$) are replaced with adjusted population centroids, which are multivariate means denoted as $\boldsymbol{\mu}_{j\,\mathrm{adj}}$ (where boldface indicates a vector). The adjusted population centroid may be viewed as a vector of p adjusted population means, i.e.,

$$\boldsymbol{\mu}_{j\,\mathrm{adj}} = \begin{bmatrix} \mu^{(1)}_{j\,\mathrm{adj}} \\ \mu^{(2)}_{j\,\mathrm{adj}} \\ \vdots \\ \mu^{(p)}_{j\,\mathrm{adj}} \end{bmatrix}$$

but this representation does not adequately convey the notion that a centroid is a point in hyperspace. When several populations are involved, $p\mu$ values exist for each population and it can be seen that each population has a centroid. Thus the test of the hypothesis $\mu_{1\,\mathrm{adj}} = \mu_{2\,\mathrm{adj}} = \cdots = \mu_{J\,\mathrm{adj}}$ is a test that all populations have equal adjusted means on each dependent variable:

$$\begin{bmatrix} \mu^{(1)}_{1\,\mathrm{adj}} \\ \mu^{(2)}_{1\,\mathrm{adj}} \\ \vdots \\ \mu^{(p)}_{1\,\mathrm{adj}} \end{bmatrix} = \begin{bmatrix} \mu^{(1)}_{2\,\mathrm{adj}} \\ \mu^{(2)}_{2\,\mathrm{adj}} \\ \vdots \\ \mu^{(p)}_{2\,\mathrm{adj}} \end{bmatrix} = \cdots = \begin{bmatrix} \mu^{(1)}_{J\,\mathrm{adj}} \\ \mu^{(2)}_{J\,\mathrm{adj}} \\ \vdots \\ \mu^{(p)}_{J\,\mathrm{adj}} \end{bmatrix}$$

The purpose of MANCOVA is to test this hypothesis. If the multivariate F-test is not significant, the analysis is terminated because there are insufficient data to conclude that the adjusted population centroids are different. If the multivariate F is significant, there may be interest in one of the follow-up procedures designed for this purpose. The simplest procedure is to compute a univariate ANCOVA F-test on each dependent variable (these results are included in *Minitab* MANCOVA output), but such tests do not evaluate the contribution of each variable to the composite used in the overall MANCOVA. This is similar to the situation in multiple regression

Table 26.1 Comparison of Basic Steps in ANCOVA and MANCOVA

ANCOVA	MANCOVA
H_0: $\mu_{1\text{ adj}} = \mu_{2\text{ adj}} = \cdots = \mu_{J\text{ adj}}$	H_0: $\boldsymbol{\mu}_{1\text{ adj}} = \boldsymbol{\mu}_{2\text{ adj}} = \cdots = \boldsymbol{\mu}_{J\text{ adj}}$
Step 1. Compute total sum of squares $\sum y_t^2 = \text{SS}_t$	*Step 1.* Compute total sum of products matrix $\mathbf{T}_{yy}$
Step 2. Compute total residual sum of squares $\sum y_t^2 - \left(\sum xy_t\right)^2 \Big/ \sum x_t^2 = \text{SSres}_t$	*Step 2.* Compute total residual sum of products matrix $\mathbf{T}_{yy} - \mathbf{T}_{yx}\mathbf{T}_{xx}^{-1}\mathbf{T}_{xy} = \mathbf{T}_{y\cdot x}$
Step 3. Compute within-group sum of squares $\sum y_w^2 = \text{SS}_w$	*Step 3.* Compute within-group sum of products matrix $\mathbf{W}_{yy}$
Step 4. Compute within-group residual sum of squares $\sum y_w^2 - \left(\sum xy_w\right)^2 \Big/ \sum x_w^2 = \text{SSres}_w$	*Step 4.* Compute within-group residual sum of products matrix $\mathbf{W}_{yy} - \mathbf{W}_{yx}\mathbf{W}_{xx}^{-1}\mathbf{W}_{xy} = \mathbf{W}_{y\cdot x}$
Step 5. Compute adjusted treatment effect sum of squares $\text{SSres}_\text{t} - \text{SSres}_\text{w} = \text{SSAT}$	*Step 5.* Compute Wilks Lambda for adjusted treatment effects $\dfrac{\lvert\mathbf{W}_{y\cdot x}\rvert}{\lvert\mathbf{T}_{y\cdot x}\rvert} = \Lambda$
Step 6. Compute F ratio for adjusted treatment effects where $F = \dfrac{\text{SSAT}/(J-1)}{\text{SSres}_w/(N-J-C)}$	*Step 6.* Compute multivariate F using Rao's approximation, which is $F = \left(\dfrac{1-\Lambda^{1/s}}{\Lambda^{1/s}}\right)\left(\dfrac{ms-v}{p(J-1)}\right)$ Where $s = \sqrt{\dfrac{p^2(J-1)^2-4}{p^2+(J-1)^2-5}}$ $m = N - C - 1 - \dfrac{p+J}{2}$ $v = \dfrac{p(J-1)-2}{2}$
Step 7. Evaluate F_{obt} with $F_{(\alpha,\,J-1,\,N-J-C)}$	*Step 7.* Evaluate multivariate F with $F_{\alpha,\,p(J-1),\,(ms-v)}$
Step 8. Carry out multiple comparisons if appropriate	*Step 8.* If multivariate F does not exceed the critical value, retain H_0: $\boldsymbol{\mu}_{1\text{ adj}} = \boldsymbol{\mu}_{2\text{ adj}} = \cdots = \boldsymbol{\mu}_{J\text{ adj}}$ and terminate the analysis. If multivariate F is significant, reject H_0 and conclude that some linear combination(s) of dependent variables differentiates groups.

Table 26.2 Raw Data for MANCOVA Examples

Group 1			Group 2			Group 3		
Y_1	Y_2	X	Y_1	Y_2	X	Y_1	Y_2	X
15	1	29	20	2	22	14	1	33
19	3	49	34	3	24	20	5	45
21	5	48	28	2	49	30	6	35
27	4	35	35	4	46	32	6	39
35	6	53	42	5	52	34	5	36
39	6	47	44	4	43	42	7	48
23	5	46	46	5	64	40	7	63
38	6	74	47	5	61	38	6	57
33	6	72	40	4	55	54	7	56
50	7	67	54	6	54	56	7	78

where tests on bivariate regression coefficients are not used to evaluate the partial regression coefficients.

Computational Example 26.1

Suppose that two dependent variables were employed in the achievement study described in Chapter 6 rather than one; hypothetical data are listed in Table 26.2. The Y_1-values are scores on a biology achievement test, the Y_2-values are scores measuring interest in the biological sciences, and the X values are scores on an aptitude test. The problem is to test the hypothesis that the three adjusted population centroids are equal. That is, H_0: $\boldsymbol{\mu}_{1\text{ adj}} = \boldsymbol{\mu}_{2\text{ adj}} = \boldsymbol{\mu}_{3\text{ adj}}$.

The preliminary starting point for MANCOVA is the computation of the total and within-group deviation sum of squares and sum of products supermatrices **T** and **W**. The total supermatrix is of the following form:

$$\mathbf{T} = \left[\begin{array}{c|c} \mathbf{T}_{yy} & \mathbf{T}_{yx} \\ \hline \mathbf{T}_{xy} & \mathbf{T}_{xx} \end{array}\right]$$

T is a symmetric matrix of order $(p + C) \times (p + C)$. That is, the number of rows (or columns) is equal to the number of dependent variables plus the number of covariates C. The submatrices $\mathbf{T}_{xx}$, $\mathbf{T}_{yx}$, $\mathbf{T}_{xy}$, and $\mathbf{T}_{yy}$ are total-deviation sum of squares and sum of products matrices (generally simply called *sum of products matrices*) associated with the variables indicated by the subscripts.

The supermatrix **W** is also symmetric and of the order $(p + C) \times (p + C)$. It differs from **T** only in that it is based on pooled within-group rather than total-deviation sums of squares and products:

$$\mathbf{W} = \left[\begin{array}{c|c} \mathbf{W}_{yy} & \mathbf{W}_{yx} \\ \hline \mathbf{W}_{xy} & \mathbf{W}_{xx} \end{array}\right]$$

It is understood that $\mathbf{W} = \mathbf{W}_1 + \mathbf{W}_2 + \cdots + \mathbf{W}_J$, where the subscripts refer to groups 1 through J.

The **T** and **W** supermatrices for the example data are presented below.

$$\mathbf{T} = \begin{array}{c} \\ y_1 \\ y_2 \\ x \end{array} \begin{array}{c} \begin{array}{ccc} y_1 & y_2 & x \end{array} \\ \left[\begin{array}{ccc|c} 3956 & 411 & & 3022 \\ 411 & 89.47 & & 466.33 \\ \hline 3022 & 466.33 & & 5826.67 \end{array}\right] \end{array}$$

$$\mathbf{W}_1 = \begin{array}{c} \\ y_1 \\ y_2 \\ x \end{array} \begin{array}{c} \begin{array}{ccc} y_1 & y_2 & x \end{array} \\ \left[\begin{array}{cc|c} 1064 & 150 & 1003 \\ 150 & 28.9 & 183 \\ \hline 1003 & 189 & 2014 \end{array}\right] \end{array}$$

$$\mathbf{W}_2 = \begin{array}{c} \\ y_1 \\ y_2 \\ x \end{array} \begin{array}{c} \begin{array}{ccc} y_1 & y_2 & x \end{array} \\ \left[\begin{array}{cc|c} 896 & 113 & 911 \\ 113 & 16 & 119 \\ \hline 911 & 119 & 1798 \end{array}\right] \end{array}$$

$$\mathbf{W}_3 = \begin{array}{c} \\ y_1 \\ y_2 \\ x \end{array} \begin{array}{c} \begin{array}{ccc} y_1 & y_2 & x \end{array} \\ \left[\begin{array}{cc|c} 1576 & 176 & 3252 \\ 176 & 30.01 & 448 \\ \hline 1338 & 146 & 1888 \end{array}\right] \end{array}$$

$$\mathbf{W} = \mathbf{W}_1 + \mathbf{W}_2 + \mathbf{W}_3 = \begin{array}{c} \\ y_1 \\ y_2 \\ x \end{array} \begin{array}{c} \begin{array}{ccc} y_1 & y_2 & x \end{array} \\ \left[\begin{array}{cc|c} 3536 & 439 & 3252 \\ 439 & 75 & 448 \\ \hline 3252 & 448 & 5700 \end{array}\right] \end{array}$$

Because the basic **T** and **W** matrices are now available, the steps described in Table 26.1 can be carried out.

1. The total sum of products matrix $\mathbf{T}_{yy}$ is found in the upper left quadrant of the supermatrix **T**.

$$\mathbf{T}_{yy} = \begin{array}{c} \\ y_1 \\ y_2 \end{array} \begin{array}{c} \begin{array}{cc} y_1 & y_2 \end{array} \\ \left[\begin{array}{cc} 3956 & 411 \\ 411 & 89.47 \end{array}\right] \end{array}$$

2. The total residual sum of products matrix is

$$\begin{aligned} \mathbf{T}_{yx} &= \mathbf{T}_{yy} - \mathbf{T}_{yx}\mathbf{T}_{xy}^{-1}\mathbf{T}_{xy} \\ &= \begin{bmatrix} 3956 & 411 \\ 411 & 89.47 \end{bmatrix} - \begin{bmatrix} 3022 \\ 466.33 \end{bmatrix} [0.000172][3022 \quad 466.33] \\ &= \begin{bmatrix} 2388.64 & 169.14 \\ 169.14 & 52.15 \end{bmatrix} \end{aligned}$$

3. The within-group sum of products matrix is found in the upper left quadrant of the supermatrix **W**:

$$\mathbf{W}_{yy} = \begin{matrix} \\ y_1 \\ y_2 \end{matrix} \begin{matrix} \begin{matrix} y_1 & y_2 \end{matrix} \\ \begin{bmatrix} 3536 & 439 \\ 439 & 75 \end{bmatrix} \end{matrix}$$

4. The within-group residual sum of products matrix is

$$\begin{aligned} \mathbf{W}_{y\cdot x} &= \mathbf{W}_{yy} - \mathbf{W}_{yx}\mathbf{W}_{xy}^{-1}\mathbf{W}_{xy} \\ &= \begin{bmatrix} 3536 & 439 \\ 439 & 75 \end{bmatrix} - \begin{bmatrix} 3252 \\ 488 \end{bmatrix} [0.000172][3252 \quad 488] \\ &= \begin{bmatrix} 1680.65 & 183.40 \\ 183.40 & 39.79 \end{bmatrix} \end{aligned}$$

5. Wilks Λ is computed as

$$\frac{\left|\mathbf{W}_{y\cdot x}\right|}{\left|\mathbf{T}_{y\cdot x}\right|} = \frac{33233.83}{95954.13} = 0.34635 = \Lambda$$

6. The F_{obt} is

$$\left(\frac{1 - (0.34635)^{1/2}}{(0.34635)}\right)\left(\frac{[(25.5)2] - 1}{2(2)}\right) = 8.74$$

where

$$\begin{aligned} s &= \sqrt{\frac{2^2(3-1)^2 - 4}{2^2 + (3-1)^2 - 5}} = 2; \\ m &= 30 - 1 - 1 - \frac{5}{2} = 25.5; \text{ and} \\ v &= \frac{2(2) - 2}{2} = 1. \end{aligned}$$

7. Using $\alpha = .05$, the critical value of F is

$$F_{(.05\, p(J-1), ms-v)} = F_{(.05,4,50)} = 2.57.$$

8. Because the obtained F exceeds the critical value of F, we conclude that the three adjusted population centroids are not equal. That is, we reject H_0: $\boldsymbol{\mu}_{1\text{ adj}} = \boldsymbol{\mu}_{2\text{ adj}} = \boldsymbol{\mu}_{3\text{ adj}}$

Separate univariate ANCOVA F-tests may then carried out on the two dependent variables. These tests require very little additional computation if the $\mathbf{T}_{y\cdot x}$ and $\mathbf{W}_{y\cdot x}$

matrices, which were employed in the overall MANCOVA F-test, are available. These matrices are

$$\mathbf{T}_{y \cdot x} = \begin{bmatrix} 2388.64 & 169.14 \\ 169.14 & 52.15 \end{bmatrix}$$

$$\mathbf{W}_{y \cdot x} = \begin{bmatrix} 1680.65 & 183.40 \\ 183.40 & 39.79 \end{bmatrix}$$

The element in the first row and first column of the $\mathbf{T}_{y \cdot x}$ matrix is the SSres_t for the first dependent variable. The corresponding element in $\mathbf{W}_{y \cdot x}$ is SSres_w for the first dependent variable. Hence, the ANCOVA F for the first dependent variable is easily computed because $\text{SSres}_t - \text{SSres}_w = \text{SS}$ adjusted treatment effects. In this case $2388.64 - 1680.65 = 707.99$, and the F-test is as follows:

Source	SS	*df*	MS	*F*
Adjusted treatments	707.99	2	354	5.48
Res_w	1680.65	26	64.64	
Res_t	2388.64	28		

The ANCOVA F-test for the second dependent variable is based on the values found in the second row and second column of the $\mathbf{T}_{y \cdot x}$ and $\mathbf{W}_{y \cdot x}$ matrices. These values are the SSres_t and SSres_w for Y_2. The difference $52.15 - 39.79 = 12.36$ is the SS adjusted treatment effects for the second dependent variable. The univariate F-test for this variable is as follows:

Source	SS	*df*	MS	*F*
Adjusted treatments	12.36	2	6.18	4.04
Res_w	39.79	26	1.53	
Res_t	52.15	28		

The critical value of F is 3.37 for $\alpha = .05$.

Minitab Data Entry and Menu Commands for MANCOVA and Univariate ANCOVA (example data from Table 26.2)

It can be seen below that the data presented in Table 26.2 are stacked in three columns and a column is added to number the groups.

```
Row   TX   Y1   Y2    X
 1     1   15    1   29
 2     1   19    3   49
 3     1   21    5   48
 4     1   27    4   35
 5     1   35    6   53
 6     1   39    6   47
```

```
 7    1  23   5  46
 8    1  38   6  74
 9    1  33   6  72
10    1  50   7  67
11    2  20   2  22
12    2  34   3  24
13    2  28   2  49
14    2  35   4  46
15    2  42   5  52
16    2  44   4  43
17    2  46   5  64
18    2  47   5  61
19    2  40   4  55
20    2  54   6  54
21    3  14   1  33
22    3  20   5  45
23    3  30   6  35
24    3  32   6  39
25    3  34   5  36
26    3  42   7  48
27    3  40   7  63
28    3  38   6  57
29    3  54   7  56
30    3  56   7  78
```

The *Minitab* menu commands follow:

Stat → ANOVA →General MANOVA → Responses: *Y*1 *Y*2 → Model: *Tx* → Covariates: *X* → *OK* → Results: Univariate analysis of variance → Display least squares means corresponding to the following terms: *TX* → OK → OK

Minitab ANCOVA and MANCOVA Output

```
General Linear Model: Y1, Y2 versus TX
Factor  Type   Levels  Values
TX      fixed       3  1, 2, 3

Analysis of Variance for Y1, using Adjusted SS for Tests

Source  DF   Seq SS   Adj SS   Adj MS      F      P
X        1  1567.36  1855.35  1855.35  28.70  0.000
TX       2   707.99   707.99   354.00   5.48  0.010
Error   26  1680.65  1680.65    64.64
Total   29  3956.00

S = 8.03992   R-Sq = 57.52%   R-Sq(adj) = 52.61%

Term        Coef  SE Coef     T      P
Constant   6.854    5.455  1.26  0.220
X         0.5705   0.1065  5.36  0.000
```

Analysis of Variance for Y2, using Adjusted SS for Tests

Source	DF	Seq SS	Adj SS	Adj MS	F	P
X	1	37.323	35.211	35.211	23.01	0.000
TX	**2**	**12.355**	**12.355**	**6.178**	**4.04**	**0.030**
Error	26	39.789	39.789	1.530		
Total	29	89.467				

S = 1.23707 R-Sq = 55.53% R-Sq(adj) = 50.40%

Term	Coef	SE Coef	T	P
Constant	0.9892	0.8393	1.18	0.249
X	0.07860	0.01639	4.80	0.000

Unusual Observations for Y2

Obs	Y2	Fit	SE Fit	Residual	St Resid
21	1.00000	4.44246	0.47092	-3.44246	-3.01 R

R denotes an observation with a large standardized residual.

Means for Covariates

Covariate	Mean	StDev
X	49.33	14.17

Least Squares Means

	Y1		Y2	
TX	Mean	SE Mean	Mean	SE Mean
1	28.479	2.5583	4.690	0.3936
2	40.331	2.5546	4.183	0.3931
3	36.190	2.5427	5.726	0.3912

MANOVA for X

s = 1 m = 0.0 n = 11.5

	Test		DF		
Criterion	Statistic	F	Num	Denom	P
Wilks'	0.45853	14.761	2	25	0.000
Lawley-Hotelling	1.18088	14.761	2	25	0.000
Pillai's	0.54147	14.761	2	25	0.000
Roy's	1.18088				

MANOVA for TX

s = 2 m = -0.5 n = 11.5

	Test		DF		
Criterion	Statistic	F	Num	Denom	P
Wilks'	**0.34638**	**8.739**	**4**	**50**	**0.000**
Lawley-Hotelling	1.62992	9.780	4	48	0.000
Pillai's	0.74267	7.679	4	52	0.000
Roy's	1.45299				

The label "analysis of variance" actually refers to univariate ANCOVA, the least squares means are the adjusted means, the "MANOVA for TX" section provides Wilk's lambda for MANCOVA, the corresponding *F,* and the associated *p*-value. The

MANOVA for X output can be ignored; it refers to a test on the within group multivariate regression of both dependent variables on X. As with univariate ANCOVA, this test need not be significant to proceed with MANCOVA.

26.5 MANCOVA THROUGH MULTIPLE REGRESSION ANALYSIS: TWO GROUPS ONLY

The computation of MANCOVA for the two-group situation can be carried out by using a multiple regression analysis computer program. Unlike the univariate case in which the regression approach can be easily applied to any number of groups, the multivariate case is conveniently analyzed through regression analysis only when two groups are involved. The similarity of the two-group MANCOVA procedure and the univariate ANCOVA procedure (described in Chapter 7) is described in Table 26.3.

Computational Example 26.2

The MANCOVA steps described in Table 26.3 have been applied to the following data:

	D	Y_1	Y_2	Y_3	Y_4	Y_5	X
Group 1	1	1	7	10	10	17	22
	1	1	6	7	6	14	17
	1	5	3	13	12	20	14
	1	7	3	5	6	18	14
	1	1	1	7	6	18	9
Group 2	0	10	13	15	17	10	24
	0	12	13	18	12	13	18
	0	9	13	18	14	10	14
	0	10	10	18	13	8	15
	0	6	6	11	13	16	6

1. The dummy variable is constructed by assigning a "one" to all subjects in the first group and a "zero" to all subjects in the second group. Column 1 is the dummy variable, columns 2 through 6 are dependent variables, and column 7 is the covariate.
2. Column 1, the dummy variable, is regressed on columns 2 through 7 to obtain $R^2_{dY,X} = 0.96943$.
3. Column 1 is regressed on column 7, the covariate, to obtain $R^2_{dX} = 0.00038$.
4. The multivariate F is

$$\frac{(0.96943 - 0.00038)/5}{(1 - 0.96943)/3} = \frac{0.19381}{0.0191} = 19.02$$

5. Evaluate F_{obt} with $F_{(\alpha,5,10-5-1-1)}$.

Table 26.3 Comparison of Steps Involved in Computing ANCOVA and MANCOVA Through Multiple Linear Regression Procedures

One-factor ANCOVA — J Groups		One-factor MANCOVA – Two Groups	
Step 1.	Construct $J-1$ dummy variables indicating group membership	*Step 1.*	Construct one dummy variable indicating group membership
Step 2.	Regress Y on $J-1$ dummy variables and covariate(s) to obtain $R^2_{yD,x}$	*Step 2.*	Regress the dummy variable on all p dependent variables and covariate(s) to obtain $R^2_{dY,X}$
Step 3.	Regress Y on covariate(s) to obtain R^2_{yX}	*Step 3.*	Regress dummy variable on covariate(s) to obtain R^2_{dX}
Step 4.	Compute F by using $\frac{\left(R^2_{yD,X}-R^2_{yX}\right)/(J-1)}{\left(1-R^2_{yD,x}\right)/(N-J-C)}=F$	*Step 4.*	Compute multivariate F by using $\frac{\left(R^2_{dY,X}-R^2_{dX}\right)/p}{\left(1-R^2_{dY,X}\right)/(N-p-C-1)}=F$
Step 5.	Evaluate F_{obt} with $F_{(\alpha,J-1,N-J-C)}$	*Step 5.*	Evaluate multivariate F_{obt} with $F_{(\alpha,p,N-p-C-1)}$

If $\alpha = .05$, then $F_{(5,\,3)} = 9.01$ and the null hypothesis is rejected. The p-value is .018. This means that a difference exists between the two population centroids defined by the optimum linear combination of the five dependent variables and the covariate. This difference is not displayed here because it is not of interest to most researchers. Univariate ANCOVA on each dependent variable is likely to be of more descriptive interest at this stage.

26.6 ISSUES ASSOCIATED WITH BONFERRONI *F* AND MANCOVA

Although the Bonferroni and MANCOVA approaches are both popular, the method most likely to be used depends on the area of application. MANCOVA and uncorrected univariate ANCOVA are favored in the behavioral sciences whereas Bonferroni and other procedures (discussed subsequently) are more often used in the medical sciences. The properties of these methods are very different. The strengths of MANOVA are revealed in exploratory studies.

There are situations in which exploratory studies are carried out when little is known about the characteristics of either the treatments or the response measures. These situations may be appropriate for the application of MANCOVA because the associated matrices contain information that can clarify the nature of the response variables. When an exploratory study is based on large samples it is often possible to learn a great deal about both the effects of the treatments and the nature of the dependent variables. But when the research is programmatic or confirmatory rather than exploratory, the selection of response measures should be based on a thorough understanding of the instrumentation.

Most experimenters select dependent variables on the basis of information that is available before the experiment is carried out. A review of previous research, knowledge of current practice, and literature on properties of measurement methods will generally lead to a set of variables of interest to the experimenter. If the response measures have been carefully researched and chosen to directly answer the researcher's question there may be little to be gained by employing a complex multivariate technique. The research question is "On which dependent variables (if any) are treatment effects present and how large are they?" The MANCOVA multivariate F-test answers the somewhat different question, "Are treatment effects significant on an optimum linear combination of the dependent variables?" The experimenter may have no particular interest in a test on a linear combination determined by the analysis if the dependent variables were carefully chosen to yield the information of experimental interest. The difference between the researcher's question and the answer provided by MANCOVA sometimes leads to problems of interpretation.

Suppose that two dependent variables are employed in a two-group experiment and the multivariate F is significant. It is quite possible for both univariate F-tests to be nonsignificant. In this case it is concluded that treatment effects have been detected, but the statement concerning the existence of effects does not apply to either of the dependent variables! Rather, it is concluded that an optimum linear combination of the two dependent variables results in a composite variable that *does* reflect the treatment effect. The linear combination responsible for the significant multivariate F can be specified (it is known as the discriminant function) but the experimenter may not be particularly interested in it. Further, under other conditions, the multivariate test may not detect a treatment effect whereas the univariate tests are significant.

The latter outcome is one of the main reasons that multivariate tests (including Hotelling's T^2, which is equivalent to MANOVA in the two-group case) are dismissed in many areas, especially clinical trials. Following O'Brien (1984), Pocock et al. (1987) state this position clearly: ". . . Hotelling's T^2 is intended to detect *any* departure from the null hypothesis and hence lacks power to detect any specific types of departure that are considered a priori to be biologically plausible. Thus, Hotelling's T^2 is quite unsuitable for analysis of clinical trials and is not considered further." (p. 488). One reason for lower power of multivariate tests is that they are nondirectional; in contrast, hypotheses regarding treatment effects in clinical trials are often stated in directional form. Procedures that often have higher power than either Bonferroni or MANCOVA are frequently adopted in these areas; several are described next.

26.7 ALTERNATIVES TO BONFERRONI AND MANCOVA

Fixed Sequence Procedure

The fixed sequence procedure requires the researcher to order the dependent variables according to the expected strength of the effect. The variable expected to be the most affected is the first in the sequence of tests. Alpha is set at α_i and a univariate test is performed on this dependent variable. If the test result is nonsignificant the testing

sequence is terminated. If it is significant the next test in the sequence is performed on the dependent variable that is ordered second in expected strength; this test is also based on α_i. The testing stops at the point in the sequence where a nonsignificant result is obtained. Each test is based on the same (unadjusted) level $\alpha_{i,}$, which is usually set at .05. This method is widely used in areas where much is known about the properties of the dependent variables and the treatments. The more that is known the better the predictions and, consequently, the higher power is likely to be. This approach has three advantages: simplicity, high power, and high credibility among federal regulatory agencies. The disadvantage is that it can fail miserably if used when there is no strong basis for predicting outcomes. I recommend that it not be used in this situation.

Unequal α Allocation Procedure

The unequal α allocation procedure has elements of Bonferroni and fixed sequence procedures. It involves ordering the dependent variables by expected effectiveness of the treatments, but the tests do not all use the same level for α_i. The purpose of ordering the dependent variables is to determine the various α_i levels to be used for the tests. Unlike the conventional Bonferroni approach, where α_f is split evenly among the set of tests (i.e., $\alpha_i = \alpha_f$/number of dvs), the unequal α allocation procedure assigns more weight to the variable(s) considered most important. This variable (sometimes called the primary endpoint) is often assigned the lion's share of α. For example, suppose the family α is set at .05 and there are four dependent variables, one of which is considered to be the most likely to be affected. If this variable by itself is considered as important as the whole set of three remaining variables the α allocation is .025 for the first variable and .0083 for each remaining variable. Unlike the fixed sequence procedure, all tests are performed regardless of the outcome of the other tests.

Global Tests (Two Groups)

Global tests require no *a priori* weighting of variable importance or expectations of likely strength of effect. Instead, each variable has the same weight in the formation of a composite variable based on all dependent variables. These tests are somewhat like the multivariate approach in the sense that they provide one overall test. O'Brien (1984) introduced several global tests for designs without covariates. The approximate global ANCOVA approach described in this section is in the spirit of O'Brien's global nonparametric test, but it is parametric in nature. I recommend a follow-up procedure based on the closure principle (Lehmacher et al. 1991) to make inferences regarding individual dependent variables following a global test result.

Overall Global ANCOVA

The first step is to standardize each dependent variable. This standardization is based on the pooled within group standard deviation. Each standard score z_{ijd} is computed using:

$$\frac{Y_{ijd} - \bar{Y}_{..d}}{s_{w_d}} = z_{ijd},$$

where

Y_{ijd} is the raw score for subject i, from group j, on dependent variable d;
$\bar{Y}_{..d}$ is the grand mean on dependent variable d; and
s_{w_d} is the pooled within groups standard deviation on dependent variable d.

It is important that the original scores be scaled in the same direction. That is, the nature of the scales must be such that superior performance is consistently associated with high scores. Be sure that high performance is not associated with low scores; if this is true for any dependent variable the scaling for this variable must be reversed. Of course the scaling direction must be established before the data are collected. After each dependent variable is standardized, the second step is to compute the sum of the standardized scores (z_{sum}) for each subject. The third step is to perform a univariate ANCOVA, using z_{sum} as the dependent variable and X as the covariate.

Tests on Individual Dependent Variables

The global test is applied to all combinations of the p dependent variables (there are $2^p - 1$ of these). If the p-value for *any* combination of the dependent variables is greater than α, the variables in that set are declared nonsignificant. A specific dependent variable is declared significant if the global tests for all combinations including that variable are significant.

26.8 EXAMPLE ANALYSES USING *MINITAB*

The methods just described are illustrated in this section using a single data set.

Minitab Input for MANCOVA

The data shown in Table 26.4 were entered in the *Minitab* worksheet. Column c20 was the treatment indicator, the five dependent variables were entered in worksheet columns c21 through c25, and the covariate was entered in c27.

Table 26.4 Fictitious Data for a Two-Group Experiment with Five Dependent Variables and One Covariate

Row	Tx	Y1	Y2	Y3	Y4	Y5	X
1	1	5	7	14	14	73	22
2	1	5	5	11	10	77	17
3	1	9	11	17	16	75	14
4	1	11	7	9	10	75	14
5	1	5	10	11	10	75	9
6	0	10	13	15	14	79	24
7	0	12	13	18	13	83	18
8	0	9	13	18	17	77	14
9	0	10	10	18	13	82	15
10	0	6	6	11	12	80	6

Commands for Computing Univariate ANCOVA on Each Dependent Variable and the MANCOVA

```
MTB > GLM C21-C25 = c20;
SUBC>   Covariates c27;
SUBC>   MANOVA;
SUBC>   Means c20.
```

Output

The output is shown below. Note that each univariate analysis is labeled as "Analysis of Variance." These are actually analyses of covariance. The relevant portion of the output for each dependent variable is in boldface. The adjusted means associated with all five dependent variables appear together in the section labeled "Least Squares Means." The output for MANCOVA begins immediately after the least squares mean section. The relevant portion is in boldface.

```
General Linear Model: Y1, Y2, Y3, Y4, Y5 versus Tx
Factor  Type   Levels  Values
Tx      fixed       2  0, 1

Analysis of Variance for Y1, using Adjusted SS for Tests

Source  DF  Seq SS  Adj SS  Adj MS     F      P
X        1   4.256   3.957   3.957  0.59  0.469
Tx       1  14.101  14.101  14.101  2.09  0.192
Error    7  47.243  47.243   6.749
Total    9  65.600

S = 2.59787   R-Sq = 27.98%   R-Sq(adj) = 7.41%

Term        Coef  SE Coef     T      P
Constant   6.320    2.589  2.44  0.045
X         0.1229   0.1605  0.77  0.469

Analysis of Variance for Y2, using Adjusted SS for Tests

Source  DF  Seq SS  Adj SS  Adj MS     F      P
X        1   6.571   6.107   6.107  0.76  0.411
Tx       1  22.036  22.036  22.036  2.76  0.141
Error    7  55.893  55.893   7.985
Total    9  84.500

S = 2.82573   R-Sq = 33.85%   R-Sq(adj) = 14.96%

Term        Coef  SE Coef     T      P
Constant   7.164    2.816  2.54  0.038
X         0.1527   0.1746  0.87  0.411

Analysis of Variance for Y3, using Adjusted SS for Tests
```

```
Source  DF   Seq SS  Adj SS  Adj MS     F      P
X        1   12.571  11.799  11.799  1.26  0.298
Tx       1   31.628  31.628  31.628  3.39  0.108
Error    7   65.401  65.401   9.343
Total    9  109.600

S = 3.05663   R-Sq = 40.33%   R-Sq(adj) = 23.28%

Term        Coef  SE Coef     T      P
Constant  10.953    3.047  3.60  0.009
X         0.2122   0.1888  1.12  0.298
```

Analysis of Variance for Y4, using Adjusted SS for Tests

```
Source  DF  Seq SS  Adj SS  Adj MS     F      P
X        1   4.231   4.007   4.007  0.66  0.445
Tx       1   7.876   7.876   7.876  1.29  0.294
Error    7  42.793  42.793   6.113
Total    9  54.900

S = 2.47251   R-Sq = 22.05%   R-Sq(adj) = 0.00%

Term        Coef  SE Coef     T      P
Constant  11.008    2.464  4.47  0.003
X         0.1237   0.1528  0.81  0.445
```

Analysis of Variance for Y5, using Adjusted SS for Tests

```
Source  DF  Seq SS  Adj SS  Adj MS      F      P
X        1   0.088   0.209   0.209   0.05  0.833
Tx       1  67.721  67.721  67.721  15.50  0.006
Error    7  30.591  30.591   4.370
Total    9  98.400

S = 2.09049   R-Sq = 68.91%   R-Sq(adj) = 60.03%

Term         Coef  SE Coef      T      P
Constant   78.032    2.084  37.45  0.000
X         -0.0282   0.1292  -0.22  0.833

Means for Covariates
Covariate   Mean  StDev
X          15.30  5.397
```

Least Squares Means

```
    ------Y1------  -------Y2------  -------Y3------
Tx   Mean  SE Mean    Mean  SE Mean    Mean  SE Mean
0   9.388   1.1619  10.985   1.2638  15.979   1.3671
1   7.012   1.1619   8.015   1.2638  12.421   1.3671
```

```
  -------Y4------  --Y5--
    Mean  SE Mean    Mean
  13.788   1.1058  80.203
  12.012   1.1058  74.997

Tx  SE Mean
0    0.9350
1    0.9350

MANOVA for X
s = 1    m = 1.5    n = 0.5

                         Test              DF
Criterion              Statistic      F  Num  Denom      P
Wilks'                   0.76553  0.184    5      3  0.951
Lawley-Hotelling         0.30628  0.184    5      3  0.951
Pillai's                 0.23447  0.184    5      3  0.951
Roy's                    0.30628

MANOVA for Tx
s = 1    m = 1.5    n = 0.5

                         Test              DF
Criterion              Statistic      F  Num  Denom      P
Wilks'                   0.07984  6.915    5      3  0.071
Lawley-Hotelling        11.52431  6.915    5      3  0.071
Pillai's                 0.92016  6.915    5      3  0.071
Roy's                   11.52431
```

An inspection of the output reveals that the adjusted means for the "0" treatment condition are consistently larger than the means for condition "1," yet the MANCOVA F is not significant (i.e., $p > .07$). It is concluded that the optimum linear combination of the five dependent variables does not provide convincing evidence of a treatment effect.

Global ANCOVA (applied to the standardized data in Table 26.5)
Analysis of Covariance for Sum z

```
Source       DF   Adj SS       MS       F      P
Covariates    1   11.388   11.388    1.20  0.310
Tx            1  106.836  106.836   11.22  0.012
Error         7   66.679    9.526
Total         9  186.311

Adjusted Means
Tx  N    z sum
0   5   3.2692
1   5  -3.2692
```

Table 26.5 Data from Table 26.4 Transformed to Standard Scores, Sum of the Transformed Scores for Each Subject, and Covariate Scores

Row	Tx	z1	z2	z3	z4	z5	zsum	X
1	1	−1.26482	−0.89831	−0.06439	0.45473	−2.34455	−4.11734	22
2	1	−1.26482	−1.61696	−1.03026	−1.19884	−0.30581	−5.41670	17
3	1	0.31621	0.53899	0.90148	1.28152	−1.32518	1.71302	14
4	1	1.10672	−0.89831	−1.67418	−1.19884	−1.32518	−3.98979	14
5	1	−1.26482	0.17966	−1.03026	−1.19884	−1.32518	−4.63944	9
6	0	0.71146	1.25764	0.25757	0.45473	0.71356	3.39496	24
7	0	1.50198	1.25764	1.22344	0.04134	2.75229	6.77668	18
8	0	0.31621	1.25764	1.22344	1.69492	−0.30581	4.18638	14
9	0	0.71146	0.17966	1.22344	0.04134	2.24261	4.39851	15
10	0	−0.86957	−1.25764	−1.03026	−0.37205	1.22324	−2.30628	6

Note that the *p*-value for the global test on the five dependent variables is .012. It can be concluded that there is strong evidence of an overall treatment effect. Follow up testing using the global approach applied to all 30 combinations of the dependent variables identifies variable five as the only individual measure on which the treatment has a statistically significant effect.

Conventional Bonferroni

The product of five (i.e., the number of dependent variables) times each univariate ANCOVA *p*-value (i.e., .192, .141, .108, .294, and .006) yields the Bonferroni-adjusted *p*-values for the five dependent variables. They are: .96, .70, .54, 1.00, and .03 for dependent variables 1 through 5, respectively. It is concluded that the evidence supports a claim of a treatment effect on only the fifth outcome measure.

Unequal Allocation of α

Suppose the grant proposal for the example research specified a family error rate of .05 and the first dependent variable was identified as the most important outcome measure; the set of remaining measures was assigned importance equal to that of the first measure. In this case the α values for the five outcomes are set at .025, .00625, .00625, .00625, and .00625. Consequently, response measure five is declared statistically significant because the univariate *p*-value for this measure is .006, which is less than the α allocated to this variable.

Fixed Sequential Testing

Suppose the sequence had been specified as variables one through five (in that order). That is, variable Y_1 was viewed as the most likely to reveal an effect and variable Y_5 the least likely. Because the first endpoint did not reveal a statistically significant effect, the testing sequence terminates without testing the remaining endpoints; it is concluded that the evidence does not support a claim of a convincing effect.

Considerations in the Choice of Analytic Procedure

The results of the example study demonstrate major inconsistencies in the conclusions reached. No effect was identified using either MANCOVA or the fixed sequence procedure (in which a poor guess was made regarding the variable to position first in the sequence). The conventional Bonferroni and the unequal α allocation procedures identified variable five as the only variable affected. The global test yielded a very convincing *p*-value of .012 (nondirectional). The inconsistencies among these conclusions illustrate that the analytic choice is important.

The choice should be based on what is known before the experiment is carried out. When little is known regarding either the likely size or direction of the treatment effects on any of the available measures, MANCOVA is a reasonable choice if

reasonable sample size is available and the dependent variables do not appear to measure completely different characteristics. This is the situation in many exploratory studies. MANCOVA also has a role in observational studies when the purpose is to describe ways in which existing groups differ. Multivariate tests tend to have lower power than univariate and global tests when sample size is small (as demonstrated in the second example) but it should be pointed out that the joint distribution of the dependent variables may be such that the multivariate test is very sensitive. It can be shown that it is possible, with certain distributions, for small samples to yield a very large multivariate F-value and very small univariate F-values. This is, however, a rare outcome.

Consider next the situation where a substantial amount is known. If the direction of the treatment effect on several variables can be anticipated but the degree cannot, I recommend the global test when the response variables are related. The global test can identify a small but consistent (across dependent variables) overall treatment effect where univariate and Bonferroni methods fail to do so. Indeed, this is the strength of the global approach. For example, Wang et al. (2010) investigated the effects of a potential pain reducing therapy on three measures. These measures included maximum pain experienced during the total experimental period, average pain experienced during the period, and a measure of immediate change in pain following brief applications of the treatment. These measures are all in the pain domain, but they measure different aspects of pain perception. Because the study was in a new area there was no basis to predict the measure most likely to be affected. No primary endpoint could be identified but it was expected that all measures would be affected in the same direction (which turned out to be true). Hence, a global measure was ideal in this situation.

Last, consider the situation where a well-researched area is involved. Suppose the planned design involves response variables that have well-known properties and predictions can be made from theory regarding likely effects of the treatments. This makes selection of a primary endpoint natural. If the treatment is expected to have a large effect on a primary outcome it is reasonable to use the fixed sequence method where the primary outcome is the first to be tested. The unequal α allocation approach is a reasonable alternative in this situation if confidence regarding the choice of the primary outcome is not high.

The conventional Bonferroni approach is best if there is strong belief that one of the dependent variables will have a large response to the treatment but there is disagreement among members of the research team regarding the specific variable. The strength of the Bonferroni approach is in detecting a large effect that shows up on only one of the measures. Because the global approach is most useful when a small but consistent effect is present on all measures and, in contrast, the Bonferroni approach is most appropriate when there is a large effect on one of the measures, a hybrid global-Bonferroni procedure is possible. Such approaches have been proposed outside the ANCOVA context (see Dmitrienko et al. 2010), but this approach can be extended to ANCOVA. Unfortunately, determining the appropriate distribution for combined statistics of this type is computationally intensive and has not been implemented in the major software packages.

26.9 SUMMARY

Studies that contain more than one dependent variable can be analyzed in several different ways. One approach is to ignore the multiplicity and simply report the results on individual dependent variables. The argument supporting this approach is that consumers of the research results can easily informally acknowledge the multiplicity.

Some researchers (and funding agencies) prefer an approach that formally acknowledges the multiplicity problem. The choice among procedures should be based on what is known regarding the treatment and response variables. When little is known, as is often the case in exploratory experiments and observational studies, MANCOVA should be considered if sample size is large and the response variables do not measure completely different constructs. If a substantial amount is known regarding the variables of interest, other methods are simpler to interpret and are likely to be more powerful.

The Bonferroni approach is especially useful when the treatment effect is strong on one outcome measure, but it has low power for detecting an effect that is small but consistent on all measures. The global approach is very useful in the latter situation. This involves constructing a composite variable based on scores standardized by within treatment standard deviations and then performing a univariate ANCOVA on the composite.

When a great deal is known about treatment outcomes that are biologically or psychologically plausible, a simple approach is to select a primary response measure and test it first in a planned (i.e., before data collection) sequence of tests that are ordered by the expected strength of the treatment effect. Each test is based on a conventional unadjusted α level. If any test in the sequence is not significant the testing is terminated. A related planned approach is to split the α unevenly across the set of planned tests; the value of α allocated to each variable is determined by the predicted strength of the effect.

PART VII

Quasi-Experiments and Misconceptions

CHAPTER 27

Nonrandomized Studies: Measurement Error Correction

27.1 INTRODUCTION

The methods described in this chapter and the next are relevant in the context of nonrandomized studies. There are many variants of nonrandomized design. Two of them are quasi-experimental: (1) the intentionally biased assignment design and (2) the nonequivalent groups design. The researcher is in control of who gets what in both of these quasi-experimental designs even though the groups to which treatments are assigned are not formed randomly. A third nonrandomized design is the observational study. It differs from the quasi-experimental designs in that the researcher does not have control of treatment assignment. That is, the researcher identifies participants who have been exposed to conditions of interest (such as those who live near suspected toxic substances and those who do not) but there is no assignment to those conditions; often the identification of the exposed subjects is subject to measurement error.

Different Methods of Analysis for Different Versions of Nonrandomized Design

Appropriate methods of analysis differ greatly among nonrandomized designs; appropriate methods are easily identified in some cases but are unclear in others. For example, there is little ambiguity in the case of the intentionally biased design. ANCOVA and related regression models are almost always appropriate for this design because the covariate traps all the information used to assign subjects to treatments. Observational studies are the most problematic; there are major methodological disagreements among researchers and statisticians regarding the analysis of these designs. Indeed, the analysis of observational studies has been one of the most active areas of statistical research for over two decades.

The Analysis of Covariance and Alternatives: Statistical Methods for Experiments, Quasi-Experiments, and Single-Case Studies, Second Edition. Bradley E. Huitema.

I recommend the approach described in this chapter when a fallible exposure (or treatment) variable is employed, the number of covariates (fallible and infallible) is small (say, a dozen or less), and reliability data are available on the covariates or can be collected in a separate study. I recommend the methods of Chapter 28 when the exposure variable is measured without error (i.e., infallible because there is no doubt about who received which condition), many covariates are available, and a large sample of potential control subjects is available.

The methods in both chapters contend with problems of model misspecification, but in completely different ways. These methods are not needed when the ANCOVA model is well specified, as is likely to be the case in randomized designs and intentionally biased assignment designs. In a sense, both approaches acknowledge that something has been left out of the conventional ANCOVA model and that the omission is likely to lead to biased treatment effect estimates. In the case of measurement error correction, the problem is viewed as one of the diluted exposure measures (e.g., estimated saturated fat consumption or estimated amount of smoke to which one is exposed) and diluted covariate measures that do not adequately control confounding.

This brief overview chapter describes (1) the effects of measurement error (in X and in Y) in randomized experiments, (2) the effects of measurement error in nonrandomized studies, and (3) an outline of a measurement error correction procedure for ANCOVA.

Classical Measurement Error

The measurement errors referred to in this chapter are of the type associated with the "classical" measurement error model. This model views an observed measure X (which may be either the exposure measure or a covariate included to remove confounding) as the sum of a true score component T and a measurement error component ME. In other words, an observed exposure or covariate score is conceptualized as

$$X = T + \text{ME}.$$

A corresponding structure applies to Y. It is assumed that ME has a mean of zero, constant variance, and is independent of the true score T. This is the simplest of many measurement error models that are discussed in the psychometric, biometric, and mathematical statistics literature. Important references on measurement error models include books by Fuller (1987) and Carroll et al. (2006), and work cited in Spiegelman (2010).

27.2 EFFECTS OF MEASUREMENT ERROR: RANDOMIZED-GROUP CASE

The effects of measurement error in X and Y for the case of randomized group experiments is briefly reviewed in this section. The effects of measurement error in the case of nonrandomized designs are described in Section 27.3.

Measurement error on the dependent variable is not usually a major problem because the slope associated with the regression of Y on X is not affected. Suppose

(unrealistically) that the regression of Y on X yields a slope of 1.0 and there is a perfect fit. If measurement error is now added to Y the slope will still be 1.0, but the regression line will no longer be a perfect fit. Greater variability on Y will be observed as a result of the measurement error and this will contribute to the error mean square. In the case of ANCOVA the effect of measurement error on Y is an increase in the MS residual within groups, but the pooled slope (b_w) is not affected. This means that the power of the F-test is decreased and the width of the confidence intervals is increased. But, because the slope is not affected, the adjusted means are not affected and no bias is introduced to the adjusted mean difference.

When measurement error is present in the covariate(s) the power of ANCOVA is reduced relative to what it would be if no error were present, but treatment effects are not biased. Hence the effect of measurement error in either the covariate or the dependent variable is similar in that power is reduced. With other designs the effects of measurement error on X are likely to be more serious.

27.3 EFFECTS OF MEASUREMENT ERROR IN EXPOSURE AND COVARIATES: NONRANDOMIZED DESIGN

When subjects are not randomly assigned to treatments, the groups are almost certainly nonequivalent before treatments are carried out and unadjusted mean effect estimates are likely to be biased. It is common to attempt to adjust for preexisting differences using ANCOVA. In this case the claim is made that the difference between adjusted means cannot be explained on the basis of the control variables specified as the covariates. This is not, however, a claim that the effect estimate is unbiased, unless the reason for the nonequivalence is that subjects have been assigned to treatments exclusively on the basis of the covariate. In the typical nonrandomized design, the reason for the nonequivalence is not known. When ANCOVA is used in this case it can be claimed that the difference between adjusted means cannot be explained by the covariates. But this is true only in one sense.

ANCOVA controls for the covariate *as measured*; this does not mean that the analysis controls for the true covariate or the construct supposedly measured by the true covariate. That is, when a fallible variable is used as the covariate (as is usually the case) the means are adjusted for the fallible variable, but fallibility in a covariate usually implies that there would be more adjustment if the variable were measured without error. The concern is that conventional ANCOVA undercorrects because the covariate is fallible; that is, the means are adjusted less than they would be if the covariate were measured without error. Therefore, there may be interest in estimating what the adjusted effect would be if true (infallible) covariate scores had been available. This is the situation in which corrections for measurement error apply. They are not required when the researcher is content to describe a difference that has been adjusted for a covariate measured in the conventional (fallible) manner. But it should be understood that conventional covariate measurements often inadequately operationalize the construct of interest.

Mayo et al. (1978) were interested in investigating whether personality characteristics are determined by sun signs. The various signs of the zodiac were classified

as negative or positive and each member of a large sample ($N = 2324$) was then classified into one of the two categories as determined by the birth date. All subjects (each of which had asked the senior author for astrological predictions) then completed a personality scale that provides a score on extraversion; the two groups were then compared on this measure. Sure enough, just as unequivocally predicted by astrology, the subjects born under the odd-numbered signs had higher scores on the extraversion scale. It was claimed that the sun sign effect was not confounded by subjects' knowledge of astrology because such knowledge was "controlled."

This design and the associated claim that confounding variables were controlled are typical of many studies. The skeptic, however, should not be content with the statement that potential confounding was controlled. The question should be: "Exactly how was it controlled?" The "control" in this study involved determining whether subjects had *knowledge of astrological principles*. This can be viewed as a fallible measure of whether the subjects had knowledge of their own sun sign (which is the key issue). Questions regarding knowledge of topics such as astrological chart reading and knowing one's sun sign may be reasonable items on a scale designed to measure the construct "astrological knowledge," but surely it is not necessary to be able to read an astrological chart in order to know one's sign. A large proportion of those classified as not having knowledge of astrological principles probably knew their sign and something about the purported characteristics of people born under the sign. This is the critical issue in the study. Misclassification of this type is a form of measurement error. If the astrological knowledge measurements in the study had been precise, a value of zero would have meant that a subject with this score had absolutely no knowledge of astrology and the result would not have been contaminated by this factor. Indeed, subsequent studies adequately controlled knowledge of the sun sign and the "effect" disappeared.

Although measurement error in continuous variables is usually the concern discussed in behavioral science presentations of measurement error, this example shows that dichotomous variables, such as indicators of exposure to some environmental condition and control variables, are also subject to the problem. The key point is that the method of measuring exposure conditions and control variables can have a major effect on the outcome of the study. It is not an idle question to ask what the outcome would have been if such variables had been measured without error.

Although measurement error correction procedures can provide better estimates of exposure effects, they do not lead to more powerful tests of exposure effects. Because my priority ordering is low bias description first and inference second, I recommend correction procedures. It appears from the general low level of enthusiasm for these methods in most areas, however, that others disagree. I view this as another manifestation of unreasonable reliance on the p-value as the critical outcome of a study.

27.4 MEASUREMENT ERROR CORRECTION IDEAS

Although the problem of measurement error has been acknowledged in the statistical literature for over a century, solutions for ANCOVA did not appear until about 1960.

Many approaches have been developed since then for all types of generalized linear models. The general idea associated with all of them is that in simple regression the uncorrected slope estimate is attenuated by measurement error; this attenuation is correctable given an estimate of the ratio of the measurement error variance on X over the total observed variance on X. The ratio

$$\frac{\sigma_T^2}{\sigma_T^2 + \sigma_{ME}^2} = \frac{\sigma_T^2}{\sigma_X^2} = \lambda,$$

where σ_T^2 is the true score variance on the covariate, σ_{ME}^2 is the measurement error variance on the covariate, and σ_X^2 is the total observed variance on the covariate. The parameter λ is sometimes called the reliability ratio.

Estimates of λ are usually called reliability coefficients in the behavioral sciences. The estimates may be based on a separate reliability study in which the population relevant to the main study is sampled and the subjects are tested on instrument X twice. Ideally the time interval separating the two measurements (X_1 and X_2) is of the same duration as occurs between pretesting and posttesting in the main experiment. Either the correlation between the two measurements or the slope from the regression of X_2 on X_1 can be used to estimate the reliability coefficient. These two approaches are equivalent under the assumption that the variance is same for X_1 and X_2. (Recall that when the variance on the predictor and the outcome variables are the same the correlation coefficient is equal to the slope.) Other approaches for estimating λ are available; some combine data from the reliability study with those from the main study. Also, a cleaver computationally intensive measurement error approach known as SIMEX (simulation extrapolation) has been developed (Carroll et al., 2006); it provides reliability estimates and descriptions of effect estimates under different assumed values of the reliability ratio. The agreement among most estimation procedures is usually high.

In the context of simple regression the slope estimate b_1 is attenuated by measurement error in X. This "naïve" or uncorrected estimate is closer to zero than would be found if no measurement error were present. Therefore it underestimates the size of the effect of X. If $\hat{\lambda}$ is available, it is possible to estimate what the slope would be if it were based on the true X scores rather than the observed fallible X scores. The corrected slope is easily computed using $b_1/\hat{\lambda}$. Because $\hat{\lambda}$ cannot be greater than 1.0, the corrected slope is always larger in absolute value than is the uncorrected value unless there is no measurement error.

Measurement Error Correction for ANCOVA Models

Because ANCOVA regression models require two or more predictor variables (i.e., the treatment or exposure indicator plus the covariates), the measurement error corrections are more involved than in the simple regression case. When more than one predictor is involved the corrected partial regression coefficients may be larger, smaller, or even of different sign than the naïve estimates; the corrections depend upon the correlations among the variables and correlations among the errors.

Just as the naïve slope β_1 in the simple regression model can be corrected using β_1/λ to obtain the measurement error corrected population slope (and the sample naïve slope b_1 can be corrected using $b_1/\hat{\lambda}$), the vector of naïve partial regression coefficients $\mathbf{B}_{\text{Naive}}$ in the multiple regression model can be corrected for measurement error using

$$\mathbf{\Lambda}^{-1}\mathbf{B}_{\text{Naive}},$$

where

$$\mathbf{\Lambda} = \left[\sum\nolimits_T + \sum\nolimits_{\text{ME}}\right]^{-1} \sum\nolimits_T,$$

where

$\sum_T$ is the variance–covariance matrix for the set of predictor variable true scores; and

$\sum_{\text{ME}}$ is the variance–covariance matrix for the measurement errors associated with the set of observed predictor variables.

The corrected variance–covariance matrix is defined as

$$\mathbf{\Lambda}^{-1}\text{Var}(\mathbf{B}_{\text{Naive}})[\mathbf{\Lambda}^{-1}]^{\text{T}},$$

where $\text{Var}(\mathbf{B}_{\text{Naive}})$ is the variance–covariance matrix for the naïve partial regression parameters in the ANCOVA regression model.

Although the corrected partial regression coefficients are straightforward to compute (given $\mathbf{\Lambda}$), the computation of the associated standard errors is quite involved. The standard error for the ANCOVA adjusted treatment effect must be modified to acknowledge (1) the measurement error correction to the naïve adjusted treatment effect estimate, (2) the sampling error in the estimate of the reliability ratio used in the correction, and (3) the imprecision in the naïve adjusted effect estimate. Methods for doing so are not described here. The estimation of corrected coefficients and the associated inference is best left to specialized software for measurement error models.

Software

Routines for measurement error corrected regression models are not available in most popular software packages. An exception is *STATA*, which contains an extensive and refined set of routines for this purpose; it runs on most operating systems. Software for the logistic regression version of the calibration method for measurement error estimation is available at http://www.hsph.harvard.edu/faculty/spiegelman/multsurr.htm. *R* routines are starting to appear but I am unaware of evaluations of them. Structural equation model routines can sometimes be used, but my preference is to rely on dedicated measurement error routines.

After the corrected model is estimated it is appropriate to apply diagnostic procedures to evaluate the extent to which it appears that the assumptions of the model are

met. Details regarding estimation procedures, reliability studies, and assumptions are available in the references cited earlier.

27.5 SUMMARY

Measurement error always has an effect on ANCOVA but it does not lead to misestimated effects in all cases. Measurement error in the dependent variable does not lead to bias in the estimates of the adjusted means. The major consequence in this case is a reduction in the power of the ANCOVA F-test.

Measurement error in the exposure variable and the covariates causes bias in the case of nonequivalent group designs; this bias can be serious. If the ANCOVA model has an infallible treatment variable and one fallible covariate, the effect of measurement error in X is to reduce the absolute size of the coefficient associated with the fallible measure. When multiple covariates are involved the effect on the coefficients depends on the covariance structure of the covariates. Measurement error corrected ANCOVA provides an estimate of the outcome that ANCOVA would yield if hypothetical true (error-free) scores were employed as the covariate rather than observed (fallible) scores. The reduction in bias is accompanied by an increase in error variance that results from the correction procedure and from uncertainty in the parameter estimates that are used to compute the corrections. Specialized measurement error software is recommended for estimating these models.

CHAPTER 28

Design and Analysis of Observational Studies

28.1 INTRODUCTION

Although randomized experiments are claimed to be the gold standard for evidence of causal effects in many areas, this holds only when the implementation of the design is pristine and problems such as postrandomization dropout and treatment adherence are well measured or controlled. Even in situations where such problems are unlikely to occur, there are strong reasons to reject a randomized design. Many important questions simply cannot be investigated using randomized designs for practical, ethical, or scientific reasons. For example, a suspected carcinogen would not qualify as an acceptable condition to be investigated in a randomized human experiment, for obvious ethical reasons.

As a second example, suppose a new drug is to be evaluated for both efficacy on a chosen outcome variable and for possible side effects. Randomized studies are often very expensive and consequently they often have relatively small sample sizes. This makes it difficult to identify rare side effects of new treatments. In contrast, less expensive but much larger observational studies often facilitate the detection of rare events.

I begin this chapter with a brief review of the randomized design, the conditions under which unbiased estimates can be obtained in the absence of random assignment, and a description of nonrandomized designs and observational studies. Rubin's causal model and methods for designing observational studies that are consistent with this model are described next. Finally, new methods for the analysis of observational studies that have been designed using propensity scores are described and illustrated using data from a large observational study.

The Analysis of Covariance and Alternatives: Statistical Methods for Experiments, Quasi-Experiments, and Single-Case Studies, Second Edition. Bradley E. Huitema.

Randomized-Group Designs

In the case of a randomized-group equal sample size design, the investigator knows exactly the treatment assignment probability for each subject. That is, the probability of assignment to the treatment rather than the control condition is .50 for each subject, regardless of the subject's observed covariate scores or unmeasured characteristics. Advantages of this type of assignment are that the groups are probabilistically equivalent on all variables before treatments are implemented. This leads to simple statistical analyses, mean comparisons that are meaningful, and causal statements that are justified.

Unbiased Effect Estimates in the Absence of Randomization

Departures from random assignment in the design usually lead to groups that are not equivalent to begin with. This is a problem because the observed mean difference on the outcome variable is difficult to interpret unless one knows how large the difference would have been without a treatment effect. This difference is known to be zero in randomized studies (assuming no missing data or other design flaws) because the probability of assignment to the treatment is known to be .50 for each subject. But in nonrandomized studies the probability of assignment to treatment is not .50 for each subject and the expected difference between means is not zero. In these situations it will not be possible to provide an unbiased estimate of the treatment effect *unless* the probability of assignment to treatment is known.

This implies that it is possible to obtain unbiased treatment effect estimates without random assignment if the assignment is known. Therefore the key to unbiased effect estimation in nonrandomized studies is a model of the selection process. One class of design without random assignment is discussed in Chapters 18 to 21; single-case designs capture the assignment process with the time variable. A second design in which the process of assignment to treatment is perfectly known is the intentionally biased assignment design. Both single-case designs and intentionally biased designs have relatively straightforward analyses *because* the assignment process is known. In contrast, the observational study is more difficult to analyze because the assignment process is usually not easily identified. Much of this chapter is devoted to this problem.

Intentionally Biased Assignment

If scores on some covariate X are available early in the design phase the investigator can pick a cutoff score on X and assign subjects to the treatment condition if they exceed this score or to the control condition if they do not. This is especially convenient if the covariate is a pretest measure of need for the treatment. Examples include LDL cholesterol level, systolic blood pressure, and depression scores. When this approach is followed the dependent variable is usually a posttest measure of the same characteristic measured by the pretest covariate. This design makes it obvious that the subjects are assigned in an intentionally biased manner; that is, they are intentionally assigned to treatment *only* if they have a high value on X, so the treatment and control groups cannot possibly be equivalent.

Because pretests are almost always highly correlated with posttests the groups are expected to differ on the posttest in the absence of treatments. Because subjects are assigned to the treatment condition *exclusively* on the basis of the observed covariate there is no ambiguity regarding the probability of treatment assignment for any subject. That is, the probability of being assigned to the control condition is zero for subjects with X scores below the cutoff and the probability is one for subjects who exceed the cutoff score. Although the comparison of means is irrelevant with this design, a straightforward method of analysis applies because the exact assignment process is known. ANCOVA (or the regression analog) using the pretest assignment variable X as the covariate provides an appropriate solution as long as the data reasonably conform to the ANCOVA assumptions.

The issues of concern are essentially the same as those described in the single-case design chapters (18–21) except that autocorrelated errors are not an issue here. A major difference between the two designs is that the predictor variable is time in the single-case design whereas it is the pretest in the biased assignment design. The focus of the analysis of both designs is on the discontinuity in the regression function that occurs near the cutoff score on the predictor variable. Because the estimate of the treatment effect is measured by the discontinuity, it is essential that the functional form of the regression be correctly specified. Hence, a close inspection of residuals is a critical step in this analysis; undetected nonlinearity can lead to large bias in treatment estimates. Excellent applied references on this design (which is called the regression-discontinuity design throughout the behavioral sciences) are Shadish et al. (2002) and Trochim (1984, 1990). A more technical presentation can be found in Imbens and Lemieux (2008).

Other Nonrandomized Designs

When the process through which subjects end up in the two groups is *not* completely known, there is no simple way (that is known to be unbiased) to adequately determine what the mean difference would be in the absence of treatments; therefore, there is no simple way to provide unbiased treatment effect estimates. Two versions of this situation are listed in Table 28.1 along with randomized-group and intentionally biased assignment designs. Note in the last column that the best guess is that ANCOVA adjusted population means will differ using nonequivalent groups unless the reason for the nonequivalence is trapped by the covariate as is indicated for the biased assignment design. Additional details on these versions are provided next.

Nonrandomized Design: Selection Based on Covariate and Additional Covariate Relevant Information

This design can be viewed as a biased assignment design gone wrong. In this situation the subjects are initially assigned to treatments on the basis of their covariate scores; then exceptions are made concerning the criteria for selection. Suppose that the biased assignment approach is involved in the assignment to treatment 1 of all subjects with X scores below the median and assignment to treatment 2 of all subjects with X scores

Table 28.1 Adjusted Difference When No Treatment Effects Are Present: Four Pretest-Posttest Designs

Design	Expected Pretest Mean Difference	Expected Posttest Mean Difference	Expected Posttest Adjusted Mean Difference
Randomized groups: covariate unaffected by treatment	0.0	0.0	0.0
Intentionally biased assignment: groups formed on basis of pretest (covariate) scores	$\neq 0.0$	$\neq 0.0$	0.0
Nonequivalent groups: formed on basis of covariate and unknown additional relevant data	$\neq 0.0$	$\neq 0.0$	$\neq 0.0$
Nonequivalent groups: formed on basis of unknown selection factors	$\neq 0.0$	$\neq 0.0$	$\neq 0.0$

at or above the median; then a second (unplanned) stage of assignment is initiated in which it is decided that certain additional information will be used to make the assignment decision. This second stage creates problems; note in the table that the expected difference between adjusted means is not zero. A hypothetical example might help to clarify why the second stage causes problems.

Suppose that a compensatory reading program is to be evaluated. Subjects are pretested on a short and fairly unreliable reading test before the program begins. The investigator then assigns students to the compensatory reading program if they score below the median on the pretest; those who score at or above the median are assigned to the traditional program. Then, for unanticipated political reasons, it is decided to consider additional information concerning students' reading skills by an examination of (1) teachers' judgements, (2) parents' evaluations, and (3) students' self-evaluations. These various sources of information are considered in deciding which students are assigned to the compensatory program and which are assigned to the traditional program. There is not, however, a clear-cut rule for combining the information from these additional sources. If ANCOVA is applied by using the originally administered pretest as the only covariate and another reading test as the dependent variable, this analysis is unlikely to yield an unbiased result. This design is not a true biased assignment design because the pretest is not the *only* basis for assigning subjects to treatments. The analysis problem here is that ANCOVA will adjust the means as they should be adjusted under the condition that the covariate *as measured* is the *only* variable involved in the assignment process. When subjects are *actually* assigned on the basis of information that differs from that provided by the reading pretest alone, the ANCOVA will underadjust the means; consequently the expected mean difference is not zero when there are no treatment effects. Again, the problem is that the covariate does not contain all the information concerning the assignment process.

The preferable alternative is to identify the process through which subjects are actually assigned to groups. In the example of the compensatory reading program it was pointed out that the different sources of information included in the assignment process were (1) the pretest, (2) teacher's evaluations, (3) parents' evaluations, and (4) students' self-evaluations. These sources were combined in some unknown manner to make assignment decisions. What is required is a procedure to capture the decision-making strategy that was employed using these sources. The question is, "How were the four sources of information combined in making the assignment decision?" If the information from the four sources was combined into some known "composite score" during the decision-making process, each subject's composite score, if identified, can be used as the covariate in a conventional ANCOVA. If the composite contains all the information used to make the assignment decision, the use of this composite with ANCOVA will yield unbiased treatment effect estimates. Unfortunately, a well-defined method of combining the data from different sources is seldom employed.

Observational Study: Unknown Selection Factors

The observational study involves groups that have been formed on the basis of unknown selection factors. One simply has dependent variable scores from subjects who either were exposed to a specific treatment condition or were not exposed to the condition. Although much pretest data may be available, the process through which subjects end up in the two groups is unknown.

28.2 DESIGN OF NONEQUIVALENT GROUP/OBSERVATIONAL STUDIES

The previous section pointed out that ANCOVA is unlikely to provide unbiased estimates of treatment effects when the covariate does not include complete information relevant to the formation of the groups. If no information other than the dependent variable scores is available there is no satisfactory analysis other than simple description. The appropriate response to studies that simply present the difference between means on Y (with no other information regarding the nature of the groups) is somewhere between "That's interesting" and "So what?".

When multiple covariates are available to describe the characteristics of the groups before treatments are applied, satisfactory analyses are sometimes possible. Several approaches are possible in this situation; two of them are described here. The traditional approach is to employ many covariates in a multiple covariance analysis. In the case of very large samples that differ little with respect to the covariates this is often a satisfactory solution. But an alternative approach should be considered.

Design a Meaningful Comparison

Rather than attempting to analyze all the observed data from two nonequivalent groups using multiple ANCOVA with many covariates, an alternative is to redesign

Table 28.2 Covariate Means for Treatment and Control Groups on 11 Covariates: LTL Data ($n_t = 310$ and $n_{fc} = 30{,}025$)

Variable	Treatment ($n = 310$)	Full Control ($n = 30{,}025$)
Gender (Proportion female)	.516	.560
Race (Proportion Caucasian)	.635	.899
Entry age	18.474	18.470
Alpha program (Proportion yes)	.087	.052
Overall ACT score	18.326	21.081
English ACT score	18.206	20.421
Mathematics ACT score	16.684	20.400
Reading ACT score	17.635	20.326
Science ACT score	19.990	22.564
High School GPA	2.637	3.059
Entry year	88.729	88.754

the study so that the control group is approximately equivalent to the treatment group; then use some version of ANCOVA to analyze the data from the redesigned study. Consider the following example.

An observational study was recently carried out to determine whether a program intended to improve college graduation rates and academic achievement actually did so (Huitema and McKean, 2005; Kosten et al., submitted). This program, known as Learning to Learn (LTL), was implemented during a period of 10 years. The "treated" program participants were 310 at risk undergraduate students and the potential control sample included 30,025 students who did not participate in the program. Eleven covariates believed to be relevant were identified by subject matter experts and measurements on them were obtained for both treatment and control samples. Table 28.2 shows the treatment and control mean scores on the covariates.

An inspection of this table reveals obvious differences between the treated and control subjects. Formal tests provide unambiguous evidence that the observed differences on most of the covariates are far outside the range of sampling error (e.g., ANOVA F-values on Gender ACT, Overall ACT, and Mathematics ACT are approximately 118, 24, and 33, respectively). It would be naïve to interpret differences between these groups on an outcome measure as providing meaningful estimates of treatment effects. In contrast, if a control group existed that had covariate means (on all 11 variables) that were the same as those in the treatment group a strong case could be made for a meaningful outcome comparison.

One approach to achieving the desired equivalence might be to consider exactly matching the groups on all of the covariates. In general, the goal of exactly matched groups is unrealistic wishful thinking. Consider the simple situation where each covariate is dichotomous. If we have a study containing 30 covariates the number of subjects required to have perfect matches is $2^{30} = 1{,}073{,}741{,}824$. To make matters worse, in the case of variables that are approximately continuously scaled the required number approaches infinity. Fortunately, perfect matches are not required for valid comparisons.

Suppose the LTL study had been set up as a randomized experiment. What would the two columns of covariate means in Table 28.2 be expected to look like in this case? Recall that random assignment does not guarantee exact equivalence in any particular application. Chance differences are always anticipated. Hence, even though randomization yields groups that are equal in expectation, the presence of sampling error produces differences between sample means. Therefore we should expect variation between the covariate means (within sampling error) even though an ideal randomized group design is used.

But obtaining groups in an observational study that differ no more than anticipated by sampling error might also seem like wishful thinking. Fortunately, this is not necessarily true. Although population covariate means on the groups initially observed in an observational study are almost certainly different, the study often can be refined (using methods described subsequently) so that the ultimate analysis will not be biased by these differences.

Traditional Design and Analysis of Observational Studies

In a traditional observational study the data analyzed are the same as those collected. That is, the design is considered fixed at the outset in the sense that the comparison groups and the associated outcome measures are established as soon as the treated and untreated participants are sampled. The conventional analysis is multiple ANCOVA (or the equivalent regression analog). For example, if the LTL data were analyzed in this manner, multiple ANCOVA using 11 covariates would be applied; of course this would require dependent variable scores on all 30,335 participants.

Experimental Design View of Observational Studies

An alternative to the conventional ANCOVA/regression approach is to view observational research from the perspective of experimental design. Rather than simply sample treated and untreated participants and then compute a model-based analysis such as ANCOVA, an attempt is made to first carefully design the study in a manner that approximates a true experiment. Because random assignment cannot be involved in forming the comparison groups, an approximation to it that has the same goal (i.e., balanced covariates) is substituted. This requires a new design aspect known as the propensity score (Rosenbaum and Rubin, 1983, 1984).

There is considerable confusion among researchers first exposed to propensity score analysis because it is misleadingly referred to as a competitor to ANCOVA for analyzing observational data. Indeed there are many published studies that claim to compare propensity score analysis with ANCOVA. This implies that both procedures are used to analyze the outcome measures. When a researcher asks, “How should I analyze my observational study data?” he or she wants to know what analysis should be applied to the dependent variable scores. Propensity score analysis is *not* the answer to the question; propensity score analysis is *not* a method of analyzing dependent variable scores. Dependent variable scores are not at all a part of propensity score modeling; indeed this is an important advantage of propensity modeling.

Table 28.3 Comparison of Steps for Experiment and Observational Study

Randomized Experiment	Designed Observational Study
1. Randomly select participants from a population of untreated subjects.	1. Randomly select participants from a treated population.
2. Obtain covariate scores if convenient.	2. Obtain scores on many covariates; this is essential.
3. Randomly assign participants to treatments.	3. Assemble (with the aid of propensity score modeling) a control group from a large reservoir of participants who are similar to the treated participants.
4. Administer treatments.	4. _____
5. Obtain outcome measurements.	5. Obtain outcome measurements.
6. Analyze outcome data using t, ANCOVA, or some robust analog.	6. Analyze outcome data using t, ANCOVA, or some robust analog.
7. Generalize results to a population of the untreated.	7. Generalize results to a population of the treated.

Propensity scores are used in the *design* of observational studies and a conventional method such as *t*, ANCOVA, or a robust analog is used in the *analysis* on the dependent variable.

A direct comparison of the randomized experiment and a designed observational study is shown in Table 28.3 to indicate the stage at which propensity scores play a role.

Ideally the first step in both designs involves random selection from a defined population; this is usually unrealistic. Typically one simply identifies an accessible group in which each participant meets the selection criteria. If a randomized experiment had been designed to evaluate the LTL program, untreated university students would have been randomly assigned to receive either the treatment or the control condition. Instead, the program was simply made available, covariate information was obtained for all who participated in the program, and outcomes were measured. The group of students who participated in the program were defined as the *treatment* group.

The second step involved obtaining covariate scores for all same year students who were not exposed to the program; this group was labeled as the *full control* group.

The full control group of 30,025 students was quite heterogeneous on most covariates. The third step of the design involved identifying the students in the full control group who were most similar to the treated students. This was accomplished using propensity scores (described in detail subsequently). After the collection of control students most similar to the treated students was assembled, it was labeled as the *matched control* group. Note in Table 28.3 that this third step is the analog to random assignment in the randomized-group design.

It can be seen that step 4 in a randomized experiment involves treatment application; of course the treatment had been applied much earlier in the sequence of steps for the observational study so a blank is listed at this stage for this design. Step 5 is the same for both designs; outcome measures are obtained.

Step 6 involves the final analysis on the outcome variable; that is, the comparison of the treatment and control groups on the dependent variable. There are several choices for this analysis regardless of the design. Usually some form of ANCOVA is desirable because groups formed using either random assignment or propensity methods will be somewhat different on the covariates and it is appropriate to adjust for them.

Step 7 involves the generalization of results. The results of the final analysis of the observational study generalize to a population of treated participants. This differs from the population to which results of a randomized-group design generalize. If a randomized-group design had been used, a sample of untreated participants would have been randomized to treatment and control conditions; consequently the results would have generalized to the population of untreated students.

This general overview of the design of observational studies is not explicit regarding the nature of the propensity score. Details regarding propensity score computation, rationale, and benefits in observational studies are described next in the context of Rubin's causal model.

Rubin's Causal Model and the Role of Propensity Scores in Causal Inference

An understanding of causal inference in both experiments and observational studies begins with an explicit definition of a causal effect as the difference between two potential outcomes (Rubin, 1974, 2005). That is, the causal effect of a treatment on an individual subject is conceptualized as the difference between the value of the outcome if the subject is treated at time *t* and the value of the outcome if the subject is not treated at time *t*. Although it is impossible to "know" the effect for an individual subject (because it is impossible for a subject to be exposed to both the treatment condition and the control condition at exactly the same time *t*) this definition of the effect is the essential conceptual foundation in Rubin's causal model.

In practice, each subject will actually be exposed to only one condition, but it must be reasonable to imagine observing the subject under both the treatment and the control conditions. Because a causal effect is conceptualized for each subject in a study, one can also conceptualize the average of these individual causal effects. The purpose of the statistical analysis is to estimate the average causal effect. This effect can be estimated in the context of experiments, quasi-experiments, and observational studies. But this is possible only if a justifiable method is available to estimate the value of the missing potential outcome for each subject.

Need for Propensity Score Methods

In the case of a randomized-group experiment the estimation of the average causal effect is straightforward. It involves simply computing the difference between the mean of the outcome scores obtained from the subjects who were assigned to and measured under the treatment condition and the mean of the outcome scores obtained from the subjects who were assigned to and measured under the control condition. This simple approach provides a valid estimate of the average causal effect *because the*

treatment assignment is independent of subject characteristics in a correctly executed randomized-group experiment. Recall that the probability that a given subject will receive the treatment in a two-group randomized design is .50. Regardless of the characteristics of any subject, the probability of assignment to the treatment condition is always .50. Although one of the two potential outcome scores is missing for each subject, it can be argued that the unobserved scores are independent of all possible subject characteristics. Consequently these data are missing completely at random; it follows that the average difference between the observed treatment and control scores is an unbiased estimate of the average causal effect. The ease of estimating the average causal effect in a randomized experiment stands in stark contrast to the difficulties of obtaining causal meaning from data provided by an observational study.

Although the difference between treatment and control means is almost always computed and reported (regardless of the design), this difference is not useful as an estimate of the average causal effect if the data have been collected using a quasi-experiment or observational design. Without random assignment, the treatment-control mean difference is unlikely to be an unbiased estimator of the average difference between potential outcomes measured under treatment and control conditions. This problem occurs because without random assignment it can no longer be argued that the missing potential observations are missing at random. Because the treatment and control groups are very likely to differ with respect to many subject characteristics it is also very likely that these characteristics will be related to differences on the outcome measure. Hence, the difference between treatment and control means confounds possible treatment effects with differences between treatment and control groups that exist before different conditions are applied.

Capturing the Assignment Mechanism: Propensity Scores

As emphasized in an earlier section, the issue of how subjects end up in or are assigned to treatment and control groups is crucial in deciding how to estimate causal effects. Single-case and intentionally biased assignment designs use time and pretest scores, respectively, in deciding treatment assignment. Because time and pretest scores capture the assignment mechanism, the estimation of causal effects is relatively straightforward with these designs because the assignment model provides an estimate of the missing potential outcome. But observational studies are another matter because the assignment mechanism is not known *a priori*.

The role of propensity score analysis is to model the complex assignment mechanism that is likely to be required to describe how participants end up in the nonequivalent groups to which they belong. Although the propensity approach used to assemble the comparison groups for an observational study is more cumbersome than is random assignment in a true experiment, it is well worth the trouble.

Definition of the Propensity Score

A propensity score is a type of conditional probability; it is defined as $p = p(\mathbf{X}) = \Pr(z^*|\mathbf{X})$,

where

z^* denotes the observed group membership;
$z^* = 1$ for treatment subjects and $z^* = 0$ for control subjects; and
X denotes a matrix of observed covariates.

The **X** (observed subject characteristics) and z^* (observed group membership) have conditionally independent distributions given their computed numerical value on the propensity score, p. So the covariate distribution in the treatment group is the same as the covariate distribution in the matched control group for all covariates observed and included in the propensity model.

The propensity score model is usually estimated using logistic regression; in this case it can be written as

$$\Pr(z* = 1|X_1,\ X_2, \ldots X_m) = \{1 + \exp[-(\beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_m X_m)]\}^{-1},$$

where

z^* is a binary (0, 1) variable indicating treatment conditions,
$X_1, X_2, \ldots, X_m$ are observed subject characteristic (covariates), and
$\beta_1, \beta_2, \ldots, \beta_m$ are the partial regression coefficients.

In practice, the propensity score estimate for individual i ($\hat{p}_i$) is obtained from the vector of fitted logistic regression coefficients (**b**) and the matrix of covariate scores ($\mathbf{X}_i$) using

$$\hat{p}_i = \left[1 + \exp(-\mathbf{b}'\mathbf{X}_i)\right]^{-1}$$

Each predicted propensity score ($\hat{p}_i$) is interpreted as the conditional probability of treatment-group membership given the set of observed subject characteristics ($\mathbf{X}_i$).

The essential idea is to compute the estimated propensity score for each treated and full control group participant and then form either subgroups or matched pairs using these scores; each subgroup or pair must include both treated and untreated subjects. This will be possible only if the treatment and control distributions of the propensity scores overlap substantially.

An important advantage of the propensity approach is that it immediately alerts the researcher when there is an insufficient basis for estimating causal effects. The first thing that should be done after estimating propensity scores is to plot them for each group on the same graph. A complete lack of overlap of the two propensity distributions indicates that there are no data on which to compute causal effects. Hence, in this case, propensity scores inform the researcher that causal inference should be abandoned because there are no data to support such an inference. When reasonable overlap exists, the researcher should proceed to form subgroups or (preferably) matched pairs of treated and control subjects.

Subgrouping

When propensity scores are used for subgrouping, propensity bins (usually five) are formed by dividing the observed propensity continuum into fifths; each bin should contain a reasonable number of both treated and control subjects. After the bins are formed the associated treatment and control outcome scores are identified for each one; then a simple test of the difference between the dependent variable means (described in the next section) is carried out. Additional detail on subgrouping can be found in Rubin (1997).

Matching

Many matching algorithms are available to match each treated participant with a control participant on the basis of the propensity score. Matching is slightly more complex than subgrouping, but it usually leads to greater bias reduction in the final (outcome) analysis.

Matching was used to design the LTL study; the specific method was a routine known as *one-to-one matching with replacement*. It has been shown that this method provides greater bias reduction than simply matching each treated subject with the closest remaining control subject (Dehejia and Wahba, 2002). Control participant j is matched to treatment participant i if

$$|\mathrm{p}_i - \mathrm{p}_j| < |\mathrm{p}_i - \mathrm{p}_k|,$$

where

p is the propensity score;
$k = 1, \ldots, j-1, j+1, \ldots, n_{fc}$, and
n_{fc} is the number of control participants in the full-control sample.

It is possible that a control participant will be matched to more than one treated participant because the matching is done with replacement. It is also possible to have two or more control participants with the same propensity score; in this case the matched participant is randomly selected from the set of participants with identical propensity scores. When this routine was applied to the LTL data the $n_{\mathrm{t}} = 310$ treated participants were matched to $n_{\mathrm{mc}} = 294$ matched-control participants. Hence, there were 16 instances in which a control was matched to more than one treated participant. The within-pair matches in this study are so close that no differences are visible using conventional graphics. Of course this study had the luxury of a large reservoir ($n_{fc} = 30{,}025$) of potential controls from which to select matches.

Properties of Observational Studies Designed Using Propensity Scores

Recall that the most important property of a true randomized group experiment is that randomization yields groups having approximately balanced covariate distributions. Remarkably, an observational study based on groups formed using propensity scores will usually turn out to be well-matched on *all* of the observed covariates. Indeed, under some conditions the covariate balance achieved using propensity score methods tends to be somewhat better than that achieved using randomization.

Table 28.4 Covariate Means for Treatment and Propensity Matched Control Groups on 11 Covariates: LTL Data ($n_t = 310$ and $n_{mc} = 294$)

Variable	Treatment ($n_t = 310$)	Matched Control ($n_{mc} = 294$)
Gender (Proportion female)	.516	.544
Race (Proportion Caucasian)	.635	.619
Entry age	18.474	18.306
Alpha program (Proportion yes)	.087	.102
Overall ACT score	18.326	18.500
English ACT score	18.206	18.422
Mathematics ACT score	16.684	16.755
Reading ACT score	17.635	17.830
Science ACT score	19.990	20.102
High School GPA	2.637	2.657
Entry year	88.729	88.759

The skeptical reader may be wondering if propensity score matching "worked" in the case of the LTL study. Recall that the data presented in Table 28.2 reveal substantial differences between treatment and control means on most covariates and that these differences cannot be explained by sampling error (in one case an ANOVA F of well over 100 was obtained). Compare the values in that table with those in Table 28.4, which are based on treatment and propensity matched control subjects. It can be seen that the covariate differences between groups in this table are very small. Neither univariate nor multivariate tests on differences between treatment and control covariate means even hint that the groups are unbalanced. Indeed the two groups are very well-matched on all observed covariates. Other recommended descriptive statistics such as g (not shown here) confirm the effectiveness of the propensity score matching. These statistics support the decision to proceed with the final (outcome) analysis on the dependent variables.

The finding of excellent balance on observed covariates does not, however, lead to the conclusion that observational studies designed using propensity score methods are of the same quality (in terms of the adequacy of causal statements) as randomized experiments. Randomized studies are clearly superior in terms of bias reduction because they lead to probabilistic balance on *all possible* (i.e., both unobserved and observed) covariates. Propensity score methods generally lead to balance of covariate distributions within sampling error on all included covariates and approximate balance for variables not observed if they are highly correlated with those in the propensity model. But sometimes the desired balance is not achieved. When this occurs the researcher is warned that the groups should not be considered adequately balanced and that outcome comparisons are not appropriate.

28.3 FINAL (OUTCOME) ANALYSIS

After either subgrouping or matching using propensity scores is completed, the analysis of the difference between treatment and control groups on the dependent variable

is carried out. The specific analysis to use depends on the whether subgrouping or matching was used in the design. One method is described below for each design, although there are many other choices.

Final Analysis for Subgrouped Designs

The following formula for t is often used to evaluate the difference between weighted treatment and control means:

$$t = \frac{\hat{\Delta}}{\sqrt{Var\left(\hat{\Delta}\right)}} = \frac{\sum_{b=1}^{B}\left(\bar{Y}_{tb} - \bar{Y}_{cb}\right)\hat{p}_b}{\sqrt{\sum_{b=1}^{B}\hat{p}_b^2\left[\frac{s_{tb}^2}{n_{tb}} + \frac{s_{cb}^2}{n_{cb}}\right]}},$$

where

$\hat{p}_b = \frac{n_b}{N}$ is the bin weight associated with the bth bin (not to be confused with the propensity score p_i);
n_b is the number of subjects (treatment and controls combined) in the bth bin;
n_{tb} is the number of subjects in the bth bin exposed to the treatment condition;
n_{cb} is the number of subjects in the bth bin exposed to the control condition;
N is the total number of treatment and control subjects included in the comparison;
S_{tb}^2 and S_{cb}^2 are the sample variances in the treatment and control conditions, respectively, of the bth bin; and
B is the number of bins.

The critical value of t is based on $N - 2B$ degrees of freedom.

Final (Outcome) Analysis for Matched Designs

Just as there are many methods of matching there are many methods of performing a final analysis on the outcome data provided by the treatment and propensity matched control groups. Hill and Reiter (2006) evaluated several procedures including matched pairs, weighted two-sample, weighted least-squares, robust sandwich variance, bootstrapping, and Hodges–Lehmann aligned rank methods. No clear winner emerged in this study.

Eighteen methods, including all of those investigated by Hill and Reiter plus several new methods and a modification of a method previously proposed by Rubin (1979) were subsequently evaluated in a large simulation study (Kosten, 2010; Kosten et al., submitted). The latter study included three response surfaces, four different amounts of overlap between treated and control groups, and three error distributions (including normal, contaminated normal, and Cauchy). When the 18 methods were evaluated with respect to the average length of the confidence intervals and the proportion of the time the intervals captured the true treatment effect the clear winner was a new robust method "***w***", which is based on Wilcoxon scores; these scores are described in McKean and Sheather (1991). This is the method I recommend; it can be estimated using the robust regression software (*RGLM*) illustrated in Chapter 14.

A nonrobust analog to the ***w*** method that performs very well when outliers are not present is also recommended. This method (denoted as ***l***) is a block- and covariate-adjusted approach based on the same design matrix as the ***w*** method; it can be estimated using an OLS based model comparison approach.

Both the ***l*** and ***w*** methods include covariates for the same reasons that it is desirable to include covariates in a randomized-group design. Because small differences between groups on covariates are likely, it makes sense to adjust for those differences.

Design Matrix for l and w Methods

Each unique control subject that has been matched to at least one treated subject defines a block. This block contains a unique control subject and all treated subjects to which it is matched. The design matrix is of the form

$$[\,\mathbf{1}\ \mathbf{Tx}\ \mathbf{Bk}_2\ \mathbf{Bk}_3 \cdots \mathbf{Bk}_{n_{uc}}\ \mathbf{X}_1\ \mathbf{X}_2 \cdots \mathbf{X}_C\,],$$

where

1 is a unity vector;

Tx is a 1–0 dummy variable where "1" indicates the treatment subjects and "0" indicates the control subjects;

$\mathbf{Bk}_2$ through $\mathbf{Bk}n_{uc}$ are the set of $(k–1)$ dummy variables (0–1) that indicate the block to which a subject belongs (e.g., a 1 in column $\mathbf{BK}_2$ indicates that the subject belongs to block 2); the number of block dummy variables is one less than the number of blocks and the number of blocks is equal to the number of unique control subjects; and

$\mathbf{X}_1$ through $\mathbf{X}_C$ are the vectors for covariates 1 through C.

Example 28.1: Design matrix Suppose (unrealistically) that there are only five treated subjects and one covariate. The design matrix and response vector must contain 10 rows. The first five subjects are treated and the last five subjects are the matched controls. The example design matrix and outcome vector are as follows:

$$\mathbf{X} = \begin{bmatrix} 1 & 1 & 1 & 0 & 0 & 27 \\ 1 & 1 & 0 & 1 & 0 & 13 \\ 1 & 1 & 0 & 0 & 1 & 19 \\ 1 & 1 & 0 & 0 & 1 & 18 \\ 1 & 1 & 0 & 0 & 0 & 15 \\ 1 & 0 & 1 & 0 & 0 & 12 \\ 1 & 0 & 0 & 1 & 0 & 24 \\ 1 & 0 & 0 & 0 & 1 & 8 \\ 1 & 0 & 0 & 0 & 0 & 16 \\ 1 & 0 & 0 & 0 & 0 & 19 \end{bmatrix} \quad \mathbf{Y} = \begin{bmatrix} 12 \\ 3 \\ 7 \\ 8 \\ 5 \\ 9 \\ 12 \\ 2 \\ 12 \\ 13 \end{bmatrix}.$$

It can be seen that the third column has a 1 in rows 1 and 6; this means that the first subject in the treated group is compared with the corresponding matched control. The same pattern exists in column 4. But note in column 5 that there are two 1s in

the treatment group and one 1 in the control group. This means that the third control subject is matched to both the third and fourth treated subjects.

Method l Estimation

The conventional OLS solution based on the design matrix shown above yields an adjusted treatment effect estimate of −3.72 and a p-value of .04. The point estimate from this solution is correct but the p-value is not because the conventional error variance estimate does not acknowledge the reduction in degrees of freedom that occurs when a matched control is used more than once. Hence, the following model comparison procedure is recommended to compute the correct F-test:

1. Regress Y on the treatment dummy variable, all dummy variables indicating blocks, and the covariate(s) (one in this example). This analysis provides the full model regression sum of squares $\mathrm{SS}_{\mathrm{Reg\ (full)}}$.
2. Regress Y on all dummy variables indicating blocks, and the covariates. This analysis provides the restricted model regression sum of squares $\mathrm{SS}_{\mathrm{Reg\ (restricted)}}$.
3. Divide the residual sum of squares from the full model by modified degrees of freedom equal to $n_{\mathrm{u}} - k - C - 1$, where $n_{\mathrm{u}} = n_t + n_{uc}$ is the number of unique subjects in the analysis, n_t is the number of treated subjects, n_{uc} is the number of unique matched control subjects, and C is the number of covariates. In the example one of the controls was matched to two treated subjects so $n_{uc} = 4$, $n_u = 9$, and $df_{\mathrm{error}} = n_u - k - C - 1 = 3$.
4. Compute the adjusted treatment effect F-statistic using

$$\frac{\mathrm{SS}_{\mathrm{Reg\ (full)}} - \mathrm{SS}_{\mathrm{Reg\ (restricted)}}}{\mathrm{SS}_{\mathrm{Res\ (full)}}/df_{\mathrm{Error}}} = F.$$

5. Evaluate the obtained F using degrees of freedom = 1, $n_{\mathrm{u}} - k - C - 1$.

The application of this approach to the example data yields:

$$\frac{129.897 - 100.574}{14.203/3} = 6.19 = F$$

An F-statistic of 6.19 based on 1 and 3 degrees of freedom yields a p-value of .09.

Method w Estimation

The recommended robust ***w*** method can be computed using the *RGLM* software mentioned in Chapter 14. The full and restricted models are estimated using the same design matrix as described above; the robust test statistic F_R is computed using:

$$\frac{D_{\varphi(\mathrm{Restricted})} - D_{\varphi(\mathrm{Full})}}{\left[\sqrt{\dfrac{n_{\mathrm{u}}}{n_{\mathrm{u}} - k - C - 1}}\right]\left[\sqrt{\dfrac{N - k - C - 1}{N}}\right](\hat{\tau}_{\varphi}/2)} = F_{\mathrm{R}}$$

where

$D_{\varphi(\text{Restricted})}$ and $D_{\varphi(\text{Full})}$ are the restricted and full model residual dispersions;
$\hat{\tau}_{\varphi}$ is the scale parameter estimate from the full model (provided by *RGLM*), and
$N = 2n_{\text{t}}$.

The obtained value of the test statistic F_{R} is evaluated using the conventional F distribution with 1 and $n_{\text{u}} - k - C - 1$ degrees of freedom. The application of the ***w*** method to the data set listed above yields a robust treatment effect estimate of -3.66 and $F_{\text{R}} = 6.078$, which are similar to the values obtained using the ***l*** method.

Results of Six Outcome Analyses Applied to LTL Data

The first three analyses of the LTL outcome are shown in Panel A of Table 28.5. All three assume homogeneous regression hyperplanes. It can be seen that there are substantial differences among the three point estimates. The ***w*** method point estimate is zero whereas the ANCOVA estimate is .13, with zero outside the corresponding confidence interval. But the test for homogeneous regression hyperplanes (assumed for ANCOVA) identified heterogeneity ($p = .0045$); an analogous test for the robust

Table 28.5 Point Estimates and Confidence Intervals from Multiple Outcome Analyses Applied to LTL Data

A

Method	Point Estimate GPA Difference (Tx – Contol)	95% Confidence Interval	Treatment and Control Sample Sizes
w	.0000	(−.0040, .0040)	$n_{\text{t}} = 310$ $n_{\text{mc}} = 294$
Conventional ANCOVA (assume homogeneous regression)	.1284	(.0624, .1943)	$n_{\text{t}} = 310$ $n_{\text{fc}} = 30{,}025$
Robust ANCOVA (assume homogeneous regression)	.0749	(.0186, .1312)	$n_{\text{t}} = 310$ $n_{\text{fc}} = 30{,}025$

B

Method	Point Estimate GPA Difference (Tx – Contol)	95% Confidence Interval	Treatment and Control Sample Sizes
w	.0000	(−.0046, .0046)	$n_{\text{t}} = 310$ $n_{\text{mc}} = 294$
Conventional ANCOVA (assume heterogeneous regression)	.0419	(−.0430, .1267)	$n_{\text{t}} = 310$ $n_{\text{fc}} = 30{,}025$
Robust ANCOVA (assume heterogeneous regression)	.0128	(−.0599, .0855)	$n_{\text{t}} = 310$ $n_{\text{fc}} = 30{,}025$

ANCOVA yielded the same *p*-value. Consequently, heterogeneous regression versions of ***w***, ANCOVA, and robust ANCOVA models were estimated using the grand mean on each covariate as the point at which the effect was estimated. The results are shown in Panel B of Table 28.5. Note that the two ANCOVA point estimates of the adjusted treatment effect are much closer to zero than they were using the homogeneous regression model, and the confidence interval associated with each method includes zero.

28.4 PROPENSITY DESIGN ADVANTAGES

There are several advantages of designing an observational study using propensity scores over the traditional approach of simply applying multiple ANCOVA or regression to the complete data set.

Practicality/Sample Size

A major but largely unrecognized advantage of the propensity approach is that it can turn a prohibitively expensive project into a feasible one. Outcome measures in many research areas (especially medicine) are very expensive to obtain. The number of subjects required in the final analysis of a propensity designed observational study may be *far* less than the number required in the case of ANCOVA. ANCOVA and regression require covariate and dependent variable scores on all subjects in the study; this is not true with the propensity approach. Only covariate information, which is often inexpensive to collect, is needed to compute propensity scores. Recall that the propensity model involves the logistic regression of group membership on the covariates for all available control and treatment subjects. But all control subjects need not be measured on the dependent variable. This is the case because propensity scores are used to identify only the most appropriate (i.e., most closely matched) subjects in the reservoir of control subjects; these are the only control subjects that need be measured on the outcome measure. In the *LTL* study, dependent variable scores were required on less than one percent of the full-control sample.

Power

The precision associated with a well-matched small study is equivalent to the precision of a very large study using heterogeneous samples. This was demonstrated in the LTL study, even when using a less than optimum final outcome analysis. The standard errors for a dozen different analyses applied to the outcome data based on the 310 matched pairs were actually smaller than the standard error from the ANCOVA applied to all 30,335 subjects (Kosten, 2010).

Transparency

The concept of matching is generally much more transparent to a non-statistically oriented audience than is ANCOVA. An ANCOVA based on two or more covariates involves a comparison of the levels of two or more hyperplanes that are neither

easy to visualize nor to explain. In contrast, the results of the propensity approach (when used to form bins) can be illustrated using two histograms. These histograms illustrate (1) the number of treatment and control observations in each bin and (2) the treatment effects in each bin. The first histogram clearly illustrates the extent to which treatment and control data are available in each bin to make a meaningful estimate of the causal effect. This histogram immediately identifies whether a study can provide meaningful causal information. The second histogram illustrates the difference between treatment and control outcome means for each subgroup (bin).

Similarly, when propensity scores are used to form matched pairs (instead of bins) a display of the difference within each pair on the outcome variable plotted against the propensity score establishes the basis of the overall comparison. Both the subgrouping approach and the matching approach are ultimately evaluated by how well the covariates are balanced. A display illustrating similar covariate means for treatment and control groups on a large number of covariates is easily understood as a logical basis for arguing that the groups are reasonably equivalent.

Simple Detection of Model Inadequacies

ANCOVA/regression models with 50 to 100 covariates are not unusual in some research areas. The identification of model inadequacies can be challenging in this environment even with complex modern diagnostic aids that are routinely recommended. But, in the case of propensity models, there is little concern regarding issues such as possible covariate redundancy, understanding the coefficients, and related estimation problems because the focus is not on the interpretation of the coefficients. Instead, there is only one question to answer in evaluating the adequacy of the model: "Are the covariates balanced in the treatment and matched control groups?" This is easily answered.

Elimination of Bias Generated by *Post Hoc* Covariate Shopping

The propensity score approach eliminates possible bias that can accompany the practice of reporting only the most positive result among many different ANCOVA models that have been fitted to a single data set. Selecting a model based on the outcome of the analysis on the dependent variable is impossible with the propensity approach because the propensity model does not include the outcome variable. In many cases the outcome data will not even be available at the time the propensity model is estimated.

Robustness to Model Misspecification

It appears that the propensity approach is more robust to model misspecifications than is ANCOVA. It has been shown by Drake (1993) that the bias associated with the fitted linear logistic propensity model is much smaller than the bias associated with the corresponding ANCOVA model when the data are generated using quadratic models. The LTL results suggest that the same may be true regarding undetected heterogeneity of regression.

More Covariates

The number of covariates that can be included in the propensity model is not limited to the relatively small number appropriate for conventional ANCOVA adjustment. This limitation on the number of covariates has implications for bias reduction. If covariate information that is useful in adjusting the treatment effect estimate for pretreatment differences is omitted from the analysis, the treatment effect estimate will be biased.

Extrapolation Problem Identified

Because ANCOVA is a model-based approach it provides adjusted treatment effect estimates regardless of the difference between groups on the covariates. If the groups differ so much that there are no data at the grand mean of the covariates the effect is still estimated at this point. Consequently the effect estimate is vacuous in the absence of convincing outside evidence that the model is valid in the covariate range between groups. Rubin (2001) has clearly demonstrated that ANCOVA can yield large overestimates or underestimates when substantial differences exist in the covariate distributions.

Propensity matching reduces concern with these problems for two reasons. First, the presence of extrapolation is immediately revealed in the design stage when attempting to form subgroups or matched pairs. If there are no overlapping treatment and control propensity scores the absence of data for a meaningful comparison is obvious. Second, when partial overlap is present, comparisons will be possible only for that portion of the propensity distribution. If the analysis is pursued it will be clear to the researcher that the results should be generalized to only the restricted subpopulation defined by the matched propensity range.

Better Covariates

In some studies the selection of covariates to include in ANCOVA may be hampered by a lack of knowledge regarding the determinants of the outcome variable. In contrast, the research team may have a good understanding of covariates that are likely to be associated with treatment group membership. These are the most important covariates for propensity modeling. Well-selected covariates in propensity modeling are likely to lead to better control of bias than ANCOVA with a poor set of covariates.

28.5 EVALUATIONS OF ANCOVA VERSUS PROPENSITY-BASED APPROACHES

The statistics literature contains many studies that (1) compare results of ANCOVA and propensity methods applied to the same empirical studies and (2) simulate data under various assumed population situations and compare results of the two methods. The general finding, based on both actual studies and simulations, is that there is high

agreement of outcomes from the two approaches, but outcome analyses based on groups formed using propensity scores tend to detect slightly fewer effects than are identified using conventional ANCOVA.

One problem with many of these studies (both types) is that the "propensity score approach" to which they refer is usually based on subclassification and the outcome analysis is usually the stratified *t* approach. These are not the best ways to design and analyze. Matching is usually more effective than subclassification and the combination of this method of design with outcome analyses designed for matched data can be expected to provide higher power than the approaches used.

A second problem is that the data used for ANCOVA and propensity/outcome analysis were often the same. This suggests that the approach of finding a large reservoir of potential control subjects who are similar to the treated subjects was not followed. Consequently the studies were not designed to capitalize on a major strength of the propensity score approach to design.

It has been argued that a comparison of methods should be based on the same data in order to have an appropriate evaluation. The adequacy of this view may be questioned. Consider a study with a total of 300 subjects. All of them will be used for both ANCOVA and for the propensity model. Suppose there are large covariate differences between the treatment group and the initial control group. This is likely to mean that there will be subjects in the initial control group who cannot be used in the outcome analysis because no treated subjects have similar propensity scores. After all, the purpose of propensity modeling is to design more meaningful comparisons through subject selection. So even though the ANCOVA and propensity approaches start out with 300 subjects, the outcome analysis following propensity modeling may have only, say 220. Although the sample size is smaller the comparison may be more meaningful because like is compared to like.

If the adjusted mean difference computed using ANCOVA ($N = 300$) is compared with the average difference between means from the propensity design ($N = 220$), should they be the same? Not necessarily.

ANCOVA estimates the difference between means for subjects falling at the *grand* covariate(s) mean. In contrast, the propensity approach identifies subjects whose outcome scores are used as estimates of the potential outcomes that are missing for the *treated* subjects. The mean difference between the outcome scores for the treated subjects and the outcome scores for the propensity identified control subjects is the causal effect estimate. This is sometimes referred to as the effect of the treatment on the treated. If the homogeneity of regression assumption is met in the ANCOVA model, then the adjusted mean difference from ANCOVA and the effect of the treatment on the treated are equal. But if heterogeneity of regression is present the two analyses do not estimate the same parameter. In this case the conventional ANCOVA model is misspecified; if an appropriate model that handles heterogeneous regression is applied the effect estimate should be similar to the estimate from the propensity designed data (as was the case in the LTL study). This may be taken as an argument for the use of picked-points analysis instead of the propensity approach because the sample size is smaller for the latter approach. However, other aspects such as transparency should be kept in mind when selecting the method to use.

28.6 ADEQUACY OF OBSERVATIONAL STUDIES

ANCOVA, measurement error corrected ANCOVA, and propensity based methods are good attempts to provide reasonable estimates of treatment effects from observational data. Regardless of the method of analysis, observational studies usually do not have the credibility of randomized experiments when it comes to support for causal statements. The main reason is that one never knows what has been left out of the model. Balanced covariates resulting from propensity modeling are encouraging indicators of an adequate model, but balance on the observed covariates does not mean that the study is free of unknown important covariates that will bias the effect estimate.

The best way to avoid major omissions in selecting covariates is to assemble the most knowledgeable research team possible. Because a goal in propensity modeling is to identify the process through which participants end up in the treatment group rather than the control, the model demands variables that are predictive of group membership. Consequently, team members who understand the differences between the characteristics of treated and untreated subjects are most likely to identify appropriate covariates to include in the propensity model. An additional source of information regarding critical covariates is the sample of subjects who were treated; ask these subjects why they decided to participate. Their responses should provide suggestions for variables that may distinguish treated and control participants.

In the case of ANCOVA, covariates that are most highly correlated with the dependent variable are of greatest interest. Different experts may be required to select covariates for each type of analysis. But, because the recommended propensity approach involves the use of ANCOVA for the final (outcome) analysis, it is ideal to identify and measure both types of covariate. Similarly, both types of covariate are useful when ANCOVA is used (Beach and Meier, 1989). Before any covariate is included one should be assured that the treatment has not affected it.

One way of attempting to evaluate the adequacy of observational studies analyzed using propensity approaches (and other methods) is to compare the results of randomized experiments with those of observational studies designed to evaluate the same treatments (e.g., Benson and Hartz, 2000; Concato et al., 2000; Ioannidis et al., 2001). Such comparisons have been reported in the literature for over two decades, but the conclusions are quite inconsistent; it could hardly be otherwise. The variation in implementations of competing analytic procedures reported in many of these studies is large. Some studies use very few covariates and linear functional forms only, whereas others use many covariates, nonlinear functions, heterogeneous regression models, and higher order interactions. In general, it appears that studies using many covariates and sophisticated models (whether propensity based or ANCOVA) tend to agree with randomized design results as long as the treated and control subjects are from similar populations. An interesting article by Shadish et al. (2008) along with associated comments by Hill (2008), Little et al. (2008), and Rubin (2008) contain references and nuances regarding the work in this area.

Because one never knows whether important hidden covariates have been left out of the model a formal approach known as sensitivity analysis is sometimes used to

estimate how much hidden bias would have to exist to change the conclusion of the statistical test on the outcome. Some studies are very sensitive to small biases whereas others are quite robust to substantial bias. These analyses are quite involved; the major reference on this approach is Rosenbaum (2002).

Advanced References and Alternative Methods

This brief chapter has provided only some of the basic ideas associated with observational study design and analysis. The definitive sources on the approaches to estimating causal effects in observational studies that are briefly described here are Imbens and Rubin (in preparation), Rosenbaum (2002, 2010) and Rubin (2006).

Several other methods are frequently recommended for the analysis of observational studies. Structural equation models (SEM) and instrumental variable (IV) analysis (sometimes viewed as a variant of SEM) are the most popular alternatives. I avoid them. The untestable asssumptions that must be made to justify causal conclusions with these approaches strike me as breathtaking. Others disagree.

28.7 SUMMARY

An observational study is frequently a necessary alternative to a randomized experiment. There are cases where randomized designs are impractical, unethical, or scientifically undesirable. When a large control sample can be drawn from the same population that has been exposed to the treatment it is often possible to select a refined control sub-sample for comparison purposes. This is accomplished using the propensity score, which is defined as the probability of receiving the treatment condition given the observed covariates. An estimate of this score can be computed for each participant in the treatment and full control groups. Each treated subject is matched to a control subject who has essentially the same propensity score. The groups of treated and control subjects resulting from the propensity score matching are usually well-matched on all the individual covariates. Ideally, an outcome analysis on the dependent variable scores is carried out using some version of ANCOVA. A robust blocked version of ANCOVA (denoted as ***w***) is recommended.

This design and analysis approach is an alternative to applying some version of multiple ANCOVA. Certain practical advantages such as the transparency of the analysis and robustness to flaws in the model are associated with the propensity based analysis. Most comparisons of these procedures find a high degree of correspondence between them in terms of the point estimates for the size of the effect estimate as well as the statistical decision.

CHAPTER 29

Common ANCOVA Misconceptions

29.1 INTRODUCTION

Major misunderstandings surround ANCOVA in terms of technical details, situations in which it should be applied, and its relationship to other methods of analysis. Most of these issues are described in other chapters, but three of them that do not conveniently fit elsewhere are described here. Twelve misconceptions that are covered in detail in earlier chapters are listed along with a brief comment regarding each one.

29.2 SS_{AT} VERSUS $SS_{Intuitive\ AT}$: SINGLE COVARIATE CASE

Several times a year I am asked the following question: "Why does the sum of squares for the adjusted treatment effects (SS_{AT}) not equal the sum of squares obtained when applying the conventional formula for the between-group sum of squares to the adjusted means? This is a very reasonable question. After all, if the expression

$$\sum_{j=1}^{J} n_j(\bar{Y}_j - \bar{Y}..)^2$$

defines the between-group sum of squares (SS_B) in ANOVA, then by analogy the sum of squares for adjusted treatments (SS_{AT}) in ANCOVA should be

$$\sum_{j=1}^{J} n_j(\bar{Y}_{j\,\mathrm{adj}} - \bar{Y}..)^2.$$

Because this expression follows intuition, I label it $SS_{Intuitive\ AT}$. If we apply this formula to the first example data in Chapter 6, we obtain a value of 723.67 rather than the correct SS_{AT}-value of 707.99; clearly, $SS_{Intuitive\ AT} \neq SS_{AT}$.

The Analysis of Covariance and Alternatives: Statistical Methods for Experiments, Quasi-Experiments, and Single-Case Studies, Second Edition. Bradley E. Huitema.

Why are they not the same? The brief answer is that the adjusted means are not independent of each other (as long as some adjustment has taken place). Recall that adjusted means are computed using $\bar{Y}_{j\,\text{adj}} = \bar{Y}_j - b_W(\bar{X}_j - \bar{X}..)$. Because the adjustments of all means are based on the same pooled within-group regression coefficient (b_W) the adjusted means will be slightly dependent unless b_W is equal to zero. More specifically, the intuitive formula does not consider the covariance between the treatment effects estimates and the covariate means. Although this covariance is assumed to be zero in a true experiment, it usually deviates slightly from zero with sample data even though the groups are randomly assigned (as in the example). The following equation shows the terms necessary to explain the difference between the intuitive and correct values:

$$\begin{aligned} SS_{AT} &= \sum_{j=1}^{J} n_j \left(\bar{Y}_{j\,\text{adj}} - \bar{Y}..\right)^2 - (b_W - b_B)^2 \left[\frac{\left(SS_{B_X}\right)^2}{SS_{T_X}}\right] \\ &= SS_{\text{Intuitive AT}} - (b_W - b_B)^2 \left[\frac{\left(SS_{B_X}\right)^2}{SS_{T_X}}\right], \end{aligned}$$

where

$\bar{Y}_{j\,\text{adj}}$ and $\bar{Y}..$ are the adjusted and grand means on the outcome, respectively;
b_W and b_B are the pooled within-group and between-group slopes, respectively; and
SS_{B_X} and SS_{T_X} are the between-group and total sum of squares on the covariate, respectively.

If these terms are computed for the example data we find:

$$\begin{aligned} SS_{AT} &= 723.6721 - [.5705 - (-1.8158)]^2 \frac{126.6667^2}{126.6667 + 5700} \\ &= 723.67 - 15.68 \\ &= 707.99 \end{aligned}$$

The difference between $SS_{\text{Intuitive AT}}$ and SS_{AT} of 15.68 points is explained by the covariance between the adjusted treatment effects and the covariate means.

If there were no differences among the covariate means it can be seen that the sum of squares between groups on the covariate would be zero, the ratio $\frac{\left(SS_{B_X}\right)^2}{SS_{T_X}}$ would equal zero, and there would be no correction of $SS_{\text{Intuitive AT}}$. That is, $SS_{\text{Intuitive AT}}$ would be equal to SS_{AT}. A more detailed description of this issue is presented in the next section in the context of multiple covariates. That description also applies to the single covariate case.

29.3 SS_{AT} VERSUS $SS_{Intuitive\ AT}$: MULTIPLE COVARIATE CASE

It was pointed out in the previous section that there is usually a dependency among adjusted treatment means and that this dependency is responsible for the discrepancy found between the adjusted treatment sum of squares (SS_{AT}) and the intuitive adjusted treatment sum of squares ($SS_{Intuitive\ AT}$). The latter was defined as

$$\sum_{j=1}^{J} n_j(\bar{Y}_{j\,\mathrm{adj}} - \bar{Y}_{..})^2.$$

That discussion was in the context of a single covariate. When multiple covariates are involved the same issue is present but the dependency among the adjusted treatment means is more complex because it is a function of all covariates in the model. Recall that in the case of multiple covariates an adjusted treatment mean can be computed using

$$\bar{Y}_{j\,\mathrm{adj}} = \bar{Y}_j - [b_{W_1}(\bar{X}_{1_j} - \bar{X}_{1..}) + b_{W_2}(\bar{X}_{2_j} - \bar{X}_{2..}) + \cdots + b_{W_C}(\bar{X}_{C_j} - \bar{X}_{C..})],$$

or, in matrix notation,

$$\bar{Y}_{j\,\mathrm{adj}} = \bar{Y}_j - \mathbf{d}_{x,j}^T \mathbf{b}_W.$$

Note that all adjusted means are based on the same vector ($\mathbf{b}_W$) of pooled within-group regression coefficients. If the covariate means and the adjusted means for the various groups covary, the resulting adjusted means will be somewhat dependent; this dependency is reflected in the discrepancy between SS_{AT} and $SS_{Intuitive\ AT}$. The difference between these two measures of variation is explainable from information regarding the covariance structure of $\bar{Y}_{j\,\mathrm{adj}}$.

The form of the population $J \times J$ variance–covariance matrix of the adjusted means is shown below:

$$\begin{bmatrix} \sigma^2_{\bar{Y}_{1\,\mathrm{adj}}} & \sigma_{\bar{Y}_{1\,\mathrm{adj}},\bar{Y}_{2\,\mathrm{adj}}} & \cdots & \sigma_{\bar{Y}_{1\,\mathrm{adj}},\bar{Y}_{J\,\mathrm{adj}}} \\ \sigma_{\bar{Y}_{2\,\mathrm{adj}},\bar{Y}_{1\,\mathrm{adj}}} & \sigma^2_{\bar{Y}_{2\,\mathrm{adj}}} & \cdots & \sigma_{\bar{Y}_{2\,\mathrm{adj}},\bar{Y}_{J\,\mathrm{adj}}} \\ \cdot & \cdot & \cdots & \cdot \\ \cdot & \cdot & \cdots & \cdot \\ \cdot & \cdot & \cdots & \cdot \\ \sigma_{\bar{Y}_{J\,\mathrm{adj}},\bar{Y}_{1\,\mathrm{adj}}} & \sigma_{\bar{Y}_{J\,\mathrm{adj}},\bar{Y}_{2\,\mathrm{adj}}} & \cdots & \sigma^2_{\bar{Y}_{J\,\mathrm{adj}}} \end{bmatrix}.$$

Each element in the main diagonal of this matrix can be estimated using

$$\mathrm{MS}_{\mathrm{Res_w}}[n^{-1} + \mathbf{d}_{x,j}^T \mathbf{W}_x^{-1} \mathbf{d}_{x,j}],$$

where

MS_{Res_w} is the mean square residual within groups (obtained from the multiple ANCOVA);

$$\mathbf{d}_{x,j} = \begin{bmatrix} \bar{X}_{1,j} - \bar{X}_{1..} \\ \bar{X}_{2,j} - \bar{X}_{2..} \\ \vdots \\ \bar{X}_{C,j} - \bar{X}_{C..} \end{bmatrix};$$

$\bar{X}_{1,j}, \bar{X}_{2,j}, \ldots \bar{X}_{C,j}$ are the means for covariates 1 through C associated with the jth group;

$\bar{X}_{1..}, \bar{X}_{2..}, \ldots, \bar{X}_{C..}$ are the grand means for covariates 1 through C;

$\mathbf{d}_{x,j}^T$ is the transpose of $\mathbf{d}_{x,j}$;

$\mathbf{W}_x^{-1}$ is the inverse of the pooled corrected within-group sum of products matrix for the covariates; that is,

$$\mathbf{W}_x = \begin{bmatrix} \sum x_1^2 & \sum x_1 x_2 & \cdots & \sum x_1 x_C \\ \sum x_2 x_1 & \sum x_2^2 & \cdots & \sum x_2 x_C \\ \cdot & \cdot & \cdots & \cdot \\ \cdot & \cdot & \cdots & \cdot \\ \cdot & \cdot & \cdots & \\ \sum x_C x_1 & \sum x_C x_2 & \cdots & \sum x_C^2 \end{bmatrix}; \quad \text{and}$$

n is the individual group sample size.

Each off-diagonal element in the variance–covariance matrix of the adjusted means can be estimated using

$$MS_{Res_w}[\mathbf{d}_{x,i}^T \mathbf{W}_x^{-1} \mathbf{d}_{x,j}],$$

where

$$\mathbf{d}_{x,i} = \begin{bmatrix} \bar{X}_{1,i} & -\bar{X}_{1..} \\ \bar{X}_{2,i} & -\bar{X}_{2..} \\ & \vdots \\ \bar{X}_{C,i} & -\bar{X}_{C..} \end{bmatrix}.$$

Although MS_{Res_w} enters into all elements in the variance–covariance matrix of the adjusted means, this factor can be eliminated for our purposes. This is advantageous because we want to demonstrate the dependency of the adjusted mean sum of squares (SS_{AT}) on both (1) the intuitive adjusted mean sum of squares, i.e.,

$$SS_{\text{Intuitive AT}} = \sum_{j=1}^{J} n_j (\bar{Y}_{j\,\text{adj}} - \bar{Y}_{..})^2,$$

and (2) the structure of the covariates. The separation of these two issues may be somewhat clearer when the MS_{Res_w} factor is deleted from the variance–covariance matrix defined above.

The elements of the resulting simplified matrix are defined as $[n^{-1} + \mathbf{d}_{x,j}^T \mathbf{W}_x^{-1} \mathbf{d}_{x,j}]$ on the main diagonal and $\mathbf{d}_{x,i}^T \mathbf{W}_x^{-1} \mathbf{d}_{x,j}$ off-diagonal. I denote the simplified matrix (i.e., the MS_{Res_w} deleted variance–covariance matrix of adjusted means) as $\mathbf{A}$.

Even though matrix $\mathbf{A}$ is relevant to dependency among the adjusted dependent variable means, it can be seen that the elements of $\mathbf{A}$ are all computed on the covariates, not the adjusted Y means. On the other hand, adjusted Y means are involved in every element of a second matrix ($\mathbf{C}$), which is defined below:

$$C = \begin{bmatrix} (\bar{Y}_{1\,\mathrm{adj}} - \bar{Y}..)^2 & (\bar{Y}_{1\,\mathrm{adj}} - \bar{Y}..)(\bar{Y}_{2\,\mathrm{adj}} - \bar{Y}..) & \cdots & (\bar{Y}_{1,\mathrm{adj}} - \bar{Y}..)(\bar{Y}_{J\,\mathrm{adj}} - \bar{Y}..) \\ (\bar{Y}_{1\,\mathrm{adj}} - \bar{Y}..)(\bar{Y}_{2\,\mathrm{adj}} - \bar{Y}..) & (\bar{Y}_{2\,\mathrm{adj}} - \bar{Y}..)^2 & \cdots & (\bar{Y}_{2\,\mathrm{adj}} - \bar{Y}..)(\bar{Y}_{\mathrm{J\,adj}} - \bar{Y}..) \\ \cdot & \cdot & \cdots & \cdot \\ \cdot & \cdot & \cdots & \cdot \\ \cdot & \cdot & \cdots & \cdot \\ (\bar{Y}_{1\,\mathrm{adj}} - \bar{Y}..)(\bar{Y}_{J\,\mathrm{adj}} - \bar{Y}..) & (\bar{Y}_{2\,\mathrm{adj}} - \bar{Y}..)(\bar{Y}_{J\,\mathrm{adj}} - \bar{Y}..) & \cdots & (\bar{Y}_{J\,\mathrm{adj}} - \bar{Y}..)^2 \end{bmatrix}.$$

If each element on the main diagonal of matrix $\mathbf{C}$ is multiplied by n, the sum of the resulting products is equal to $SS_{Intuitive\ AT}$. Correspondingly, the trace of the product $\mathbf{A}^{-1}\mathbf{C}$ is equal to SS_{AT}. It turns out that if there are no differences between group means on the covariates (and consequently all elements of each $\mathbf{d}x,j = 0$) then $\mathrm{tr}\ \mathbf{A}^{-1}\mathbf{C} = SS_{Intuitive\ AT} = SS_{AT} = SS_B$.

Hence, in very large randomized-groups experiments the $SS_{Intuitive\ AT}$, SS_{AT}, and the between-group sum of squares (SS_B from ANOVA) will be essentially the same. The reason $SS_{AT} = SS_B$ in this case is that there is no adjustment among means when covariate means are equal; we expect them to be essentially equal in a large randomized-group experiment. The reason $SS_{Intuitive\ AT} = SS_{AT}$ in this case is that there can be no dependency between the covariate means and the adjusted means if there is no variance among the means on the covariates.

If an experiment is based on small samples it is likely that $SS_{Intuitive\ AT} \neq SS_{AT} \neq SS_B$. The means are likely to be adjusted because sampling error on the covariate will lead to covariate mean differences that directly affect the adjustment. SS_{AT} is likely to differ from $SS_{Intuitive\ AT}$ because the covariance between covariate means and adjusted treatment effect estimates is usually nonzero as a result of sampling error.

Example 29.1 A demonstration of the influence of the covariance between adjusted treatment effects estimates and covariate means on the adjusted-treatment sum of squares is provided below for the multiple covariate example presented in Chapter 10. The adjusted treatment sum of squares (SS_{AT} shown in the multiple ANCOVA summary table in Chapter 10) is 624.74 whereas the intuitive value is 640.76. The relationship of these sum of squares to matrices $\mathbf{A}$ and $\mathbf{C}$ (defined above) is described below:

First, matrix $\mathbf{A}$ is computed. The computation of the elements of matrix $\mathbf{A}$ is as follows:

Diagonal elements $[n^{-1} + \mathbf{d}_{x,j}^T \mathbf{W}_x^{-1} \mathbf{d}_{x,j}]$:

$$a_{1,1} = 10^{-1} + [2.667 \quad .100] \begin{bmatrix} 5700 & 591 \\ 591 & 149.4 \end{bmatrix}^{-1} \begin{bmatrix} 2.667 \\ .100 \end{bmatrix} = .1016;$$

$$a_{2,2} = 10^{-1} + [-2.333 \quad -.200] \begin{bmatrix} 5700 & 591 \\ 591 & 149.4 \end{bmatrix}^{-1} \begin{bmatrix} -2.333 \\ -.200 \end{bmatrix} = .1010; \quad \text{and}$$

$$a_{3,3} = 10^{-1} + [-.333 \quad .100] \begin{bmatrix} 5700 & 591 \\ 591 & 149.4 \end{bmatrix}^{-1} \begin{bmatrix} -.333 \\ .100 \end{bmatrix} = .1002.$$

Off-diagonal elements $\mathbf{d}_{x,i}^T \mathbf{W}_x^{-1} \mathbf{d}_{x,j}$:

$$a_{1,2} = [2.667 \quad .100] \begin{bmatrix} 5700 & 591 \\ 591 & 149.4 \end{bmatrix}^{-1} \begin{bmatrix} 2.333 \\ -.200 \end{bmatrix} = -.0012;$$

$$a_{1,3} = [2.667 \quad .100] \begin{bmatrix} 5700 & 591 \\ 591 & 149.4 \end{bmatrix}^{-1} \begin{bmatrix} -.333 \\ .100 \end{bmatrix} = -.0004; \quad \text{and}$$

$$a_{2,3} = [-2.333 \quad -.200] \begin{bmatrix} 5700 & 591 \\ 591 & 149.4 \end{bmatrix}^{-1} \begin{bmatrix} -.333 \\ .100 \end{bmatrix} = .0002.$$

Hence,

$$\mathbf{A} = \begin{bmatrix} .1016 & -.0012 & -.0004 \\ -.0012 & .1010 & .0002 \\ -.0004 & .0002 & .1002 \end{bmatrix}.$$

Matrix $\mathbf{C}$ is

$$\mathbf{C} = \begin{bmatrix} (\bar{Y}_{1\,\text{adj}} - \bar{Y}..)^2 & (\bar{Y}_{1\,\text{adj}} - \bar{Y}..)(\bar{Y}_{2\,\text{adj}} - \bar{Y}..) & \cdots & (\bar{Y}_{1\,\text{adj}} - \bar{Y}..)(\bar{Y}_{J\,\text{adj}} - \bar{Y}..) \\ (\bar{Y}_{1\,\text{adj}} - \bar{Y}..)(\bar{Y}_{2\,\text{adj}} - \bar{Y}..) & (\bar{Y}_{2\,\text{adj}} - \bar{Y}..)^2 & \cdots & (\bar{Y}_{2\,\text{adj}} - \bar{Y}..)(\bar{Y}_{J\,\text{adj}} - \bar{Y}..) \\ \cdot & \cdot & \cdots & \cdot \\ \cdot & \cdot & \cdots & \cdot \\ \cdot & \cdot & \cdots & \cdot \\ (\bar{Y}_{1\,\text{adj}} - \bar{Y}..)(\bar{Y}_{J\,\text{adj}} - \bar{Y}..) & (\bar{Y}_{2\,\text{adj}} - \bar{Y}..)(\bar{Y}_{J\,\text{adj}} - \bar{Y}..) & \cdots & (\bar{Y}_{j\,\text{adj}} - \bar{Y}..)^2 \end{bmatrix}$$

$$= \begin{bmatrix} 36.2533 & -31.3840 & -4.8693 \\ -31.3840 & 27.1687 & 4.2153 \\ -4.8693 & 4.2153 & .6540 \end{bmatrix}.$$

If the covariances of the covariate means and the adjusted mean effects were all zero the off-diagonal elements of $\mathbf{A}$ would be zero and each element on the main diagonal would be $1/n = .10$. It would then be found that $\text{SS}_{\text{AT}} = \text{SS}_{\text{Intuitive AT}}$. But we

see that the off-diagonal values are not equal to zero and the diagonal elements are not each equal to $1/n = .10$. This implies that the covariate means and the adjusted mean effects covary.

A more direct approach is to inspect the diagonal elements of the inverse of **A**, which is shown below:

$$\mathbf{A}^{-1} = \begin{bmatrix} 9.84405 & .11688 & .03906 \\ .11688 & 9.90242 & -.01930 \\ .03906 & -.01930 & 9.98023 \end{bmatrix}.$$

If no covariance exists between the adjusted mean effects and the covariate means the diagonal elements of this matrix will equal n, but it can be seen that these elements differ slightly from 10. This slight discrepancy indicates that there is some dependency and therefore the adjusted treatment sum of squares must differ from the intuitive adjusted sum of squares.

If we compute the product $\mathbf{A}^{-1}\mathbf{C}$ we obtain the following matrix:

$$\begin{bmatrix} 353.021 & -305.605 & -47.4152 \\ -306.446 & 265.286 & 41.1596 \\ -46.575 & 40.319 & 6.2556 \end{bmatrix}.$$

The trace of this matrix equals SS_{AT}. The fact that the trace of this matrix is less than it would have been with no dependency implies that dependency causes a reduction in the "effective" sample size that is used in computing the intuitive sum of squares. That is, the formula for the intuitive adjusted treatment sum of squares will provide the correct adjusted treatment sum of squares if the sample size term n is replaced with ajj, the jth diagonal element of $\mathbf{A}^{-1}$. The use of both n and ajj is shown below for the example data:

$$SS_{Intuitive\ AT} = \sum_{j=1}^{J} n_j(\overline{Y}_{j\,adj} - \overline{Y}..)^2 = [10(\overline{Y}_{1\,adj} - \overline{Y}..)^2] + 10(\overline{Y}_{2\,adj} - \overline{Y}..)^2 + 10(\overline{Y}_{3\,adj} - \overline{Y}..)^2] = 640.76,$$

whereas

$$SS_{AT} = \sum_{j=1}^{J} a_{ii}(\overline{Y}_{j\,adj} - \overline{Y}..)^2 = [9.84405(\overline{Y}_{1\,adj} - \overline{Y}..)^2] + 9.90242(\overline{Y}_{2\,adj} - \overline{Y}..)^2 + 9.98023(\overline{Y}_{3\,adj} - \overline{Y}..)^2] = 624.74$$

Note that the three "effective sample sizes" (ajj) shown immediately above constitute the main diagonal elements of $\mathbf{A}^{-1}$ and that the trace of $\mathbf{A}^{-1}\mathbf{C} = SS_{AT}$. Alternatively, if the covariances of the covariate means and the adjusted mean effects had

all been zero the trace of $\mathbf{A}^{-1}\mathbf{C}$ would have been equal to $SS_{\text{Intuitive AT}}$; in this case $SS_{AT} = SS_{\text{Intuitive AT}}$.

In summary, the intuitive adjusted treatment sum of squares does not generally equal the correct adjusted treatment sum of squares because covariances between the adjusted mean effects and the means of the covariates are not usually zero. These covariances are nonzero because each mean adjustment is based on the same vector of pooled within-group regression coefficients. The result is a minor modification in the effective sample size that can be used to compute the adjusted treatment sum of squares directly from the adjusted means.

29.4 ANCOVA VERSUS ANOVA ON RESIDUALS

It is often claimed that ANCOVA is the same thing as ANOVA computed "on residuals of fitting *X*." This is not true. As stated, this claim is not clear regarding exactly what the nature is of the regression model. There are several different ways of regressing *Y* on *X* (i.e., total, within groups, and between groups) and consequently several ways to compute residuals. Regardless of the method it is clear that the claim cannnot be true. Recall from the introductory chapter on ANCOVA (Chapter 6) that the adjusted treatment sum of squares defined for ANCOVA is the *difference* between the residual sum of squares from fitting the total regression of *Y* on *X* and the residual sum of squares from fitting the within-group regression of *Y* on *X*. Consequently, sum of squares computed on residuals from each different type of residual will differ from those computed from ANCOVA. The *F* ratios computed from ANOVA on residuals (any type) are not distributed as *F*, whereas the ANCOVA *F* does follow the theoretical *F* distribution. Hence, probability values are correct using ANCOVA, but not using ANOVA on residuals. A nice diagramatic explanation of the various residuals and the difference between ANOVA on them and ANCOVA is provided by Maxwell et al. (1985).

29.5 ANCOVA VERSUS *Y/X* RATIO

A common approach used to control confounding in some areas is to standardize the dependent variable by dividing *Y* by *X*. Suppose *Y* is the amount of food consumed, *X* is a measure of body size, and the two treatments are levels of social stress. The ratio *Y/X* is then used to describe results and ANOVA on this ratio is used as the inferential approach. This method of controlling for confounding is enticing because, unlike ANCOVA, it seems simple to both compute and understand. Indeed, it is simple to compute. But the properties of this approach are not always well understood. There is a substantial literature on this problem, contributed by both nonstatistians (e.g., Gould, 1966) and statisticians (e.g., Packard and Boardman, 1988).

The *Y/X* ratio makes sense only when the relationship between *X* and *Y* is isometric. This means that the simple regression of *Y* on *X* has the following properties: (1) the relationship is linear and (2) the value of the intercept is zero. If the relationship

is nonlinear and/or the intercept is nonzero, the ratio is likely to be misleading. The relationship between many psychological and physiological variables with measures of size is known to be nonlinear and intercepts are often far from zero (e.g., Smith, 1984). If either the linearity assumption or the zero intercept assumption do not hold the relationship it is said to be allometric.

When isometry is present (which is seldom) the *Y/X* ratio is independent of *X* and the ratio does not change from one level of *X* to another. This situation yields meaningful interpretations of the ratio and it provides an adequate measure. But when the relationship is allometric the ratio is not independent of *X*; so the ratio changes from one level of *X* to another. This is an undesirable property because a lack of independence means that the effect of *X* has not been removed from the ratio. In other words, this method does not remove the confounding effect of *X* on *Y*. As a consequence, ANOVA on *Y/X* may identify spurious effects in some situations and fail to identify strong effects when they are present in other situations. The solution to such problems is to avoid computing Y/X ratios and use ANCOVA instead.

29.6 OTHER COMMON MISCONCEPTIONS

A dozen ANCOVA misconceptions are listed here with a brief correction. Each one is discussed in detail in earlier chapters.

1. ANCOVA is not useful in randomized-group experiments.
   (This is the major reason for using ANCOVA with these designs.)
2. The covariate should be tested for statistical significance as a preliminary step to determine whether the covariate should be included in the model.
   (There is no need for this. If you want to examine a treatment effect that is independent of the covariate then include it.)
3. ANOVA is preferable to ANCOVA if the within-group regression slopes are heterogeneous.
   (No. Use heterogeneous regression slopes procedures described in Chapter 11.)
4. ANCOVA is the best method for observational studies.
   (Consider redesigning and analyzing using alternatives described in Chapter 28).
5. There is no reason to include a covariate in the model if there are no differences between the covariate means because no mean adjustment will take place in this case.
   (*Adjustment of means is not the only thing that ANCOVA does. See Chapter 6.)*
6. Test the difference between covariate means to see if it is worthwhile to use the covariate.
   (Unless sample size is very small include the variable you want to control, whether the difference is significant or not.)

7. Before carrying out ANCOVA it is necessary to test the differences between covariate means to see if randomization has worked.
   (In a well-designed study there is no need for this.)
8. If the covariate is affected by the treatments, the tainted covariate should be ignored in any subsequent data analysis and ANOVA should be used instead.
   (This is throwing away information. Use quasi-ANCOVA.)
9. Blocking is usually superior to ANCOVA.
   (Consider using both. See Chapter 23.)
10. Gain score analysis and two-factor ANOVA (with one repeated factor) are better than ANCOVA for the analysis of multiple-group pretest–posttest designs.
    (ANCOVA has higher power. See Chapter 25.)
11. ANCOVA does not apply if the covariate is scaled dichotomously.
    (Both continuous and dichotomous variables qualify.)
12. ANCOVA assumes that the within-group correlation between the covariate and the dependent variable is the same within each treatment group.
    (No. The assumption is that the slopes are homogeneous, not the correlations.)

29.7 SUMMARY

There are many misconceptions surrounding ANCOVA. Most of them are discussed in earlier chapters. Three that do not conveniently fit elsewhere are the focus of this chapter. First, it is often believed that the adjusted treatment sum of squares in ANCOVA can be computed by applying the between-group SS computation procedure (used in ANOVA) to the adjusted means. An explanation is provided for why this is not true. Second, it is sometimes claimed that ANCOVA is equivalent to applying ANOVA to the residuals from regressing Y on X. The point is reiterated that ANCOVA involves a comparison of two types of residuals (total and within group); this is not equivalent to ANOVA on one type of residuals. The third misconception is that ANCOVA is equivalent to ANOVA computed on the ratio Y/X. This approach will almost always produce incorrect descriptive and inferential results. It will be satisfactory only in the unlikely case that the relationship between X and Y is isometric.

CHAPTER 30

Uncontrolled Clinical Trials

30.1 INTRODUCTION

The design discussed in this chapter goes by many different names. Many medical researchers label it as an uncontrolled clinical trial, a clinical trial, a before–after design, a screening design, or an intervention study. Biologists sometimes call it an uncontrolled experiment and organizational researchers often label it as the "PPWC" (pretest–posttest without control) design. It is usually called a one-group pretest–posttest design in the behavioral science quasi-experimental design literature. Regardless of the name attached or the content area of application, it should be recognized that this is not a very good design for evaluating most behavioral and medical interventions.

Over four decades ago Campbell and Stanley (1966) pointed out that this design is "worth doing where nothing better can be done" (p. 7); then they concluded that it is a bad example of designs frequently used. During the past four decades many research method textbooks have echoed this lukewarm appraisal, but the design continues to be exceedingly popular throughout the behavioral and medical sciences. Indeed, this approach is so prevalent that entire meta-analyses have been based studies using this design alone (e.g., Wu et al., 2007).

The essentials of the most common version of this design are seemingly straightforward. A pretest measure is administered to a single group of patients, an intervention is applied, and then the same measurement approach is applied again (the posttest). Knowledge of only these essential characteristics of the design is quite insufficient for selecting and interpreting an adequate analysis. One also needs information regarding the exact sampling method, the specific hypotheses to be investigated, approximate knowledge regarding the population distribution from which the sample was drawn, and information regarding threats to internal validity. Unfortunately, these issues and their relevance to reasonable interpretations of results based on this design are very infrequently acknowledged. An inappropriate analysis applied to this design usually

The Analysis of Covariance and Alternatives: Statistical Methods for Experiments, Quasi-Experiments, and Single-Case Studies, Second Edition. Bradley E. Huitema.

has much more serious consequences than in the case of many other designs. Before carrying out any statistical analysis it is important to have confidence that the data justify it. The internal validity of the study should be considered before carrying out formal inference; if a convincing case cannot be made that a comparison is meaningful there is no need to ask whether the difference is outside sampling error.

30.2 INTERNAL VALIDITY THREATS OTHER THAN REGRESSION

The difference between the pretest and posttest sample means is the basic descriptive measure of the study outcome. It is also the focus of the inferential test or confidence interval. The parameter it estimates is the difference between the population pretest and posttest means. Although it is common to label this difference as the intervention effect, let's instead call this parameter the naïve intervention effect; denote it as Δ_{Naive}.

Conceptualize Δ_{Naive} as the sum of three components:

$$\Delta_{\text{Naïve}} = \Delta_{\text{TX}} + \Delta_{\text{DT}} + \Delta_{\text{RTM}},$$

where

- Δ_{TX} is the true treatment effect (TX) for a specific population; it is the difference between the population pretest and posttest means that would be obtained if there were no confounding design threats and no regression effect (i.e., it is the treatment effect independent of any other cause for the mean difference);
- Δ_{DT} is the difference between pre- and postpopulation means that is explained by design threats (DT) such as history, maturation, instrumentation, and testing, which are independent of the treatment effect and the regression effect; and
- Δ_{RTM} is the difference between the pre and post population means that is explained by the regression toward the mean effect (RTM), which is independent of the treatment effect and design threats.

Note that the naïve effect is not the same as the true treatment effect unless there are no design threats and no regression toward the mean effect. Also note that sampling error does not appear in the equation because Δ_{Naive} refers to the difference between population means rather than sample means.

It is important to distinguish the pretest–posttest mean comparison in this design from the two-group comparison in a well-executed randomized-group experiment. In the latter design it is likely that the design threats (i.e., history, maturation, testing, and instrumentation) are not of concern; similarly, regression toward the mean is not an issue with an appropriate randomized-group design. Consequently, the mean difference in a sound randomized two-group experiment is realistically interpreted as an estimate of the true treatment effect whereas the pretest–posttest difference in the one-group pre–post design usually should not be interpreted as a true treatment effect estimate because design threats and regression effects are both likely to be present. (See Campbell and Kenny (1999) for an excellent general discussion of regression effects.)

When the conventional paired-sample analysis is applied to the naïve effect estimate (as is virtually always the case), it should be acknowledged that this approach ignores two of the three components that may contribute to the sample mean difference; this is the reason I label the difference as the naïve treatment effect estimator. Whereas the conventional analysis acknowledges neither design threats nor regression toward the mean effects, a more satisfactory analysis will acknowledge these threats and provide evidence regarding them.

Although the design threats listed here and the regression toward the mean effect are given equal footing in the classic discussions of internal validity threats (e.g., Campbell and Stanley, 1966), I have separated regression effects from the others because the methods of contending with these problems are different. Analytic solutions are available for the regression problem whereas additional design elements are required to illuminate the other threats. Design and regression effects are likely to occur under some conditions but not under others. A brief overview of some of the conditions under which these problems are most likely to occur is presented next.

Design Threats

History, maturation, testing, and instrumentation are explanations for the difference between pretest and posttest means that are independent of possible intervention effects. The effects of these problems (as well as regression) are controlled in a *randomized two-group* pretest–posttest experiment. Because there is no control group in the one-group pre–post design, the only way to eliminate these design threats as plausible explanations for the pre–post difference is to have evidence outside the experiment indicating that these threats are not of concern (or to build a model that includes and controls for such effects).

History

History refers to environmental events (other than the planned intervention) that occur between the administration of the pretest and the administration of the posttest that are alternative explanations for the observed mean difference. The chance that such events are operating increases with the length of the interval between pretest and posttest measurements. If data outside the experiment reveal that observed unplanned events have no effect on variables similar to the pre–post measure, then history may be evaluated as an unlikely cause of the pre–post difference. In some studies it is possible to keep a log of unplanned events that are potential threats. If evidence is available to support the notion that there is little effect of these events on variables similar to the measures used on the study, it should be presented and discussed along with the main outcome. Although evidence of this type is relevant, it does not rule out contaminating effects of unknown and unmeasured historical events.

Maturation

Because organisms change with the passage of time in the absence of external events, such change must be acknowledged when interpreting pre–post differences. In some cases existing data that are external to the pre–post measures collected during an

experiment can inform the interpretation of observed change. For example, in the case of physical growth, much normative information and statistical modeling has been published regarding human growth rates. Information of this type can be useful in the interpretation of differences in pre–post intervention studies. In some cases, it may be reasonable to subtract maturation effects estimated from historical data from the naïve intervention effect estimate. This approach may also be used when performance decreases across time, as often occurs in studies of vigilance, maximum performance, and sleep deprivation.

Testing

Under some conditions the process of administering the pretest has a subsequent effect on the posttest score. That is, the posttest score differs from what it would have been if there had been no pretest. This is especially likely to occur if (a) the measure used as the pretest was constructed using a spiral-omnibus format (a term used in psychometrics to describe a test with items that become progressively more difficult and sample various content areas), and (b) the interval between the administration of the pretest and the administration of the posttest is short. On the other hand, if the pre- and postmeasures are based on a nonreactive or routinely applied approach (such as body temperature, body weight, heart rate, or respiration rate) or if the interval between measurements is long, then it is less likely that there will be a testing effect. Evidence regarding such effects can sometimes be obtained from existing studies, but this is not typical.

Instrumentation

Changes in the calibration of the measurement procedure that occur between pretest and posttest measurements easily can produce changes that are misinterpreted as intervention effects. An egregious example of this problem occurred several years ago in a physiological psychology laboratory where olfactory bulb responding was measured before and after a surgical intervention was applied to rats. The pre–post mean difference was exceedingly large and opposite in direction relative to theoretical prediction. It was later discovered that a student (unaffiliated with the experiment) had dropped the sensitive instrument between pre- and posttesting; a dramatic change in calibration resulted.

In the realm of politics it is common to encounter claims that reductions in crime rate can be attributed to policies of some new politician. Upon closer inspection it is usually discovered that the method of measuring crime was changed substantially along with the political change; in cases such as this the improved crime statistics should often be attributed to changes in measurement method rather than enlightened policy. Similar examples can be found in medicine, education, and clinical psychology. Calibration changes often take place when explicit measurement conventions do not exist or when observer drift occurs with insufficiently trained and monitored observers.

In summary, the adequacy of both descriptive and inferential analyses depends on the truth of the assumption that the pretest and posttest population means are the same ($\mu_{\text{Pre}} = \mu_{\text{Post}}$) in the absence of an intervention; therefore, it is important to have convincing justifications for claiming that potentially confounding sources of

pretest–posttest change are not operating. Unfortunately, such justifications usually require collecting information that is external to the experiment. This evidence should be included in the discussion section accompanying the interpretation of the results of the study. If such information is unavailable it should be pointed out that its absence introduces ambiguity in the interpretation of results and may eliminate interest in formal inferential analyses.

When strong evidence is available to argue that design threats are not present, it may be of interest to pursue formal statistical analysis. The next sections briefly describe some of the issues that should be considered before carrying out an analysis.

30.3 PROBLEMS WITH CONVENTIONAL ANALYSES

Popular statistics books read by those who are expected to carry out or evaluate pretest–posttest research have continued to recommend the correlated sample *t*-test (or the corresponding Wilcoxon test) as the appropriate analysis of one-group pretest–posttest designs; this is in spite of warnings regarding this design and analysis that appeared in Campbell and Stanley (1966) many years ago. Although savvy researchers are aware of regression artifact problems that are likely with one version of this design, an examination of the methodological literature comes up short in terms of guidance regarding appropriate analyses for this situation. The selection of an analysis that is better than the conventional one requires an understanding of both sampling methods and relevant hypotheses. It is helpful to distinguish among four sampling methods illustrated in Figure 30.1 and described below.

Sampling Methods

Only two of the sampling methods described here are associated with reasonably straightforward methods of analysis. These sampling methods are minor modifications of approaches that were originally described in the area of sample survey methodology long ago (Cohen, 1955; Senn and Brown, 1985). Although the types of sampling design originally described in Cohen's work were not developed in the context of the one-group pretest–posttest intervention design, these distinctions are relevant to this design and to appropriate estimation methods.

(a) Simple Random Sampling

A simple random sample is drawn from the population of interest. All participants included in the sample are tested to obtain pretest scores (Y_1), exposed to the intervention, and tested again to obtain the posttest scores (Y_2). There is no selection of a subgroup from this complete random sample. This is labeled as a "complete sample" in the lower right of Figure 30.1.

(b) Screening Sampling: Intentionally Biased Subsample Selected from Simple Random Sample

A simple random sample is drawn from the population. All participants in this sample are then measured on the pretest Y_1; only those participants with $Y_1 \geq cs$ (where cs is the cut score on a measure such as systolic blood pressure) are selected, exposed to

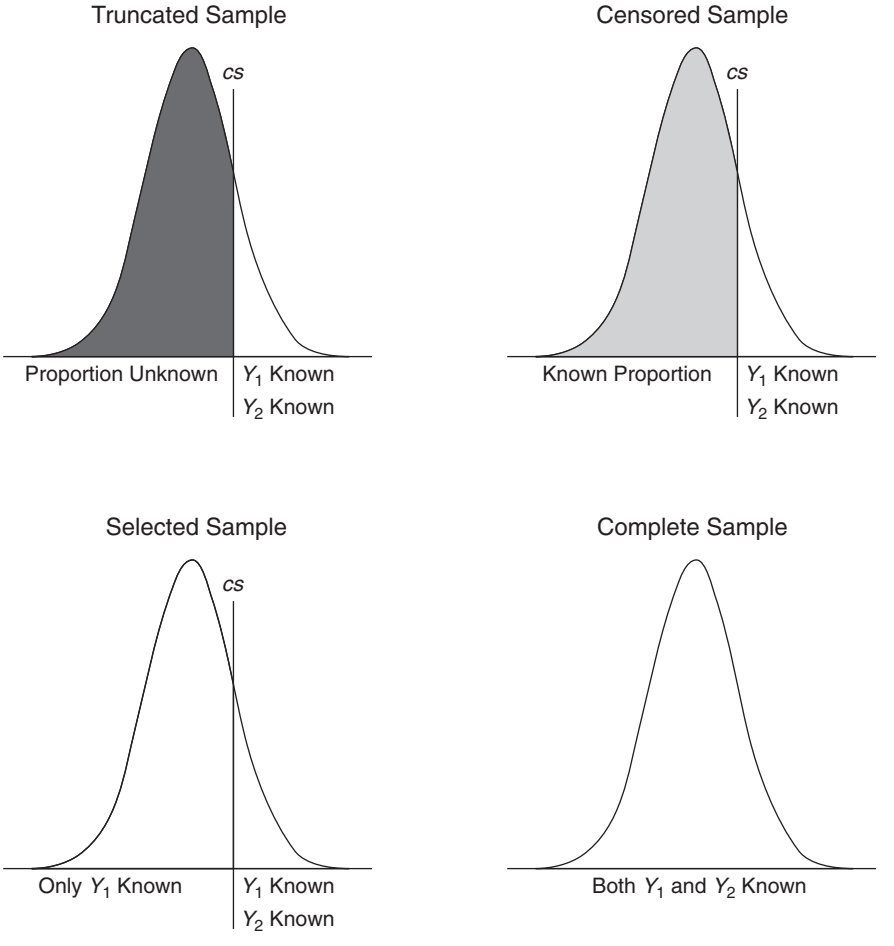


Figure 30.1 Four types of sampling from a univariate distribution.

the intervention, and measured on the posttest Y_2. (Alternatively, the subgroup may be selected from the lower end of the Y_1 distribution, as in compensatory education studies). This is labeled as a "selected" sample in the lower left of Figure 30.1.

(c) Censored Sampling

A *cs* and a sample size *n* are decided upon *before* selecting participants. Each contacted individual is measured on Y_1. Those who meet the *cs* criterion are included in the sample; the pretest scores for those who fail to meet the criterion are not recorded and they are not measured on Y_2, *but they are counted*. Sampling continues until *n* participants who meet the criterion $Y_1 \geq cs$ are obtained. The intervention is applied to the selected participants and measurements on Y_2 are obtained. The proportion of contacted individuals who fall below the criterion is known.

(d) Truncated Sampling

A *cs* and a sample size *n* are decided upon *before* selecting participants from the population. Subject recruiting is carried out by announcing that the study seeks participants who have high need for the treatment. Each individual who contacts the investigator is then measured to obtain Y_1. Recruiting of participants who meet the criterion $Y_1 \geq cs$ is continued until the desired *n* is obtained. The intervention is applied to the selected participants, and measurements on Y_2 are obtained. The number not selected is unknown because the recruiting method implicitly or explicitly announces that only those who have need for the treatment will be selected.

Differences among these four sampling designs have profound implications for analytic choice. Design *a* is likely to be adequately analyzed using the conventional correlated samples *t* test as long as the previously discussed design threats can be ruled out. Design *b* is widely used and is usually inappropriately analyzed using the correlated samples *t* test. More satisfactory analyses for design *b* are described below. Adequate methods of analysis for sampling designs *c* and *d* are not currently available.

30.4 CONTROLLING REGRESSION EFFECTS

The general problem with applying the conventional analysis to designs *b*, *c*, and *d* is that the difference between the subsample pretest and posttest means is unlikely to be zero in the absence of a treatment effect. Because the sub-sample posttest mean is an inappropriate counterfactual, it should be replaced with an estimate that has been corrected for RTM. Alternatively, the naïve difference should be corrected to provide the treatment effect that is independent of the RTM effect. That is, if there are no other design threats, the treatment effect can be conceptualized as $\Delta_{\text{Naive}} - \Delta_{\text{RTM}} = \Delta_{\text{TX}}$.

The naïve effect is estimated using $\hat{\Delta}_{\text{Naive}} = (\overline{Y}_{\text{Post}(>\text{css})} - (\overline{Y}_{\text{Pre}(>\text{css})})$, where the means are based on the sub-sample falling above the cut score (>css). In the case of design *b* the RTM effect is estimated using $\hat{\Delta}_{\text{RTM}} = (1 - \hat{\rho})(\hat{\mu} - \overline{Y}_{\text{Pre}(\text{css})})$, where $\hat{\rho}$ is the estimate of the correlation between pretest and posttest scores, $\hat{\mu}$ is the estimate of the overall population mean, and it is assumed that the treatment effect is additive. For example, suppose a treatment for mild hypertension is applied to a subsample of patients identified using a pretest. If the sub-sample pretest and posttest blood pressure means are 150 and 138, respectively, the naïve treatment effect estimate is -12. If the correlation between pretest and posttest blood pressure measurements is .80 and the estimate of the population mean is 100, the RTM effect estimate is $(1 - .80)(100 - 150) = -10$. Hence, the corrected treatment effect estimate $\hat{\Delta}_{\text{TX}}$ is $\hat{\Delta}_{\text{Naive}} - \hat{\Delta}_{\text{RTM}} = -2$. The naïve difference greatly overestimates the treatment effect. The treatment appears to reduce blood pressure by only 2 two points even though the posttest mean is 12 points lower than the pretest mean.

There will often be interest in a formal test of significance (or confidence interval) for the treatment effect estimate $\hat{\Delta}_{\text{TX}}$. There are several approaches (none of which is well-known) for testing $H_0 : \Delta_{\text{TX}} = 0$.

Alternative models for analyzing design variant *b* have been proposed by Mee and Chau (1991), Chen and Cox (1992), and Naranjo and McKean (2001). All three

models acknowledge and correct for the problem of regression toward the mean. The Mee-Chau model is additive; the intervention effect is assumed to be the same for all participants. The parameters of this model can be estimated using ordinary least-squares. The Chen–Cox model is multiplicative; intervention effects are assumed to vary as a function of the pretest score. Estimation is based on a variant of the pseudo maximum likelihood approach of Gong and Samaniego (1981). A problem with both of these tests is that they assume *either* an additive *or* a multiplicative effect; the researcher usually does not know the nature of the effect before the study is carried out.

The Naranjo–McKean approach is a dual effects model that includes parameters measuring both additive and multiplicative effects. Estimation is based on maximum likelihood, but inference is based on a bootstrap percentile confidence interval method. Available simulation experiments and application-based evidence regarding these models (Chen et al., 1998; Naranjo and McKean, 2001) lead to the following conclusions: the Mee-Chau test is inferentially valid in the case of additive data, the Chen–Cox test is excessively liberal with respect to type I error rate, and the Naranjo–McKean approach is valid with respect to tests on both the additive and multiplicative parameters.

30.5 NARANJO–MCKEAN DUAL EFFECTS MODEL

The Naranjo–McKean (2001) dual effects model includes parameters required for a comprehensive description of the characteristics of one-group pretest–posttest data. This approach is appropriate for data collected under sampling methods *a* and *b* (described above). The model can be written as follows:

$$Y_2 = \mu + \rho(Y_1 - \mu) - \delta - \eta(Y_1 - \mu) + \varepsilon$$

where Y_2 is the posttest score, Y_1 is the pretest score, μ is the common (pretest and posttest) population mean (assumed in the absence of a treatment effect), ρ is the population correlation between pretest and posttest scores, δ is the additive intervention effect (independent of the regression to the mean effect), η is a measure of the multiplicative intervention effect that is proportional to the pretest score, and ε is the error.

Because all intervention models of this type can be estimated only if a population mean is assumed, it is necessary that an estimate of this value be available. This is easily obtained in applications using conventional measures for which normative data are available. Also, estimates can be computed from the complete pretest sample distributions that exist under sampling methods *a* and *b*.

A value of $\delta > 0$ measures the amount of increase attributable to the intervention if the pretest sample mean is below the pretest population mean; it is the amount of decrease attributable to the intervention if the pretest sample mean is above the pretest population mean.

In most cases there will be interest in both additive and multiplicative effects. Let $(\rho - \eta)/\rho = \gamma$. If the model is additive then $\gamma = 1$. If the model is multiplicative then $\gamma \neq 1$. Usually there will be interest in testing the following null hypotheses:

$$H_0 : \delta = 0 \text{ and } H_0 : \gamma = 1.$$

Inferential procedures based on bootstrap methodology are applied to test these hypotheses.

When the model is multiplicative, ($\gamma \neq 1$), we conclude that the intervention effect varies with the pretest score. (This is much like the case of heterogeneity of regression slopes in a two-group ANCOVA.) If a multiplicative model is identified there is likely to be interest in evaluating the intervention effect at certain points on the pretest dimension; perhaps at a very low point, at some intermediate point, and at a point that is far above the mean of the observed pretest data. No method is currently available to address these relevant questions, but methods of this type and associated software are under development and will appear in the statistical literature in the near future.

Software

Information regarding software that performs the Naranjo–McKean dual effects procedure is available at joseph.mckean@wmich.edu.

30.6 SUMMARY

The uncontrolled clinical trial (i.e., one-group pretest–posttest design) should be avoided, if possible. It is vulnerable to many sources of invalidity that are often difficult to rule out. If the sampling method is known to be of a simple form it is possible to identify an analysis that will correct for regression artifacts, but such an analysis is relevant only if design threats that are common with this design can be ruled out. Ruling out design threats involves the collection of data that are external to the experiment. The most adequate analytic approach appears to be the Naranjo–McKean dual effects model; it corrects for regression artifacts and provides both additive and multiplicative effect estimates.

Appendix: Statistical Tables

Table 1a Bonferroni *t* Critial Values for $\alpha = 0.05$

	C'									
df	1	2	3	4	5	6	7	8	9	10
1	12.71	25.45	38.19	50.92	63.66	76.39	89.12	101.86	114.59	127.32
2	4.30	6.21	7.65	8.86	9.92	10.89	11.77	12.59	13.36	14.09
3	3.18	4.18	4.86	5.39	5.84	6.23	6.58	6.90	7.18	7.45
4	2.78	3.50	3.96	4.31	4.60	4.85	5.07	5.26	5.44	5.60
5	2.57	3.16	3.53	3.81	4.03	4.22	4.38	4.53	4.66	4.77
6	2.45	2.97	3.29	3.52	3.71	3.86	4.00	4.12	4.22	4.32
7	2.36	2.84	3.13	3.34	3.50	3.64	3.75	3.86	3.95	4.03
8	2.31	2.75	3.02	3.21	3.36	3.48	3.58	3.68	3.76	3.83
9	2.26	2.69	2.93	3.11	3.25	3.36	3.46	3.55	3.62	3.69
10	2.23	2.63	2.87	3.04	3.17	3.28	3.37	3.45	3.52	3.58
11	2.20	2.59	2.82	2.98	3.11	3.21	3.29	3.37	3.44	3.50
12	2.18	2.56	2.78	2.93	3.05	3.15	3.24	3.31	3.37	3.43
13	2.16	2.53	2.75	2.90	3.01	3.11	3.19	3.26	3.32	3.37
14	2.14	2.51	2.72	2.86	2.98	3.07	3.15	3.21	3.27	3.33
15	2.13	2.49	2.69	2.84	2.95	3.04	3.11	3.18	3.23	3.29
16	2.12	2.47	2.67	2.81	2.92	3.01	3.08	3.15	3.20	3.25
17	2.11	2.46	2.65	2.79	2.90	2.98	3.06	3.12	3.17	3.22
18	2.10	2.45	2.64	2.77	2.88	2.96	3.03	3.09	3.15	3.20
19	2.09	2.43	2.63	2.76	2.86	2.94	3.01	3.07	3.13	3.17
20	2.09	2.42	2.61	2.74	2.85	2.93	3.00	3.06	3.11	3.15
21	2.08	2.41	2.60	2.73	2.83	2.91	2.98	3.04	3.09	3.14
22	2.07	2.41	2.59	2.72	2.82	2.90	2.97	3.02	3.07	3.12
23	2.07	2.40	2.58	2.71	2.81	2.89	2.95	3.01	3.06	3.10
24	2.06	2.39	2.57	2.70	2.80	2.88	2.94	3.00	3.05	3.09
25	2.06	2.38	2.57	2.69	2.79	2.86	2.93	2.99	3.03	3.08
26	2.06	2.38	2.56	2.68	2.78	2.86	2.92	2.98	3.02	3.07
27	2.05	2.37	2.55	2.68	2.77	2.85	2.91	2.97	3.01	3.06
28	2.05	2.37	2.55	2.67	2.76	2.84	2.90	2.96	3.00	3.05

(*Continued*)

Table 1a Bonferroni *t* Critial Values for $\alpha = 0.05$ (*Continued*)

					C'					
df	1	2	3	4	5	6	7	8	9	10
29	2.05	2.36	2.54	2.66	2.76	2.83	2.89	2.95	3.00	3.04
30	2.04	2.36	2.54	2.66	2.75	2.82	2.89	2.94	2.99	3.03
40	2.02	2.33	2.50	2.62	2.70	2.78	2.84	2.89	2.93	2.97
50	2.01	2.31	2.48	2.59	2.68	2.75	2.81	2.85	2.90	2.94
75	1.99	2.29	2.45	2.56	2.64	2.71	2.77	2.81	2.86	2.89
∞	1.96	2.24	2.39	2.50	2.58	2.64	2.69	2.73	2.77	2.81
df	15	20	25	30	35	40	45	50	55	60
1	190.98	254.65	318.31	381.97	445.63	509.30	572.96	636.62	700.28	763.94
2	17.28	19.96	22.33	24.46	26.43	28.26	29.97	31.60	33.14	34.62
3	8.58	9.46	10.21	10.87	11.45	11.98	12.47	12.92	13.35	13.75
4	6.25	6.76	7.17	7.53	7.84	8.12	8.38	8.61	8.83	9.03
5	5.25	5.60	5.89	6.14	6.35	6.54	6.71	6.87	7.01	7.15
6	4.70	4.98	5.21	5.40	5.56	5.71	5.84	5.96	6.07	6.17
7	4.36	4.59	4.79	4.94	5.08	5.20	5.31	5.41	5.50	5.58
8	4.12	4.33	4.50	4.64	4.76	4.86	4.96	5.04	5.12	5.19
9	3.95	4.15	4.30	4.42	4.53	4.62	4.71	4.78	4.85	4.91
10	3.83	4.00	4.14	4.26	4.36	4.44	4.52	4.59	4.65	4.71
11	3.73	3.89	4.02	4.13	4.22	4.30	4.37	4.44	4.49	4.55
12	3.65	3.81	3.93	4.03	4.12	4.19	4.26	4.32	4.37	4.42
13	3.58	3.73	3.85	3.95	4.03	4.10	4.16	4.22	4.27	4.32
14	3.53	3.67	3.79	3.88	3.96	4.03	4.09	4.14	4.19	4.23
15	3.48	3.62	3.73	3.82	3.90	3.96	4.02	4.07	4.12	4.16
16	3.44	3.58	3.69	3.77	3.85	3.91	3.96	4.01	4.06	4.10
17	3.41	3.54	3.65	3.73	3.80	3.86	3.92	3.97	4.01	4.05
18	3.38	3.51	3.61	3.69	3.76	3.82	3.87	3.92	3.96	4.00
19	3.35	3.48	3.58	3.66	3.73	3.79	3.84	3.88	3.93	3.96
20	3.33	3.46	3.55	3.63	3.70	3.75	3.80	3.85	3.89	3.93
21	3.31	3.43	3.53	3.60	3.67	3.73	3.78	3.82	3.86	3.90
22	3.29	3.41	3.50	3.58	3.64	3.70	3.75	3.79	3.83	3.87
23	3.27	3.39	3.48	3.56	3.62	3.68	3.72	3.77	3.81	3.84
24	3.26	3.38	3.47	3.54	3.60	3.66	3.70	3.75	3.78	3.82
25	3.24	3.36	3.45	3.52	3.58	3.64	3.68	3.73	3.76	3.80
26	3.23	3.35	3.43	3.51	3.57	3.62	3.67	3.71	3.74	3.78
27	3.22	3.33	3.42	3.49	3.55	3.60	3.65	3.69	3.73	3.76
28	3.21	3.32	3.41	3.48	3.54	3.59	3.63	3.67	3.71	3.74
29	3.20	3.31	3.40	3.47	3.52	3.58	3.62	3.66	3.70	3.73
30	3.19	3.30	3.39	3.45	3.51	3.56	3.61	3.65	3.68	3.71
40	3.12	3.23	3.31	3.37	3.43	3.47	3.51	3.55	3.58	3.61
50	3.08	3.18	3.26	3.32	3.38	3.42	3.46	3.50	3.53	3.56
75	3.03	3.13	3.20	3.26	3.31	3.35	3.39	3.43	3.45	3.48
∞	2.94	3.02	3.09	3.14	3.19	3.23	3.26	3.29	3.32	3.34

This table was produced by Dr. J. W. McKean.

Table 1b Bonferroni *t* Critial Values for $\alpha = 0.01$

	C'									
df	1	2	3	4	5	6	7	8	9	10
1	63.66	127.32	190.98	254.65	318.31	381.97	445.63	509.30	572.96	636.62
2	9.92	14.09	17.28	19.96	22.33	24.46	26.43	28.26	29.97	31.60
3	5.84	7.45	8.58	9.46	10.21	10.87	11.45	11.98	12.47	12.92
4	4.60	5.60	6.25	6.76	7.17	7.53	7.84	8.12	8.38	8.61
5	4.03	4.77	5.25	5.60	5.89	6.14	6.35	6.54	6.71	6.87
6	3.71	4.32	4.70	4.98	5.21	5.40	5.56	5.71	5.84	5.96
7	3.50	4.03	4.36	4.59	4.79	4.94	5.08	5.20	5.31	5.41
8	3.36	3.83	4.12	4.33	4.50	4.64	4.76	4.86	4.96	5.04
9	3.25	3.69	3.95	4.15	4.30	4.42	4.53	4.62	4.71	4.78
10	3.17	3.58	3.83	4.00	4.14	4.26	4.36	4.44	4.52	4.59
11	3.11	3.50	3.73	3.89	4.02	4.13	4.22	4.30	4.37	4.44
12	3.05	3.43	3.65	3.81	3.93	4.03	4.12	4.19	4.26	4.32
13	3.01	3.37	3.58	3.73	3.85	3.95	4.03	4.10	4.16	4.22
14	2.98	3.33	3.53	3.67	3.79	3.88	3.96	4.03	4.09	4.14
15	2.95	3.29	3.48	3.62	3.73	3.82	3.90	3.96	4.02	4.07
16	2.92	3.25	3.44	3.58	3.69	3.77	3.85	3.91	3.96	4.01
17	2.90	3.22	3.41	3.54	3.65	3.73	3.80	3.86	3.92	3.97
18	2.88	3.20	3.38	3.51	3.61	3.69	3.76	3.82	3.87	3.92
19	2.86	3.17	3.35	3.48	3.58	3.66	3.73	3.79	3.84	3.88
20	2.85	3.15	3.33	3.46	3.55	3.63	3.70	3.75	3.80	3.85
21	2.83	3.14	3.31	3.43	3.53	3.60	3.67	3.73	3.78	3.82
22	2.82	3.12	3.29	3.41	3.50	3.58	3.64	3.70	3.75	3.79
23	2.81	3.10	3.27	3.39	3.48	3.56	3.62	3.68	3.72	3.77
24	2.80	3.09	3.26	3.38	3.47	3.54	3.60	3.66	3.70	3.75
25	2.79	3.08	3.24	3.36	3.45	3.52	3.58	3.64	3.68	3.73
26	2.78	3.07	3.23	3.35	3.43	3.51	3.57	3.62	3.67	3.71
27	2.77	3.06	3.22	3.33	3.42	3.49	3.55	3.60	3.65	3.69
28	2.76	3.05	3.21	3.32	3.41	3.48	3.54	3.59	3.63	3.67
29	2.76	3.04	3.20	3.31	3.40	3.47	3.52	3.58	3.62	3.66
30	2.75	3.03	3.19	3.30	3.39	3.45	3.51	3.56	3.61	3.65
40	2.70	2.97	3.12	3.23	3.31	3.37	3.43	3.47	3.51	3.55
50	2.68	2.94	3.08	3.18	3.26	3.32	3.38	3.42	3.46	3.50
75	2.64	2.89	3.03	3.13	3.20	3.26	3.31	3.35	3.39	3.43
∞	2.58	2.81	2.94	3.02	3.09	3.14	3.19	3.23	3.26	3.29
df	15	20	25	30	35	40	45	50	55	60
1	955	1273	1592	1910	2228	2547	2865	3183	3501	3820
2	38.71	44.70	49.98	54.76	59.15	63.23	67.07	70.70	74.15	77.45
3	14.82	16.33	17.60	18.71	19.70	20.60	21.43	22.20	22.92	23.60
4	9.57	10.31	10.92	11.44	11.90	12.31	12.69	13.03	13.35	13.65
5	7.50	7.98	8.36	8.69	8.98	9.24	9.47	9.68	9.87	10.05
6	6.43	6.79	7.07	7.31	7.52	7.71	7.87	8.02	8.16	8.29

(*Continued*)

Table 1b Bonferroni *t* Critial Values for $\alpha = 0.01$ (*Continued*)

	C'									
df	1	2	3	4	5	6	7	8	9	10
7	5.80	6.08	6.31	6.50	6.67	6.81	6.94	7.06	7.17	7.27
8	5.37	5.62	5.81	5.97	6.11	6.23	6.34	6.44	6.53	6.62
9	5.08	5.29	5.46	5.60	5.72	5.83	5.92	6.01	6.09	6.16
10	4.85	5.05	5.20	5.33	5.44	5.53	5.62	5.69	5.76	5.83
11	4.68	4.86	5.00	5.12	5.22	5.31	5.38	5.45	5.52	5.57
12	4.55	4.72	4.85	4.96	5.05	5.13	5.20	5.26	5.32	5.38
13	4.44	4.60	4.72	4.82	4.91	4.98	5.05	5.11	5.17	5.22
14	4.35	4.50	4.62	4.71	4.79	4.87	4.93	4.99	5.04	5.08
15	4.27	4.42	4.53	4.62	4.70	4.77	4.83	4.88	4.93	4.97
16	4.21	4.35	4.45	4.54	4.62	4.68	4.74	4.79	4.84	4.88
17	4.15	4.29	4.39	4.47	4.55	4.61	4.66	4.71	4.76	4.80
18	4.10	4.23	4.33	4.42	4.49	4.55	4.60	4.65	4.69	4.73
19	4.06	4.19	4.28	4.36	4.43	4.49	4.54	4.59	4.63	4.67
20	4.02	4.15	4.24	4.32	4.39	4.44	4.49	4.54	4.58	4.62
21	3.99	4.11	4.20	4.28	4.34	4.40	4.45	4.49	4.53	4.57
22	3.96	4.08	4.17	4.24	4.31	4.36	4.41	4.45	4.49	4.53
23	3.93	4.05	4.14	4.21	4.27	4.33	4.37	4.42	4.45	4.49
24	3.91	4.02	4.11	4.18	4.24	4.29	4.34	4.38	4.42	4.45
25	3.88	4.00	4.08	4.15	4.21	4.27	4.31	4.35	4.39	4.42
26	3.86	3.97	4.06	4.13	4.19	4.24	4.28	4.32	4.36	4.39
27	3.84	3.95	4.04	4.11	4.16	4.22	4.26	4.30	4.33	4.37
28	3.83	3.94	4.02	4.09	4.14	4.19	4.24	4.28	4.31	4.34
29	3.81	3.92	4.00	4.07	4.12	4.17	4.22	4.25	4.29	4.32
30	3.80	3.90	3.98	4.05	4.11	4.15	4.20	4.23	4.27	4.30
40	3.69	3.79	3.86	3.92	3.98	4.02	4.06	4.09	4.13	4.15
50	3.63	3.72	3.79	3.85	3.90	3.94	3.98	4.01	4.04	4.07
75	3.55	3.64	3.71	3.76	3.81	3.85	3.88	3.91	3.94	3.96
∞	3.40	3.48	3.54	3.59	3.63	3.66	3.69	3.72	3.74	3.76

This table was produced by Dr. J. W. McKean.

Table 2a Studentized Range Distribution Critical Values for $\alpha = .05$

	J																		
df	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
5	3.64	4.60	5.22	5.67	6.03	6.33	6.58	6.80	6.99	7.17	7.32	7.47	7.60	7.72	7.83	7.93	8.03	8.12	8.21
6	3.46	4.34	4.90	5.30	5.63	5.90	6.12	6.32	6.49	6.65	6.79	6.92	7.03	7.14	7.24	7.34	7.43	7.51	7.59
7	3.34	4.16	4.68	5.06	5.36	5.61	5.82	6.00	6.16	6.30	6.43	6.55	6.66	6.76	6.85	6.94	7.02	7.10	7.17
8	3.26	4.04	4.53	4.89	5.17	5.40	5.60	5.77	5.92	6.05	6.18	6.29	6.39	6.48	6.57	6.65	6.73	6.80	6.87
9	3.20	3.95	4.41	4.76	5.02	5.24	5.43	5.59	5.74	5.87	5.98	6.09	6.19	6.28	6.36	6.44	6.51	6.58	6.64
10	3.15	3.88	4.33	4.65	4.91	5.12	5.30	5.46	5.60	5.72	5.83	5.93	6.03	6.11	6.19	6.27	6.34	6.40	6.47
11	3.11	3.82	4.26	4.57	4.82	5.03	5.20	5.35	5.49	5.61	5.71	5.81	5.90	5.98	6.06	6.13	6.20	6.27	6.33
12	3.08	3.77	4.20	4.51	4.75	4.95	5.12	5.27	5.39	5.51	5.61	5.71	5.80	5.88	5.95	6.02	6.09	6.15	6.21
13	3.06	3.73	4.15	4.45	4.69	4.88	5.05	5.19	5.32	5.43	5.53	5.63	5.71	5.79	5.86	5.93	5.99	6.05	6.11
14	3.03	3.70	4.11	4.41	4.64	4.83	4.99	5.13	5.25	5.36	5.46	5.55	5.64	5.71	5.79	5.85	5.91	5.97	6.03
15	3.01	3.67	4.08	4.37	4.59	4.78	4.94	5.08	5.20	5.31	5.40	5.49	5.57	5.65	5.72	5.78	5.85	5.90	5.96
16	3.00	3.65	4.05	4.33	4.56	4.74	4.90	5.03	5.15	5.26	5.35	5.44	5.52	5.59	5.66	5.73	5.79	5.84	5.90
17	2.98	3.63	4.02	4.30	4.52	4.70	4.86	4.99	5.11	5.21	5.31	5.39	5.47	5.54	5.61	5.67	5.73	5.79	5.84
18	2.97	3.61	4.00	4.28	4.49	4.67	4.82	4.96	5.07	5.17	5.27	5.35	5.43	5.50	5.57	5.63	5.69	5.74	5.79
19	2.96	3.59	3.98	4.25	4.47	4.65	4.79	4.92	5.04	5.14	5.23	5.31	5.39	5.46	5.53	5.59	5.65	5.70	5.75
20	2.95	3.58	3.96	4.23	4.45	4.62	4.77	4.90	5.01	5.11	5.20	5.28	5.36	5.43	5.49	5.55	5.61	5.66	5.71
21	2.94	3.56	3.94	4.21	4.42	4.60	4.74	4.87	4.98	5.08	5.17	5.25	5.33	5.40	5.46	5.52	5.58	5.63	5.68
22	2.93	3.55	3.93	4.20	4.41	4.58	4.72	4.85	4.96	5.06	5.14	5.23	5.30	5.37	5.43	5.49	5.55	5.60	5.65
23	2.93	3.54	3.91	4.18	4.39	4.56	4.70	4.83	4.94	5.03	5.12	5.20	5.27	5.34	5.41	5.46	5.52	5.57	5.62
24	2.92	3.53	3.90	4.17	4.37	4.54	4.68	4.81	4.92	5.01	5.10	5.18	5.25	5.32	5.38	5.44	5.49	5.55	5.59
25	2.91	3.52	3.89	4.15	4.36	4.53	4.67	4.79	4.90	4.99	5.08	5.16	5.23	5.30	5.36	5.42	5.47	5.52	5.57
26	2.91	3.51	3.88	4.14	4.35	4.51	4.65	4.77	4.88	4.98	5.06	5.14	5.21	5.28	5.34	5.40	5.45	5.50	5.55
27	2.90	3.51	3.87	4.13	4.33	4.50	4.64	4.76	4.86	4.96	5.04	5.12	5.19	5.26	5.32	5.38	5.43	5.48	5.53
28	2.90	3.50	3.86	4.12	4.32	4.49	4.62	4.74	4.85	4.94	5.03	5.11	5.18	5.24	5.30	5.36	5.41	5.46	5.51
29	2.89	3.49	3.85	4.11	4.31	4.47	4.61	4.73	4.84	4.93	5.01	5.09	5.16	5.23	5.29	5.34	5.40	5.44	5.49
30	2.89	3.49	3.85	4.10	4.30	4.46	4.60	4.72	4.82	4.92	5.00	5.08	5.15	5.21	5.27	5.33	5.38	5.43	5.47
31	2.88	3.48	3.84	4.09	4.29	4.45	4.59	4.71	4.81	4.90	4.99	5.06	5.13	5.20	5.26	5.31	5.36	5.41	5.46
32	2.88	3.48	3.83	4.09	4.28	4.45	4.58	4.70	4.80	4.89	4.98	5.05	5.12	5.18	5.24	5.30	5.35	5.40	5.45
33	2.88	3.47	3.83	4.08	4.28	4.44	4.57	4.69	4.79	4.88	4.97	5.04	5.11	5.17	5.23	5.29	5.34	5.39	5.43
34	2.87	3.47	3.82	4.07	4.27	4.43	4.56	4.68	4.78	4.87	4.96	5.03	5.10	5.16	5.22	5.27	5.33	5.37	5.42
35	2.87	3.46	3.81	4.07	4.26	4.42	4.56	4.67	4.77	4.86	4.95	5.02	5.09	5.15	5.21	5.26	5.31	5.36	5.41
36	2.87	3.46	3.81	4.06	4.25	4.41	4.55	4.66	4.76	4.85	4.94	5.01	5.08	5.14	5.20	5.25	5.30	5.35	5.40
37	2.87	3.45	3.80	4.05	4.25	4.41	4.54	4.66	4.76	4.85	4.93	5.00	5.07	5.13	5.19	5.24	5.29	5.34	5.39
38	2.86	3.45	3.80	4.05	4.24	4.40	4.53	4.65	4.75	4.84	4.92	4.99	5.06	5.12	5.18	5.23	5.28	5.33	5.38
39	2.86	3.45	3.79	4.04	4.24	4.39	4.53	4.64	4.74	4.83	4.91	4.98	5.05	5.11	5.17	5.22	5.27	5.32	5.37
40	2.86	3.44	3.79	4.04	4.23	4.39	4.52	4.63	4.73	4.82	4.90	4.98	5.04	5.11	5.16	5.22	5.27	5.31	5.36
41	2.86	3.44	3.79	4.03	4.23	4.38	4.51	4.63	4.73	4.82	4.90	4.97	5.04	5.10	5.15	5.21	5.26	5.30	5.35
42	2.85	3.44	3.78	4.03	4.22	4.38	4.51	4.62	4.72	4.81	4.89	4.96	5.03	5.09	5.15	5.20	5.25	5.30	5.34
43	2.85	3.43	3.78	4.03	4.22	4.37	4.50	4.62	4.72	4.80	4.88	4.96	5.02	5.08	5.14	5.19	5.24	5.29	5.33
44	2.85	3.43	3.78	4.02	4.21	4.37	4.50	4.61	4.71	4.80	4.88	4.95	5.02	5.08	5.13	5.19	5.24	5.28	5.33
45	2.85	3.43	3.77	4.02	4.21	4.36	4.49	4.61	4.70	4.79	4.87	4.94	5.01	5.07	5.13	5.18	5.23	5.28	5.32
46	2.85	3.42	3.77	4.01	4.20	4.36	4.49	4.60	4.70	4.79	4.87	4.94	5.00	5.06	5.12	5.17	5.22	5.27	5.31
47	2.85	3.42	3.77	4.01	4.20	4.36	4.48	4.60	4.69	4.78	4.86	4.93	5.00	5.06	5.11	5.17	5.22	5.26	5.31
48	2.84	3.42	3.76	4.01	4.20	4.35	4.48	4.59	4.69	4.78	4.86	4.93	4.99	5.05	5.11	5.16	5.21	5.26	5.30
49	2.84	3.42	3.76	4.00	4.19	4.35	4.48	4.59	4.69	4.77	4.85	4.92	4.99	5.05	5.10	5.16	5.20	5.25	5.29
50	2.84	3.42	3.76	4.00	4.19	4.34	4.47	4.58	4.68	4.77	4.85	4.92	4.98	5.04	5.10	5.15	5.20	5.24	5.29
51	2.84	3.41	3.76	4.00	4.19	4.34	4.47	4.58	4.68	4.76	4.84	4.91	4.98	5.04	5.09	5.15	5.19	5.24	5.28
52	2.84	3.41	3.75	4.00	4.18	4.34	4.47	4.58	4.67	4.76	4.84	4.91	4.97	5.03	5.09	5.14	5.19	5.23	5.28
53	2.84	3.41	3.75	3.99	4.18	4.33	4.46	4.57	4.67	4.76	4.83	4.90	4.97	5.03	5.08	5.14	5.18	5.23	5.27
54	2.84	3.41	3.75	3.99	4.18	4.33	4.46	4.57	4.67	4.75	4.83	4.90	4.96	5.02	5.08	5.13	5.18	5.22	5.27
55	2.83	3.41	3.75	3.99	4.18	4.33	4.46	4.57	4.66	4.75	4.83	4.90	4.96	5.02	5.08	5.13	5.17	5.22	5.26
56	2.83	3.40	3.74	3.99	4.17	4.32	4.45	4.56	4.66	4.74	4.82	4.89	4.96	5.02	5.07	5.12	5.17	5.22	5.26
57	2.83	3.40	3.74	3.98	4.17	4.32	4.45	4.56	4.66	4.74	4.82	4.89	4.95	5.01	5.07	5.12	5.17	5.21	5.25
58	2.83	3.40	3.74	3.98	4.17	4.32	4.45	4.56	4.65	4.74	4.82	4.89	4.95	5.01	5.06	5.11	5.16	5.21	5.25

(*Continued*)

Table 2a Studentized Range Distribution Critical Values for $\alpha = .05$ (*Continued*)

	J																		
df	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
59	2.83	3.40	3.74	3.98	4.17	4.32	4.44	4.55	4.65	4.73	4.81	4.88	4.95	5.00	5.06	5.11	5.16	5.20	5.25
60	2.83	3.40	3.74	3.98	4.16	4.31	4.44	4.55	4.65	4.73	4.81	4.88	4.94	5.00	5.06	5.11	5.15	5.20	5.24
61	2.83	3.40	3.74	3.98	4.16	4.31	4.44	4.55	4.64	4.73	4.81	4.88	4.94	5.00	5.05	5.10	5.15	5.20	5.24
62	2.83	3.40	3.73	3.97	4.16	4.31	4.44	4.55	4.64	4.73	4.80	4.87	4.94	4.99	5.05	5.10	5.15	5.19	5.23
63	2.83	3.39	3.73	3.97	4.16	4.31	4.43	4.54	4.64	4.72	4.80	4.87	4.93	4.99	5.05	5.10	5.14	5.19	5.23
64	2.83	3.39	3.73	3.97	4.15	4.30	4.43	4.54	4.64	4.72	4.80	4.87	4.93	4.99	5.04	5.09	5.14	5.18	5.23
65	2.82	3.39	3.73	3.97	4.15	4.30	4.43	4.54	4.63	4.72	4.79	4.86	4.93	4.99	5.04	5.09	5.14	5.18	5.22
66	2.82	3.39	3.73	3.97	4.15	4.30	4.43	4.54	4.63	4.72	4.79	4.86	4.92	4.98	5.04	5.09	5.13	5.18	5.22
67	2.82	3.39	3.73	3.96	4.15	4.30	4.42	4.53	4.63	4.71	4.79	4.86	4.92	4.98	5.03	5.08	5.13	5.18	5.22
68	2.82	3.39	3.72	3.96	4.15	4.30	4.42	4.53	4.63	4.71	4.79	4.86	4.92	4.98	5.03	5.08	5.13	5.17	5.21
69	2.82	3.39	3.72	3.96	4.15	4.29	4.42	4.53	4.62	4.71	4.78	4.85	4.92	4.97	5.03	5.08	5.13	5.17	5.21
70	2.82	3.39	3.72	3.96	4.14	4.29	4.42	4.53	4.62	4.71	4.78	4.85	4.91	4.97	5.03	5.08	5.12	5.17	5.21
71	2.82	3.39	3.72	3.96	4.14	4.29	4.42	4.52	4.62	4.70	4.78	4.85	4.91	4.97	5.02	5.07	5.12	5.16	5.21
72	2.82	3.38	3.72	3.96	4.14	4.29	4.41	4.52	4.62	4.70	4.78	4.85	4.91	4.97	5.02	5.07	5.12	5.16	5.20
73	2.82	3.38	3.72	3.96	4.14	4.29	4.41	4.52	4.62	4.70	4.77	4.84	4.91	4.96	5.02	5.07	5.11	5.16	5.20
74	2.82	3.38	3.72	3.95	4.14	4.29	4.41	4.52	4.61	4.70	4.77	4.84	4.90	4.96	5.02	5.07	5.11	5.16	5.20
75	2.82	3.38	3.72	3.95	4.14	4.28	4.41	4.52	4.61	4.70	4.77	4.84	4.90	4.96	5.01	5.06	5.11	5.15	5.19
76	2.82	3.38	3.71	3.95	4.13	4.28	4.41	4.52	4.61	4.69	4.77	4.84	4.90	4.96	5.01	5.06	5.11	5.15	5.19
77	2.82	3.38	3.71	3.95	4.13	4.28	4.41	4.51	4.61	4.69	4.77	4.84	4.90	4.96	5.01	5.06	5.11	5.15	5.19
78	2.82	3.38	3.71	3.95	4.13	4.28	4.40	4.51	4.61	4.69	4.76	4.83	4.90	4.95	5.01	5.06	5.10	5.15	5.19
79	2.81	3.38	3.71	3.95	4.13	4.28	4.40	4.51	4.60	4.69	4.76	4.83	4.89	4.95	5.00	5.05	5.10	5.14	5.19
80	2.81	3.38	3.71	3.95	4.13	4.28	4.40	4.51	4.60	4.69	4.76	4.83	4.89	4.95	5.00	5.05	5.10	5.14	5.18
81	2.81	3.38	3.71	3.95	4.13	4.28	4.40	4.51	4.60	4.68	4.76	4.83	4.89	4.95	5.00	5.05	5.10	5.14	5.18
82	2.81	3.38	3.71	3.94	4.13	4.27	4.40	4.51	4.60	4.68	4.76	4.83	4.89	4.95	5.00	5.05	5.09	5.14	5.18
83	2.81	3.37	3.71	3.94	4.13	4.27	4.40	4.50	4.60	4.68	4.76	4.82	4.89	4.94	5.00	5.05	5.09	5.14	5.18
84	2.81	3.37	3.71	3.94	4.12	4.27	4.40	4.50	4.60	4.68	4.75	4.82	4.88	4.94	5.00	5.04	5.09	5.13	5.18
85	2.81	3.37	3.71	3.94	4.12	4.27	4.39	4.50	4.60	4.68	4.75	4.82	4.88	4.94	4.99	5.04	5.09	5.13	5.17
86	2.81	3.37	3.71	3.94	4.12	4.27	4.39	4.50	4.59	4.68	4.75	4.82	4.88	4.94	4.99	5.04	5.09	5.13	5.17
87	2.81	3.37	3.70	3.94	4.12	4.27	4.39	4.50	4.59	4.68	4.75	4.82	4.88	4.94	4.99	5.04	5.09	5.13	5.17
88	2.81	3.37	3.70	3.94	4.12	4.27	4.39	4.50	4.59	4.67	4.75	4.82	4.88	4.94	4.99	5.04	5.08	5.13	5.17
89	2.81	3.37	3.70	3.94	4.12	4.27	4.39	4.50	4.59	4.67	4.75	4.81	4.88	4.93	4.99	5.04	5.08	5.13	5.17
90	2.81	3.37	3.70	3.94	4.12	4.27	4.39	4.50	4.59	4.67	4.75	4.81	4.88	4.93	4.98	5.03	5.08	5.12	5.16
91	2.81	3.37	3.70	3.94	4.12	4.26	4.39	4.49	4.59	4.67	4.74	4.81	4.87	4.93	4.98	5.03	5.08	5.12	5.16
92	2.81	3.37	3.70	3.94	4.12	4.26	4.39	4.49	4.59	4.67	4.74	4.81	4.87	4.93	4.98	5.03	5.08	5.12	5.16
93	2.81	3.37	3.70	3.93	4.12	4.26	4.39	4.49	4.58	4.67	4.74	4.81	4.87	4.93	4.98	5.03	5.08	5.12	5.16
94	2.81	3.37	3.70	3.93	4.11	4.26	4.38	4.49	4.58	4.67	4.74	4.81	4.87	4.93	4.98	5.03	5.07	5.12	5.16
95	2.81	3.37	3.70	3.93	4.11	4.26	4.38	4.49	4.58	4.66	4.74	4.81	4.87	4.92	4.98	5.03	5.07	5.12	5.16
96	2.81	3.37	3.70	3.93	4.11	4.26	4.38	4.49	4.58	4.66	4.74	4.81	4.87	4.92	4.98	5.03	5.07	5.11	5.15
97	2.81	3.37	3.70	3.93	4.11	4.26	4.38	4.49	4.58	4.66	4.74	4.80	4.87	4.92	4.97	5.02	5.07	5.11	5.15
98	2.81	3.37	3.70	3.93	4.11	4.26	4.38	4.49	4.58	4.66	4.74	4.80	4.86	4.92	4.97	5.02	5.07	5.11	5.15
99	2.81	3.37	3.70	3.93	4.11	4.26	4.38	4.49	4.58	4.66	4.73	4.80	4.86	4.92	4.97	5.02	5.07	5.11	5.15
100	2.81	3.36	3.70	3.93	4.11	4.26	4.38	4.48	4.58	4.66	4.73	4.80	4.86	4.92	4.97	5.02	5.07	5.11	5.15
105	2.80	3.36	3.69	3.93	4.11	4.25	4.37	4.48	4.57	4.65	4.73	4.79	4.86	4.91	4.96	5.01	5.06	5.10	5.14
110	2.80	3.36	3.69	3.92	4.10	4.25	4.37	4.48	4.57	4.65	4.72	4.79	4.85	4.91	4.96	5.01	5.05	5.10	5.14
115	2.80	3.36	3.69	3.92	4.10	4.24	4.37	4.47	4.56	4.64	4.72	4.79	4.85	4.90	4.95	5.00	5.05	5.09	5.13
120	2.80	3.36	3.68	3.92	4.10	4.24	4.36	4.47	4.56	4.64	4.71	4.78	4.84	4.90	4.95	5.00	5.04	5.09	5.13
125	2.80	3.35	3.68	3.91	4.09	4.24	4.36	4.46	4.56	4.64	4.71	4.78	4.84	4.89	4.95	4.99	5.04	5.08	5.12
130	2.80	3.35	3.68	3.91	4.09	4.24	4.36	4.46	4.55	4.63	4.71	4.77	4.83	4.89	4.94	4.99	5.03	5.08	5.12
135	2.80	3.35	3.68	3.91	4.09	4.23	4.35	4.46	4.55	4.63	4.70	4.77	4.83	4.89	4.94	4.99	5.03	5.07	5.11
140	2.80	3.35	3.68	3.91	4.09	4.23	4.35	4.46	4.55	4.63	4.70	4.77	4.83	4.88	4.93	4.98	5.03	5.07	5.11
145	2.80	3.35	3.68	3.91	4.08	4.23	4.35	4.45	4.54	4.63	4.70	4.76	4.82	4.88	4.93	4.98	5.02	5.07	5.11
150	2.79	3.35	3.67	3.90	4.08	4.23	4.35	4.45	4.54	4.62	4.70	4.76	4.82	4.88	4.93	4.98	5.02	5.06	5.10
∞	2.77	3.31	3.63	3.86	4.03	4.17	4.29	4.39	4.47	4.55	4.62	4.68	4.74	4.80	4.85	4.89	4.93	4.97	5.01

This table was produced by Dr. J. W. McKean.

Table 2b Studentized Range Distribution Critical Values for $\alpha = .01$

	J																		
df	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
5	5.70	6.98	7.80	8.42	8.91	9.32	9.67	9.97	10.24	10.48	10.70	10.89	11.08	11.24	11.40	11.55	11.68	11.81	11.93
6	5.24	6.33	7.03	7.56	7.97	8.32	8.61	8.87	9.10	9.30	9.48	9.65	9.81	9.95	10.08	10.21	10.32	10.43	10.54
7	4.95	5.92	6.54	7.00	7.37	7.68	7.94	8.17	8.37	8.55	8.71	8.86	9.00	9.12	9.24	9.35	9.46	9.55	9.65
8	4.75	5.64	6.20	6.62	6.96	7.24	7.47	7.68	7.86	8.03	8.18	8.31	8.44	8.55	8.66	8.76	8.85	8.94	9.03
9	4.60	5.43	5.96	6.35	6.66	6.91	7.13	7.33	7.49	7.65	7.78	7.91	8.03	8.13	8.23	8.33	8.41	8.49	8.57
10	4.48	5.27	5.77	6.14	6.43	6.67	6.87	7.05	7.21	7.36	7.49	7.60	7.71	7.81	7.91	7.99	8.08	8.15	8.23
11	4.39	5.15	5.62	5.97	6.25	6.48	6.67	6.84	6.99	7.13	7.25	7.36	7.46	7.56	7.65	7.73	7.81	7.88	7.95
12	4.32	5.05	5.50	5.84	6.10	6.32	6.51	6.67	6.81	6.94	7.06	7.17	7.26	7.36	7.44	7.52	7.59	7.66	7.73
13	4.26	4.96	5.40	5.73	5.98	6.19	6.37	6.53	6.67	6.79	6.90	7.01	7.10	7.19	7.27	7.35	7.42	7.48	7.55
14	4.21	4.89	5.32	5.63	5.88	6.08	6.26	6.41	6.54	6.66	6.77	6.87	6.96	7.05	7.13	7.20	7.27	7.33	7.39
15	4.17	4.84	5.25	5.56	5.80	5.99	6.16	6.31	6.44	6.55	6.66	6.76	6.84	6.93	7.00	7.07	7.14	7.20	7.26
16	4.13	4.79	5.19	5.49	5.72	5.92	6.08	6.22	6.35	6.46	6.56	6.66	6.74	6.82	6.90	6.97	7.03	7.09	7.15
17	4.10	4.74	5.14	5.43	5.66	5.85	6.01	6.15	6.27	6.38	6.48	6.57	6.66	6.73	6.81	6.87	6.94	7.00	7.05
18	4.07	4.70	5.09	5.38	5.60	5.79	5.94	6.08	6.20	6.31	6.41	6.50	6.58	6.65	6.73	6.79	6.85	6.91	6.97
19	4.05	4.67	5.05	5.33	5.55	5.73	5.89	6.02	6.14	6.25	6.34	6.43	6.51	6.58	6.65	6.72	6.78	6.84	6.89
20	4.02	4.64	5.02	5.29	5.51	5.69	5.84	5.97	6.09	6.19	6.28	6.37	6.45	6.52	6.59	6.65	6.71	6.77	6.82
21	4.00	4.61	4.99	5.26	5.47	5.65	5.79	5.92	6.04	6.14	6.23	6.32	6.39	6.47	6.53	6.60	6.65	6.71	6.76
22	3.99	4.59	4.96	5.22	5.43	5.61	5.75	5.88	5.99	6.10	6.19	6.27	6.35	6.42	6.48	6.54	6.60	6.66	6.71
23	3.97	4.57	4.93	5.20	5.40	5.57	5.72	5.84	5.95	6.05	6.14	6.23	6.30	6.37	6.44	6.50	6.55	6.61	6.66
24	3.96	4.55	4.91	5.17	5.37	5.54	5.69	5.81	5.92	6.02	6.11	6.19	6.26	6.33	6.39	6.45	6.51	6.56	6.61
25	3.94	4.53	4.89	5.14	5.35	5.51	5.65	5.78	5.89	5.98	6.07	6.15	6.22	6.29	6.35	6.41	6.47	6.52	6.57
26	3.93	4.51	4.87	5.12	5.32	5.49	5.63	5.75	5.86	5.95	6.04	6.12	6.19	6.26	6.32	6.38	6.43	6.48	6.53
27	3.92	4.49	4.85	5.10	5.30	5.46	5.60	5.72	5.83	5.92	6.01	6.09	6.16	6.22	6.29	6.34	6.40	6.45	6.50
28	3.91	4.48	4.83	5.08	5.28	5.44	5.58	5.70	5.80	5.90	5.98	6.06	6.13	6.20	6.26	6.31	6.37	6.42	6.47
29	3.90	4.47	4.81	5.06	5.26	5.42	5.56	5.67	5.78	5.87	5.96	6.03	6.10	6.17	6.23	6.28	6.34	6.39	6.44
30	3.89	4.45	4.80	5.05	5.24	5.40	5.54	5.65	5.76	5.85	5.93	6.01	6.08	6.14	6.20	6.26	6.31	6.36	6.41
31	3.88	4.44	4.79	5.03	5.23	5.38	5.52	5.63	5.74	5.83	5.91	5.99	6.05	6.12	6.18	6.23	6.29	6.33	6.38
32	3.87	4.43	4.77	5.02	5.21	5.37	5.50	5.61	5.72	5.81	5.89	5.96	6.03	6.10	6.16	6.21	6.26	6.31	6.36
33	3.87	4.42	4.76	5.00	5.20	5.35	5.48	5.60	5.70	5.79	5.87	5.94	6.01	6.08	6.13	6.19	6.24	6.29	6.33
34	3.86	4.41	4.75	4.99	5.18	5.34	5.47	5.58	5.68	5.77	5.85	5.93	5.99	6.06	6.11	6.17	6.22	6.27	6.31
35	3.85	4.40	4.74	4.98	5.17	5.32	5.45	5.57	5.67	5.75	5.84	5.91	5.98	6.04	6.10	6.15	6.20	6.25	6.29
36	3.85	4.40	4.73	4.97	5.16	5.31	5.44	5.55	5.65	5.74	5.82	5.89	5.96	6.02	6.08	6.13	6.18	6.23	6.27
37	3.84	4.39	4.72	4.96	5.15	5.30	5.43	5.54	5.64	5.72	5.80	5.88	5.94	6.00	6.06	6.11	6.16	6.21	6.26
38	3.83	4.38	4.71	4.95	5.13	5.29	5.41	5.53	5.62	5.71	5.79	5.86	5.93	5.99	6.05	6.10	6.15	6.20	6.24
39	3.83	4.37	4.70	4.94	5.12	5.28	5.40	5.51	5.61	5.70	5.78	5.85	5.91	5.97	6.03	6.08	6.13	6.18	6.22
40	3.82	4.37	4.70	4.93	5.11	5.26	5.39	5.50	5.60	5.69	5.76	5.83	5.90	5.96	6.02	6.07	6.12	6.16	6.21
41	3.82	4.36	4.69	4.92	5.11	5.26	5.38	5.49	5.59	5.67	5.75	5.82	5.89	5.95	6.00	6.06	6.10	6.15	6.19
42	3.82	4.35	4.68	4.91	5.10	5.25	5.37	5.48	5.58	5.66	5.74	5.81	5.88	5.94	5.99	6.04	6.09	6.14	6.18
43	3.81	4.35	4.67	4.91	5.09	5.24	5.36	5.47	5.57	5.65	5.73	5.80	5.86	5.92	5.98	6.03	6.08	6.12	6.17
44	3.81	4.34	4.67	4.90	5.08	5.23	5.35	5.46	5.56	5.64	5.72	5.79	5.85	5.91	5.97	6.02	6.07	6.11	6.15
45	3.80	4.34	4.66	4.89	5.07	5.22	5.34	5.45	5.55	5.63	5.71	5.78	5.84	5.90	5.96	6.01	6.06	6.10	6.14
46	3.80	4.33	4.66	4.89	5.07	5.21	5.34	5.44	5.54	5.62	5.70	5.77	5.83	5.89	5.95	6.00	6.04	6.09	6.13
47	3.80	4.33	4.65	4.88	5.06	5.21	5.33	5.44	5.53	5.61	5.69	5.76	5.82	5.88	5.94	5.99	6.03	6.08	6.12
48	3.79	4.32	4.64	4.87	5.05	5.20	5.32	5.43	5.52	5.61	5.68	5.75	5.81	5.87	5.93	5.98	6.02	6.07	6.11
49	3.79	4.32	4.64	4.87	5.05	5.19	5.31	5.42	5.51	5.60	5.67	5.74	5.80	5.86	5.92	5.97	6.01	6.06	6.10
50	3.79	4.32	4.63	4.86	5.04	5.19	5.31	5.41	5.51	5.59	5.67	5.73	5.80	5.85	5.91	5.96	6.01	6.05	6.09
51	3.78	4.31	4.63	4.86	5.03	5.18	5.30	5.41	5.50	5.58	5.66	5.73	5.79	5.85	5.90	5.95	6.00	6.04	6.08
52	3.78	4.31	4.63	4.85	5.03	5.17	5.29	5.40	5.49	5.58	5.65	5.72	5.78	5.84	5.89	5.94	5.99	6.03	6.07
53	3.78	4.30	4.62	4.85	5.02	5.17	5.29	5.39	5.49	5.57	5.64	5.71	5.77	5.83	5.88	5.93	5.98	6.02	6.07
54	3.78	4.30	4.62	4.84	5.02	5.16	5.28	5.39	5.48	5.56	5.64	5.70	5.77	5.82	5.88	5.93	5.97	6.02	6.06
55	3.77	4.30	4.61	4.84	5.01	5.16	5.28	5.38	5.47	5.56	5.63	5.70	5.76	5.82	5.87	5.92	5.96	6.01	6.05
56	3.77	4.29	4.61	4.83	5.01	5.15	5.27	5.38	5.47	5.55	5.62	5.69	5.75	5.81	5.86	5.91	5.96	6.00	6.04
57	3.77	4.29	4.60	4.83	5.00	5.15	5.27	5.37	5.46	5.54	5.62	5.68	5.75	5.80	5.86	5.90	5.95	5.99	6.03
58	3.77	4.29	4.60	4.83	5.00	5.14	5.26	5.37	5.46	5.54	5.61	5.68	5.74	5.80	5.85	5.90	5.94	5.99	6.03

(*Continued*)

Table 2b Studentized Range Distribution Critical Values for $\alpha = .01$ (*Continued*)

	J																		
df	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
59	3.76	4.29	4.60	4.82	5.00	5.14	5.26	5.36	5.45	5.53	5.61	5.67	5.73	5.79	5.84	5.89	5.94	5.98	6.02
60	3.76	4.28	4.59	4.82	4.99	5.13	5.25	5.36	5.45	5.53	5.60	5.67	5.73	5.78	5.84	5.89	5.93	5.97	6.01
61	3.76	4.28	4.59	4.81	4.99	5.13	5.25	5.35	5.44	5.52	5.60	5.66	5.72	5.78	5.83	5.88	5.93	5.97	6.01
62	3.76	4.28	4.59	4.81	4.98	5.12	5.24	5.35	5.44	5.52	5.59	5.66	5.72	5.77	5.83	5.87	5.92	5.96	6.00
63	3.76	4.27	4.59	4.81	4.98	5.12	5.24	5.34	5.43	5.51	5.59	5.65	5.71	5.77	5.82	5.87	5.91	5.96	6.00
64	3.75	4.27	4.58	4.80	4.98	5.12	5.24	5.34	5.43	5.51	5.58	5.65	5.71	5.76	5.81	5.86	5.91	5.95	5.99
65	3.75	4.27	4.58	4.80	4.97	5.11	5.23	5.33	5.42	5.50	5.58	5.64	5.70	5.76	5.81	5.86	5.90	5.95	5.99
66	3.75	4.27	4.58	4.80	4.97	5.11	5.23	5.33	5.42	5.50	5.57	5.64	5.70	5.75	5.80	5.85	5.90	5.94	5.98
67	3.75	4.26	4.57	4.79	4.97	5.11	5.22	5.33	5.42	5.50	5.57	5.63	5.69	5.75	5.80	5.85	5.89	5.94	5.98
68	3.75	4.26	4.57	4.79	4.96	5.10	5.22	5.32	5.41	5.49	5.56	5.63	5.69	5.74	5.80	5.84	5.89	5.93	5.97
69	3.75	4.26	4.57	4.79	4.96	5.10	5.22	5.32	5.41	5.49	5.56	5.62	5.68	5.74	5.79	5.84	5.88	5.93	5.97
70	3.74	4.26	4.57	4.79	4.96	5.10	5.21	5.31	5.40	5.48	5.56	5.62	5.68	5.74	5.79	5.83	5.88	5.92	5.96
71	3.74	4.26	4.56	4.78	4.95	5.09	5.21	5.31	5.40	5.48	5.55	5.62	5.68	5.73	5.78	5.83	5.87	5.92	5.96
72	3.74	4.25	4.56	4.78	4.95	5.09	5.21	5.31	5.40	5.48	5.55	5.61	5.67	5.73	5.78	5.83	5.87	5.91	5.95
73	3.74	4.25	4.56	4.78	4.95	5.09	5.20	5.30	5.39	5.47	5.54	5.61	5.67	5.72	5.77	5.82	5.87	5.91	5.95
74	3.74	4.25	4.56	4.78	4.95	5.08	5.20	5.30	5.39	5.47	5.54	5.61	5.66	5.72	5.77	5.82	5.86	5.90	5.94
75	3.74	4.25	4.56	4.77	4.94	5.08	5.20	5.30	5.39	5.47	5.54	5.60	5.66	5.72	5.77	5.81	5.86	5.90	5.94
76	3.74	4.25	4.55	4.77	4.94	5.08	5.20	5.30	5.38	5.46	5.53	5.60	5.66	5.71	5.76	5.81	5.85	5.90	5.94
77	3.74	4.25	4.55	4.77	4.94	5.08	5.19	5.29	5.38	5.46	5.53	5.59	5.65	5.71	5.76	5.81	5.85	5.89	5.93
78	3.73	4.24	4.55	4.77	4.94	5.07	5.19	5.29	5.38	5.46	5.53	5.59	5.65	5.71	5.76	5.80	5.85	5.89	5.93
79	3.73	4.24	4.55	4.76	4.93	5.07	5.19	5.29	5.38	5.45	5.52	5.59	5.65	5.70	5.75	5.80	5.84	5.88	5.92
80	3.73	4.24	4.55	4.76	4.93	5.07	5.18	5.28	5.37	5.45	5.52	5.59	5.64	5.70	5.75	5.80	5.84	5.88	5.92
81	3.73	4.24	4.54	4.76	4.93	5.07	5.18	5.28	5.37	5.45	5.52	5.58	5.64	5.70	5.75	5.79	5.84	5.88	5.92
82	3.73	4.24	4.54	4.76	4.93	5.06	5.18	5.28	5.37	5.44	5.52	5.58	5.64	5.69	5.74	5.79	5.83	5.87	5.91
83	3.73	4.24	4.54	4.76	4.92	5.06	5.18	5.28	5.36	5.44	5.51	5.58	5.64	5.69	5.74	5.79	5.83	5.87	5.91
84	3.73	4.23	4.54	4.75	4.92	5.06	5.17	5.27	5.36	5.44	5.51	5.57	5.63	5.69	5.74	5.78	5.83	5.87	5.91
85	3.73	4.23	4.54	4.75	4.92	5.06	5.17	5.27	5.36	5.44	5.51	5.57	5.63	5.68	5.73	5.78	5.82	5.87	5.90
86	3.73	4.23	4.54	4.75	4.92	5.06	5.17	5.27	5.36	5.43	5.50	5.57	5.63	5.68	5.73	5.78	5.82	5.86	5.90
87	3.72	4.23	4.53	4.75	4.92	5.05	5.17	5.27	5.35	5.43	5.50	5.57	5.62	5.68	5.73	5.77	5.82	5.86	5.90
88	3.72	4.23	4.53	4.75	4.92	5.05	5.17	5.27	5.35	5.43	5.50	5.56	5.62	5.68	5.73	5.77	5.82	5.86	5.89
89	3.72	4.23	4.53	4.75	4.91	5.05	5.16	5.26	5.35	5.43	5.50	5.56	5.62	5.67	5.72	5.77	5.81	5.85	5.89
90	3.72	4.23	4.53	4.74	4.91	5.05	5.16	5.26	5.35	5.43	5.49	5.56	5.62	5.67	5.72	5.77	5.81	5.85	5.89
91	3.72	4.23	4.53	4.74	4.91	5.05	5.16	5.26	5.35	5.42	5.49	5.56	5.61	5.67	5.72	5.76	5.81	5.85	5.89
92	3.72	4.22	4.53	4.74	4.91	5.04	5.16	5.26	5.34	5.42	5.49	5.55	5.61	5.67	5.71	5.76	5.80	5.85	5.88
93	3.72	4.22	4.52	4.74	4.91	5.04	5.16	5.25	5.34	5.42	5.49	5.55	5.61	5.66	5.71	5.76	5.80	5.84	5.88
94	3.72	4.22	4.52	4.74	4.90	5.04	5.15	5.25	5.34	5.42	5.49	5.55	5.61	5.66	5.71	5.76	5.80	5.84	5.88
95	3.72	4.22	4.52	4.74	4.90	5.04	5.15	5.25	5.34	5.41	5.48	5.55	5.60	5.66	5.71	5.75	5.80	5.84	5.88
96	3.72	4.22	4.52	4.74	4.90	5.04	5.15	5.25	5.34	5.41	5.48	5.54	5.60	5.66	5.71	5.75	5.79	5.84	5.87
97	3.72	4.22	4.52	4.73	4.90	5.04	5.15	5.25	5.33	5.41	5.48	5.54	5.60	5.65	5.70	5.75	5.79	5.83	5.87
98	3.72	4.22	4.52	4.73	4.90	5.03	5.15	5.25	5.33	5.41	5.48	5.54	5.60	5.65	5.70	5.75	5.79	5.83	5.87
99	3.71	4.22	4.52	4.73	4.90	5.03	5.15	5.24	5.33	5.41	5.48	5.54	5.60	5.65	5.70	5.74	5.79	5.83	5.87
100	3.71	4.22	4.52	4.73	4.90	5.03	5.14	5.24	5.33	5.40	5.47	5.54	5.59	5.65	5.70	5.74	5.79	5.83	5.86
105	3.71	4.21	4.51	4.72	4.89	5.02	5.14	5.23	5.32	5.40	5.47	5.53	5.58	5.64	5.69	5.73	5.78	5.82	5.85
110	3.71	4.21	4.51	4.72	4.88	5.02	5.13	5.23	5.31	5.39	5.46	5.52	5.58	5.63	5.68	5.72	5.77	5.81	5.84
115	3.70	4.20	4.50	4.71	4.88	5.01	5.12	5.22	5.31	5.38	5.45	5.51	5.57	5.62	5.67	5.72	5.76	5.80	5.84
120	3.70	4.20	4.50	4.71	4.87	5.01	5.12	5.21	5.30	5.37	5.44	5.50	5.56	5.61	5.66	5.71	5.75	5.79	5.83
125	3.70	4.20	4.49	4.70	4.87	5.00	5.11	5.21	5.29	5.37	5.44	5.50	5.55	5.61	5.66	5.70	5.74	5.78	5.82
130	3.70	4.19	4.49	4.70	4.86	5.00	5.11	5.20	5.29	5.36	5.43	5.49	5.55	5.60	5.65	5.69	5.74	5.78	5.81
135	3.69	4.19	4.49	4.70	4.86	4.99	5.10	5.20	5.28	5.36	5.43	5.49	5.54	5.60	5.64	5.69	5.73	5.77	5.81
140	3.69	4.19	4.48	4.69	4.86	4.99	5.10	5.19	5.28	5.35	5.42	5.48	5.54	5.59	5.64	5.68	5.72	5.76	5.80
145	3.69	4.19	4.48	4.69	4.85	4.98	5.09	5.19	5.27	5.35	5.42	5.48	5.53	5.58	5.63	5.68	5.72	5.76	5.80
150	3.69	4.18	4.48	4.69	4.85	4.98	5.09	5.19	5.27	5.34	5.41	5.47	5.53	5.58	5.63	5.67	5.71	5.75	5.79
∞	3.64	4.12	4.40	4.60	4.76	4.88	4.99	5.08	5.16	5.23	5.29	5.35	5.40	5.45	5.49	5.54	5.57	5.61	5.65

This table was produced by Dr. J. W. McKean.

Table 3 Durbin–Watson Statistic: 5% Significance Points of d_L and d_U

	$m=1$		$m=2$		$m=3$		$m=4$		$m=5$		$m=6$		$m=7$		$m=8$		$m=9$		$m=10$	
n	d_L	d_U	d_L	d_U	d_L	d_U	d_L	d_U	d_L	d_U	d_L	d_U	d_L	d_U	d_L	d_U	d_L	d_U	d_L	d_U
6	0.610	1.400	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—
7	0.700	1.356	0.467	1.896	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—
8	0.763	1.332	0.559	1.777	0.367	2.287	—	—	—	—	—	—	—	—	—	—	—	—	—	—
9	0.824	1.320	0.629	1.699	0.455	2.128	0.296	2.588	—	—	—	—	—	—	—	—	—	—	—	—
10	0.879	1.320	0.697	1.641	0.525	2.016	0.376	2.414	0.243	2.822	—	—	—	—	—	—	—	—	—	—
11	0.927	1.324	0.758	1.604	0.595	1.928	0.444	2.283	0.315	2.645	0.203	3.004	—	—	—	—	—	—	—	—
12	0.971	1.331	0.812	1.579	0.658	1.864	0.512	2.177	0.380	2.506	0.268	2.832	0.171	3.149	—	—	—	—	—	—
13	1.010	1.340	0.861	1.562	0.715	1.816	0.574	2.094	0.444	2.390	0.328	2.692	0.230	2.985	0.147	3.266	—	—	—	—
14	1.045	1.350	0.905	1.551	0.767	1.779	0.632	2.030	0.505	2.296	0.389	2.572	0.286	2.848	0.200	3.111	0.127	3.360	—	—
15	1.077	1.361	0.946	1.543	0.814	1.750	0.685	1.977	0.562	2.220	0.447	2.471	0.343	2.727	0.251	2.979	0.175	3.216	0.111	3.438
16	1.106	1.371	0.982	1.539	0.857	1.728	0.734	1.935	0.615	2.157	0.502	2.388	0.398	2.624	0.304	2.860	0.222	3.090	0.155	3.304
17	1.133	1.381	1.015	1.536	0.897	1.710	0.779	1.900	0.664	2.104	0.554	2.318	0.451	2.537	0.356	2.757	0.272	2.975	0.198	3.184
18	1.158	1.391	1.046	1.535	0.933	1.696	0.820	1.872	0.710	2.060	0.603	2.258	0.502	2.461	0.407	2.668	0.321	2.873	0.244	3.073
19	1.180	1.401	1.074	1.536	0.967	1.685	0.859	1.848	0.752	2.023	0.649	2.206	0.549	2.396	0.456	2.589	0.369	2.783	0.290	2.974
20	1.201	1.411	1.100	1.537	0.998	1.676	0.894	1.828	0.792	1.991	0.691	2.162	0.595	2.339	0.502	2.521	0.416	2.704	0.336	2.885
21	1.221	1.420	1.125	1.538	1.026	1.669	0.927	1.812	0.829	1.964	0.731	2.124	0.637	2.290	0.546	2.461	0.461	2.633	0.380	2.806
22	1.239	1.429	1.147	1.541	1.053	1.664	0.958	1.797	0.863	1.940	0.769	2.090	0.677	2.246	0.588	2.407	0.504	2.571	0.424	2.735
23	1.257	1.437	1.168	1.543	1.078	1.660	0.986	1.785	0.895	1.920	0.804	2.061	0.715	2.208	0.628	2.360	0.545	2.514	0.465	2.670
24	1.273	1.446	1.188	1.546	1.101	1.656	1.013	1.775	0.925	1.902	0.837	2.035	0.750	2.174	0.666	2.318	0.584	2.464	0.506	2.613
25	1.288	1.454	1.206	1.550	1.123	1.654	1.038	1.767	0.953	1.886	0.868	2.013	0.784	2.144	0.702	2.280	0.621	2.419	0.544	2.560
26	1.302	1.461	1.224	1.553	1.143	1.652	1.062	1.759	0.979	1.873	0.897	1.992	0.816	2.117	0.735	2.246	0.657	2.379	0.581	2.513
27	1.316	1.469	1.240	1.556	1.162	1.651	1.084	1.753	1.004	1.861	0.925	1.974	0.845	2.093	0.767	2.216	0.691	2.342	0.616	2.470
28	1.328	1.476	1.255	1.560	1.181	1.650	1.104	1.747	1.028	1.850	0.951	1.959	0.874	2.071	0.798	2.188	0.723	2.309	0.649	2.431
29	1.341	1.483	1.270	1.563	1.198	1.650	1.124	1.743	1.050	1.841	0.975	1.944	0.900	2.052	0.826	2.164	0.753	2.278	0.681	2.396

(*Continued*)

Table 3 Durbin–Watson Statistic: 5% Significance Points of d_L and d_U (*Continued*)

	$m=1$		$m=2$		$m=3$		$m=4$		$m=5$		$m=6$		$m=7$		$m=8$		$m=9$		$m=10$	
n	d_L	d_U	d_L	d_U	d_L	d_U	d_L	d_U	d_L	d_U	d_L	d_U	d_L	d_U	d_L	d_U	d_L	d_U	d_L	d_U
30	1.352	1.489	1.284	1.567	1.214	1.650	1.143	1.739	1.071	1.833	0.998	1.931	0.926	2.034	0.854	2.141	0.782	2.251	0.712	2.363
31	1.363	1.496	1.297	1.570	1.229	1.650	1.160	1.735	1.090	1.825	1.020	1.920	0.950	2.018	0.879	2.120	0.810	2.226	0.741	2.333
32	1.373	1.502	1.309	1.574	1.244	1.650	1.177	1.732	1.109	1.819	1.041	1.909	0.972	2.004	0.904	2.102	0.836	2.203	0.769	2.306
33	1.383	1.508	1.321	1.577	1.258	1.651	1.193	1.730	1.127	1.813	1.061	1.900	0.994	1.991	0.927	2.085	0.861	2.181	0.796	2.281
34	1.393	1.514	1.333	1.580	1.271	1.652	1.208	1.728	1.144	1.808	1.079	1.891	1.015	1.978	0.950	2.069	0.885	2.162	0.821	2.257
35	1.402	1.519	1.343	1.584	1.283	1.653	1.222	1.726	1.160	1.803	1.097	1.884	1.034	1.967	0.971	2.054	0.908	2.144	0.845	2.236
36	1.411	1.525	1.354	1.587	1.295	1.654	1.236	1.724	1.175	1.799	1.114	1.876	1.053	1.957	0.991	2.041	0.930	2.127	0.868	2.216
37	1.419	1.530	1.364	1.590	1.307	1.655	1.249	1.723	1.190	1.795	1.131	1.870	1.071	1.948	1.011	2.029	0.951	2.112	0.891	2.197
38	1.427	1.535	1.373	1.594	1.318	1.656	1.261	1.722	1.204	1.792	1.146	1.864	1.088	1.939	1.029	2.017	0.970	2.098	0.912	2.180
39	1.435	1.540	1.382	1.597	1.328	1.658	1.273	1.722	1.218	1.789	1.161	1.859	1.104	1.932	1.047	2.007	0.990	2.085	0.932	2.164
40	1.442	1.544	1.391	1.600	1.338	1.659	1.285	1.721	1.230	1.786	1.175	1.854	1.120	1.924	1.064	1.997	1.008	2.072	0.952	2.149
45	1.475	1.566	1.430	1.615	1.383	1.666	1.336	1.720	1.287	1.776	1.238	1.835	1.189	1.895	1.139	1.958	1.089	2.022	1.038	2.088
50	1.503	1.585	1.462	1.628	1.421	1.674	1.378	1.721	1.335	1.771	1.291	1.822	1.246	1.875	1.201	1.930	1.156	1.986	1.110	2.044
55	1.528	1.601	1.490	1.641	1.452	1.681	1.414	1.724	1.374	1.768	1.334	1.814	1.294	1.861	1.253	1.909	1.212	1.959	1.170	2.010
60	1.549	1.616	1.514	1.652	1.480	1.689	1.444	1.727	1.408	1.767	1.372	1.808	1.335	1.850	1.298	1.894	1.260	1.939	1.222	1.984
65	1.549	1.616	1.514	1.652	1.480	1.689	1.444	1.727	1.408	1.767	1.372	1.808	1.335	1.850	1.298	1.894	1.260	1.939	1.222	1.984
70	1.583	1.641	1.554	1.672	1.525	1.703	1.494	1.735	1.464	1.768	1.433	1.802	1.401	1.838	1.369	1.874	1.337	1.910	1.305	1.948
75	1.598	1.652	1.571	1.680	1.543	1.709	1.515	1.739	1.487	1.770	1.458	1.801	1.428	1.834	1.399	1.867	1.369	1.901	1.339	1.935
80	1.611	1.662	1.586	1.688	1.560	1.715	1.534	1.743	1.507	1.772	1.480	1.801	1.453	1.831	1.425	1.861	1.397	1.893	1.369	1.925
85	1.624	1.671	1.600	1.696	1.575	1.721	1.550	1.747	1.525	1.774	1.500	1.801	1.474	1.829	1.448	1.857	1.422	1.886	1.396	1.916
90	1.635	1.679	1.612	1.703	1.589	1.726	1.566	1.751	1.542	1.776	1.518	1.801	1.494	1.827	1.469	1.854	1.445	1.881	1.420	1.909
95	1.645	1.687	1.623	1.709	1.602	1.732	1.579	1.755	1.557	1.778	1.535	1.802	1.512	1.827	1.489	1.852	1.465	1.877	1.442	1.903
100	1.654	1.694	1.634	1.715	1.613	1.736	1.592	1.758	1.571	1.780	1.550	1.803	1.528	1.826	1.506	1.850	1.484	1.874	1.462	1.898
150	1.720	1.746	1.706	1.760	1.693	1.774	1.679	1.788	1.665	1.802	1.651	1.817	1.637	1.832	1.622	1.847	1.608	1.862	1.594	1.877
200	1.758	1.778	1.748	1.789	1.738	1.799	1.728	1.810	1.718	1.820	1.707	1.831	1.697	1.841	1.686	1.852	1.675	1.863	1.665	1.874

	$m = 11$		$m = 12$		$m = 13$		$m = 14$		$M = 15$		$m = 16$		$m = 17$		$m = 18$		$m = 19$		$m = 20$	
n	d_L	d_U	d_L	d_U	d_L	d_U	d_L	d_U	d_L	d_U	d_L	d_U	d_L	d_U	d_L	d_U	d_L	d_U	d_L	d_U
16	0.098	3.503	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—
17	0.138	3.378	0.087	3.557	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—
18	0.177	3.265	0.123	3.441	0.078	3.603	—	—	—	—	—	—	—	—	—	—	—	—	—	—
19	0.220	3.159	0.160	3.335	0.111	3.496	0.070	3.642	—	—	—	—	—	—	—	—	—	—	—	—
20	0.263	3.063	0.200	3.234	0.145	3.395	0.100	3.542	0.063	3.676	—	—	—	—	—	—	—	—	—	—
21	0.307	2.976	0.240	3.141	0.182	3.300	0.132	3.448	0.091	3.583	0.058	3.705	—	—	—	—	—	—	—	—
22	0.349	2.897	0.281	3.057	0.220	3.211	0.166	3.358	0.120	3.495	0.083	3.619	0.052	3.731	—	—	—	—	—	—
23	0.391	2.826	0.322	2.979	0.259	3.128	0.202	3.272	0.153	3.409	0.110	3.535	0.076	3.650	0.048	3.753	—	—	—	—
24	0.431	2.761	0.362	2.908	0.297	3.053	0.239	3.193	0.186	3.327	0.141	3.454	0.101	3.572	0.070	3.678	0.044	3.773	—	—
25	0.470	2.702	0.400	2.844	0.335	2.983	0.275	3.119	0.221	3.251	0.172	3.376	0.130	3.494	0.094	3.604	0.065	3.702	0.041	3.790
26	0.508	2.649	0.438	2.784	0.373	2.919	0.312	3.051	0.256	3.179	0.205	3.303	0.160	3.420	0.120	3.531	0.087	3.632	0.060	3.724
27	0.544	2.600	0.475	2.730	0.409	2.859	0.348	2.987	0.291	3.112	0.238	3.233	0.191	3.349	0.149	3.460	0.112	3.563	0.081	3.658
28	0.578	2.555	0.510	2.680	0.445	2.805	0.383	2.928	0.325	3.050	0.271	3.168	0.222	3.283	0.178	3.392	0.138	3.495	0.104	3.592
29	0.612	2.515	0.544	2.634	0.479	2.755	0.418	2.874	0.359	2.992	0.305	3.107	0.254	3.219	0.208	3.327	0.166	3.431	0.129	3.528
30	0.643	2.477	0.577	2.592	0.512	2.708	0.451	2.823	0.392	2.937	0.337	3.050	0.286	3.160	0.238	3.266	0.195	3.368	0.156	3.465
31	0.674	2.443	0.608	2.553	0.545	2.665	0.484	2.776	0.425	2.887	0.370	2.996	0.317	3.103	0.269	3.208	0.224	3.309	0.183	3.406
32	0.703	2.411	0.638	2.517	0.576	2.625	0.515	2.733	0.457	2.840	0.401	2.946	0.349	3.050	0.299	3.153	0.253	3.252	0.211	3.348
33	0.731	2.382	0.668	2.484	0.606	2.588	0.546	2.692	0.488	2.796	0.432	2.899	0.379	3.000	0.329	3.100	0.283	3.198	0.239	3.293
34	0.758	2.355	0.695	2.454	0.634	2.554	0.575	2.654	0.518	2.754	0.462	2.854	0.409	2.954	0.359	3.051	0.312	3.147	0.267	3.240
35	0.783	2.330	0.722	2.425	0.662	2.521	0.604	2.619	0.547	2.716	0.492	2.813	0.439	2.910	0.388	3.005	0.340	3.099	0.295	3.190
36	0.808	2.306	0.748	2.398	0.689	2.492	0.631	2.586	0.575	2.680	0.520	2.774	0.467	2.868	0.417	2.961	0.369	3.053	0.323	3.142

(*Continued*)

Table 3 Durbin–Watson Statistic: 5% Significance Points of d_L and d_U (*Continued*)

	$m = 11$		$m = 12$		$m = 13$		$m = 14$		$M = 15$		$m = 16$		$m = 17$		$m = 18$		$m = 19$		$m = 20$	
n	d_L	d_U	d_L	d_U	d_L	d_U	d_L	d_U	d_L	d_U	d_L	d_U	d_L	d_U	d_L	d_U	d_L	d_U	d_L	d_U
37	0.831	2.285	0.772	2.374	0.714	2.464	0.657	2.555	0.602	2.646	0.548	2.738	0.495	2.829	0.445	2.920	0.397	3.009	0.351	3.097
38	0.854	2.265	0.796	2.351	0.739	2.438	0.683	2.526	0.628	2.614	0.575	2.703	0.522	2.792	0.472	2.880	0.424	2.968	0.378	3.054
39	0.875	2.246	0.819	2.329	0.763	2.413	0.707	2.499	0.653	2.585	0.600	2.671	0.549	2.757	0.499	2.843	0.451	2.929	0.404	3.013
40	0.896	2.228	0.840	2.309	0.785	2.391	0.731	2.473	0.678	2.557	0.626	2.641	0.575	2.724	0.525	2.808	0.477	2.829	0.430	2.974
45	0.988	2.156	0.938	2.225	0.887	2.296	0.838	2.367	0.788	2.439	0.740	2.512	0.692	2.586	0.644	2.659	0.598	2.733	0.553	2.807
50	1.064	2.103	1.019	2.163	0.973	2.225	0.927	2.287	0.882	2.350	0.836	2.414	0.792	2.479	0.747	2.544	0.703	2.610	0.660	2.675
55	1.129	2.062	1.087	2.116	1.045	2.170	1.003	2.225	0.961	2.281	0.919	2.338	0.877	2.396	0.836	2.454	0.795	2.512	0.754	2.571
60	1.184	2.031	1.145	2.079	1.106	2.127	1.068	2.177	1.029	2.227	0.990	2.278	0.951	2.330	0.913	2.382	0.874	2.434	0.836	2.487
65	1.231	2.006	1.195	2.049	1.160	2.093	1.124	2.138	1.088	2.183	1.052	2.229	1.016	2.276	0.980	2.323	0.944	2.371	0.908	2.419
70	1.272	1.987	1.239	2.026	1.206	2.066	1.172	2.106	1.139	2.148	1.105	2.189	1.072	2.232	1.038	2.275	1.005	2.318	0.971	2.362
75	1.308	1.970	1.277	2.006	1.247	2.043	1.215	2.080	1.184	2.118	1.153	2.156	1.121	2.195	1.090	2.235	1.058	2.275	1.027	2.315
80	1.340	1.957	1.311	1.991	1.283	2.024	1.253	2.059	1.224	2.093	1.195	2.129	1.165	2.165	1.136	2.201	1.106	2.238	1.076	2.275
85	1.369	1.946	1.342	1.977	1.315	2.009	1.287	2.040	1.260	2.073	1.232	2.105	1.205	2.139	1.177	2.172	1.149	2.206	1.121	2.241
90	1.395	1.937	1.369	1.966	1.344	1.995	1.318	2.025	1.292	2.055	1.266	2.085	1.240	2.116	1.213	2.148	1.187	2.179	1.160	2.211
95	1.418	1.930	1.394	1.956	1.370	1.984	1.345	2.012	1.321	2.040	1.296	2.068	1.271	2.097	1.247	2.126	1.222	2.156	1.197	2.186
100	1.439	1.923	1.416	1.948	1.393	1.974	1.371	2.000	1.347	2.026	1.324	2.053	1.301	2.080	1.277	2.108	1.253	2.135	1.229	2.164
150	1.579	1.892	1.564	1.908	1.550	1.924	1.535	1.940	1.519	1.956	1.504	1.972	1.489	1.989	1.474	2.006	1.458	2.023	1.443	2.040
200	1.654	1.885	1.643	1.896	1.632	1.908	1.621	1.919	1.610	1.931	1.599	1.943	1.588	1.955	1.576	1.967	1.565	1.979	1.554	1.991

Note: m is the number of predictors in the model not counting the intercept. For example, the four-parameter model discussed in Chapters 18 and 19 has an intercept and three predictors so the appropriate column is headed $m = 3$.

Table 4 Random Numbers

44507	24991	94455	88127	50325	41204	51909	24997	72017	89632
04758	43926	28608	41275	40064	79797	73058	15054	99131	48898
99663	31230	85271	46997	85748	42618	34113	86803	65763	83729
94542	54445	57440	48094	25998	01934	14532	32048	92871	02993
62244	42702	81736	32834	87559	61333	20763	10073	13978	84515
13431	28316	61151	52698	00415	10449	28673	29237	67710	34287
14476	99842	56185	25293	59017	48819	17325	06653	36375	05611
70271	58447	57390	19716	75997	62025	05480	34396	08531	01869
42196	80744	05143	02394	34162	67288	30289	42652	93916	17957
80195	66132	48244	69203	83131	16918	98872	58326	81233	53274
53713	38541	33788	73726	03498	55638	06734	70905	80427	61479
18684	01959	74479	51380	64581	43192	40981	35955	63345	81217
13641	37716	93287	83446	52527	02732	02125	86511	51926	09616
77742	95365	63632	57876	29266	64967	17172	53507	05181	18201
05991	66491	57647	73000	26029	11539	29290	95711	38727	68868
41047	31938	83562	68340	28295	87122	10950	63388	63547	84855
71403	85980	07484	86990	66197	44671	90578	51054	54688	31079
45596	48708	50671	77192	37832	93376	33312	04402	56745	13083
82559	78778	01584	92545	71846	38214	99208	05445	78201	46782
47506	38035	27926	15153	83363	19031	76697	95007	35758	17151
70543	10916	87967	58317	98465	56499	52036	73908	81801	71209
42079	34607	26599	58408	30975	24166	67088	35008	87882	22039
84219	89487	63240	59301	72455	84972	60590	98755	80966	93178
29906	84782	05389	57258	38982	19040	82443	98746	93212	51929
79652	20559	49893	59863	96356	21200	46644	70360	58959	82927
04544	44952	71709	87883	32369	14523	70271	55973	45846	32594
85178	34981	56729	08012	92980	54225	20568	31461	79341	55929
57162	44482	44227	07667	59131	00007	00049	07492	18225	53494
20362	95973	30255	94418	81472	24820	00033	73868	94376	60802
05969	73218	54875	00153	29951	13951	29247	84061	18838	11362
50667	25450	06137	14662	38525	03354	95763	17356	58014	42275
17226	77609	92152	55513	97232	01720	89035	47906	97948	17579
25668	78071	30134	38750	34057	92889	68199	94606	78137	90191
44240	03450	16920	82934	17947	00202	16116	73010	61995	27451
82297	34414	88532	75214	39524	53378	68742	92652	46937	78292
46493	90984	55621	38276	96460	70141	22536	02927	15152	65201
68466	70549	11451	08902	71775	19519	14015	95057	00284	06159
81045	95612	64204	89036	34140	87751	59798	64857	93873	47960
01046	54879	21034	58143	27524	37003	26169	53551	49569	34504
36589	10073	11580	35791	81481	83307	96366	34919	96156	55987
22906	90648	82313	36576	18176	90892	81992	64607	70616	12013
58152	02996	52667	96673	66481	30326	32884	37762	06298	86378
75806	19197	33297	89617	48782	23900	22258	16121	84915	74287
97166	77829	80136	99980	79631	43375	48276	48505	27228	74718
32677	31335	80259	01592	69523	98513	87666	62293	63669	04884
96791	52350	88781	10033	01117	25031	22272	22745	84317	07101
83933	05912	76686	83218	23920	81070	74040	33709	81053	34407
33271	46204	73684	45559	11307	52450	85862	26079	22429	78450
82437	69096	81952	10057	18734	03043	66065	99860	58895	00852
78368	25901	66006	56006	38974	38151	66413	91178	11938	56728

(*Continued*)

Table 4 Random Numbers (*Continued*)

11947	45801	88860	55623	56424	67730	12486	17188	28516	54001
19821	99200	80143	10817	51042	33277	03100	63518	31196	60084
84455	54188	61689	49290	44794	26614	26192	96329	53388	94645
28671	69683	52680	20321	20892	53578	46372	36082	63185	95248
56514	14599	94180	11047	54885	13430	98637	20679	26972	88137
40766	03571	29944	72664	52334	14618	98310	32760	65937	78604
56860	44483	72204	48966	97036	53781	03484	91662	46059	11175
79059	61802	99592	66920	93660	82129	39311	11425	76038	01863
89934	64293	24560	54020	92869	19415	06805	32375	78928	23056
11672	21626	69013	73666	59780	27910	78609	66016	92813	41216
27796	70535	70439	55287	73439	46586	72679	03266	92218	96979
46770	71056	76333	08664	76706	96282	83775	86428	69100	64580
99382	11399	65085	13727	25102	31307	81679	81645	14915	88038
03509	82625	54258	11676	48592	63351	38972	15514	86903	46021
85677	66375	52022	81141	26219	20759	87769	26924	50442	96483
17590	47907	90408	64250	43395	96099	60757	27625	80848	89757
06108	00695	05893	91471	30276	09918	44770	50983	28430	26255
59865	85004	37862	02579	50008	76623	11616	88212	15064	61659
54797	76621	32714	51463	97044	95006	98252	10314	49602	33558
95648	98415	12014	17006	05401	52883	56367	88057	11733	49952
77061	46028	92090	38444	25548	01654	07883	91203	36265	63780
70974	06268	18991	57128	97941	48275	94701	35959	20280	14401
49070	17428	12864	06965	04979	46450	92239	82592	51090	05533
58608	38292	45992	00309	51487	93539	51808	35805	58275	15842
96758	17329	29750	84825	29629	40525	72582	90522	21596	37692
99086	08135	50709	46089	00783	31968	60420	22980	24615	56847
80788	37625	06451	44670	98073	45260	92482	19490	28485	88843
29187	87198	36701	65034	03081	72858	43882	91235	16529	03257
75528	18326	98081	73611	39015	09363	45722	75837	80964	32628
64958	02207	96367	18388	10264	64468	14684	25852	65331	35555
61660	20811	94115	50474	39634	60792	91490	44569	74458	04425
63761	14729	65648	34498	67825	77812	64166	33513	42205	07026
82240	19793	68420	11447	80102	61521	30946	88203	18749	65557
75835	51401	97868	34580	43443	23744	70055	66077	18860	99023
16024	84279	46428	02250	05301	65909	82574	62304	97315	01811
75701	33106	99267	31418	39184	47810	53387	87456	61877	21111
41382	58470	44806	98582	88727	09691	61525	95304	54887	78715
12566	62597	12068	04707	81372	99367	63312	51782	41761	33848
58676	07557	31504	35506	01297	44645	97330	47684	58803	35907
45341	34281	12228	85533	15372	90718	23141	40745	49572	98435
01378	98319	33490	72751	09993	51458	61853	90398	17073	56594
51942	57064	62098	04267	89393	24936	63007	88985	09972	43549
90878	99833	23149	11340	24756	29098	06170	33057	68590	55258
00036	15089	52890	98970	98181	87008	58579	47004	31240	08663
26037	83932	00830	39834	68747	09907	39754	15009	72328	67438
01612	03756	70643	14987	72697	87485	93281	72719	16664	82772
17976	73847	80995	77024	81605	01599	06187	39702	04769	57404
50068	56735	80790	52153	47732	08955	05036	95737	77945	32831
07758	17528	37281	34863	78317	26881	77573	77993	20787	82766
47803	71550	07697	29874	21955	62555	27615	61044	21883	11868

This table was produced by Dr. J. W. McKean.

Table 5 Bonferroni z Critical Values (C' = number of planned comparisons)

C'	.10 .20	.05 .10	.025 .05	.01 .02	.005 = α for Directional Test .01 = α for Nondirectional Test
1	1.282	1.645	1.960	2.326	2.576
2	1.645	1.960	2.241	2.576	2.807
3	1.834	2.128	2.394	2.713	2.935
4	1.960	2.241	2.498	2.807	3.023
5	2.054	2.326	2.576	2.878	3.090
6	2.128	2.394	2.638	2.935	3.144
7	2.189	2.450	2.690	2.983	3.189
8	2.241	2.498	2.734	3.023	3.227
9	2.287	2.539	2.773	3.059	3.261
10	2.326	2.576	2.807	3.090	3.291
11	2.362	2.609	2.838	3.118	3.317
12	2.394	2.638	2.865	3.144	3.341
13	2.423	2.665	2.891	3.167	3.364
14	2.450	2.690	2.914	3.189	3.384
15	2.475	2.713	2.935	3.209	3.403
16	2.498	2.734	2.955	3.227	3.421
17	2.519	2.754	2.974	3.245	3.437
18	2.539	2.773	2.991	3.261	3.452
19	2.558	2.790	3.008	3.276	3.467
20	2.576	2.807	3.023	3.291	3.481
21	2.593	2.823	3.038	3.304	3.494
22	2.609	2.838	3.052	3.317	3.506
23	2.624	2.852	3.065	3.330	3.518
24	2.638	2.865	3.078	3.341	3.529
25	2.652	2.878	3.090	3.353	3.540
26	2.665	2.891	3.102	3.364	3.550
27	2.678	2.902	3.113	3.374	3.560
28	2.690	2.914	3.124	3.384	3.570
29	2.702	2.925	3.134	3.394	3.579
30	2.713	2.935	3.144	3.403	3.588
31	2.724	2.945	3.154	3.412	3.596
32	2.734	2.955	3.163	3.421	3.605
33	2.744	2.965	3.172	3.429	3.613
34	2.754	2.974	3.180	3.437	3.620
35	2.764	2.983	3.189	3.445	3.628
36	2.773	2.991	3.197	3.452	3.635
37	2.782	3.000	3.205	3.460	3.642
38	2.790	3.008	3.213	3.467	3.649
39	2.799	3.016	3.220	3.474	3.656
40	2.807	3.023	3.227	3.481	3.662
41	2.815	3.031	3.234	3.487	3.669
42	2.823	3.038	3.241	3.494	3.675
43	2.830	3.045	3.248	3.500	3.681
44	2.838	3.052	3.254	3.506	3.687
45	2.845	3.059	3.261	3.512	3.692
46	2.852	3.065	3.267	3.518	3.698
47	2.859	3.072	3.273	3.524	3.703
48	2.865	3.078	3.279	3.529	3.709
49	2.872	3.084	3.285	3.535	3.714
50	2.878	3.090	3.291	3.540	3.719
51	2.884	3.096	3.296	3.545	3.724

This table was produced by Dr. J. W. McKean.

Table 6 Critical Values of the Chi-Square Distribution

df	$\alpha = .05$	$\alpha = .01$
1	3.84	6.63
2	5.99	9.21
3	7.81	11.34
4	9.49	13.28
5	11.07	15.09
6	12.59	16.81
7	14.07	18.48
8	15.51	20.09
9	16.92	21.67
10	18.31	23.21
11	19.68	24.72
12	21.03	26.22
13	22.36	27.69
14	23.68	29.14
15	25.00	30.58
16	26.30	32.00
17	27.59	33.41
18	28.87	34.81
19	30.14	36.19
20	31.41	37.57
21	32.67	38.93
22	33.92	40.29
23	35.17	41.64
24	36.42	42.98
25	37.65	44.31
26	38.89	45.64
27	40.11	46.96
28	41.34	48.28
29	42.56	49.59
30	43.77	50.89
40	55.76	63.69
50	67.50	76.15
60	79.08	88.38
70	90.53	100.42
80	101.88	112.33
90	113.14	124.12
100	124.34	135.81

Table 7 Critical Values of the *t*-Distribution

	Proportion in One Tail		
	.05	.025	.005
	Proportion in Two Tails		
df	.10	.05	.01
1	6.314	12.706	63.657
2	2.920	4.303	9.925
3	2.353	3.182	5.841
4	2.132	2.776	4.604
5	2.015	2.571	4.032
6	1.943	2.447	3.707
7	1.895	2.365	3.499
8	1.860	2.306	3.355
9	1.833	2.262	3.250
10	1.812	2.228	3.169
11	1.796	2.201	3.106
12	1.782	2.179	3.055
13	1.771	2.160	3.012
14	1.761	2.145	2.977
15	1.753	2.131	2.947
16	1.746	2.120	2.921
17	1.740	2.110	2.898
18	1.734	2.101	2.878
19	1.729	2.093	2.861
20	1.725	2.086	2.845
21	1.721	2.080	2.831
22	1.717	2.074	2.819
23	1.714	2.069	2.807
24	1.711	2.064	2.797
25	1.708	2.060	2.787
26	1.706	2.056	2.779
27	1.703	2.052	2.771
28	1.701	2.048	2.763
29	1.699	2.045	2.756
30	1.697	2.042	2.750
40	1.684	2.021	2.704
60	1.671	2.000	2.660
120	1.658	1.980	2.617
∞	1.645	1.960	2.576

Table 8 Critical Values of the *F* Distribution for Evaluating Cook's Distance ($\alpha = .50$)

	Numerator *df*				
Denominator *df*	1	2	3	4	5
1	1.00	1.50	1.71	1.82	1.89
2	.67	1.00	1.13	1.21	1.25
3	.58	.88	1.00	1.06	1.10
4	.55	.83	.94	1.00	1.04
5	.53	.80	.91	.96	1.00
6	.52	.78	.89	.94	.98
7	.51	.77	.87	.93	.96
8	.50	.76	.86	.92	.95
9	.49	.75	.85	.91	.94
10	.49	.74	.84	.90	.93
12	.48	.74	.84	.89	.92
15	.48	.73	.83	.88	.91
20	.47	.72	.82	.87	.90
24	.47	.71	.81	.86	.90
30	.47	.71	.81	.86	.89
60	.46	.70	.80	.85	.88
120	.46	.70	.79	.84	.88
∞	.46	.69	.79	.84	.87
Denominator *df*	6	7	8	9	10
1	1.94	1.98	2.00	2.03	2.04
2	1.28	1.30	1.32	1.33	1.34
3	1.13	1.15	1.16	1.17	1.18
4	1.06	1.08	1.09	1.10	1.11
5	1.02	1.04	1.05	1.06	1.07
6	1.00	1.02	1.03	1.04	1.05
7	.98	1.00	1.01	1.02	1.03
8	.97	.99	1.00	1.01	1.02
9	.96	.98	.99	1.00	1.01
10	.95	.97	.98	.99	1.00
12	.94	.96	.97	.98	.99
15	.93	.95	.96	.97	.98
20	.92	.94	.95	.96	.97
24	.92	.93	.94	.95	.96
30	.91	.93	.94	.95	.96
60	.90	.92	.93	.94	.94
120	.90	.91	.92	.93	.94
∞	.89	.91	.92	.93	.93

Table 8 Critical Values of the *F* Distribution for Evaluating Cook's Distance ($\alpha = .50$) (*Continued*)

	Numerator *df*				
Denominator *df*	12	15	20	30	120
1	2.07	2.09	2.12	2.15	2.18
2	1.36	1.38	1.39	1.41	1.43
3	1.20	1.21	1.23	1.24	1.26
4	1.13	1.14	1.15	1.16	1.18
5	1.09	1.10	1.11	1.12	1.14
6	1.06	1.07	1.08	1.10	1.12
7	1.04	1.05	1.07	1.08	1.10
8	1.03	1.04	1.05	1.07	1.08
9	1.02	1.03	1.04	1.05	1.07
10	1.01	1.02	1.03	1.05	1.06
12	1.00	1.01	1.02	1.03	1.04
15	.99	1.00	1.01	1.02	1.04
20	.98	.99	1.00	1.01	1.03
24	.97	.98	.99	1.01	1.02
30	.97	.98	.99	1.00	1.02
60	.96	.97	.98	.99	1.01
120	.95	.96	.97	.98	1.00
∞	.94	.96	.97	.98	.99

Table 9a Critical Values of the *F* Distribution ($\alpha = .05$)

	Numerator *df*				
Denominator *df*	1	2	3	4	5
1	161	200	216	225	230
2	18.5	19.0	19.2	19.2	19.3
3	10.1	9.55	9.28	9.12	9.01
4	7.71	6.94	6.59	6.39	6.26
5	6.61	5.79	5.41	5.19	5.05
6	5.99	5.14	4.76	4.53	4.39
7	5.59	4.74	4.35	4.12	3.97
8	5.32	4.46	4.07	3.84	3.69
9	5.12	4.26	3.86	3.63	3.48
10	4.96	4.10	3.71	3.48	3.33
11	4.84	3.98	3.59	3.36	3.20
12	4.75	3.89	3.49	3.26	3.11
13	4.67	3.81	3.41	3.18	3.03
14	4.60	3.74	3.34	3.11	2.96
15	4.54	3.68	3.29	3.06	2.90
16	4.49	3.63	3.24	3.01	2.85
17	4.45	3.59	3.20	2.96	2.81
18	4.41	3.55	3.16	2.93	2.77
19	4.38	3.52	3.13	2.90	2.74
20	4.35	3.49	3.10	2.87	2.71
21	4.32	3.47	3.07	2.84	2.68
22	4.30	3.44	3.05	2.82	2.66
23	4.28	3.42	3.03	2.80	2.64
24	4.26	3.40	3.01	2.78	2.62
25	4.24	3.39	2.99	2.76	2.60
30	4.17	3.32	2.92	2.69	2.53
60	4.00	3.15	2.76	2.53	2.37
120	3.92	3.07	2.68	2.45	2.29
∞	3.84	3.00	2.60	2.37	2.21
Denominator *df*	6	7	8	9	10
1	234	237	239	241	242
2	19.3	19.4	19.4	19.4	19.4
3	8.94	8.89	8.85	8.81	8.79
4	6.16	6.09	6.04	6.00	5.96
5	4.95	4.88	4.82	4.77	4.74
6	4.28	4.21	4.15	4.10	4.06
7	3.87	3.79	3.73	3.68	3.64
8	3.58	3.50	3.44	3.39	3.35
9	3.37	3.29	3.23	3.18	3.14
10	3.22	3.14	3.07	3.02	2.98
11	3.09	3.01	2.95	2.90	2.85

Table 9a Critical Values of the *F* Distribution ($\alpha = .05$) (*Continued*)

	Numerator *df*				
Denominator *df*	6	7	8	9	10
12	3.00	2.91	2.85	2.80	2.75
13	2.92	2.83	2.77	2.71	2.67
14	2.85	2.76	2.70	2.65	2.60
15	2.79	2.71	2.64	2.59	2.54
16	2.74	2.66	2.59	2.54	2.49
17	2.70	2.61	2.55	2.49	2.45
18	2.66	2.58	2.51	2.46	2.41
19	2.63	2.54	2.48	2.42	2.38
20	2.60	2.51	2.45	2.39	2.35
21	2.57	2.49	2.42	2.37	2.32
22	2.55	2.46	2.40	2.34	2.30
23	2.53	2.44	2.37	2.32	2.27
24	2.51	2.42	2.36	2.30	2.25
25	2.49	2.40	2.34	2.28	2.24
30	2.42	2.33	2.27	2.21	2.16
60	2.25	2.17	2.10	2.04	1.99
120	2.18	2.09	2.02	1.96	1.91
∞	2.10	2.01	1.94	1.88	1.83

Denominator *df*	12	15	20	30
1	244	246	248	250
2	19.4	19.4	19.4	19.5
3	8.74	8.70	8.66	8.62
4	5.91	5.86	5.80	5.75
5	4.68	4.62	4.56	4.50
6	4.00	3.94	3.87	3.81
7	3.57	3.51	3.44	3.38
8	3.28	3.22	3.15	3.08
9	3.07	3.01	2.94	2.86
10	2.91	2.84	2.77	2.70
11	2.79	2.72	2.65	2.57
12	2.69	2.62	2.54	2.47
13	2.60	2.53	2.46	2.38
14	2.53	2.46	2.39	2.31
15	2.48	2.40	2.33	2.25
16	2.42	2.35	2.28	2.19
17	2.38	2.31	2.23	2.15
18	2.34	2.27	2.19	2.11
19	2.31	2.23	2.16	2.07
20	2.28	2.20	2.12	2.04
21	2.25	2.18	2.10	2.01

(*Continued*)

Table 9a Critical Values of the *F* Distribution ($\alpha = .05$) (*Continued*)

	Numerator *df*			
Denominator *df*	12	15	20	30
22	2.23	2.15	2.07	1.98
23	2.20	2.13	2.05	1.96
24	2.18	2.11	2.03	1.94
25	2.16	2.09	2.01	1.92
30	2.09	2.01	1.93	1.84
60	1.92	1.84	1.75	1.65
120	1.83	1.75	1.66	1.55
∞	1.75	1.67	1.57	1.46

Table 9b Critical Values of the *F* Distribution ($\alpha = .01$)

	Numerator *df*				
Denominator *df*	1	2	3	4	5
1	4052	5000	5403	5625	5764
2	98.5	99.0	99.2	99.2	99.3
3	34.1	30.8	29.5	28.7	28.2
4	21.2	18.0	16.7	16.0	15.5
5	16.3	13.3	12.1	11.4	11.0
6	13.7	10.9	9.78	9.15	8.75
7	12.2	9.55	8.45	7.85	7.46
8	11.3	8.65	7.59	7.01	6.63
9	10.6	8.02	6.99	6.42	6.06
10	10.0	7.56	6.55	5.99	5.64
11	9.65	7.21	6.22	5.67	5.32
12	9.33	6.93	5.95	5.41	5.06
13	9.07	6.70	5.74	5.21	4.86
14	8.86	6.51	5.56	5.04	4.69
15	8.68	6.36	5.42	4.89	4.56
16	8.53	6.23	5.29	4.77	4.44
17	8.40	6.11	5.18	4.67	4.34
18	8.29	6.01	5.09	4.58	4.25
19	8.18	5.93	5.01	4.50	4.17
20	8.10	5.85	4.94	4.43	4.10
21	8.02	5.78	4.87	4.37	4.04
22	7.95	5.72	4.82	4.31	3.99
23	7.88	5.66	4.76	4.26	3.94
24	7.82	5.61	4.72	4.22	3.90
25	7.77	5.57	4.68	4.18	3.85
30	7.56	5.39	4.51	4.02	3.70
60	7.08	4.98	4.13	3.65	3.34
120	6.85	4.79	3.95	3.48	3.17
∞	6.63	4.61	3.78	3.32	3.02

Table 9b Critical Values of the *F* Distribution ($\alpha = .01$) (*Continued*)

	Numerator *df*				
Denominator *df*	6	7	8	9	10
1	5859	5928	5981	6022	6056
2	99.3	99.4	99.4	99.4	99.4
3	27.9	27.7	27.5	27.3	27.2
4	15.2	15.0	14.8	14.7	14.5
5	10.7	10.5	10.3	10.2	10.1
6	8.47	8.26	8.10	7.98	7.87
7	7.19	6.99	6.84	6.72	6.62
8	6.37	6.18	6.03	5.91	5.81
9	5.80	5.61	5.47	5.35	5.26
10	5.39	5.20	5.06	4.94	4.85
11	5.07	4.89	4.74	4.63	4.54
12	4.82	4.64	4.50	4.39	4.30
13	4.62	4.44	4.30	4.19	4.10
14	4.46	4.28	4.14	4.03	3.94
15	4.32	4.14	4.00	3.89	3.80
16	4.20	4.03	3.89	3.78	3.69
17	4.10	3.93	3.79	3.68	3.59
18	4.01	3.84	3.71	3.60	3.51
19	3.94	3.77	3.63	3.52	3.43
20	3.87	3.70	3.56	3.46	3.37
21	3.81	3.64	3.51	3.40	3.31
22	3.76	3.59	3.45	3.35	3.26
23	3.71	3.54	3.41	3.30	3.21
24	3.67	3.50	3.36	3.26	3.17
25	3.63	3.46	3.32	3.22	3.13
30	3.47	3.30	3.17	3.07	2.98
60	3.12	2.95	2.82	2.72	2.63
120	2.96	2.79	2.66	2.56	2.47
∞	2.80	2.64	2.51	2.41	2.32

Denominator *df*	12	15	20	30
1	6106	6157	6209	6261
2	99.4	99.4	99.4	99.5
3	27.1	26.9	26.7	26.5
4	14.4	14.2	14.0	13.8
5	9.89	9.72	9.55	9.38
6	7.72	7.56	7.40	7.23
7	6.47	6.31	6.16	5.99
8	5.67	5.52	5.36	5.20
9	5.11	4.96	4.81	4.65
10	4.71	4.56	4.41	4.25

(*Continued*)

Table 9b Critical Values of the *F* Distribution ($\alpha = .01$) (*Continued*)

	Numerator *df*			
Denominator *df*	12	15	20	30
11	4.40	4.25	4.10	3.94
12	4.16	4.01	3.86	3.70
13	3.96	3.82	3.66	3.51
14	3.80	3.66	3.51	3.35
15	3.67	3.52	3.37	3.21
16	3.55	3.41	3.26	3.10
17	3.46	3.31	3.16	3.00
18	3.37	3.23	3.08	2.92
19	3.30	3.15	3.00	2.84
20	3.23	3.09	2.94	2.78
21	3.17	3.03	2.88	2.72
22	3.12	2.98	2.83	2.67
23	3.07	2.93	2.78	2.62
24	3.03	2.89	2.74	2.58
25	2.99	2.85	2.70	2.54
30	2.84	2.70	2.55	2.39
60	2.50	2.35	2.20	2.03
120	2.34	2.19	2.03	1.86
∞	2.18	2.04	1.88	1.70

References

Abebe, A., McKean, J. W., and Huitema, B. E. (2008). *Robust Propensity Score Analysis for Causal Inference in Observational Studies*. Salt Lake City, UT: American Statistical Association.

Abebe, A., McKean, J. W., and Kloke, J. D. (In preparation). Iterated Reweighted Rank- Based Estimates for GEE Models.

Abelson, R. P., and Tukey, J. W. (1963). Efficient utilization of non-numerical information in quantitative analysis: General theory and the case of simple order. *Annals of Mathematical Statistics*, *34*, 1347–1369.

American Psychological Association (2001). *Publication Manual for the American Psychological Association* (5th edn.). Washington, DC: Author.

Atiqullah, M. (1964). The robustness of the covariance analysis of a one-way classification. *Biometrika*, *51*, 365–373.

Awosoga, O. A. (2009). Meta-analyses of multiple baseline time-series design intervention models for dependent and independent series. Unpublished Doctoral Dissertation. Kalamazoo: Western Michigan University.

Bancroft, T. A. (1964). Analysis and inference for incompletely specified models involving the use of preliminary tests of significance. *Biometrics*, *20*, 427–439.

Barlow, D., Nock, M., and Hersen, M. (2009). *Single Case Experimental Designs: Strategies for Studying Behavior for Change*. Boston: Pearson Allyn and Bacon.

Bathke, A., and Brunner, E. (2003). A nonparametric alternative to analysis of covariance. In M. G. Akritas and D. N. Politis (Eds.), *Recent Advances and Trends in Nonparametric Statistics* (pp. 109–120). Amsterdam: Elsevier.

Beach, M. L., and Meier, P. (1989). Choosing covariates in the analysis of clinical trials. *Controlled Clinical Trials*, *10*, 161S–175S.

Begg, C. B. (1990). Significance of tests of covariate imbalance. *Controlled Clinical Trials*, *11*, 223–225.

Benson, K., and Hartz, A. J. (2000). A comparison of observational studies and randomized controlled trials: Special articles. *New England Journal of Medicine*, *342*, 1878–1886.

The Analysis of Covariance and Alternatives: Statistical Methods for Experiments, Quasi-Experiments, and Single-Case Studies, Second Edition. Bradley E. Huitema.

Berger, V. W. (2005a). Quantifying the magnitude of baseline imbalances resulting from selection bias in randomized clinical trials. *Biometrical Journal*, *47*, 119–127.

Berger, V. W. (2005b). *Selection Bias and Covariate Imbalances in Randomized Clinical Trials*. Chichester, West Sussex, England: Wiley.

Berger, V. W., and Exner, D. V. (1999). Detecting selection bias in randomized clinical trials. *Controlled Clinical Trials*, *20*, 319–327.

Berger, V. W., and Weinstein, S. (2004). Ensuring the comparability of comparison groups: Is randomization enough? *Controlled Clinical Trials*, *25*, 515–524.

Bernard, C. (1865). *An Introduction to the Study of Experimental Medicine*. First English translation by Henry Copley Greene, 1927; reprinted in 1949. London: Macmillan & Co., Ltd.

Borich, G. D., Godbout, R. D., and Wunderlich, K. W. (1976). *The Analysis of Aptitude-Treatment Interactions: Computer Programs and Applications*. San Francisco: Jossey-Bass.

Borich, G. D., and Wunderlich, K. W. (1973). A note on some statistical considerations for using Johnson-Neyman regions of significance. *Annual meeting of the American Psychological Association*. Montreal, Quebec.

Box, G. E. P., Jenkins, G. M., and Reinsel, G. C. (2008). *Time Series Analysis: Forecasting and Control* (4th edn.). Hoboken, NJ: Wiley.

Box, G. E. P., and Tiao, G. C. (1965). Achange in level of a nonstationary time series. *Biometrika*, *52*, 181–192.

Box, G. E. P., and Tiao, G. C. (1975). Intervention analysis with applications to economic and environmental problems. *Journal of the American Statistical Association*, *70*, 70–79.

Bradley, R. A., and Strivastava, S. S. (1979). Correlation in polynomial regression. *American Statistician*, *33*, 11–14.

Browne, R. H. (2010a). The *t*-test *p* value and its relationship to the effect size and $P(X > Y)$. *American Statistician*, *64*, 30–33.

Browne, R. H. (2010b). Correction: The *t*-test *p* value and its relationship to the effect size and $P(X > Y)$. *American Statistician*, *64*, 195.

Bryant, J. L., and Brunvold, N. T. (1980). Multiple comparison procedures in the ANCOVA. *Journal of the American Statistical Association*, *75*, 874–880.

Bryant, J. L., and Paulson, A. S. (1976). An extension of Tukey's method of multiple comparisons to experimental designs with random concomitant variables. *Biometrika*, *63*, 631–638.

Budescu, D. V. (1980). A note on polynomial regression. *Multivariate Behavioral Research*, *15*, 497–506.

Burk, D. (1980). Cancer mortality linked with artificial fluoridation in Birmingham, England. Paper presented at the *4th International Symposium on the Prevention and Detection of Cancer*. Wembley, UK.

Cahen, L., and Linn, R. L. (1971). Regions of significant criterion differences in aptitude-treatment-interaction research. *American Educational Research Journal*, *8*, 521–530.

Campbell, D. T., and Kenny, D. A. (1999). *A Primer on Regression Artifacts*. New York: Guilford Press.

Campbell, D. T., and Stanley, J. (1966). *Experimental and Quasi-Experimental Designs for Research*. Chicago: Rand McNally.

Carroll, R. J., Ruppert, D., Stefanski, L. A., and Crainiceanu, C. M. (2006). *Measurement Error in Nonlinear Models: A Modern Perspective* (2nd edn.). Boca Raton, FL: Chapman & Hall/CRC.

Chen, S., and Cox, C. (1992). Use of baseline data for estimation of treatment effects in the presence of regression to the mean. *Biometrics*, *48*, 593–598.

Chen, S., Cox, C., and Cui, L. (1998). A more flexible regression-to-the-mean model with possible stratification. *Biometrics*, *54*, 939–947.

Cobb, G. W. (1998). *Introduction to Design and Analysis of Experiments*. New York: Springer.

Cochran, W. G. (1968). Errors of measurement in statistics. *Technometrics*, *10*, 637–666.

Cohen, A. C. (1955). Restriction and selection in samples from bivarate normal distributions. *Journal of the American Statistical Association*, *50*, 884–893.

Cohen, J. (1988). *Statistical Power Analysis for the Behavioral Sciences* (2nd edn.). Hillsdale, NJ: Lawrence Erlbaum.

Cohen, P., Cohen, J., West, S., and Aiken, L. S. (2003). *Applied Multiple Regression/Correlation Analysis for the Behavioral Sciences* (3rd edn.). Hillsdale, NJ: Lawrence Erlbaum.

Concato, J., Shah, N., and Horwitz, R. (2000). Randomized controlled trials, observational studies, and the hierarchy of research designs. *New England Journal of Medicine*, *342*, 1887–1892.

Conover, W. J., and Inman, R. L. (1982). Analysis of covariance using the rank transformation. *Biometrics*, *38*, 715–724.

Cook-Mozaffare, P., Bulusu, L., and Doll, R. (1981). Fluoridation of water supplies and cancer mortality. I. A search for an effect in the U.K. on risk of death from cancer. *Journal of Epidemiology and Community Health*, *35*, 227–232.

Cook-Mozaffare, P., and Doll, R. (1981). Fluoridation of water supplies and cancer mortality. II. Mortality trends after fluoridation. *Journal of Epidemiology and Community Health*, *35*, 233–238.

Cooper, J. O., Heron, T. E., and Heward, W. L. (2006). *Applied Behavior Analysis* (2nd edn.). Englewood Cliffs, NJ: Prentice Hall.

Cramer, E. M., and Appelbaum, M. I. (1978). The validity of polynomial regression in the random regression model. *Review of Educational Research*, *48*, 511–515.

DeGracie, J. S., and Fuller, W. A. (1972). Estimation of the slope and analysis of covariance when the concomitant variable is measured with error. *Journal of the American Statistical Association*, *67*, 930–937.

Dehejia, R. H., and Wahba, S. (2002). Propensity score-matching methods for nonexperimental causal studies. *The Review of Economics and Statistics*, *841*, 151–161.

Dixon, S. L., and McKean, J. W. (1996). Rank based analysis of the heteroscedastic linear model. *Journal of the American Statistical Association*, *91*, 699–712.

Dmitrienko, A., Tamhane, A. C., and Bretz, F. (2010). *Multiple Testing Problems in Pharmaceutical Statistics*. New York: CRC Press.

Drake, C. (1993). Effects of misspecification of the propensity score on estimators of treatment effect. *Biometrics*, *49*, 1231–1336.

Dunn, O. J. (1961). Multiple comparisons among means. *Journal of the American Statistical Association*, *56*, 52–64.

Durbin, J., and Watson, G. S. (1950). Testing for serial correlation in least squares regression: I. *Biometrika*, *37*, 409–428.

Durbin, J., and Watson, G. S. (1951). Testing for serial correlation in least squares regression: II. *Biometrika*, *38*, 159–178.

Dyer, O. (2003). GMC reprimands doctor for research fraud. *British Medical Journal*, *326*, 730.

Dyer, K., Schwartz, I. S., and Luce, S. C. (1984). A supervision program for increasing functional activities for severely handicapped students in a residential setting. *Journal of Applied Behavior Analysis*, *17*, 249–259.

Efron, B., and Tibshirani, R. J. (1993). *An Introduction to the Bootstrap*. New York: Chapman & Hall.

Elashoff, J. D. (1969). Analysis of covariance: A delicate instrument. *American Educational Research Journal*, *6*, 383–401.

Enders, W. (2010). *Applied Econometric Time Series* (3rd edn.) Hoboken, NJ: Wiley.

Erlander, S., and Gustavsson, J. (1965). Simultaneous confidence regions in normal regression analysis with an application to road accidents. *Review of the International Statistical Institute*, *33*, 364–377.

Fante, R. M., Dickinson, A. M., and Huitema, B. E. (Submitted). A comparison of three training methods on the acquisition and retention of automotive product knowledge.

Farthing, M. J. G. (2004). "Publish and be damned" . . . the road to research misconduct. *Journal of the Royal College of Physicians of Edinburgh*, *34*, 301–304.

Forsythe, A. B. (1977). Post-hoc decision to use a covariate. *Journal of Chronic Disease*, *30*, 61–64.

Fredericks, H. D. (1969). A comparison of Doman-Delacato method and a behavior modification method upon the coordination of mongoloids. Unpublished Doctoral Dissertation. University of Oregon, Eugene.

Fuller, W. A. (1987). *Measurement Error Models*. New York: Wiley.

Fuller, W. A., and Hidiroglou, M. A. (1978). Regression estimation after correction for attenuation. *Journal of the American Statistical Association*, *73*, 99–104.

Gelman, A., and Hill, J. (2007). *Data Analysis Using Regression and Multilevel/Hierarchical Models*. New York: Cambridge University Press.

Glass, G. V., Willson, V. L., and Gottman, J. (1975). *Design and Analysis of Time-Series Experiments*. Boulder, CO: Colorado Associated University Press.

Gong, G., and Samaniego, F. J. (1981). Pseudo maximum likelihood estimation: Theory and applications. *Annals of Statistics*, *9*, 861–869.

Gould, S. J. (1966). Allometry and size in ontogeny and physiology. *Biological Reviews*, *41*, 587–640.

Greenhouse, G. R. (2003). The growth and future of biostatistics. *Statistics in Medicine*, *22*, 3323–3335.

Grice, G. R., and Hunter, J. J. (1964). Stimulus intensity effects depend upon the type of experimental design. *Psychological Review*, *71*, 247–256.

Guyatt, G. H., Heyting, A., Jaeschke, R., Keller, J., Adachi, J. D., and Roberts, R. S. (1990). N of 1 randomized trials for investigating new drugs. *Controlled Clinical Trials*, *11*, 88–100.

Hamilton, B. L. (1976). A Monte Carlo test of the robustness of parametric and nonparametric analysis of covariance against unequal regression slopes. *Journal of the American Statistical Association*, *71*, 864–869.

Hayes, A. F., and Matthes, J. (2009). Computational procedures for probing interactions in OLS and logistic regression: SPSS and SAS implementations. *Behavior Research Methods*, *41*, 924–936.

Hayes, R. J., and Moulton, L. H. (2009). *Cluster Randomized Trials*. New York: Chapman & Hall.

Hettmansperger, T. J., and McKean, J. W. (1998). *Robust Nonparametric Statistical Methods*. London: Arnold.

Hill, J. (2008). Comment. *Journal of the American Statistical Association*, *103*, 1346–1350.

Hill, J., and Reiter, J. P. (2006). Interval estimation for treatment effects using propensity score matching. *Statistics in Medicine*, *25*, 2230–2256.

Hochberg, Y., and Varon-Salomon, Y. (1984). On simultaneous pairwise comparisons in analysis of covariance. *Journal of the American Statistical Association*, *79*, 863–866.

Hollingsworth, H. H. (1980). An analytic investigation of the effects of heterogeneous regression slopes in analysis of covariance. *Educational and Psychological Measurement*, *40*, 611–618.

Hosmer, D. W., and Lemeshow, S. (2000). *Applied Logistic Regression* (2nd edn.). New York: Wiley.

Howell, D. (2010). *Statistical Methods for Psychology* (7th edn.). New York: Wadsworth.

Huitema, B. E. (1980). *The Analysis of Covariance and Alternatives*. New York: Wiley.

Huitema, B. E. (1985). Autocorrelation in applied behavior analysis: A myth. *Behavioral Assessment*, *7*, 109–120.

Huitema, B. E. (1986a). Autocorrelation in behavioral research: Wherefore Art Thou? In A. Poling and R. W. Fuqua (Eds.), *Research Methods in Applied Behavior Analysis: Issues and Advances* (pp. 187–208). New York: Plenum Press.

Huitema, B. E. (1986b). Statistical analysis and single-subject designs: Some misunderstandings. In A. Poling and R. W. Fuqua (Eds.), *Research Methods in Applied Behavior Analysis: Issues and Advances* (pp. 209–232). New York: Plenum.

Huitema, B. E. (1988). Autocorrelation: 10 years of confusion. *Behavioral Assessment*, *10*, 253–294.

Huitema, B. E. (2004). Analysis of interrupted time-series experiments using ITSE: A critique. *Understanding Statistics: Statistical Issues in Psychology, Education, and the Social Sciences*, *3*, 27–46.

Huitema, B. E. (2008). A four phase time-series intervention model with parameters measuring immediate and delayed effects. Unpublished manuscript. Kalamazoo: Western Michigan University.

Huitema, B. E. (2009). Reversed ordinal logistic regression: A method for the analysis of experiments with ordered treatment levels. Unpublished manuscript. Kalamazoo: Western Michigan University.

Huitema, B. E., McKean, J. W., and McKnight, S. (1994, June). New methods of intervention analysis: Simple and Complex. Presented at the meeting of the *Association for Behavior Analysis*. Atlanta, GA.

Huitema, B. E., and McKean, J. W. (1998). Irrelevant autocorrelation in least-squares intervention models. *Psychological Methods*, *3*, 104–116.

Huitema, B. E., and McKean, J. W. (2000a). A simple and powerful test for autocorrelated errors in OLS intervention models. *Psychological Reports*, *87*, 3–20.

Huitema, B. E., and McKean, J. W. (2000b). Design specification issues in time-series intervention models. *Educational and Psychological Measurement*, *60*, 38–58.

Huitema, B. E., and McKean, J. W. (2005). Propensity score methodology combined with modified ANCOVA. Paper presented at the *6th International Conference on Health Policy Research: Methodological Issues in Health Services and Outcomes Research*. Boston, MA.

Huitema, B. E., and McKean, J. W. (2007a). Identifying autocorrelation generated by various error processes in interrupted time-series regression designs. *Educational and Psychological Measurement*, *67*, 447–459.

Huitema, B. E., and McKean, J. W. (2007b). An improved portmanteau test for autocorrelated errors in interrupted time-series regression. *Behavior Research Methods, Instruments, & Computers*, *39*, 343–349.

Huitema, B. E., McKean, J. W., and Laraway, S. (2008). Time-series intervention analysis using ITSACORR: Fatal flaws. *Journal of Modern Applied Statistical Methods*, *6*, 367–379.

Huitema, B. E., McKean, J. W., and Laraway, S. (In press). Erratum for: Time-series intervention analysis using ITSACORR: Fatal flaws. *Journal of Modern Applied Statistical Methods*.

Hunka, S. (1994). Using Mathematica to solve Johnson-Neyman problems. *Mathematica in Education*, *3*, 32–36.

Hunka, S. (1995). Identifying regions of significance in ANCOVA problems having nonhomogeneous regressions. *British Journal of Mathematical and Statistical Psychology*, *48*, 161–188.

Hunka, S., and Leighton, J. (1997). Defining Johnson-Neyman Regions of significance in the three-covariate ANCOVA using Mathematica. *Journal of Educational and Behavioral Statistics*, *22*, 361–387.

Imbens, G., and Lemieux, T. (2008). Regression discontinuity design: A guide to practice. *Journal of Econometrics*, *142*, 615–635.

Imbens, G., and Rubin, D. B. (In preparation.) *Causal Inference in Statistics, and in the Social and Biomedical Sciences*. New York: Cambridge University Press.

Ioannidis, J., Haidich, A., and Pappa, M., et al. (2001). Comparison of evidence of treatment effects in randomized and nonrandomized studies. *Journal of the American Medical Association*, *286*, 821–830.

Johnson, P. O., and Neyman, J. (1936). Tests of certain linear hypotheses and their application to some educational problems. *Statistical Research Memoirs*, *1*, 57–93.

Johnston, J. M., and Pennypacker, H. S. (2008). *Strategies and Tactics of Behavioral Research* (3rd edn.). New York: Routledge.

Jonckheere, A. R. (1954). A distribution free *k*-sample test against ordered alternatives. *Biometrika*, *41*, 133–145.

Jones, R. R., Vaught, R. S., and Weinrott, M. R. (1977). Time-series analysis in operant research. *Journal of Applied behavior Analysis*, *10*, 151–167.

Karpman, M. B. (1980). ANCOVA—a one covariate Johnson-Neyman algorithm. *Educational and Psychological Measurement*, *40*, 791–793.

Karpman, M. B. (1983). The Johnson-Neyman tecnmique using SPSS or BMDP. *Educational and Psychological Measurement*, *43*, 137–147.

Karpman, M. B. (1986). Comparing two non-parallel regression lines with the parametric alternative to analysis of covariance using SPSS-X or SAS—The Johnson- Neyman technique. *Educational and Psychological Measurement*, *46*, 639–644.

Kazdin, A. E. (1982). *Single-Case Research Designs: Methods for Clinical and Applied Settings*. New York: Oxford University Press.

Keppel, G., and Wickens, T. D. (2004). *Design and Analysis: a Researcher's Handbook* (4th edn.). Upper Saddle River, NJ: Pearson Prentice-Hall.

Kirk, R. (1995). *Experimental Design* (3rd edn.). Pacific Grove, CA: Brooks/Cole.

Kocher, A. T. (1974). An investigation of the effects of non-homogeneous within-group regression coefficients upon the F test of analysis of covariance. Paper presented at the *annual meeting of the American Educational Research Association*, Chicago.

Koehler, M. J., and Levin, J. R. (2000). RegRand: Statistical software for the multiple-baseline design. *Behavior Research Methods Instruments, & Computers*, *32*, 367–371.

Kosten, S. F. (2010). Robust interval estimation of a treatment effect in observational studies using propensity score matching. Unpublished Doctoral Dissertation. Kalamazoo: Western Michigan University.

Kosten, S. F., McKean, J. W., and Huitema, B. E. (Submitted). Robust and nonrobust interval estimation of treatment effects in observational studies designed using propensity score matching. Manuscript submitted for publication.

Kramer, C. Y. (1956). Extensions of multiple range tests to group means with unequal numbers of replications. *Biometrics*, *12*, 307–310.

Kreft, I., and de Leeuw, J. (1998). *Introducing Multilevel Modeling*. London: Sage.

Lautenschlager, G. J. (1987). JOHN-NEY: An interactive program for computing the Johnson-Neyman confidence region for nonsignificant prediction differences. *Applied Psychological Measurement*, *11*, 194–195.

Lehmacher, W., Wassmer, G., and Reitmeir, P. (1991). Procedures for two-sample comparisons with multiple endpoints controlling the experimentwise error rate. *Biometrics*, *47*, 511–521.

Lehmann, E. L., and D'Abrera, H. J. M. (1975). *Nonparametrics: Statistical Methods Based On Ranks*. San Francisco: Holden-Day.

Levy, K. (1980). A Monte Carlo study of analysis of covariance under violations of the assumptions of normality and equal regression slopes. *Educational and Psychological Measurement*, *40*, 835–840.

Little, R. J., Long, Q., and Lin, X. (2008). Comment. *Journal of the American Statistical Association*, *103*, 1344–1346.

Long, J. S. (1997). *Regression Models for Categorical and Limited Dependent Variables*. Thousand Oaks, CA: Sage Publications.

Lord, F. M. (1960). Large-sample covariance analysis when the control variable is fallible. *Journal of the American Statistical Association*, *55*, 307–321.

Lundervold, D. A., and Belwood, M. F. (2000). The best kept secret in counseling: Single-case ($N = 1$) experimental designs. *Journal of Counseling & Development*, *78*, 92–102.

Maddox, J., Randi, J., and Stewart, W. W. (1988). "High-dilution" experiments a delusion. *Nature*, *334*, 287–290.

Madsen, L. G., and Bytzer, P. (2002). Single subject trials as a research instrument in gastrointestinal pharmacology. *Alimentary Pharmacology & Therapeutics*, *16*, 189–196.

Manly, B. F. J. (1992). *The Design and Analysis of Research Studies.* London: Cambridge University Press.

Maxwell, S. E., and Delaney, H. D. (2004). *Designing Experiments and Analyzing Data: A Model Comparison Perspective* (2nd edn.). Mahwah, NJ: Lawrence Erlbaum.

Maxwell, S. E., Delaney, H. D., and Manheimer, J. M. (1985). ANOVA of residuals and ANCOVA: Correcting an illusion by using model comparisons and graphs. *Journal of Educational Statistics*, *10*, 197–209.

Mayo, J., White, O., and Eysenck, H. J. (1978). An empirical study of the relation between astrological factors and personality. *Journal of Social Psychology*, *105*, 229–236.

McKean, J. W. (2004). Robust analysis of linear models. *Statistical Science*, *19*, 562–570.

McKean, J. W., Naranjo, J., and Huitema, B. E. (2001). A robust method for the analysis of experiments with ordered treatment levels. *Psychological Reports*, *89*, 267–273.

McKean, J. W., Naranjo, J., and Sheather, S. J. (1999). Diagnostics for comparing robust and least squares fits. *Journal of Nonparametric Statistics*, *11*, 161–188.

McKean, J. W., and Sheather, S. J. (1991). Small sample properties of robust analyses of liner models based on R-estimates: A Survey. In W. Stahel and S. Weisberg (Eds.), *Directions in Robust Statistics and Diagnostics*, Part II (pp. 1–19). New York: Springer-Verlag.

McKean, J. W., Sheather, S. J., and Hettmansperger, T. P. (1993). The use and interpretation of residuals based on robust estimation. *Journal of the American Statistical Association*, *88*, 1254–1263.

McKean, J. W., and Vidmar, T. J. (1994). A comparison of two rank-based methods for the analysis of linear models. *American Statistician*, *48*, 220–229.

McKnight, S., McKean, J. W., and Huitema, B. E. (2000). A double bootstrap method to analyze linear models with autoregressive error terms. *Psychological Methods*, *5*, 87–101.

Mee, R. W., and Chau, T. C. (1991). Regression toward the mean and the paired sample *t* test. *The American Statistician*, *45*, 39–41.

Mendro, R. L. (1975). A Monte Carlo Study of the Robustness of the Johnson-Neyman Technique. *Annual meeting of the American Educational Research Association*, Washington, DC.

Methot, L. L. (1995). Autocorrelation in single-subject data: A meta-analytic view. Unpublished Doctoral Dissertation. Kalamazoo: Western Michigan University.

Milliken, G. A., and Johnson, D. E. (2002). *Analysis of Messy Data, Volume III: Analysis of Covariance*. New York: Chapman & Hall/CRC.

Montgomery, D. C., Jennings, C. L., and Kulahci, M. (2008). *Introduction to Time Series Analysis and Forecasting*. Hoboken, NJ: Wiley.

Morgan, D. L., and Morgan, R. K. (2001). Single Participant Research Design: Bringing science to managed care. *American Psychologist*, *56*, 119–127.

Morgan, O. W., Griffiths, C., and Majeed, A. (2007). Interrupted time-series analysis of regulations to reduce paracetamol (acetaminophen) poisoning. *Public Library of Science (PLoS) Medicine*, *4*, e105.

Naranjo, J. D., and McKean, J. W. (2001). Adjusting for regression effect in uncontrolled studies. *Biometrics*, *57*, 178–181.

O'Brien, P. C. (1984). Procedures for comparing samples with multiple endpoints. *Biometrics*, *40*, 1079–1087.

Olejnik, S. F., and Algina, J. (1985). A review of nonparametric alternatives to analysis of covariance. *Evaluation Review*, *9*, 51–83.

O'Neill, R. (2006). Patient-reported outcome instruments: Overview and comments on the FDA draft guidance. *42nd Annual meeting of the Drug Information Association*, Philadelphia.

Packard, G. C., and Boardman, T. J. (1988). The misuse of ratios, indices, and percentages in ecophysiological research. *Physiological Zoology*, *61*, 1–9.

Parsonson, B. S., and Baer, D. M. (1986). The graphic analysis of data. In A. Poling and R. W. Fuqua (Eds.), *Research Methods in Applied Behavior Analysis: Issues and Advances* (pp. 157–186). New York: Plenum.

Partridge, L., and Farquhar, M. (1981). Sexual activity reduces lifespan of male fruitflies. *Nature*, *294*, 580–581.

Permutt, T. (1990). Testing for imbalance of covariates in controlled experiments. *Statistics in Medicine*, *9*, 1455–1462.

Pocock, S. J., Geller, N. L., and Tsiatis, A. (1987). The analysis of multiple endpoints in clinical trials. *Biometrics*, *43*, 487–498.

Poling, A. D., Methot, L. L., and LeSage, M. G. (1995). *Fundamentals of Behavior Analytic Research*. New York: Plenum.

Potthoff, R. F. (1964). On the Johnson-Neyman technique and some extensions thereof. *Psychometrika*, *29*, 241–256.

Preacher, K. J., Curran, P. J., and Bauer, D. J. (2006). Computational tools for probing interactions in multiple linear regression, multilevel modeling, and latent curve analysis. *Journal of Educational & Behavioral Statistics*, *31*, 437–448.

Rantz, W. G. (2007). The effects of feedback on the accuracy of completing flight checklists. Unpublished master's thesis. Kalamazoo: Western Michigan University.

Rantz, W. (2009). Comparing the accuracy of performing digital and paper cdhecklists using a feedback package during normal workload conditions in simulated flight. Unpublished doctoral dissertation. Western Michigan University, Kalamazoo.

Raudenbush, S. W., and Bryk, A. S. (2002). *Hierarchical Linear Models: Applications and Data Analysis Methods*. Thousand Oaks, CA: Sage.

Rheinheimer, D. C., and Penfield, D. A. (2001). The effects of type I error rate and power of the ANCOVA *F* test and selected alternatives under nonnormality and variance heterogeneity. *Journal of Experimental Education*, *4*, 373–391.

Rogosa, D. (1977). Some results for the Johnson-Neyman technique. Unpublished Doctoral Dissertation. Stanford, California: Stanford University.

Rogosa, D. (1980). Comparing nonparallel regression lines. *Psychological Bulletin*, *88*, 307–321.

Rogosa, D. (1981). On the relationship between the Johnson-Neyman region of significance and statistical tests of parallel within-group regressions. *Educational and psychological Measurement*, *41*, 73–84.

Rosenbaum, P. R. (2002). *Observational Studies* (2nd edn.). New York: Springer-Verlag.

Rosenbaum, P. R. (2010). *Design of Observational Studies*. New York: Springer.

Rosenbaum, P. R., and Rubin, D. B. (1983). The central role of the propensity score in observational studies for causal effects. *Biometrika*, *70*, 41–55.

Rosenbaum, P. R., and Rubin, D. B. (1984). Reducing bias in observational studies using subclassification on the propensity score. *Journal of the American Statistical Association*, *79*, 516–524.

Rothman, K. J. (1990). No adjustments are needed for multiple comparisons. *Epidemiology*, *1*, 43–46.

Rubin, D. B. (1974). Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of Educational Psychology*, *66*, 688–701.

Rubin, D. B. (1979). Using multivariate matched sampling and regression adjustment to control bias in observational studies. *Journal of the American Statistical Association*, *74*, 318–328.

Rubin, D. B. (1997). Estimating causal effects from large data sets using propensity scores. *Annals of Internal Medicine*, *127*, 757–763.

Rubin, D. B. (2001). Using propensity scores to help design observational studies: Application to the tobacco litigation. *Health Services & Outcomes Research Methodology*, *2*, 169–188.

Rubin, D. B. (2005). Causal inference using potential outcomes: Design, modeling, decisions. *Journal of the American Statistical Association*, *100*, 322–331.

Rubin, D. B. (2006). *Matched Sampling for Causal Effects*. New York: Cambridge University Press.

Rubin, D. B. (2008). Comment. *Journal of the American Statistical Association*, *103*, 1350–1353.

Samuels, M. L. (1991). Statistical reversion toward the mean: more universal than regression toward the mean. *American Statistician*, *45*, 344–346.

Scheffé, H. (1959). *The Analysis of Variance*. New York: Wiley.

Scherrer, M. D., and Wilder, D. A. (2008). Training to increase safe tray carrying among cocktail servers. *Journal of Applied Behavior Analysis*, *41*, 131–135.

Schluchter, M. D., and Forsythe, A. B. (1985). Post-hoc selection of covariates in randomized experiments. *Communication in Statistical Theory and Methods*, *14*, 679–699.

Schmittlein, D. C. (1989). Surprising inferences from unsurprising observations: do conditional expectations really regress to the mean? *American Statistician*, *43*, 176–183.

Schultz, K. F. (1995). Subverting randomization in controlled trials. *Journal of the American Medical Association*, *274*, 1456–1458.

Senn, S. J. (1994). Testing for baseline balance in clinical trials. *Statistics in Medicine*, *13*, 1715–1726.

Senn, S. J. (1995). In defense of analysis of covariance: a reply to Chambless and Roeback [letter comment]. *Statistics in Medicine*, *14*, 2283–2285.

Senn, S. J. (1998). Applying results of randomized trials to patients. *N* of 1 trials are needed. *British Medical Journal*, *317*, 537–538.

Senn, S. J. (2005). Comment. *Biometrical Journal*, *47*, 133–135.

Senn, S. J., and Brown, R. A. (1985). Estimating treatment effects in clinical trials subject to regression to the mean. *Biometrics*, *41*, 555–560.

Shadish, W. R., Clark, M. H., and Steiner, P. M. (2008). Can nonrandomized experiments yield accurate answers? A randomized experiment comparing random and nonrandom assignments. *Journal of the American Statistical Association*, *103*, 1334–1343.

Shadish, W. R., Cook, T. D., and Campbell, D. T. (2002). *Experimental and Quasi-Experimental Designs for Generalized Causal Inference*. New York: Houghton Mifflin.

Shields, J. L. (1978). An empirical investigation of the effects of heteroscedasticity and heterogeneity of variance on the analysis of covariance and the Johnson-Neyman technique. Technical paper No. 292, U.S. Army Research Institute for the Behaavioral and Social Sciences, Alexandria, Virginia.

Shirley, E. A. C. (1981). A distribution-free method for analysis of covariance based on ranked data. *Applied Statistics*, *30*, 158–162.

Shurbutt, J., Van Houten, R., Turner, S., and Huitema, B. E. (2009). An analysis of the effects of LED rectangular rapid flash beacons (RRFB) on yielding to pedestrians using multilane crosswalks. *Transportation Research Record*, *2140*, 85–95.

Sideridis, G. D., and Greenwood, C. R. (1997). Is human behavior autocorrelated? An empirical analysis. *Journal of Behavioral Education*, *7*, 273–293.

Sidman, M. (1960). *Tactics of Scientific Research*. New York: Basic Books.

Skinner, B. F. (1956). A case history in scientific method. *American Psychologist*, *11*, 221–233.

Slowiak, J. M., Huitema, B. E., and Dickinson, A. M. (2008). Reducing wait time in a hospital pharmacy to promote customer service. *Quality Management in Health Care*, *17*, 112–127.

Smith, R. J. (1984). Allometric scaling in comparative biology: problems of concept and method. *American Journal of Physiology*, *246*, 152–160.

Spiegelman, D. (2010). Approaches to uncertainty in exposure assessment in environmental epidemiology. *Annual Review of Public Health*, *31*, 149–163.

Steering Committee of the Physicians' Health Study Research Group. (1988). Preliminary report: Findings from the aspirin component of the ongoing physicians' health study. *New England Journal of Medicine*, *318*, 262–264.

Stevens, J. P. (2009). *Applied Multivariate Statistics for the Social Sciences*. New York: Taylor and Francis.

Stoline, M. R., Huitema, B. E., and Mitchell, B. (1980). Intervention time series model with different pre- and postintervention first order autoregressive parameters. *Psychological Bulletin*, *88*, 46–53.

Stone, R. (1993). The assumptions on which causal inferences rest. *Journal of the Royal Statistical Society*, *55*, 455–466.

Stricker, G., and Trierweiler, S. J. (1995). The local clinical scientist: A bridge between science and practice. *American Psychologist*, *50*, 995–1002.

Sullivan, L. M., and D'Agostino, R. B., Sr. (2002). Robustness and power of analysis of covariance applied to data distorted from normality by floor effects: non-homogeneous regression slopes. *Journal of Statistical Computation and Simulation*, *72*, 141–165.

Tamhane, A. C., and Logan, B. R. (2004). On O'Brien's OLS and GLS tests for multiple endpoints. In Y. Benjamini, F. Bretz, and S. Sarkar (Eds.), *Recent Developments in Multiple Comparison Procedures*, IMS Lecture Notes and Monograph Series, (pp. 76–88). Bethesda, MD: Institute of Mathematical Statisties.

Terpstra, T. J. (1952). The asymptotic normality and consistency of Kendall's test against trend, when ties are present in one ranking. *Indagationes Mathematicae*, *14*, 327–333.

Terpstra, J. T., and McKean, J. W. (2004). Rank-based analysis of liner models using R. Technical Report 151. Statistical Computation Lab, Western Michigan University.

Thigpen, C. C., and Paulson, A. S. (1974). A multiple range test for analysis of covariance. *Biometrika*, *61*, 479–484.

Thorndike, R. L. (1942). Regression fallacies in the matched groups experiment. *Psychometrika*, *7*, 85–102.

Trochim, W. M. K. (1984). *Research Design for Program Evaluation: The Regression-Discontinuity Approach*. Newbury Park, CA: Sage.

Trochim, W. M. K. (1990). The regression-discontinuity design. In L. Sechrest, E. Perrin, and J. Bunker (Eds.), *Research Methodology: Strengthening Causal Interpretations of Nonexperimental Data* (pp. 199–140). Rockville, MD: Public Health Service, Agency for Health Care Policy and Research.

Tryon, P. V., and Hettmansperger, T. P. (1973). A class of non-parametric tests for homogeneity against ordered alternatives. *Annals of Statistics*, *1*, 1061–1070.

Van Den Noortgate, W., and Onghena, P. (2003). Hierarchical linear models for the quantitative integration of effect sizes in single-case research. *Behavior Research Methods, Instruments, & Computers*, *35*, 1–10.

Wang, S., Huitema, B. E., Bruyere, R., Weintrub, K., Megregian, P., and Steinhorn, D. M. (2010). Pain perception and touch healing in healthy adults: A preliminary prospective randomized controlled study. *Journal of Alternative Medicine Research*, *2*, 75–82.

Watcharotone, K. (2010). On robustification of some procedures used in analysis of covariance. Unpublished doctoral dissertation. Kalamazoo: Western Michigan University.

Watcharotone, K., McKean, J. W., and Huitema, B. E. (2010). Robust procedures for heterogeneous regression ANCOVA. Manuscript submitted for publication.

Westfall, P. H., Tobias, R. D., Rom, D., Wolfinger, R. D., and Hochberg, Y. (1999). *Multiple Comparisons and Multiple Tests Using the SAS System*. Cary, NC: SAS Institute.

Wilcox, R. R. (2005). *Introduction to Robust Estimation and Hypothesis Testing*. Burlington, MA: Elsevier Academic Press.

Wu, M-J, Becker, B. J., and Netz, Y. (2007). Effects of physical activity on psychological change in advanced age: a multivariate meta-analysis. *Journal of Modern Applied Statistical Methods*, *6*, 2–7.

Wu, Y. B. (1984). The effects of heterogeneous regression slopes on the robustness of two test statistics in the analysis of covariance. *Educational and Psychological Measurement*, *44*, 647–663.

Wunderlich, K. W., and Borich, G. K. (1974). Curvilinear extensions to Johnson-Neyman regions of significance and some applications to educational research. *Annual meeting of the American Educational Research Association*. Chicago.

Index

The Analysis of Covariance and Alternatives: Statistical Methods for Experiments, Quasi-Experiments, and Single-Case Studies, Second Edition. Bradley E. Huitema.

WILEY SERIES IN PROBABILITY AND STATISTICS

ESTABLISHED BY WALTER A. SHEWHART AND SAMUEL S. WILKS

The ***Wiley Series in Probability and Statistics*** is well established and authoritative. It covers many topics of current research interest in both pure and applied statistics and probability theory. Written by leading statisticians and institutions, the titles span both state-of-the-art developments in the field and classical methods.

Reflecting the wide range of current research in statistics, the series encompasses applied, methodological and theoretical statistics, ranging from applications and new techniques made possible by advances in computerized practice to rigorous treatment of theoretical approaches.

This series provides essential and invaluable reading for all statisticians, whether in academia, industry, government, or research.

† ABRAHAM and LEDOLTER • Statistical Methods for Forecasting
AGRESTI • Analysis of Ordinal Categorical Data, *Second Edition*
AGRESTI • An Introduction to Categorical Data Analysis, *Second Edition*
AGRESTI • Categorical Data Analysis, *Second Edition*
ALTMAN, GILL, and McDONALD • Numerical Issues in Statistical Computing for the Social Scientist
AMARATUNGA and CABRERA • Exploration and Analysis of DNA Microarray and Protein Array Data
ANDĚL • Mathematics of Chance
ANDERSON • An Introduction to Multivariate Statistical Analysis, *Third Edition*
* ANDERSON • The Statistical Analysis of Time Series
ANDERSON, AUQUIER, HAUCK, OAKES, VANDAELE, and WEISBERG • Statistical Methods for Comparative Studies
ANDERSON and LOYNES • The Teaching of Practical Statistics
ARMITAGE and DAVID (editors) • Advances in Biometry
ARNOLD, BALAKRISHNAN, and NAGARAJA • Records
* ARTHANARI and DODGE • Mathematical Programming in Statistics
* BAILEY • The Elements of Stochastic Processes with Applications to the Natural Sciences
BALAKRISHNAN and KOUTRAS • Runs and Scans with Applications
BALAKRISHNAN and NG • Precedence-Type Tests and Applications
BARNETT • Comparative Statistical Inference, *Third Edition*
BARNETT • Environmental Statistics
BARNETT and LEWIS • Outliers in Statistical Data, *Third Edition*
BARTOSZYNSKI and NIEWIADOMSKA-BUGAJ • Probability and Statistical Inference
BASILEVSKY • Statistical Factor Analysis and Related Methods: Theory and Applications
BASU and RIGDON • Statistical Methods for the Reliability of Repairable Systems
BATES and WATTS • Nonlinear Regression Analysis and Its Applications

*Now available in a lower priced paperback edition in the Wiley Classics Library.
†Now available in a lower priced paperback edition in the Wiley–Interscience Paperback Series.

BECHHOFER, SANTNER, and GOLDSMAN • Design and Analysis of Experiments for Statistical Selection, Screening, and Multiple Comparisons

BEIRLANT, GOEGEBEUR, SEGERS, TEUGELS, and DE WAAL • Statistics of Extremes: Theory and Applications

BELSLEY • Conditioning Diagnostics: Collinearity and Weak Data in Regression

† BELSLEY, KUH, and WELSCH • Regression Diagnostics: Identifying Influential Data and Sources of Collinearity

BENDAT and PIERSOL • Random Data: Analysis and Measurement Procedures, *Fourth Edition*

BERNARDO and SMITH • Bayesian Theory

BERRY, CHALONER, and GEWEKE • Bayesian Analysis in Statistics and Econometrics: Essays in Honor of Arnold Zellner

BHAT and MILLER • Elements of Applied Stochastic Processes, *Third Edition*

BHATTACHARYA and WAYMIRE • Stochastic Processes with Applications

BIEMER, GROVES, LYBERG, MATHIOWETZ, and SUDMAN • Measurement Errors in Surveys

BILLINGSLEY • Convergence of Probability Measures, *Second Edition*

BILLINGSLEY • Probability and Measure, *Third Edition*

BIRKES and DODGE • Alternative Methods of Regression

BISGAARD and KULAHCI • Time Series Analysis and Forecasting by Example

BISWAS, DATTA, FINE, and SEGAL • Statistical Advances in the Biomedical Sciences: Clinical Trials, Epidemiology, Survival Analysis, and Bioinformatics

BLISCHKE AND MURTHY (editors) • Case Studies in Reliability and Maintenance

BLISCHKE AND MURTHY • Reliability: Modeling, Prediction, and Optimization

BLOOMFIELD • Fourier Analysis of Time Series: An Introduction, *Second Edition*

BOLLEN • Structural Equations with Latent Variables

BOLLEN and CURRAN • Latent Curve Models: A Structural Equation Perspective

BOROVKOV • Ergodicity and Stability of Stochastic Processes

BOSQ and BLANKE • Inference and Prediction in Large Dimensions

BOULEAU • Numerical Methods for Stochastic Processes

BOX • Bayesian Inference in Statistical Analysis

BOX • Improving Almost Anything, *Revised Edition*

BOX • R. A. Fisher, the Life of a Scientist

BOX and DRAPER • Empirical Model-Building and Response Surfaces

* BOX and DRAPER • Evolutionary Operation: A Statistical Method for Process Improvement

BOX and DRAPER • Response Surfaces, Mixtures, and Ridge Analyses, *Second Edition*

BOX, HUNTER, and HUNTER • Statistics for Experimenters: Design, Innovation, and Discovery, *Second Editon*

BOX, JENKINS, and REINSEL • Time Series Analysis: Forcasting and Control, *Fourth Edition*

BOX, LUCEÑO, and PANIAGUA-QUIÑONES • Statistical Control by Monitoring and Adjustment, *Second Edition*

BRANDIMARTE • Numerical Methods in Finance: A MATLAB-Based Introduction

† BROWN and HOLLANDER • Statistics: A Biomedical Introduction

BRUNNER, DOMHOF, and LANGER • Nonparametric Analysis of Longitudinal Data in Factorial Experiments

BUCKLEW • Large Deviation Techniques in Decision, Simulation, and Estimation

CAIROLI and DALANG • Sequential Stochastic Optimization

CASTILLO, HADI, BALAKRISHNAN, and SARABIA • Extreme Value and Related Models with Applications in Engineering and Science

CHAN • Time Series: Applications to Finance with R and S-Plus®, *Second Edition*

CHARALAMBIDES • Combinatorial Methods in Discrete Distributions

CHATTERJEE and HADI • Regression Analysis by Example, *Fourth Edition*

*Now available in a lower priced paperback edition in the Wiley Classics Library.

†Now available in a lower priced paperback edition in the Wiley–Interscience Paperback Series.

CHATTERJEE and HADI • Sensitivity Analysis in Linear Regression
CHERNICK • Bootstrap Methods: A Guide for Practitioners and Researchers, *Second Edition*
CHERNICK and FRIIS • Introductory Biostatistics for the Health Sciences
CHILÈS and DELFINER • Geostatistics: Modeling Spatial Uncertainty
CHOW and LIU • Design and Analysis of Clinical Trials: Concepts and Methodologies, *Second Edition*
CLARKE • Linear Models: The Theory and Application of Analysis of Variance
CLARKE and DISNEY • Probability and Random Processes: A First Course with Applications, *Second Edition*
* COCHRAN and COX • Experimental Designs, *Second Edition*
COLLINS and LANZA • Latent Class and Latent Transition Analysis: With Applications in the Social, Behavioral, and Health Sciences
CONGDON • Applied Bayesian Modelling
CONGDON • Bayesian Models for Categorical Data
CONGDON • Bayesian Statistical Modelling, *Second Edition*
CONOVER • Practical Nonparametric Statistics, *Third Edition*
COOK • Regression Graphics
COOK and WEISBERG • An Introduction to Regression Graphics
COOK and WEISBERG • Applied Regression Including Computing and Graphics
CORNELL • A Primer on Experiments with Mixtures
CORNELL • Experiments with Mixtures, Designs, Models, and the Analysis of Mixture Data, *Third Edition*
COVER and THOMAS • Elements of Information Theory
COX • A Handbook of Introductory Statistical Methods
* COX • Planning of Experiments
CRESSIE • Statistics for Spatial Data, *Revised Edition*
CRESSIE and WIKLE • Statistics for Spatio-Temporal Data
CSÖRGŐ and HORVÁTH • Limit Theorems in Change Point Analysis
DANIEL • Applications of Statistics to Industrial Experimentation
DANIEL • Biostatistics: A Foundation for Analysis in the Health Sciences, *Eighth Edition*
* DANIEL • Fitting Equations to Data: Computer Analysis of Multifactor Data, *Second Edition*
DASU and JOHNSON • Exploratory Data Mining and Data Cleaning
DAVID and NAGARAJA • Order Statistics, *Third Edition*
* DEGROOT, FIENBERG, and KADANE • Statistics and the Law
DEL CASTILLO • Statistical Process Adjustment for Quality Control
DEMARIS • Regression with Social Data: Modeling Continuous and Limited Response Variables
DEMIDENKO • Mixed Models: Theory and Applications
DENISON, HOLMES, MALLICK and SMITH • Bayesian Methods for Nonlinear Classification and Regression
DETTE and STUDDEN • The Theory of Canonical Moments with Applications in Statistics, Probability, and Analysis
DEY and MUKERJEE • Fractional Factorial Plans
DILLON and GOLDSTEIN • Multivariate Analysis: Methods and Applications
DODGE • Alternative Methods of Regression
* DODGE and ROMIG • Sampling Inspection Tables, *Second Edition*
* DOOB • Stochastic Processes
DOWDY, WEARDEN, and CHILKO • Statistics for Research, *Third Edition*
DRAPER and SMITH • Applied Regression Analysis, *Third Edition*
DRYDEN and MARDIA • Statistical Shape Analysis
DUDEWICZ and MISHRA • Modern Mathematical Statistics

*Now available in a lower priced paperback edition in the Wiley Classics Library.
†Now available in a lower priced paperback edition in the Wiley–Interscience Paperback Series.

DUNN and CLARK • Basic Statistics: A Primer for the Biomedical Sciences, *Third Edition*
DUPUIS and ELLIS • A Weak Convergence Approach to the Theory of Large Deviations
EDLER and KITSOS • Recent Advances in Quantitative Methods in Cancer and Human Health Risk Assessment
* ELANDT-JOHNSON and JOHNSON • Survival Models and Data Analysis
ENDERS • Applied Econometric Time Series
† ETHIER and KURTZ • Markov Processes: Characterization and Convergence
EVANS, HASTINGS, and PEACOCK • Statistical Distributions, *Third Edition*
EVERITT • Cluster Analysis, *Fifth Edition*
FELLER • An Introduction to Probability Theory and Its Applications, Volume I, *Third Edition,* Revised; Volume II, *Second Edition*
FISHER and VAN BELLE • Biostatistics: A Methodology for the Health Sciences
FITZMAURICE, LAIRD, and WARE • Applied Longitudinal Analysis, *Second Edition*
* FLEISS • The Design and Analysis of Clinical Experiments
FLEISS • Statistical Methods for Rates and Proportions, *Third Edition*
† FLEMING and HARRINGTON • Counting Processes and Survival Analysis
FUJIKOSHI, ULYANOV, and SHIMIZU • Multivariate Statistics: High-Dimensional and Large-Sample Approximations
FULLER • Introduction to Statistical Time Series, *Second Edition*
† FULLER • Measurement Error Models
GALLANT • Nonlinear Statistical Models
GEISSER • Modes of Parametric Statistical Inference
GELMAN and MENG • Applied Bayesian Modeling and Causal Inference from Incomplete-Data Perspectives
GEWEKE • Contemporary Bayesian Econometrics and Statistics
GHOSH, MUKHOPADHYAY, and SEN • Sequential Estimation
GIESBRECHT and GUMPERTZ • Planning, Construction, and Statistical Analysis of Comparative Experiments
GIFI • Nonlinear Multivariate Analysis
GIVENS and HOETING • Computational Statistics
GLASSERMAN and YAO • Monotone Structure in Discrete-Event Systems
GNANADESIKAN • Methods for Statistical Data Analysis of Multivariate Observations, *Second Edition*
GOLDSTEIN • Multilevel Statistical Models, *Fourth Edition*
GOLDSTEIN and LEWIS • Assessment: Problems, Development, and Statistical Issues
GOLDSTEIN and WOOFF • Bayes Linear Statistics
GREENWOOD and NIKULIN • A Guide to Chi-Squared Testing
GROSS, SHORTLE, THOMPSON, and HARRIS • Fundamentals of Queueing Theory, *Fourth Edition*
GROSS, SHORTLE, THOMPSON, and HARRIS • Solutions Manual to Accompany Fundamentals of Queueing Theory, *Fourth Edition*
* HAHN and SHAPIRO • Statistical Models in Engineering
HAHN and MEEKER • Statistical Intervals: A Guide for Practitioners
HALD • A History of Probability and Statistics and their Applications Before 1750
HALD • A History of Mathematical Statistics from 1750 to 1930
† HAMPEL • Robust Statistics: The Approach Based on Influence Functions
HANNAN and DEISTLER • The Statistical Theory of Linear Systems
HARMAN and KULKARNI • An Elementary Introduction to Statistical Learning Theory
HARTUNG, KNAPP, and SINHA • Statistical Meta-Analysis with Applications
HEIBERGER • Computation for the Analysis of Designed Experiments
HEDAYAT and SINHA • Design and Inference in Finite Population Sampling

*Now available in a lower priced paperback edition in the Wiley Classics Library.
†Now available in a lower priced paperback edition in the Wiley–Interscience Paperback Series.

HEDEKER and GIBBONS • Longitudinal Data Analysis
HELLER • MACSYMA for Statisticians
HERITIER, CANTONI, COPT, and VICTORIA-FESER • Robust Methods in Biostatistics
HINKELMANN and KEMPTHORNE • Design and Analysis of Experiments, Volume 1: Introduction to Experimental Design, *Second Edition*
HINKELMANN and KEMPTHORNE • Design and Analysis of Experiments, Volume 2: Advanced Experimental Design
HOAGLIN, MOSTELLER, and TUKEY • Fundamentals of Exploratory Analysis of Variance
* HOAGLIN, MOSTELLER, and TUKEY • Exploring Data Tables, Trends and Shapes
* HOAGLIN, MOSTELLER, and TUKEY • Understanding Robust and Exploratory Data Analysis
HOCHBERG and TAMHANE • Multiple Comparison Procedures
HOCKING • Methods and Applications of Linear Models: Regression and the Analysis of Variance, *Second Edition*
HOEL • Introduction to Mathematical Statistics, *Fifth Edition*
HOGG and KLUGMAN • Loss Distributions
HOLLANDER and WOLFE • Nonparametric Statistical Methods, *Second Edition*
HOSMER and LEMESHOW • Applied Logistic Regression, *Second Edition*
HOSMER, LEMESHOW, and MAY • Applied Survival Analysis: Regression Modeling of Time-to-Event Data, *Second Edition*
HUBER • Data Analysis: What Can Be Learned From the Past 50 Years
HUBER • Robust Statistics
† HUBER and RONCHETTI • Robust Statistics, *Second Edition*
HUBERTY • Applied Discriminant Analysis, *Second Edition*
HUBERTY and OLEJNIK • Applied MANOVA and Discriminant Analysis, *Second Edition*
HUITEMA • The Analysis of Covariance and Alternatives: Statistical Methods for Experiments, Quasi-Experiments, and Single-Case Studies, *Second Edition*
HUNT and KENNEDY • Financial Derivatives in Theory and Practice, *Revised Edition*
HURD and MIAMEE • Periodically Correlated Random Sequences: Spectral Theory and Practice
HUSKOVA, BERAN, and DUPAC • Collected Works of Jaroslav Hajek—with Commentary
HUZURBAZAR • Flowgraph Models for Multistate Time-to-Event Data
IMAN and CONOVER • A Modern Approach to Statistics
JACKMAN • Bayesian Analysis for the Social Sciences
† JACKSON • A User's Guide to Principle Components
JOHN • Statistical Methods in Engineering and Quality Assurance
JOHNSON • Multivariate Statistical Simulation
JOHNSON and BALAKRISHNAN • Advances in the Theory and Practice of Statistics: A Volume in Honor of Samuel Kotz
JOHNSON and BHATTACHARYYA • Statistics: Principles and Methods, *Fifth Edition*
JOHNSON, KEMP, and KOTZ • Univariate Discrete Distributions, *Third Edition*
JOHNSON and KOTZ • Distributions in Statistics
JOHNSON and KOTZ (editors) • Leading Personalities in Statistical Sciences: From the Seventeenth Century to the Present
JOHNSON, KOTZ, and BALAKRISHNAN • Continuous Univariate Distributions,Volume 1, *Second Edition*
JOHNSON, KOTZ, and BALAKRISHNAN • Continuous Univariate Distributions, Volume 2, *Second Edition*
JOHNSON, KOTZ, and BALAKRISHNAN • Discrete Multivariate Distributions
JUDGE, GRIFFITHS, HILL, LÜTKEPOHL, and LEE • The Theory and Practice of Econometrics, *Second Edition*
JUREČKOVÁ and SEN • Robust Statistical Procedures: Aymptotics and Interrelations
JUREK and MASON • Operator-Limit Distributions in Probability Theory

*Now available in a lower priced paperback edition in the Wiley Classics Library.
†Now available in a lower priced paperback edition in the Wiley–Interscience Paperback Series.

KADANE • Bayesian Methods and Ethics in a Clinical Trial Design
KADANE AND SCHUM • A Probabilistic Analysis of the Sacco and Vanzetti Evidence
KALBFLEISCH and PRENTICE • The Statistical Analysis of Failure Time Data, *Second Edition*
KARIYA and KURATA • Generalized Least Squares
KASS and VOS • Geometrical Foundations of Asymptotic Inference
† KAUFMAN and ROUSSEEUW • Finding Groups in Data: An Introduction to Cluster Analysis
KEDEM and FOKIANOS • Regression Models for Time Series Analysis
KENDALL, BARDEN, CARNE, and LE • Shape and Shape Theory
KHURI • Advanced Calculus with Applications in Statistics, *Second Edition*
KHURI, MATHEW, and SINHA • Statistical Tests for Mixed Linear Models
* KISH • Statistical Design for Research
KLEIBER and KOTZ • Statistical Size Distributions in Economics and Actuarial Sciences
KLEMELÄ • Smoothing of Multivariate Data: Density Estimation and Visualization
KLUGMAN, PANJER, and WILLMOT • Loss Models: From Data to Decisions, *Third Edition*
KLUGMAN, PANJER, and WILLMOT • Solutions Manual to Accompany Loss Models: From Data to Decisions, *Third Edition*
KOSKI and NOBLE • Bayesian Networks: An Introduction
KOTZ, BALAKRISHNAN, and JOHNSON • Continuous Multivariate Distributions, Volume 1, *Second Edition*
KOTZ and JOHNSON (editors) • Encyclopedia of Statistical Sciences: Volumes 1 to 9 with Index
KOTZ and JOHNSON (editors) • Encyclopedia of Statistical Sciences: Supplement Volume
KOTZ, READ, and BANKS (editors) • Encyclopedia of Statistical Sciences: Update Volume 1
KOTZ, READ, and BANKS (editors) • Encyclopedia of Statistical Sciences: Update Volume 2
KOVALENKO, KUZNETZOV, and PEGG • Mathematical Theory of Reliability of Time-Dependent Systems with Practical Applications
KOWALSKI and TU • Modern Applied U-Statistics
KRISHNAMOORTHY and MATHEW • Statistical Tolerance Regions: Theory, Applications, and Computation
KROESE, TAIMRE, and BOTEV • Handbook of Monte Carlo Methods
KROONENBERG • Applied Multiway Data Analysis
KULINSKAYA, MORGENTHALER, and STAUDTE • Meta Analysis: A Guide to Calibrating and Combining Statistical Evidence
KUROWICKA and COOKE • Uncertainty Analysis with High Dimensional Dependence Modelling
KVAM and VIDAKOVIC • Nonparametric Statistics with Applications to Science and Engineering
LACHIN • Biostatistical Methods: The Assessment of Relative Risks, *Second Edition*
LAD • Operational Subjective Statistical Methods: A Mathematical, Philosophical, and Historical Introduction
LAMPERTI • Probability: A Survey of the Mathematical Theory, *Second Edition*
LANGE, RYAN, BILLARD, BRILLINGER, CONQUEST, and GREENHOUSE • Case Studies in Biometry
LARSON • Introduction to Probability Theory and Statistical Inference, *Third Edition*
LAWLESS • Statistical Models and Methods for Lifetime Data, *Second Edition*
LAWSON • Statistical Methods in Spatial Epidemiology, *Second Edition*
LE • Applied Categorical Data Analysis
LE • Applied Survival Analysis
LEE • Structural Equation Modeling: A Bayesian Approach
LEE and WANG • Statistical Methods for Survival Data Analysis, *Third Edition*
LEPAGE and BILLARD • Exploring the Limits of Bootstrap
LEYLAND and GOLDSTEIN (editors) • Multilevel Modelling of Health Statistics
LIAO • Statistical Group Comparison
LINDVALL • Lectures on the Coupling Method

*Now available in a lower priced paperback edition in the Wiley Classics Library.
†Now available in a lower priced paperback edition in the Wiley–Interscience Paperback Series.

LIN • Introductory Stochastic Analysis for Finance and Insurance
LINHART and ZUCCHINI • Model Selection
LITTLE and RUBIN • Statistical Analysis with Missing Data, *Second Edition*
LLOYD • The Statistical Analysis of Categorical Data
LOWEN and TEICH • Fractal-Based Point Processes
MAGNUS and NEUDECKER • Matrix Differential Calculus with Applications in Statistics and Econometrics, *Revised Edition*
MALLER and ZHOU • Survival Analysis with Long Term Survivors
MALLOWS • Design, Data, and Analysis by Some Friends of Cuthbert Daniel
MANN, SCHAFER, and SINGPURWALLA • Methods for Statistical Analysis of Reliability and Life Data
MANTON, WOODBURY, and TOLLEY • Statistical Applications Using Fuzzy Sets
MARCHETTE • Random Graphs for Statistical Pattern Recognition
MARDIA and JUPP • Directional Statistics
MARKOVICH • Nonparametric Analysis of Univariate Heavy-Tailed Data: Research and Practice
MARONNA, MARTIN and YOHAI • Robust Statistics: Theory and Methods
MASON, GUNST, and HESS • Statistical Design and Analysis of Experiments with Applications to Engineering and Science, *Second Edition*
McCULLOCH, SEARLE, and NEUHAUS • Generalized, Linear, and Mixed Models, *Second Edition*
McFADDEN • Management of Data in Clinical Trials, *Second Edition*
* McLACHLAN • Discriminant Analysis and Statistical Pattern Recognition
McLACHLAN, DO, and AMBROISE • Analyzing Microarray Gene Expression Data
McLACHLAN and KRISHNAN • The EM Algorithm and Extensions, *Second Edition*
McLACHLAN and PEEL • Finite Mixture Models
McNEIL • Epidemiological Research Methods
MEEKER and ESCOBAR • Statistical Methods for Reliability Data
MEERSCHAERT and SCHEFFLER • Limit Distributions for Sums of Independent Random Vectors: Heavy Tails in Theory and Practice
MENGERSEN, ROBERT, and TITTERINGTON • Mixtures: Estimation and Applications
MICKEY, DUNN, and CLARK • Applied Statistics: Analysis of Variance and Regression, *Third Edition*
* MILLER • Survival Analysis, *Second Edition*
MONTGOMERY, JENNINGS, and KULAHCI • Introduction to Time Series Analysis and Forecasting
MONTGOMERY, PECK, and VINING • Introduction to Linear Regression Analysis, *Fourth Edition*
MORGENTHALER and TUKEY • Configural Polysampling: A Route to Practical Robustness
MUIRHEAD • Aspects of Multivariate Statistical Theory
MULLER and STOYAN • Comparison Methods for Stochastic Models and Risks
MURRAY • X-STAT 2.0 Statistical Experimentation, Design Data Analysis, and Nonlinear Optimization
MURTHY, XIE, and JIANG • Weibull Models
MYERS, MONTGOMERY, and ANDERSON-COOK • Response Surface Methodology: Process and Product Optimization Using Designed Experiments, *Third Edition*
MYERS, MONTGOMERY, VINING, and ROBINSON • Generalized Linear Models. With Applications in Engineering and the Sciences, *Second Edition*
† NELSON • Accelerated Testing, Statistical Models, Test Plans, and Data Analyses
† NELSON • Applied Life Data Analysis
NEWMAN • Biostatistical Methods in Epidemiology
OCHI • Applied Probability and Stochastic Processes in Engineering and Physical Sciences

*Now available in a lower priced paperback edition in the Wiley Classics Library.
†Now available in a lower priced paperback edition in the Wiley–Interscience Paperback Series.

OKABE, BOOTS, SUGIHARA, and CHIU • Spatial Tesselations: Concepts and Applications of Voronoi Diagrams, *Second Edition*
OLIVER and SMITH • Influence Diagrams, Belief Nets and Decision Analysis
PALTA • Quantitative Methods in Population Health: Extensions of Ordinary Regressions
PANJER • Operational Risk: Modeling and Analytics
PANKRATZ • Forecasting with Dynamic Regression Models
PANKRATZ • Forecasting with Univariate Box-Jenkins Models: Concepts and Cases
PARDOUX • Markov Processes and Applications: Algorithms, Networks, Genome and Finance
* PARZEN • Modern Probability Theory and Its Applications
PEÑA, TIAO, and TSAY • A Course in Time Series Analysis
PIANTADOSI • Clinical Trials: A Methodologic Perspective
PORT • Theoretical Probability for Applications
POURAHMADI • Foundations of Time Series Analysis and Prediction Theory
POWELL • Approximate Dynamic Programming: Solving the Curses of Dimensionality, *Second Edition*
PRESS • Bayesian Statistics: Principles, Models, and Applications
PRESS • Subjective and Objective Bayesian Statistics, *Second Edition*
PRESS and TANUR • The Subjectivity of Scientists and the Bayesian Approach
PUKELSHEIM • Optimal Experimental Design
PURI, VILAPLANA, and WERTZ • New Perspectives in Theoretical and Applied Statistics
† PUTERMAN • Markov Decision Processes: Discrete Stochastic Dynamic Programming
QIU • Image Processing and Jump Regression Analysis
* RAO • Linear Statistical Inference and Its Applications, *Second Edition*
RAO • Statistical Inference for Fractional Diffusion Processes
RAUSAND and HØYLAND • System Reliability Theory: Models, Statistical Methods, and Applications, *Second Edition*
RAYNER • Smooth Tests of Goodnes of Fit: Using R, *Second Edition*
RENCHER • Linear Models in Statistics
RENCHER • Methods of Multivariate Analysis, *Second Edition*
RENCHER • Multivariate Statistical Inference with Applications
* RIPLEY • Spatial Statistics
* RIPLEY • Stochastic Simulation
ROBINSON • Practical Strategies for Experimenting
ROHATGI and SALEH • An Introduction to Probability and Statistics, *Second Edition*
ROLSKI, SCHMIDLI, SCHMIDT, and TEUGELS • Stochastic Processes for Insurance and Finance
ROSENBERGER and LACHIN • Randomization in Clinical Trials: Theory and Practice
ROSS • Introduction to Probability and Statistics for Engineers and Scientists
ROSSI, ALLENBY, and MCCULLOCH • Bayesian Statistics and Marketing
† ROUSSEEUW and LEROY • Robust Regression and Outlier Detection
ROYSTON and SAUERBREI • Multivariate Model Building: A Pragmatic Approach to Regression Analysis Based on Fractional Polynomials for Modeling Continuous Variables
* RUBIN • Multiple Imputation for Nonresponse in Surveys
RUBINSTEIN and KROESE • Simulation and the Monte Carlo Method, *Second Edition*
RUBINSTEIN and MELAMED • Modern Simulation and Modeling
RYAN • Modern Engineering Statistics
RYAN • Modern Experimental Design
RYAN • Modern Regression Methods, *Second Edition*
RYAN • Statistical Methods for Quality Improvement, *Third Edition*
SALEH • Theory of Preliminary Test and Stein-Type Estimation with Applications
SALTELLI, CHAN, and SCOTT (editors) • Sensitivity Analysis
* SCHEFFE • The Analysis of Variance

*Now available in a lower priced paperback edition in the Wiley Classics Library.
†Now available in a lower priced paperback edition in the Wiley–Interscience Paperback Series.

SCHIMEK • Smoothing and Regression: Approaches, Computation, and Application
SCHOTT • Matrix Analysis for Statistics, *Second Edition*
SCHOUTENS • Levy Processes in Finance: Pricing Financial Derivatives
SCHUSS • Theory and Applications of Stochastic Differential Equations
SCOTT • Multivariate Density Estimation: Theory, Practice, and Visualization
* SEARLE • Linear Models
† SEARLE • Linear Models for Unbalanced Data
† SEARLE • Matrix Algebra Useful for Statistics
† SEARLE, CASELLA, and McCULLOCH • Variance Components
SEARLE and WILLETT • Matrix Algebra for Applied Economics
SEBER • A Matrix Handbook For Statisticians
† SEBER • Multivariate Observations
SEBER and LEE • Linear Regression Analysis, *Second Edition*
† SEBER and WILD • Nonlinear Regression
SENNOTT • Stochastic Dynamic Programming and the Control of Queueing Systems
* SERFLING • Approximation Theorems of Mathematical Statistics
SHAFER and VOVK • Probability and Finance: It's Only a Game!
SHERMAN • Spatial Statistics and Spatio-Temporal Data: Covariance Functions and Directional Properties
SILVAPULLE and SEN • Constrained Statistical Inference: Inequality, Order, and Shape Restrictions
SINGPURWALLA • Reliability and Risk: A Bayesian Perspective
SMALL and MCLEISH • Hilbert Space Methods in Probability and Statistical Inference
SRIVASTAVA • Methods of Multivariate Statistics
STAPLETON • Linear Statistical Models, *Second Edition*
STAPLETON • Models for Probability and Statistical Inference: Theory and Applications
STAUDTE and SHEATHER • Robust Estimation and Testing
STOYAN, KENDALL, and MECKE • Stochastic Geometry and Its Applications, *Second Edition*
STOYAN and STOYAN • Fractals, Random Shapes and Point Fields: Methods of Geometrical Statistics
STREET and BURGESS • The Construction of Optimal Stated Choice Experiments: Theory and Methods
STYAN • The Collected Papers of T. W. Anderson: 1943–1985
SUTTON, ABRAMS, JONES, SHELDON, and SONG • Methods for Meta-Analysis in Medical Research
TAKEZAWA • Introduction to Nonparametric Regression
TAMHANE • Statistical Analysis of Designed Experiments: Theory and Applications
TANAKA • Time Series Analysis: Nonstationary and Noninvertible Distribution Theory
THOMPSON • Empirical Model Building
THOMPSON • Sampling, *Second Edition*
THOMPSON • Simulation: A Modeler's Approach
THOMPSON and SEBER • Adaptive Sampling
THOMPSON, WILLIAMS, and FINDLAY • Models for Investors in Real World Markets
TIAO, BISGAARD, HILL, PEÑA, and STIGLER (editors) • Box on Quality and Discovery: with Design, Control, and Robustness
TIERNEY • LISP-STAT: An Object-Oriented Environment for Statistical Computing and Dynamic Graphics
TSAY • Analysis of Financial Time Series, *Third Edition*
UPTON and FINGLETON • Spatial Data Analysis by Example, Volume II: Categorical and Directional Data
† VAN BELLE • Statistical Rules of Thumb, *Second Edition*

*Now available in a lower priced paperback edition in the Wiley Classics Library.
†Now available in a lower priced paperback edition in the Wiley–Interscience Paperback Series.

VAN BELLE, FISHER, HEAGERTY, and LUMLEY • Biostatistics: A Methodology for the Health Sciences, *Second Edition*

VESTRUP • The Theory of Measures and Integration

VIDAKOVIC • Statistical Modeling by Wavelets

VINOD and REAGLE • Preparing for the Worst: Incorporating Downside Risk in Stock Market Investments

WALLER and GOTWAY • Applied Spatial Statistics for Public Health Data

WEERAHANDI • Generalized Inference in Repeated Measures: Exact Methods in MANOVA and Mixed Models

WEISBERG • Applied Linear Regression, *Third Edition*

WEISBERG • Bias and Causation: Models and Judgment for Valid Comparisons

WELSH • Aspects of Statistical Inference

WESTFALL and YOUNG • Resampling-Based Multiple Testing: Examples and Methods for *p*-Value Adjustment

* WHITTAKER • Graphical Models in Applied Multivariate Statistics

WINKER • Optimization Heuristics in Economics: Applications of Threshold Accepting

WONNACOTT and WONNACOTT • Econometrics, *Second Edition*

WOODING • Planning Pharmaceutical Clinical Trials: Basic Statistical Principles

WOODWORTH • Biostatistics: A Bayesian Introduction

WOOLSON and CLARKE • Statistical Methods for the Analysis of Biomedical Data, *Second Edition*

WU and HAMADA • Experiments: Planning, Analysis, and Parameter Design Optimization, *Second Edition*

WU and ZHANG • Nonparametric Regression Methods for Longitudinal Data Analysis

YANG • The Construction Theory of Denumerable Markov Processes

YOUNG, VALERO-MORA, and FRIENDLY • Visual Statistics: Seeing Data with Dynamic Interactive Graphics

ZACKS • Stage-Wise Adaptive Designs

* ZELLNER • An Introduction to Bayesian Inference in Econometrics

ZELTERMAN • Discrete Distributions—Applications in the Health Sciences

ZHOU, OBUCHOWSKI, and MCCLISH • Statistical Methods in Diagnostic Medicine, *Second Edition*

*Now available in a lower priced paperback edition in the Wiley Classics Library.

†Now available in a lower priced paperback edition in the Wiley–Interscience Paperback Series.