Data Mining Using SAS Applications

George Fernandez

CHAPMAN & HALL/CRC
A CRC Press Company
Boca Raton London New York Washington, D.C.

Library of Congress Cataloging-in-Publication Data

Fernandez, George, 1952-
Data mining using SAS applications / George Fernandez.
p. cm.
Includes bibliographical references and index.
ISBN 1-58488-345-6 (alk. paper)
1. Commercial statistics--Computer programs. 2. SAS (Computer file) I. Title.

HF1017 .F476 2002
005.3′042--dc21 2002034917

This book contains information obtained from authentic and highly regarded sources. Reprinted material is quoted with permission, and sources are indicated. A wide variety of references are listed. Reasonable efforts have been made to publish reliable data and information, but the authors and the publisher cannot assume responsibility for the validity of all materials or for the consequences of their use.

Direct all inquiries to CRC Press LLC, 2000 N.W. Corporate Blvd., Boca Raton, Florida 33431.

Trademark Notice: Product or corporate names may be trademarks or registered trademarks, and are used only for identification and explanation, without intent to infringe.

International Standard Book Number 1-58488-345-6
Library of Congress Card Number 2002034917
Printed in the United States of America 1 2 3 4 5 6 7 8 9 0
Printed on acid-free paper

Preface

Objective

The objective of this book is to introduce data mining concepts, describe methods in data mining from sampling to decision trees, demonstrate the features of user-friendly data mining SAS tools, and, above all, allow readers to download data mining SAS macro-call files and help them perform complete data mining. The user-friendly SAS macro approach integrates the statistical and graphical analysis tools available in SAS systems and offers complete data mining solutions without writing SAS program codes or using the point-and-click approach. Step-by-step instructions for using SAS macros and interpreting the results are provided in each chapter. Thus, by following the step-by-step instructions and downloading the user-friendly SAS macros described in the book, data analysts can perform complete data mining analysis quickly and effectively.

Why Use SAS Software?

SAS Institute, the industry leader in analytical and decision support solutions, offers a comprehensive data mining solution that allows users to explore large quantities of data and discover relationships and patterns that lead to intelligent decision making. *Enterprise Miner*, SAS Institute's data mining software, offers an integrated environment for businesses that need to conduct comprehensive data mining. SAS provides additional data mining capabilities such as neural networks, memory-based reasoning, and association/sequence discovery that are not presented in this book. These additional features can be obtained through *Enterprise Miner*.

Including complete SAS codes in this book for performing comprehensive data mining solutions would not be very effective because a majority of business and statistical analysts are not experienced SAS programmers. Quick results from data mining are not feasible, as many hours of modifying code and debugging program errors are required when analysts are required to work with SAS program codes. An alternative to the point-and-click menu interface modules and the high-priced SAS *Enterprise Miner* is the user-friendly SAS macro applications for performing several data mining tasks that are included in this book. This macro approach integrates statistical and graphical tools available in SAS systems and provides user-friendly data analysis tools that allow data analysts to complete data mining tasks quickly, without writing SAS programs, by running the SAS macros in the background.

Coverage

The following types of analyses can be performed using the user-friendly SAS macros:

- Converting PC databases to SAS data
- Sampling techniques to create training and validation samples
- Exploratory graphical techniques
  - Univariate analysis of continuous response
  - Frequency data analysis for categorical data
- Unsupervised learning
  - Principal component
  - Factor and cluster analysis
  - *k*-mean cluster analysis
  - Bi-plot display
- Supervised learning: prediction
  - Multiple regression models
    - Partial and VIF plots, plots for checking data and model problems
    - Lift charts
    - Scoring
    - Model validation techniques
  - Logistic regression
    - Partial delta logit plots, ROC curves false positive/negative plots
    - Lift charts
    - Model validation techniques
- Supervised learning: classification
  - Discriminant analysis
    - Canonical discriminant analysis — bi-plots
    - Parametric discriminant analysis
    - Nonparametric discriminant analysis
    - Model validation techniques
  - CHAID — decisions tree methods
    - Model validation techniques

Why Do I Believe the Book Is Needed?

During the last decade, there has been an explosion in the field of data warehousing and data mining for knowledge discovery. The challenge of understanding data has led to the development of a new data mining tool. Data mining books that are currently available mainly address data mining principles but provide no instructions and explanations to carry out a data mining project. Also, many data analysts are interested in expanding their expertise in the field of data mining and are looking for "how-to" books on data mining that do not require expensive software such as *Enterprise Miner*. Business school instructors are currently incorporating data mining into their MBA curriculum and are looking for "how-to" books on data mining using available software. This book on data mining using SAS macro applications easily fills the gap and complements the existing data mining book market.

Key Features of the Book

- *No SAS programming experience is required.* This essential "how-to" guide is especially suitable for data analysts to practice data mining techniques for knowledge discovery. Thirteen user-friendly SAS macros to perform data mining are described, and instructions are given in regard to downloading the macro-call file and running the macros from the website that has been set up for this book. No experience in modifying SAS macros or programming with SAS is needed to run these macros.
- *Complete analysis can be performed in less than 10 minutes.* Complete predictive modeling, including data exploration, model fitting, assumption checks, validation, and scoring new data, can be performed on SAS datasets in less than 10 minutes.
- *Expensive SAS Enterprise Miner is not required.* The user-friendly macros work with the standard SAS modules: BASE, STAT, GRAPH, and IML. No additional SAS modules are required.
- *No experience in SAS ODS is required.* Options are included in the SAS macros for saving data mining output and graphics in RTF, HTML, and PDF format using the new ODS features of SAS.
- *More than 100 figures are included.* These data mining techniques stress the use of visualization for a thorough study of the structure of data and to check the validity of statistical models fitted to data. These figures allow readers to visualize the trends and patterns present in their databases.

Textbook or a Supplementary Lab Guide

This book is suitable for adoption as a textbook for a statistical methods course in data mining and data analysis. This book provides instructions and tools for performing complete exploratory statistical method, regression analysis, multivariate methods, and classification analysis quickly. Thus, this book is ideal for graduate-level statistical methods courses that use SAS software. Some examples of potential courses include:

- Advanced business statistics
- Research methods
- Advanced data analysis

Potential Audience

- This book is suitable for data analysts who need to apply data mining techniques using existing SAS modules for successful data mining, without investing a lot of time to research and buy new software products or to learn how to use additional software.
- Experienced SAS programmers can utilize the SAS macro source codes available in the companion CD-ROM and customize it to fit in their business goals and different computing environments.
- Graduate students in business and the natural and social sciences can successfully complete data analysis projects quickly using these SAS macros.
- Large business enterprises can use data mining SAS macros in pilot studies involving the feasibility of conducting a successful data mining endeavor, before making a significant investment in full-scale data mining.
- Finally, any SAS users who want to impress their supervisors can do so with quick and complete data analysis presented in PDF, RTF, or HTML formats.

Additional Resources

- *Book website:* A website has been set up at

  http://www.ag.unr.edu/gf/dm.html

  Users can find information regarding downloading the sample data files used in the book and the necessary SAS macro-call files. Readers are encouraged to visit this site for information on any errors in the book, SAS macro updates, and links for additional resources.
- *Companion CD-ROM:* For experienced SAS programmers, a companion CD-ROM is available for purchase that contains sample datasets, macro-call

files, and the actual SAS macro source code files. This information allows programmers to modify the SAS code to suit their needs and to use it on various platforms. An active Internet connection is not required to run the SAS macros when the companion CD-ROM is available.

Acknowledgments

I am indebted to many individuals who have directly and indirectly contributed to the development of this book. Many thanks to my graduate advisor, Prof. Creighton Miller, Jr., at Texas A&M University, and to Prof. Rangesan Narayanan at the University of Nevada–Reno, both of whom in one way or another have positively influenced my career all these years. I am grateful to my colleagues and my former and current students who have presented me with consulting problems over the years that have stimulated me to develop this book and the accompanying SAS macros. I would also like to thank the University of Nevada–Reno College of Agriculture–Biotechnology–Natural Resources, Nevada Agricultural Experimental Station, and the University of Nevada Cooperative Extension for their support during the time I spent writing the book and developing the SAS macros.

I am also grateful to Ann Dougherty for reviewing the initial book proposal, as well as Andrea Meyer and Suchitra Injati for reviewing some parts of the material. I have received constructive comments from many CRC Press anonymous reviewers on this book, and their advice has greatly improved this book. I would like to acknowledge the contributions of the CRC Press staff, from the conception to the completion of this book. My special thanks go to Jasmin Naim, Helena Redshaw, Nadja English, and Naomi Lynch of the CRC Press publishing team for their tremendous efforts to produce this book in a timely fashion. A special note of thanks to Kirsty Stroud for finding me in the first place and suggesting that I work on this book, thus providing me with a chance to share my work with fellow SAS users. I would also like to thank the SAS Institute for providing me with an opportunity to learn about this powerful software over the past 23 years and allowing me to share my SAS knowledge with other users.

I owe a great debt of gratitude to my family for their love and support as well as their great sacrifice during the last 12 months. I cannot forget to thank my dad, Pancras Fernandez, and my late grandpa, George Fernandez, for their love and support, which helped me to take on challenging projects and succeed. I would like to thank my son, Ryan Fernandez, for helping me create the table of contents.

A very special thanks goes to my daughter, Ramya Fernandez, for reviewing this book from beginning to end and providing me with valuable suggestions. Finally, I would like to thank the most important person in my life, my wife, Queency Fernandez, for her love, support, and encouragement, which gave me the strength to complete this project within the deadline.

George Fernandez

Contents

Chapter 1

Data Mining: A Gentle Introduction

1.1 Introduction

Data mining, or knowledge discovery in databases (KDD), is a powerful information technology tool with great potential for extracting previously unknown and potentially useful information from large databases. Data mining automates the process of finding relationships and patterns in raw data and delivers results that can be either utilized in an automated decision support system or assessed by decision makers. Many successful organizations practice data mining for intelligent decision-making.[1] Data mining allows the extraction of nuggets of knowledge from business data that can help enhance customer relationship management (CRM)[2] and can help estimate the return on investment (ROI).[3] Using powerful analytical techniques, data mining enables institutions to turn raw data into valuable information to gain a critical competitive advantage

With data mining, the possibilities are endless. Although data mining applications are popular among forward-thinking businesses, other disciplines that maintain large databases could reap the same benefits from properly carried out data mining. Some of the potential applications of data mining include characterizations of genes in animal and plant genomics, clustering and segmentation in remote sensing of satellite image data, and predictive modeling in wildfire incidence databases.

The purpose of this chapter is to introduce data mining concepts, provide some examples of data mining applications, list the most commonly used data mining techniques, and briefly discuss the data mining applications available in the SAS software. For a thorough discussion of data mining concepts, methods, and applications, see Two Crows Corporation[4] and Berry and Linoff.[5,6]

1.2 Data Mining: Why Now?

1.2.1 Availability of Large Databases and Data Warehousing

Data mining derives its name from the fact that analysts search for valuable information among gigabytes of huge databases. For the past two decades, we have seen an explosive rate of growth in the amount of data being stored in an electronic format. The increase in the use of electronic data gathering devices such as point-of-sale, web logging, or remote sensing devices has contributed to this explosion of available data. The amount of data accumulated each day by various businesses and scientific and governmental organizations around the world is daunting.

Data warehousing collects data from many different sources, reorganizes it, and stores it within a readily accessible repository that can be utilized for productive decision making using data mining. A data warehouse (DW) should support relational, hierarchical, and multidimensional database management systems and is designed specifically to meet the needs of data mining. A DW can be loosely defined as any centralized data repository that makes it possible to extract archived operational data and overcome inconsistencies between different data formats. Thus, data mining and knowledge discovery from large databases become feasible and productive with the development of cost-effective data warehousing.

1.2.2 Price Drop in Data Storage and Efficient Computer Processing

Data warehousing has become easier and more efficient and cost effective as data processing and database development have become less expensive. The need for improved and effective computer processing can now be met in a cost-effective manner with parallel multiprocessor computer technology. In addition to the recent enhancement of exploratory graphical statistical methods, the introduction of new machine learning methods based on logic programming, artificial intelligence, and genetic algorithms opened the doors for productive data mining. When data mining tools are implemented on high-performance, parallel-processing systems, they can analyze massive databases in minutes. Faster processing means that users can automatically experiment with more models to understand complex data. The high speed makes it more practical for users to analyze huge quantities of data.

1.2.3 New Advancements in Analytical Methodology

Data mining algorithms embody techniques that have existed for at least 10 years but have only recently been implemented as mature, reliable, understandable tools that consistently outperform older methods. Advanced analytical models and algorithms, such as data visualization and exploration, segmentation and clustering, decision trees, neural networks, memory-based reasoning, and market basket analysis, provide superior analytical depth. Thus, quality data mining is now feasible with the availability of advanced analytical solutions.

1.3 Benefits of Data Mining

For businesses that use data mining effectively, the payoffs can be huge. By applying data mining effectively, businesses can fully utilize data about customers' buying patterns and behavior and gain a greater understanding of customers' motivations to help reduce fraud, forecast resource use, increase customer acquisition, and curb customer attrition. Successful implementation of data mining techniques sweeps through databases and identifies previously hidden patterns in one step. An example of pattern discovery is the analysis of retail sales data to identify seemingly unrelated products that are often purchased together. Other pattern discovery applications include detecting fraudulent credit card transactions and identifying anomalous data that could represent data entry keying errors. Some of the specific benefits associated with successful data mining include:

- Increase customer acquisition and retention.
- Uncover and reduce fraud (determining if a particular transaction is out of the normal range of a person's activity and flagging that transaction for verification).
- Improve production quality and minimize production losses in manufacturing.
- Increase up-selling (offering customers a higher level of services or products, such as a gold credit card vs. a regular credit card) and cross-selling (selling customers more products based on what they have already bought).
- Sell products and services in combinations based on market basket analysis (by determining what combinations of products are purchased at a given time).

1.4 Data Mining: Users

Data mining applications have recently been deployed successfully by a wide range of companies.[1] While the early adopters of data mining belong mainly to information-intensive industries such as as financial services and direct mail marketing, the technology is applicable to any institution seeking to leverage a large data warehouse to extract information that can be used in intelligent decision making. Data mining applications reach across industries and business functions. For example, telecommunications, stock exchange, credit card, and insurance companies use data mining to detect fraudulent use of their services; the medical industry uses data mining to predict the effectiveness of surgical procedures, diagnostic medical tests, and medications; and retailers use data mining to assess the effectiveness of discount coupons and sales promotions. Data mining has many varied fields of application, some of which are listed below:

- *Retail/marketing.* An example of pattern discovery in retail sales is to identify seemingly unrelated products that are often purchased together. Market basket analysis is an algorithm that examines a long list of transactions in order to determine which items are most frequently purchased together. The results can be useful to any company that sells products, whether in a store, by catalog, or directly to the customer.
- *Banking.* A credit card company can leverage its customer transaction database to identify customers most likely to be interested in a new credit product. Using a small test mailing, the characteristics of customers with an affinity for the product can be identified. Data mining tools can also be used to detect patterns of fraudulent credit card use, including detecting fraudulent credit card transactions and identifying anomalous data that could represent data entry keying errors. It identifies loyal customers, predicts customers likely to change their credit card affiliation, determines credit card spending by customer groups, uncovers hidden correlations among various financial indicators, and identifies stock trading trends from historical market data.
- *Healthcare insurance.* Through claims analysis (i.e., identifying medical procedures that are claimed together), data mining can predict which customers will buy new policies, defines behavior patterns of risky customers, and identifies fraudulent behavior.
- *Transportation.* State and federal departments of transportation can develop performance and network optimization models to predict the life-cycle costs of road pavement.
- *Product manufacturing companies.* Manufacturers can apply data mining to improve their sales process to retailers. Data from consumer panels, shipments, and competitor activity can be applied to understand the reasons for brand and store switching. Through this analysis, a manufacturer can select promotional strategies that best reach their target customer segments. Data mining can determine distribution schedules among outlets and analyze loading patterns.
- *Healthcare and pharmaceutical industries.* A pharmaceutical company can analyze its recent sales records to improve targeting of high-value physicians and determine which marketing activities will have the greatest impact in the next few months. The ongoing, dynamic analysis of the data warehouse

allows the best practices from throughout the organization to be applied in specific sales situations.

- *Internal Revenue Service (IRS) and Federal Bureau of Investigation (FBI).* As examples, the IRS uses data mining to track federal income tax frauds, and the FBI uses data mining to detect any unusual patterns or trends in thousands of field reports to look for any leads in terrorist activities.

1.5 Data Mining Tools

All data mining methods used now have evolved from advances in artificial intelligence (AI), statistical computation, and database research. Data mining methods are not considered as replacements of traditional statistical methods but as extensions of the use of statistical and graphical techniques. Once it was thought that automated data mining tools would eliminate the need for statistical analysts to build predictive models, but the value that an analyst provides cannot be automated out of existence. Analysts are still necessary to assess model results and validate the plausibility of the model predictions. Because data mining software lacks the human experience and intuition to recognize the difference between a relevant and irrelevant correlation, statistical analysts will remain in high demand.

1.6 Data Mining Steps

1.6.1 Identification of Problem and Defining the Business Goal

One of the main causes of data mining failure is not defining the business goals based on short- and long-term problems facing the enterprise. The data mining specialist should define the business goal in clear and sensible terms as far as specifying what the enterprise hopes to achieve and how data mining can help. Well-identified business problems lead to formulated business goals and data mining solutions geared toward measurable outcomes.[4]

1.6.2 Data Processing

The key to successful data mining is using the appropriate data. Preparing data for mining is often the most time-consuming aspect of any data mining endeavor. Typical data structure suitable for data mining should contain observations (e.g., customers and products) in rows and variables (e.g., demographic data and sales history) in columns. Also, the measurement levels (interval or categorical) of each variable in the dataset should be clearly defined. The steps involved in preparing the data for data mining are as follows:

- *Preprocessing.* This is the data cleansing stage, where certain information that is deemed unnecessary and likely to slow down queries is removed. Also, the data are checked to ensure use of a consistent format in dates, zip codes, currency, units of measurements, etc. Inconsistent formats in the database are always a possibility because the data are drawn from several sources. Data entry errors and extreme outliers should be removed from the dataset because influential outliers can affect the modeling results and subsequently limit the usability of the predicted models.
- *Data integration.* Combining variables from many different data sources is an essential step because some of the most important variables are stored in different data marts (customer demographics, purchase data, business transaction). The uniformity in variable coding and the scale of measurements should be verified before combining different variables and observations from different data marts.
- *Variable transformation.* Sometimes expressing continuous variables in standardized units (or in log or square-root scale) is necessary to improve the model fit that leads to improved precision in the fitted models. Missing value imputation is necessary if some important variables have large proportions of missing values in the dataset. Identifying the response (target) and the predictor (input) variables and defining their scale of measurement are important steps in data preparation because the type of modeling is determined by the characteristics of the response and the predictor variables.
- *Splitting databases.* Sampling is recommended in extremely large databases because it significantly reduces the model training time. Randomly splitting the data into training, validation, and testing categories is very important in calibrating the model fit and validating the model results. Trends and patterns observed in the training dataset can be expected to generalize the complete database if the training sample used sufficiently represents the database.

1.6.3 Data Exploration and Descriptive Analysis

Data exploration includes a set of descriptive and graphical tools that allow exploration of data visually both as a prerequisite to more formal data analysis and as an integral part of formal model building. It facilitates discovering the unexpected, as well as confirming the expected. The purpose of data visualization is pretty simple: to let the user understand the structure and dimension of the complex data matrix. Because data mining usually involves extracting "hidden" information from a database, the understanding process can get a bit complicated. The key is to put users in a context in which they feel comfortable and then let them poke and prod until they uncover what they did not see before. Understanding is undoubtedly the most fundamental motivation behind visualizing the model.

Simple descriptive statistics and exploratory graphics displaying the distribution pattern and the presence of outliers are useful in exploring continuous variables. Descriptive statistical measures such as the mean, median, range, and standard deviation of continuous variables provide information regarding their distributional properties and the presence of outliers. Frequency histograms display the distributional properties of the continuous variable. Box plots provide an excellent visual summary of many important aspects of a distribution. The box plot is based on a five-number summary plot, which is based on the median, quartiles, and extreme values. One-way and multi-way frequency tables of categorical data are useful in summarizing group distributions and relationships between groups, as well as checking for rare events. Bar charts show frequency information for categorical variables and display differences among the various groups in the categorical variable. Pie charts compare the levels or classes of a categorical variable to each other and to the whole. They use the size of pie slices to graphically represent the value of a statistic for a data range.

1.6.4 Data Mining Solutions:Unsupervised Learning Methods

Unsupervised learning methods are used in many fields under a wide variety of names. No distinction between the response and predictor variable is made in unsupervised learning methods. The most commonly practiced unsupervised methods are latent variable models (principal component and factor analyses), disjoint cluster analyses, and market basket analysis:

- *Principal component analysis (PCA).* In PCA, the dimensionality of multivariate data is reduced by transforming the correlated variables into linearly transformed uncorrelated variables.
- *Factor analysis (FA).* In FA, a few uncorrelated hidden factors that explain the maximum amount of common variance and are responsible for the observed correlation among the multivariate data are extracted.
- *Disjoint cluster analysis (DCA).* DCA is used for combining cases into groups or clusters such that each group or cluster is homogeneous with respect to certain attributes.
- *Association and market basket analysis.* Market basket analysis is one of the most common and useful types of data analysis for marketing. The purpose of market basket analysis is to determine what products customers purchase together. Knowing what products consumers purchase as a group can be very helpful to a retailer or to any other company.

1.6.5 Data Mining Solutions: Supervised Learning Methods

The supervised predictive models include both classification and regression models. Classification models use categorical responses while regression models use continuous and binary variables as targets. In regression we want to approximate the regression function, while in classification problems we want to approximate the probability of class membership as a function of the input variables. Predictive modeling is a fundamental data mining task. It is an approach that reads training data composed of multiple input variables and a target variable. It then builds a model that attempts to predict the target on the basis of the inputs. After this model is developed, it can be applied to new data similar to the training data but not containing the target.

- *Multiple linear regression (MLR).* In MLR, the association between the two sets of variables is described by a linear equation that predicts the continuous response variable from a function of predictor variables.
- *Logistic regressions.* This type of regression uses a binary or an ordinal variable as the response variable and allows construction of more complex models than the straight linear models do.
- *Neural net (NN) modeling.* Neural net modeling can be used for both prediction and classification. NN models enable construction of trains and validate multiplayer feed-forward network models for modeling large data and complex interactions with many predictor variables. NN models usually contain more parameters than a typical statistical model, the results are not easily interpreted, and no explicit rationale is given for the prediction. All variables are considered to be numeric and all nominal variables are coded as binary. Relatively more training time is needed to fit the NN models.
- *Classification and regression tree (CART).* These models are useful in generating binary decision trees by splitting the subsets of the dataset using all predictor variables to create two child nodes repeatedly beginning with the entire dataset. The goal is to produce subsets of the data that are as homogeneous as possible with respect to the target variable. Continuous, binary, and categorical variables can be used as response variables in CART.
- *Discriminant function analysis.* This is a classification method used to determine which predictor variables discriminate between two or more naturally occurring groups. Only categorical variables are allowed to be the response variable and both continuous and ordinal variables can be used as predictors.
- *Chi-square automatic interaction detector (CHAID) decision tree.* This is a classification method used to study the relationships between a categorical response measure and a large series of possible predictor variables that may interact with each other. For qualitative predictor variables, a series of chi-square analyses are conducted between the response and predictor variables to see if splitting the sample based on these predictors leads to a statistically significant discrimination in the response.

1.6.6 Model Validation

Validating models obtained from training datasets by independent validation datasets is an important requirement in data mining to confirm the usability of the developed model. Model validation assesses the quality of the model fit and protects against over-fitted or under-fitted models. Thus, model validation could be considered as the most important step in the model building sequence.

1.6.7 Interpretation and Decision Making

Decision making is critical for any successful business. No matter how good a person may be at making decisions, making an intelligent decision can be difficult. The patterns identified by the data mining solutions can be transformed into knowledge, which can then be used to support business decision making.

1.7 Problems in the Data Mining Process

Many of the so-called data mining solutions currently available on the market today do not integrate well, are not scalable, or are limited to one or two modeling techniques or algorithms. As a result, highly trained quantitative experts spend more time trying to access, prepare, and manipulate data from disparate sources and less time modeling data and applying their expertise to solve business problems. The data mining challenge is compounded even further as the amount of data and complexity of the business problems increase. Often, the database is designed for purposes other than data mining, so properties or attributes that would simplify the learning task are not present and cannot be requested from the real world.

Data mining solutions rely on databases to provide the raw data for modeling, and this raises problems in that databases tend to be dynamic, incomplete, noisy, and large. Other problems arise as a result of the adequacy and relevance of the information stored. Databases are usually contaminated by errors so it cannot be assumed that the data they contain are entirely correct. Attributes, which rely on subjective or measurement judgments, can give rise to errors in such a way that some examples may even be misclassified. Errors in either the values of attributes or class information are known as noise. Obviously, where possible, it is desirable to eliminate noise from the classification information, as this affects the overall accuracy of the generated rules; therefore, adopting a software system that provides a complete data mining solution is crucial in the competitive environment.

1.8 SAS Software: The Leader in Data Mining

SAS Institute,[7] the industry leader in analytical and decision support solutions, offers a comprehensive data mining solution that allows users to explore large quantities of data and discover relationships and patterns that lead to proactive decision making. The SAS data mining solution provides business technologists and quantitative experts the necessary tools to obtain the enterprise knowledge necessary for their organizations to achieve a competitive advantage.

1.8.1 SEMMA: The SAS Data Mining Process

The SAS data mining solution is considered a process rather than a set of analytical tools. Beginning with a statistically representative sample of the data, SEMMA makes it easy to apply exploratory statistical and visualization techniques, select and transform the most significant predictive variables, model the variables to predict outcomes, and confirm the accuracy of a model. The acronym SEMMA refers to a methodology that clarifies this process:[8]

- *Sample* the data by extracting a portion of a dataset large enough to contain the significant information, yet small enough to manipulate quickly.
- *Explore* the data by searching for unanticipated trends and anomalies in order to gain understanding and ideas.
- *Modify* the data by creating, selecting, and transforming the variables to focus the model selection process.
- *Model* the data by allowing the software to search automatically for a combination of data that reliably predicts a desired outcome.
- *Assess* the data by evaluating the usefulness and reliability of the findings from the data mining process.

By assessing the results gained from each stage of the SEMMA process, users can determine how to model new questions raised by previous results and thus proceed back to the exploration phase for additional refinement of the data. The SAS data mining solution integrates everything necessary for discovery at each stage of the SEMMA process: These data mining tools indicate patterns or exceptions, and mimic human abilities for comprehending spatial, geographical, and visual information sources. Complex mining techniques are carried out in a totally code-free environment, allowing analysts to concentrate on visualization of the data, discovery of new patterns, and new questions to ask.

1.8.2 SAS Enterprise Miner for Comprehensive Data Mining Solutions

Enterprise Miner,[9,10] SAS Institute's enhanced data mining software, offers an integrated environment for businesses that want to conduct comprehensive data mining. *Enterprise Miner* combines a rich suite of integrated data mining tools, empowering users to explore and exploit huge databases for strategic business advantages. In a single environment, *Enterprise Miner* provides all the tools necessary to match robust data mining techniques to specific business problems, regardless of the amount or source of data or complexity of the business problem.

It should be noted, however, that the annual licensing fee for using *Enterprise Miner* is extremely high, so small businesses, nonprofit institutions, and academic universities are unable to take advantage of this powerful analytical tool for data mining. Trying to provide complete SAS codes here for performing comprehensive data mining solutions would not be very effective because a majority of business and statistical analysts are not experienced SAS programmers. Also, quick results from data mining are not feasible because many hours of modifying code and debugging program errors are required when analysts are required to work with SAS program codes.

1.9 User-Friendly SAS Macros for Data Mining

Alternatives to the point-and-click menu interface modules and high-priced SAS *Enterprise Miner* are the user-friendly SAS macro applications for performing several data mining tasks that are included in this book. This macro approach integrates the statistical and graphical tools available in SAS systems and provides user-friendly data analysis tools that allow data analysts to complete data mining tasks quickly, without writing SAS programs, by running the SAS macros in the background. Detailed instructions and help files for using the SAS macros are included in each chapter. Using this macro approach, analysts can effectively and quickly perform complete data analysis, which allows them to spend more time exploring data and interpreting graphs and output rather than debugging program errors. The main advantages of using these SAS macros for data mining include:

- Users can perform comprehensive data mining tasks by inputting the macro parameters in the macro-call window and by running the SAS macro.
- SAS codes required for performing data exploration, model fitting, model assessment, validation, prediction, and scoring are included in each macro so complete results can be obtained quickly.

- Experience in the SAS output delivery system (ODS) is not required because options for producing SAS output and graphics in RTF, WEB, and PDF are included within the macros.
- Experience in writing SAS program codes or SAS macros is not required to use these macros.
- The SAS enhanced data mining software *Enterprise Miner* is not required to run these SAS macros.
- All SAS macros included in this book use the same simple user-friendly format, so minimal training time is needed to master usage of these macros.
- Experienced SAS programmers can customize these macros by modifying the SAS macro codes included.
- Regular updates to the SAS macros will be posted in the book website, so readers can always take advantage of the updated features in the SAS macros by downloading the latest versions.

The fact that these SAS macros do not use *Enterprise Miner* is something of a limitation in that SAS macros could not be included for performing neural net, CART, and market basket analysis, as these data mining tools require the use of *Enterprise Miner.*

1.10 Summary

Data mining is a journey — a continuous effort to combine business knowledge with information extracted from acquired data. This chapter briefly introduces the concept and applications of data mining, which is the secret and intelligent weapon that unleashes the power hidden in data. The SAS Institute, the industry leader in analytical and decision support solutions, provides the powerful software *Enterprise Miner* to perform complete data mining solutions; however, because of the high price tag for *Enterprise Miner*, application of this software is not feasible for all business analysts and academic institutions. As alternatives to the point-and-click menu interface modules and *Enterprise Miner*, user-friendly SAS macro applications for performing several data mining tasks are included in this book. Instructions are given in the book for downloading and applying these user-friendly SAS macros for producing quick and complete data mining solutions.

References

1. SAS Institute, Inc., *Customer Success Stories* (http://www.sas.com/news/success/solutions.html).

2. SAS Institute, Inc., *Customer Relationship Management* (http://www.sas.com/solutions/crm/index.html).
3. SAS Institute, Inc., SAS *Enterprise Miner* Product Review (http://www.sas.com/products/miner/miner_review.pdf).
4. Two Crows Corporation, *Introduction to Data Mining and Knowledge Discovery*, 3rd ed., Potomac, MD, 1999 (http://www.twocrows.com/intro-dm.pdf).
5. Berry, M.J.A. and Linoff, G.S., *Data Mining Techniques: For Marketing, Sales, and Customer Support,* John Wiley & Sons, New York, 1997.
6. Berry, M.J.A. and Linoff, G.S., *Mastering Data Mining: The Art and Science of Customer Relationship Management,* 2nd ed., John Wiley & Sons, New York, 1999.
7. SAS Institute, Inc., *The Power To Know* (http://www.sas.com).
8. SAS Institute, Inc., *Data Mining Using Enterprise Miner Software: A Case Study Approach,* 1st ed., SAS Institute, Inc., Cary, NC, 2000.
9. SAS Institute, Inc., *The Enterprise Miner* (http://www.sas.com/products/miner/index.html).
10. SAS Institute, Inc., *The Enterprise Miner Standalone Tutorial* (http://www. sas.com/service/tutorials/v8/em/mainmenu.htm).

Suggested Reading and Case Studies

Exclusive Core, Inc., *Data Mining Case Study: Retail Marketing* (http://www. exclusiveore.com/casestudies/case%20study_telco.pdf).

Exclusive Core, Inc., *Data Mining Case Study: Telecom Churn Study* (http://www. exclusiveore.com/casestudies/case%20study_telco.pdf).

Exclusive Core, Inc., *Data Warehousing and OLAP Case Study: Fast Food* (http://www.exclusiveore.com/casestudies/case%20study_fastfood.pdf).

Gerritsen, R., *A Data Mining Case Study: Assessing Loan Risks* (http://www. exclusiveore.com/casestudies/dm%20at%20usda%20(itpro).pdf).

Linoff, G.S. and Berry, M.J.A., *Mining the Web: Transforming Customer Data into Customer Value*, John Wiley & Sons, New York, 2002.

Megaputer Intelligence, *Data Mining Case Studies* (http://www.megaputer.com/company/pacases.php3).

Pyle, D., *Data Preparation for Data Mining*, Morgan Kaufmann, San Francisco, CA, 1999.

Rud, O.P., *Data Mining Cookbook: Modeling Data for Marketing, Risk, and Customer Relationship Management*, John Wiley & Sons, New York, 2000.

SAS Institute, Inc., *Data Mining and the Case for Sampling: Solving Business Problems Using SAS Enterprise Miner Software*, SAS Institute, Inc., Cary, NC (http://www.ag.unr.edu/gf/dm/sasdm.pdf).

SAS Institute, Inc., *Using Data Mining Techniques for Fraud Detection: Solving Business Problems Using SAS Enterprise Miner Software* (http://www.ag.unr.edu/gf/dm/dmfraud.pdf).

Small, R.D., Debunking data mining myths, *Information Week*, January 20, 1997 (http://www.twocrows.com/iwk9701.htm).

Soukup, T. and Davidson, I., *Visual Data Mining: Techniques and Tools for Data Visualization and Mining*, John Wiley & Sons, New York, 2002.

Thuraisingham, B., *Data Mining: Technologies, Techniques, Tools, and Trends*, CRC Press, Boca Raton, FL, 1998.

Way, R., *Using SAS/INSIGHT Software as an Exploratory Data Mining Platform* (http://www2.sas.com/proceedings/sugi24/Infovis/p160-24.pdf).

Westphal, C. and Blaxton, T., *Data Mining Solutions*, John Wiley & Sons, New York, 1998.

Chapter 2

Preparing Data for Data Mining

2.1 Introduction

Data are the backbone of data mining and knowledge discovery; however, real-world business data usually are not available in data-mining-ready form. The biggest challenge for data miners, then, is preparing data suitable for modeling. Many businesses maintain central data storage and access facilities called data warehouses. Data warehousing is defined as a process of centralized data management and allows analysts to access, update, and maintain the data for analysis and reporting. Thus, data warehouse technology improves the efficiency of extracting and preparing data for data mining. Popular data warehouses use relational databases (e.g., Oracle, Informix, Sybase), and the PC data format (spreadsheets and MS Access). Roughly 70% of data mining operation time is spent on preparing the data obtained from different sources; therefore, considerable time and effort should be spent on preparing data tables to be suitable for data mining modeling.

2.2 Data Requirements in Data Mining

Summarized data are not suitable for data mining because information about individual customers or products is not available. For example, to identify profitable customers, individual customer records that include demographic information are necessary to profile or cluster customers based on their purchasing patterns. Similarly, to identify the characteristics of profitable customers in a predictive model, target (outcome or response) and input (predictor) variables should be included. Therefore, for solving specific business objectives, suitable data must be extracted from data warehouses or new data collected that meet the data mining requirements.

2.3 Ideal Structures of Data for Data Mining

The rows (observations or cases) and columns (variables) format, similar to a spreadsheet worksheet file, is required for data mining. The rows usually contain information regarding individual customers or consumer products. The columns describe the attributes (variables) of individual cases. The variables can be continuous or categorical. Total sales per product, number of units purchased by each customer, and annual income per customer are some examples of continuous variables. Gender, race, and age group are considered categorical variables. Knowledge about the possible maximum and minimum values for the continuous variables can help to identify and exclude extreme outliers from the data. Similarly, knowledge about the possible levels for categorical variables can help to detect data entry errors and anomalies in the data.

Constant values in continuous (e.g., zip code) or categorical (state code) fields should not be included in any predictive or descriptive data mining modeling because these values are unique for each case and do not help to discriminate or group individual cases. Similarly, unique information about customers, such as phone numbers and Social Security numbers, should also be excluded from predictive data mining; however, these unique value variables can be used as ID variables to identify individual cases and exclude extreme outliers. Also, it is best not to include highly correlated (correlation coefficient >0.95) continuous predictor variables in predictive data mining, as they can produce unstable predictive models that work only with the particular sample used.

2.4 Understanding the Measurement Scale of Variables

The measurement scale of the target and input variables determines the type of modeling technique that is appropriate for a specific data mining project; therefore, understanding the nature of the measurement scale of variables used in modeling is an important data mining requirement. The variables can be generally classified into continuous or categorical.

Continuous variables are numeric variables that describe quantitative attributes of the cases and have a continuous scale of measurement. Means and standard deviations are commonly used to quantify the central tendency and dispersion. Total sales per customers and total manufacturing costs per products are examples of interval scales. An interval-scale target variable is a requirement for multiple regression and neural net modeling.

Categorical variables can be further classified as:

- *Nominal*, a categorical variable with more than two levels. Mode is the preferred estimate for measuring the central tendency, and frequency analysis is the common form of descriptive technique. Different kinds of accounts in banking, telecommunication services, and insurance policies are some examples of nominal variables. Discriminant analysis and decision tree methods are suitable for modeling nominal target variables.
- *Binary*, a categorical variable with only two levels. Sale vs. no sale and good vs. bad credit are some examples of binary variables. Logistic regression is suitable for modeling binary target variables.
- *Ordinal*, a categorical or discrete rank variable with more than two levels. Ordinal logistic regression is suitable for modeling ordinal variables.

2.5 Entire Database vs. Representative Sample

To find trends and patterns in business data, data miners can use the entire database or randomly selected samples from the entire database. Although using the entire database is currently feasible with today's high-powered computing environment, using randomly selected representative samples in model building is more attractive due to the following reasons:

- Using random samples allows the modeler to develop the model from training or calibration samples, validate the model with a holdout "validation" dataset, and test the model with another independent test sample.
- Mining a representative random sample is easier and more efficient and can produce accurate results similar to those produced when using the entire database.
- When samples are used, data exploration and visualization help to gain insights that lead to faster and more accurate models.
- Representative samples require a relatively shorter time to cleanse, explore, and develop and validate models. They are therefore more cost effective than using entire databases.

2.6 Sampling for Data Mining

The sample used in modeling should represent the entire database because the main goal in data mining is to make predictions about the entire database. The size and other characteristics of the selected sample determine whether the sample used in modeling is a good representation of the entire database. The following types of sampling are commonly practiced in data mining:[1]

- *Simple random sampling.* This is the most common sampling method in data mining. Each observation or case in the database has an equal chance of being included in the sample.
- *Cluster sampling.* The database is divided into clusters at the first stage of sample selection and a few of those clusters are randomly selected based on random sampling. All the records from those randomly selected clusters are included in the study.
- *Stratified random sampling.* The database is divided into mutually exclusive strata or subpopulations; random samples are then taken from each stratum proportional to its size.

2.6.1 Sample Size

The number of input variables, the functional form of the model (liner, nonlinear, models with interactions, etc.) and the size of the databases can influence the sample size requirement in data mining. By default, the SAS *Enterprise Miner* software takes a simple random sample of 2000 cases from the data table and divides it into TRAINING (40%), VALIDATION (30%), and TEST (30%) datasets.[2] If the number of cases is less than 2000, the entire database is used in the model building. Data analysts can use these sampling proportions as a guideline in determining sample sizes; however, depending on the data mining objectives and the nature of the database, data miners can modify sample size proportions.

2.7 SAS Applications Used in Data Preparation

SAS software has many powerful features available for extracting data from different database management systems (DBMS). Some of the features are described in the following section. Readers are expected to have a basic knowledge in using SAS to perform the following operations. *The Little SAS Book*[3] can serve as an introductory SAS guide to become familiar with the SAS systems and SAS programming.

2.7.1 Converting Relational DBMS into SAS Datasets

2.7.1.1 Instructions for Extracting SAS Data from Oracle Database Using the SAS SQL Pass-Through Facility

If you have SAS/ACCESS software installed for your DBMS, you can extract DBMS data by using the PROC SQL (SAS/BASE) pass-through facility. The following SAS code can be modified to create an SAS data "SAS_data_name" from the Oracle database "tbl_name" to extract all the variables by inputting the username, password, file path, oracle filename, and the SAS dataset name:

```
PROC SQL;
CONNECT TO oracle(USER = <user> ORAPW = <pass-
word> PATH = "mypath");
CREATE TABLE sas_data_name AS
SELECT *
FROM CONNECTION TO oracle
(SELECT * FROM tbl_name);
DISCONNECT FROM oracle;
QUIT;
```

Users can find additional SAS sample files in the SAS online site, which provides instructions and many examples to extract data using the SQL pass-through facility.[4]

2.7.1.2 Instructions for Creating SAS Dataset from Oracle Database Using SAS/ACCESS and the LIBNAME Statement

In SAS version 8.0, an Oracle database can be identified directly by associating it with the LIBNAME statement if the SAS/ACCESS software is installed. The following SAS code illustrates the DATA step with LIBNAME that refers to the Oracle database:

```
LIBNAME myoralib ORACLE
USER = <user>
PASSWORD = <password>
PATH = "mypath"
SCHEMA = hrdept
PRESERVE_COL_NAMES = yes;
PROC CONTENTS DATA = myoralib.orafilename;
TITLE "The list of variable names and charac-
teristics in the Oracle data";
RUN;
```

2.7.2 Converting PC-Based Data Files

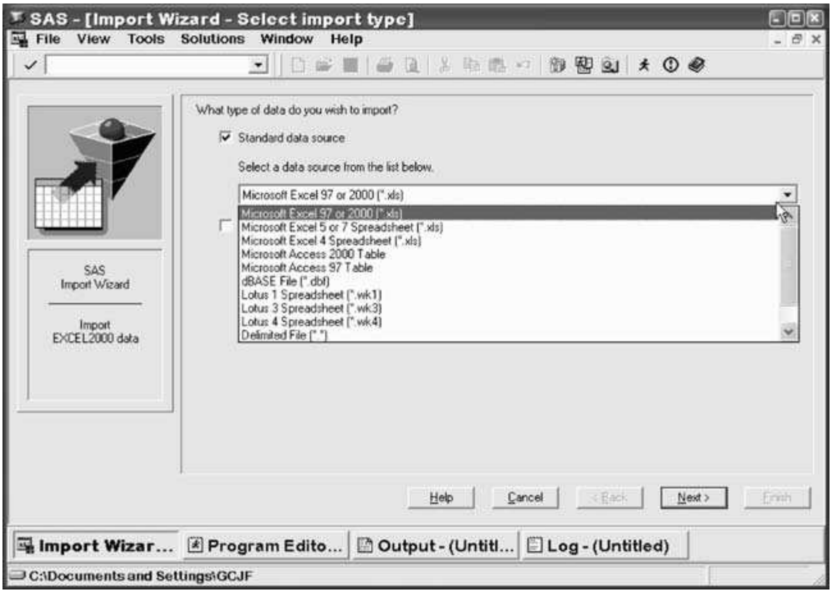
MS Excel, Access, dBase, Lotus worksheets, and tab-delimited and comma-separated are some of the popular PC data files used in data mining. These file types can be easily converted to SAS datasets by using the PROC ACCESS or PROC IMPORT procedures in SAS. A graphical user interface (GUI)-based import wizard is also available in SAS to convert a single PC file type to an SAS dataset, but, before converting the PC file types, the following points should be considered:

- Be aware that the maximum number of rows and columns allowed in an Excel worksheet is 65,536 × 246.
- Check to see that the first row of the worksheet contains the names of the variables stored in the columns. Select names that are valid SAS variable names (one word, maximum length of 8 characters). Also, do not have any blank rows in the worksheet.
- Save only one data table per worksheet. Name the data table to "sheet1" if you are importing an MS Access table.
- Be sure to close the Excel file before trying to convert it in SAS, as SAS cannot read a worksheet file that is currently open in Excel. Trying to do so will cause a sharing violation error.
- Assign a LIBNAME before importing the PC file into an SAS dataset to create a permanent SAS data file. For information on the LIBNAME statement and making permanent SAS data files, refer to *The Little SAS Book*.[3]
- Make sure that each column in a worksheet contains either numeric or character variables. Do not mix numeric and character values in the same column. The results of most Excel formulas should import into SAS without a problem.

2.7.2.1 Instructions for Converting PC Data Formats to SAS Datasets Using the SAS Import Wizard

The SAS import wizard available in the SAS/ACCESS module can be used to import or export Excel 4, 5, 7 (95), 98, and 2000 files, as well as Microsoft Access files in version 8.0. The GUIs in the import wizard guide users through menus and provide step-by-step instructions for transferring data between external data sources and SAS datasets. The types of files that can be imported depend on the operating system and the SAS/ACCESS engines installed. The steps involved in using the import wizard for importing a PC file follow:

1. *Select the PC file type.* The import wizard can be activated by using the pull-down menu, selecting FILE, and then clicking IMPORT. For a list of available data sources from which to choose, click the drop-down arrow (Figure 2.1). Select the file format in which your data are stored. To read an Excel file, click the black triangle and choose the type of Excel file (4.0, 5.0, 7.0 (95), 97, and 2000 spreadsheets). You can also select other PC file types, such as MS Access (97 and 2000 tables), dBASE (5.0, IV, III+, and III files), Lotus (1–2–3 WK1, WK3, and WK4 files), or text files such as tab-delimited and comma-separated files. After selecting the file type, click the NEXT button to continue.
2. *Select the PC file location.* In the import wizard's Select file window, type the full path for the file or click BROWSE to find the file. Then click the NEXT button to go to the next screen. On the second screen, after the Excel file is chosen, the OPTIONS button becomes active. The OPTIONS button allows the user to choose which worksheet to read (if the file has multiple sheets), to specify whether or not the first row of the spreadsheet contains the variable names, and to choose the range of the worksheet to be read. Generally, these options can be ignored.


SAS - [Import Wizard - Select import type]

What type of data do you wish to import?

☑ Standard data source

Select a data source from the list below.

Microsoft Excel 97 or 2000 (*.xls)

- Microsoft Excel 97 or 2000 (*.xls)
- Microsoft Excel 5 or 7 Spreadsheet (*.xls)
- Microsoft Excel 4 Spreadsheet (*.xls)
- Microsoft Access 2000 Table
- Microsoft Access 97 Table
- dBASE File (*.dbf)
- Lotus 1 Spreadsheet (*.wk1)
- Lotus 3 Spreadsheet (*.wk3)
- Lotus 4 Spreadsheet (*.wk4)
- Delimited File (*.*)

SAS Import Wizard

Import EXCEL2000 data

C:\Documents and Settings\GCJF

Figure 2.1 Screen copy of SAS IMPORT (version 8.2) showing the valid file types that can be imported to SAS datasets.

3. *Select the temporary or permanent SAS dataset name.* The third screen prompts for the SAS data file name. Select the LIBRARY (the alias name for the folder) and member (SAS dataset name) for your SAS data file. For example, to create a temporary data file called "fraud", choose "WORK" for the LIBRARY and "fraud" as the valid SAS dataset name for the member. When you are ready, click FINISH, and SAS will convert the specified Excel spreadsheet into an SAS data file.
4. *Perform a final check.* Check the LOG window for a message indicating that SAS has successfully converted the Excel file to an SAS dataset. Also, compare the number of observations and variables in the SAS dataset with the source Excel file to make sure that SAS did not import any empty rows or columns.

2.7.2.2 Converting PC Data Formats to SAS Datasets Using the EXCELSAS Macro

The EXCELSAS macro application can be used as an alternative to the SAS import wizard to convert PC file types to SAS datasets. The SAS procedure PROC IMPORT is the main tool if the EXCELSAS macro is used with post-SAS version 8.0. PROC IMPORT can import a wide variety of types and versions of PC files. However, if the EXCELSAS macro is used in SAS version 6.12, then PROC ACCESS will be selected as the main tool for importing only limited PC file formats. See Section 2.7.2.3 for more details regarding the various PC data formats that can be imported using the EXCELSAS macro. The advantages for using the EXCELSAS macro over the import wizard include:

- Multiple PC files can be converted in a single operation.
- A sample printout of the first 10 observations is produced in the output file.
- The characteristics of the numeric and character variables and number of observations in the converted SAS data file are reported in the output file.
- Descriptive statistics of all the numeric variables and the frequency information of all character variables are reported in the output file.
- Options for saving the output tables in WORD, HTML, PDF, and TXT formats are available.

Software requirements for using the EXCELSAS macro include:

- The SAS/CORE, SAS/BASE, and SAS/ACCESS interface to PC file formats must be licensed and installed at your site.

- The EXCELSAS macro has been tested only in the Windows (Windows 98 and later) environment. However, to import DBF, CSV, and tab-delimited files in the Unix platform, the EXCELSAS macro could be used with minor modification in the macro-call file (see the steps below).
- An active Internet connection is required for downloading the EXCELSAS macro from the book website if the companion CD-ROM is not available.
- SAS version 8.0 or above is recommended for full utilization.

2.7.2.3 Steps Involved in Running the EXCELSAS Macro

1. Prepare the PC data file by following the recommendations given in Section 2.7.2.
2. *If the companion CD-ROM is not available*, first verify that the Internet connection is active. Open the Excelsas.sas macro-call file in the SAS PROGRAM EDITOR window. The Appendix provides instructions for downloading the macro-call and sample data files from the book website. *If the companion CD-ROM is available*, the Excelsas.sas macro-call file can be found in the mac-call folder on the CD-ROM. Open the Excelsas.sas macro-call file in the SAS PROGRAM EDITOR window. Click the RUN icon to submit the macro-call file Excelsas.sas to open the MACRO window called EXCELSAS.
3. Input the appropriate parameters in the macro-call window by following the instructions provided in the EXCELSAS macro help file (see Section 2.7.2.4). After inputting all the required macro parameters, check whether the cursor is in the last input field (#6) and that the RESULTS VIEWER window is closed, then hit the ENTER key (not the RUN icon) to submit the macro.
4. Examine the LOG window for any macro execution errors only in the DISPLAY mode. If any errors in the LOG window are found, activate the PROGRAM EDITOR window, resubmit the Excelsas.sas macro-call file, check the macro input values, and correct any input errors. Otherwise, activate the PROGRAM EDITOR window, resubmit the Excelsas.sas macro-call file, and change the macro input (#6) value from DISPLAY to any other desirable format (see Section 2.7.2.4). The PC file will be imported to a temporary (if macro input #4 is blank or WORK) or permanent (if a LIBNAME is specified in macro input option #4) SAS dataset. The output, including the first 10 observations of the imported SAS data, characteristics of numeric and character variables, simple statistics for numeric variables, and frequency information for the character variables, will be saved in the user-specified format in the user-specified folder as a single file.

2.7.2.4 Help File for SAS Macro EXCELSAS: Description of Macro Parameters

1. **Macro-call parameter:** Input PC file type (required parameter).
   Descriptions and explanation: Include the type of PC file being imported.
   Options/explanations:
   Pre-version 8.0
   Excel — (XLS) files; Excel 95, Excel5, Excel4
   Lotus — (WK4) files
   dBase — (III and IV) files
   Version 8.0 and after
   Excel — (XLS) files; all types of Excel
   Lotus — (WK4) files
   dBase — (III and IV) files
   Access — (mdb) files; 97 and 2000 files
   Tab — (TAB) tab-delimited files
   CSV — (CSV) comma-delimited files
2. **Macro-call parameter:** Input folder name containing the PC file (required parameter).
   Descriptions and explanation: Input the location (path) of folder name containing the PC file. If the field is left blank, SAS will look in the default HOME folder.
   Options/explanations:
   Possible values
   a:\ — A drive
   c:\excel\ — folder named "Excel" in the C drive (be sure to include the back-slash at the end of folder name)
3. **Macro-call parameter:** Input PC file names (required statement).
   Descriptions and explanation: List the names of PC files (without the file extension) being imported. The same file name will be used for naming the imported SAS dataset. If multiple PC files are listed, all of the files can be imported in one operation.
   Options/examples:
   BASEBALL CRIME
   customer99
   Use a short file name (eight characters or less in pre-8.0 versions).
4. **Macro-call parameter:** Optional LIBNAME.
   Descriptions and explanation: To save the imported PC file as a permanent SAS dataset, input the preassigned library (LIBNAME) name. The predefined LIBNAME will tell SAS in which folder to

save the permanent dataset. If this field is left blank, a temporary data file will be created.

Option/example:

SASUSER

The permanent SAS dataset is saved in the library called SASUSER.

5. **Macro-call parameter:** Folder to save SAS output (optional).

   Descriptions and explanation: To save the SAS output files in a specific folder, input the full path of the folder. The SAS dataset name will be assigned to the output file. If this field is left blank, the output file will be saved in the default folder.

   Options/explanations:

   Possible values

   c:\output\ — folder named "OUTPUT"

   s:\george\ — folder named "George" in network drive S

   Be sure to include the back-slash at the end of the folder name.

6. **Macro-call parameter**: Display or save SAS output (required statement).

   Descriptions and explanation: Option for displaying all output files in the OUTPUT window or saving as a specific format in a folder specified in option #5.

   Options/explanations:

   Possible values

   DISPLAY: Output will be displayed in the OUTPUT window. System messages will be displayed in LOG window.

   WORD: Output will be saved in the user-specified folder and viewed in the results VIEWER window as a single RTF format (version 8.0 and later) or saved only as a text file in pre-8.0 versions.

   WEB: Output will be saved in the user-specified folder and viewed in the results VIEWER window as a single HTML file (version 8.0 and later) or saved only as a text file in pre-8.0 versions.

   PDF: Output will be saved in the user-specified folder and viewed in the results VIEWER window as a single PDF file (version 8.2 and later) or saved only as a text file in pre-8.2 versions.

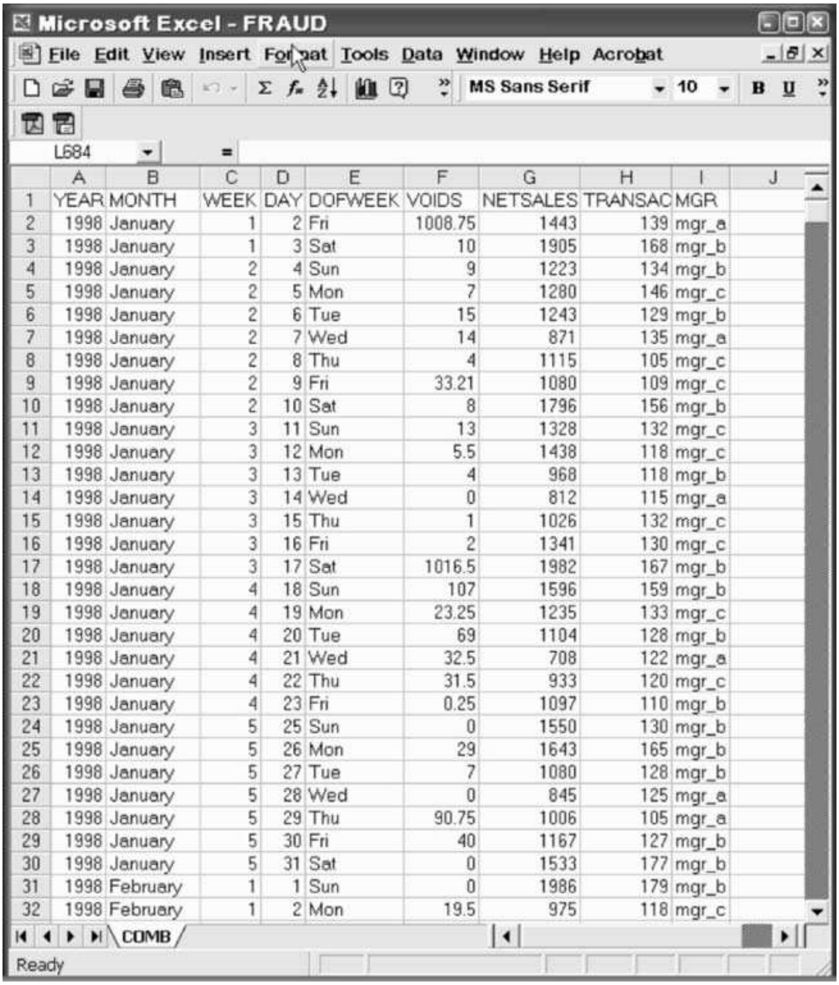
   TXT: Output will be saved as a TXT file in all SAS versions. No output will be displayed in the OUTPUT window.

   Note: All system messages will be deleted from the LOG window at the end of macro execution if DISPLAY is not selected as the macro input in option #6.

2.7.2.5 Importing an Excel File Called "fraud" to a Permanent SAS Dataset Called "fraud"

Source file	fraud.xls; MS Excel sheet 2000
Variables	Daily retail sales, number of transactions, net sales, and manager on duty in a small convenience store
Number of observations	923

1. Open the Excel file "fraud" and make sure that all the specified data requirements reported in Section 2.7.2 are satisfied. The screen copy of the Excel file with the required format is shown in Figure 2.2. Close the "fraud" worksheet file and exit from Excel.
2. Open the EXCELSAS macro-call window in SAS (see Figure 2.3); input the appropriate macro-input values by following the suggestions given in the help file in Section 2.7.2.4. Submit the EXCELSAS macro to import the "fraud" Excel worksheet to a permanent SAS dataset called "fraud".
3. A printout of the first 10 observations including all variables in the SAS dataset "fraud" is displayed (Table 2.1). Examine the printout to see whether SAS imported all the variables from the Excel worksheet correctly.
4. Examine the PROC CONTENTS display of all the variables in the SAS dataset called "fraud". Table 2.2 shows the characteristics of all numeric variables, and Table 2.3 shows the character variables.
5. Examine the simple descriptive statistics for all the numeric variables (Table 2.4). Note that the variables YEAR, WEEK, and DAY are treated as numeric. Total number of observations in the dataset is 923. Confirm that three observations in VOIDS and TRANSAC and two observations in NETSALES are missing in the Excel file. Also, examine the minimum and the maximum numbers for all the numeric variables and verify that no unusual or extreme values are present.
6. Examine the frequency information (Tables 2.5 to 2.7) for all the character variables. Make sure that character variable levels are entered consistently. SAS systems consider uppercase and lowercase data values differently. For example, *April*, *april*, and *APRIL* are considered different data values. The frequency information for MGR (manager on duty) indicated that managers mgr_a and mgr_e were on duty relatively fewer times than the other three managers (Table 2.8). This information should be considered in modeling.

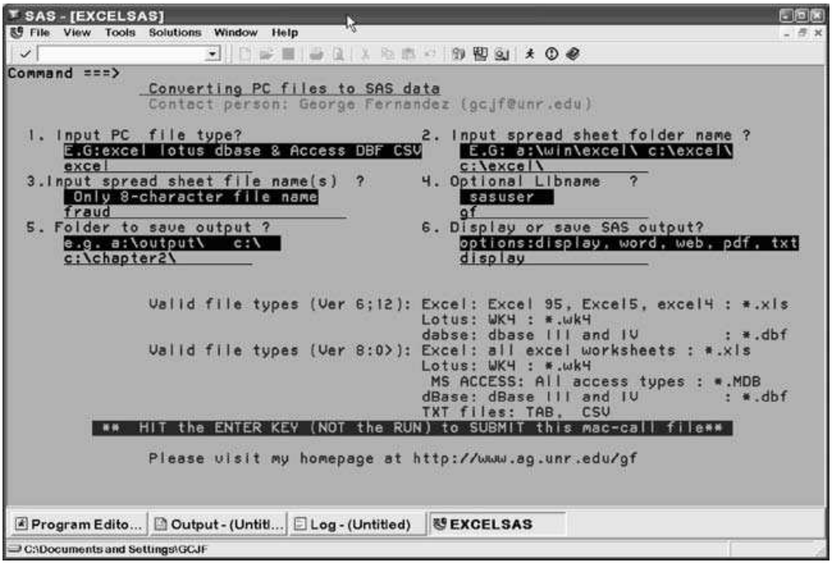


	A	B	C	D	E	F	G	H	I	J
1	YEAR	MONTH	WEEK	DAY	DOFWEEK	VOIDS	NETSALES	TRANSAC	MGR	
2	1998	January	1	2	Fri	1008.75	1443	139	mgr_a	
3	1998	January	1	3	Sat	10	1905	168	mgr_b	
4	1998	January	2	4	Sun	9	1223	134	mgr_b	
5	1998	January	2	5	Mon	7	1280	146	mgr_c	
6	1998	January	2	6	Tue	15	1243	129	mgr_b	
7	1998	January	2	7	Wed	14	871	135	mgr_a	
8	1998	January	2	8	Thu	4	1115	105	mgr_c	
9	1998	January	2	9	Fri	33.21	1080	109	mgr_c	
10	1998	January	2	10	Sat	8	1796	156	mgr_b	
11	1998	January	3	11	Sun	13	1328	132	mgr_c	
12	1998	January	3	12	Mon	5.5	1438	118	mgr_c	
13	1998	January	3	13	Tue	4	968	118	mgr_b	
14	1998	January	3	14	Wed	0	812	115	mgr_a	
15	1998	January	3	15	Thu	1	1026	132	mgr_c	
16	1998	January	3	16	Fri	2	1341	130	mgr_c	
17	1998	January	3	17	Sat	1016.5	1982	167	mgr_b	
18	1998	January	4	18	Sun	107	1596	159	mgr_b	
19	1998	January	4	19	Mon	23.25	1235	133	mgr_c	
20	1998	January	4	20	Tue	69	1104	128	mgr_b	
21	1998	January	4	21	Wed	32.5	708	122	mgr_a	
22	1998	January	4	22	Thu	31.5	933	120	mgr_c	
23	1998	January	4	23	Fri	0.25	1097	110	mgr_b	
24	1998	January	5	25	Sun	0	1550	130	mgr_b	
25	1998	January	5	26	Mon	29	1643	165	mgr_b	
26	1998	January	5	27	Tue	7	1080	128	mgr_b	
27	1998	January	5	28	Wed	0	845	125	mgr_a	
28	1998	January	5	29	Thu	90.75	1006	105	mgr_a	
29	1998	January	5	30	Fri	40	1167	127	mgr_b	
30	1998	January	5	31	Sat	0	1533	177	mgr_b	
31	1998	February	1	1	Sun	0	1986	179	mgr_b	
32	1998	February	1	2	Mon	19.5	975	118	mgr_c	

Figure 2.2 Screen copy of MS Excel 2000 worksheet "fraud.xls" opened in Office 2000; shows the required structure of the PC spreadsheet.

2.7.3 SAS Macro Applications: Random Sampling from the Entire Database Using the SAS Macro RANSPLIT

The RANSPLIT macro can be used to obtain TRAINING, VALIDATION, and TEST samples from the entire database. The SAS data step and the RANUNI function are the main tools in the RANSPLIT macro. The advantages of using the RANSPLIT macro are:


SAS - [EXCELSAS]
File View Tools Solutions Window Help
Program Edito... Output - (Untitl... Log - (Untitled) EXCELSAS
C:\Documents and Settings\GCJF

```
Command ===>
            Converting PC files to SAS data
            Contact person: George Fernandez (gcjf@unr.edu)

  1. Input PC  file type?                    2. Input spread sheet folder name ?
     E.G:excel lotus dbase & Access DBF CSV     E.G: a:\win\excel\ c:\excel\
     excel                                      c:\excel\
  3.Input spread sheet file name(s)  ?       4. Optional Libname   ?
     Only 8-character file name                 sasuser
     fraud                                      gf
  5. Folder to save output ?                 6. Display or save SAS output?
     e.g. a:\output\   c:\                      options:display, word, web, pdf, txt
     c:\chapter2\                               display


            Valid file types (Ver 6;12): Excel: Excel 95, Excel5, excel4 : *.xls
                                         Lotus: WK4 : *.wk4
                                         dabse: dbase III and IV        : *.dbf
            Valid file types (Ver 8:0>): Excel: all excel worksheets : *.xls
                                         Lotus: WK4 : *.wk4
                                          MS ACCESS: All access types : *.MDB
                                         dBase: dBase III and IV        : *.dbf
                                         TXT files: TAB,  CSV
         **  HIT the ENTER KEY (NOT the RUN) to SUBMIT this mac-call file**

            Please visit my homepage at http://www.ag.unr.edu/gf
```

Figure 2.3 Screen copy of EXCELTOSAS macro-call window showing the macro-call parameters required to import PC file types to SAS datasets.

- The distribution pattern among the TRAINING, VALIDATION, and TEST samples for user-specified numeric variables can be examined graphically by box plots to confirm that all three sample distributions are similar.
- A sample printout of the first 10 observations can be examined from the TRAINING sample.
- Options for saving the output tables and graphics in WORD, HTML, PDF, and TXT formats are available.

Software requirements for using the RANSPLIT macro include:

- SAS/CORE, SAS/BASE, and SAS/GRAPH must be licensed and installed at the site.
- SAS version 8.0 and above is recommended for full utilization.
- An active Internet connection is required for downloading the RANSPLIT macro from the book website if the companion CD-ROM is not available.

Table 2.1 Macro EXCELSAS: PROC PRINT Output, First 10 Observations

YEAR	*MONTH*	*WEEK*	*DAY*	*DOFWEEK*	*VOIDS*	*NETSALES*	*TRANSAC*	*MGR*
1998	January	1	2	Fri	1008.75	1443	139	mgr_a
1998	January	1	3	Sat	10.00	1905	168	mgr_b
1998	January	2	4	Sun	9.00	1223	134	mgr_b
1998	January	2	5	Mon	7.00	1280	146	mgr_c
1998	January	2	6	Tue	15.00	1243	129	mgr_b
1998	January	2	7	Wed	14.00	871	135	mgr_a
1998	January	2	8	Thu	4.00	1115	105	mgr_c
1998	January	2	9	Fri	33.21	1080	109	mgr_c
1998	January	2	10	Sat	8.00	1796	156	mgr_b
1998	January	3	11	Sun	13.00	1328	132	mgr_c

Table 2.2 Macro EXCELAS: PROC CONTENTS Output, Numeric Variable Description

Obs	*NAME*	*TYPE*	*LENGTH*	*VARNUM*	*LABEL*	*NPOS*	*NOBS*	*ENGINE*
1	DAY	1	8	4	DAY	16	923	V8
5	NETSALES	1	8	7	NETSALES	32	923	V8
6	TRANSAC	1	8	8	TRANSAC	40	923	V8
7	VOIDS	1	8	6	VOIDS	24	923	V8
8	WEEK	1	8	3	WEEK	8	923	V8
9	YEAR	1	8	1	YEAR	0	923	V8

Table 2.3 Macro EXCELSAS: PROC CONTENTS Output, Character Variable Descriptions

Obs	*NAME*	*TYPE*	*LENGTH*	*VARNUM*	*LABEL*	*FORMAT*	*NPOS*	*NOBS*	*ENGINE*
2	DOFWEEK	2	3	5	DOFWEEK	$	57	923	V8
3	MGR	2	5	9	MGR	$	60	923	V8
4	MONTH	2	9	2	MONTH	$	48	923	V8

Table 2.4 Macro EXCELSAS: PROC MEANS Output, Simple Statistics and Numeric Variables

Variable	*Label*	*N*	*Mean*	*Std Dev*	*Minimum*	*Maximum*
Year	YEAR	923	1998.88	0.7867336	1998.00	2000.00
Week	WEEK	923	3.0270856	1.3215726	1.0000000	6.0000000
Day	DAY	923	15.7941495	8.7590603	1.0000000	31.0000000
Voids	VOIDS	923	69.6595543	183.0534292	0	1752.45
Net sales	NETSALES	923	1324.33	471.5667690	7.0000000	4114.00
Transactions	TRANSAC	923	132.2576087	33.0792886	10.0000000	259.0000000

Table 2.5 Macro EXCELSAS: PROC FREQ Output, Frequency and Character Variable MONTH

MONTH	*Frequency*
April	89
August	88
December	57
February	83
January	83
July	87
June	88
March	92
May	88
November	53
October	58
September	57

Table 2.6 Macro EXCELSAS: PROC FREQ Output: Frequency and Character Variable DOFWEEK

DOFWEEK	*Frequency*
Fri	133
Mon	133
Sat	130
Sun	137
Thu	128
Tue	129
Wed	133

Table 2.7 Macro EXCELSAS: PROC FREQ Output, Frequency and Character Variable MGR

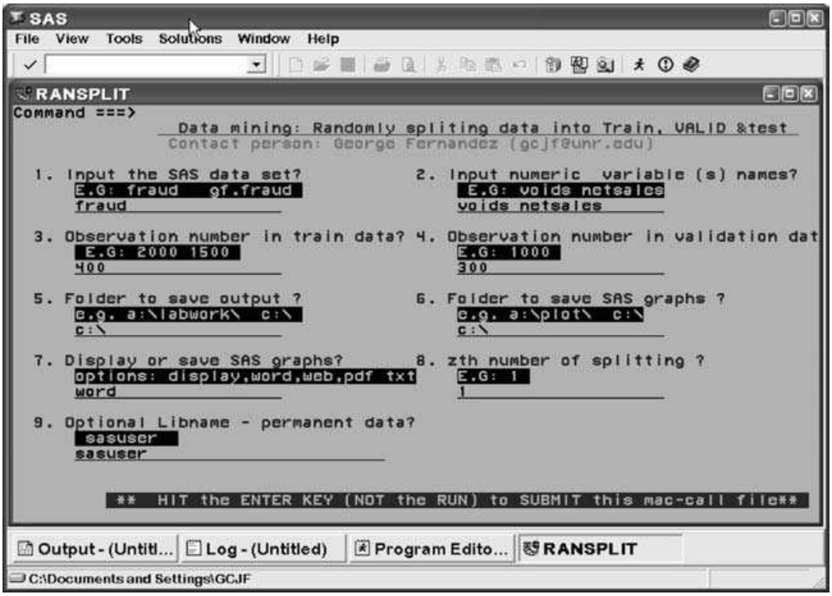
MGR	*Frequency*
mgr_a	38
mgr_b	204
mgr_c	258
mgr_d	408
mgr_e	15

Table 2.8 Macro RANSPLIT: PROC PRINT Output, First 10 Observations, Training Data

Obs	*YEAR*	*MONTH*	*WEEK*	*DAY*	*DOFWEEK*	*VOIDS*	*NETSALES*	*TRANSAC*	*MGR*
1	1999	May	4	17	Mon	45.00	1148.00	117	mgr_c
2	1998	December	2	7	Mon	12.50	1208.25	130	mgr_c
3	1998	July	3	15	Wed	0.00	930.25	89	mgr_d
4	2000	July	3	10	Mon	0.00	1900.97	163	mgr_b
5	1999	November	2	11	Thu	1601.50	785.00	113	mgr_d
6	1998	January	2	7	Wed	14.00	871.00	135	mgr_a
7	1999	August	2	8	Sun	16.00	1439.20	126	mgr_c
8	1999	February	4	25	Thu	4.75	751.50	83	mgr_d
9	1998	February	2	8	Sun	7.00	2103.00	187	mgr_b
10	2000	August	5	27	Sun	5.00	1329.94	121	mgr_b

2.7.3.1 Steps Involved in Running the RANSPLIT Macro

1. Prepare the SAS dataset (permanent or temporary) and examine the variables.
2. *If the companion CD-ROM is not available*, first verify that the Internet connection is active. Open the Ransplit.sas macro-call file in the SAS PROGRAM EDITOR window. The Appendix provides instructions for downloading the macro-call and sample data files from the book website. *If the companion CD-ROM is available*, the Ransplit.sas macro-call file can be found in the mac-call folder on the CD-ROM. Open the Ransplit.sas macro-call file in the SAS PROGRAM EDITOR window. Click the RUN icon to submit the macro-call file Ransplit.sas to open the macro-call window called RANSPLIT (Figure 2.4).
3. Input the appropriate parameters in the macro-call window by following the instructions provided in the RANSPLIT macro help file (Section 2.7.3.2). After inputting all the required macro parameters, be sure the cursor is in the last input field and that the RESULTS VIEWER window is closed, then hit the ENTER key (not the RUN icon) to submit the macro.



```
Command ===>
          Data mining: Randomly spliting data into Train, VALID &test
          Contact person: George Fernandez (gcjf@unr.edu)

 1. Input the SAS data set?                  2. Input numeric  variable (s) names?
    E.G: fraud    gf.fraud                      E.G: voids netsales
    fraud                                       voids netsales

 3. Observation number in train data?        4. Observation number in validation dat
    E.G: 2000 1500                              E.G: 1000
    400                                         300

 5. Folder to save output ?                  6. Folder to save SAS graphs ?
    e.g. a:\labwork\  c:\                       e.g. a:\plot\  c:\
    c:\                                         c:\

 7. Display or save SAS graphs?              8. zth number of splitting ?
    options: display,word,web,pdf txt           E.G: 1
    word                                        1

 9. Optional Libname - permanent data?
    sasuser
    sasuser

      **  HIT the ENTER KEY (NOT the RUN) to SUBMIT this mac-call file**
```

Figure 2.4 Screen copy of RANSPLIT macro-call window showing the macro-call parameters required to split the database into TRANING, VALIDATION, and TEST samples.

4. Examine the LOG window (only in DISPLAY mode) for any macro execution errors. If any errors appear in the LOG window, activate the PROGRAM EDITOR window, resubmit the Ransplit.sas macro-call file, check the macro input values, and correct any input errors.
5. Save the output files. If no errors are found in the LOG window, activate the PROGRAM EDITOR window, resubmit the Ransplit.sas macro-call file, and change the macro input value from DISPLAY to any other desirable format (see Section 2.7.3.2). If the sample size input for the validation sample is blank, a random sample with a user-specified sample size will be saved as TRAINING and the leftover observations in the database will be saved as VALIDATION datasets. If sample sizes are specified for both TRAINING and VALIDATION input, random samples with user-specified sample sizes will be saved as TRAINING and VALIDATION samples and the leftover observations will be saved as the TEST sample. The new SAS datasets will be saved as temporary (if macro input option #9 is blank or WORK) or permanent files (if a LIBNAME is specified in macro input option #9). The printout of the first 10 observations of the TRAINING SAS data and box plots illustrating distribution patterns among the TRAINING, VALIDATION, and TEST samples for user-specified numeric variables can be saved in a user-specified format in the user-specified folder.

2.7.3.2 Help File for SAS Macro RANSPLIT: Description of Macro Parameters

1. **Macro-call parameter:** Input the SAS data set (required parameter).
   Descriptions and explanation: Include the SAS dataset name, temporary or permanent (LIBNAME.sas_data_name) of the database you would like to draw samples from.
   Options/explanations:
   fraud (temporary SAS data called "fraud")
   gf.fraud (permanent SAS data called "fraud" saved in the predefined SAS library called "GF")
2. **Macro-call parameters:** Input numeric variable names (optional parameter).
   Descriptions and explanation: Input names of the numeric variables. Distribution aspects of the specified numeric variables are compared among different samples by box plots.
   Options/example: fraud net sales

3. **Macro-call parameter:** Input observation number in train data (required statement).

   Descriptions and explanation: Input the desired sample size number for the TRAINING data. Usually 40% of the database equivalent to 2000 observations is selected.

   Options/example: 2000 1400 400

4. **Macro-call parameter:** Observation number in validation data (optional parameter).

   Descriptions and explanation: Input the desired sample size number for the VALIDATION data. Usually 30% of the database equivalent to roughly 1000 observations is selected for validation. The leftover observations in the database after the TRAINING and VALIDATION samples are selected will be included in the TEST sample. If this field is left blank, all of the leftover observations in the database after the TRAINING sample is selected will be included in the VALIDATION set.

   Options/example: 1000 300

5. **Macro-call parameter:** Folder to save SAS output (optional statement).

   Descriptions and explanation: To save the SAS output files in a specific folder, input the full path of the folder. If this field is left blank, the output file will be saved in the default folder.

   Options/explanations:

   Possible values

   c:\output\ — folder named "OUTPUT"

   s:\george\ — folder named "George" in network drive S

   Be sure to include the back-slash at the end of the folder name.

6. **Macro-call parameter:** Folder to save SAS graphics (optional).

   Descriptions and explanation: To save the SAS graphics files in EMF format suitable for including in PowerPoint presentations, specify the output format as TXT in version 8.0 or later. In pre-8.0 versions, all graphic format files will be saved in a user-specified folder. If the graphics folder field is left blank, the graphics file will be saved in the default folder.

   Options/explanations:

   Possible values

   c:\output\ — folder named "OUTPUT"

7. **Macro-call parameter:** Display or save SAS output (required statement).

Descriptions and explanation: Option for displaying all output files in the OUTPUT window or save as a specific format in a folder specified in option #5.

Options/explanations:

Possible values

DISPLAY: Output will be displayed in the OUTPUT window. All SAS graphics will be displayed in the GRAPHICS window. System messages will be displayed in the LOG window.

WORD: Output and all SAS graphics will be saved together in the user-specified folder and will be displayed in the VIEWER window as a single RTF format file (version 8.0 and later) or saved only as a text file, and all graphics files in CGM format will be saved separately in a user-specified folder (macro input option #6) in pre-8.0 version SAS.

WEB: Output and graphics are saved in the user-specified folder and are viewed in the results VIEWER window as a single HTML file (version 8.1 and later) or saved only as a text file, and all graphics files in GIF format will be saved separately in a user-specified folder (macro input option #5) in pre-8.0 versions.

PDF: Output and graphics are saved in the user-specified folder and are viewed in the results VIEWER window as a single PDF file (version 8.2 and later) or saved only as a text file, and all graphics files in the PNG format will be saved separately in a user-specified folder (macro input option #6) in pre-8.2 versions.

TXT: Output will be saved as a TXT file in all SAS versions. No output will be displayed in the OUTPUT window. All graphic files will be saved in the EMF format in version 8.0 and later or CGM format in pre-8.0 versions in a user-specified folder (macro input option #6).

Note: System messages are deleted from the LOG window if DISPLAY is not selected as the input.

8. **Macro-call parameter:** *z*th number of run (required statement).

Descriptions and explanation: SAS output files will be saved by forming a file name from the original SAS dataset name and the counter number provided in macro input option #8. For example, if the original SAS dataset name is “fraud” and the counter number included is 1, the SAS output files will be saved as “fraud1.*” in the user-specified folder. By changing the counter numbers, users can avoid replacing the previous SAS output files with the new outputs.

Options/explanations: Numbers 1 to 10 and any letters are valid.

9. **Macro-call parameter:** Optional LIBNAME for creating permanent SAS data.

   Descriptions and explanation: To save the TRAINING, VALIDATION, and TEST datasets as permanent SAS datasets and input the preassigned library (LIBNAME) name. The predefined LIBNAME will tell SAS in which folder to save the permanent datasets. If this field is left blank, temporary WORK data files will be created for all samples.

   Options/example:

   SASUSER

   The permanent SAS dataset is saved in the library called SASUSER.

2.7.3.3 Drawing TRAINING (400), VALIDATION (300), and TEST (All Leftover Observations) Samples from the Permanent SAS Dataset Called "fraud"

Source file	Permanent SAS data set "fraud" located in the library "GF"
Variables	Daily retail sales, number of transactions, net sales, and manager on duty in a small convenience store
Number of observations	923

1. Open the RANSPLIT macro-call window in SAS (see Figure 2.4), input the appropriate macro input values by following the suggestions given in the help file in Section 2.7.3.2. Submit the RANSPLIT macro, and SAS will randomly split the entire database into three samples and save these TRAIN (400 observations), VALIDATION (300 observations), and TEST (leftover observations) as permanent SAS datasets in the LIBRARY called "GF".
2. The output file shows a list of the first 10 observations from the train dataset (Table 2.8). This dataset will be used in calibrating or training the models. Examine the contents and characteristics of the variables in the SAS data set called "fraud".
3. The distribution pattern among the TRAINING, VALIDATION, and TEST samples for one of the numeric variables NETSALES can be graphically examined by the box plot (Figure 2.5) created by the RANSPLIT SAS macro. A box plot shows the distribution pattern and the central tendency of the data. The line between the lowest adjacent limit and the bottom of the box represents one fourth of the data. One fourth of the data fall between the bottom of the

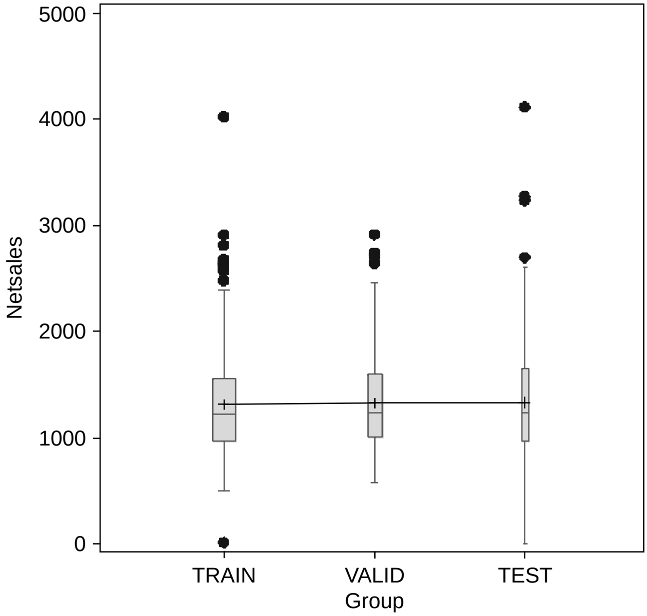


Figure 2.5 A box plot illustrating the distribution pattern among the TRAINING, VALIDATION, and TEST samples for the continuous variable NETSALES generated by running the SAS macro RANSPLIT.

box and the median, and another one fourth between the median and the top of the box. The line between the top of the box and the upper adjacent limit represents the final one fourth of the observations. For more information about interpreting the box plot, see Chapter 3. The box plot confirmed that the distribution showed a similar pattern for NETSALES among the TRAINING, VALIDATION, and TEST samples and confirmed that the random sampling was successful.

2.8 Summary

Data mining and knowledge discovery are driven by massive amounts of data. Business databases are growing at exponential rates because of the multitude of data that exist. Today, organizations are accumulating vast and growing amounts of data in various formats and in different databases. Dynamic data access is critical for data navigation applications, and the ability to store large databases is critical to data mining. The data may exist in a variety of formats such as relational databases, mainframe systems, or flat files; therefore, in data mining, it is common to work with data from several different sources. Roughly 70% of the time spent data mining is in preparing the data. The methods of extracting and preparing suitable data for data mining are covered in this chapter. Calibrating the prediction model using the TRAINING sample, validating the model using the VALIDATION sample, and fine-tuning the model using the TEST data are briefly addressed. The steps involved in applying the user-friendly SAS macro applications for importing PC worksheet files into SAS datasets and randomly splitting the entire database into TRAIN, VALIDATION, and TEST data are shown by using the example of a small business dataset called "fraud".

References

1. SAS Institute, Inc., *Data Mining and the Case for Sampling: Solving Business Problems Using SAS Enterprise Miner Software,* SAS Institute Best Practice paper, SAS Institute, Inc., Cary, NC (http://www.ag.unr.edu/gf/dm/sasdm.pdf).
2. SAS Institute, Inc., *Data Mining Using Enterprise Miner Software: A Case Study Approach*, 1st ed., SAS Institute, Cary, NC, 2000.
3. Delwiche, L.D. and Slaughter, S.J., *The Little SAS Book: A Primer*, 2nd ed., SAS Institute, Cary, NC, 1998.
4. SAS Institute, Inc., *SAS Sample for Extracting Data Using SQL Pass-Through Facility,* SAS Institute, Cary, NC (ftp://ftp.sas.com/techsup/download/sample/samp_lib/orlsampUsing_the_SQL_Passthru_Facility_html).

Suggested Reading

An, A. and Watts, D., *New SAS Procedures for Analysis of Sample Survey Data,* SAS Institute, Inc., Cary, NC (http://support.sas.com/rnd/app/papers/survey.pdf).

Michael, B.J.A. and Linoff, G., *Mastering Data Mining: The Art and Science of Customer Relationship Management*, John Wiley & Sons, New York, 2000, chap. 2.

Paules, M., Canete, P., and Yeh, S., Automatically converting data set specifications in Excel to a SAS program used to assign data set attributes: an approach to global data mart building process, in *Proc. SAS Users Group International (SUGI)25*, SAS Institute, Cary, NC, 2000 (http://www2.sas.com/proceedings/sugi25/25/po/25p215.pdf).

SAS Institute, Inc., *Getting Started with SQL Procedure Version 6*, 1st ed., SAS Institute, Cary, NC, 1994.

SAS Institute, Inc., *The Quality Data Warehouse: Serving the Analytical Needs of the Manufacturing Enterprise,* SAS Institute White Papers, SAS Institute, Cary, NC (http://www.datawarehouse.com/iknowledge/whitepapers/SAS_289.pdf).

Chapter 3

Exploratory Data Analysis

3.1 Introduction

The goal of exploratory data analysis (EDA) is to examine the underlying structure of the data and learn about the systematic relationships among many variables. EDA includes a set of descriptive and graphical tools for exploring data visually both as a prerequisite to more formal data analysis and as an integral part of formal model building. It facilitates discovering the unexpected, as well as confirming the expected. Although the two terms are used almost interchangeably, EDA is not identical to statistical graphical analysis. As an important step in data mining, EDA employs graphical and descriptive statistical techniques for studying a dataset, detecting outliers and anomalies, and testing the underlying model assumptions. Thus, thorough data exploration is an important prerequisite for any successful data mining project. For additional information on EDA, see Chambers et al.[1] and Cleveland and McGill.[2]

3.2 Exploring Continuous Variables

Simple descriptive statistics and exploratory graphics displaying the distribution pattern and the presence of outliers are useful in exploring continuous variables. Commonly used descriptive statistics and exploratory graphics suitable for analyzing continuous variables are described next.

3.2.1 *Descriptive Statistics*

Simple descriptive statistics of continuous variables are useful in summarizing central tendency, quantifying variability, detecting extreme outliers, and checking for distributional assumptions. The SAS procedures MEANS, SUMMARY, and UNIVARIATE provide a wide range of summary and exploratory statistics. For additional information on statistical theory, formulae, and computational details, readers should refer to Schlotzhauer and Littel[3] and SAS Institute.[4]

3.2.1.1 Measures of Location or Central Tendency

- **Arithmetic mean.** The most commonly used measure of central tendency, the mean is equal to the sum of the variable divided by the number of observations; however, mean can be heavily influenced by a few extreme values in the tails of a distribution.
- **Median.** The median is the mid-value of a ranked continuous variable and the number that separates the bottom 50% of the data from the top 50%; thus, half of the values in a sample will have values that are equal to or larger than the median, and half will have values that are equal to or smaller than the median. The median is less sensitive to extreme outliers than the mean; therefore, it is a better measure than the mean for highly skewed distributions. For example, the median salary is usually more informative than the mean salary when summarizing average salary. The mean value is higher than the median in positively skewed distributions and lower than the median in negatively skewed distributions.
- **Mode.** The most frequent observation in a distribution, mode is the most commonly used measure of central tendency with the nominal data.
- **Geometric mean.** The geometric mean is an appropriate measure of central tendency when averages of rates or index numbers are required. It is the *n*th root of the product of a positive variable. For example, to estimate the average rate of return of a 3-year investment that earns 10% the first year, 50% the second year, and 30% the third year, the geometric mean of these three rates should be used.
- **Harmonic mean.** Harmonic mean is the reciprocal of the average of the reciprocals. The harmonic mean of N positive numbers (x_1, x_2, …, x_n) is equal to $N/(1/x_1 + 1/x_2 + \ldots + 1/x_n)$. The harmonic mean is used to estimate the mean of sample sizes and rates. For example, when averaging rate of speed, which is measured by miles per hour, harmonic mean is the appropriate measure rather than arithmetic mean in averaging the rate.

3.2.1.2 Robust Measures of Location

- **Winsorized mean.** The Winsorized mean compensates for the presence of extreme values in the mean computation by setting the tail values equal to a certain percentile value. For example, when estimating a 95% Winsorized mean, the bottom 2.5% of the values are set equal to the value corresponding to the 2.5th percentile, while the upper 2.5% of the values are set equal to the value corresponding to the 97.5th percentile.
- **Trimmed mean.** The trimmed mean is calculated by excluding a given percentage of the lowest and highest values and then computing the mean of the remaining values. For example, by excluding the lower and upper 2.5% of the scores and taking the mean of the remaining scores, a 5% trimmed mean is computed. The median is considered as the mean trimmed 100% and the arithmetic mean is the mean trimmed 0%. A trimmed mean is not as affected by extreme outliers as an arithmetic mean. Trimmed means are commonly used in sports ratings to minimize the effects of extreme ratings possibly caused by biased judges.

3.2.1.3 Five-Number Summary Statistics

The five-number summary of a continuous variable consists of the minimum value, the first quartile, the median, the third quartile, and the maximum value. The median, or second quartile, is the mid-value of the sorted data. The first quartile is the 25th percentile and the third quartile is the 75th percentile of the sorted data. The range between the first and third quartiles includes half of the data. The difference between the third quartile and the first quartile is called the inter-quartile range (IQR). Thus, these five numbers display the full range of variation (from minimum to maximum), the common range of variation (from first to third quartile), and a typical value (the median).

3.2.1.4 Measures of Dispersion

- **Range.** Range is the difference between the maximum and minimum values. It is easy to compute because only two values, the minimum and maximum, are used in the estimation; however, a great deal of information is ignored, and the range is greatly influenced by outliers.
- **Variance.** Variance is the average measure of the variation. It is computed as the average of the square of the deviation from the average; however, because variance relies on the squared

differences of a continuous variable from the mean, a single outlier has greater impact on the size of the variance than does a single value near the mean.

- **Standard deviation.** Standard deviation is the square root of the variance. In a normal distribution, about 68% of the values fall within one standard deviation of the mean, and about 95% of the values fall within two standard deviations of the mean. Both variance and standard deviation measurements take into account the difference between each value and the mean. Consequently, these measures are based on a maximum amount of information.
- **Inter-quartile range.** The IQR is a robust measure of dispersion. It is the difference between the 75th percentile (Q3) and the 25th percentile (Q1). The IQR is hardly affected by extreme scores; therefore, it is a good measure of spread for skewed distributions. In normally distributed data, the IQR is approximately equal to 1.35 times the standard deviation.

3.2.1.5 Standard Errors and Confidence Interval Estimates

- **Standard error.** Standard error is the standard deviation of the sampling distribution of a given statistic. Standard errors show the amount of sampling fluctuation that exists in the estimated statistics in repeated sampling. Confidence interval estimation and statistical significance testing are dependent on the magnitude of the standard errors. The standard error of a statistic depends on the sample size. In general, the larger the sample size, the smaller the standard error.
- **Confidence interval.** The confidence interval is an interval estimate that quantifies the uncertainty caused by sampling error. It provides a range of values, which are likely to include an unknown population parameter, as the estimated range is being calculated from a given set of sample data. If independent samples are taken repeatedly from the same population, and a confidence interval is calculated for each sample, then a certain percentage of the intervals will include the unknown population parameter. The width of the confidence interval provides some idea about the uncertainty of the unknown parameter estimates. A very wide interval may indicate that more data must be collected before making inferences about the parameter.

3.2.1.6 Detecting Deviation from Normally Distributed Data

- **Skewness.** Skewness is a measure that quantifies the degree of asymmetry of a distribution. A distribution of a continuous variable is symmetric if it looks the same to the left and right of the center point. Data from positively skewed (skewed to the right) distributions have values that are clustered together below the mean but have a long tail above the mean. Data from negatively skewed (skewed to the left) distributions have values that are clustered together above the mean but have a long tail below the mean. The skewness estimate for a normal distribution equals zero. A negative skewness estimate indicates that the data are skewed left (the left tail is heavier than the right tail), and a positive skewness estimate indicates that the data are skewed right (the right tail is heavier than the left tail).

- **Kurtosis.** Kurtosis is a measure to quantify whether the data are peaked or flat relative to a normal distribution. Datasets with large kurtosis have a distinct peak near the mean, decline rather rapidly, and have heavy tails. Datasets with low kurtosis have a flat top near the mean rather than a sharp peak. Kurtosis can be both positive and negative. Distributions with positive kurtosis have typically heavy tails. Kurtosis and skewness estimates are very sensitive to the presence of outliers. These estimates may be influenced by a few extreme observations in the tails of the distribution; therefore, these statistics are not a robust measure of non-normality. The Shapiro–Wilks test[5] and the d'Agostino–Pearson omnibus test[6] are commonly used for detecting non-normal distributions.

3.2.2 Graphical Techniques Used in EDA of Continuous Data

Graphical techniques convert complex and messy information in large databases into meaningful displays; no quantitative analogs can give the same insight as well-chosen graphics in data exploration. The SAS/GRAPHICS procedures GCHART and GPLOT, SAS/BASE procedure UNIVARIATE, and SAS/QC procedure SHEWHART provide many types of graphical displays to explore continuous variables.[7] This section provides a brief description of some useful graphical techniques used in EDA of continuous data.

- **Frequency histogram.** The horizontal frequency histogram displays classes on the vertical axis and frequencies of the classes on the horizontal axis (see Figure 3.1 for an example of a histogram). The frequency of each class is represented by a horizontal bar that has a height equal to the frequency of that class.
- **Box plot.** A box plot provides an excellent visual summary of many important aspects of a distribution. The box plot is based on the

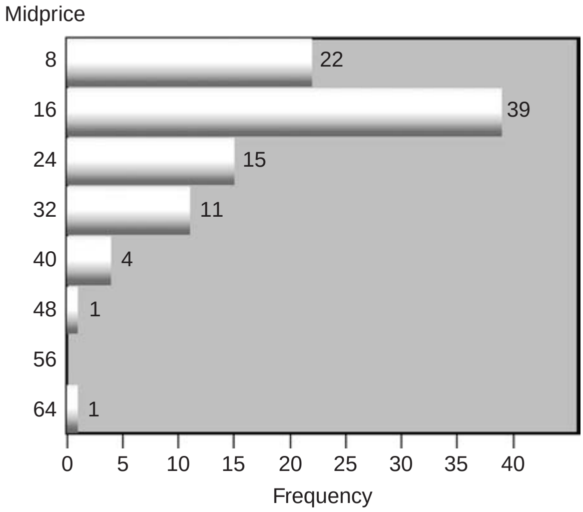


Figure 3.1 Frequency histogram illustrating the distribution pattern of car midprice. This graphic was generated by using the graphic device Activex when the WORD file type was selected in the SAS macro UNIVAR.

five-number summary plot, which is based on the median, quartiles, and extreme values. The box stretches from the lower hinge (first quartile) to the upper hinge (the third quartile) and therefore contains the middle half of the scores in the distribution. The median is shown as a line across the box (see Figure 3.2 for an example of a box plot). Therefore, one quarter of the distribution is between this line and the top of the box, and one quarter of the distribution is between this line and the bottom of the box. A box plot may be useful in detecting skewness to the right or to the left.

- **Normal probability plot.** The normal probability plot is a graphical technique for assessing whether or not a dataset is approximately normally distributed. The data are plotted against a theoretical normal distribution in such a way that the points should form an approximate straight line. Departures from this straight line indicate departures from normality. A normal probability plot, also known as a normal Q–Q plot (or normal quantile–quantile plot), is the plot of the ordered data values (*y* axis) against the associated quantiles of the normal distribution (*x* axis). For data from a normal distribution, the points of the plot should lie close to a straight line. Normal probability plots may also be useful in detecting skewness to the right or left (see Figure 3.3 for an example of a normal probability plot). If outliers are present, the normality test may reject the null

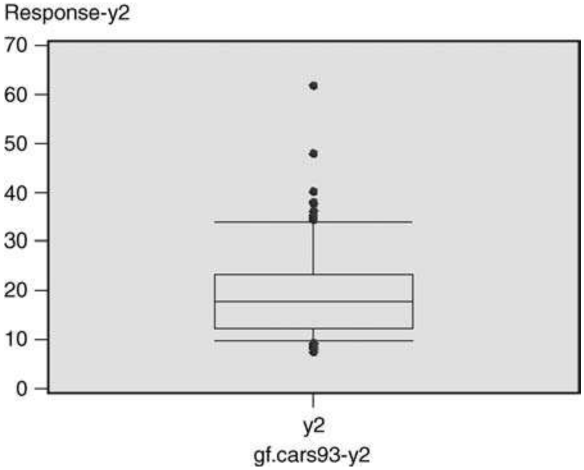


Figure 3.2 A box-plot display illustrating the five-number summary statistics of car mid-price. This graphic was generated by using the graphic device Activex when the WORD file type was selected in the SAS macro UNIVAR.

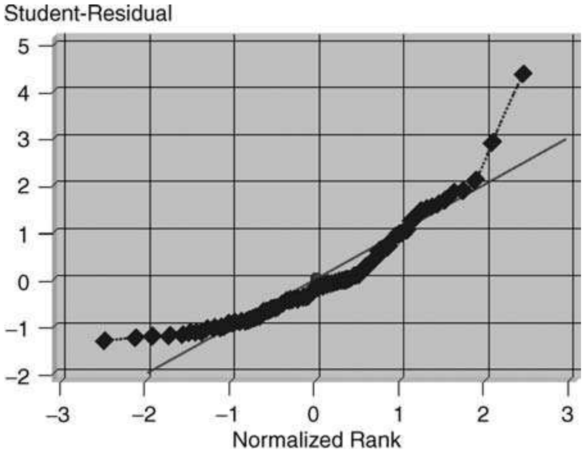


Figure 3.3 Normal probability display illustrating the right-skewed distribution of car mid-price. This graphic was generated by using the graphic device Activex when the WORD file type was selected in the SAS macro UNIVAR.

hypothesis that the distribution is normal even when the remainder of the data do in fact come from a normal distribution. Often, the effect of an assumption violation on the normality test result depends on the extent of the violation. Some small violations may have little practical effect on the analysis, while serious violations may render the normality test result incorrect or uninterpretable.

3.3 Data Exploration: Categorical Variables

One-way and multi-way frequency tables of categorical data are useful in summarizing group distributions and relationships between groups and for checking for rare events. The SAS procedure FREQ provides wide range of frequency tables and exploratory statistics. For additional information on statistical theory, formulae, and computational details, readers should refer to SAS Institute.[8]

3.3.1 Descriptive Statistical Estimates

- **Cross tabulation.** Cross tabulation uses a two-way table to show the frequencies for each level in one categorical variable across the levels of other categorical variables. One of the categorical variables is associated with the columns of the contingency table, and the other categorical variable is associated with the rows of the contingency table. This table is commonly used to display the correlation between two categorical variables.
- **Pearson's chi-square test for independence.** For a contingency table, Pearson's chi-square test for independence tests the null hypothesis that the row classification factor and the column classification factor are independent by comparing observed and expected frequencies. The expected frequencies are calculated by assuming that the null hypothesis is true. The chi-square test statistic is the sum of the squares of the differences between the observed and expected frequencies, with each squared difference being divided by the corresponding expected frequency.

3.3.2 Graphical Displays for Categorical Data

The graphical techniques employed in this chapter to display categorical data are quite simple, consisting of bar, block, and pie charts. The SAS/GRAPH procedure GCHART provides many types of graphical displays to explore categorical variables.[7] This section provides a brief description of some simple graphical techniques used in EDA of categorical data. For advanced methods in exploring categorical data, see Friendly.[9]

- **Bar charts.** Bar charts display a requested statistic based on the values of one or more variables. They are useful for displaying exact magnitudes emphasizing differences among the charted values and for comparing a number of discontinuous values against the same scale. Thus, bar charts allow us to see the differences between events, rather than trends. Stacked bar and block charts are effective in showing relationships between two-way and three-way tables. See Figures 3.4 and 3.5 for examples of stacked block and bar charts.
- **Pie charts.** Pie charts compare the levels or classes of a categorical variable to each other and to the whole. Sizes of the pie slices graphically represent the values of a statistic for a data range. Pie charts are useful for examining how the values of a variable contribute to the whole and for comparing the values of several variables. Donut charts, which are modified pie charts, are useful in displaying differences between groups in two-way data (see Figure 3.6 for a sample donut chart).

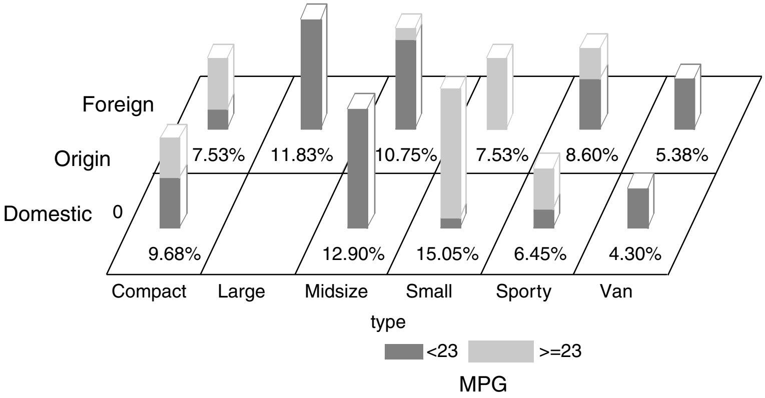


Figure 3.4 Stacked block chart illustrating the three-way relationship between car type, car origin, and the fuel efficiency (MPG). This graphic was generated by using the graphic device Activex when the WORD file type was selected in the SAS macro FREQ.

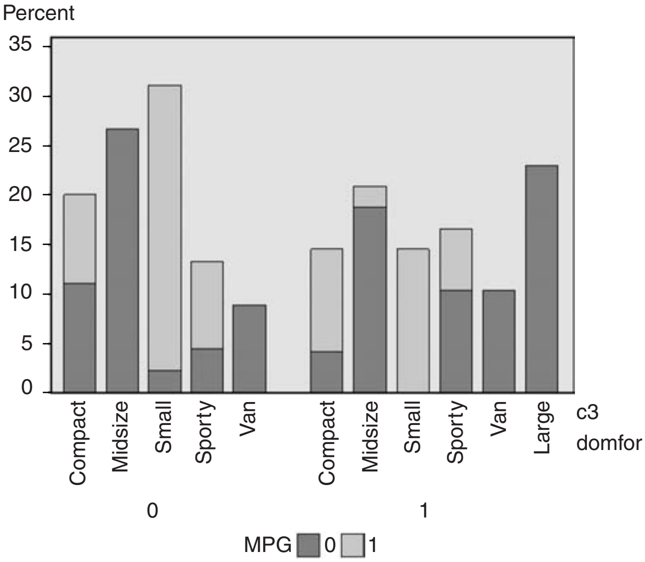


Figure 3.5 Stacked vertical bar chart illustrating the three-way relationship between car type, car origin, and the fuel efficiency (MPG). This graphic was generated by using the graphic device Activex when the WORD file type was selected in the SAS macro FREQ.

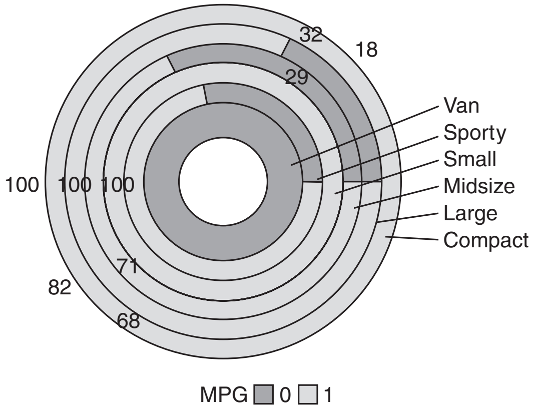


Figure 3.6 Donut chart illustrating the relationship between car type and the fuel efficiency (MPG). This graphic was generated by using the graphic device Activex when the WORD file type was selected in the SAS macro FREQ.

3.4 SAS Macro Applications Used in Data Exploration

SAS software has many statistical and graphical features for exploring numeric and categorical variables in large datasets. Some of the features are described in the following section. Readers are expected to have a basic knowledge in using SAS to perform the following operations. *The Little SAS Book*[10] can be used as an introductory SAS guide to become familiar with the SAS systems and SAS programming.

3.4.1 Exploring Categorical Variables Using the SAS Macro FREQ

The FREQ macro application is an enhanced version of SAS PROC FREQ with graphical capabilities for exploring categorical data. Since the release of SAS version 8.0, many additional statistical capabilities are available for data exploration in the PROC FREQ macro.[8] The advantages of using the FREQ SAS macro over PROC FREQ include:

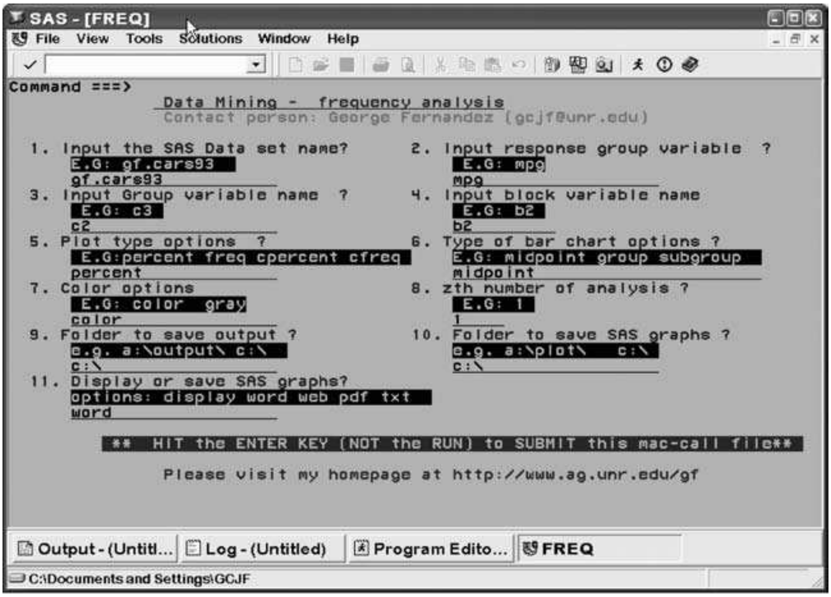
- Vertical, horizontal, block, and pie charts for exploring one-way and two-way frequency tables are automatically produced.
- Options for saving the output tables and graphics in WORD, HTML, PDF, and TXT formats are available.

Software requirements for using the FREQ macro include:

- SAS/CORE, SAS/BASE, and SAS/GRAPH must be licensed and installed at the site.
- The FREQ macro has only been tested in the Windows (Windows 98 and later) environment.
- SAS versions 8.0 and above are recommended for full utilization.
- An active Internet connection is required for downloading the FREQ macro from the book website if the companion CD-ROM is not available.

3.4.1.1 Steps Involved in Running the FREQ Macro

1. Create a temporary or permanent SAS data file.
2. *If the companion CD-ROM is not available*, first verify that the Internet connection is active. Open the FREQ.sas macro-call file in the SAS PROGRAM EDITOR window. Instructions are given in the Appendix regarding downloading the macro-call and sample data files from the book website. *If the companion CD-ROM is available*, you will find



SAS - [FREQ]
File View Tools Solutions Window Help
Output - (Untitl...
Log - (Untitled)
Program Edito...
FREQ
C:\Documents and Settings\GCJF

```
Command ===>
               Data Mining -  frequency analysis
               Contact person: George Fernandez (gcjf@unr.edu)

 1. Input the SAS Data set name?       2. Input response group variable  ?
    E.G: gf.cars93                         E.G: mpg
    gf.cars93                              mpg
 3. Input Group variable name  ?       4. Input block variable name
    E.G: c3                                E.G: b2
    c2                                     b2
 5. Plot type options  ?               6. Type of bar chart options ?
    E.G:percent freq cpercent cfreq        E.G: midpoint group subgroup
    percent                                midpoint
 7. Color options                      8. zth number of analysis ?
    E.G: color  gray                       E.G: 1
    color                                  1
 9. Folder to save output ?           10. Folder to save SAS graphs ?
    e.g. a:\output\ c:\                    e.g. a:\plot\   c:\
    c:\                                    c:\
11. Display or save SAS graphs?
    options: display word web pdf txt
    word

     **  HIT the ENTER KEY (NOT the RUN) to SUBMIT this mac-call file**

          Please visit my homepage at http://www.ag.unr.edu/gf
```

Figure 3.7 Screen copy of FREQ macro-call window showing the macro-call parameters required for exploring categorical variable.

the FREQ.sas macro-call file in the mac-call folder in the CD-ROM. Open the FREQ.sas macro-call file in the SAS PROGRAM EDITOR window. Click the RUN icon to submit the macro-call file FREQ.sas to open the macro window called FREQ (Figure 3.7).

3. Input the appropriate parameters in the macro-call window by following the instructions provided in the FREQ macro help file in Section 3.4.1.2. After inputting all the required macro parameters, be sure the cursor is in the last input field (#11) and the RESULTS VIEWER window is closed, then hit the ENTER key (not the RUN icon) to submit the macro.
4. Examine the LOG window only in the DISPLAY mode for any macro execution errors. If any errors in the LOG window are found, activate the PROGRAM EDITOR window, resubmit the FREQ.sas macro-call file, check the macro input values, and correct any input errors. Otherwise, activate the PROGRAM EDITOR window, resubmit the FREQ.sas macro-call file, and change the macro input (#11) value from DISPLAY to any other desirable format (see Section 3.4.1.2). The output, including exploratory graphics and frequency statistics, will be saved as the user-specified format in the user-specified folder as a single file for the file formats WORD, WEB, or PDF. If TXT is selected as the

file format in the #11 macro input field, SAS output and graphics files will be saved as separate files.

3.4.1.2 Help File for SAS Macro: FREQ, Description of Macro Parameters

1. **Macro-call parameter:** Input SAS dataset name (required parameter).
   Descriptions and explanation: Include the name of the temporary (member name) or permanent (libname.member_name) SAS dataset name on which the data exploration is to be performed.
   Options/examples:
   Permanent SAS dataset — gf.cars93 (LIBNAME: gf; SAS dataset name: cars93)
   Temporary SAS dataset — cars93 (SAS dataset name)
2. **Macro-call parameter:** Input response group variable name (required parameter).
   Descriptions and explanation: Input name of the categorical variables to be treated as the output variables in a two- or three-way analysis. For creating one-way tables and charts, however, input the categorical variable names and leave macro input fields #3 and #4 blank.
   Option/example:
   mpg (name of a target categorical variable)
3. **Macro-call parameter:** Input GROUP variable name (optional statement).
   Descriptions and explanation: Input the name of the first-level categorical variable for a two-way analysis.
   Option/example:
   c2
4. **Macro-call parameter:** Input BLOCK variable name (optional statement).
   Descriptions and explanation: Input the name of the second level categorical variable for a three-way analysis.
   Option/example:
   b2
5. **Macro-call parameter:** Plot type options (required statement).
   Descriptions and explanation: Select the type of frequency/percentage statistics desired in the charts.
   Options/explanations:
   Percent: report percentages
   Freq: report frequencies
   Cpercent: report cumulative percentages
   Cfreq: report cumulative frequencies

6. **Macro-call parameter:** Type of patterns used in bars (required statement).
   Descriptions and explanation: Select the pattern specifications in different bar charts.
   Options/explanations:
   Midpoint: Changes patterns when the midpoint value changes. If the GROUP= option is specified, the respective midpoint patterns are repeated for each group report percentage.
   Group: Changes patterns when the group variable changes. All bars within each group use the same pattern, but a different pattern is used for each group.
   Subgroup: Changes patterns when the value of the subgroup variable changes. The SUBGROUP= option must have been specified. Without SUBGROUP=, all bars will have the same pattern.
7. **Macro-call parameter:** Color options (required statement).
   Descriptions and explanation: Input whether color or black-and-white charts are required.
   Options/explanations:
   Color: preassigned colors used in charts
   Gray: preassigned gray shades used in charts
8. **Macro-call parameter:** *z*th number of run (required statement).
   Descriptions and explanation: SAS output files will be saved by forming a file name from the original SAS dataset name and the counter number provided in macro input field #8. For example, if the original SAS dataset "name" is "gf.cars93" and the counter number included is 1, the SAS output files will be saved as "gf.cars931.*" in the user-specified folder. By changing the counter numbers, the users can avoid replacing the previous SAS output files with the new outputs.
   Options/explanations: Any numbers or letters are valid.
9. **Macro-call parameter:** Folder to save SAS output (optional statement).
   Descriptions and explanation: To save the SAS output files in a specific folder, input the full path of the folder. The SAS dataset name will be assigned to the output file. If this field is left blank, the output file will be saved in the default folder.
   Options/explanations:
   Possible values
   c:\output\ — folder named "OUTPUT"
   s:\george\ — folder named "George" in mapped network drive S
   Be sure to include the back-slash at the end of the folder name.
10. **Macro-call parameter:** Folder to save SAS graphics (optional statement)

Descriptions and explanation: To save the SAS graphics files in the EMF format suitable for inclusion in PowerPoint presentations, specify the output format as TXT in version 8.0 or later. In pre-8.0 versions, all graphic format files will be saved in a user-specified folder. If the graphics folder field is left blank, the graphics file will be saved in the default folder.

Options/explanations:

Possible values

c:\output\ — folder named OUTPUT

11. **Macro-call parameter:** Display or save SAS output (required statement).

Descriptions and explanation: Option for displaying all output files in the OUTPUT window or saving files as a specific format in a folder specified in option #9.

Options/explanations:

Possible values

DISPLAY: Output will be displayed in the OUTPUT window. All SAS graphics will be displayed in the GRAPHICS window. System messages will be displayed in the LOG window.

WORD: Output and all SAS graphics will be saved together in the user-specified folder and will be displayed in the VIEWER window as a single RTF format file (version 8.0 and later). In pre-8.0 versions, SAS output will be saved as a text file, and all graphics files will be saved separately in the CGM format in a user-specified folder (macro input option #10).

WEB: Output and graphics are saved in the user-specified folder and are viewed in the results VIEWER window as a single HTML file (version 8.0 and later). In pre-8.0 versions, SAS output will be saved as a text file, and all graphics files will be saved separately in GIF format in a user-specified folder (macro input option #10).

PDF: Output and graphics are saved in the user-specified folder and are viewed in the results VIEWER window as a single PDF (version 8.2 and later) file. In pre-8.2 versions, SAS output will be saved as a text file, and all graphics files will be saved separately in the PNG format in a user-specified folder (macro input option #10).

TXT: Output will be saved as a TXT file in all SAS versions. No output will be displayed in the OUTPUT window. All graphic files will be saved in the EMF format in version 8.0 and later or CGM format in pre-8.0 versions in a user-specified folder (macro input option #10).

Note: System messages are deleted from the LOG window if DISPLAY is not selected as the input.

3.4.1.3 Case Study 1: Exploring Categorical Variables in a Permanent SAS Dataset gf.cars93

Source file	gf.cars93
Categorical variables	**MPG** (fuel efficiency: 0, below 26 mpg; 1, over 26 mpg) **b2** (origin of cars: 0, foreign; 1, American) **c3** (type of vehicle: compact, large, midsize, small, sporty, van)
Number of observations	93
Data source	Lock[11]

Open the FREQ macro-call window in SAS (Figure 3.7) and input the appropriate macro input values following the suggestions given in the help file (Section 3.4.1.2). Input MPG (miles per gallon) as the target categorical variable in macro input option #2. Input b2 (origin) as the group variable in macro input option #3. To account for the differences in car types, input c3 (car type) as the block variable in macro input option #4. After inputting other graphical and file saving parameters, submit the FREQ macro-call window, and SAS will output frequency statistics and exploratory charts for MPG categorical variables by car origin and car type. Only selected output and graphics generated by the FREQ macro are described below.

The one-way frequency and percentage statistics for car origin and car type are presented in Tables 3.1 and 3.2. Two-way percentage statistics for car type and MPG are illustrated in a donut chart in Figure 3.6. Table 3.3 is a two-way frequency table for car type × MPG group for foreign-made cars. The variation in frequency distribution by car type × car origin × MPG group is illustrated as a stacked block chart in Figure 3.4 and as a stacked vertical bar chart in Figure 3.5. No large car is found among the 44 foreign-made cars. Regardless of origin, a majority of the compact and small cars are more fuel efficient than the mid-size, sporty, large, and van-type vehicles.

Table 3.1 Macro FREQ: PROC FREQ Output, Frequency, and Percentage Values for Origin

Origin (b2)	*Frequency*	*Percent*
Foreign (0)	45	48.39
Domestic (1)	48	51.61

Table 3.2 Macro FREQ: PROC FREQ Output, Frequency, and Percentage Values for Car Type

Type (c3)	*Frequency*	*Percent*
Compact	16	17.20
Large	11	11.83
Midsize	22	23.66
Small	21	22.58
Sporty	14	15.05
Van	9	9.68

Table 3.3 Macro FREQ: PROC FREQ Output, Two-Way Frequency Table for Car Type × Miles per Gallon (MPG) for American-Made Cars (Origin = b2 = 1)

	Frequency (Percent)		
Type (c3)	*0*	*1*	*Total*
Compact	2 (4.17)	5 (10.42)	7 (14.58)
Large	11 (22.92)	0 (0.00)	11 (22.92)
Midsize	9 (18.75)	1 (2.08)	10 (20.83)
Small	0 (0.00)	7 (14.58)	7 (14.58)
Sporty	5 (10.42)	3 (6.25)	8 (16.67)
Van	5 (10.42)	0 (0.00)	5 (10.42)
Total	32 (66.67)	16 (33.33)	48 (100.00)

For the proportion of foreign-made cars, the 95% confidence intervals and exact confidence intervals are given in Table 3.4. The hypothesis test that the foreign-made car proportion in the database is not equal to 0.5 could not be rejected at the 5% level (*P* value 0.7557 in Table 3.5). The null hypothesis that car type and fuel efficiency (MPG) are independent is rejected at the 5% level based on chi-square test (*P* value <0.0001 in Table 3.6).

3.4.2 EDA Analysis of Continuous Variables Using SAS Macro UNIVAR

The UNIVAR macro is a powerful SAS application for exploring and visualizing continuous variables. SAS procedures UNIVARIATE, GCHART, GPLOT, and SHEWHART are the main tools utilized in the UNIVAR macro.

Table 3.4 Macro FREQ: PROC FREQ Output, 95% Confidence Intervals for Proportion of Foreign-Made Cars

Binomial Proportion for b2 = 0	
Proportion	0.4839
Asymptotic standard error (ASE)	0.0518
95% lower confidence limit	0.3823
95% upper confidence limit	0.5854
Exact confidence limits:	
95% lower confidence limit	0.3789
95% upper confidence limit	0.5899

Table 3.5 Macro FREQ: PROC FREQ Output, Hypothesis Testing That the Proportion of Foreign-Made Cars = 0.5

Test of H_0: Proportion = 0.5	
ASE under H_0	0.0518
Z-Statistic	– 0.3111
One-side P-value (Pr < *Z*)	0.3779
Two-side P-value (Pr > |*Z*|)	0.7557

Table 3.6 Macro FREQ, PROC FREQ Output, Hypothesis Testing That Car Type and Miles per Gallon (MPG) Are Independent Using a Chi-Square Test

Statistic	*Degrees of Freedom*	*Value*	*Probability*
Chi-square	5	41.0718	<.0001
Likelihood ratio chi-square	5	46.5097	<.0001
Mantel–Haenszel chi-square	1	15.8144	<.0001

The advantages of using the UNIVAR macro over the PROC UNIVARIATE include:

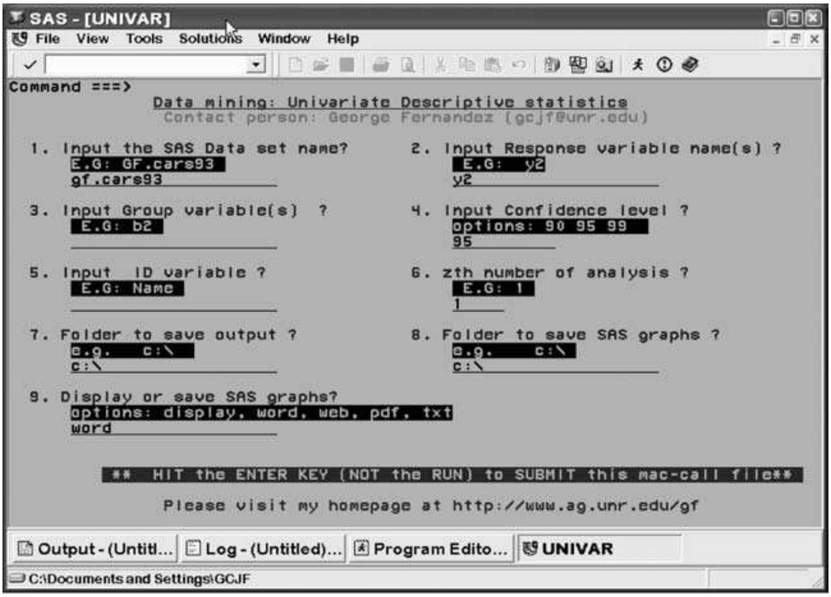
- Additional central tendency estimates such as geometric and harmonic means are reported.
- Test statistics and *P* values for testing significant skewness and kurtosis are reported.
- Outliers are detected based on (a) greater than 1.5 × IQR, and (b) Student residual value greater than 2.5 criteria. New datasets are also created after excluding the outliers based on the 1.5 × IQR criterion.
- Separate output and exploratory graphs for all observations, 5% trimmed data, and outlier excluded data are created for each specified continuous variable.
- Options for saving the output tables and graphics in WORD, HTML, PDF, and TXT formats are available.

Software requirements for using the UNIVAR macro include:

- SAS/CORE, SAS/BASE, SAS/GRAPH, and SAS/QC must be licensed and installed at the site.
- SAS versions 8.0 and above are recommended for full utilization.
- An active Internet connection is required for downloading the UNIVAR macro from the book website if the companion CD-ROM is not available.

3.4.2.1 Steps Involved in Running the UNIVAR Macro

1. Prepare the SAS dataset (permanent or temporary) and determine whether continuous variables exist.
2. *If the companion CD-ROM is not available*, first verify that the Internet connection is active. Open the UNIVAR.sas macro-call file in the SAS PROGRAM EDITOR window. Instructions are given in the Appendix regarding downloading the macro-call and sample data files from the book website. *If the companion CD-ROM is available*, the UNIVARs.sas macro-call file will found in the mac-call folder in the CD-ROM. Open the UNIVAR.sas macro-call file in the SAS PROGRAM EDITOR window. Click the RUN icon to submit the macro-call file UNIVAR.sas to open the macro–call window called UNIVAR (Figure 3.8).
3. Input the appropriate parameters in the macro-call window by following the instructions provided in the UNIVAR macro help file in Section 3.4.2.2. After inputting all the required macro parameters, be sure the cursor is in the last input field and the RESULTS VIEWER


SAS - [UNIVAR]
File View Tools Solutions Window Help
Output - (Untitl... Log - (Untitled)... Program Edito... UNIVAR
C:\Documents and Settings\GCJF

```
Command ===>
            Data mining: Univariate Descriptive statistics
             Contact person: George Fernandez (gcjf@unr.edu)

 1. Input the SAS Data set name?          2. Input Response variable name(s) ?
    E.G: GF.cars93                           E.G:  y2
    gf.cars93                                y2

 3. Input Group variable(s)  ?            4. Input Confidence level ?
    E.G: b2                                  options: 90 95 99
                                             95

 5. Input  ID variable ?                  6. zth number of analysis ?
    E.G: Name                                E.G: 1
                                             1

 7. Folder to save output ?               8. Folder to save SAS graphs ?
    e.g.   c:\                               e.g.    c:\
    c:\                                      c:\

 9. Display or save SAS graphs?
    options: display, word, web, pdf, txt
    word

      ** HIT the ENTER KEY (NOT the RUN) to SUBMIT this mac-call file**

           Please visit my homepage at http://www.ag.unr.edu/gf
```

Figure 3.8 Screen copy of UNIVAR macro call window showing the macro-call parameters required for exploring continuous variables.

window is closed, then hit the ENTER key (not the RUN icon) to submit the macro.

4. Examine the LOG window (only in DISPLAY mode) for any macro execution errors. If you see any errors in the LOG window, activate the PROGRAM EDITOR window, resubmit the UNIVAR.sas macro-call file, check the macro input values, and correct any input errors.
5. Save the output files. If no errors are found in the LOG window, activate the PROGRAM EDITOR window, resubmit the UNIVAR.sas macro-call file, and change the macro input value from DISPLAY to any other desirable format (see Section 3.4.2.2). The printout of all descriptive statistics and exploratory graphs can be saved under the user-specified format file in the user-specified folder.

3.4.2.2 Help File for SAS UNIVAR Macro: Description of Macro Parameters

1. **Macro-call parameter:** Input SAS dataset name (required parameter).
   Descriptions and explanation: Include the name of the temporary (member_name) or permanent (libname.member_name) SAS dataset on the EDA that will be performed.

Options/examples:
Permanent SAS dataset — gf.cars93 (LIBNAME: gf; SAS dataset name: cars93)
Temporary SAS dataset — cars93 (SAS dataset name)

2. **Macro-call parameter:** Input response variable names (required parameter).
   Descriptions and explanation: Input continuous variable names to explore. If multiple variable names are included, exploratory analysis will be performed, and a new dataset excluding the outliers will be created for each variable specified. For example, if you input two continuous variables Y1 and Y2 from a temporary SAS dataset, "cars93", and leave the group variable field (#3) blank, separate exploratory analysis will be performed on these two variables. Also, two new SAS datasets, "cars931" and "cars932", will be created after excluding the outliers.
   Option/example:
   Y2 (name of a continuous variable)
3. **Macro-call parameter:** Input GROUP variable names (optional statement).
   Descriptions and explanation: If you would like to perform data exploration by group variables, specify the group variable names in this field. If you include group variable names, exploratory analysis will be performed and separate datasets excluding the outliers will be created for each variable specified and for each level within the group variable. For example, if one continuous variable, Y2, is input from a temporary SAS dataset, "cars93", and a group variable name such as "b2" (origin with two levels) is input in macro input field #3, separate exploratory analyses will be performed for each level within a group variable. Also, two new SAS datasets, "cars9311" and "cars9312", will be created after excluding the outliers.
   Option/example:
   b2 (origin with two levels)
4. **Macro-call parameter:** Input confidence levels (required statement).
   Descriptions and explanation: Input the level of confidence you would like to assign to your parameter estimates.
   Options/example:
   Available options: 90, 95, 99
5. **Macro-call parameter:** Input ID variable (optional statement).
   Descriptions and explanation: Input the name of the variable to be treated as the ID. If you leave this field blank, a character variable will be created from the observational number and will be treated as the ID variable.

Option/example:
Car ID model

6. **Macro-call parameter:** *z*th number of run (required statement).
   Descriptions and explanation: SAS output files will be saved by forming a file name from the original SAS dataset name and the counter value provided in this field. For example, if the original SAS dataset name is "gf.cars93" and the counter number included is 1, the SAS output files will be saved as "gf.cars931.*" in the user-specified folder. By changing the counter value, users can avoid replacing the previous SAS output files with the new outputs.
   Options/examples:
   Any numbers or letters are valid.
7. **Macro-call parameter:** Folder to save SAS output (optional statement).
   Descriptions and explanation: To save the SAS output files in a specific folder, input the full path of the folder. The SAS dataset name will be assigned to the output file. If this field is left blank, the output file will be saved in the default folder.
   Options/examples:
   Possible values
   c:\output\ — folder named OUTPUT
   s:\george\ — folder named "George" in network drive S
   Be sure to include the back-slash at the end of the folder name.
8. **Macro-call parameter:** Folder to save SAS graphics? (Optional statement)
   Descriptions and explanation: To save the SAS graphics files in EMF format suitable for inclusion in PowerPoint presentations, specify the output format as TXT in Version 8.0 or later. In pre-8.0 versions, all graphic format files will be saved in a user-specified folder. If the graphics folder field is left blank, the graphics file will be saved in the default folder.
   Options/example:
   Possible values
   c:\output\ — folder named OUTPUT
9. **Macro-call parameter:** Display or save SAS output (required statement).
   Descriptions and explanation: Option for displaying all output files in the OUTPUT window or saving files as a specific format in a folder specified in option #7.
   Options/examples:
   Possible values

DISPLAY: Output will be displayed in the OUTPUT window. All SAS graphics will be displayed in the GRAPHICS window. System messages will be displayed in the LOG window.

WORD: Output and all SAS graphics will be saved together in the user-specified folder and will be displayed in the VIEWER window as a single RTF format file (version 8.0 and later). In pre-8.0 versions, SAS output will be saved as a text file, and all graphics files will be saved separately in CGM format in a user-specified folder (macro input option #8).

WEB: Output and graphics are saved in the user-specified folder and are viewed in the results VIEWER window as a single HTML file (version 8.0 and later). In pre-8.0 versions, SAS output will be saved as a text file, and all graphics files will be saved separately in GIF format in a user-specified folder (macro input option #8).

PDF: Output and graphics are saved in the user-specified folder and are viewed in the results VIEWER window as a single PDF (version 8.2 and later) file. In pre-8.2 versions, SAS output will be saved as a text file, and all graphics files will be saved separately in the PNG format in a user-specified folder (macro input option #8).

TXT: Output will be saved as a TXT file in all SAS versions. No output will be displayed in the OUTPUT window. All graphic files will be saved in the EMF format (version 8.0 and later) or CGM format (pre-8.0 version) in a user-specified folder (macro input option #8).

Note: System messages are deleted from the LOG window if DISPLAY is not selected as the input.

3.4.2.3 Case Study 2: Data Exploration and Continuous Variables

Data name	Permanent SAS dataset "CARS93" located in the library "GF"
Continuous variables	Y2 (mid-price)
Number of observations	93
Data source	Lock[11]

Open the UNIVAR macro call window in SAS (Figure 3.8) and input the appropriate macro-input values by following the suggestions given in the help file (Section 3.4.2.2). Input Y2 (mid-price) as the response variable (option #2). Leave the #3 (group variable) field blank because data exploration on mid-price is performed for the entire dataset. Submit the

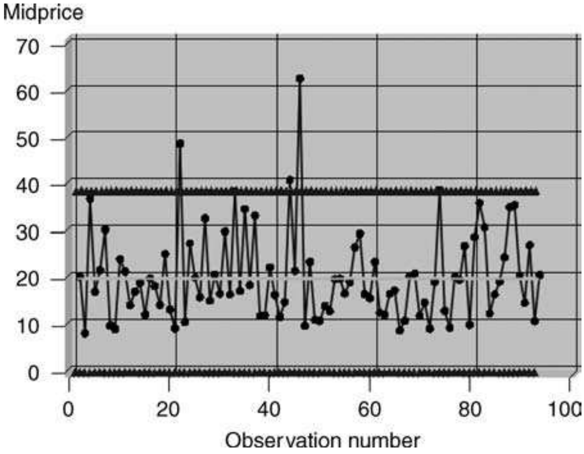


Figure 3.9 Control chart illustrating the variation in car mid-price. This graphic was generated by using the graphic device Activex when the WORD file type was selected in the SAS macro UNIVAR.

UNIVAR macro, and SAS will output descriptive statistics and exploratory graphs and create a new dataset by excluding the outliers. Only selected output and graphics generated by the UNIVAR macro are described below.

The histogram of the mid-price value clearly shows a right-skewed distribution (Figure 3.1). More than 50% of the mid-prices are below $18,000. The variation in mid-price is clearly illustrated by the control chart (Figure 3.9). The control chart displays the variation in continuous response variables by the observation number. The mean and mean ± 2 standard deviation boundary lines are useful in detecting extreme observations and their location in the dataset. The five different mean values for the mid-price of the 93 cars aregiven in Table 3.7. A large positive difference between (1) the mean and the median and (2) the mean and the geometric mean clearly shows the mid-price data are right skewed. Excluding a total of 5% extreme observations (4 observations out of 93) in both ends reduces the mean price from 19.50 to 18.97 (Table 3.8).

Table 3.7 Macro UNIVAR: PROC PRINT Output, Five Different Mean Statistics

Mean Y2	*Median Y2*	*Most Frequent Value Y2*	*Geometric Mean*	*Harmonic Mean*	*Sample Size*
19.5097	17.7	15.9	17.5621	15.9329	93

Table 3.8 Macro UNIVAR: PROC PRINT Output, Three Different Mean Statistics for 5% Trimmed Data

Mean Y2	*Median Y2*	*Most Frequent Value Y2*	*Sample Size Y2*
18.9798	17.7	15.9	89

However, the median price is not affected by deleting the extreme observations. The box-plot display illustrates the five-number summary statistics graphically and shows the presence of outliers (Figure 3.2). The five-number summary statistics for mid-prices before and after 5% trimming are given in Tables 3.9 and 3.10, respectively. The median mid-car price is 17.7 (thousand), and 50% of the mid-prices varies between 12.5 and 23.3 (thousand). Excluding the 5% extreme values in both ends reduces the range and the standard deviation but not the IQR of the mid-price (Tables 3.11 and 3.12). The 95% Winsorized mean computed after replacing the high extreme values with 97.5th percentile values and the low extreme values with 2.5th percentile values for mid-price are reported in Table 3.13. The Winsorized mean (19.05) value falls between the arithmetic mean (19.5) and the 5% trimmed mean (18.97).

Table 3.9 Macro UNIVAR: PROC PRINT Output, Five-Number Summary

Smallest Value Y2	*Lower Quartile Y2*	*Median Y2*	*Upper Quartile Y2*	*Largest Value Y2*
7.4	12.2	17.7	23.3	61.9

Table 3.10 Macro UNIVAR: PROC PRINT Output, Five-Number Summary for 5% Trimmed Data

Smallest Value Y2	*Lower Quartile Y2*	*Median Y2*	*Upper Quartile Y2*	*Largest Value Y2*
8.3	12.5	17.7	22.7	40.1

Confidence intervals (95%) for mean and dispersion estimates for the mid-price assuming normality are presented in Table 3.14. Confidence intervals assuming normality and distribution-free confidence intervals for all quantile values are reported in Table 3.15. Because the mid-price distribution is significantly right skewed, distribution-free confidence intervals provide more reliable information on the interval estimates.

Table 3.11 Macro UNIVAR: PROC PRINT Output, Measures of Dispersion

Range Y2	*Inter-Quartile Range Y2*	*Standard Deviation Y2*
54.5	11.1	9.65943

Table 3.12 Macro UNIVAR: PROC PRINT Output, Measures of Dispersion for 5% Trimmed Data

Range Y2	*Inter-Quartile Range Y2*	*Standard Deviation Y2*
31.8	10.2	8.03187

Table 3.13 Macro UNIVAR: Winsorized Mean and Related Statistics

Percent Winsorized in Tail	*Number Winsorized in Tail*	*Winsorized Mean*	*Standard Error Winsorized Mean*	*95% Confidence Limits*	
				Lower	*Upper*
5.38	5	19.05161	0.953873	17.15406	20.94917

Table 3.14 Macro UNIVAR: Confidence Interval (95%) Estimates Assuming Normality

		95% Confidence Limits	
Parameter	*Estimate*	*Lower*	*Upper*
Mean	19.50968	17.52034	21.49901
Standard deviation	9.65943	8.44274	11.28908
Variance	93.30458	71.27983	127.44339

Table 3.15 Macro UNIVAR: Confidence Intervals (95%) for Quantile Values

		95% Confidence Limits			
		Assuming Normality		*Distribution Free*	
Quantile (%)	*Estimate*	*Lower*	*Upper*	*Lower*	*Upper*
100 (max)	61.9	—	—	—	—
99	61.9	38.61584	46.33797	40.1	61.9
95	37.7	32.68462	38.82574	33.9	61.9
90	33.9	29.47349	34.86857	28.0	37.7
75 (Q3)	23.3	23.96988	28.38610	19.9	28.7
50 (median)	17.7	17.52034	21.49901	15.8	19.5
25 (Q1)	12.2	10.63325	15.04948	11.1	14.9
10	9.8	4.15078	9.54587	8.4	11.1
5	8.4	0.19361	6.33473	7.4	9.8
1	7.4	–7.31862	0.40351	7.4	8.3
0 (min)	7.4	—	—	—	—

Table 3.16 Macro UNIVAR: Extreme Observations Based on IQR Criteria

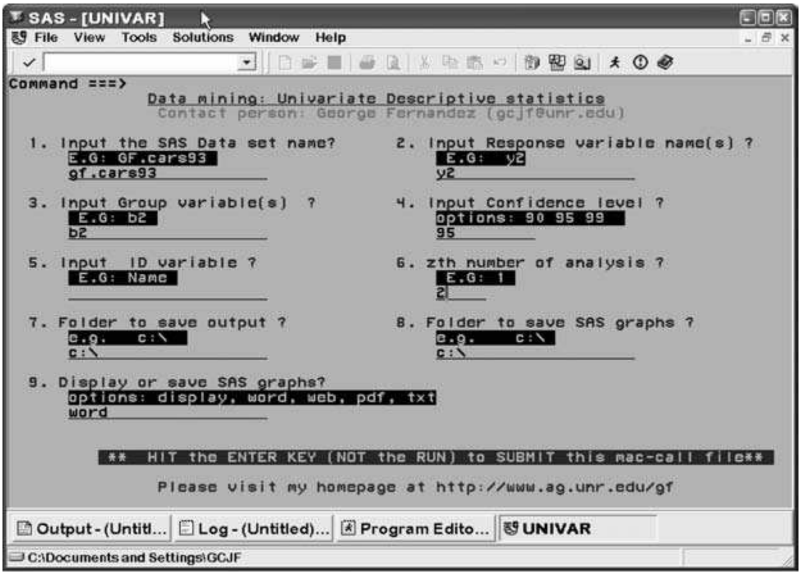
ID	*Y2 Mid-Price*
21	47.9
43	40.1
45	61.9

Table 3.17 Macro UNIVAR: Extreme Observations Based on Studentized Residual

ID	*Y2 Mid-Price*	*Student Residual*	*Outlier*
21	47.9	2.95506	*
45	61.9	4.41228	****

* Outlier.
**** Extreme outlier.

Observation numbers 45 and 21 are identified as large extreme values based on both IQR and Student residual statistics (Tables 3.16 and 3.17). In addition, observation number 43 is also identified as extreme based on IQR criterion (Table 3.16). The presence of large extreme values for



Command ===>

Data mining: Univariate Descriptive statistics
Contact person: George Fernandez (gcjf@unr.edu)

1. Input the SAS Data set name?
E.G: GF.cars93
gf.cars93

2. Input Response variable name(s) ?
E.G: y2
y2

3. Input Group variable(s) ?
E.G: b2
b2

4. Input Confidence level ?
options: 90 95 99
95

5. Input ID variable ?
E.G: Name

6. zth number of analysis ?
E.G: 1
2

7. Folder to save output ?
e.g. c:\
c:\

8. Folder to save SAS graphs ?
e.g. c:\
c:\

9. Display or save SAS graphs?
options: display, word, web, pdf, txt
word

** HIT the ENTER KEY (NOT the RUN) to SUBMIT this mac-call file**

Please visit my homepage at http://www.ag.unr.edu/gf

Figure 3.10 Screen copy of UNIVAR macro-call window showing the macro-call parameters required for exploring continuous variables by a group variable.

mid-price is the main cause for significant skewness and kurtosis and confirms that the distribution is not normal based on the d'Agostino–Pearson omnibus normality test (Table 3.18). The upward curvature in the normal probability plot of the Student residual (standardized deviation from the mean) value confirms that the distribution is right skewed (Figure 3.3).

3.4.2.4 Case Study 3: Exploring Continuous Data by a Group Variable

Data name	Permanent SAS dataset "CARS93" located in the library "GF"
Continuous variables	Y2 mid-price
Number of observations	93
Group variable	b2 (Origin of cars: 0, foreign; 1, American)
Data source	Lock[11]

Open the UNIVAR macro-call window in SAS (Figure 3.10) and input the appropriate macro input values by following the suggestions given in the help file in Section 3.4.2.2. Input the Y2 (mid-price) as the response

Table 3.18 Macro UNIVAR: Test Statistics Checking for Normal Distribution

Skewness	*P Value for Skewness*	*Kurtosis Statistic*	*P Value for Kurtosis*	*Chi-Square Value for Kurtosis*	*P Value, d'Agostino–Pearson Omnibus Normality Test*
1.50824	0.000000945	6.18369	0.000459508	36.3105	1.304e-8

variable in macro-call field #2. Because data exploration on mid-price is performed by the origin of vehicle, input b2 (origin) as the group variable in macro-call field #3. Submit the UNIVAR macro, and SAS will generate descriptive statistics and exploratory graphs and create new SAS datasets by excluding the outliers by the group variable origin. Only the features of selected output and graphics are described below.

Approximately 50% of the 93 cars in the dataset are foreign made. The mean, median, and geometric mean estimates for the foreign cars are greater than those of domestic cars (Table 3.19). The five-number summary statistics, variation, and presence of outliers for both domestic and foreign cars are illustrated in a box-plot display (Figure 3.11).The observed lowest price is 7.4 (thousand) for an American-made car and the highest car price is 61.9 (thousand) for a foreign-made car (Table 3.20). The median price of foreign cars is approximately 3000 more than the median price of domestic cars.

The variation in mid-priced cars for foreign cars is greater than the variation for American cars as illustrated by the combined control chart and box plots in Figure 3.12 and by the range, IQR, and standard deviation estimates (Table 3.21). One observation is identified as the outlier in both foreign and domestic car types by the 3-sigma cut-off value in the individual control charts. However, many extreme values are detected based on individual IQR criterion in the box plots (Figure 3.11). Observation number 23, the most expensive car in the database, is identified as an outlier among the foreign cars based on 2.5-Student's and 1.5-IQR criteria. Among the domestic cars, two observations are identified as outliers based on the 2.5-Student's criterion (Table 3.22). Because the IQR value for the domestic cars is relatively smaller than the value for the foreign cars, the 1.5-IQR criterion detects many observations as outliers (Table 3.23). The distributional pattern differences between the foreign and the domestic car types are illustrated in Figure 3.13. The normal probability plots of Student's residual for both foreign and domestic cars clearly indicate the nature of right-skewed distribution for mid-price (Figure 3.14). Other statistics, such as confidence interval estimates, trimmed mean statistics, Winsorized mean estimates, extreme outliers, and normality check statistics generated by the l UNIVAR macro are not shown.

Table 3.19 Macro UNIVAR: PROC PRINT Output, Different Mean Statistics for Mid-Price by Origin

Mean Y2	*Median Y2*	*Most Frequent Value Y2*	*Geometric Mean*	*Harmonic Mean*	*Sample Size*
Origin (b2): Domestic (1)					
18.5729	16.3	11.1	17.1994	16.0164	48
Origin (b2): Foreign (0)					
20.5089	19.1	8.4	17.9573	15.8448	45

Table 3.20 Macro UNIVAR: PROC PRINT Output, Five-Number Summary Statistics for Mid-Price by Origin

Smallest Value Y2	*Lower Quartile Y2*	*Median Y2*	*Upper QuartileY2*	*Largest Value Y2*
Origin (b2): Foreign (0)				
8	11.6	19.1	26.7	61.9
Origin (b2): Domestic (1)				
7.4	13.45	16.3	20.75	40.1

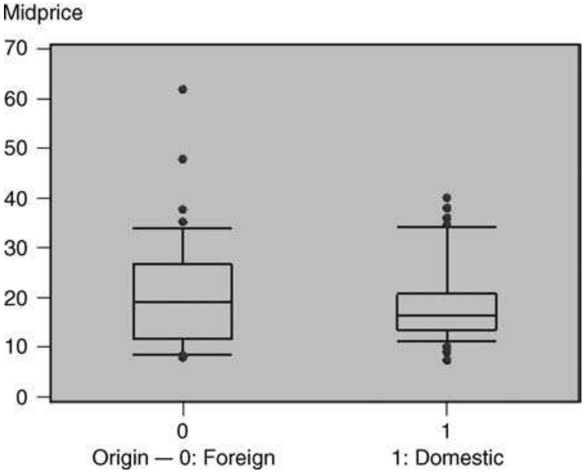


Figure 3.11 A box-plot display comparing the five-number summary statistics of car mid-price by car origin. This graphic was generated by using the graphic device Activex when the WORD file type was selected in the SAS macro UNIVAR.

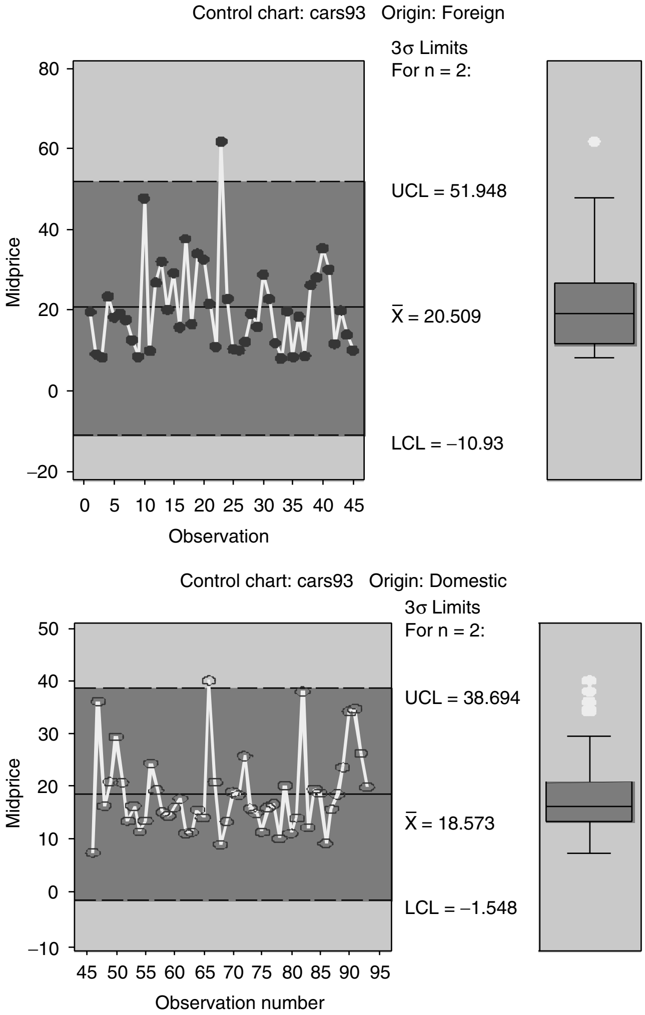


Figure 3.12 A combination of control chart and box-plot display comparing the variation and five-number summary statistics of car mid-price by car origin. This graphic was generated by using the graphic device Activex when the WORD file type was selected in the SAS macro UNIVAR.

Table 3.21 Macro UNIVAR: PROC PRINT Output, Estimates on Dispersion Statistics for Mid-Price by Origin

Range Y2	*Inter-Quartile Range Y2*	*Standard Deviation Y2*
Origin (b2): Foreign (0)		
53.9	15.1	11.3068
Origin (b2): Domestic (1)		
32.7	7.3	7.81691

Table 3.22 Macro UNIVAR: PROC PRINT Output, List of Outliers for Mid-Price by Origin

ID	*Y2*	*Student's t*	*Outlier*
Origin (b2): Foreign (0)			
23	61.9	3.70211	***
Origin (b2): Domestic (1)			
82	38	2.51156	*
66	40.1	2.78305	*

* Outlier.
*** Extreme outlier.

Table 3.23 Macro UNIVAR: PROC PRINT Output, List of Outliers Based on 1.5-IQR for Mid-Price in Domestic Cars

ID	*Mid-Price*
47	36.1
66	40.1
82	38
90	34.3
91	34.7

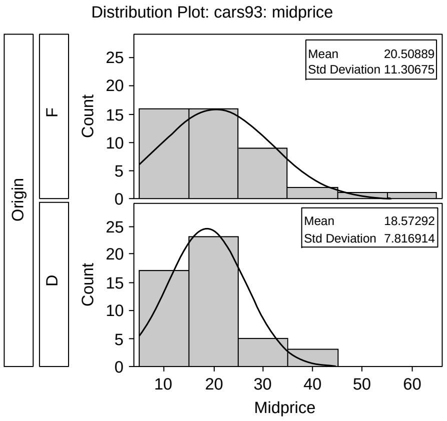


Figure 3.13 Display of frequency histograms comparing the right-skewed distribution of car mid-price by car origin. This graphic was generated by using the graphic device Activex when the WORD file type was selected in the SAS macro UNIVAR.

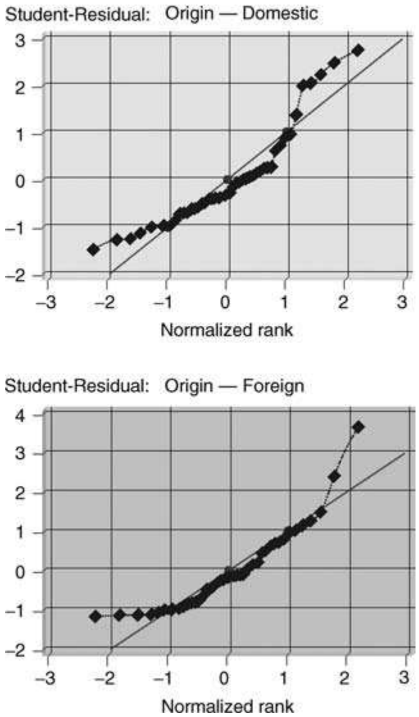


Figure 3.14 Display of normal probability plots comparing the right-skewed distribution of car mid-price by car origin. This graphic was generated by using the graphic device Activex when the WORD file type was selected in the SAS macro UNIVAR.

3.5 Summary

The methods of exploring continuous and categorical data using SAS macro applications are covered in this chapter. Both descriptive summary statistics and graphical analysis are used to explore continuous variables to learn about the data structure, detect outliers, and test the distributional assumptions. Frequency analysis and multi-way graphical charts are used to explore the relationshipst between categorical variables. Steps involved in using the user-friendly SAS macro applications FREQ and UNIVAR for exploring categorical and continuous variables are shown by using the "cars93" dataset.

References

1. Chambers, J.M., Cleveland, W.S., Kleiner, B., and Tukey, P.A., *Graphical Methods for Data Analysis,* Duxbury Press, Boston, MA, 1983.
2. Cleveland, W. and McGill, R., Graphical perception and graphical methods for analyzing scientific data, *Science*, 229, 828–833, 1985.
3. Schlotzhauer, S.D. and Littel, R.C., *SAS System for Elementary Statistical Analysis*, SAS Institute, Inc., Cary, NC, 1990.
4. SAS Institute, Inc., *Statistical Computation: UNIVARIATE Procedure Version 8*, online documentation, SAS Institute, Inc., Cary, NC (http://v8doc.sas.com/sashtml/proc/zte-comp.htm-z0296771; accessed March 2002).
5. Shapiro, S.S. and Wilk, M.B., An analysis of variance test for normality (complete samples), *Biometrika*, 52(3), 591–611, 1965.
6. d'Agostino, R.B., Belanger, A., and d'Agostino, R.B., Jr., A suggestion for using powerful and informative tests of Normality, *Am. Statistician*, 44, 316–321, 1990.
7. SAS Institute, Inc., Online documentation for SAS version 8.2, SAS Institute, Inc., Cary, NC (http://v8doc.sas.com/sashtml/; accessed March 2002).
8. SAS Institute, Inc., *Statistical Computation: FREQ Procedure Version 8*, online documentation, SAS Institute, Inc., Cary, NC (http://v8doc.sas.com/sashtml/proc/zeq-comp.htm; accessed March 2002).
9. Friendly, M., Visualizing Categorical Data: Data, Stories, and Pictures, paper presented at the SAS Users Group International 25th Annual Conference (http://www.math.yorku.ca/SCS/vcd/vcdstory.pdf; accessed March 2002).
10. Delwiche, L.D. and Slaughter, S.J., *The Little SAS Book: A Primer*, 2nd ed., SAS Institute, Inc., Cary, NC, 1998.
11. Lock, R.H., New car data, *J. Statistics Educ.*, 1(1), 1993 (http://www.amstat.org/publications/jse/v1n1/datasets.lock.html; accessed March 2002).

Suggested Reading

Curtis, N.E., *Are Histograms Giving You Fits? New SAS Software for Analyzing Distributions*, SAS Institute, Inc., Cary, NC (http://www.sas.com/rnd/app/papers/distributionanalysis.pdf; accessed March 2002).

Leiby, B.E., Saving paper and making it pretty: a SAS macro for simplifying and enhancing preliminary analysis of continous data, in *North East SAS Users Group (NESUG) Proc.* (http://www.pace.edu/nesug/proceedings/nesug00/ps/Ps7013.pdf; accessed March 2002).

Stokes, M.E., *Recent Advances in Categorical Data Analysis*, SAS Institute, Inc., Cary, NC (http://support.sas.com/rnd/app/papers/categorical.pdf; accessed March 2002).

Chapter 4

Unsupervised Learning Methods

4.1 Introduction

Analysis of multivariate data plays a key role in data mining and knowledge discovery. Multivariate data consists of many different attributes or variables recorded for each observation. If there are *p* variables in a database, each variable could be regarded as constituting a different dimension, in a *p*-dimensional hyperspace. This multidimensional hyperspace is often difficult to visualize; thus, the main objectives of unsupervised learning methods are to reduce dimensionality, scoring all observations based on a composite index and clustering similar observations together based on multiple attributes. Also, summarizing multivariate attributes by two or three that can be displayed graphically with minimal loss of information is useful in knowledge discovery.

The main difference between supervised and unsupervised learning methods is the underlying model structure. In supervised learning, relationships between input and the target variables are being established, but in unsupervised learning no variable is defined as a target or response variable. In fact, for most types of unsupervised learning, the targets are the same as the inputs. All the variables are assumed to be influenced by a few hidden factors or latent variables in unsupervised learning. Because of this feature, it is better to study large complex models with unsupervised learning than with supervised learning.

Unsupervised learning methods are used in many fields under a wide variety of names. The most commonly practiced unsupervised methods are latent variable models (principal component and factor analyses) and disjoint cluster analyses. In principal component analysis, dimensionality of multivariate data is reduced by transforming the correlated variables into linearly transformed uncorrelated variables. In factor analysis, a few uncorrelated hidden factors that explain the maximum amount of common variance and are responsible for the observed correlation among the multivariate data are extracted. The relationship between the multiple attributes and the extracted hidden factors are then investigated. Disjoint cluster analysis is used for combining cases into groups or clusters such that each group or cluster is homogeneous with respect to certain attributes. Each cluster is also assumed to be different from other groups with respect to the similar characteristics. This implies that observations belonging to the same cluster are as similar to each other as possible, while observations belonging to different clusters are as dissimilar as possible.

The non-mathematical description and application of these unsupervised learning methods are discussed briefly in this chapter. For a mathematical account of principal component analysis, factor analysis, and disjoint cluster analysis, readers are encouraged to refer to Sharma[1] and Johnson and Wichern.[2]

4.2 Applications of Unsupervised Learning Methods

The applications of unsupervised learning methods in data mining have increased tremendously in recent years. Many of the application examples are multifaceted, as more than one unsupervised learning method can be used to solve one specific objective. To clearly differentiate the purpose of performing different unsupervised learning methods, specific examples for each unsupervised learning method application are described in this section.

- **Principal component analysis (PCA).** A business analyst is interested in ranking 2000 mutual funds based on the last two years' monthly performance of 20 financial indicators and ratios. It would be very difficult to score each mutual fund based on 20 indicators and interpret the findings. Thus, the analyst performed PCA on a standardized 2000 × 20 data matrix and extracted the first two principal components. The first two components accounted for 74% variation contained in the 20 variables. Thus, the analyst used the first two principal components to develop score cards and rank the mutual funds.

- **Exploratory factor analysis.** An online-based major book company collects and maintains a large database on annual purchase patterns for various categories of books, audio CDs, DVDs, and CD-ROMs for individual customers. The CEO for marketing was interested in using this database to find similarities among different consumer purchasing patterns so that advertising strategies could be implemented for the various groups of purchasing patterns. A business analyst used exploratory factor analysis and extracted hidden factors responsible for the observed correlation among the various purchase patterns. Using the correlation structure between the original purchase patterns and the hidden factors, the marketing CEO was able to design appropriate advertising strategies. The customers were also ranked based on the factor scores, and different levels of sales promotions were adopted for each customer based on the factor score values.
- **Disjoint cluster analysis.** A major bank collects and maintains a large database on customer banking patterns for bank services, such as checking accounts, savings accounts, certificates of deposit, loans, and credit cards. Based on the banking attributes, the bank CEO would like to segment bank customers into very active, moderate, and passive groups based on data for the last three years for the banking patterns of individual customers. The bank's business analyst performed a disjoint cluster analysis on the customer × banking indicator data and extracted non-overlapping disjoint cluster groups. The marketing division used the segmented customer group information and tried differential advertising strategies for the various cluster groups.

4.3 Principal Component Analysis

Because it is difficult to visualize multidimensional space, principal component analysis, a popular multivariate technique, is primarily used to reduce the dimensionality of p multiple attributes to two or three dimensions. PCA summarizes the variation in a correlated multi-attribute to a set of uncorrelated components, each of which is a particular linear combination of the original variables. The extracted uncorrelated components are called principal components (PCs) and are estimated from the eigenvectors of the covariance or correlation matrix of the original variables; therefore, the objective of PCA is to achieve parsimony and reduce dimensionality by extracting the smallest number of components that account for most of the variation in the original multivariate data and to summarize the data with little loss of information.

In PCA, uncorrelated PCs are extracted by linear transformations of the original variables so that the first few PCs contain most of the variations in the original dataset. These PCs are extracted in decreasing order of importance so that the first PC accounts for as much of the variation as possible and each successive component accounts for a little less. Following PCA, the analyst tries to interpret the first few principal components in terms of the original variables and thereby have a greater understanding of the data. To reproduce the total system variability of the original p variables, we need all p PCs. However, if the first few PCs account for a large proportion of the variability (80–90%), we have achieved our objective of dimension reduction. Because the first principal component accounts for the covariation shared by all attributes, this may be a better estimate than simple or weighted averages of the original variables. Thus, PCA can be useful when a very high degree of correlation is present in the multiple attributes.

In PCA, the extractions of PC can be made using either original multivariate datasets or the covariance or the correlation matrix, if the original dataset is not available. To derive PCs, the correlation matrix is commonly used when the variables in the dataset are measured using different units (e.g., annual income, educational level, numbers of cars owned per family) or when the variables have different variances. Using the correlation matrix is equivalent to standardizing the variables to zero mean and unit standard deviation. The statistical theory, methods, and computation aspects of PCA are presented in details elsewhere.[3]

4.3.1 PCA Terminology

- **Eigenvalues.** Eigenvalues measure the amount of the variation explained by each PC and will be largest for the first PC and smaller for the subsequent PCs. An eigenvalue greater than 1 indicates that PCs account for more variance than accounted for by one of the original variables in standardized data. This is commonly used as a cutoff point for which PCs are retained.
- **Eigenvectors.** Eigenvectors provide the weights to compute the uncorrelated PCs, which are the linear combinations of the centered standardized or centered unstandardized original variables.
- **PC scores.** PC scores are the derived composite scores computed for each observation based on the eigenvectors for each PC. The means of PC scores are equal to zero, as these are linear combinations of the centered variables. These uncorrelated PC scores can be used in subsequent analyses to check for multivariate normality,[4] to detect multivariate outliers,[4] or as a remedial measure in regression analysis with severe multi-collinearity.[5]

- **Estimating the number of PCs.** Several criteria are available for determining the number of PCs to be extracted, but these are just empirical guidelines rather than definite solutions. In practice, we seldom use a single criterion to decide on the number of PCs to extract. Some of the most commonly used guidelines are the Kaiser–Guttman rule, the scree and parallel analysis plots, and interpretability.[6]
  - **Kaiser-Guttman rule.** The Kaiser–Guttman rule states that the number of PCs to be extracted should be equal to the number of PCs having an eigenvalue greater than 1.0.
  - **Scree test.** Plotting the eigenvalues against the corresponding PCs produces a scree plot that illustrates the rate of change in the magnitude of the eigenvalues for the PCs. The rate of decline tends to be fast at first, then it levels off. The "elbow", or the point at which the curve bends, is considered to indicate the maximum number of PCs to extract. One less PC than the number at the elbow might be appropriate if an overly defined solution is sought. However, scree plots may not give such a clear indication of the number of PCs at all times.
  - **Parallel analysis.** To aid in determining the number of PCs to be extracted in standardized data, another graphical method known as parallel analysis is suggested to enhance the interpretation the scree plot (see Sharma[7] for a description of the computational details of performing parallel analysis). In parallel analysis, eigenvalues are extracted in repeated sampling from a totally independent multivariate dataset with the exact same dimensions as the data of interest. Because the variables are not correlated in this simulated dataset, all the extracted eigenvalues should have a value equal to 1. However, due to sampling error, the first half of the PC will have eigenvalues greater than 1, and the second half of the PCs will have eigenvalues less than 1. The average eigenvalues for each PC computed from the repeated sampling is overlaid on the same scree plot of the actual data. The optimum number of PCs is selected at the cutoff point where the scree plot and the parallel analysis curve intersect. An example of a scree/parallel analysis plot is presented in Figure 4.3.
- **Interpretability.** Another very important criterion for determining the number of PCs is the interpretability of the PCs extracted. The number of PCs extracted should be evaluated not only according to empirical criteria, but also according to the criterion of meaningful interpretation.

- **PC loadings.** PC loadings are correlation coefficients between the PC scores and the original variables. They measure the importance of each variable in accounting for the variability in the PCs. It is often possible to interpret the first few PCs in terms of overall effect or contrast between groups of variables based on the structures of PC loadings. A high correlation between the first principal component (PC1) and a variable indicates that the variable is associated with the direction of the maximum amount of variation in the dataset. More than one variable might have a high correlation with PC1. A strong correlation between a variable and the second principal component (PC2) indicates that the variable is responsible for the next largest variation in the data perpendicular to PC1, and so on. Conversely, if a variable does not correlate to any PC axis or correlates only with the last PC or one before the last PC, this usually suggests that the variable has little or no contribution to the variation in the dataset. Therefore, PCA may often indicate which variables in a dataset are important and which ones may be of little consequence. Some of these low-performance variables might therefore be removed from consideration in order to simplify the overall analyses.

4.4 Exploratory Factor Analysis

Factor analysis is a multivariate statistical technique concerned with extraction of a small number of hidden factors that are responsible for the correlation among the set of observable variables. It is assumed that observed variables are correlated because they share one or more underlying factors. Basically, factor analysis tells us what variables can be grouped or go together. Factor analysis transforms a correlation matrix or a highly correlated multivariate data matrix into a few major factors so that the variables within the factors are more highly correlated with each other than with variables in the other factors.

Exploratory factor analysis (EFA) is the most common form of factor analysis that seeks to uncover the underlying structure of a relatively large set of variables. The *a priori* assumption is that any variable may be associated with any factor. No prior theories regarding factor structures are available; we use estimated factor loadings to intuit the factor structure of the data.

If the extracted hidden factors account for most of the variation in the data matrix, the partial correlations among the observed variables will be close to zero. Thus, the hidden factors determine the values of the observed variables. Each observed variable could be expressed as a weighted composite of a set of latent factors. Given the assumption that the residuals are uncorrelated across the observed variables, the correlations among the observed variables are accounted for by the factors. Any correlation between a pair of observed variables can be explained in terms of their relationships with the latent factors. The statistical theory, methods, and computation aspects of EFA are presented in detail elsewhere.[8,9]

4.4.1 Exploratory Factor Analysis vs. Principal Component Analysis

Many different methods of factor analysis are available in the SAS system, and PCA is one of the most common. While EFA and PCA analyses are functionally very similar and both are used for data reduction and summarization, they are quite different in terms of the underlying assumptions. In EFA, the variance of a single variable can be partitioned into common and unique variances. The common variance is considered shared by other variables included in the model, and the unique variance that includes the error component is unique to that particular variable. Thus, EFA analyzes only the common variance of the observed variables, whereas PCA summarizes the total variance and makes no distinction between common and unique variance.

The selection of PCA over EFA is dependent upon the objective of the analysis and the assumptions about the variation in the original variables. EFA and PCA are considered similar because the objectives of both analyses are to reduce the original variables into fewer components: factors or principal components. However, they are also different in that the extracted components serve different purposes. In EFA, a small number of factors are extracted to account for the intercorrelations among the observed variables and to identify the latent factors that explain why the variables are correlated with each other; however, the objective of PCA is to account for the maximum portion of the variance present in the original variables with a minimum number of PCs.

If the observed variables are measured relatively error free (e.g., customer age, number of years of education, or number of family members) or if it is assumed that the error and unique variance represent a small portion of the total variance in the original set of the variables, then PCA is appropriate. But, if the observed variables are only indicators of the latent factor or if the error variance represents a significant portion of the total variance, then the appropriate technique is EFA.

4.4.2 Exploratory Factor Analysis Terminology

4.4.2.1 Communalities and Uniqueness

The proportion of variance in the observed variable that is attributed to the factors is called *communality*, and the proportion of variance in the observed variable that is not accounted by factors is called *uniqueness*. Thus, communalities and uniqueness sum to one. Many different methods are available to estimate communalities. When the communalities are assumed to be equal to 1.0 (i.e., all the variables are completely predicted by the factors), then this factor analysis is equal to PCA; however, the objective of PCA is dimension reduction rather than explaining observed correlations with underlying factors.

A second method for estimating prior communalities is to use the squared multiple correlation (SMC) in a regression model. In estimating the SMC, each variable is treated as the response in a regression model in which all the other variables are considered predictor or input variables. The estimated R^2 from this multiple regression is used as the prior communality estimate for SMC and in factor extraction. Similarly, the SMCs for all observed variables are estimated and used as prior communality estimates. After the factor analysis is completed, the actual communality values are re-estimated and reported as the final communality estimates.

4.4.2.2 Factor Analysis Methods

A variety of different factor extraction methods are available in the SAS PROC FACTOR procedure, including principal component, principal factor, iterative principal factor, unweighted least-squares factor, maximum-likelihood factor, alpha factor, image analysis, and Harris component analysis. The two most commonly employed factor analysis techniques are principal factor and maximum-likelihood factor. The various factor analysis techniques employ different criteria for extracting factors. Discussions on choosing different methods of factor extraction can be found in Sharma.[8]

4.4.2.3 Sampling Adequacy Check in Factor Analysis

Kaiser–Meyer–Olkin (KMO) statistics predict if data are likely to factor well, based on correlation and partial correlation among the variables. A KMO statistic is reported for each variable, and the sum of these statistics is the KMO overall statistic. The KMO varies from 0 to 1.0, and the overall KMO should be 0.60 or higher to proceed with successful factor analysis.

If the overall KMO statistic is less than 0.60, then drop the variables with the lowest individual KMO statistic values, until the overall KMO rises above 0.60. To compute the overall KMO, find the numerator, which is the sum of squared correlations of all variables in the analysis (except for the 1.0 self-correlations of variables with themselves), then calculate the denominator, which is the sum of squared correlations plus the sum of squared partial correlations of each *i*th variable with each *j*th variable, controlling for others in the analysis. The partial correlation values should not be very large for successful factor extraction.

4.4.2.4 Estimating the Number of Factors

Methods described in Section 4.3.1 for extracting the optimum number of PCs could also be used in factor analysis. Some of the most commonly used guidelines in estimating the number of factors are the modified Kaiser–Guttman rule, percentage of variance, scree test, size of the residuals, and interpretability.[8]

4.4.2.5 Modified Kaiser–Guttman Rule

The modified Kaiser–Guttman rule states that the number of factors to be extracted should be equal to the number of factors having an eigenvalue greater than 1.0. This rule should be adjusted downward when the common factor model is chosen. It has been suggested that the eigenvalue criterion should be lower and around the average of the initial communality estimates.

4.4.2.6 Percentage of Variance

Another criterion related to the eigenvalue is the percentage of the common variance (defined by the sum of communality estimates) explained by successive factors. For example, if the cutoff value is set at 75% of the common variance, then factors will be extracted until the sum of eigenvalues for the retained factors exceeds 75% of the common variance, defined as the sum of initial communality estimates.

4.4.2.7 Scree/Parallel Analysis Plot

Similar to PC analysis, scree plot and parallel analysis could be used to detect the optimum number of factors for standardized data (see Section 4.3.1); however, the parallel analysis suggested under PC is not valid for the maximum-likelihood-based EFA method.

4.4.2.8 Chi-Square Test in Maximum-Likelihood Factor Analysis Method

This test is comprised of two separate hypotheses tests. The first test — test of H_0: no common factors — tests the null hypothesis that no common factors can sufficiently explain the intercorrelations among the variables included in the analysis. This test should be statistically significant ($p < .05$); a non-significant value for this test statistic suggests that the intercorrelations may not be strong enough to warrant performing a factor analysis, as the results from such an analysis could probably not be replicated.

The second chi-square test statistic — test of H_0: *N* factors are sufficient — is the test of the null hypothesis that *N* common factors are sufficient to explain the intercorrelations among the variables, where *N* is the number of factors specified. This test is useful for testing the hypothesis that a given number of factors are sufficient to account for the data. In this instance, the goal is a small chi-square value relative to its degrees of freedom. This outcome results in a large *p* value ($p > .05$). One downside of this test is that the chi-square test is very sensitive to sample size: Given large degrees of freedom, this test will normally reject the null hypothesis of the residual matrix being a null matrix, even when the factor analysis solution is very good. Therefore, be careful in interpreting the significance value of this test. Some datasets do not lend themselves to good factor solutions, regardless of the number of factors extracted.

4.4.2.9 A Priori Hypothesis

The *a priori* hypothesis can provide a criterion for deciding the number of factors to be extracted. If a theory or previous research suggests a certain number of factors and the analyst wants to confirm the hypothesis or replicate the previous study, then a factor analysis with the prespecified number of factors can be run. Ultimately, the criterion for determining the number of factors should be replicability of the solution. It is important to extract only factors that can be expected to replicate themselves when a new sample of subjects is employed.

4.4.2.10 Interpretability

Another very important criterion for determining the number of factors is the interpretability of the factors extracted. Factor solutions should be evaluated not only according to empirical criteria, but also according to the criterion of theoretical meaningfulness.

4.4.2.11 Eigenvalues

Eigenvalues measure the amount of variation in the total sample accounted for by each factor and reveal the explanatory importance of the factors with respect to the variables. If a factor has a low eigenvalue, then it is contributing little to the explanation of variances in the variables and may be excluded as redundant. Note that the eigenvalue is not the percent of variance explained but rather a measure of "amount" used for comparison with other eigenvalues. The eigenvalue of a factor may be computed as the sum of its squared factor loadings for all the variables. Note that the eigenvalues associated with the unrotated and rotated solution will differ, although the sum of all eigenvalues will be the same.

4.4.2.12 Factor Loadings

Factor loadings are the basis for assigning labels to the various factors, and they represent the correlation or linear association between a variable and the latent factors. Factor loadings are represented by a $p \times k$ matrix of correlations between the original variables and their factors, where p is the number of variables and k is the number of factors retained. Factor loadings greater than 0.40 in absolute value are frequently used to make decisions regarding significant loading. As the sample size and the number of variables increase, the criterion may have to be adjusted slightly downward; as the number of factors increase, the criterion may have to be adjusted upward. The procedure described next outlines the steps of interpreting a factor matrix.

Once all significant loadings are identified, we can assign some meaning to the factors based on the factor loadings patterns. First, examine the significant loading for each factor. In general, the larger the absolute size of the factor loading for a variable, the more important the variable is in interpreting the factor. The sign of the loading also must be considered in labeling the factors. By considering the loading of all variables on a factor, including the size and sign of the loading, we can determine what the underlying factor may represent.

The squared factor loading is the percent of variance in that variable explained by the factor. To get the percent of variance in all the variables accounted for by each factor, add the sum of the squared factor loadings for that factor and divide by the number of variables. Note that the number of variables equals the sum of their variances, as the variance of a standardized variable is 1. This is the same as dividing the factor's eigenvalue by the number of variables. The ratio of the squared factor loadings for a given variable shows the relative importance of the different factors in explaining the variance of the given variable.

4.4.2.13 Factor Rotation

The idea of simple structure and ease of interpretation form the basis for rotation. The goal of factor rotation is to rotate the factors simultaneously in order to have as many zero loadings on each factor as possible. The sum of eigenvalues is not affected by rotation, but rotation will alter the eigenvalues of particular factors. The rotated factor pattern matrix is calculated by post-multiplying the original factor pattern matrix by the orthogonal transformation matrix.

The simplest case of rotation is an orthogonal rotation in which the angle between the reference axes of factors is maintained at 90 degrees. More complicated forms of rotation allow the angle between the reference axes to be other than a right angle (i.e., factors can be correlated with each other). These types of rotational procedures are referred to as oblique rotations. Orthogonal rotation procedures are more commonly used than oblique rotation procedures. In some situations, theory may mandate that underlying latent factors be uncorrelated with each other, and oblique rotation procedures would not be appropriate. In other situations, when the correlations between the underlying factors are not assumed to be zero, oblique rotation procedures may yield simpler and more interpretable factor patterns. In all cases, interpretation is easiest if we achieve what is called *simple structure*. In simple structure, each variable is highly associated with one and only one factor. If that is the case, we can name factors for the observed variables highly associated with them.

VARIMAX is the most widely used orthogonal rotation method, and PROMAX is the most popular oblique rotation method. VARIMAX rotation produces factors that have high correlations with one smaller set of variables and little or no correlation with another set of variables. Each factor will tend to have either large or small loadings of particular variables on it. A VARIMAX solution yields results that make it as easy as possible to identify each variable with a single factor.

PROMAX rotation is a nonorthogonal rotation method that is computationally faster and therefore is recommended for very large datasets. PROMAX rotation begins with a VARIMAX rotation and makes the larger loadings closer to 1.0 and the smaller loadings closer to 0, resulting in an easy-to-interpret simple factor structure. When an oblique rotation method is performed, the output also includes a factor pattern matrix, which is a matrix of standardized regression coefficients for each of the original variables on the rotated factors. The meaning of the rotated factors is inferred from the variables significantly loaded on their factors. One downside of an oblique rotation method is that if the correlations among the factors are substantial, then it is sometimes difficult to distinguish among factors by examining the factor loadings. In such situations, investigate the factor pattern matrix, which displays the variance explained by each factor and the final communality estimates.

4.4.2.14 Standardized Factor Scores

Standardized factor scores are the scores of all the cases on all the factors, where cases are the rows and factors are the columns. Factor scores can quantify individual cases on a latent factor using a *z*-score scale, which ranges from approximately –3.0 to +3.0. The SAS FACTOR procedure can provide the estimated scoring confidents, which are then used in PROC SCORE to produce a matrix of estimated factor scores. These scores can then be input into an SAS dataset for further analysis.[10]

4.5 Disjoint Cluster Analysis

Cluster analysis is a multivariate statistical method to group cases or data points into clusters or groups suggested by the data, not defined *a priori*, so that the degree of association is strong between members of the same cluster and weak between members of different clusters. When exploring and describing large datasets, it is sometimes useful to summarize the information by assigning each observation to a cluster with similar characteristics. Clustering can be used to reduce the size of the data and to induce groupings. As a result, cluster analysis can reveal similarities in multivariate data that may have been impossible to find otherwise. Cluster analysis tries to identify a set of groups that both minimize within-group variation and maximize between-group variation. Each group or cluster is homogeneous with respect to the multiple attributes. The statistical theory, methods, and computation aspects of cluster analysis are presented in detail elsewhere.[11,12]

4.5.1 Types of Cluster Analysis

4.5.1.1 Hierarchical Cluster Analysis

Hierarchical cluster analysis (HCA) is a nested method of grouping observations in which one cluster may be entirely contained within another cluster, but no other kind of overlap between clusters is allowed; therefore, HCA is not suitable for clustering large databases. Prior knowledge of the number of clusters is not required, and once a cluster is assigned it cannot be reassigned.

4.5.1.2 Disjoint Cluster Analysis

Disjoint (nonhierarchical) cluster analysis (DCA) assigns each observation in only one cluster. First, the observations are arbitrarily divided into clusters, and then observations are reassigned one by one to different clusters on the basis of their similarity to the other observations in the cluster. The process continues until no items must be reassigned. Disjoint clustering is generally more efficient for large datasets, and *a priori* knowledge of the number of clusters is required to perform DCA.

4.5.1.3 k-Mean Cluster Analysis

This type of analysis is the most common form of disjoint cluster analysis and is often used in data mining. In *k*-mean clustering, the cluster centers are derived from the means of the observations assigned to each cluster when the algorithm is run to complete convergence. In the *k*-means model, each iteration reduces the variation within the clusters and maximizes the differences among the distinct clusters until convergence is achieved. A set of points called *cluster seeds* is selected as a first guess of the means of the clusters. Each observation is assigned to the nearest seed to form temporary clusters. The seeds are then replaced by the means of the temporary clusters, and the process is repeated until no further changes occur in the clusters.

4.5.1.4 Optimum Number of Population Clusters

While no perfect way to determine the number of clusters exists, we could use some statistics to help in the process. Cubic clustering criterion (CCC), pseudo F statistic (PSF), and pseudo T^2 statistic (PST2) are useful in determining the number of clusters in the data. An overlay plot of PST2 and PSF (Y axis) vs. the number of clusters (X axis) plot can be used to select the number of potential clusters in the data. Starting from large values on the X axis, move left until a big jump up in the PST2 value occurs. Select the cluster number when the PST2 value has a relatively big jump. Similarly, a relatively large PSF value indicates an optimum cluster number when checking the PSF value from right to left on the X axis of the overlay plot. Values of the CCC greater than 2 indicate good clusters; values between 0 and 2 indicate potential clusters. They should be considered with caution, however, because large negative values can indicate outliers. It may be advisable to look for consensus among the three statistics — that is, the local peaks of the CCC and PSF combined with a big jump in the value of the PST2 statistic. It must be emphasized that these criteria are appropriate only for compact or slightly elongated clusters that are roughly multivariate normal.[13]

4.5.2 FASTCLUS: SAS Procedure To Perform Disjoint Cluster Analysis

FASTCLUS performs a DCA on the basis of distances computed from one or more quantitative variables.[13] The observations are divided into clusters such that every observation belongs to only one cluster. The clusters do not form a tree structure as they do in the hierarchical clustering procedure. FASTCLUS finds disjoint clusters of observations using a *k*-means method applied to coordinate data. By default, the FASTCLUS procedure uses Euclidean distances, so the cluster centers are based on least-squares estimation. The initialization method used by the FASTCLUS procedure makes it sensitive to outliers. PROC FASTCLUS can be an effective procedure for detecting outliers because outliers often appear as clusters with only one member.

The FASTCLUS procedure is intended for use with large datasets with 100 or more observations. With small datasets, the results may be highly sensitive to the order of the observations in the dataset.[13] Another potential problem is that the choice of the number of clusters (*k*) may be critical. Quite different kinds of clusters may emerge when *k* is changed. Good initialization of the cluster centroids is also important because some clusters may be left empty if their centroids initially lie far from the distribution of data.

Before using disjoint clustering, decide whether the multiple attributes should be standardized in some way, as variables with large variances tend to have more effect on the resulting clusters than those with small variances. If all variables are measured in the same units, standardization may not be necessary; otherwise, some form of standardization is strongly recommended. Removing all clusters with small frequencies improves cluster separation and provides visually sharper cluster outlines in scatterplots.

4.6 Bi-Plot Display of PCA, EFA, and DCA Results

Bi-plot[14] display is a visualization technique for investigating the interrelationships between observations and variables in multivariate data. To display a bi-plot, the data should be considered as a matrix, in which the column represents the variable space, while the row represents the observational space. The term *bi-plot* means it is a plot of two dimensions, with the observation and variable spaces plotted simultaneously. In principal component analysis (PCA), relationships between PC scores and PCA loadings associated with any two PCs can be illustrated in a bi-plot display. Similarly, in exploratory factor analysis, relationships between the factor scores and factor loadings associated with any two factors can be displayed simultaneously. In disjoint cluster analysis (DCA), the success of cluster groupings can be verified by plotting cluster grouping and the first two canonical discriminate function scores.[13]

4.7 PCA and EFA Using SAS Macro FACTOR

The FACTOR macro is a powerful SAS application for performing principal component and factor analysis on multivariate attributes. The SAS procedure, PROC FACTOR, is the main tool used in the macro as both PCA and EFA can be performed using PROC FACTOR.[15,16] In addition to using PROC FACTOR as the main tool, other SAS procedures, such as CORR, GPLOT, BOXPLOT, and IML modules, are also incorporated in the FACTOR macro. The advantages of using the FACTOR macro over the PROC FACTOR include:

- The scatterplot matrix (PROC GPLOT) and simple descriptive statistics (PROC CORR) of all multivariate attributes and the significance of their correlations (PROC CORR) are reported.
- Test statistics and *P* values for testing multivariate skewness and kurtosis (SAS/IML) are reported.
- The quantile–quantile (Q–Q) plot for detecting deviation from multivariate normality and the outlier detection plot (PROC GPLOT) for detecting multivariate outliers are generated.
- Bi-plot displays (PROC GPLOT) showing the interrelationships between the principal components or factor scores and the correlations among the multiple attributes are produced for all combinations of selected principal components or factors.
- Options for saving the output tables and graphics in WORD, HTML, PDF, and TXT formats are available.

Software requirements for using the FACTOR macro include:

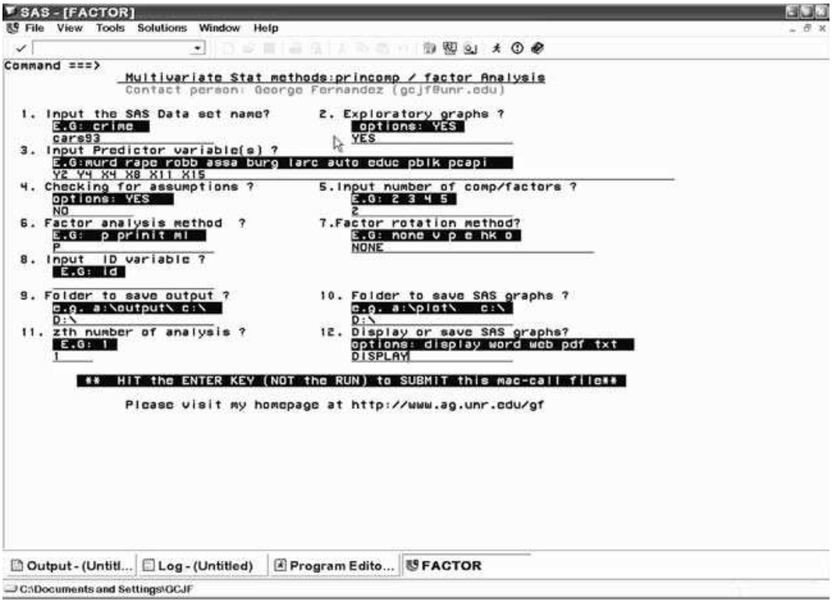
- SAS/CORE, SAS/BASE, SAS/STAT, SAS/GRAPH, and SAS/IML must be licensed and installed at the site.
- SAS version 8.0 and above is recommended for full utilization.
- An active Internet connection is required for downloading the FACTOR macro from the book website if the companion CD-ROM is not available.

4.7.1 Steps Involved in Running the FACTOR Macro

1. Create an SAS dataset (permanent or temporary) from $n \times p$ coordinate data containing p correlated continuous variables and

n observations. If an *n* × *p* coordinate dataset is not available and only a correlation matrix is available, then create a special correlation SAS dataset.

2. *If the companion CD-ROM is not available*, first verify that the Internet connection is active. Open the FACTOR.sas macro-call file in the SAS PROGRAM EDITOR window. Instructions are given in the Appendix regarding downloading the macro-call and sample data files from the book website. *If the companion CD-ROM is available*, the FACTOR.sas macro-call file will be found in the maccall folder on the CD-ROM. Open the FACTOR.sas macro-call file in the SAS PROGRAM EDITOR window. Click the RUN icon to submit the macro-call file FACTOR.sas to open the macro-call window called FACTOR (Figure 4.1).
3. Input the appropriate parameters in the macro-call window by following the instructions provided in the FACTOR macro help file in Section 4.7.2. Users can choose either the scatterplot analysis option or PCA/EFA analysis option. Options for checking for multivariate normality assumptions and detecting the presence of outliers are also available. After inputting all the required macro



```
Command ===>
              Multivariate Stat methods:princomp / factor Analysis
              Contact person: George Fernandez (gcjf@unr.edu)

 1. Input the SAS Data set name?          2. Exploratory graphs ?
    E.G: crime                               options: YES
    cars93                                   YES
 3. Input Predictor variable(s) ?
    E.G:murd rape robb assa burg larc auto educ pbik pcapi
    Y2 Y4 X4 X8 X11 X15
 4. Checking for assumptions ?            5.Input number of comp/factors ?
    options: YES                             E.G: 2 3 4 5
    NO                                       2
 6. Factor analysis method  ?             7.Factor rotation method?
    E.G:  p prinit ml                        E.G: none v p e hk o
    P                                        NONE
 8. Input  ID variable ?
    E.G: id

 9. Folder to save output ?               10. Folder to save SAS graphs ?
    e.g. a:\output\ c:\                      e.g. a:\plot\   c:\
    D:\                                      D:\
11. zth number of analysis ?              12. Display or save SAS graphs?
    E.G: 1                                   options: display word web pdf txt
    1                                        DISPLAY

     ** HIT the ENTER KEY (NOT the RUN) to SUBMIT this mac-call file**

              Please visit my homepage at http://www.ag.unr.edu/gf
```

Figure 4.1 Screen copy of FACTOR macro-call window showing the macro-call parameters required for performing PCA.

parameters, be sure the cursor is in the last input field and the RESULTS VIEWER window is closed, then hit the ENTER key (not the RUN icon) to submit the macro.

4. Examine the LOG window (only in DISPLAY mode) for any macro execution errors. If any errors appear in the LOG window, activate the PROGRAM EDITOR window, resubmit the FACTOR.sas macro-call file, check the macro input values, and correct any input errors.
5. Save the output files. If no errors are found in the LOG window, activate the PROGRAM EDITOR window, resubmit the FACTOR.sas macro-call file, and change the macro input value from DISPLAY to any other desirable format (see Section 4.7.2). PCA or EFA SAS output files and exploratory graphs can be saved in user-specified formats in a user-specified folder.

4.7.2 Help File for SAS Macro FACTOR

1. **Macro-call parameter:** Input SAS dataset name (required parameter).
   Descriptions and explanation: Input the name of the temporary (member_name) or permanent (libname.member_name) SAS dataset for the raw data on which PCA or EFA will be performed. If only the correlation matrix is available and PCA or EFA is to be performed on this correlation matrix, create an SAS special correlation matrix data.
   Options/examples:
   - **Permanent SAS dataset** — gf.cars93 (LIBNAME: gf; SAS dataset name: cars93)
   - **Temporary SAS dataset** — cars93 (SAS dataset name)
2. **Macro-call parameter:** Exploratory graphs (optional parameter).
   Descriptions and explanation: This macro-call parameter is used to select the type of analysis (exploratory graphics and descriptive statistics or PCA/EFA).
   Options/examples:
   - **Yes** — Only simple descriptive statistics, correlation matrix, and scatterplot matrix of all variables are produced. PCA/EFA is not performed.
   - **Blank** — If the macro input field is left blank, no descriptive statistics, correlation matrix, or scatterplot matrix are produced. Only PCA or EFA is performed, depending on the macro input in option #6.
3. **Macro-call parameter:** Input continuous multi-attribute variable names (required parameter for PCA or EFA on raw data).
   If this field is left blank, the macro is expected to perform PCA or FACTOR analysis on a special correlation matrix. The dataset

specified in macro input option #1 should be in the form of a correlation matrix.

Descriptions and explanation: Input continuous multi-attribute names from the SAS dataset that should be included in the PCA/EFA. If no raw data are available, but the correlation matrix is, and PCA or FACTOR analysis is to be performed on this correlation matrix, then leave this field blank. Make sure, however, that the SAS dataset type specified in macro input option #1 is a special correlation matrix.

Options/examples:

Y2 Y4 X4 X8 X11 X15 (names of continuous multi-attributes)

4. **Macro-call parameter:** Check for assumptions (optional statement).

   Descriptions and explanation: To check for multivariate normality and for the presence of any extreme multivariate outliers or influential data, input YES. If this field is to be left blank, this step is omitted.

   Options/examples:

   Yes — Statistical estimates for multivariate skewness, multivariate kurtosis, and their statistical significance are produced. In addition Q–Q plots for checking multivariate normality and multivariate outlier detection plots are also produced.

   Blank — If the macro input field is left blank, no statistical estimates for checking for multivariate normality or detecting outliers are performed.

5. **Macro-call parameter:** Input number of PC or factors (required options).

   Descriptions and explanation: Input the number of principal components or factors to be extracted. The number should be from 1 to the total number of multi-attributes.

   Options/examples:

   2 3 3 4

6. **Macro-call parameter:** Input the factor analysis method (required option).

   Descriptions and explanation: Various factor analysis methods are available in SAS PROC FACTOR, but, to perform PCA, input factor analysis method P. This macro will use the default prior communality estimate 1. To perform factor analysis, select any one of the factor analysis methods other than P.

   Options/examples:

   P — PCA analysis with the default prior communality estimate 1. The scree plot analysis also includes parallel analysis plot.

PRINIT — Iterative PCA with the default prior communality estimate SMC. The scree plot analysis also includes parallel analysis plot.

ML — Maximum-likelihood factor analysis with the default prior communality estimate SMC. The scree plot analysis does not include parallel analysis plot.

For other EFA methods, refer to the PROC FACTOR section in the SAS online manual.[16]

7. **Macro-call parameter:** Input the factor rotation method (required option).

   Descriptions and explanation: Different factor rotation methods are available in SAS PROC FACTOR, but, to perform PCA, input the factor rotation method None. To perform rotated factor analysis, select one of the following factor rotation methods.

   Options/examples:

   None — default

   V — Varimax, orthogonal rotation

   P — Promax, oblique rotation

   For other rotation methods, refer to the PROC FACTOR section in the SAS online manual.[16]

8. **Macro-call parameter:** Input ID variable (optional statement).

   Descriptions and explanation: Input the name of the variable you would like to treat as ID. If this field is left blank, a character variable will be created from the observational number and will be used as the ID variable.

   Option/example:

   Car ID model

9. **Macro-call parameter:** Folder to save SAS output (optional statement).

   Descriptions and explanation: To save the SAS output files in a specific folder, input the full path of the folder. The SAS dataset name will be assigned to the output file. If this field is left blank, the output file will be saved in the default folder.

   Options/examples:

   Possible values

   c:\output\ — folder named “OUTPUT”

   s:\george\ — folder named “George” in network drive S

   Be sure to include the back-slash at the end of the folder name.

10. **Macro-call parameter:** Folder to save SAS graphics (optional statement).

    Descriptions and explanation: To save the SAS graphics files in an EMF format suitable for including in PowerPoint presentations, specify output format as TXT in version 8.0 or later. In pre-8.0

versions, all graphic format files will be saved in a user-specified folder. If the graphics folder field is left blank, the graphics file will be saved in the default folder.

Options/examples:

Possible values

c:\output\ — folder named "OUTPUT"

11. **Macro-call parameter:** *z*th number of run (required statement).

    Descriptions and explanation: SAS output files will be saved by forming a file name from the original SAS dataset name and the counter value provided in this field. For example, if the original SAS dataset name is "gf.cars93" and counter number included is 1, the SAS output files will be saved as "gf.cars931.*" in the user-specified folder. By changing the counter value, users can avoid replacing previous SAS output files with new outputs.

12. **Macro-call parameter:** Display or save SAS output (required statement).

    Descriptions and explanation: Option for displaying output files in the OUTPUT window or options for saving as a specific file format in a folder specified in option #9.

    Options/examples:

    Possible values

    DISPLAY: Output will be displayed in the OUTPUT window, all SAS graphics will be displayed in the GRAPHICS window, and system messages will be displayed in the LOG window.

    WORD: Output and all SAS graphics will be saved together in the user-specified folder and will be displayed in the VIEWER window as a single RTF format file if MS WORD is installed on the system (version 8.0 and later) or saved only as a text file in pre-8.0 versions. All graphics files (CGM) will be saved separately in a user-specified folder (macro input option #10).

    WEB: Output and graphics are saved in the user-specified folder and are viewed in the results VIEWER window as a single HTML file (version 8.0 and later) or saved only as a text file in pre-8.0 versions. All graphics files (GIF) will be saved separately in a user-specified folder (macro input option #10).

    PDF: Output and graphics are saved in the user-specified folder and are viewed in the results VIEWER window as a single PDF file (version 8.2 and later) or saved only as a text file in pre-8.2 versions. All graphics files (PNG) will be saved separately in a user-specified folder (macro input option #10).

TXT: Output will be saved as a TXT file in all SAS versions. No output will be displayed in the OUTPUT window. All graphic files will be saved in the EMF format (version 8.0 and later) or CGM format (pre-8.0 versions) in a user-specified folder (macro input option #9).

Note: System messages are deleted from the LOG window if DISPLAY is not selected as the input.

4.7.3 Case Study 1: Principal Component Analysis of 1993 Car Attribute Data

4.7.3.1 Study Objectives

1. **Variable reduction:** Reduce the dimensions (six) of highly correlated, multi-attribute, coordinate data into a smaller number of dimensions (two or three) without losing much of the variation in the dataset.
2. **Scoring observations:** Group or rank the observations in the dataset based on composite scores generated by an optimally weighted linear combination of the original variables.
3. **Interrelationships:** Investigate the interrelationship between the observations and the multiple attributes, and group similar observations and similar variables.

4.7.3.2 Data Descriptions

Data name	Permanent SAS dataset "CARS93" located in the library "GF"
Multi-attributes	Y2: mid-price Y4: city gas mileage/gallon (mpg) X4: horsepower (hp) X8: passenger capacity X11: width of the vehicle X15: physical weight
Number of observations	92
Data source	Lock[17]

Open the FACTOR.sas macro-call file in the SAS PROGRAM EDITOR window and click RUN to open the FACTOR macro-call window (Figure 4.1). Input the appropriate macro-input values by following the suggestions given in the help file (Section 4.7.2).

4.7.3.3 Exploratory Analysis

Input Y2, Y4, X4, X11, and X15 as the multi-attributes in macro input option #3. Input YES in field #2 to perform data exploration and create a scatterplot matrix. Submit the FACTOR macro, and SAS will output descriptive statistics, correlation matrices, and scatterplot matrices. Only selected output and graphics generated by the FACTOR macro are interpreted below.

The descriptive simple statistics of all multi-attributes generated by the SAS PROC CORR are presented in Table 4.1. The number of observations (*N*) per variable is useful in checking for missing values for any given attribute and providing information on the size of the $n \times p$ coordinate data. The estimates of central tendency (mean) and the variability (standard deviation) provide information on the nature of multiple attributes that can be used to decide whether to use standardized or unstandardized data in the PCA analysis. The minimum and maximum values describe the range of variation in each attribute and help to check for any extreme outliers.

The degree of linear association among the variables measured by the Pearson correlation coefficient (*r*) and their statistical significance are presented in Table 4.2. The value of the *r* ranges from 0 to 0.87. The statistical significance of the *r* varies from no correlation (p = 0.967) to a highly significant correlation (p < 0.0001). Among the 15 possible pairs of correlations, 13 pairs of correlations were highly significant, indicating that these data are suitable for performing PCA analysis. The scatterplot matrix among the five attributes presented in Figure 4.2 reveals the strength of correlation, presence of any outliers, and nature of the bi-directional variation. In addition, each scatterplot shows the linear regression line, 95% mean confidence interval band, and a horizontal line (Y-bar line), which passes through the mean of the Y variable. If this Y-bar line intersects the confidence band lines (i.e., the confidence band region does not enclose the Y-bar line), then the correlation between the X and Y variables is statistically significant. For example, among the15 scatterplots presented in Figure 4.2, only in two of them (Y2 vs. X8; X4 vs. X8) do the Y-bar lines not intersect the confidence band. Only these two correlations are statistically not significant (Table 4.2).

In the PCA analysis, the dimensions of standardized multi-attributes define the number of eigenvalues. An eigenvalue greater than 1 indicates that a PC accounts for more of the variance than one of the original variables in standardized data. This can be confirmed by visually examining the improved scree plot (Figure 4.3) of eigenvalues and the parallel analysis of eigenvalues. This added scree plot shows the rate of change in the magnitude of the eigenvalues for an increasing number of PCs. The rate of decline levels off at a given point in the scree plot that indicates the optimum number of PCs to extract. Also, the intersection point between the scree plot and the parallel analysis plot reveals that the first two eigenvalues that account for 86.2% of the total variation could be retained as the significant PC (Table 4.3).

Table 4.1 Macro FACTOR: PROC CORR Output, Simple Statistics

Variable	N	Mean	Standard Deviation	Sum	Minimum	Maximum	Label
Y2	92	19.63152	9.64023	1806	7.40000	61.90000	mid-price
Y4	92	22.29348	5.60717	2051	15.00000	46.00000	ctympg
X4	92	144.50000	52.25666	13294	55.00000	300.00000	hp
X8	92	5.09783	1.03838	469.00000	2.00000	8.00000	pcap
X11	92	69.41304	3.78299	6386	60.00000	78.00000	width
X15	92	3081	587.86751	283455	1695	4105	weight

Table 4.2 Macro FACTOR: PROC CORR Output, Pearson Correlations, and Their Statistical Significance Levels[2]

Variable	Y2	Y4	X4	X8	X11	X15
Y2 (mid-price)	1.00000	–0.58837[a]	0.78498	0.04514	0.44994	0.64142
		<.0001[b]	<.0001	0.6692	<.0001	<.0001
Y4 (ctympg)	–0.58837	1.00000	–0.66755	–0.40888	–0.71759	–0.84055
	<.0001		<.0001	<.0001	<.0001	<.0001
X4 (hp)	0.78498	–0.66755	1.00000	–0.00435	0.64054	0.73446
	<.0001	<.0001		0.9671	<.0001	<.0001
X8 (pcap)	0.04514	–0.40888	–0.00435	1.00000	0.48475	0.54683
	0.6692	<.0001	0.9671		<.0001	<.0001
X11 (width)	0.44994	–0.71759	0.64054	0.48475	1.00000	0.87408
	<.0001	<.0001	<.0001	<.0001		<.0001
X15 (weight)	0.64142	–0.84055	0.73446	0.54683	0.87408	1.00000
	<.0001	<.0001	<.0001	<.0001	<.0001	

[a] Correlation coefficient.

[b] Satistical signficance (P-value).

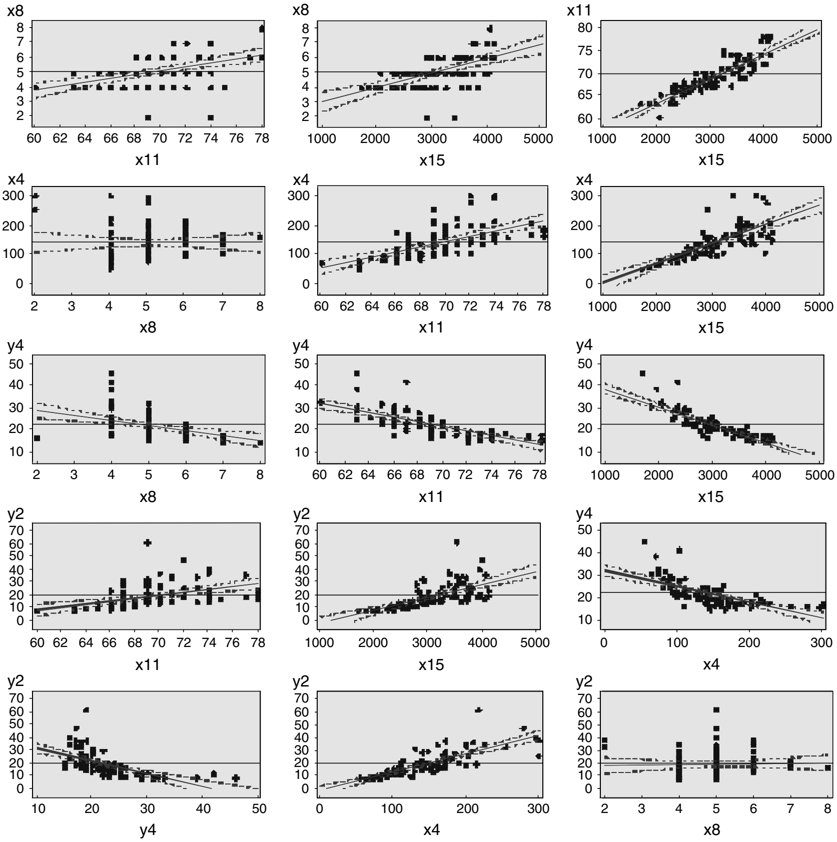


Figure 4.2 Scatter plot matrix illustrating the degree of linear correlation among the five attributes. The graphics file was generated by selecting the TXT file type in the SAS macro FACTOR. The SAS graphics driver used was EMF.

If the data are standardized (i.e., normalized to zero mean and 1 standard deviation), the sum of the eigenvalue is equal to the number of variables used. The magnitude of the eigenvalue is usually expressed in percent of the total variance. The information in Table 4.3 indicates that the first eigenvalue accounts for about 65% of the variation; the second, for 20%. The proportions drop off gradually for the rest of the eigenvalues. Cumulatively, the first two eigenvalues together account for 86% of the variation in the dataset. A two-dimensional view (of the six-dimensional dataset) can be created by projecting all data points onto the plane defined by the axes of the first two PCs. This two-dimensional view will retain 85% of the information from the six-dimensional plot.

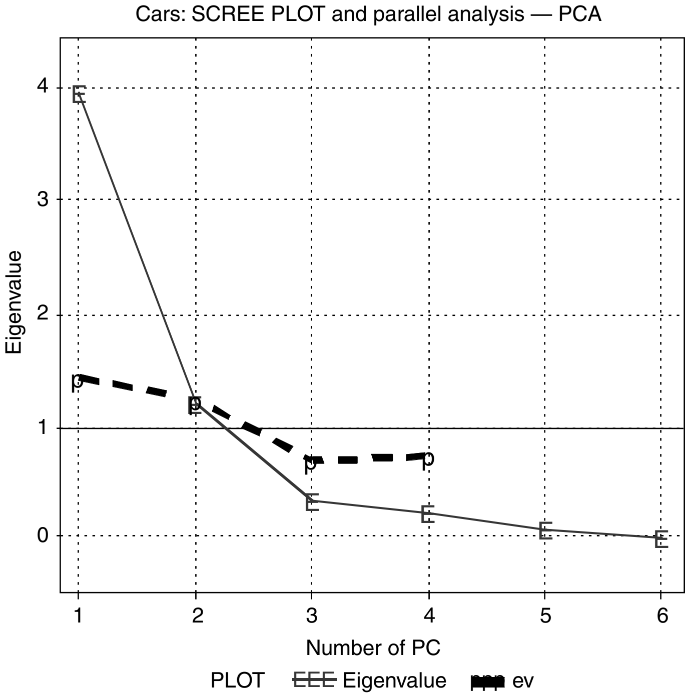


Figure 4.3 Intersection of scree plot illustrating the relationship between number of PC and the rate of decline of eigenvalue and the parallel analysis plot. The graphics file was generated by selecting the TXT file type in the SAS macro FACTOR. The SAS graphics driver used was EMF.

Table 4.3 Macro FACTOR: Eigenvalues in Principal Component Analysis[a]

	Eigenvalue	*Difference*	*Proportion*	*Cumulative*
1	3.94813234	2.72409793	0.6580	0.6580
2	1.22403441	0.84393957	0.2040	0.8620
3	0.38009485	0.11230439	0.0633	0.9254
4	0.26779046	0.14448041	0.0446	0.9700
5	0.12331005	0.06667217	0.0206	0.9906
6	0.05663789	—	0.0094	1.0000

[a] Eigenvalues of the correlation matrix: total = 6; average = 1.

The new variables PC1 and PC2 are linear combinations of the six standardized variables, and the magnitude of the eigenvalues accounts for the variation in the new PC scores. The eigenvectors presented in Table 4.4 provide the weights for transforming the six standardized variables into

Table 4.4 Macro FACTOR: Eigenvectors in Principal Component Analysis

		Eigenvectors	
Variables		*1*	*2*
Y2	mid-price	0.37710	–0.44197
Y4	ctympg	–0.44751	–0.05229
X4	hp	0.41811	–0.42603
X8	pcap	0.22974	0.75215
X11	width	0.43939	0.19469
X15	weight	0.48669	0.12956

PCs. For example, PC1 is derived by performing the following linear transformation using these eigenvectors:

$$\begin{gathered} PC1 = 0.37710Y1 - 0.44751Y2 + 0.41811X4 \\ + 0.22974X8 + 0.43939X11 + 0.48669X15 \end{gathered}$$

The sum of the squares of eigenvectors for a given PC is equal to one.

Principal component loadings presented in Table 4.5 are the correlation coefficients between the first two PC scores and the original variables. They measure the importance of each variable in accounting for the variability in the PC. That is, the larger the loadings in absolute term, the more influential the variables are in forming the new PC and vice versa. A high correlation between PC1 and mid-price, ctympg, hp, width, and weight indicates that these variables are associated with the direction of the maximum amount of variation in this dataset. The first PC loading patterns suggest that heavy, big, very powerful, and highly priced cars

Table 4.5 Macro FACTOR: PC Loadings for the First Two PCs (Factors)

	Variables	*Factor1*	*Factor2*
Y2	mid-price	0.74930	–0.48898
Y4	ctympg	–0.88919	–0.05785
X4	hp	0.83079	–0.47135
X8	pcap	0.45649	0.83215
X11	width	0.87307	0.21540
X15	weight	0.96704	0.14334

are less fuel efficient. A strong correlation between pcap and PC2 indicates that this variable is mainly attributed to the passenger capacity of the vehicle responsible for the next largest variation in the data perpendicular to PC1.

A partial list of the first two PC scores presented in Table 4.6 are the scores computed by the linear combination of the standardized variables using the eigenvectors as the weights. The cars that have small negative scores for the PC1 are less expensive, small, and less powerful, but they

Table 4.6 Macro FACTOR: Standardized and Sorted PC Scores for the First Two Components

Observation	*ID*	*Factor1*	*Factor2*
1	39	–2.60832	–0.36804
2	82	–2.15032	–0.37408
3	81	–2.04321	–0.46095
4	32	–1.93080	–0.22925
5	42	–1.67883	–0.51745
6	85	–1.53146	0.31244
7	74	–1.46923	–0.13646
8	44	–1.44650	0.38442
9	87	–1.41928	–0.30308
10	47	–1.32023	–0.37438
11	23	–1.23447	0.37825
12	40	–1.18520	–0.31955
13	30	–1.16499	0.25891
14	62	–1.14394	0.38416
15	56	–1.03193	0.24003
16	80	–1.01518	0.35356
—[a]	—	—	—
83	19	1.24816	–3.61769
84	28	1.25067	–1.81951
85	11	1.30762	–0.11547
86	31	1.34511	0.81154
87	57	1.40631	–2.25287
88	7	1.50488	0.81694
89	16	1.53423	2.52361
90	52	1.71018	0.00869
91	48	1.82993	–1.87121
92	10	1.87482	–1.58474

[a] Partial list.

are highly fuel efficient. Similarly, expensive, large, and powerful cars with low fuel efficiency are listed at the end of the table with large positive PC1 scores.

A bi-plot display of both PC (PC1 and PC2) scores and PC loadings (Figure 4.4) is very effective for studying the relationships within observations and between variables and the interrelationships between observations and the variables. The X–Y axis of the bi-plot of the PCA represents the standardized PC1 and PC2 scores, respectively. In order to display the relationships among the variables, the PC loading values for each PC are overlaid on the same plot after being multiplied by the corresponding maximum value of the PCs. For example, the PC1 loading values are multiplied by the maximum value of the PC1 score, and the PC2 loadings

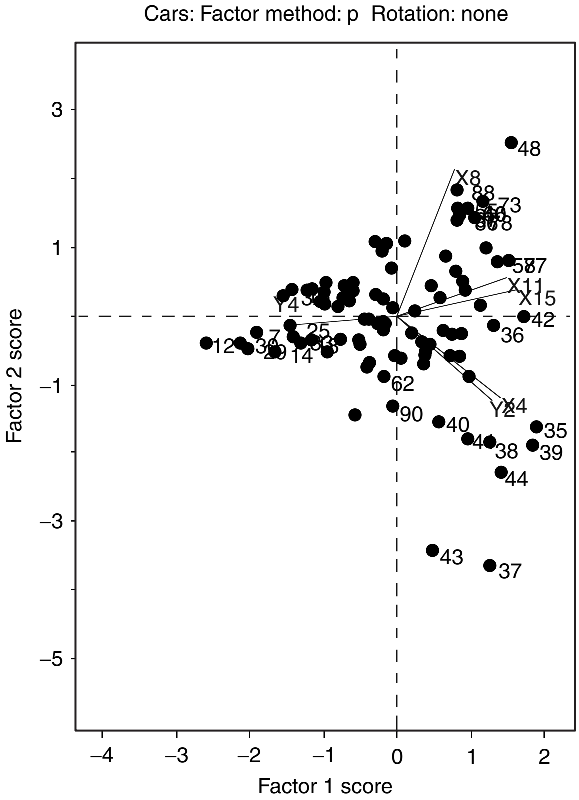


Figure 4.4 Bi-plot display of interrelationship between the first two PC scores and PC loadings. The graphics file was generated by selecting TXT file type in the SAS macro FACTOR. The SAS graphics driver used was EMF.

are multiplied by the maximum value of the PC2 scores. This transformation places both the variables and the observations on the same scale in the bi-plot display because the range of PC loadings is usually shorter than that for the PC scores.

Cars having larger (>75% percentile) or smaller (<25% percentile) PC scores are only identified by their ID numbers on the bi-plot to avoid crowding too many ID values. Cars with similar characteristics are displayed together in the bi-plot observational space because they have similar PC1 and PC2 scores. For example, small compact cars with relatively higher gas mileage, such as Geo Metro (ID 12) and Ford Fiesta (ID 7), are grouped closer. Similarly, cars with different attributes are displayed far apart because their PC1, PC2, or both PC scores are different. For example, small compact cars with relatively higher gas mileage, such as Geo Metro (ID 12), and large expensive cars with relatively lower gas mileage, such as Lincoln Town Car (ID 42), are located far apart.

Correlations among the multivariate attributes used in the PCA are revealed by the angles between any two PC loading vectors. For each variable, a PC load vector is created by connecting the X–Y origin (0,0) and the multiplied value of PC1 and PC2 loadings in the bi-plot. The angles between any two variable vectors will be:

1. Narrower (<45°) if the correlations between these two attributes are positive and larger — for example, Y2 (mid-price) and X4 (horsepower).
2. Wider (around 90°) if the correlation is not significant — for example, Y2 (mid-price) and X8 (passenger capacity).
3. Closer to 180° (>135°) if the correlations between these two attributes are negative and stronger — for example, Y4 (city gas mileage) and X15 (physical weight).

Similarly, stronger associations between some cars (observations) and specific multi-attributes (variables) — for example, gas mileage (Y4) and Geo Metro (ID 12); passenger capacity (X8) and Chevrolet Astro Van (ID 48) — are illustrated by the closer positioning near the tip of the variable vector.

4.7.4 Case Study 2: Maximum-Likelihood Factor Analysis with VARIMAX Rotation of 1993 Car Attribute Data

4.7.4.1 Study Objectives

1. **Suitability of multi-attribute data for factor analysis:** Check the multi-attribute data for multivariate outliers and normality and for sampling adequacy for factor analysis.

2. **Latent factor:** Identify the latent common factors that are responsible for the significant correlations among the five observed car attributes.
3. **Factor scores:** Group or rank the cars in the dataset based on the extracted common factor scores generated by an optimally weighted linear combination of the observed attributes.
4. **Interrelationships:** Investigate the interrelationship between the cars and the multiple attributes and group similar cars and similar attributes together based on factor scores and factor loadings.

4.7.4.2 Data Descriptions

Data name	Permanent SAS dataset "CARS93" located in the library "GF"
Multi-attributes	Y2: mid-price Y4: city gas mileage/gallon (mpg) X4: horsepower (hp) X11: width of the vehicle X15: physical weight
Number of observations	92
Data source	Lock[17]

4.7.4.3 Checking for Multivariate Normality Assumptions and Performing Maximum-Likelihood Factor Analysis

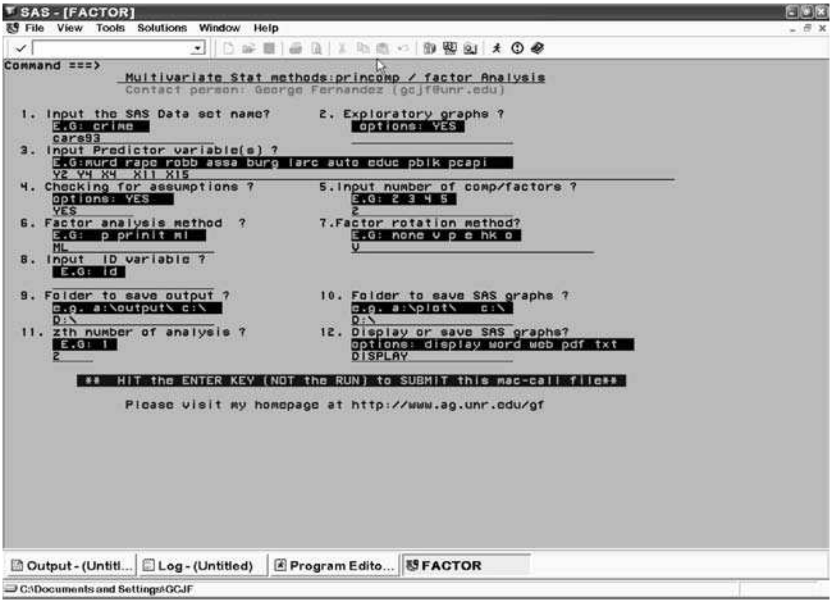
Multivariate normality is a requirement in maximum-likelihood factor analysis, particularly for testing the hypothesis on a number of factors. Checking for multivariate normality and influential observations in multivariate data provides valuable insight into the distribution aspects and influential effects on correlation among the multi-attributes as well. Input Y2, Y4, X4, X11, and X15 as the multi-attributes in macro input field #3 (Figure 4.5). Leave the #2 field blank to skip EDA and to perform ML factor analysis. To check for both multivariate normality and the presence of influential outliers, input YES in macro input field #4. Also, input ML in macro input field #6 for factor analysis method and V in macro input field #7 for factor rotation method. Submit the FACTOR macro, and SAS will output multivariate normality graphs, outlier detection statistics, and complete results from factor analysis output. Only selected output and graphics generated by the FACTOR macro are presented and interpreted below.

Checking for multivariate normality is performed by computing Mardia's multivariate skewness, kurtosis measures, and by chi-square tests.[18]

Table 4.7 Macro FACTOR: Multivariate Normality Test Statistics

Multivariate skewness	18.822
Skewness chi-square	288.598
Skewness *p* value	0.000
Multivariate kurtosis	54.784
Kurtosis *z* value	11.340
Kurtosis *p* value	0.000

The estimates of multivariate normality test statistics and the corresponding *p* values are presented in Table 4.7. Moderately large multivariate skewness and kurtosis values and highly significant *p* values clearly indicate that the distribution of five multi-attributes used in the EFA does not satisfy multivariate normality. This is confirmed by assessing visually the chi-squared quantile–quantile plot of the squared Mahalanobis distances (Figure 4.6). A strong departure from the 45° reference line clearly indicates that the distribution is not multivariate normal. Performing EFA and interpreting factor scores are not affected by a violation in multivariate normality assumptions, but the hypothesis tests on a number of factors



```
Command ===>
              Multivariate Stat methods:princomp / factor Analysis
              Contact person: George Fernandez (gcjf@unr.edu)

  1. Input the SAS Data set name?        2. Exploratory graphs ?
     E.G: crime                             options: YES
     cars93
  3. Input Predictor variable(s) ?
     E.G:murd rape robb assa burg larc auto educ pbik pcapi
     Y2 Y4 X4  X11 X15
  4. Checking for assumptions ?          5.Input number of comp/factors ?
     options: YES                           E.G: 2 3 4 5
     YES                                    2
  6. Factor analysis method  ?           7.Factor rotation method?
     E.G:  p prinit ml                      E.G: none v p e hk o
     ML                                     V
  8. Input  ID variable ?
     E.G: id

  9. Folder to save output ?            10. Folder to save SAS graphs ?
     e.g. a:\output\ c:\                    e.g. a:\plot\   c:\
     D:\                                    D:\
 11. zth number of analysis ?           12. Display or save SAS graphs?
     E.G: 1                                 options: display word web pdf txt
     2                                      DISPLAY

    **  HIT the ENTER KEY (NOT the RUN) to SUBMIT this mac-call file**

              Please visit my homepage at http://www.ag.unr.edu/gf
```

Figure 4.5 Screen copy of FACTOR macro-call window showing the macro-call parameters required for performing EFA.

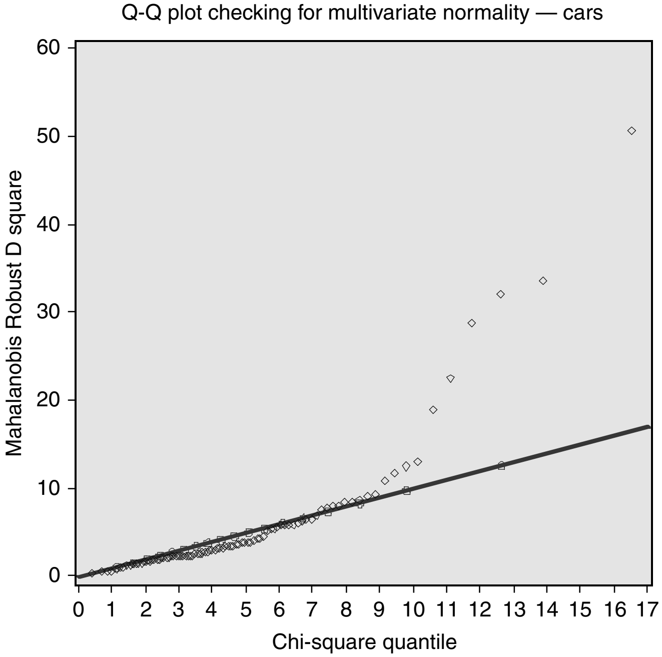


Figure 4.6 Quantile plot for assessing multivariate normality ML factor analysis. The graphics file was generated by selecting the TXT file type in the SAS macro FACTOR. The SAS graphics driver used was EMF.

extracted in the maximum-likelihood factor method are affected by the severe departure from multivariate normality.

Checking for the presence of multivariate influential observations is performed by computing the robust distance square statistic (RDSQ) and the difference between RDSQ and the quantiles of the expected chi-square value (DIFF). Multivariate influential observations are identified when the DIFF values exceed 2.5. The estimates of RDSQ and DIFF values for the eight multivariate influential observations are presented in Table 4.8. Very small, compact, inexpensive cars; some sports cars; and very expensive luxury cars are identified as the most influential observations. The presence of multivariate influential observations is also visually assessed by a graphical display of DIFF values vs. the quantile of chi-square values (Figure 4.7). The impact of these extreme influential observations on EFA outcomes can be verified by excluding these extreme observations one at a time and examining the effects on hypothesis tests and factor loadings.

Table 4.8 Macro FACTOR: Multivariate Influential/Outlier Observations

ID	*Model*	*RDSQ*	*Chi-Square*	*DIFF = RDSQ − Chi-Square > 2.5*
44	Mercedes-Benz 300E	50.6050	16.5511	34.0539
12	Geo Metro	33.6640	13.8929	19.7711
14	Honda Civic	32.0837	12.6235	19.4601
38	Dodge Stealth	28.8156	11.7724	17.0432
37	Chevrolet Corvette	22.5587	11.1273	11.4314
43	Mazda RX-7	18.8527	10.6057	8.2470
39	Infiniti Q45	12.9521	10.1665	2.7856
88	Volkswagen Eurovan	12.4927	9.7863	2.7064

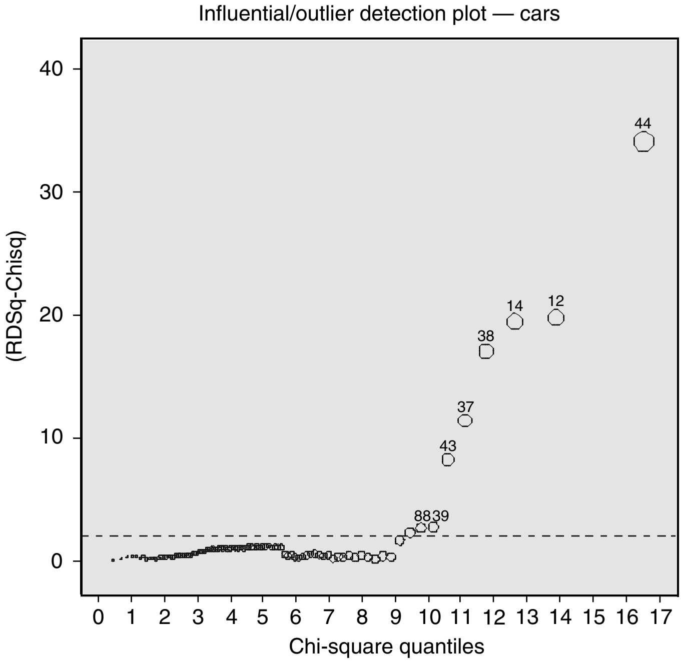


Figure 4.7 Multivariate outlier detection plot based on robust squared distances. The graphics file was generated by selecting the TXT file type in the SAS macro FACTOR. The SAS graphics driver used was EMF.

4.7.4.4 Assessing the Appropriateness of Common Factor Analysis

A *common factor* is an unobservable, hypothetical variable that contributes to the variance of at least two of the observed variables. A *unique factor* is an unobservable, hypothetical variable that contributes to the variance of only one of the observed variables. The model for common factor analysis has one unique factor for each observed variable. If the multivariate data are appropriate for common factor analysis, the partial correlations between any two attributes controlling all the other variables should be small compared to the original correlations presented in Table 4.1. The partial correlation between the variables Y2 and X4, for example, is 0.63 (Table 4.9), slightly less than the original correlation of –0.78. The partial correlation between Y2 and Y4 is –0.02, which is much smaller in absolute value than the original correlation of –0.58; this is a very good indication of appropriateness for common factor analysis. In general, all pairs of partial correlations are smaller than the original simple correlation. Only 2 out of the 10 all-possible partial correlations are significant based on the default criteria (Table 4.9).

In addition to examining the partial correlation coefficients, Kaiser's measure of sampling adequacy (MSA) and the residual correlations also provide very effective measures of assessing the suitability of performing common factor analysis. Kaiser's MSA value ranges between 0 to 1. For each variable and for all variables together, it provides a composite index quantifying how much smaller the partial correlations are in relation to the original simple correlations. Values of 0.8 or 0.9 are considered good, while MSA below 0.5 are unacceptable.[19] All five attributes used in this study have acceptable MSA values, and the overall MSA value (0.77) (Table 4.10) also falls in the acceptable region, confirming the suitability of the data for performing common factor analysis.

Table 4.9 Macro FACTOR: Maximum-Likelihood Factor Method — Partial Correlations

	Partial Correlations Controlling All Other Variables				
Variable	*Y2*	*Y4*	*X4*	*X11*	*X15*
Y2 (mid-price)	100*	–2	63*	–37	32
Y4 (ctympg)	–2	100*	–9	5	–52
X4 (hp)	63*	–9	100*	23	3
X11 (width)	–37	5	23	100*	72*
X15 (weight)	32	–52	3	72*	100*

Note: Prior communality estimates = squared multiple correlation (SMC). Printed values are multiplied by 100 and rounded to the nearest integer. Values greater than 0.563661 are flagged by an asterisk.

Table 4.11 shows the prior communality estimates for five attributes used in this analysis. The squared multiple correlations (SMCs), which are given in Table 4.11, represent the proportion of variance of each of the five attributes shared by all remaining attributes. A small communality estimate might indicate that the attribute may have to be modified or even dropped from the analysis. The Y2 attribute has the prior communality estimate of 0.676, which means about 67% of the variance of the Y2 mid-price is shared by all other attributes included in the analysis, indicating that this attribute shares a common variance among the other attributes. Similarly, the prior communality estimates of the other four attributes are also high (>0.65), indicating that all five attributes contribute to a common variance in the data. The sum of all prior communality estimates, 5.505, is the estimate of the common variance among all attributes. This initial estimate of the common variance constitutes about 42% of the total variance present among all five attributes.

The amount of variances accounted for by each factor is presented in Table 4.12 as eigenvalues. The number of eigenvalues reported should equal the number of variables included in the study. Following the column of eigenvalues are three measures of the relative size and importance of each eigenvalue. The first of these displays the difference between the eigenvalue and its successor. The last two columns display the individual and cumulative proportions that the corresponding factor contributes to the total variation. The first two eigenvalues account for almost all the variability (95% and 5%) in the five-dimensional dataset.

Table 4.10 Macro FACTOR: Maximum-Likelihood Factor Method — Kaiser's Measures of Sampling Adequacy (MSA)

		Variable		
Y2	*Y4*	*X4*	*X11*	*X15*
0.71585505	0.87898105	0.81570844	0.72834723	0.73345010

Note: Prior communality estimates = squared multiple correlation (SMC). Overall MSA = 0.77110942.

Table 4.11 Macro FACTOR: Maximum-Likelihood Factor Method

		Variable		
Y2	*Y4*	*X4*	*X11*	*X15*
0.67654286	0.71337325	0.72459532	0.79670149	0.88441941

Note: Prior communality estimates = squared multiple correlation (SMC).

Table 4.12 Macro FACTOR: Maximum-Likelihood Factor Method — Eigenvalues

	Eigenvalue	*Difference*	*Proportion*	*Cumulative*
1	17.9789723	16.4208803	0.9572	0.9572
2	1.5580919	1.4695273	0.0830	1.0402
3	0.0885647	0.3817004	0.0047	1.0449
4	-0.2931357	0.2570327	–0.0156	1.0293
5	-0.5501684		–0.0293	1.0000

Note: Prior communality estimates = squared multiple correlation (SMC). Preliminary eigenvalues: total = 18.7823248; average = 3.75646495.

4.7.4.5 Determining the Number of Latent Factors

In common factor analysis, the dimensions of standardized multi-attributes define the number of extractable eigenvalues. An eigenvalue greater than 1 indicates that the factor accounts for more variance than one of the original variables in the standardized data. This can be confirmed by visually examining the scree plot (Figure 4.8) of eigenvalues. A scree plot of the eigenvalues could be used to determine the number of meaningful factors. The scree plot shows the rate of change in the magnitude of the eigenvalues for an increasing number of factors. The rate of decline levels off where the scree plot bends, indicating that two is the optimum number of factors to extract.

According to the Kaiser–Guttman rule, only two factors can be extracted because only the first two factors have an eigenvalue greater than 1 (Table 4.12). These two large positive eigenvalues together account for 104% of the common variance, which is close to 100%. The occurrence of negative eigenvalues happens only due to the restriction that the sum of eigenvalues be set equal to the estimated common variance (communality estimates) and not the total variance in common factor analysis. A large first eigenvalue (17.79) and a much smaller second eigenvalue (1.55) suggest the presence of a dominant global factor.

Assuming multivariate normality and larger samples, chi-square tests can be performed in the maximum-likelihood factor method to test the number of meaningful factors. The probability levels for the chi-square test are <0.0001 for the hypothesis of no common factors and 0.0284 for two common factors (Table 4.13); therefore, the two-factor model seems to be an adequate representation and reconfirms the results from the scree plot and the Kaiser–Guttman rule.

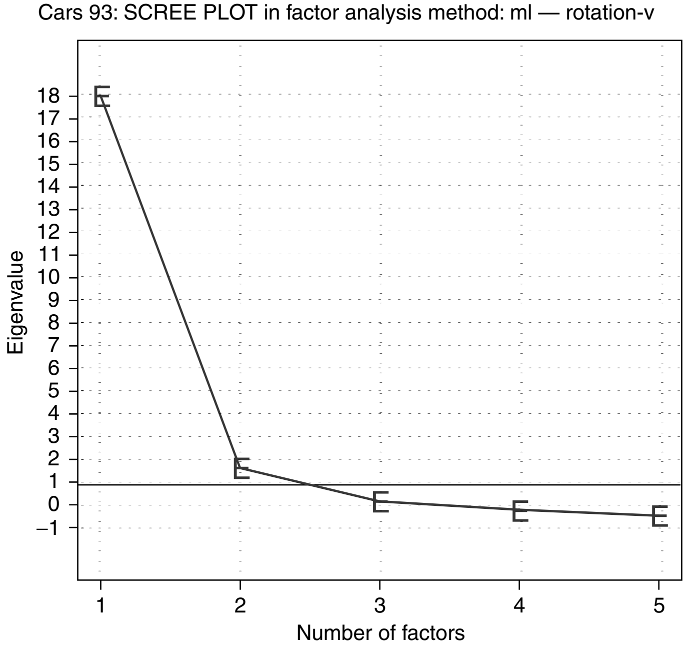


Figure 4.8 Scree plot illustrating the relationship between number of factors and the rate of decline of eigenvalue in ML factor analysis. The graphics file was generated by selecting the TXT file type in the SAS macro FACTOR. The SAS graphics driver used was EMF.

Table 4.13 Macro FACTOR: Maximum-Likelihood Factor Method (Significant tests on the number of factors based on maximum-likelihood factor analysis.)

Test	*Degrees of Freedom*	*Chi-Square*	*Pr > Chi-Square*
H_0: No common factors	10	406.8335	<.0001
H_A: At least one common factor	—	—	—
H_0: Two factors are sufficient	1	4.8059	0.0284
H_A: More factors are needed	—	—	—

Note: Prior communality estimates = squared multiple correlation (SMC).

4.7.4.6 Interpreting Common Factors

Table 4.14 shows the initial unrotated factor loadings, which consist of the correlations between the five attributes and the two retained factors. The correlations greater than 0.649 are flagged by an asterisk. Some loadings are split, where X15 and Y4 are significantly (>0.59) loaded on more than one factor. A commonly used rule is that there should be at least three variables per common factor. In this case study, the five variables we used are not adequate to extract two common factors following the rule of three variables per factor.

Table 4.15 shows the factor loadings of the two extracted factors after the VARIMAX rotation. The VARIMAX rotation is an orthogonal rotation in which the angle between the reference axes of factors is maintained

Table 4.14 Macro FACTOR: Maximum-Likelihood Factor Method — Initial Factor Loadings

Variable	*Factor1*	*Factor2*
Y2 (mid-price)	100*	0
X4 (hp)	78*	32
X11 (width)	45	78*
X15 (weight)	64	75*
Y4 (ctympg)	–59	–61

Note: Prior communality estimates = squared multiple correlation (SMC). Printed values are multiplied by 100 and rounded to the nearest integer. Values greater than 0.649803 are flagged by an asterisk.

Table 4.15 Macro FACTOR: Maximum Likelihood Factor Method — Factor Loadings after VARIMAX Rotation

	Rotated Factor Pattern	
Variable	*Factor1*	*Factor2*
X15 (weight)	90*	42
X11 (width)	87*	22
Y4 (ctympg)	–75*	–40
Y2 (mid-price)	27	96*
X4 (hp)	51	67*

Note: Prior communality estimates = squared multiple correlation (SMC). Printed values are multiplied by 100 and rounded to the nearest integer. Values greater than 0.649803 are flagged by an asterisk.

at 90°. The rotated loading is usually somewhat simpler to interpret; that is, the rotated Factor1 can now be interpreted as a size factor. The size variables X15, X11, and Y4 load higher on Factor1. The large and heavy cars show a negative correlation with the gas mileage, as well. The rotated Factor2 seems to measure price performance. The mid-price and the HP variables load heavily on Factor2. Even though the variance explained by rotated Factor1 is less than Factor2 (Table 4.16), the cumulative variance explained by both common factors remains the same after orthogonal rotation, like the VARIMAX rotation. Also note that the VARIMAX rotation, as with any orthogonal rotation, has not changed the final communalities.

4.7.4.7 Checking the Validity of Common Factor Analysis

The final communality estimates are the proportion of variance of the variables accounted for by the common factors (Table 4.17). When the factors are orthogonal, the final communalities are calculated by taking the sum of the squares of each row of the factor loadings. The final communality estimates show that all the variables are well accounted for by the two factors, with final communality estimates ranging from 0.71 for X4 to 1 for Y2. The variable Y2 has a communality of 1.0 and therefore has an infinite weight that is displayed next to the final communality estimate infinite value. The first eigenvalue is also infinite. Infinite values are ignored in computing the total of the eigenvalues and the total final communality. The final communality estimates are all fairly close to the prior communality estimates reported in Table 4.11. Only the communality for the variables Y2 and X15 increases appreciably. Inspection of the partial correlation matrix yields similar results: The correlations among the five attributes after removing the factor contribution are very low, the largest being 0.18 (Table 4.18). The root-mean-square off-diagonal partial is also very low at 0.071876 (Table 4.19). The residual matrix provides an indication of how well the factor model fits the data. The off-diagonal elements of the residual correlation matrix (Table 4.20) are all close to 0.02, indicating that the correlations among the five attributes can be

Table 4.16 Macro FACTOR: Maximum-Likelihood Factor Method — Percentage Variance Explained by Each Factor after VARIMAX Rotation

Factor	*Weighted*	*Unweighted*
Factor1	21.8130985	2.57624328
Factor2	28.6798365	1.64618997

Note: Prior communality estimates = squared multiple correlation (SMC).

Table 4.17 Macro FACTOR: Maximum-Likelihood Factor Method — Final Communality Estimates and Variable Weights

Variable	*Communality*	*Weight*
Y2	1.00000000	Infinity
Y4	0.72155670	3.5905001
X4	0.71577764	3.5193611
X11	0.80882398	5.2301970
X15	0.97627493	42.1528768

Note: Prior communality estimates = squared multiple correlation (SMC). Total communality: weighted = 50.492935; unweighted = 4.222433.

Table 4.18 Macro FACTOR: Maximum-Likelihood Factor Method — Final Partial Correlations Controlling Factors

Variable	*Y2*	*Y4*	*X4*	*X11*	*X15*
Y2 (mid-price)	0	0	0	0	0
Y4 (ctympg)	0	100*	–4	11	–3
X4 (hp)	0	–4	100*	18	–8
X11 (width)	0	11	18	100*	0
X15 (weight)	0	–3	–8	0	100*

Note: Prior communality estimates = squared multiple correlation (SMC). Printed values are multiplied by 100 and rounded to the nearest integer. Values greater than 0.405133 are flagged by an asterisk.

reproduced fairly accurately from the retained factors. The overall root-mean-square off-diagonal residual is 0.015867 (Table 4.21), which is the average of the sum of off-diagonal values in the residual matrix. These results confirm that the SMCs provide good and optimal communality estimates.

It is possible to estimate the factor scores, or a subject's relative standing on each of the factors, if the original subject-x variable coordinate data is available. Standardized rotated factor score values for part of the observations in the dataset are presented in Table 4.22. These factor scores can be used to rank the cars or develop scorecards based on size and price factors.

Table 4.19 Macro FACTOR: Maximum-Likelihood Factor Method — Root-Mean-Square (RMS) Off-Diagonal Partials

Y2	*Y4*	*X4*	*X11*	*X15*
0.00000000	0.05927935	0.09939887	0.10358809	0.04130537

Note: Prior communality estimates = squared multiple correlation (SMC). Overall RMS = 0.07187601.

Table 4.20 Macro FACTOR: Maximum-Likelihood Factor Method — Residual Correlations with Uniqueness on the Diagonal

Variable	*Y2*	*Y4*	*X4*	*X11*	*X15*
Y2 (mid-price)	0	0	0	0	0
Y4 (ctympg)	0	28*	–1	2	0
X4 (hp)	0	–1	28*	4	–1
X11 (width)	0	2	4	19*	0
X15 (weight)	0	0	–1	0	2

Note: Prior communality estimates = squared multiple correlation (SMC). Printed values are multiplied by 100 and rounded to the nearest integer. Values greater than 0.089545 are flagged by an asterisk.

Table 4.21 Macro FACTOR: Maximum-Likelihood Factor Method — Root-Mean-Square (RMS) Off-Diagonal Residuals

Y2	*Y4*	*X4*	*X11*	*X15*
0.00000000	0.01366703	0.02192589	0.02408045	0.00338542

Note: Prior communality estimates = squared multiple correlation (SMC). Overall RMS = 0.01586733.

Table 4.22 Macro FACTOR: Maximum-Likelihood Factor Method — VARIMAX Rotated Factor Scores Prior Communality Estimates: SMC

Observation	*ID*	*Factor1*	*Factor2*
1	12	–2.40523	–0.53880
2	7	–1.89236	–0.78951
3	30	–1.84432	–0.67368
4	29	–1.71996	–0.72990
5	32	–1.58444	–0.61692
6	44	–1.42890	4.95033
7	14	–1.23635	–0.46630
8	6	–1.21426	–0.46169
9	33	–1.20450	–0.79825
10	18	–1.14461	–0.71803
11	4	–1.09596	–0.81775
12	82	–1.08325	1.27865
13	21	–1.07670	–0.70466
14	16	–1.01238	–0.97029
15	68	–1.00938	1.60266
—[a]	—	—	—
81	55	1.28150	–0.42536
82	60	1.31290	–0.33720
83	67	1.32480	–0.42667
84	70	1.34488	–0.88456
85	78	1.34845	–0.39019
86	46	1.52128	–0.78299
87	88	1.66693	–0.45746
88	58	1.83741	–0.37577
89	45	1.84486	–0.60400
90	77	1.97584	–0.11285
91	73	2.04305	–0.62696
92	48	2.22178	–0.94602

[a] Partial list.

4.7.4.8 Investigating Interrelationships Between Multiple Attributes and Observations

Bi-plot display of both factor (Factor1 and Factor2) scores and factor loadings (Figure 4.9) is very effective for studying the relationships within observations and between variables and the interrelationships between observations and variables. The X–Y axis of the bi-plot of rotated factor analysis represents

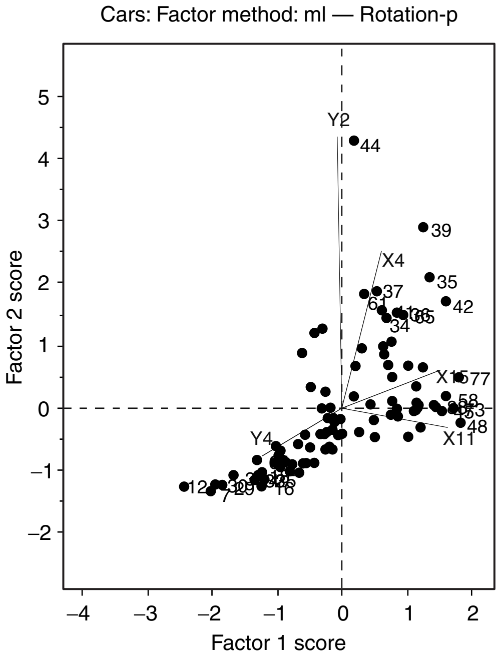


Figure 4.9 Bi-plot display of interrelationship between the first two ML factor scores and ML factor loadings. The graphics file was generated by selecting the TXT file type in the SAS macro FACTOR. The SAS graphics driver used was EMF.

the standardized Factor1 and Factor2 scores, respectively. In order to display the relationships among the variables, the factor loading for each factor is overlaid on the same plot after being multiplied by the corresponding maximum value of factor score. For example, Factor1 loading values are multiplied by the maximum value of the Factor1 scores, and Factor2 loadings are multiplied by the maximum value of the Factor2 scores. This transformation places both the variables and the observations on the same scale in the bi-plot display, as the range of factor loadings is usually shorter (–1 to +1) than for the factor scores.

Cars having larger (>75% percentile) or smaller (<25% percentile) factor scores are identified only by their ID numbers on the bi-plot to avoid overcrowded labeling. Cars with similar characteristics are displayed together in the bi-plot observational space because they have similar Factor1 and Factor2 scores. For example, small compact cars with relatively higher gas mileage, such as the Geo Metro (ID 12) and Ford Fiesta (ID 7), are grouped closer together. Similarly, cars with different attributes are displayed far apart because their Factor1, Factor2, or both factor scores are different. For example, small compact cars with relatively higher gas mileage, such as Geo Metro (ID 12), and large expensive cars with relatively lower gas mileage, such as the Lincoln Town Car (ID 42), are separated far apart in the bi-plot display.

Correlations among the multivariate attributes used in the factor analysis are revealed by the angles between any two factor loading vectors. For each variable, a factor loading vector is created by connecting the origin (0,0) and the multiplied value of the Factor1 and Factor2 loadings on the bi-plot. The angles between any two variable vectors will be:

1. Narrower (<45°) if the correlations between these two attributes are positive and larger — for example, Y2 (mid-price) and X4 (horsepower).
2. Wider (around 90°) if the correlation is not significant — for example, Y2 (mid-price) and X11 (width).
3. Closer to 180° (>135°) if the correlations between these two attributes are negative and stronger — for example, Y4 (city gas mileage) and X15 (weight).

Similarly, a stronger association between some cars (observations) and specific multi-attributes (variables) — for example, gas mileage (Y4) and Geo Metro (ID 12); mid-price (Y2) and Mercedes-Benz 300E (ID 44) — are illustrated by the closer positioning near the tip of the variable vector.

4.8 Disjoint Cluster Analysis Using SAS Macro DISJCLUS

The DISJCLUS macro is a powerful SAS application for exploring and visualizing multivariate data and for performing disjoint cluster analysis using the *k*-means algorithms. The SAS procedure FASTCLUS[20,21] is the main tool used in the DISJCLUS macro. In addition, the additional SAS/STAT procedures CLUSTER, STEPDISC, CANDISC, GPLOT, BOXPLOT, and IML modules are also utilized in the DISJCLUS macro to perform enhanced disjoint cluster analysis.

The FASTCLUS procedure is used to extract a user-specified number of clusters based on *k*-means cluster analysis. To verify the user-specified optimum cluster number, the cubic clustering criterion (CCC), pseudo F statistic (PSF), and pseudo T^2 statistic (PST2) for a number of clusters ranging from 1 to 20 are generated using Ward's method of cluster analysis in PROC CLUSTER. To perform variable selection that significantly discriminates the clusters, the backward selection method in stepwise discriminant analysis using the STEPDISC procedure is used. To test the hypothesis that significant differences exist among the clusters for multiple attributes, canonical discriminant analysis based on the CANDISC is used. SAS IML is also used to test the hypothesis of multivariate normality, which is a requirement for canonical discriminant analysis hypothesis testing. PROC GPLOT is used to generate scatterplots by cluster, quantile–quantile plots for testing of multivariate normality, and diagnostic plots based on CCC, PSF, and PST2 for selecting the optimum cluster numbers and bi-plot display of canonical discriminant analysis. The BOXPLOT procedure is used to show the between-cluster differences for intra-cluster distances and canonical discriminant functions. For more details of these SAS procedures and the available options, readers are encouraged to refer to the online references.[20,21] The advantages of using the DISJCLUS macro over the PROC FASTCLUS include:

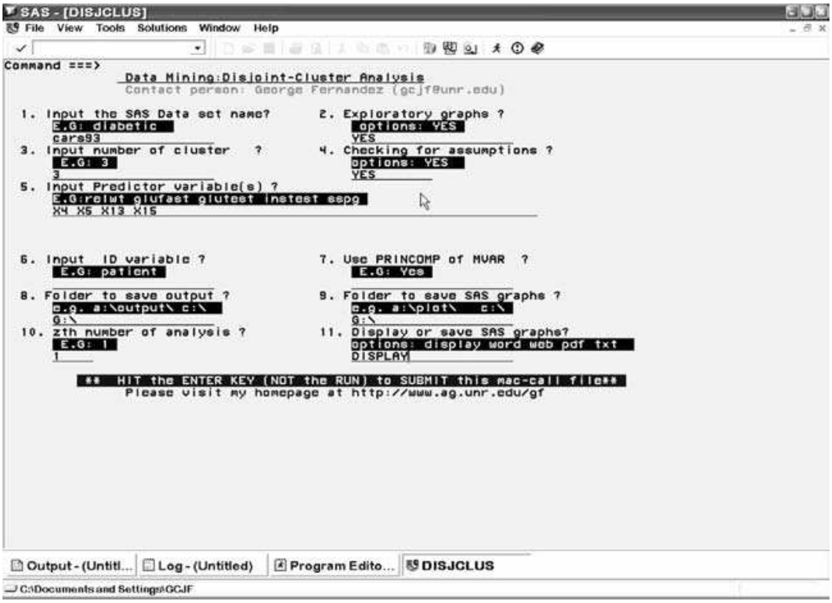
- A scatterplot matrix of all multivariate attributes by cluster groups is displayed.
- Test statistics and *P* values for testing multivariate skewness and kurtosis after accounting for the variation among the cluster groups are reported.
- Quantile–quantile (Q–Q) plots for detecting deviation from multivariate normality and plots for detecting multivariate outliers after accounting for the variation among the cluster groups are produced.
- Graphical displays of CCC, PSF, and PST2 by cluster numbers ranging from 1 to 20 verify the user-specified number of clusters in the DCA as the optimum cluster solution.
- Significance of the variables used in DCA in discriminating the clusters is verified by performing a step-wise discriminant analysis.
- DISJCLUS macro offers options for performing a disjoint cluster analysis on standardized multi-attributes, as variables with large variances tend to have more influence on the resulting clusters than those with small variances; also, DCA based on principal components of highly correlated multi-attributes is also available in the DISJCLUS macro.
- The DISJCLUS macro offers options for detecting statistical significance among cluster groups by performing canonical discriminant analysis.
- The DISJCLUS macro offers options for displaying interrelationships between the canonical discriminant scores for cluster groups and correlations among the multi-attributes in bi-plot graphs.
- Options for saving the output tables and graphics in WORD, HTML, PDF, and TXT formats are available.

Software requirements for using the DISJCLUS macro include:

- SAS/CORE, SAS/BASE, SAS/STAT, SAS/GRAPH, and SAS/IML must be licensed and installed at the site to perform the complete analysis presented here.
- SAS version 8.0 and above is recommended for full utilization.
- An active Internet connection is required for downloading the DISJCLUS macro from the book website if the companion CD-ROM is not available.

4.8.1 Steps Involved in Running the DISJCLUS Macro

1. Create an SAS dataset (permanent or temporary) containing continuous variables.
2. *If the companion CD-ROM is not available*, first verify that the Internet connection is active. Open the DISJCLUS.sas macro-call file in the SAS PROGRAM EDITOR window. Instructions are given in the Appendix regarding downloading the macro-call and sample data files from the book website. *If the companion CD-ROM is available*, the DISJCLUS.sas macro-call file will be found in the mac-call folder in the CD-ROM. Open the DISJCLUS.sas macro-call file in the SAS PROGRAM EDITOR window. Click the RUN icon to submit the macro-call file DISJCLUS.sas to open the macro-call window called DISJCLUS (Figure 4.10).
3. Input the appropriate parameters in the macro-call window by following the instructions provided in the DISJCLUS macro help file in Section 4.8.2. Users can choose the cluster exploration option, disjoint cluster analysis, and checking for multivariate normality assumptions option. After inputting all the required macro parameters, be sure the cursor is in the last input field and the RESULTS VIEWER window is closed, then hit the ENTER key (not the RUN icon) to submit the macro.
4. Examine the LOG window (only in DISPLAY mode) for any macro execution errors. If any errors are reported in the LOG window, activate the PROGRAM EDITOR window, resubmit the DISJCLUS.sas macro-call file, check the macro input values, and correct any input errors.
5. Save the output files. If no errors are found in the LOG window, activate the PROGRAM EDITOR window, resubmit the DISJCLUS.sas macro-call file, and change the macro input value from DISPLAY to any other desirable format (see Section 4.8.2). The printout of the disjoint cluster analysis and exploratory graphs can be saved as a user-specified format file in the user-specified folder.



```
Command ===>
                Data Mining:Disjoint-Cluster Analysis
                Contact person: George Fernandez (gcjf@unr.edu)

 1. Input the SAS Data set name?           2. Exploratory graphs ?
    E.G: diabetic                             options: YES
    cars93                                    YES
 3. Input number of cluster   ?            4. Checking for assumptions ?
    E.G: 3                                    options: YES
    3                                         YES
 5. Input Predictor variable(s) ?
    E.G:relwt glufast glutest instest sspg
    X4 X5 X13 X15


 6. Input  ID variable ?                   7. Use PRINCOMP of MVAR  ?
    E.G: patient                              E.G: Yes

 8. Folder to save output ?                9. Folder to save SAS graphs ?
    e.g. a:\output\ c:\                       e.g. a:\plot\   c:\
    G:\                                       G:\
10. zth number of analysis ?              11. Display or save SAS graphs?
    E.G: 1                                    options: display word web pdf txt
    1                                         DISPLAY

    ** HIT the ENTER KEY (NOT the RUN) to SUBMIT this mac-call file**
         Please visit my homepage at http://www.ag.unr.edu/gf
```

Figure 4.10 Screen copy of DISJCLUS macro-call window showing the macro-call parameters required for performing complete disjoint cluster analysis.

4.8.2 Help File for SAS Macro DISJCLUS

1. **Macro-call parameter:** Input SAS dataset name (required parameter).
   Descriptions and explanation: Include the name of the temporary (member_name) or permanent (libname.member_name) SAS dataset for which a disjoint cluster analysis will be performed. It should be in the form of coordinate data: rows (cases) × columns (variables).
   Options/examples:
   > **Permanent SAS dataset:** gf.cars93 (LIBNAME: gf; SAS dataset name: cars93)
   > **Temporary SAS dataset:** cars93 (SAS dataset name)
2. **Macro-call parameter:** Exploratory cluster analysis (optional parameter).
   Descriptions and explanation: Displays the results of cluster groupings in a simple two-variable scatterplot display; verifies the optimum cluster number by CCC, pseudo *F* statistic (PSF), and pseudo T^2 statistics (PST2); and selects variables using the backward elimination method in stepwise discriminant analysis.

Options/examples:

> **Yes:** (1) Results of the disjoint cluster analysis are displayed in a simple scatterplot matrix if the number of multi-attributes is less than eight. (2) Plots of CCC, PSF, and PST2 against cluster numbers ranging from 1 to 20 for verifying the optimum cluster number are produced. (3) Results of backward elimination variable selection in stepwise discriminant analysis are produced. *Note:* Verification of cluster groupings by canonical discriminant analysis or checking for multivariate normality is not performed if YES is selected in this field.
>
> **Blank:** Only disjoint cluster analysis and canonical discriminant analyses are performed. Exploratory cluster analysis is not performed.

3. **Macro-call parameter:** Input the number of disjoint clusters (required options).

   Descriptions and explanation: Input the number of disjoint clusters you would like to extract.

   Options/examples:

   3 10 25

4. **Macro-call parameter:** Check for assumptions (optional statement).

   Descriptions and explanation: To check for multivariate normality assumption and detect for any extreme outliers or influential data, input YES. Multivariate normality is a requirement for canonical discriminant analysis but is not a requirement for DCA. If this field is left blank, this step will be omitted.

   Options/examples:

   Yes: Statistical estimates of multivariate skewness and kurtosis and their statistical significance, Q–Q plots for checking for multivariate normality, and multivariate outlier detection plots are produced.

   Blank: If the macro input field is left blank, no statistical estimates for checking for multivariate normality and detecting for outliers are performed.

5. **Macro-call parameter:** Input continuous multi-attribute variable names (required parameter for performing disjoint cluster analysis on coordinate data).

   Descriptions and explanation: Input continuous multi-attribute names from the SAS dataset that are to be included in the DISJCLUS analysis.

   Options/examples:

   X4 X8 X11 X15 (names of continuous multi-attributes)

6. **Macro-call parameter:** Input ID variable (optional statement).

Descriptions and explanation: Input the name of the variable to be treated as the ID. If this field is left blank, a character variable will be created from the observational number and will be used as the ID variable.

Option/example:

Car ID model

7. **Macro-call parameter:** Use PRINCOMP of MVAR (optional statement).

   Descriptions and explanation: Input YES if severe multicollinearity exists among the multi-attributes and to perform disjoint cluster using all principal components of the variables specified in macro input option #5.

   Options/examples:

   Yes: Cluster analysis based on PCA.

   Blank: Cluster analysis based on standardized data.

8. **Macro-call parameter:** Folder to save SAS output (optional statement).

   Descriptions and explanation: To save the SAS output files in a specific folder, input the full path of the folder. The SAS dataset name will be assigned to the output file. If this field is left blank, the output file will be saved in the default folder.

   Options/examples:

   Possible values

   c:\output\ — folder named "OUTPUT"

   s:\george\ — folder named "George" in network drive S

   Be sure to include the back-slash at the end of the folder name.

9. **Macro-call parameter:** Folder to save SAS graphics (optional statement).

   Descriptions and explanation: To save the SAS graphics files in the EMF format suitable for inclusion in PowerPoint presentations, specify the output format as TXT in version 8.0 or later. In pre-8.0 versions, all graphic format files will be saved in a user-specified folder. If the graphics folder field is left blank, the graphics file will be saved in the default folder.

   Options/examples:

   Possible values

   c:\output\ — folder named "OUTPUT"

10. **Macro-call parameter:** *z*th number of run (required statement).

    Descriptions and explanation: SAS output files will be saved by forming a file name from the original SAS dataset name and the counter value provided in this field. For example, if the original

SAS dataset name is “gf.cars93” and the counter number included is 1, the SAS output files will be saved as “gf.cars931.*” in the user-specified folder. By changing the counter value, users can avoid replacing the previous SAS output files with new outputs.

11. **Macro-call parameter:** Display or save SAS output (required statement).

    Descriptions and explanation: Option for displaying all output files in the OUTPUT window or saving as a specific format in a folder specified in macro input option #8.

    Options/examples:

    Possible values

    DISPLAY: Output will be displayed in the OUTPUT window, all SAS graphics will be displayed in the GRAPHICS window, and system messages will be displayed in LOG window.

    WORD: Output and all SAS graphics will be saved together in the user-specified folder and displayed in the VIEWER window as a single RTF format file (version 8.0 and later if MS Word is installed on the system). In pre-8.0 versions, the SAS output is saved only as a text file and all graphics files (CGM) are saved separately in a user-specified folder (macro input option #9).

    WEB: Output and graphics are saved in the user-specified folder and are viewed in the results VIEWER window as a single HTML file (version 8.0 and later). In pre-8.0 versions, the SAS output is saved only as a text file and all graphics files (GIF) are saved separately in a user-specified folder (macro input option #9).

    PDF: Output and graphics are saved in the user-specified folder and are viewed in the results VIEWER window as a single PDF file (version 8.2 and later if Adobe Acrobat Reader is installed on the system). In pre-8.2 versions, the SAS output is saved only as a text file and all graphics files (PNG) are saved separately in a user-specified folder (macro input option #9).

    TXT: Output will be saved as a TXT file in all SAS versions; no output will be displayed in the OUTPUT window. All graphic files will be saved in the EMF format (version 8.0) or CGM format (pre-8.0 versions) in a user-specified folder (macro input option #9).

    Note: System messages are deleted from the LOG window if DISPLAY is not selected in macro input option #11.

4.8.3 Case Study 3: Disjoint Cluster Analysis of 1993 Car Attribute Data

4.8.3.1 Study Objectives

1. **Checking for suitability of car93 data for *k-mean cluster analysis:*** Perform the preliminary *k*-mean cluster analysis with the user-specified cluster number and check the multi-attribute data for multivariate outliers and normality after adjusting for between-cluster variations. This is a requirement for hypothesis tests on cluster differences based on canonical discriminant analysis. Perform stepwise discriminant analysis using backward elimination routine to confirm that the multi-attributes used in the cluster analysis are statistically significant.
2. ***k-Mean clustering:*** Extract disjoint clusters from the multivariate dataset and characterize the cluster differences.
3. **Canonical discriminant analysis:** Verify the results of disjoint cluster groupings using MANOVA test statistics and canonical discriminant analysis.
4. **Interrelationships:** Investigate the interrelationships between the cluster groupings and the multi-attributes based on canonical discriminant function scores and structure loadings.

4.8.3.2 Data Descriptions

Data name	Permanent SAS dataset "CARS93" located in the library "GF"
Multi-attributes	X4: hp; maximum horsepower X5: rpm; revolutions per minute at maximum horsepower X13: rseatrm; rear seat room (inches) X15: weight (pounds)
Number of observations	92
Data source	Lock[17]

Open the DISJCLUS macro-call window in SAS (Figure 4.10) and input the appropriate macro input values by following the suggestions given in the help file (Section 4.8.2).

4.8.3.3 Scatterplot Matrix of Cluster Separation, Variable Selection, and Optimum Cluster Number Estimation

Input X4, X5, X13, and X15 as the multi-attributes in (macro input option #2). Input YES in field #3 to perform scatterplot matrix analysis of cluster groupings and stepwise variable selections. Submit the DISJCLUS macro to output the disjoint cluster analysis results and scatterplot matrix plot. Only selected output and graphics generated by the DISJCLUS macro are interpreted below.

4.8.3.4 Scatterplot Matrix of Cluster Separation

Characteristics of standardized multi-attributes included in the disjoint cluster analysis are presented in Table 4.23. The total standard deviation (STD) for all attributes is equal to 1, because all the variables are standardized to a zero mean and unit standard deviation prior to clustering. The pooled within-cluster standard deviation (within-STD) describes the average within-cluster variability. A relatively homogenous cluster should have a smaller within-STD. The percentage of variability in each standardized attribute attributed to the cluster differences is given by the R^2 statistic. For example, 69% of the variation in X15 (weight) could be attributed to between-cluster differences. The ratio of between-cluster variance to within-cluster variance is provided in the last column: $R^2/(1 - R^2)$. Variable X15 accounts for most of the variation among the clusters, and X13 accounts for the least amount of between-cluster variations. The Overall row provides the average estimates of between and within-cluster estimates pooled across all variables. The estimated pseudo F statistic for disjoint clustering is estimated by the following formula:

$$[(\ [R^2)/(c\ \text{-}\ 1)]\ )/(\ [(1\text{-}R^2)/(n\ \text{-}\ c)]\ )$$

where R^2 is the observed overall R, c is the number of clusters, and n is the number of observations in the dataset. The pseudo F statistic is an

Table 4.23 Macro FASTCLUS: Characteristics of Standardized Multi-Attributes Used in Disjoint Cluster Analysis

Variable	*Total Standard Deviation (TSD)*	*Within-STD*	*R-Squared (R^2)*	*$R^2/(1 - R^2)$*
X4	1.00000	0.58358	0.666920	2.002280
X5	1.00000	0.70214	0.517838	1.073990
X13	1.00000	0.84483	0.302305	0.433290
X15	1.00000	0.55941	0.693936	2.267295
Overall	1.00000	0.68092	0.546592	1.205518

Note: Statistic used for variables was the pseudo *F* statistic (PSF) = 53.65.

overall indicator of the measure of fit. The general goal is to maximize the pseudo *F* statistic when selecting the optimum number of clusters or significant attributes used in clustering.

Cluster means and standard deviations for the three extracted clusters are presented in Tables 4.24 and 4.25, respectively. Based on the cluster mean values, we can come to the following conclusions: Clusters 1 and 2 are mainly differentiated by X4 and X15. Variables X4 and X5 separate clusters 1 and 3. All four attributes equally separate clusters 2 and 3. Cluster standard deviation provides measures of within-cluster homogeneity, and, overall, clusters 2 and 3 are more homogeneous than cluster 1.

The scatter plot matrix among the four attributes presented in Figure 4.11 reveals the strength of correlation, presence of any outlying observations, and nature of bi-directional variation in separating the cluster groupings. Variables X15 and X4 appear to have strong correlations in the scatterplot. The best bi-variable cluster separation is observed in the scatterplot between X4 and X5. Even though clusters 2 and 3 are more homogeneous than cluster 1 based on within-cluster standard deviations, a few extreme observations belonging to cluster 2 show up in the scatterplot based on attribute X4.

Table 4.24 Macro FASTCLUS: Cluster Mean Values of Standardized Multi-Attributes Used in Disjoint Cluster Analysis

	Cluster Means			
Cluster	*X4*	*X5*	*X13*	*X15*
1	1.700924242	0.900060957	–0.052278189	0.790546535
2	–0.661424429	0.399999715	–0.473951479	–0.810304619
3	0.280466657	–0.953149346	0.714086862	0.868975379

Table 4.25 Macro FASTCLUS: Cluster Standard Deviation of Standardized Multi-Attributes Used in Disjoint Cluster Analysis

	Cluster Standard Deviations			
Cluster	*X4*	*X5*	*X13*	*X15*
1	0.901492494	0.639618576	1.031975011	0.501535524
2	0.530477353	0.742979503	0.745301788	0.596992931
3	0.495572337	0.661732590	0.913930226	0.521749822

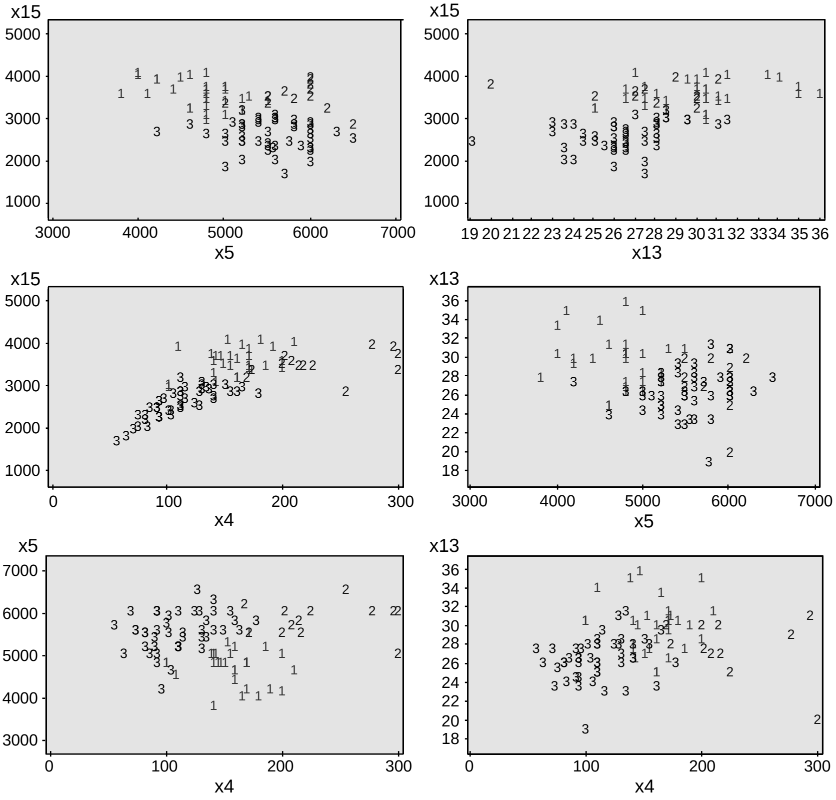


Figure 4.11 Scatter plot matrix illustrating the success of disjoint clustering in a two-dimensional display of multiple attributes. The graphics file was generated by selecting the TXT file type in the SAS macro DISJCLUS. The SAS graphics driver used was EMF.

4.8.3.5 Significant Variable Selection

The backward elimination method in the stepwise discriminant (PROC STEPDISC) analysis is used to select the significant variables for effective clustering. After trying out many combinations of variables, the final four attributes presented in Table 4.26 are selected as the best subset of variables for effective clustering of the cars93 data. The backward elimination variable selection results presented in Table 4.26 show that, based on the STEPDISC default cutoff values, none of the four variables can be dropped from the analysis. However, X5 and X4 can be identified as the most significant variables with the largest partial R^2 and F statistic responsible for effective clustering.

Table 4.26 Macro FASTCLUS: Summary Statistics of Backward Elimination Method in Testing for Cluster Separations Using Stepwise Discriminant Analysis

Variable	*Label*	*Partial R^2*	*F Value*	*Pr > F*
X4	hp	0.1712	8.68	0.0004
X5	rpm	0.2275	12.37	<.0001
X13	rseatrm	0.0511	2.26	0.1105
X15	weight	0.1297	6.26	0.0029

Note: Statistics for removal, DF = 2, 84. No variables can be removed.

4.8.3.6 Determining Optimum Number of Clusters

To determine the optimum cluster solutions, Ward's method of cluster analysis available in the PROC CLUSTER is used to generate clustering criteria such as the cubic clustering criterion (CCC), pseudo F statistic (PSF), and pseudo T^2 statistic (PST2). A trend plot of CCC (Y axis) vs. the number of clusters (X axis) plot presented in Figure 4.12 can be used to select the optimum number of clusters in the data. According to Khattree and Naik,[19] values of the CCC greater than 2 or 3 indicate good clusters. Values between 0 and 2 indicate potential clusters but should be evaluated with caution; large negative values may indicate the presence of outliers. In the case of the cars93 data, however, all CCC values are negative, indicating the presence of multivariate outliers; however, the CCC values take a big jump at the three-cluster solution. Thus, three clusters are selected tentatively.

An overlay plot of PST2 and PSF (Y axis) vs. the number of clusters (X axis) plot presented in Figure 4.13 can also be used to select the optimum number of clusters in the data. Starting from the large values in the X axis, when we move left a big jump in the PST2 value occurs for cluster number 2. Similarly, a relatively large PSF occurs at cluster number 2 when we move from right to left in the X axis of the overlay plot. The PST2 and PSF statistics both indicate two clusters as the optimum solution. Based on these results, we select two as the number of potential clusters. These graphical methods suggest between two and three clusters as being the optimal number. Thus, the three-cluster solution is selected for k-mean cluster analysis and is subsequently verified by canonical discriminant analysis. The results of canonical discriminant analysis are presented in the next section.

In the second phase of analysis, the results of checking for multivariate normality, cluster solution verification by canonical discriminant analysis and the bi-plot analysis results are discussed. To perform the second phase

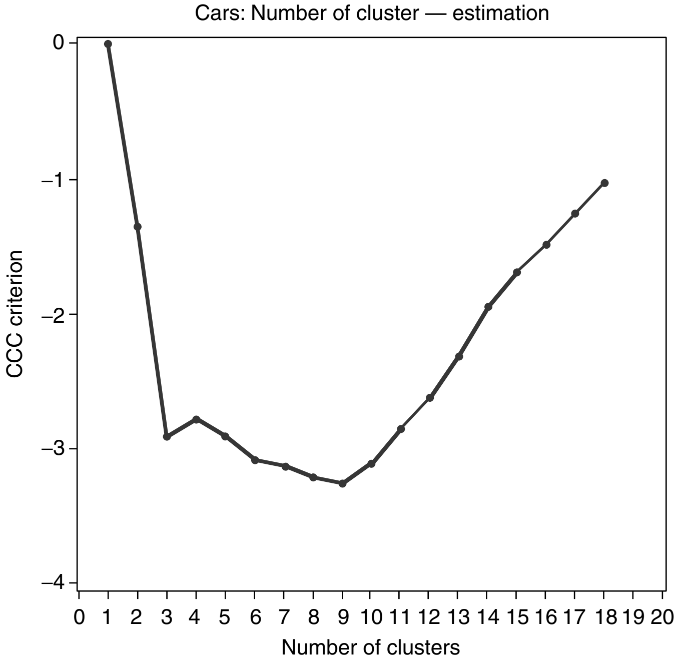


Figure 4.12 Trend plot showing the relationship between the number of clusters formed and the cubic clustering criterion (CCC) used in deciding the number of optimal clusters. The graphics file was generated by selecting the TXT file type in the SAS macro DISJCLUS. The SAS graphics driver used was EMF.

of the analysis, open the DISJCLUS macro-call window in SAS (Figure 4.10) and input the appropriate macro-input values by following the suggestions given in the help file (Section 4.8.2; input X4, X5, X13, and X15 as the multi-attributes in macro-input option #2). Leave macro field #3 blank to verify cluster solution by canonical discriminant analysis (CDA) and to skip scatterplot matrix analysis of cluster groupings and stepwise variable selections. Also, check for both multivariate normality and the presence of influential observations and then input YES in macro field #4. Submit the DISJCLUS macro to output the CDA results and multivariate normality check. Only selected output and graphics generated by the DISJCLUS macro are interpreted below.

The main objective of CDA is to extract a set of linear combinations of the quantitative variables that best reveal the differences among the groups or clusters. Given two or more clusters and several quantitative attributes, the CDA extracts linear combinations of the quantitative variables (canonical variables) that maximize between-class variation. Thus, the extracted canonical functions can be used to verify the differences

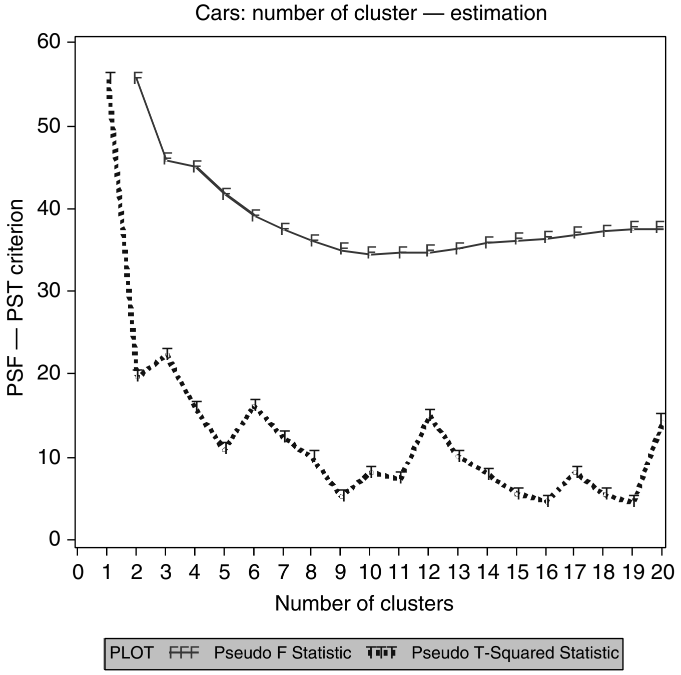


Figure 4.13 Overlaid plot showing the relationship between the number of clusters formed and the pseudo *F* and *T* squared statistics used in deciding the number of optimal clusters. The graphics file was generated by selecting the TXT file type in the SAS macro DISJCLUS. The SAS graphics driver used was EMF.

among the clusters. CDA is a parametric technique, and the validity of hypothesis testing in CDA depends on the assumption of multivariate normality. More detailed accounts of CDA are presented in Chapter 6.

4.8.3.7 Checking for Multivariate Normality

Checking for multivariate normality is performed by computing Mardia's multivariate skewness and kurtosis measures, as well as chi-square testing,[18] after adjusting for between-cluster variations. The estimates of multivariate normality test statistics and the corresponding *p* values are presented in Table 4.27. Smaller multivariate skewness and kurtosis values and non-significant *p* values clearly indicate that the distribution of four multi-attributes after adjusting for the differences between clusters does satisfy multivariate normality. This is further confirmed by visually assessing the chi-squared quantile–quantile plot of the squared Mahalanobis distances (Figure 4.14) which follows closely the 45° reference line.

Table 4.27 Macro FASTCLUS: Checking for Multivariate Normality, a Requirement for Testing for Cluster Separations Using Canonical Discriminant Analysis

Multivariate Normality Test Statistics	
1.850	Multivariate skewness
27.744	Skewness chi-square
0.116	Skewness *p* value
26.046	Multivariate kurtosis
1.401	Kurtosis *z* value
0.161	Kurtosis *p* value

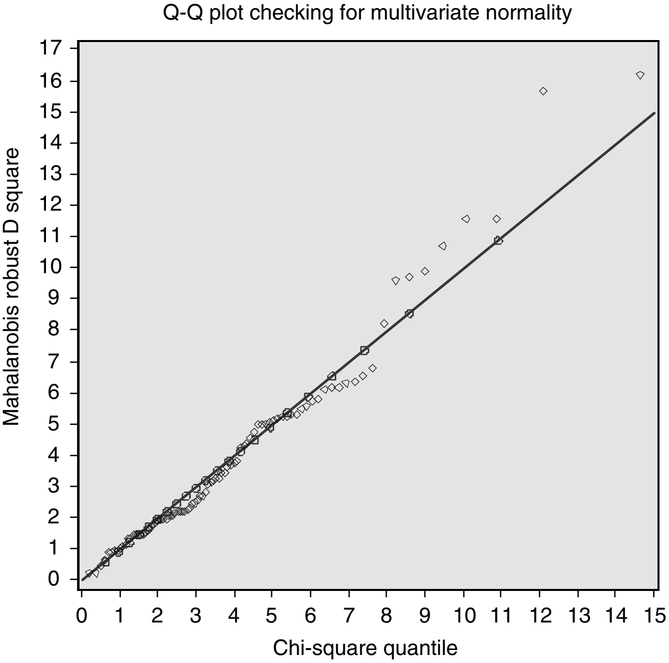


Figure 4.14 Quantile–quantile (Q-Q) plot for assessing multivariate normality after accounting for the differences in clusters. The graphics file was generated by selecting the TXT file type in the SAS macro DISJCLUS. The SAS graphics driver used was EMF.

4.8.3.8 Checking for the Presence of Multivariate Influential Observations

After adjusting for between-cluster differences, detecting for influential observations is performed by computing the robust distance square statistic (RDSQ) and the difference between RDSQ and the quantiles of the expected chi-square value (DIFF). Multivariate influential observations are identified when the DIFF values exceed 2.5. The estimates of RDSQ and DIFF values for the one multivariate influential observation are presented in Table 4.28. Observation number 38 belongs to cluster 2 and is identified as the most influential observation. The presence of one multivariate influential observation is also visually assessed by a graphical display of DIFF values vs. the quantile of chi-square values (Figure 4.15). The impact of this one extreme influential observation on DCA analysis outcomes can be verified by excluding this extreme observation and examining the effects on hypothesis tests in CDA.

4.8.3.9 Validating Cluster Differences Using CDA

A matrix of square distance values between the clusters estimated using the CDA is performed in Table 4.29. Variations within cluster distances are shown in Figure 4.16. The differences between cluster distances show that clusters 2 and 3 are separated far apart relative to the distances between clusters 1 and 2 and 1 and 3. In addition, highly significant *p* values for testing the Mahalanobis distances reveal that the distances between the clusters are statistically significant (Table 4.30).

4.8.3.10 Checking for Significant Cluster Groupings by CDA

Canonical discriminant analysis extracts canonical discriminat functions, which are linear combinations of the original variables. Assuming that the multivariate normality assumption is not seriously violated, cluster differences for combined canonical functions can be examined by multivariate ANOVA (MANOVA). CDA analysis produces an output of four multivariate

Table 4.28 Macro FASTCLUS: Checking for the Presence of Multivariate Influential Observations Using Canonical Discriminant Analysis

Cluster	*ID*	*Robust Distance Square Statistic (RDSQ)*	*Chi-Square*	*DIFF*
2	38	19.2509	12.0939	3.55929

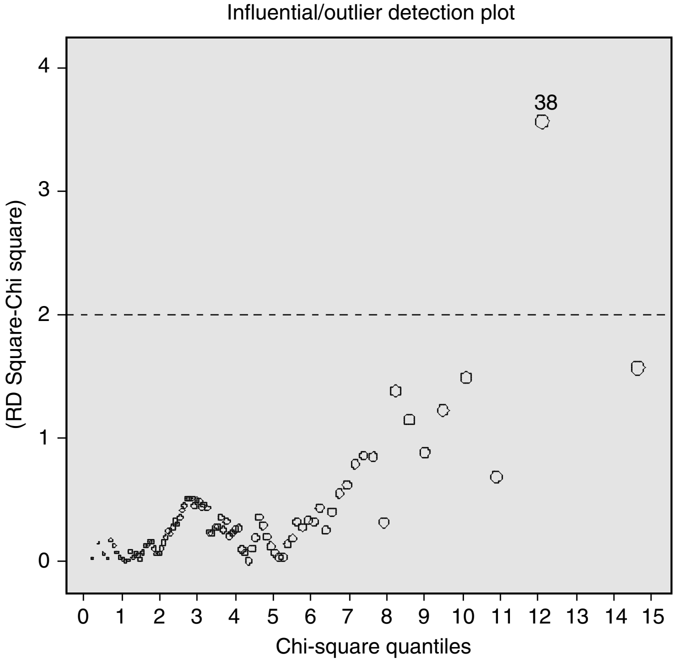


Figure 4.15 Multivariate outlier detection plot based on robust squared distances after accounting for the differences among the clusters. The graphics file was generated by selecting the TXT file type in the SAS macro DISJCLUS. The SAS graphics driver used was EMF.

Table 4.29 Macro FASTCLUS: A Table of Square Distance Matrix Between the Extracted Clusters Estimated Using Canonical Discriminant Analysis

	Squared Distance to Cluster		
From Cluster	*1*	*2*	*3*
1	0	12.11870	11.95887
2	12.11870	0	16.08562
3	11.95887	16.08562	0

ANOVA test statistics and the *p* values for testing the significance of between-cluster differences for canonical functions. All four MANOVA tests clearly indicate that large significant differences exist for at least one of the clusters (Table 4.31). The characteristics of each canonical function and their significance can be performed using the canonical correlations.

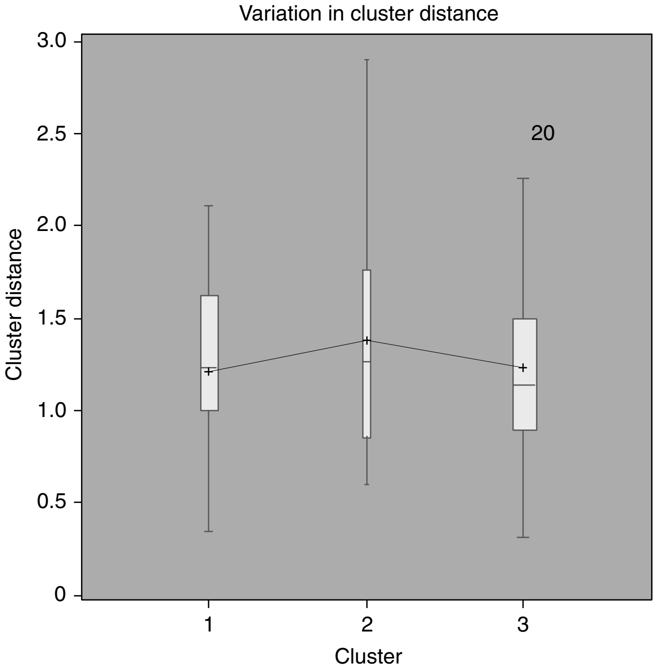


Figure 4.16 Box-plot display of intra-cluster distances by clusters assessing the homogeneity within clusters. The graphics file was generated by selecting the TXT file type in the SAS macro DISJCLUS. The SAS graphics driver used was EMF.

Table 4.30 Macro FASTCLUS: *p* Values Indicating the Statistical Significance of Square Distance Matrix Between the Extracted Clusters Using Canonical Discriminant Analysis

	Probability > Mahalanobis Distance for Squared Distance to Cluster		
From CLUSTER	*1*	*2*	*3*
1	1.0000	<.0001	<.0001
2	<.0001	1.0000	<.0001
3	<.0001	<.0001	1.0000

These maximal multiple correlations between the first canonical function and the cluster differences are called the first canonical correlations. The second canonical correlation is obtained by finding the linear combination uncorrelated with the first canonical variable that has the highest possible multiple correlations with the groups. In CDA, the process of extracting canonical functions is repeated until extraction

Table 4.31 Macro FASTCLUS: MANOVA Results Testing the Statistical Significance of Extracted Clusters Estimated Using Canonical Discriminant Analysis

Statistic	*Value*	*F Value*	*Numerator of Degrees of Freedom*	*Denominator of Degrees |of Freedom*	*Pr > F*
Wilks' lambda	0.11275384	27.69	12	168	<.0001
Pillai's trace	1.21554491	21.95	12	170	<.0001
Hotelling –Lawley trace	4.95723585	34.41	12	127.62	<.0001
Roy's greatest root	4.27636805	60.58	6	85	<.0001

of the maximum number of canonical functions, which is equal to the number of groups minus one or the number of variables in the analysis, whichever is smaller. Approximately 71% of the cluster variations can be accounted for by the first canonical function. The second independent functions account for the rest of the variation between clusters (Table 4.32). Both canonical correlations are statistically significant (Table 4.33). The degree of cluster separation by each canonical function and the distributional properties of canonical functions by clusters can be visually examined in box plots (Figures 4.17 and 4.18). The first canonical discriminant function discriminates cluster 3 from the other two clusters more effectively (Figure 4.17), while the second canonical discriminant function separates cluster 2 from the other two clusters (Figure 4.18).

The structure coefficients presented in Table 4.34 indicate the simple correlations between the variables and the canonical discriminant functions. These structure loadings are commonly used when interpreting the meaning of the canonical variable because the structure loadings appear to be more stable and they allow for the interpretation of canonical variables in a manner that is analogous to factor analysis. The first canonical function has larger positive correlations with hp, weight, and rseatrm and negative loadings for rpm. The second canonical function has significant positive loadings for hp and rpm.

Table 4.32 Macro FASTCLUS: Estimates of Canonical Correlations Between the Extracted Clusters and the Canonical Functions Using Canonical Discriminant Analysis

				*Eigenvalues of Inv(E) * H = $CanR^2/(1 - CanR^2)$*			
Canonical Correlation	*Adjusted Canonical Correlation*	*Approximate Standard Error*	*Squared Canonical Correlation*	*Eigenvalue*	*Difference*	*Proportion*	*Cumulative*
0.857405	0.849361	0.028075	0.735143	2.7756	1.6473	0.7110	0.7110
0.728117	0.727755	0.049804	0.530155	1.1284		0.2890	1.0000

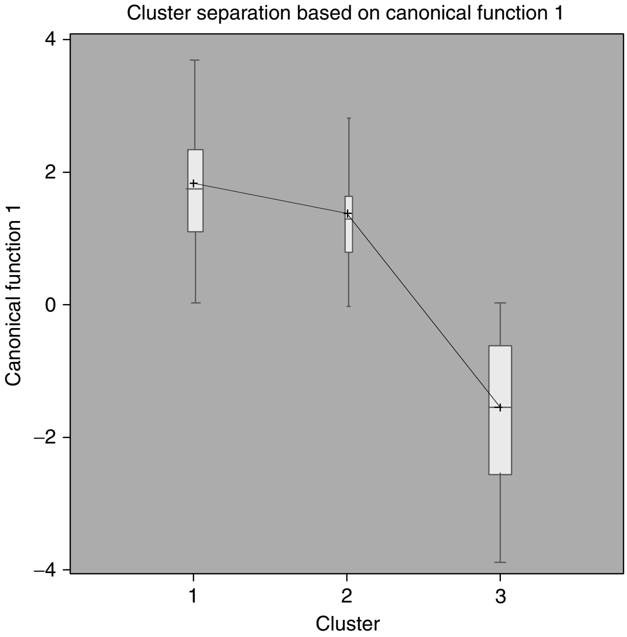


Figure 4.17 Box-plot display of canonical discriminant analysis function 1 by clusters assessing the success of cluster separation. The graphics file was generated by selecting the TXT file type in the SAS macro DISJCLUS. The SAS graphics driver used was EMF.

4.8.3.11 Bi-Plot Display of Canonical Discriminant Function Scores and the Cluster Groupings

The extracted canonical function scores can be used to plot pairs of canonical functions in two-dimensional bi-plots to aid visual interpretation of cluster differences. Interrelationships among the four multi-attributes and discriminations of the three clusters are presented in Figure 4.19. The first canonical function, which has the largest loadings on all four attributes, successfully discriminated the small, less powerful cars in cluster 3 from the other two cluster groups, 1 and 2. The second canonical function, which has larger loadings on X4 and X5, successfully discriminated the cluster 2 group containing very powerful sports cars from clusters 3 and 2. The cluster mean value for each cluster is shown by the + sign in the bi-plot. The circle around the + sign in each cluster shows the 95% confidence circle for the cluster mean values. The larger circle for cluster group 2 indicates that the within-cluster variability is relatively higher for cluster 2. The smaller cluster size might be one of the reasons for this large variation in cluster 2.

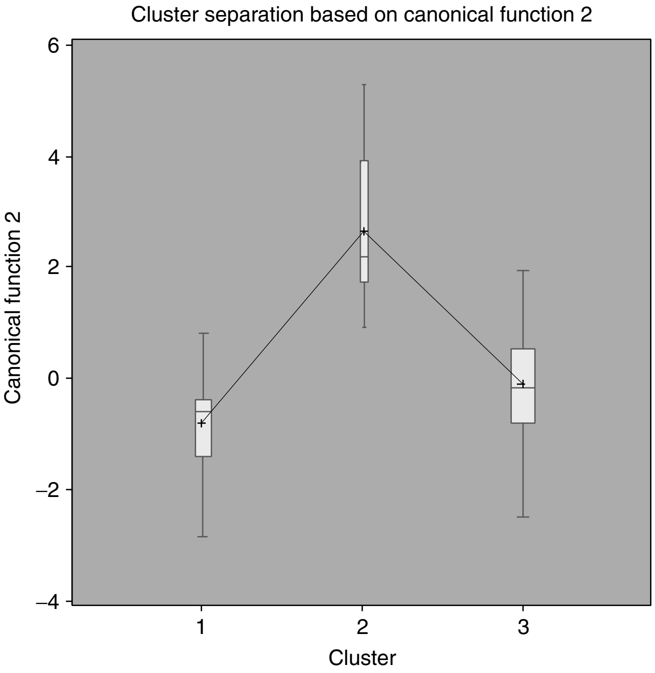


Figure 4.18 Box-plot display of the second canonical discriminant analysis function by clusters assessing the success of cluster separation. The graphics file was generated by selecting the TXT file type in the SAS macro FACTOR. The SAS graphics driver used was EMF.

Table 4.33 Macro FASTCLUS: Testing for the Significance of Canonical Correlations between the Extracted Clusters and the Canonical Functions in the Canonical Discriminant Analysis

Likelihood Ratio	*Approximate F Value*	*Numerator of Degrees of Freedom*	*Denominator of Degrees of Freedom*	*Pr > F*
0.12444178	38.53	8	168	<.0001
0.46984534	31.97	3	85	<.0001

Note: Test of H_0: The canonical correlations in the current row and all that follow are zero.

A partial list of cluster memberships, sorted canonical function scores, and the original attributes are presented in Table 4.35. These canonical function scores can be used as the basis for scorecard development in selecting observations within each cluster.

Table 4.34 Macro FASTCLUS: Correlation Coefficients Between the Multi-Attributes and the Canonical Functions in the Canonical Discriminant Analysis

		Total Canonical Structure	
Variable	*Label*	*Can1*	*Can2*
X4	hp	0.751624	0.643557
X5	rpm	–0.573941	0.736941
X13	rseatrm	0.602103	–0.259843
X15	weight	0.976968	0.118031

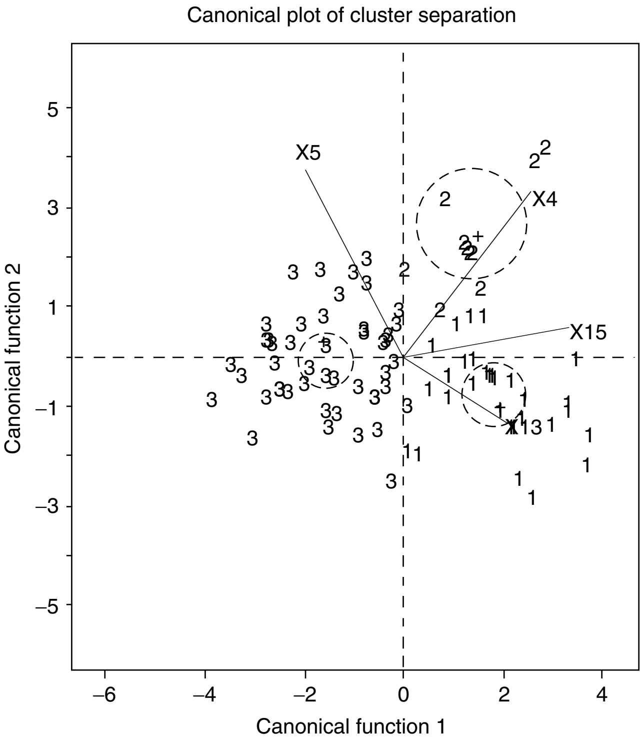


Figure 4.19 Bi-plot display of interrelationship between the first two canonical discriminant functions and cluster groupings. The graphic file was generated by selecting the TXT file type in the SAS macro DISJCLUS. The SAS graphics driver used was EMF.

Table 4.35 Macro FASTCLUS: Table of Canonical Function Scores and Original Multi-Attributes by Disjoint Clusters Derived from Canonical Discriminant Analysis

Observation	*ID*	*Can1*	*Can2*	*X4*	*X5*	*X13*	*X15*
Cluster 1							
1	53	0.01795	–1.88285	100	4800	30.5	2970
2	54	0.22754	–1.93346	100	4800	30.5	3080
3	51	0.47224	–0.62882	141	5000	30.5	3085
4	74	0.53410	0.24159	160	5200	28.5	3200
5	49	0.82925	–0.36899	160	4600	25	3240
6	81	0.82925	–0.36899	160	4600	25	3240
7	59	0.83399	–0.81358	140	4800	27.5	3325
8	61	0.99980	0.68009	172	5500	31	3405
9	50	1.19370	–0.13259	153	5300	31	3515
10	64	1.28770	0.79878	185	5200	27.5	3510
11	55	1.31764	–0.56679	142	5000	26.5	3705
12	89	1.37073	–0.06814	170	4800	26.5	3495
13	79	1.58336	0.82579	200	5000	28.5	3450
—[a]	—	—	—	—	—	—	—
18	60	1.88125	–1.08000	145	4800	30	3735
25	73	2.37714	–0.85743	151	4800	27	4100
26	88	2.55747	–2.83492	109	4500	34	3960
27	45	2.94244	–1.37543	170	4200	29.5	3910
28	36	3.25171	–1.08822	200	4100	35	3620
29	58	3.26142	–0.93881	190	4200	30	3950
30	42	3.44300	–0.02979	210	4600	31.5	4055
31	48	3.63142	–2.16622	165	4000	33.5	4025
32	77	3.67409	–1.58960	180	4000	30.5	4105
Cluster 2							
33	37	.	.	300	5000	.	3380
34	43	.	.	255	6500	.	2895
35	91	–0.03995	1.75543	168	6200	30	3245
36	57	0.68323	0.93581	172	5500	28	3375
37	41	0.78466	3.15466	225	6000	25	3515
38	71	1.18103	2.28448	202	6000	27.5	3730
39	56	1.23350	2.19906	214	5800	30	3490
40	40	1.28696	2.07761	208	5700	27	3640
41	44	1.33150	2.06196	217	5500	27	3525
42	34	1.48800	1.38289	200	5500	30	3560
43	38	1.65309	5.28198	300	6000	20	3805
44	39	2.58339	3.92161	278	6000	29	4000
45	35	2.80209	4.21129	295	6000	31	3935

Table 4.35 (continued)

Observation	*ID*	*Can1*	*Can2*	*X4*	*X5*	*X13*	*X15*
Cluster 3							
46	12	–3.90210	–0.85716	55	5700	27.5	1695
47	30	–3.49880	–0.17268	70	6000	27.5	1965
48	29	–3.31461	–0.39557	73	5600	23.5	2045
49	7	–3.05767	–1.62723	63	5000	26	1845
50	32	–2.81480	–0.80053	82	5200	24	2055
51	20	–2.79309	0.66827	100	5750	19	2450
52	4	–2.78466	0.31250	92	6000	26.5	2270
53	21	–2.78026	0.34132	92	6000	26	2295
54	6	–2.73703	0.30100	92	6000	26.5	2295
55	18	–2.62430	–0.10829	92	5550	23.5	2285
56	33	–2.56014	–0.63449	81	5500	26	2240
57	25	–2.55056	–0.67243	74	5600	25.5	2350
58	16	–2.36008	–0.68281	81	5500	26	2345
—[a]	—	—	—	—	—	—	—
77	31	–0.84110	0.47243	135	5400	23	2950
78	76	–0.80701	1.45741	155	6000	28	2910
79	90	–0.76717	1.94288	178	5800	26	2810
80	69	–0.61767	–0.80922	110	5200	28	2880
81	23	–0.59861	–0.81382	110	5200	28	2890
82	9	–0.54901	–1.47111	105	4600	24	2850
83	83	–0.46174	0.25539	130	5600	27	3085
84	68	–0.42794	–0.31673	130	5100	26	2920
85	92	–0.42186	–0.58795	114	5400	29.5	2985
86	87	–0.39711	0.32907	134	5800	31.5	2985
87	15	–0.36124	0.44313	140	5600	28	3040
88	8	–0.29328	–2.48557	96	4200	27.5	2690
89	85	–0.26296	–0.13173	130	5400	28.5	3030
90	72	–0.19919	0.64588	150	5600	28.5	3050
91	66	–0.12547	0.94880	164	5600	29.5	2970
92	47	0.02575	–0.99448	110	5200	28.5	3195

[a] Partial list.

A partial list of cluster memberships, sorted canonical function scores, and the original attributes are presented in Table 4.35. These canonical function scores can be used as the basis for scorecard development in selecting observations within each cluster.

4.9 Summary

The methods of performing unsupervised learning methods in reducing dimensionality, latent factor extraction, and cluster segmentation of continuous multi-attribute data using SAS macro applications are covered in this chapter. Both descriptive summary statistics and graphical analysis are used to summarize total variation, reduce dimensionality of multi-attributes, check for multivariate outliers and normality, segment observations, and explore interrelationships between attributes and observations in bi-plots. Steps involved in using user-friendly SAS macro applications FACTOR and DISJCLUS for performing unsupervised learning methods are presented using the "cars93" dataset.

References

1. Sharma, S., *Applied Multivariate Techniques*, John Wiley & Sons, New York, 1996, chaps. 4, 5, 7.
2. Johnson, R.A. and Wichern, D.W., *Applied Multivariate Statistical Analysis*, 5th ed., Prentice-Hall, Englewood Cliffs, NJ, 2002, chaps. 8, 9, 12.
3. Khattree, R. and Naik, D.N., *Multivariate Data Reduction and Discrimination with SAS Software*, 1st ed., SAS Institute, Inc., Cary, NC, 2000, chap. 2.
4. Khattree, R. and Naik, D.N., *Applied Multivariate Statistics with SAS Software*, 1st ed., SAS Institute, Inc., Cary, NC, 1995, chap. 2.
5. Neter, J., Kutner, M.H., Nachtsheim, C.J., and Wasserman, W., *Applied Linear Regression Models*, Irwin, Homewood, IL, 1996, chap. 6.
6. Hatcher, L., *A Step-By-Step Approach to Using the SAS System for Factor Analysis and Structural Equation Modeling*, 1st ed., SAS Institute, Cary, NC, 1994, chap. 1.
7. Sharma, S., *Applied Multivariate Techniques*, John Wiley & Sons, New York, 1996, chap. 4.
8. Sharma, S., *Applied Multivariate Techniques*, John Wiley & Sons, New York, 1996, chap. 5.
9. Johnson, R.A. and Wichern, D.W., *Applied Multivariate Statistical Analysis*, 5th ed., Prentice-Hall, Englewood Cliffs, NJ, 2002, chap. 9.
10. Khattree, R. and Naik, D.N., *Multivariate Data Reduction and Discrimination with SAS Software*, 1st ed., SAS Institute, Inc., Cary, NC, 2000, chap. 4.
11. Sharma, S., *Applied Multivariate Techniques*, John Wiley & Sons, New York, 1996, chap. 7.
12. Johnson, R.A. and Wichern, D.W., *Applied Multivariate Statistical Analysis*, 5th ed., Prentice-Hall, Englewood Cliffs, NJ, 2002, chap. 12.
13. Khattree, R. and Naik, D.N., *Multivariate Data Reduction and Discrimination with SAS Software*, 1st ed., SAS Institute, Inc., Cary, NC, 2000, chap. 6.
14. Gabriel, K.R., Bi-plot display of multivariate matrices for inspection of data and diagnosis, in *Interpreting Multivariate Data*, V. Barnett, Ed., Wiley, London, 1981.

15. SAS Institute, Inc., *Comparison of the PRINCOMP and FACTOR Procedures*, SAS online documentation, SAS Institute, Inc., Cary, NC (http://v8doc.sas.com/sashtml/stat/chap6/sect2.htm; accessed May 2002).
16. SAS Institute, Inc., *The FACTOR Procedure*, SAS online documentation, SAS Institute, Inc., Cary, NC (http://v8doc.sas.com/sashtml/stat/chap26/index.htm).
17. Lock, R.H., New car data, *J. Statistics Educ.*, 1(1), 1993 (http://www.amstat.org/publications/jse/v1n1/datasets.lock.html; accessed March 2002).
18. Khattree, R. and Naik, D.N., *Applied Multivariate Statistics with SAS Software*, 1st ed., SAS Institute, Inc., Cary, NC, 1995, chap 1.
19. Khattree, R. and Naik, D.N., *Multivariate Data Reduction and Discrimination with SAS Software*, 1st ed., SAS Institute, Inc., Cary, NC, 2000, chap. 6.
20. SAS Institute, Inc., *Introduction to Clustering Procedures*, SAS online documentation, SAS Institute, Inc., Cary, NC (http://v8doc.sas.com/sashtml/stat/chap8/index.htm; accessed May 2002).
21. SAS Institute, Inc., *The FASTCLUS Procedure*, SAS online documentation, SAS Institute, Inc., Cary, NC (http://v8doc.sas.com/sashtml/stat/chap27/sect2.htm; accessed May 2002).

Suggested Reading

Chong Chong, H.Y., Andrews, S., Winograd, D., Jannasch-Pennell, A., and DiGangi, S.A., *Teaching Factor Analysis in Terms of Variable Space and Subject Space Using Multimedia Visualization* (www.amstat.org/publications/jse/v10n1/yu.html).

Santos, J.R.A., Lippke, L., and Pope, P., PROC FACTOR: A tool for extracting hidden gems from a mountain of variables, in *Proc. 23rd Annu. SAS Users Group Int. Conf.*, SAS Institute, Cary, NC, 1998, pp. 1330–1335.

Sarle, W.S., Cluster analysis by least squares, in *Proc. 7th Annu. SAS Users Group Int. Conf.*, SAS Institute, Cary, NC, 1982, pp. 651–653.

Sarle, W.S., *Cubic Clustering Criterion*, SAS Technical Report A-108, SAS Institute, Inc., Cary, NC, 1983.

Wulder, M., *Principal Components and Factor Analysis* (http://www.pfc.forestry.ca/profiles/wulder/mvstats/pca_fa_e.html).

Chapter 5

Supervised Learning Methods: Prediction

5.1 Introduction

The goal of supervised predictive models is to find a model or mapping that will correctly associate the inputs with the targets. Automated data collection, data warehousing, and ever-faster computing combine to make it possible to fit many variations of predictive models. The combination of many predictors, large databases, and powerful software makes it easy to build such models that hold the potential to reveal hidden structures. Thus, supervised predictive models play a key role in data mining and knowledge discovery.

The supervised predictive models include both classification and regression models. Classification models use categorical response, while regression models use continuous and binary variables as targets. In regression we want to approximate the regression function, while in classification problems we want to approximate the probability of class membership as a function of the input variables.

In modeling, data give the model a chance to learn and find a solution that identifies essential patterns that are not overly specific to the sample data. One way to accomplish this is to fit the model to the data. The variable to be predicted and its predictors are carefully monitored and the observations are randomly selected in the training datasets that are used to build such models. The training data provide the predictive model a chance to identify essential patterns that are specific to the entire database. After training, the fitted model must be validated with data independent of the training set to provide a way to measure the ability of the model to generalize what it has learned. As in all other data mining techniques, these supervised predictive methods are not immune to badly chosen training data; therefore, the observations for the training set *must* be carefully chosen; the golden rule is "garbage in, garbage out."

Two supervised predictive model techniques, multiple linear regression (MLR) and binary logistic regression (BLR), are discussed in this chapter. The main objective of predictive modeling is to model the relationship between several predictor variables (regressor, input, independent, explanatory) and a response variable (target, output, dependent). The association between these two sets of variables is described by a linear equation in the case of MLR or by a nonlinear logistic function in the case of BLR that predicts the response variable from a function of predictor variables. In most situations, predictive models merely provide useful approximations of the true unknown model. However, even in cases where theory is lacking, a predictive model may provide an excellent predictive equation if the model is carefully formulated from a large representative database. In general, predictive models allow the analysts to:

- Determine which predictor variables are associated with the response.
- Determine the form of the relationship between the response and predictor variables.
- Estimate the best fitted predictive model.
- Estimate the model parameters and their confidence intervals.
- Test hypotheses about the model parameters.
- Estimate the predicted scores for new cases.

A non-mathematical description and application of these supervised predictive methods are included in this chapter; for a mathematical account of MLR and BLR, readers are encouraged to refer to Neter et al.[1] and Montgomery and Peck.[2]

5.2 Applications of Supervised Predictive Methods

Predictive modeling is a powerful tool being incorporated more and more into data mining. MLR could be used to build business intelligence solutions for problems such as predicting insurance claim losses, balances of credit card holders, amount of online orders, and cell phone usage.

MLR is also ideal for solving predictive modeling questions involving such continuous outcomes as:

- How much will a customer spend on his or her next purchase?
- How large a balance will a credit card holder carry?
- How many minutes will someone use on long distance this month?

Given binary (yes/no) outcomes, BLR can estimate the probability that a treatment for a serious illness will succeed or the pr obability that a policyholder will file an insurance claim. In addition, BLR models could also be used to answer the following:

- Will a software customer upgrade current software?
- Will particular homeowners refinance their mortgages in the next quarter?
- Will a customer respond to a direct mail offer?

It is clear from these applications that predictive modeling can be one of the most powerful tools for decision making.

5.3 Multiple Linear Regression Modeling

In MLR, the association between two sets of variables is described by a linear equation that predicts the response variable from a function of predictor variables. The estimated MLR model contains regression parameters that are estimated by using the least-squares criterion in such a way that prediction is optimized. In most situations, MLR models merely provide useful approximations of the true unknown model; however, even in cases where theory is lacking, an MLR model may provide an excellent predictive equation if the model is carefully formulated from a large representative database. The major conceptual limitation of MLR modeling based on observational studies is that one can only ascertain relationships and never be sure about the underlying causal mechanism. Significant regression relationship does not imply cause-and-effect relationships in uncontrolled observational studies; however, MLR modeling is considered to be the most widely used technique by all disciplines. The statistical theory, methods, and computation aspects of MLR ar e presented in detail elsewhere.[1,2]

5.3.1 MLR Key Concepts and Terminology

5.3.1.1 Overall Model Fit

In MLR, the statistical significance of the overall fit is determined by an F test by comparing the regression model variance to the error variance. The R^2 estimate is an indicator of how well the model fits the data (e.g., an R^2 close to 1.0 indicates that model has accounted for almost all of the variability with the variables specified in the model). The concept of R^2 can be visually examined in an overlay plot of ordered and centered response variables (describing the total variation) and the corresponding residuals (describing the residual variation) vs. the ascending observation sequence. The area in total variation not covered by the residual variation illustrates model explained variation. Other model and data violations also show up in the explained variation plot (see Figures 5.13 and 5.24 for examples of explained variation plots). Whether a given R^2 value is considered to be large or small depends on the context of the particular study. The R^2 is not recommended for selecting the best model because it does not account for the presence of redundant predictor variables; however, the $R^2_{(adjusted)}$ is recommended for model selection because the sample size and number of predictors are used in adjusting the R^2 estimate. Caution must be taken with interpretation of R^2 for models with no intercept term. As a general rule, no intercept models should be fit except when theoretical justification exists and the data appear to fit a no-intercept framework.

5.3.1.2 Regression Parameter Estimates

In MLR, the regression model is estimated by the least-squares criterion by finding the best-fitted line, which minimizes the error in the regression. The regression model contains a Y intercept and regression coefficients (β_i) for each predictor variables. The β_i measure the partial contributions of each predictor variable to the prediction of the response. Thus, the β_i estimate the amount by which the mean response changes when the predictor is changed by one unit while all the other predictors are unchanged. However, if the model includes interactions or higher order terms, it may not be possible to interpret individual regression coefficients. For example, if the equation includes both linear and quadratic terms for a given variable, we cannot physically change the value of the linear term without also changing the value of the quadratic term. To interpret the direction of the relationship between the predictor variable and the response, look at the signs (plus or minus) of the regression or β coefficients. If a β coefficient is positive, then the relationship of this variable with the response is positive, and if the β coefficient is negative then the relationship is negative. In an observational study where the true model form is unknown, interpretation of parameter estimates becomes even more complicated. A parameter estimate can be interpreted as the expected difference in response between two observations that differ by one unit on the predictor in question and have the same values for all other predictors. We cannot make inferences about changes in an observational study because we have not actually changed anything. It may not even be possible, in principle, to change one predictor independently of all the others, nor can you draw conclusions about causality without experimental manipulation.

5.3.1.3 Standardized Regression Coefficients

Two regression coefficients in the same model can be directly compared only if the predictors are measured in the same units. Sometimes standardized regression coefficients are used to compare the effects of predictors measured in different units. Standardizing the variables (zero mean with unit standard deviation) effectively makes standard deviation the unit of measurement. This makes sense only if the standard deviation is a meaningful quantity, which is usually the case only if the observations are sampled from well-defined databases.

5.3.1.4 Significance of Regression Parameters

The statistical significance of regression parameters is determined based on the partial sums of squares (SS2) and the *t* statistics derived by dividing the parameter estimates by its standard error. If higher order model terms such as quadratic and cross products are included in the regression model, the *p* values based on SS2 are incorrect for the linear and main effects of the parameters. Under these circumstances, correct significance tests for the linear and main effects could be determined using the sequential sums of squares (SS1).

Although *p* values based on a *t* test provide the statistical significance of a given variable in predicting the response in that sample, it does not necessarily measure the importance of a predictor. An important predictor can have a large (nonsignificant) *p* value if the sample is small, if the predictor is measured over a narrow range, if there are large measurement errors, or if another closely related predictor is included in the equation.

An unimportant predictor can have a very small *p* value in a large sample. Computing a confidence interval for a parameter estimate provides more useful information than just looking at the *p* value, but confidence intervals do not solve problems of measurement errors in predictors or highly correlated predictors.

5.3.1.5 Model Estimation in MLR with Categorical Variables

When categorical variables are used as predictors, separate regression models are estimated for each level or a combination of levels within all categorical variables included in the model. One of the levels is treated as the baseline and differences in the intercept and slope estimates for all other levels compared with the base level are estimated and tested for significance. The main effects of the categorical variables and the interaction between the categorical variables and continuous predictors must be specified in the model statement to estimate differences in the intercepts and slopes, respectively.[3] MLR models with categorical variables can be modeled more efficiently in the SAS general linear model (GLM) procedure,[3] where the GLM generates the suitable design matrix when the categorical variables are listed in the class statement. The influence of the categorical variables on the response could be examined graphically in scatterplots between the response and each predictor variable by each categorical variable. The significance of the categorical variable and the need for fitting the heterogeneous slope model could be checked visually by examining the interaction between the predictor and the categorical variable (see Figures 5.20 and 5.21 for examples of regression diagnostic plots suitable for testing the significance of categorical variable). For additional details regarding the fitting of MLR with categorical variables, refer to Freund and Littell[3] and SAS Institute.[4]

5.3.1.6 Predicted and Residual Scores

After the model has been fit, predicted and residual values are usually estimated. The regression line expresses the best prediction of the response for the given predictor variable. The deviation of a particular observed value from the regression line (its predicted value) is called the *residual value*. The smaller the variability of the residual values around the regression line relative to the overall variability, the better the prediction. Standard errors of residuals and the studentized residual, which is the ratio of the residual to its standard error, are also useful in modeling. The *studentized residual* is useful in detecting outliers because an observation greater than 2.5 in absolute terms could be treated as an outlier. The *predicted residual* for the *i*th observation is defined as the residual for the *i*th observation based on the regression model that results from dropping the *i*th observation from the parameter estimates. The sum of squares of predicted residual errors is called the PRESS statistic. Another R^2 statistic, called the R^2 prediction, is useful in estimating the predictive power of the regression based on the PRESS. A big drop in the R^2 prediction value from the R^2 or a negative R^2 prediction value is an indication of a very unstable regression model with low predictive potential. There are two kinds of interval estimates for the predicted value. For a given level of confidence, the *confidence interval* provides an interval estimate for the mean value of the response, whereas the *prediction interval* is an interval estimate for an individual value of a response. These interval estimates are useful in developing scorecards for observations in the database.

5.3.2 Exploratory Analysis Using Diagnostic Plots

Simple scatter plots are very useful in exploring the relationship between a response and a single predictor variable in a simple linear regression, but these plots are not effective in revealing the complex relationships of predictor variables or data problems in multiple linear regressions. However, partial scatter plots are considered useful in detecting influential observations and multiple outliers, nonlinearity and model specification errors, multicollinearity, and heteroscedasticity problems.[5] These partial plots illustrate the partial effects or the effects of a given predictor variable after adjusting for all other predictor variables in the regression model. Two types of partial scatterplots are considered superior in detecting regression model problems. The mechanics of these two partial plots are described using two variable MLR models:

1. Augmented partial residual plot[6] between response (Y) and the predictor variable X_1

   Step 1. Fit a quadratic regression model:

$$Y_i = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_1^2 + \varepsilon_i \qquad \text{(Eq. 5.1)}$$

   Step 2. Add the X_1 linear ($\beta_1 X_1$) and the X_1 quadratic ($\beta_3 X_1^2$) components back to the residual (ε_i):

$$APR = \varepsilon_i + \beta_1 X_1 + \beta_3 X_1^2 \qquad \text{(Eq. 5.2)}$$

Step 3. Fit a simple linear regression between augmented partial residual (APR) and the predicated variable X_1.

This augmented partial residual plot is considered very effective in detecting outliers, non-linearity, and heteroscedasticity.

2. Partial leverage plot (PL)[7] between response (Y) and the predictor variable X_1

Step 1. Fit two MLR models; remove the effects of all other predictors from the response (Y_i) and the predictor (X_1) in question:

$$Y_i = \beta_0 + \beta_2 X_2 + \varepsilon_i \qquad \text{(Eq. 5.3)}$$

$$X_{1i} = a_0 + b_2 X_2 + e_i \qquad \text{(Eq. 5.4)}$$

Step 2. Add the Y mean to the Y residual (ε_i) and X_1 mean to the X_1 residual (e_i) and compute the partial regression estimate for Y (PR_Y), and the partial leverage estimate for X_1 (PL_{X1}):

$$PR_y = \varepsilon_i + Y_{mean} \qquad \text{(Eq. 5.5)}$$

$$PL_{x1} = e_i + X_{1mean} \qquad \text{(Eq. 5.6)}$$

Step 3. Fit a simple linear regression between PR_Y and PL_{X1}. Include the 95% confidence band around the regression line and draw a horizontal line through the Y_{mean}. The slope of the regression line will be equal to the regression coefficient for X_1 in the MLR, and the residual from the PL plot will be equal to the residual from the MLR.

Also, based on the position of the horizontal line through the response mean and the confidence curves, the following conclusions can be made regarding the significance of the slope:

Confidence curve crosses the horizontal line = significant slope

Confidence curve asymptotic to horizontal line = borderline significance

Confidence curve does not cross the horizontal line = non-significant slope

Thus, the PL plot is effective in showing the statistical significance of the predictor variable.

3. Variance inflation factor (VIF) plot: Both augmented partial residual and partial leverage plots in the original format fail to

> detect the presence of multicollinearity. Stine[8] proposed overlaying the partial residual and partial leverage points on the same plot to detect the multicollinearity. Thus, by overlaying the augmented partial residual and partial leverage points with the centered X_i values on the X axis, the degree of multicollinearity can be detected clearly by the amount of shrinkage of partial regression residuals. Because the overlaid plot is mainly useful in detecting multicollinearity, this plot is referred to as the VIF plot.[5]

See Figures 5.2 through 5.10 for examples of these three regression diagnostic plots.

5.3.3 Model Selection[9]

Multiple linear regression modeling from large databases containing many predictor variables presents big challenges to data analysts in selecting the best model. The regression model assumes that we have specified the correct model, but many times theory or intuitive reasoning does not suggest such a model. It is customary to use an automated procedure that uses information on data to select a suitable subset of variables. SAS software offers nine model selection methods in the regression procedure to help the analyst select the best one.[9] This section discusses one of the selection methods, maximum R^2 improvement (MAXR), implemented within the SAS macro REGDIAG.

The maximum R^2 improvement technique does not settle for a single model. Instead, it compares all possible combinations and tries to find the "best" variable subsets for one-variable models, two-variable models, and so forth. The MAXR method may require much more computer time than the STEPWISE methods. In addition, R^2, $R^2_{\text{(adjusted)}}$, the Akaike's information criterion (AIC), the root-mean-square error (RMSE), and the Mallows C_p statistics are generated for each model generated in the model-selection methods. The R^2 is defined as the proportion of variance of the response that is predictable from the predictor variables. The $R^2_{\text{(adjusted)}}$ statistic is an alternative measure to R^2 that is adjusted for the number of parameters and the sample size. RMSE is the measure of MLR model error standard deviation. AIC is the MLR model variance statistic adjusted for the sample size and number of parameters. Minimum RMSE and AIC and maximum R^2 and $R^2_{\text{(adjusted)}}$ are characteristics of an optimum subset for a given number of variables. For a subset with p parameters, including the intercept, the C_p statistic is a measure of total squared error estimated by adding the model error variance and the bias component introduced by not including important variables. If the C_p/p ratio is plotted against p, Mallows recommends selecting the model where C_p/p first approaches 1.[9,10] Parameter estimates are unbiased for the best model because C_p/p approximately equals 1 (see Figure 5.11 for an example of a C_p model selection plot).

5.3.4 Violations of Regression Model Assumptions[11]

If sample regression data violates one or more of the MLR assumptions, the results of the analysis may be incorrect or misleading. The assumptions for a valid MLR are:

- Model parameters are correctly specified.
- Residuals from the regression are independent and have zero mean, constant variance, and normal distribution.
- Influential outliers are absent.
- Multicollinearity is not present.

5.3.4.1 Model Specification Error

When important predictor variables or significant higher order model terms (quadratic and cross-product) are omitted from the regression model, the residual error term no longer has the random error property. The augmented partial residual plot is very efficient in detecting the need for a nonlinear (quadratic) term. The need for an interaction between any two predictor variables could be evaluated in the "interaction test" plot. Simple scatterplots between a predictor and the response variable by an indicator variable could indicate the need for an interaction term between a predictor and the indicator variable. The significance of any omitted predictor variable can only be evaluated by including it in the model and following the usual diagnostic routine.

5.3.4.2 Serial Correlation Among the Residual

In a time series or spatially correlated data, the residuals are usually not independent and positively correlated among the adjacent observations. This condition is known as serial correlation or first-order autocorrelation. When the first-order autocorrelation is severe (>0.3), the standard error for the parameter estimates are underestimated. The significance of the first-order autocorrelation could be evaluated by the Durbin–Watson test and an approximate test based on the $2/n$ critical value criteria. The cyclic pattern observed in case of significant positive autocorrelation can be evaluated by examining the trend plot between residuals by the observation sequence (see Figure 5.14 for an example of an autocorrelation detection plot). The SAS AUTOREG procedure available in the SAS/ETS[12] module provides an effective method of adjusting for autocorrelation. A user-friendly SAS macro available in the author's SAS macro page[13] adjusts for autocorrelation using the ETS/AUTOREG procedure.

5.3.4.3 Influential Outliers

The presence of significant outliers produces biased regression estimates and reduces the predictive power of the regression model. An influential outlier may act as a high-leverage point, distorting the fitted equation and perhaps fitting the model poorly. The SAS/REG procedure has many powerful influential diagnostic statistics.[14] If the absolute value of the student residual for a given observation is greater than 2.5, then it could be treated as a significant outlier. High-leverage data points are highly influential and have significant hat values. The DFFITS statistic shows the impact of each data point by estimating the change in the predicted value in standardized units when the *i*th observation is excluded from the model; a DFFITS statistic greater than 1.5 could be used as a cutoff value in influential observation detection. An outlier detection bubble plot between the Student and hat value identifies the outliers if they fall outside the 2.5 boundary line and indicates influential points if the diameter of the bubble plot, which is proportional to DFFITS, is relatively big. Robust regressions using iterative weighted regression methods are available to minimize the impact of influential outliers.[15] A user-friendly SAS macro available in the author's SAS macro page[13] adjusts for influential observations based on robust regression by the HUBER and TUKEY methods.

5.3.4.4 Multicollinearity

When a predictor variable is nearly a linear combination of other predictors in the model, the affected estimates are unstable and have high standard errors. If multicollinearity among the predictors is strong, the partial regression estimates may have the wrong sign or size and are unstable. If a predictor involved in a collinear relationship is removed from the model, the sign and size of the remaining predictor can change dramatically. The fitting of higher order polynomials of a predictor variable with a mean not equal to zero can create difficult multicollinearity problems.

The PROC REG provides VIF and COLLINOINT options for detecting multicollinearity. The condition indices >30 indicate the presence of severe multicollinearity. The VIF option provides the variance inflation factors, which measure the inflation in the variances of the parameter estimates due to multicollinearity that exists among the predictor variables. A VIF value greater than 10 is usually considered significant. The presence of severe multicollinearity could be detected graphically in the VIF plot when the partial leverage points shrink and form a cluster near the mean of the predictor variable relative to the partial residual. One of the remedial measures for multicollinearity is to redefine highly correlated variables. For example, if X and Y are highly correlated, they could be replaced in a linear regression by $X + Y$ and $X - Y$ without changing the fit of the model or statistics for other predictors. User-friendly SAS macros available in the author's SAS macro page[13] adjust for multicollinearity based on ridge regression[15] and incomplete principal component regression[15] methods.

5.3.4.5 Heteroscedasticity in Residual Variance

Nonconstancy of error variance occurs in MLR if the residual variance is not constant and shows a trend with the change in the predicted value. The standard error of the parameters becomes incorrect, resulting in incorrect significance tests and confidence interval estimates. A fan pattern, like the profile of a megaphone, with a noticeable flare either to the right or to the left in the residual plot against predicted value is the indication of significant heteroscedasticity. The Breusch–Pagan test,[11] based on the significance of linear model using the squared absolute residual as the response and all combination of variables as predictors, is recommended for detecting heteroscedasticity. However, the presence of significant outliers and non-normality may confound with heteroscedasticity and may interfere with the detection. If both nonlinearity and unequal variances are present, employing a transformation on response may have the effect of simultaneously improving the linearity and promoting equality of the variances. User-friendly SAS macros available in the author's SAS macro page[13] adjust for heteroscedasticity based on Box–Cox[11] regression and heterogeneity regression models using the MIXED model approach.[16]

5.3.4.6 Non-Normality of Residuals

Multiple linear regression models are fairly robust against violation of non-normality, especially in large samples. Signs of non-normality are significant skewness (lack of symmetry) and/or kurtosis light-tailedness or heavy-tailedness. The normal probability plot (normal quantile–quantile [Q–Q] plot), along with the normality test,[17] can provide information on the normality of the residual distribution. In the case of non-normality, fitting generalized linear models based on the SAS GENMOD[18] procedure or employing a transformation on response or one or more predictor variables may result in a more powerful test. However, if only a small number of data points (<32) is available, non-normality can be difficult to detect. If the sample size is large (>300), the normality test may detect statistically significant but trivial departures from normality that will have no real effect on the multiple linear regression tests (see Figures 5.14 to 5.15 for examples of model violation detection plots).

5.3.5 Regression Model Validation

A regression model estimated using the training dataset could be validated by applying the model to independent validation data and by comparing the model fit. If both models produce a similar R^2 and show comparable predictive models, then the estimated regression model could be used for prediction with reasonable accuracy. Model validation could be further strengthened if both training and the validation residual plots show similar pattern (see Figures 5.27 and 5.28 for examples of comparing prediction and residual patterns between the training and validation datasets).

5.4 Binary Logistic Regression Modeling

Logistic regression is a powerful modeling technique used extensively in data mining applications. It allows more advanced analyses than the chi-square test, which tests for the independence of two categorical variables. It allows analysts to use binary responses (yes/no, true/false) and both continuous and categorical predictors in modeling. Logistic regression does allow an ordinal variable (e.g., a rank order of the severity of injury from 0 to 4) as the response variable, but only binary logistic regression (BLR) is discussed in this book. BLR allows construction of more complex models than the straight linear models, so interactions among the continuous and categorical predictors can also be explored.

Binary logistic regression uses maximum-likelihood estimation (MLE) after converting the binary response into a logit value (the natural log of the odds of the response occurring or not) and estimates the probability of a given event occurring. MLE relies on large-sample asymptotic property, which means that the reliability of the estimates declines when only a few cases for each observed combination of X variables are available. In BLR, changes in the log odds of the response, not changes in the response itself, are modeled. BLR does not assume linearity of the relationship between the predictors, and the residuals do not require normality or homoscedasticity. The success of the BLR could be evaluated by investigating the classification table, showing correct and incorrect classifications of the binary response. Also, goodness-of-fit tests such as model chi-square are available as indicators of model appropriateness, as is the Wald statistic to test the significance of individual parameters. The BLR model assumes the following:

- Inclusion of all relevant variables in the regression model
- No multicollinearity among the continuous predictor variable
- Independent error terms
- Predictor variables measured without errors
- No overdispersion

The statistical theory, methods, and computation aspects of BLR are presented in detail elsewhere.[19–21]

5.4.1 Terminology and Key Concepts

- **Probability.** Probability is the chance of an occurrence of an event. The probabilities are frequency of one category divided by the total. Note that they always sum to 1. The odds are the ratio of the probabilities of the binary event. Thus, if there is a 25% chance of rain, then the odds of rain will be 0.25/0.75 = 1/3.
- **Odds ratio.**[22,23] Odds and probability describe how often something happens relative to its opposite happening. Odds can range from zero to plus infinity, with the odds of 1 indicating neutrality, or no difference. Briefly, an odds ratio is the ratio of two odds, and relative risk is the ratio of two probabilities. The odds ratio is the ratio of two odds and is a comparative measure (effect size) between two levels of a categorical variable or a unit change in the continuous variable. An odds ratio of 1.0 indicates the two variables are statistically independent. The odds ratio of summer to winter means that the odds for summer are the denominator and the odds for winter are the numerator, and the odds ratio describes the change in the odds of rain from summer months to winter months. In this case, the odds ratio is a measure of the strength and direction of the relationship between rain and season. If the 95% confidence interval of the odds ratio includes the value of 1.0, the predictor is not considered significant. Odds ratios can be computed for both categorical and continuous data. Also, odds ratios for negative effects can vary only from 0 to 0.999, while for the case of increase it can vary from 1.001 to infinity. With these measures,

one must be very careful when coding the values for the response and the predictor variables. Reversing the coding for the binary response inverts the interpretation. The interpretation of the odds ratio is valid only when a unit change in the predictor variable is relevant. If a predictor variable is involved in a significant quadratic relationship or is interacting with other predictors, then the interpretation of the odds ratio is not valid.

- **Logits.** Logits are used in the BLR equation to estimate (predict) the log odds that the response equals 1 and contain exactly the same information as odds ratios. They range from minus infinity to plus infinity, but, because they are logarithms, the numbers usually range from –5 to +5, even when dealing with very rare occurrences. Unlike the odds ratio, a logit is symmetrical and therefore can be compared more easily. A positive logit means that, when that independent variable increases, the odds that the dependent variable equals 1 increase. A negative logit means that when the independent variable decreases, the odds that the dependent variable equals 1 decrease. The logit can be converted easily into an odds ratio of the response simply by using the exponential function (raising the natural log *e* to the β_1 power). For instance, if the logit β_1 = 2.303, then its log odds ratio (the exponential function, $e^{\beta 1}$) is 10, and we may say that when the response variable increases one unit the odds that the response event = 1 increase by a factor of 10, when other variables are controlled.
- **Percent increase in odds.** Once the logit has been transformed back into an odds ratio, it may be expressed as a percent increase in odds. Let the logit coefficient for "current asset/net sales" be 1.52, where the response variable is bankruptcy. The odds ratio, which corresponds to a logit of +1.52, is approximately 4.57 ($e^{1.52}$). Therefore, we can conclude that for each additional unit increase in "current asset/net sales" the odds of bankruptcy increase by 357% — (4.57 – 1) × 100% — while controlling for other inputs in the model. But, saying that the probability of bankruptcy increases by 357% is incorrect.
- **Standardized logit coefficients.** Standardized logit coefficients correspond to β (standardized regression) coefficients. These coefficients may be used to compare the relative importance of the predictor variables. Odds ratios are preferred for this purpose, however, because when we use standardized logit coefficients we are measuring the relative importance of the predictor in terms of effect on the log odds of the response variable, which is less intuitive than using the actual odds of the response variable, which is measured when odds ratios are used.

5.4.1.1 Testing the Model Fit[24,25]

- **Wald statistic.** The Wald statistic is used to test the significance of individual logistic regression coefficients to test the null hypothesis that a particular logit (effect) coefficient is zero. It is the ratio of the unstandardized logit coefficient to its standard error.
- **Log-likelihood ratio tests.** Log-likelihood ratio tests are an alternative to the Wald statistic. If the log-likelihood test statistic is significant, the Wald statistic can be ignored. Log-likelihood tests are also useful in model selection. Models are run with and without the variable in question, for instance, and the difference in –2 log likelihood (–2LL) between the two models is assessed by the chi-square statistic, with the degrees of freedom being equal to the difference in the number of parameters between the two models.
- **Deviance.** Because –2LL has approximately a chi-square distribution, –2LL can be used for assessing the significance of logistic regression, analogous to use of the sum of squared errors in ordinary least-squares (OLS) regression. The –2LL statistic is the scaled deviance statistic for logistic regression. Deviance measures error associated with the model after all the predictors are included in the model. It thus describes the unexplained variance in the response. Deviance for the null model describes the error associated with the model when only the intercept is included in the model — that is, –2LL for the model that accepts the null hypothesis that all the β coefficients are 0.

5.4.1.2 Assessing the Model Fit

- **Hosmer and Lemeshow's goodness of fit test.**[26,27] This test divides subjects into deciles based on predicted probabilities and then computes a chi-square from observed and expected frequencies. Then a *p* value is computed from the chi-square distribution with 8 degrees of freedom to test the fit of the logistic model. If the Hosmer and Lemeshow (H–L) goodness of fit test statistic is 0.05 or less, we reject the null hypothesis that no difference exists between the observed and model-predicted values of the response. If the H–L goodness of fit test statistic is greater than 0.05, we fail to reject the null hypothesis that there is no difference, implying that the model's estimates fit the data at an acceptable level. This does not mean that the model necessarily explains much of the variance in the dependent, only that however much or little it does explain is significant. As with other tests, as the sample size gets larger, the power of the H–L test to detect differences from the null hypothesis improves.

- **Brier score.**[27] This is a unitless measure of predictive accuracy computed from the classification table based on a cutpoint probability of 0.5. It ranges from 0 to 1. The smaller the score the better the predictive ability of the model. Brier score is useful in model selection and assessing model validity based on an independent validation dataset.
- **Adjusted generalized coefficient of determination (*R^2*).** This is a model assessment statistic similar to the R^2 in OLS regression and can reach a maximum value of 1. The statistic is computed using the ratio between the –2LL statistic for the null model and full model adjusted the sample size.[27,28]
- ***c statistic and ROC curve.*** The *c* statistic and receiver operating characteristic (ROC) curve measures the classification power of the logistic equation. The area under the ROC curve varies from 0.5 (the predictions of the model are no better than chance) to 1.0 (the model always assigns higher probabilities to correct cases than to incorrect cases). The *c* statistic is the percent of all possible pairs of cases in which the model assigns a higher probability to a correct case than to an incorrect case.[29] The receiver operating characteristic (ROC) curve is a graphical display of the predictive accuracy of the logistic curve. The ROC curve is constructed by plotting the sensitivity (measure of accuracy of predicting events) vs. 1-specificity (measure of error in predicting non-events).[29,30] The area under the ROC curve is equal to the *c* statistic. The ROC curve rises quickly and the area under the ROC is larger for a model with high predictive accuracy (see Figure 5.36 [top] for an example of a ROC curve). An overlay plot between the percentages of false positives and false negatives vs. the cutpoint probability could reveal the optimum cutpoint probability when both false positives and false negatives could be minimized (see Figure 5.36 [bottom] for an example of an overlay plot of false positives and negatives).

5.4.2 Exploratory Analysis Using Diagnostic Plots

Simple logit plots are very useful in exploring the relationship between a binary response and a single continuous predictor variable in a BLR with a single predictor variable, but these plots are not effective in revealing the complex relationships among the predictor variables or data problems in BLR with many predictors. The partial delta logit plots proposed here, however, are useful in detecting significant predictors, nonlinearity, and multicollinearity. The partial delta logit plot illustrates the effects of a given continuous predictor variable after adjusting for all other predictor variables on the change in the logit estimate when the variable in question is dropped from the BLR. By overlaying the simple logit and partial delta logit plots, many features of the BLR could be revealed. The mechanics of these two logit plots are described using a two-variable BLR model.

1. Determine a simple logit model for the binary response of the predictor variable X_1.

   Step 1. Fit a simple BLR model:

   $$Logit\ (P_i) = \beta_0 + \beta_1 X_1 \qquad \text{(Eq. 5.7)}$$

2. Obtain the delta logit (Δlogit) estimate for a given predictor.

   Step 1. Fit the full BLR model with a quadratic term for X_1:

   $$Logit_{(full)}(P_i) = \beta_0 + \beta_1 X_1 + \beta_1 X_2 + \beta_3 X_1^2 \qquad \text{(Eq. 5.8)}$$

   Step 2. Fit the reduced BLR model:

   $$Logit_{(reduced)}\ (P_i) = \beta_0 + \beta_1 X_2 \qquad \text{(Eq. 5.9)}$$

   Step 3. Estimate the delta logit (Δlogit) — difference in logit between the full and the reduced model:

   $$\Delta\text{logit} = \text{Logit}_{(\text{full})} - \text{Logit}_{(\text{reduced})} \qquad \text{(Eq. 5.10)}$$

   Step 4. Compute the partial residual (*PR*) for X_1 and add X_1-mean

   $$X_i = a_0 + b_2 X_2 + e_i \qquad \text{(Eq. 5.11)}$$

   $$PR_{x1} = e_i + X_{1mean} \qquad \text{(Eq. 5.12)}$$

   Step 5. Overlay the simple logit and partial delta logit plots:
   Simple logit plot — Logit (P_i) vs. X_1
   Partial delta logit plot — Δlogit vs. PR_{X1}

5.4.2.1 Interpretation

Positive or negative slope in the partial delta logit plot shows the significance of the predictor variable in question. A quadratic trend in the partial delta logit plot confirms the need for a quadratic term for X_i in BLR. Clustering of delta logit points near the mean of X_i in the partial delta logit plot confirms the presence of multicollinearity among the predictors. Large differences between the simple logit and the partial delta logit line illustrate the difference between the simple and partial effects for a given variable X_i (see Figures 5.32 to 5.35 for some examples of these diagnostic plots).

5.4.3 Model Selection

SAS software offers four model selection methods in the LOGISTIC procedure to help the analyst select the best model.[31] The forward selection (FS) method is included within the SAS macro LOGISTIC and is an exploratory technique; however, the results of any model selection method should be validated by an independent validation dataset.

5.4.4 Checking for Violations of Regression Model Assumptions

If logistic regression data violate one or more of the BLR assumptions, the results of the analysis may be incorrect or misleading. The assumptions for a valid MLR are:

- Model parameters are correctly specified.
- Influential outliers are absent.
- Multicollinearity is not present.
- Overdispersion is not present.

5.4.4.1 Model Specification Error

When important predictor variables or significant higher order model terms (quadratic and cross-product) are omitted from the BLR, the predicted probability will be biased. The partial delta logit plot is effective for detecting the need for a quadratic term.

5.4.4.2 Influential Outlier

The presence of significant outliers produces biased logit estimates and reduces the predictive power of the BLR. An influential outlier may act as a high-leverage point, distorting the fitted equation and perhaps fitting the model poorly. The SAS LOGISTIC procedure has many powerful influential diagnostic statistics.[32] The DIFDEV statistic detects an ill-fitted observation that is responsible for the differences between the data and the predicted probabilities. It shows the change in the deviance due to deleting given observations. Observations with a DIFDEV greater than 4 could be examined for outliers. High-leverage data points are highly influential and have significant hat values. The *cbar* statistic measures the impact on change in the confidence intervals as a result of deleting given observations. The leverage statistic *h* is useful for identifying cases with high leverage effects. The leverage statistic varies from 0 (no influence on the model) to 1 (completely determines the model). The leverage of any given case may be compared to the average leverage, which equals p/n, where $p = (k + 1)/n$, where k is the number of predictors and n is the sample size. Displaying DIFDEV, hat, and *cbar* statistics on the same plot is an effective way of identifying both outliers and influential observations. An example of the influential outlier detection plot is given in Figure 5.37.

5.4.4.3 Multicollinearity

When a predictor variable is nearly a linear combination of other predictors in the model, the affected estimates are unstable and have high standard errors. If multicollinearity among the predictors is strong, the partial logit estimates may have the wrong sign or size and are unstable. If a predictor involved in a collinear relationship is removed from the model, the sign and size of the remaining predictor can change dramatically. PROC LOGISTIC does not have options for detecting multicollinearity. However, the partial delta logit plot proposed in this chapter provides graphical methods for detecting multicollinearity.

5.4.4.4 Overdispersion[33]

The expected variance of the binary response in BLR is $np(1 - p)$, where n is the sample size and p is the probability of the binary event. When the observed variance exceeds the expected variance we have overdispersion. Model- and data-specific errors can contribute to overdispersion. Pearson chi-square and deviance chi-square tests are available for detecting significant overdispersion and adjusting the standard error of the parameters by the degree of overdispersion. In the case of a moderate discrepancy, standard errors will be underestimated and we should use the adjusted standard error, which will make the confidence intervals wider. Large discrepancies, however, indicate a need to respecify the model, that the sample was not random, or that other serious design problems are present.

5.5 Multiple Linear Regression Using SAS Macro REGDIAG

The REGDIAG macro is a powerful SAS application useful for performing multiple linear regression analysis. Options are available for obtaining various regression diagnostic graphs and tests. The SAS procedures REG and GLM are the main tools used in the macro.[34,35] In addition to these SAS procedures, GPLOT, RSREG, and BOXPLOT procedures are also utilized in the REGDIAG macro. The advantages of using the REGDIAG macro over the PROCS GLM and REG include:

- Regression diagnostic plots, such as augmented partial residual plots, partial leverage plots, and VIF plots, for each predictor variable are automatically generated.
- Plots for checking for violations of model assumptions (residual plots for detecting heteroscedasticity and autocorrelation, outlier detection plots, normal probability and distribution plots of residual) are also generated.
- Test statistics and *p* values for testing normally distributed errors and model specification errors are automatically produced, as are Breusch–Pagan tests for detecting heteroscedasticity.
- In the case of simple linear regression, plots of linear and quadratic regression plots with a 95% confidence band are generated automatically. If the MLR is fit, an overall model fit plot is produced.
- Options are available for validating the MLR model obtained from a training dataset using an independent validation dataset by comparing fitted lines and residual.
- Options for saving the output tables and graphics in WORD, HTML, PDF, and TXT formats are available.

Software requirements for using the REGDIAG macro include:

- SAS/CORE, SAS/BASE, SAS/STAT, and SAS/GRAPH must be licensed and installed at the site.
- SAS version 8.0 and above is recommended for full utilization.
- An active Internet connection is required for downloading the REGDIAG macro from the book website if the companion CD-ROM is not available.

5.5.1 Steps Involved in Running the REGDIAG Macro

1. Create an SAS dataset (permanent or temporary) containing at least one continuous response (target) variable and many continuous and/or categorical predictor (input) variables. (It is highly recommended to disable the SAS ENHANCED EDITOR in the latest SAS versions and open only one PROGRAM EDITOR window).
2. *If the companion CD-ROM is not available*, first verify that the Internet connection is active. Open the REGDIAG.sas macro-call file in the SAS PROGRAM EDITOR window. Instructions are given in the Appendix regarding downloading the macro-call and sample data files from the book website. *If the companion CD-ROM is available*, the REGDIAG.sas macro-call file can be found in the mac-call folder on the CD-ROM. Open the REGDIAG.sas macro-call file in the SAS PROGRAM EDITOR window. Click the RUN icon to submit the macro-call file REGDIAG.sas to open the macro-call window called REGDIAG (Figure 5.1).
3. Input the appropriate parameters in the macro-call window by following the instructions provided in the REGDIAG macro help file in Section 5.5.2. Users can choose whether or not to include regression diagnostic plots for each predictor variable and to exclude large extreme observations from model fitting. After inputting all the required macro parameters, be sure the cursor is in the last input field and the RESULTS VIEWER window is closed, then hit the ENTER key (not the RUN icon) to submit the macro.
4. Examine the LOG window (only in DISPLAY mode) for any macro execution errors. If any errors are reported in the LOG window, activate the PROGRAM EDITOR window, resubmit the REGDIAG.sas macro-call file, check the macro input values, and correct any input errors.
5. If no errors are found in the LOG window, activate the PROGRAM EDITOR window, resubmit the REGDIAG.sas macro-call file, and change the macro input value from DISPLAY to any other desirable format (see Section 5.5.2.). The SAS output files from complete regression modeling and exploratory graphs can be saved as user-specified-format files in the user-specified folder.

5.5.2 Help File for SAS Macro REGDIAG

1. **Macro-call parameter:** Input SAS dataset name (required parameter).
   Descriptions and explanation: Include the name of the temporary (member_name) or permanent (libname.member_name) SAS dataset on which the regression analysis will be performed.

Options/examples:

Permanent SAS dataset: gf.sales (LIBNAME: gf; SAS dataset name: sales)

Temporary SAS dataset: sales (SAS dataset name)

2. **Macro-call parameter:** Input continuous response variable name (required parameter).

   Descriptions and explanation: Input the continuous response variable name from the SAS dataset to model as the target variable.

   Option/example:

   Y (name of a continuous response)

3. **Macro-call parameter:** Input group variables (optional statement).

   Descriptions and explanation: To include categorical variables from the SAS dataset as predictors in regression modeling, input the names of these variables. The REGDIAG macro will use PROC GLM for regression modeling and use the categorical variable names in the GLM CLASS statement. If this field is left blank, the REGDIAG macro will use PROC REG for regression modeling and fit the regression model using the continuous variables specified in macro input field #5.

   Options/examples:

   month manager (regression modeling using PROC GLM)

   Blank (if the macro input field is left blank, PROC REG is used)

4. **Macro-call parameter:** Input the alpha level (required parameter).

   Descriptions and explanation: Input the alpha level for computing the confidence interval estimates for parameter estimates.

   Options/examples:

   0.05 0.01 0.10

5. **Macro-call parameter:** Input the predictor variables (optional statement).

   Descriptions and explanation: Input the continuous predictor variable names. If macro input field #3 is left blank, PROC REG is used to fit the MLR modeling using these variables as predictors. The checking of significant quadratic and cross-product effects for all predictor variables will be performed using PROC RSREG. Model selection based on all possible combinations of predictor variables will be performed using the continuous variables listed in this macro input. If YES is entered in the regression diagnostic plot option in macro input field #14, regression diagnostic plots for each variable will be generated. If macro input field #3 is *not* blank and categorical variable names are inputted, PROC GLM is used to fit the MLR modeling using these variables as predictors and the categorical variables (listed in macro input field #3) as indicator variables. If YES is entered in the regression diagnostics option in macro input field #14, partial plots are not generated, but simple scatterplots for each predictor variable by each categorical variable and box plots of response by each categorical variable are generated.

Option/example:

X1 X2 X3 mpg murder (names of continuous predictor variables)

6. **Macro-call parameter:** Input model terms (required parameter).
   Descriptions and explanation: This macro input field is equivalent to the right side of the equal sign in the PROC REG and PROC GLM model statement. In the case of fitting MLR with continuous variables using PROC REG, input names of the continuous variables and quadratic and cross-product terms. Please note that the quadratic and cross-product terms should be created in the SAS dataset before specifying them in this macro field; however, to fit an MLR with indicator variables using PROC GLM, input continuous predictor variable names, the categorical variable names specified in macro input field #3, and any possible quadratic and interaction terms.
   Options/examples:
   X1 X2 X1X2 X1SQ (MLR using PROC REG, where X1 and X2 are the linear predictors; X1X2, the interaction term; X1SQ, the quadratic term for X1)
   X1 SOURCE X1*SOURCE (MLR with indicator variable SOURCE using PROC GLM, where X1 is the linear predictor; SOURCE, the indicator variable; X1*SOURCE, the interaction term between X1 and SOURCE)
   (For details about specifying model statements, refer to SAS online manuals on PROC REG[34] and PROC GLM.[35])
7. **Macro-call parameter:** Input model terms (optional statement).
   Descriptions and explanation: Input any optional SAS PROC GLM or REG model options. Depending on the type of model being fit (GLM or REG), add these model options for additional statistics:
   REG options (default options included in the macro: VIF, STB):
   INFLUENCE: Additional influential statistics
   COLINOINT: Multicollinearity diagnostic test statistics
   SS2: Type II SS
   NOINT: No intercept model
   GLM options (default options included in the macro: SOLUTION):
   NOINT: No intercept model

(For details about specifying model options, refer to the SAS online manuals on PROC REG[34] and PROC GLM.[35])

8. **Macro-call parameter:** Input ID variable (optional statement).

   Descriptions and explanation: If a unique ID variable can be used to identify each record in the database, input that variable name here. This will be used as the ID variable so that any influential outlier observations can be detected. If no ID variable is available in the dataset, leave this field blank. This macro can create an ID variable based on the observation number from the database.

   Option/example:

   ID NUM

9. **Macro-call parameter:** Adjust for extreme influential observations (optional parameter).

   Descriptions and explanation: If YES is input for this option, the macro will fit the regression model after excluding extreme observations. Any standardized residual values falling outside ±3.5 will be treated as an outlier and will be excluded from the analysis. A printout of all excluded observations is also produced.

   Options/examples:

   Yes: Extreme outliers will be excluded from the analysis.

   Blank: All observations in the dataset will be used.

10. **Macro-call parameter:** Input validation dataset name (optional parameter).

    Descriptions and explanation: To validate the regression model obtained from a training dataset by using an independent validation dataset, input the name of the SAS validation dataset. This macro fits a separate regression line for the validation data and both regression models can be compared visually. Both regression models and residuals from both the models can be compared to validate the regression model. Include the name of the temporary (member_name) or permanent (libname.member_name) SAS dataset to be treated as the validation data.

    Options/examples:

    Permanent SAS dataset: gf.valid (LIBNAME: gf; SAS dataset name: valid)

    Temporary SAS dataset: valid (SAS dataset name)

11. **Macro-call parameter:** Display or save SAS output and graphs (required statement).

    Descriptions and explanation: Option for displaying all output files in the OUTPUT window or saving as a specific format in the folder specified in macro input option #12.

Options/examples:

Possible values

DISPLAY: Output will be displayed in the OUTPUT window, all SAS graphics will be displayed in the GRAPHICS window, and system messages will be displayed in the LOG window.

WORD: Output and all SAS graphics will be saved together in the user-specified folder and will be displayed in the VIEWER window (if MS WORD is installed in the computer) as a single RTF format file for version 8.0 and later. SAS output files will be saved as text files in pre-8.0 versions, and all graphics files (CGM) will be saved separately in a user-specified folder (macro input option #12).

WEB: Output and graphics are saved in the user-specified folder and are viewed in the results VIEWER window as a single HTML file (version 8.0 and later) or as a text file in pre-8.0 versions. All graphics files (GIF) will be saved separately in a user-specified folder (macro input option #12).

PDF: If Adobe Acrobat Reader is installed in the computer, output and graphics are saved in the user-specified folder and are viewed in the results VIEWER window as a single PDF file (version 8.2 and later only). All graphics files (PNG) will be saved separately in a user-specified folder (macro input option #12) as a text file in pre-8.2 versions.

TXT: Output will be saved as a TXT file in all SAS versions. No output will be displayed in the OUTPUT window. All graphic files will be saved in the EMF format (version 8.0 and later) or in the CGM format (pre-8.0 versions) in the user-specified folder (macro input option #12) folder.

Note: System messages are deleted from the LOG window if DISPLAY is not selected as the input.

12. **Macro-call parameter:** Folder to save SAS graphics and output files (optional statement).

Descriptions and explanation: To save the SAS graphics files in an EMF format suitable for inclusion in PowerPoint presentations, specify output format as TXT in version 8.0 or later. In pre-8.0 versions, all graphic format files will be saved in a user-specified folder. Similarly, output files in WORD, HTML, PDF, and TXT formats will be saved in the user-specified folder. If this macro field is left blank, the graphics and output files will be saved in the default folder.

Option/example:

c:\output\ — folder named “OUTPUT”

13. **Macro-call parameter:** *z*th number of analysis (required statement).

    Descriptions and explanation: SAS output files will be saved by forming a file name from the original SAS dataset name and the counter value provided in this field. For example, if the original SAS dataset name is "sales" and the counter number included is 1, then the SAS output files will be saved as "sales1" in the user-specified folder. By changing the counter value, users can avoid replacing previous SAS output files with new outputs.

14. **Macro-call parameter:** Regression diagnostic plots (optional parameter).

    Descriptions and explanation: If YES is input and no categorical variables are included in macro call parameter #2, the macro will produce regression diagnostic plots (augmented partial residuals, partial leverage plots, VIF diagnostic plots) and detect significant interactions for each predictor variable. If YES is input and categorical variables are included in macro call parameter #2, regression diagnostic plots (X–Y scatterplots by categorical variable, box plots of response by categorical variable, and regression plots for detecting significant interaction) for regression models with indicator variables are produced. If this macro field is left blank, no diagnostic plots are produced.

    Options/examples:

    > **Yes:** Diagnostic plots are produced for each predictor variable.
    >
    > **Blank:** No diagnostic plots are produced.

5.6 Lift Chart Using SAS Macro LIFT

The LIFT macro is a powerful SAS application for producing a LIFT chart (if–then analysis) in regression models. Options are available for graphically comparing the predicted response from the full model and the predicted response from the reduced model (keeping the variable of interest at a constant level). SAS procedures REG, GLM, and LOGISTIC can be used in the macro. In addition to these SAS procedures, the GPLOT procedure is also utilized in the LIFT macro. No procedures or options are currently available in the SAS system to produce LIFT charts automatically.

Software requirements for using the LIFT macro include:

- SAS/CORE, SAS/BASE, SAS/STAT, and SAS/GRAPH must be licensed and installed at the site.
- SAS version 8.0 and above is recommended for full utilization.
- An active Internet connection is required for downloading the LIFT macro from the book website if the companion CD-ROM is not available.

5.6.1 Steps Involved in Running the Lift Macro

1. Create an SAS dataset (permanent or temporary) containing at least one response (target) variable and many continuous or categorical predictor (input) variables.
2. *If the companion CD-ROM is not available*, first verify that the Internet connection is active. Open the LIFT.sas macro-call file in the SAS PROGRAM EDITOR window. Instructions are given in the Appendix regarding downloading the macro-call and sample data files from the book website. *If the companion CD-ROM is available*, the LIFT.sas macro-call file can be found in the mac-call folder in the CD-ROM. Open the LIFT.sas macro-call file in the SAS PROGRAM EDITOR window. Click the RUN icon to submit the macro-call file LIFT.sas to open the macro-call window called LIFT (Figure 5.16).
3. Input the appropriate parameters in the macro-call window by following the instructions provided in the LIFT macro help file in Section 5.6.2. After inputting all the required macro parameters, be sure the cursor is in the last input field and the RESULTS VIEWER window is closed, then hit the ENTER key (not the RUN icon) to submit the macro.
4. Examine the LOG window (only in DISPLAY mode) for any macro execution errors. If any errors are reported in the LOG window, activate the PROGRAM EDITOR window, resubmit the LIFT.sas macro-call file, check the macro input values, and correct any input errors.
5. If no errors are found in the LOG window, activate the PROGRAM EDITOR window, resubmit the LIFT.sas macro-call file, and change the macro input value from DISPLAY to any other desirable format (see Section 5.6.2). The SAS output files from the LIFT analysis and LIFT charts can be saved as user-specified-format files in the user-specified folder.

5.6.2 Help File for Using SAS Macro LIFT

1. **Macro-call parameter:** Input SAS dataset name (required parameter).

Descriptions and explanation: Include the name of the temporary (member_name) or permanent (libname.member_name) SAS dataset on which the regression analysis is to be performed.

Options/examples:

Permanent SAS dataset: gf.sales (LIBNAME: gf; SAS dataset name: sales)

Temporary SAS dataset: sales (SAS dataset name)

2. **Macro-call parameter:** Input class variable names (optional statement).

   Descriptions and explanation: To include categorical variables from the SAS dataset as predictors in regression modeling, input the names of these variables. Use this with SAS procedures GLM and LOGISTIC (version 8.0 and later).

   Options/examples:

   month manager (regression modeling using PROC GLM or LOGISTIC)

   Blank (if the macro input field is left blank, PROC REG or LOGISTIC)

3. **Macro-call parameter:** Input response variable name (required parameter).

   Descriptions and explanation: Input the response variable name from the SAS dataset to be modeled as the target variable.

   Options/examples:

   Y (name of a continuous response; PROC REG or GLM)

   Y (name of a binary response; PROC LOGISTIC)

4. **Macro-call parameter:** Input variable name of interest (required option).

   Descriptions and explanation: Input the name of the predictor variable to control in the if–then analysis. The variable of interest can be continuous (PROC REG, GLM, or LOGISTIC) or a binary categorical variable (GLM or LOGISTIC).

   Option/example:

   X1 Manager

5. **Macro-call parameter:** Input the SAS PROC name (required statement).

   Descriptions and explanation: Input the SAS procedure name for fitting the regression models.

   Options/examples:

   REG: MLR with continuous response and continuous predictor variables

   GLM: MLR with continuous response and continuous and categorical predictor variables

LOGISTIC: Logistic regression with binary response and continuous and categorical predictor variables

(For details about specifying model statements, refer to the SAS online manuals on PROC REG,[34] PROC GLM,[35] and PROC LOGISTIC.[36])

6. **Macro-call parameter:** LIFT variable fixed value (required statement).

   Descriptions and explanation: Input the fixed value for the variable of interest.

   Options/examples:

   50000 (fixed value for variable X1; applicable in REG, GLM, and LOGISTIC)

   D (fixed categorical level when the variable of interest is categorical; valid only in GLM and LOGISTIC)

7. **Macro-call parameter:** Input model terms (required statement).

   Descriptions and explanation: This macro input is equivalent to the right side of the equal sign in the PROC REG, PROC GLM, or PROC LOGISTIC model statement. In the case of fitting MLR with continuous variables using PROC REG, input names of the continuous variables, quadratic, and cross-product terms. Note that quadratic and cross-product terms should be created in the SAS dataset before specifying them in the macro field; however, to fit a regression model with indicator variables using PROC GLM/LOGISTIC (version 8.0), input continuous predictor variable names, categorical variable names you specified in macro input field #3, and any possible quadratic and interaction terms.

   Options/examples:

   X1 X2 X3 X2X3 X2SQ (MLR using PROC REG, where X1, X2, and X3 are linear predictors; X2X3, the interaction term; and X2SQ, the quadratic term for X2)

   X1 SOURCE X1*SOURCE (regression with indicator variable SOURCE using PROC GLM/LOGISTIC, where X1 is the linear predictor; SOURCE, the indicator variable; and X1*SOURCE, the interaction term between X1 and SOURCE)

8. **Macro-call parameter:** Other optional statements (optional statement).

   Descriptions and explanation: Input any other optional statements associated with the SAS procedures; see the example below.

   Options/examples:

   1.LSMEANS source/pdiff ; contrast 'source a vs. b' source –1 1 0

   (only valid in GLM and LOGISTIC)

   Note that if you are using more than one statement, use a ";" at the end of the first statement.

2.Test x1–x2=0; Restrict intercept=0

(only valid in PROC REG)

Note that if you are using more than one statement, use a ";" at the end of the first statement.

(For details about specifying other PROC statements, refer to the SAS online manuals on PROC REG,[34] PROC GLM,[35] and PROC LOGISTIC.[36])

9. **Macro-call parameter:** Input ID variable (optional statement).

   Descriptions and explanation: For a unique ID variable that can be used to identify each record in the database, input that variable name here. This will be used as the ID variable so that any outlier/influential observations can be detected. If no ID variable is available in the dataset, leave this field blank. This macro can create an ID variable based on the observation number from the database.

   Option/example:

   ID NUM

10. **Macro-call parameter:** Input weight variable (optional parameter).

    Descriptions and explanation: To perform a weighted regression analysis (weights to adjust for influential observations or heteroscedasticity) and the weight exists in the dataset, input the name of the weight variable. If this parameter is left blank, a weight of 1 will be used for all observations.

    Options/examples:

    Wt (name of the weight variable)

    Blank (1 will be used as the weight)

11. **Macro-call parameter:** Input any optional procedure options (optional parameter).

    Descriptions and explanation: To include any optional PROC options (REG and GLM: NOPRINT; LOGISTIC: DESCENDING), specify these options here. If PROC LOGISTIC is being used and the DESCENDING option is not specified, by default PROC LOGISTIC will model the probability of a non-event as 0.

    Options/examples:

    NOPRINT (REG and GLM)

    DESCENDING (LOGISTIC)

    (For details about specifying PROC options, refer to the SAS online manuals on PROC REG,[34] PROC GLM,[35] and PROC LOGISTIC.[36])

12. **Macro-call parameter:** Folder to save SAS graphics and output files (optional statement).

    Descriptions and explanation: To save the SAS graphics files in an EMF format suitable for inclusion in PowerPoint presentations, specify output format as TXT in version 8.0 or later. In pre-8.0 versions, all graphic format files will be saved in a user-specified folder. Similarly, output files in WORD, HTML, PDF, and TXT formats will be saved in the user-specified folder. If this macro field is left blank, the graphics and output files will be saved in the default folder.

Option/example:

c:\output\ — folder named "OUTPUT"

13. **Macro-call parameter:** *i*th number of analysis (required statement).
    Descriptions and explanation: SAS output files will be saved by forming a file name from the original SAS dataset name and the counter value provided in this field. For example, if the original SAS dataset name is "sales" and the counter number included is 1, the SAS output files will be saved as "sales1" in the user-specified folder. By changing the counter value, users can avoid replacing previous SAS output files with new outputs.
14. **Macro-call parameter:** Display or save SAS output and graphs (required statement).
    Descriptions and explanation: Option for displaying all output/graphics files in the OUTPUT/GRAPHICS window or saving as a specific format in a folder specified in macro input option #12.
    Options/examples: See Section 5.5.2 (macro input option #11) for detailed explanations for these options:

    DISPLAY
    WORD
    WEB
    PDF
    TXT

    Note: System messages are deleted from the LOG window if DISPLAY is not selected as the input.

5.7 Scoring New Regression Data Using the SAS Macro RSCORE

The RSCORE macro is a powerful SAS application for scoring new datasets using the established regression model estimates. Options are available for checking the residuals graphically if observed response variables are available in the new dataset; otherwise, only predicted scores are produced and saved in a SAS dataset. In addition to the SAS REG procedure, GPLOT is also utilized in the RSCORE macro. No procedures or options are currently available in the SAS systems to compare the residuals from the new data automatically.

Software requirements for using the RSCORE macro include:

- SAS/CORE, SAS/BASE, SAS/STAT, and SAS/GRAPH must be licensed and installed at the site.
- SAS version 8.0 and above is recommended for full utilization.
- An active Internet connection is required for downloading the RSCORE macro from the book website if the companion CD-ROM is not available.

5.7.1 Steps Involved in Running the RSCORE Macro

1. Create a new scoring SAS dataset (permanent or temporary) containing one response (optional) variable and continuous predictor (input) variables. This new dataset should contain all the predictor variables that were used to develop the original regression model.
2. Verify that the regression parameter estimates are available in an SAS dataset. If the REGDIAG macro was used to fit the original regression model, it should have created an SAS dataset called "regest". Check the "regest" data and create a temporary dataset to use for scoring the new dataset.
3. *If the companion CD-ROM is not available*, first verify that the Internet connection is active. Open the RSCORE.sas macro-call file in the SAS PROGRAM EDITOR window. Instructions are given in the Appendix regarding downloading the macro-call and sample data files from the book website. *If the companion CD-ROM is available*, the RSCORE.sas macro-call file can be found in the mac-call folder on the CD-ROM. Open the RSCORE.sas macro-call file in the SAS PROGRAM EDITOR window. Click the RUN icon to submit the macro-call file RSCORE.sas and open the macro-call window called RSCORE (Figure 5.18).
4. Input the appropriate parameters in the macro-call window by following the instructions provided in the RSCORE macro help file in Section 5.7.2. After inputting all the required macro parameters, be sure the cursor is in the last input field and the RESULTS VIEWER window is closed, then hit the ENTER key (not the RUN icon) to submit the macro.
5. Examine the LOG window (only in DISPLAY mode) for any macro execution errors. If any errors are reported in the LOG window, activate the PROGRAM EDITOR window, resubmit the RSCORE.sas macro-call file, check the macro input values, and correct any input errors.
6. If no errors are found in the LOG window, activate the PROGRAM EDITOR window, resubmit the RSCORE.sas macro-call file, and change the macro input value from DISPLAY to any other desirable format (see Section 5.7.2). The SAS output files from RSCORE analysis and RSCORE charts can be saved as user-specified-format files in the user-specified folder.

5.7.2 Help File for Using SAS Macro RSCORE

1. **Macro-call parameter:** Input name of the new scoring SAS dataset name (required parameter).

   Descriptions and explanation: Input the name of the temporary (member name) or permanent (libname.member_name) SAS dataset to be scored using the established regression model.

   Options/examples:

   Permanent SAS dataset: gf.new (LIBNAME: gf; SAS dataset name: new)

   Temporary SAS dataset: new (SAS dataset name)

2. **Macro-call parameter:** Input the regression parameter estimate data name (required parameter).

   Descriptions and explanation: Input the name of the temporary (member name) or permanent (libname.member_name) SAS dataset to be used in scoring the new dataset specified in macro input option #1.

   Options/examples:

   Permanent SAS dataset: gf.regest (LIBNAME: gf; SAS dataset name: regest)

   Temporary SAS dataset: regest (SAS dataset name)

3. **Macro-call parameter:** Input response variable name (optional parameter).

   Descriptions and explanation: If a response variable is available in the "new" dataset, input the name of the response variable. The RSCORE macro can also estimate the residual and compute the R^2 for prediction. If a response value is not available, leave this field blank.

   Options/examples:

   Y (name of a continuous response)

4. **Macro-call parameter:** Input model terms (required statement).

   Descriptions and explanation: Input the regression model used to develop the original regression estimates. This must be identical to the PROC REG model statements specified when estimating the regression model using the REGDIAG macro.

   Options/examples:

   X1 X2 X3 X2X3 X2SQ (X1, X2, and X3 are linear predictors; X2X3 is the interaction term; and X2SQ is the quadratic term for X2)

5. **Macro-call parameter:** Display or save SAS output and graphs (required statement).

Descriptions and explanation: Option for displaying all output/graphics files in the OUTPUT/GRAPHICS window or saving as a specific format in a folder specified in macro input option #6.
Options/examples: See Section 5.5.2 (macro input option #11) for detailed explanations for these options:

DISPLAY
WORD
WEB
PDF
TXT

Note: System messages are deleted from the LOG window if DISPLAY is not selected as the input.

6. **Macro-call parameter:** Folder to save SAS graphics and output files (optional statement).
Descriptions and explanation: To save the SAS graphics files in an EMF format suitable for inclusion in PowerPoint presentations, specify the output format as TXT in version 8.0 or later. In pre-8.0 versions, all graphic format files will be saved in a user-specified folder. Similarly, output files in WORD, HTML, PDF, and TXT formats will be saved in the user-specified folder. If this macro field is left blank, the graphics and output files will be saved in the default folder.
Option/example:
c:\output\ — folder named "OUTPUT"
7. **Macro-call parameter:** *i*th number of analysis (required statement).
Descriptions and explanation: SAS output files will be saved by forming a file name from the original SAS dataset name and using the counter value provided in this field. For example, if the original SAS dataset name is "sales" and the counter number included is 1, the SAS output files will be saved as "sales1" in the user-specified folder. By changing the counter value, users can avoid replacing the previous SAS output files with new outputs.
8. **Macro-call parameter:** Optional input ID variable (optional statement).
Descriptions and explanation: If a unique ID variable can be used to identify each record in the database, input that variable name here. This will be used as the ID variable so that any outlier/influential observations can be detected. If no ID variable is available in the dataset, leave this field blank. This macro can create an ID variable based on the observation number from the database.
Option/example:
ID NUM

5.8 Logistic Regression Using SAS Macro LOGISTIC

The LOGISTIC macro is a powerful SAS application for performing complete and user-friendly logistic regressions with and without categorical predictor variables. Options are available for performing various logistic regression diagnostic graphs and tests. The SAS procedure, LOGISTIC is the main tool used in the macro. In addition to these SAS procedures, GPLOT is also utilized in the LOGISTIC macro to obtain diagnostic graphs. The advantages of using the LOGISTIC macro over the PROC LOGISTIC include:

- Production of logistic regression diagnostic plots such as overlaid partial delta logit and simple logit plots (PROC GPLOT) for each continuous predictor variable
- Production of outlier detection plots, ROC curves, and false positive and negative plots for assessing the overall model fit
- Option for excluding extreme outliers (delta deviance >4) and then performing logistic regression using the remaining data points
- Option for validating the fitted logistic model obtained from a training dataset using an independent validation dataset by comparing Brier scores
- Options for saving the output tables and graphics in WORD, HTML, PDF, and TXT formats

Software requirements for using the LOGISTIC macro include:

- SAS/CORE, SAS/BASE, SAS/STAT, and SAS/GRAPH must be licensed and installed at the site.
- SAS version 8.0 and above is recommended for full utilization.
- An active Internet connection is required for downloading the LOGISTIC macro from the book website if the companion CD-ROM is not available.

5.8.1 Steps Involved in Running the LOGISTIC Macro

1. Create an SAS dataset (permanent or temporary) containing at least one binary response (target) variable and many continuous and/or categorical predictor (input) variables. Code 0 as non-event and 1 as an event in the data file. The LOGISTIC macro will model the probability of the event. (Disabling the SAS-enhanced editor in the latest SAS versions is highly recommended; open only one PROGRAM EDITOR window during the execution of this macro.)

2. *If the companion CD-ROM is not available*, first verify that the Internet connection is active. Open the LOGISTIC.sas macro-call file in the SAS PROGRAM EDITOR window. Instructions are given in the Appendix regarding downloading the macro-call and sample data files from the book website. *If the companion CD-ROM is available*, the LOGISTIC.sas macro-call file can be found in the mac-call folder on the CD-ROM. Open the LOGISTIC.sas macro-call file in the SAS PROGRAM EDITOR window. Click the RUN icon to submit the macro-call file LOGISTIC.sas to open the macro-call window called LOGISTIC (Figure 5.20).
3. Input the appropriate parameters in the macro-call window by following the instructions provided in the LOGISTIC macro help file in Section 5.8.2. Users can choose to exclude large extreme observations from analysis. After inputting all the required macro parameters, be sure the cursor is in the last input field and the RESULTS VIEWER window is closed, then hit the ENTER key (not the RUN icon) to submit the macro.
4. Examine the LOG window (only in DISPLAY mode) for any macro execution errors. If any errors are reported in the LOG window, activate the PROGRAM EDITOR window, resubmit the LOGISTIC.sas macro-call file, check the macro input values, and correct any input errors.
5. If no errors are found in the LOG window, activate the PROGRAM EDITOR window, resubmit the LOGISTIC.sas macro-call file, and change the macro input value from DISPLAY to any other desirable format (see Section 5.8.2). The SAS output files from complete logistic modeling and diagnostic graphs can be saved in a user-specified-format in the user-specified folder.

5.8.2 Help File for SAS Macro LOGISTIC

1. **Macro-call parameter:** Input SAS dataset name (required parameter).
   Descriptions and explanation: Include the name of the temporary (member_name) or permanent (libname.member_name) SAS dataset on which the logistic regression analysis will be performed.
   Options/examples:

   Permanent SAS dataset: gf.bank (LIBNAME: gf; SAS dataset name: bank)

   Temporary SAS dataset: bank (SAS dataset name)
2. **Macro-call parameter:** Input binary response variable name (required parameter).

Descriptions and explanation: Input the binary response variable name from the SAS dataset to be modeled as the target variable.

Option/example:

Y (name of the binary response)

3. **Macro-call parameter:** Input group variables (optional statement).

   Descriptions and explanation: To include categorical variables from the SAS dataset as predictors in logistic regression modeling, input the names of these variables.

   Options/examples:

   month manager

   Blank (categorical predictors are not used)

4. **Macro-call parameter:** Input continuous predictor variable names (optional statement).

   Descriptions and explanation: For each continuous predictor variable specified, an overlaid partial delta logit/simple logit plot will be produced automatically. This diagnostic plot is useful in checking for nonlinearity, the significance of the parameter estimate, and the presence of multicollinearity among the predictor variables. If categorical variables are also included in macro input option #3, the logistic parameter estimates are adjusted for these categorical variables. No interaction terms between categorical and the continuous predictors are included in the diagnostic plots when computing the delta logit values. Leave this field blank when only categorical variables are in the model statement.

   Options/examples:

   X1 X2 X3

5. **Macro-call parameter:** Input model terms (required option).

   Descriptions and explanation: This macro input field is equivalent to the right side of the equal sign in the PROC LOGISTIC model statement. You can include main effects of categorical variables, linear effects of continuous variables, and any interactions among these effects.

   Options/examples:

   X1 X2 X1X2 X1SQ (continuous predictor variables, including linear predictors X1 and X2, an interaction term X1X2, and the quadratic term for X1, X1SQ)

   X1 SOURCE X1*SOURCE (logistic regression with categorical variables SOURCE, linear predictor X1, and an interaction term between X1 and SOURCE, X1*SOURCE)

   (For details about specifying model statements, refer to the SAS online manuals on PROC LOGISTIC.[36])

6. **Macro-call parameter:** Exploratory analysis (optional statement).

Descriptions and explanation: To perform exploratory analysis, input YES in this field, which will result in partial delta logit plots and variable selection by forward selection method for continuous predictors, or predicted probability plots for the categorical variable model. If this field is left blank, this macro skips the exploratory analysis step and goes directly to the logistic regression step.

Options/examples:

YES: Perform exploratory analysis

Blank: Skip exploratory analysis

7. **Macro-call parameter:** Overdispersion correction (required statement).

   Descriptions and explanation: To not adjust for overdispersion, input NONE. But, if the test for overdispersion indicates that a high degree of overdispersion exists, adjust for it by inputting either DEVIANCE or PEARSON.

   Options/examples:

   NONE: No overdispersion adjustment

   DEVIANCE: Overdispersion adjustment by DEVIANCE factor

   PEARSON: Overdispersion adjustment by PEARSON factor

8. **Macro-call parameter:** Customized odds ratios/parameter test (optional statement).

   Descriptions and explanation: Input appropriate statements to obtain customized odds ratio estimates (UNITS option) or test the parameter estimates for specific values (TEST option).

   Options/examples:

   Units x1=0.5 –0.5 (to obtain customized odds ratio estimate for X1 predictor when X1 is increased or decreased by 0.5 unit)

   Test x1=0.05 (to test the hypothesis that the X1 parameter estimate is equal to 0.5)

   Units x1=0.5 –0.5 ; Test x1=0.05 (to obtain both customized odds ratio and performing parameter test)

   Note the ";" after the first statement. When more than one statement is specified, include a ";" at the end of each statement (except for the last statement).

9. **Macro-call parameter:** Input validation dataset name (optional parameter).

   Descriptions and explanation: To validate the logistic regression model obtained from a training dataset by using an independent validation dataset, input the name of the SAS validation dataset. This macro estimates predicted probabilities for the validation dataset using the model estimates derived from the training data. The success of this prediction could be verified by checking the deviance residual and the Brier scores for the validation dataset. Input the name of the temporary (member name) or permanent (libname.member_name) SAS dataset to be treated as the validation data.

Options/examples:

Permanent SAS dataset: gf.valid (LIBNAME: gf; SAS dataset name: valid)

Temporary SAS dataset: valid (SAS dataset name)

10. **Macro-call parameters:** Input ID variable (optional statement).

    Descriptions and explanation: If a unique ID variable can be used to identify each record in the database, input that variable name here. This will be used as the ID variable so that any outlier/influential observations can be detected. If no ID variable is available in the dataset, this field can be left blank. This macro can create an ID variable based on the observation number from the database.

    Option/example:

    ID NUM

11. **Macro-call parameter:** *z*th number of analysis (required statement).

    Descriptions and explanation: SAS output files will be saved by forming a file name from the original SAS dataset name and the counter value provided in this field. For example, if the original SAS dataset name is "sales" and the counter number included is 1, the SAS output files will be saved as "sales1" in the user-specified folder. By changing the counter value, users can avoid replacing previous SAS output files with new outputs.

12. **Macro-call parameter:** Display or save SAS output and graphs (required statement).

    Descriptions and explanation: Option for displaying all output files in the OUTPUT window or saving as a specific format in a folder specified in macro input option #12.

    Options/examples: See Section 5.5.2 for explanation of these formats:

    DISPLAY
    WORD
    WEB
    PDF
    TXT

    Note: System messages are deleted from the LOG window if DISPLAY is not selected as the input.

13. **Macro-call parameter:** Folder to save SAS graphics and output files (optional statement).

Descriptions and explanation: To save the SAS graphics files in an EMF format suitable for inclusion in PowerPoint presentations, specify the output format as TXT in version 8.0 or later. In pre-8.0 versions, all graphic format files will be saved in a user-specified folder. Similarly, output files in WORD, HTML, PDF, and TXT formats will be saved in the user-specified folder. If this macro field is left blank, the graphics and output files will be saved in the default folder.

Option/example:

c:\output\ — folder named "OUTPUT"

14. **Macro-call parameter:** Adjust for extreme influential observations (optional parameter).

    Descriptions and explanation: If YES is input for this option, the macro will fit the logistic regression model after excluding extreme (delta deviance, >4.0) observations. An output of all excluded observations is also produced.

    Options/examples:

    YES: Extreme outliers will be excluded from the analysis
    Blank: All observations in the dataset will be used

15. **Macro-call parameter:** Input cutoff *p* value (required option).

    Descriptions and explanation: Input the cutoff *p* value for classifying the predicted probability as event or non-event.

    Options/examples:

    0.45 0.5 0.55 0.60

5.9 Scoring New Logistic Regression Data Using the SAS Macro LSCORE

The LSCORE macro is a powerful SAS application for scoring new datasets using the established logistic regression model estimates. Options are available for checking the residuals graphically if the observed binary response variable is available in the new dataset; otherwise, only predicted probability scores are produced and saved in an SAS dataset. In addition to the SAS LOGISTIC procedure, GPLOT is also utilized in the LSCORE macro. No procedure or options are currently available in SAS systems to compare the residuals from the scoring data automatically.

Software requirements for using the LSCORE macro include:

- SAS/CORE, SAS/BASE, SAS/STAT, and SAS/GRAPH must be licensed and installed at the site.
- SAS version 8.0 and above is recommended for full utilization.

- An active Internet connection is required for downloading the LSCORE macro from the book website if the companion CD-ROM is not available.

5.9.1 Steps Involved in Running the LSCORE Macro

1. Create a new scoring SAS dataset (permanent or temporary) containing binary response (optional) variables and continuous or categorical predictor (input) variables. This new score dataset should contain all the predictor variables that were used to develop the original logistic regression model. (Disabling the SAS-enhanced editor in the latest SAS versions is highly recommended; open only one PROGRAM EDITOR window.)
2. Verify that the logistic regression parameter estimates are available in an SAS dataset. If the LOGISTIC macro was used to fit the original binary logistic regression model, it should have created an SAS dataset called "estim". Check the contents of the "estim" dataset and create a temporary dataset from this to use for scoring the new dataset.
3. *If the companion CD-ROM is not available*, first verify that the Internet connection is active. Open the LSCORE.sas macro-call file in the SAS PROGRAM EDITOR window. Instructions are given in the Appendix regarding downloading the macro-call and sample data files from the book website. *If the companion CD-ROM is available*, the LSCORE.sas macro-call file can be found in the mac-call folder in the CD-ROM. Open the LSCORE.sas macro-call file in the SAS PROGRAM EDITOR window. Click the RUN icon to submit the macro-call file LSCORE.sas to open the macro-call window called LSCORE.
4. Input the appropriate parameters in the macro-call window by following the instructions provided in the LSCORE macro help file in Section 5.9.2. After inputting all the required macro parameters, be sure the cursor is in the last input field and the RESULTS VIEWER window is closed, then hit the ENTER key (not the RUN icon) to submit the macro.
5. Examine the LOG window (only in DISPLAY mode) for any macro execution errors. If any errors are reported in the LOG window, activate the PROGRAM EDITOR window, resubmit the LSCORE.sas macro-call file, check the macro input values, and correct any input errors.
6. If no errors are found in the LOG window, activate the PROGRAM EDITOR window, resubmit the LSCORE.sas macro-call file, and change the macro input value from DISPLAY to any other desirable format (see Section 5.9.2). SAS output files from the LSCORE

analysis and LSCORE charts can be saved as user-specified-format files in the user-specified folder.

5.9.2 Help File for Using SAS Macro LSCORE

1. **Macro-call parameter:** Input name of the new scoring SAS dataset name (required parameter).
   Descriptions and explanation: Include the name of the temporary (member name) or permanent (libname.member name) SAS dataset to be used for estimating scores using the established logistic regression model estimates.
   Options/examples:
      Permanent SAS dataset: gf.new (LIBNAME: gf; SAS dataset name: new)
      Temporary SAS dataset: new (SAS dataset name)
2. **Macro-call parameter:** Input optional categorical variables (optional statement).
   Descriptions and explanation: If categorical variables are in the new "score" dataset and categorical variables were used in the original logistic regression as predictors, input the names of these variables.
   Options/examples:
      month manager (categorical variables)
      Blank: No categorical variables were used in the original model building
3. **Macro-call parameter:** Input optional binary response variable name (optional parameter).
   Descriptions and explanation: If a binary response variable is available in the "new" dataset, input the name of the response variable. The LSCORE macro can also estimate the residual and investigate the model fit graphically. If a binary response value is not available, leave this field blank.
   Options/examples:
      Y (name of a binary response)
4. **Macro-call parameter:** Input the logistic regression parameter estimate data name (required parameter).
   Descriptions and explanation: Include the name of the temporary (member name) or permanent (libname.member_name) SAS dataset to be used in scoring the new logistic regression dataset specified in macro input option #1.
   Options/examples:
      Permanent SAS dataset: gf.estim (LIBNAME: gf; SAS dataset name: estim)

Temporary SAS dataset: estim (SAS dataset name)

5. **Macro-call parameter:** Input model terms (required statement).
   Descriptions and explanation: Input the logistic regression model used to develop the original parameter estimates. This model statement must be identical to the PROC LOGISTIC model statements specified when estimating the regression model using the LOGISTIC macro.
   Options/examples:
   X1 X2 X3 X2X3 X2SQ (X1, X2, and X3 are linear predictors; X2X3 is an interaction term; X2SQ is a quadratic term for X2)
6. **Macro-call parameter:** Input the cutoff probability value for classifying predicted scores (required statement).
   Descriptions and explanation: Input a cutoff value for classifying the predicted scores as event or non-event in the new logistic regression data.
   Options/examples:
   0.45 0.5 0.55 0.6
7. **Macro-call parameter:** Input ID variable (optional statement).
   Descriptions and explanation: If a unique ID variable can be used to identify each record in the database, input that variable name here. This will be used as the ID variable so that any outlier/influential observations can be detected. If no ID variable is available in the dataset, leave this field blank. This macro can create an ID variable based on the observation number from the database.
   Option/example:
   ID NUM
8. **Macro-call parameter:** Display or save SAS output and graphs (required statement).
   Descriptions and explanation: Option for displaying all output and graphics files in the OUTPUT/GRAPHICS window or saving as a specific format in a folder specified in macro input option #9.
   Options/examples: See Section 5.5.2 for explanation of these formats:
   DISPLAY
   WORD
   WEB
   PDF
   TXT
   Note: System messages are deleted from the LOG window if DISPLAY is not selected as the input.
9. **Macro-call parameter:** Folder to save SAS graphics and output files (optional statement).

Descriptions and explanation: To save the SAS graphics files in an EMF format suitable for inclusion in PowerPoint presentations, specify the output format as TXT in version 8.0 or later. In pre-8.0 versions, all graphic format files will be saved in a user-specified folder. Similarly, output files in WORD, HTML, PDF, and TXT formats will be saved in the user-specified folder. If this macro field is left blank, the graphics and output files will be saved in the default folder.

Option/example:

c:\output\ — folder named "OUTPUT"

10. **Macro-call parameter:** *i*th number of analysis (required statement).
Descriptions and explanation: SAS output files will be saved by forming a file name from the original SAS dataset name and the counter value provided in this field. For example, if the original SAS dataset name is "new" and the counter number included is 1, the SAS output files will be saved as "new1" in the user-specified folder. By changing the counter value, users can avoid replacing previous SAS output files with new outputs.

5.10 Case Study 1: Modeling Multiple Linear Regression

5.10.1 Study Objectives

1. **Data exploration using diagnostic plots:** Check for linear and nonlinear relationships and significant outliers (augmented partial residual plot), significant regression relationship (partial leverage plot), multicollinearity (VIF plot), and detection of significant interaction (interaction plot) for each predictor variable.
2. **Variable selection using R^2 method of all possible model selection:** Perform all possible regression models for each subset (one variable, two variable, three variable, and so on) and investigate the two best regressions models for each subset. Select the best subset and the best model using the C_p, adjusted R^2, and AIC statistics.
3. **Model selection:** Check for the significance of linear, quadratic, and cross products and the overall significance for each predictor variable.
4. **Regression model fitting and prediction:** Perform hypothesis testing on overall regression and on each parameter estimate. Estimate confidence intervals for parameter estimates, predict scores, and estimate their confidence intervals.
5. **Checking for any violations of regression assumptions:** Perform statistical tests and graphical analysis to detect influential

outliers, multicollinearity among predictor variables, heteroscedasticity in residuals, and departure from normally distributed residual.

6. **Save "_score_" and "regest" datasets for future use:** These two datasets are created and saved as temporary SAS datasets in the work folder. The "_score_" dataset contains the observed variables, predicted scores (including observations with missing response value), residuals, and confidence-interval estimates. This dataset could be used as the base for developing the scorecards for each observation. Also, the second SAS dataset called "regest" contains the parameter estimates, which could be used in the RSCORE macro for scoring various datasets containing the same variables.
7. **If–then analysis and lift charts:** Perform if–then analysis and construct a lift chart to estimate the differences in the predicted response when one of the continuous predictor variables is fixed at a given value.

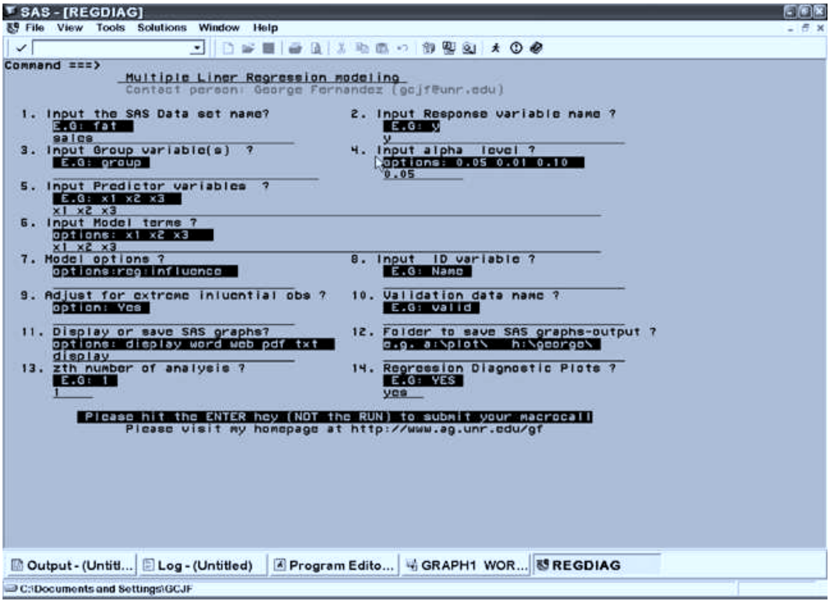
5.10.2 Data Descriptions

Data name	Permanent SAS dataset "sales" located in the library "gf"
Response (Y) and predictor variables (X)	Y: response (product sales in $1000/year) X1: advertising cost/year X2: advertising cost ratio between competing product and own product X3: personal disposable income/year
ID variable	ID: year (1 to 116)
Number of observations	116
Source	Simulated data with a controlled trend and minimum amount of error

Open the REGDIAG.sas macro-call file in the SAS PROGRAM EDITOR window and click RUN to open the REGDIAG macro-call window (Figure 5.1). Input the appropriate macro-input values by following the suggestions given in the help file (Section 5.5.2).

5.10.3 Exploratory Analysis/Diagnostic Plots

Input dataset name, response, predictor variable names, model terms, and the alpha level. Leave the group variable option blank because all the predictors used are continuous. To perform data exploration and to create regression diagnostic plots, input YES in macro input field #14. Submit



```
Command ===>
                Multiple Liner Regression modeling
                Contact person: George Fernandez (gcjf@unr.edu)

 1. Input the SAS Data set name?          2. Input Response variable name ?
    E.G: fat                                 E.G: y
    sales                                    y
 3. Input Group variable(s)  ?            4. Input alpha  level ?
    E.G: group                               options: 0.05 0.01 0.10
                                             0.05
 5. Input Predictor variables  ?
    E.G: x1 x2 x3
    x1 x2 x3
 6. Input Model terms ?
    options: x1 x2 x3
    x1 x2 x3
 7. Model options ?                       8. Input  ID variable ?
    options:reg:influence                    E.G: Name

 9. Adjust for extreme inluential obs ?  10. Validation data name ?
    option: Yes                              E.G: valid

11. Display or save SAS graphs?          12. Folder to save SAS graphs-output ?
    options: display word web pdf txt        e.g. a:\plot\   h:\george\
    display
13. zth number of analysis ?             14. Regression Diagnostic Plots ?
    E.G: 1                                   E.G: YES
    1                                        yes

      Please hit the ENTER key (NOT the RUN) to submit your macrocall
           Please visit my homepage at http://www.ag.unr.edu/gf
```

Figure 5.1 Screen copy of REGDIAG macro-call window showing the macro-call parameters required for performing multiple linear regression (MLR).

the REGDIAG macro and three regression diagnostic plots for each predictor variable produced.

Simple linear regression and augmented partial residual (APR) plots for all three predictor variables are presented in Figures 5.2 to 5.4. The linear/quadratic regression parameter estimates for the simple and multiple linear regressions and their significance levels are also displayed in the titles. The simple linear regression line describes the relationship between the response and the predictor variable in a simple linear regression. The APR line shows the quadratic regression effect of the *i*th predictor on the response variable after accounting for the linear effects of other predictors on the response. The APR plot is very effective in detecting significant outliers and nonlinear relationships. Significant outliers or influential observations are identified and marked on the APR plot if the absolute Student value exceeds 2.5 or the DFFITS statistic exceeds 1.5. These influential statistics are derived from the MLR model involving all predictor variables. If the correlations among the predictor variables are negligible, the simple and partial regression lines should have similar slopes.

The APR plots for the three predictor variables show significant linear relationships between the three predictors and annual sales. The advertising cost/year (Figure 5.2) and the personal disposable income (Figure 5.4) had very significant positive effects on the product sales.

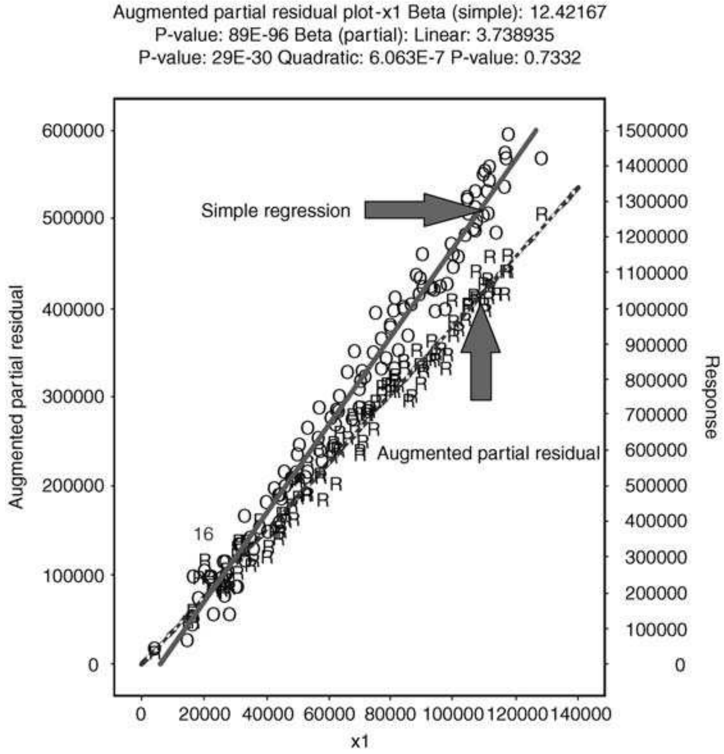


Figure 5.2 Regression diagnostic plot using SAS macro REGDIAG: overlay plot of simple linear regression and augmented partial residual plots for X1.

A big difference in the magnitude of the partial (adjusted) and the simple (unadjusted) regression effects for advertising (X1) and advertising ratio (X2) on product sales were clearly evident (Figures 5.2 to 5.3). Quadratic effects of all three predictor variables on sales were not significant at the 5% level. One significant outlier (observation #16) was detected in this APR plot. These results were expected because the sales data are a simulated dataset with fairly controlled parameter estimates and error values.

Partial leverage (PL) plots for all three predictor variables are presented in Figures 5.5 to 5.7. The PL display shows three curves: (1) the vertical reference line that goes through the response variable mean; (2) the partial regression line that quantifies the slope of the partial regression coefficient of the *i*th variable in the MLR; and (3) the 95% confidence band for the partial regression line. The partial regression parameter estimates for the *i*th variable in the multiple linear regression and their significance levels are also displayed in the titles. The slope of the partial regression coefficient

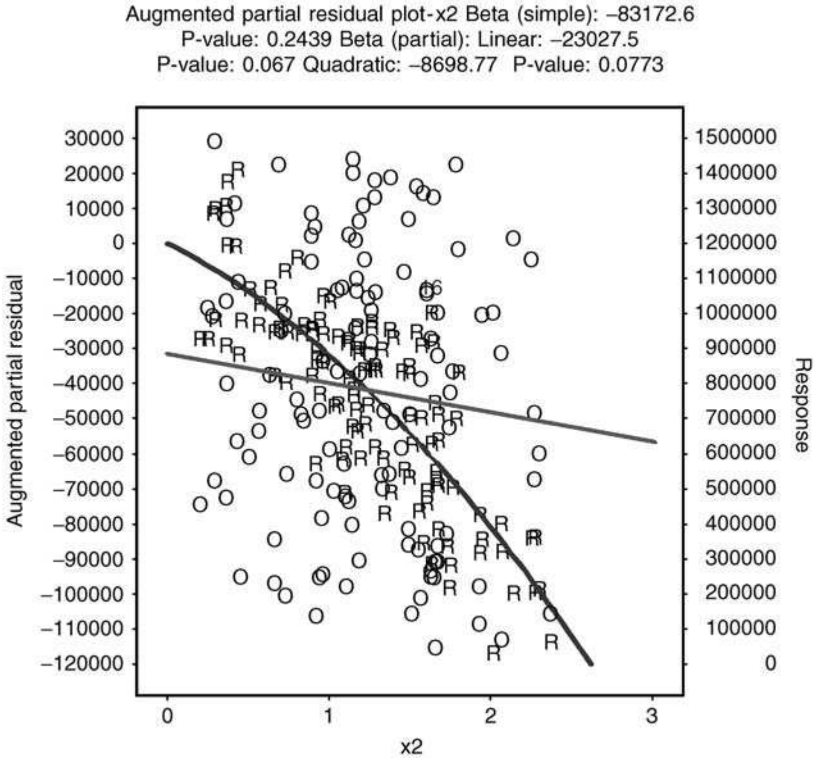


Figure 5.3 Regression diagnostic plot using SAS macro REGDIAG: overlay plot of simple linear regression and augmented partial residual plots for X2.

is considered statistically significant at the 5% level if the response mean line intersects the 95% confidence band. If the response mean line lies within the 95% confidence band without intersecting it, then the partial regression coefficient is considered insignificant.

The PL plots for the three predictor variables showed significant linear relationships between the three predictors and annual sales. The advertising cost/year (Figure 5.5) and the personal disposable income (Figure 5.7) had very significant positive effects on the product sales. The advertising cost ratio between competing product and own product (X2) showed a significant negative effect on the product sales. These results were expected because the sales data were simulated with specified parameter estimates and error values.

The VIF plots for all three predictor variables are presented in Figures 5.8 to 5.10. The VIF plot displays two overlaid curves: (1) the relationship between (partial residual + response mean) and the *i*th predictor variable, and (2) the relationship between the (partial leverage + response mean)

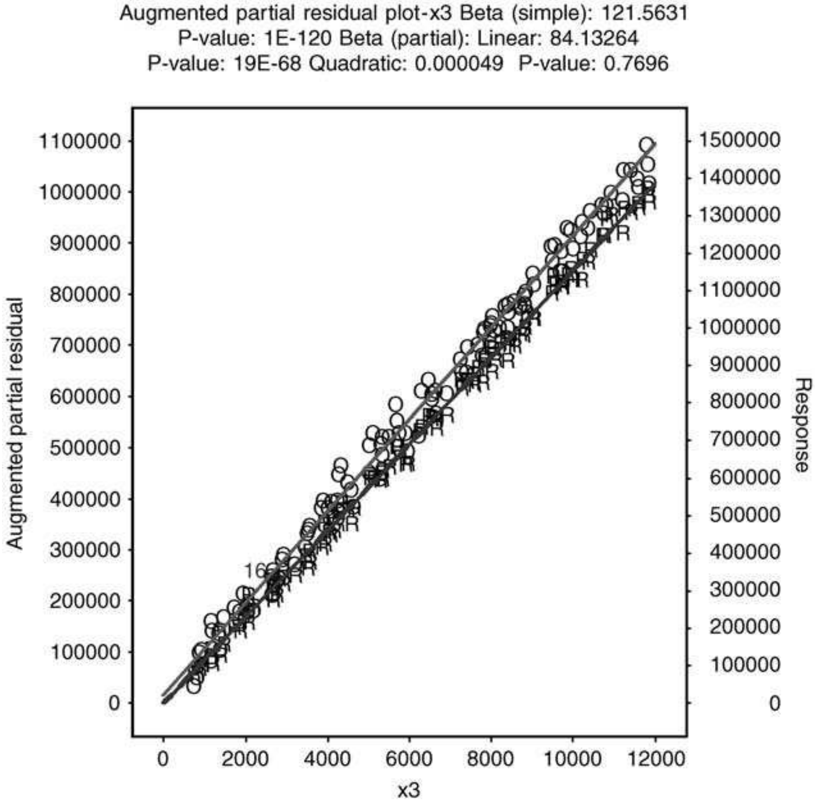


Figure 5.4 Regression diagnostic plot using SAS macro REGDIAG: overlay plot of simple linear regression and augmented partial residual plots for X3.

and the (partial *i*th predictor value + mean of the *i*th predictor value). The slope of both regression lines should be equal to the partial regression coefficient estimate for the *i*th predictor. When the degree of multicollinearity is not high, both the partial residual (symbol R) and the partial leverage (symbol E) values should be evenly distributed around the regression line. In the presence of severe multicollinearity, however, the partial leverage (E) values shrink and are distributed around the mean of the *i*th predictor variable. Also, the partial regression for the *i*th variable shows a nonsignificant relationship in the partial leverage plots, whereas the partial residual plot shows a significant trend for the *i*th variable. Furthermore, the degree of multicollinearity can be measured by the VIF statistic in an MLR model, and the VIF statistic for each predictor variable is displayed on the title statement of the VIF plot.

The VIF plots for X1 and X3 showed a moderate level of multicollinearity where the crowding of the partial leverage values were evident

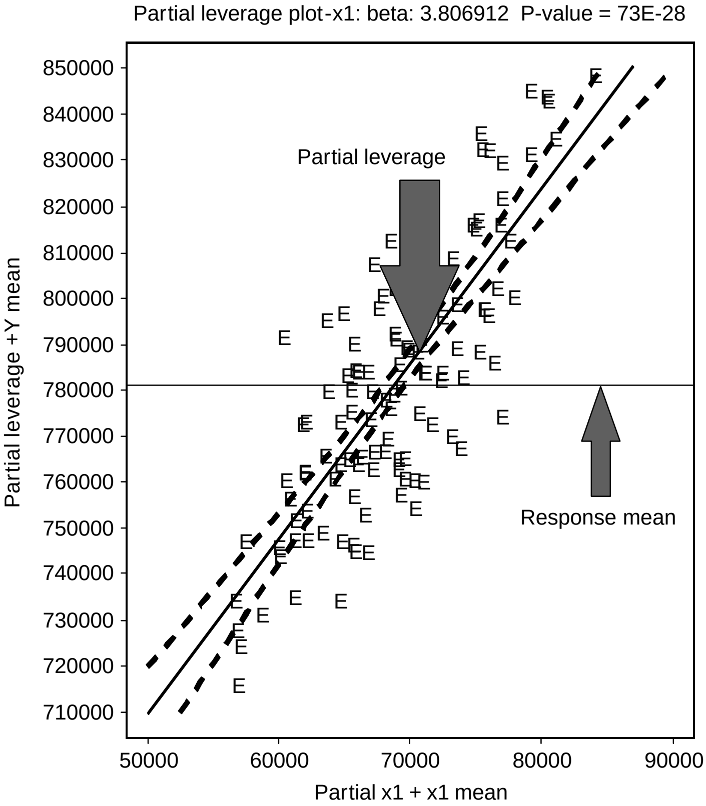


Figure 5.5 Regression diagnostic plot using SAS macro REGDIAG: partial leverage plot for X1.

in Figures 5.8 and 5.10. The VIF estimates for these two predictors were larger than 25; however, this significant multicollinearity did not impact the regression parameter estimates or their statistical significance levels. Because we used simulated data with specified parameter estimates, sample size, and small error variation, the impact of the multicollinearity on regression estimates was not significant.

The significance of all possible two-factor interactions can be verified graphically by examining a regression plot between full model residuals + interaction term in question (Y axis) with the interaction term in question (X axis). If a significant trend in the partial interaction plot is observed, then the interaction term is considered significant. The significance level of the interaction term is also displayed in the title. None of the interaction terms is not significant (plots not shown).

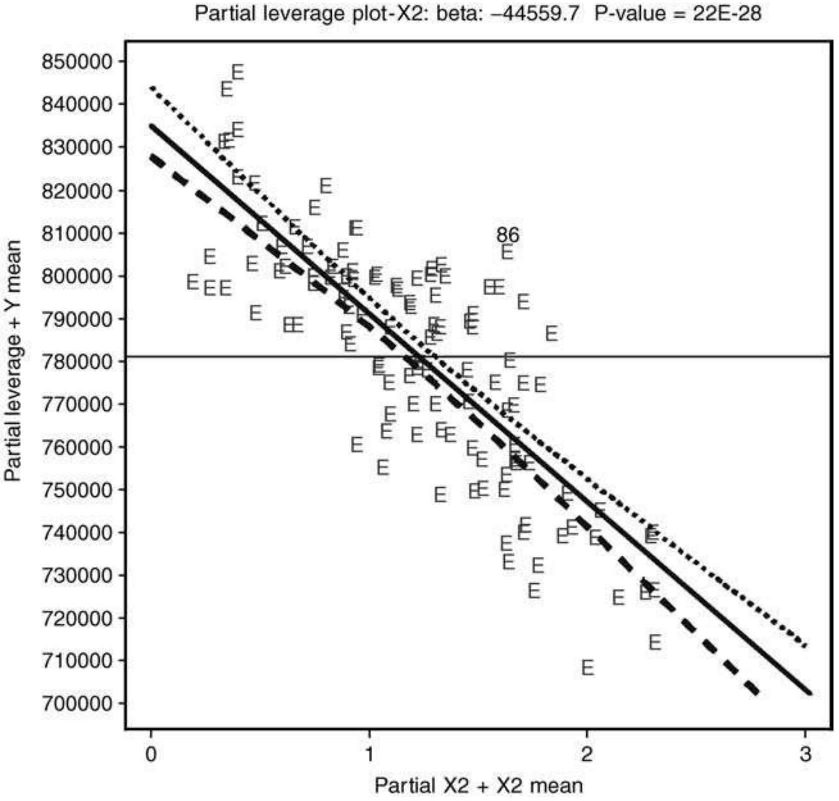


Figure 5.6 Regression diagnostic plot using SAS macro REGDIAG: partial leverage plot for X2.

5.10.4 Variable Selection/Regression Analysis

Input the dataset name, response, predictor variable names, model terms, and alpha level. Leave the group variable option blank because all the predictors used are continuous. Leave macro input field #14 blank to skip regression diagnostics and to run MLR. Submit the REGDIAG macro and to get SAS summary outputs and graphics for all possible variable selection, outputs from model specification errors, regression estimates and predictions, and graphics for checking for violations of errors.

5.10.5 Variable Selection Using R^2 Selection Method

The REGDIAG macro utilizes all possible regression models using the R^2 selection methods and outputs the best two models for all subsets (Table

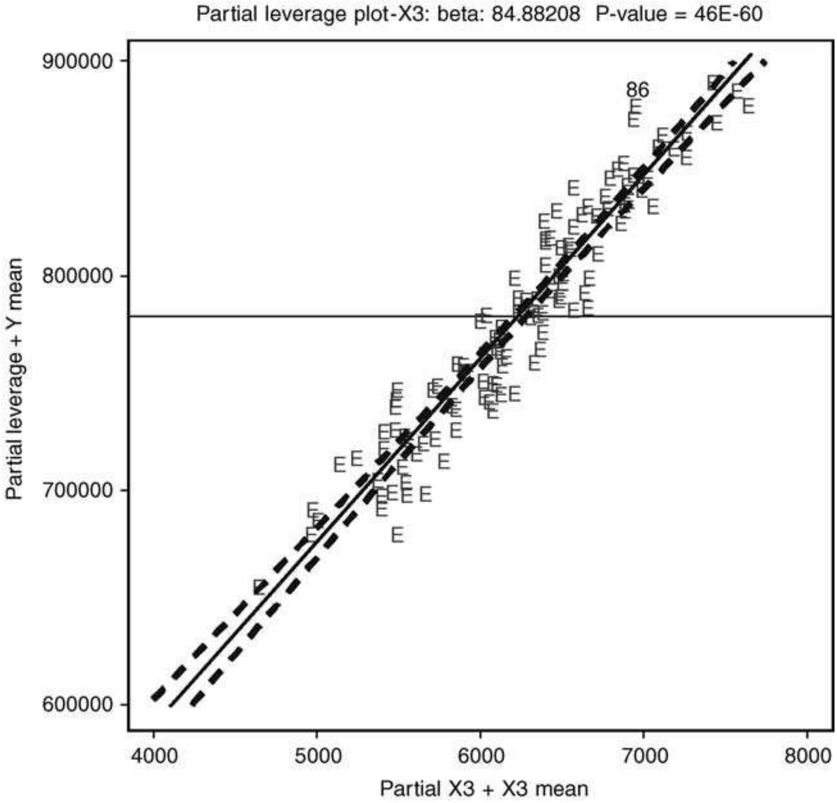


Figure 5.7 Regression diagnostic plot using SAS macro REGDIAG: partial leverage plot for X3.

5.1). Because three predictor variables were used in the model selection, the full model had three predictors. Two subsets and one variable subset are possible with three predictor variables. By comparing the R^2, $R^2_{\text{(adjusted)}}$, RMSE, C_p, and AIC values between the full model and all subsets, we can conclude that the full model is superior to all other subsets. Even though the drop in R^2 and $R^2_{\text{(adjusted)}}$ from the full model and the subsets was not large, Mallows' C_p value and the RMSE increased significantly from the full model to the subsets. Mallows' C_p measures the total squared error for a subset equal to the total error variance plus the bias introduced by not including the important variables in the subset. The C_p plot (Figure 5.11) shows the C_p statistic against the number of predictor variables for the full model and the best two models for each subset. Additionally, the RMSE statistics for the full model and the best two regression models in each subset are also shown in the C_p plot. Furthermore, the diameter of the bubbles in the C_p plot is proportional to the magnitude of the RMSE.

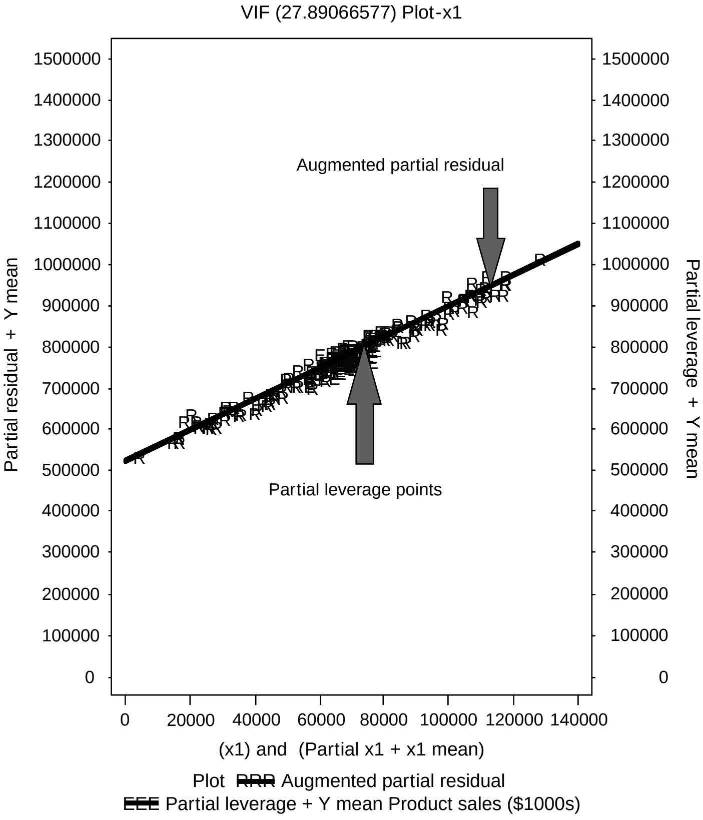


Figure 5.8 Regression diagnostic plot using SAS macro REGDIAG: variance inflation factor (VIF) plot for X1.

Consequently, dropping any variable from the three-variable model is not recommended because the C_p, RMSE, and AIC values jump up so high. However, the R^2 and $R^2_{(\text{adjusted})}$ values did not change much from the full model and subsets. These results clearly indicate that C_p, RMSE, and AIC statistics are better indicators for variable selection than R^2 and $R^2_{(\text{adjusted})}$. Thus, the C_p plot and the summary table of model selection statistics produced by the REGDIAG macro can be used effectively in selecting the best subset in regression models with many (5 to 25) predictor variables.

5.10.6 Checking for Model Specification Errors

Failure to include significant quadratic or interaction terms results in model specification errors in MLR. The APR plot discussed in the regression

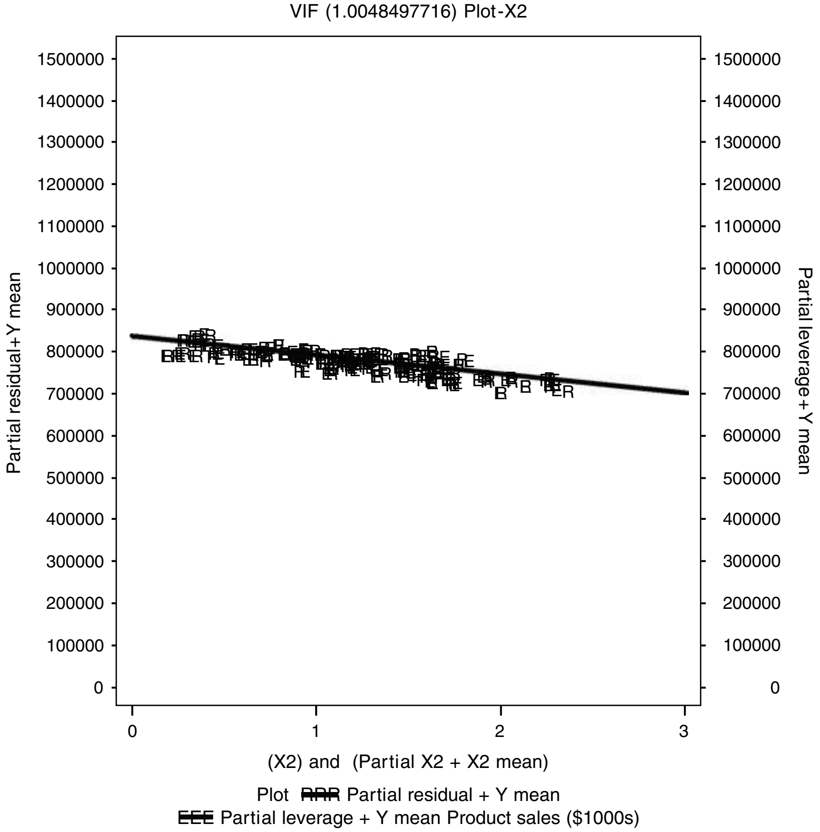


Figure 5.9 Regression diagnostic plot using SAS macro REGDIAG: variance inflation factor (VIF) plot for X2.

diagnostic plots section is effective in detecting the quadratic trend but not the interaction effects. The interaction diagnostic plot produced in data exploration could be used to select the significant interaction required. The REGDIAG macro also utilizes the RSREG procedure and tests the overall significance of linear, quadratic, and two-factor cross products using the sequential sum of squares (type I SS). If the regression data contain replicated observations, a lack-of-fit test could be performed to check for model specification errors by testing the significance of the deviation from the regression using the experimental error as the error term. The statistical significance of model specification errors is presented in Table 5.2.

The linear model accounted for 99% of the variation in response and was highly significant (*p* value, <0.0001). Both quadratic and the two-factor cross products were not significant. If either quadratic or

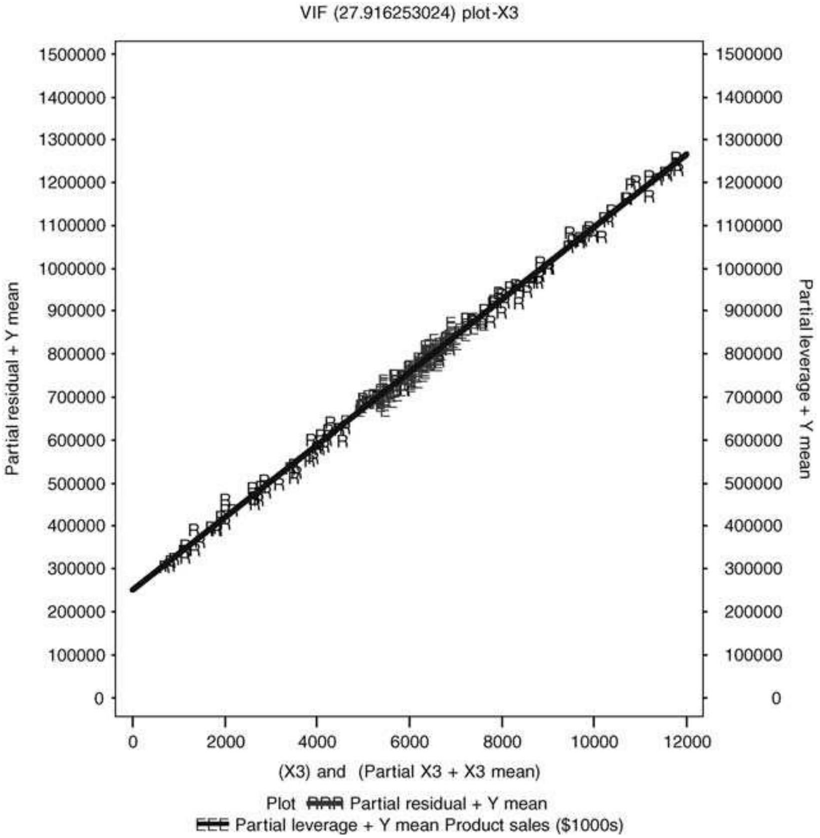


Figure 5.10 Regression diagnostic plot using SAS macro REGDIAG: variance inflation factor (VIF) plot for X3.

cross-product effects are significant and specific significant higher order terms are to be further selected, PROC RSREG also includes a table of significance levels for all possible higher order terms (table not shown here). If the higher order terms are significant and you would like to determine the most significant variable or estimate the amount of reduction in the model SS when you drop one variable completely, the RSREG procedure includes a table with useful information (Table 5.3). For each variable, the SS accounts for the linear, quadratic, and all-possible combination of two-factor interactions. Even though all three variables are highly significant, dropping the X3 variable results in a significant reduction in the model SS. Thus, the most significant factor in determining the product sales is personal disposable income (X3).

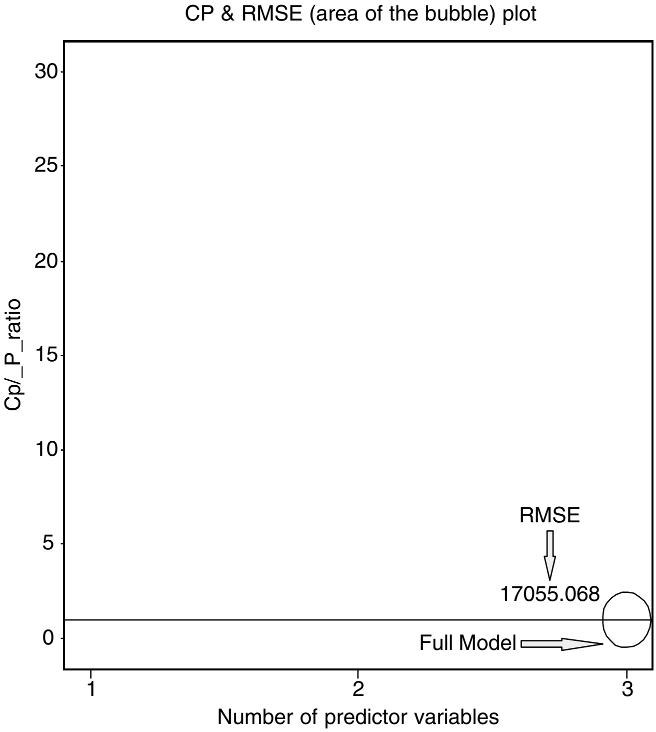


Figure 5.11 Model selection using SAS macro REGDIAG: Mallow's C_P plot for selecting the best model.

5.10.7 Regression Model Fitting

Simple correlation estimates among the predictor variables and the response are given in Table 5.4. The cost of advertising (X1) and personal income (X3) had a very high correlation (>0.9) with the sales amount. Similarly, these two predictors (X1 and X3) were also highly correlated. This high degree of correlation may cause some degree of multicollinearity in the model estimates. The observed correlation between the ratio of advertising (X2) and the sales was very small.

The overall regression model fit was highly significant based on the *F* test in the ANOVA model (Table 5.5). This result indicates that at least one of the regression coefficient slopes was not equal to zero. The R^2 and the $R^2_{(\text{adjusted})}$ estimates were almost identical and very high (0.998), indicating that no redundant predicted variable existed in the fitted regression model. The RMSE estimate is the error standard deviation and is a very useful indicator for model selection using the same data. The overall

Table 5.1 Macro REGDIAG: Variable Selection Summary Report Method, R^2 Selection

Number in Model	*R^2*	*Adjusted R^2*	*C_p*	*Akaike Information Criterion (AIC)*	*Root-Mean-Square Error (RMSE)*	*Schwartz's Bayesian Criterion (SBC)*	*Variables in Model*
1	0.9918	0.9917	394.1154	2435.5433	35936	2441.05045	X3
1	0.9766	0.9764	1329.937	2556.9927	60656	2562.49993	X1
2	0.9949	0.9948	204.4576	2382.3371	28451	2390.59784	X2, X3
2	0.9948	0.9947	211.1202	2384.7692	28751	2393.02992	X1, X3
3	0.9982	0.9981	4.0000	2264.5844	17055	2275.59879	X1, X2, X3

Table 5.2 Macro REGDIAG: Model Specification Error Summary Table Produced from the RSREG Procedure

Regression	*Degrees of Freedom*	*Type I Sum of Squares*	*R^2*	*F Value*	*Pr > F*
Linear	3	1.7873811E13	0.9982	20804.6	<.0001
Quadratic	3	912813729	0.0001	1.06	0.3683
Cross products	3	1309424767	0.0001	1.52	0.2125
Total model	9	1.7876034E13	0.9983	6935.74	<.0001

Table 5.3 Macro REGDIAG: Overall Significance (Linear, Quadratic, and Cross Product) of All Predictor Variables

RSCORE	Degrees of Freedom	Sum of Squares (SS)	Mean Square	F Value	Pr > F	Label
X1	4	57603157452	14400789363	50.29	<.0001	Advertising ($1000)
X2	4	61016412516	15254103129	53.27	<.0001	Advertising ratio[a]
X3	4	295097651218	73774412804	257.61	<.0001	Personal disposable income ($billion)

[a] Advertising cost ratio between competing product and own product.

Table 5.4 Macro REGDIAG: Simple Linear Correlations Coefficients Among All Predictors and the Response Variable

		Correlation			
Variable	Label	X1	X2	X3	Y
X1	Advertising ($1000)	1.0000	–0.0442	0.9819	0.9882
X2	Advertising ratio[a]	–0.0442	1.0000	–0.0536	–0.1091
X3	Personal disposable income ($billion)	0.9819	–0.0536	1.0000	0.9959
Y	Product sales ($1000)	0.9882	–0.1091	0.9959	1.0000

[a] Advertising cost ratio between competing product and own product.

Table 5.5 Macro REGDIAG: Testing for Overall Regression Model Fit by ANOVA

		Analysis of Variance			
Source	*Degrees of Freedom*	*Sum of Squares (SS)*	*Mean Square*	*F Value*	*Pr > F*
Model	3	1.787381E13	5.957937E12	20482.8	<.0001
Error	112	32578039964	290875357	—	—
Corrected total	115	1.790639E13	—	—	—

response mean (total sales), its percentage of coefficient variation (CV), sum of all residuals, and the SS residual are also presented in Table 5.6. Problems with rounding off and numerical accuracy in the matrix inversion can also be detected if the sum of all residuals is not equal to zero or error sum of squares (SSE) and SS residuals are not equal. No big differences between the SS residual and the PRESS statistic indicate the absence of significant influential observations in the data. The predictive potential of the fitted model can be determined by estimating the $R^2_{(\text{prediction})}$ by substituting PRESS for SSE in the formula for the R^2 estimation. The predictive power of the estimated regression model is considered high if the $R^2_{(\text{prediction})}$ estimate is large and closer to the model R^2.

The overall model fit is illustrated in Figure 5.12 by displaying the relationship between the observed response variable and predicted values. The N, R^2, $R^2_{(\text{adjusted})}$, and RMSE statistics that were useful in comparing regression models and the regression model are also included on the

Table 5.6 Macro REGDIAG: Testing for Overall Regression Model Fit Indicators

Root-mean-square error (RMSE)	17055
Dependent mean	781025
Coefficient variation (CV)	2.18368
R^2	0.9982
$R^2_{(\text{adjusted})}$	0.9981
Sum of residuals	0
Sum of squared residuals	32578039964
Predicted residual sum of squares (PRESS)	34897862201

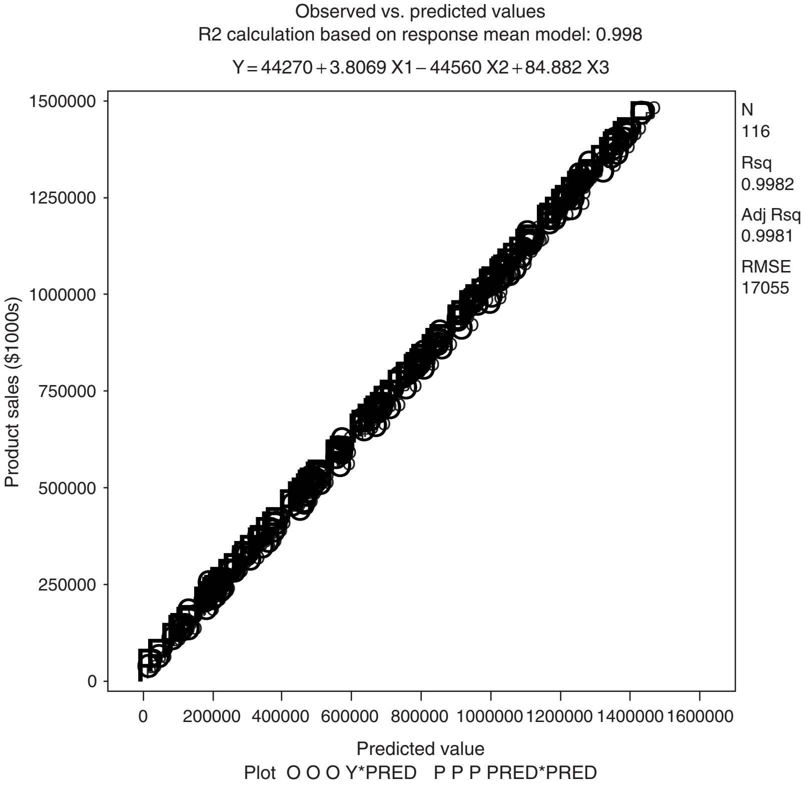


Figure 5.12 Assessing the multiple linear regression (MLR) model fit using SAS macro REGDIAG: overall model fit plot.

plot. If the data contained replicated observations, the deviation from the model includes both pure error and deviation from the regression. The R^2 estimates can be computed from a regression model using the means of the replicated observations as the response. Consequently, the R^2 computed based on the means ($R^2_{\text{(mean)}}$) is also displayed in the title statement. When no replicated data are available, $R^2_{\text{(mean)}}$ and the R^2 estimate reported by the PROC REG will be identical.

Figure 5.13 shows the total and the unexplained variation in the response variable after accounting for the regression model graphically. The ordered and the centered response variables vs. the ordered sequence display the total variability in the response. This ordered response shows a linear trend without any sharp edges at both ends because the response variable has a normal distribution. The unexplained variability in the response variable is given by the residual distribution. The residual variation shows a random distribution without any sudden peaks, trends, or

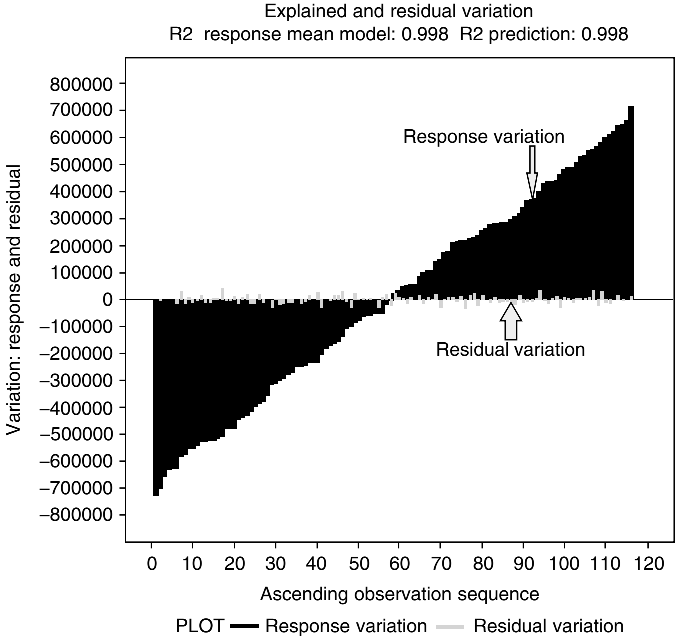


Figure 5.13 Assessing the multiple linear regression (MLR) model fit using SAS macro REGDIAG: explained variation (R^2) plot.

patterns, illustrating that the regression model assumptions are not violated. The differences between the total and residual variability show the amount of variation in the response accounted for by the regression model and are estimated by the R^2 statistic. The estimates of $R^2_{\text{(mean)}}$ and the $R^2_{\text{(prediction)}}$ described previously are also displayed in the title statement. These estimates and the graphical display of explained and unexplained variation help us evaluate the quality of the model fit.

The regression parameter estimates, the standard error, and their p values for testing the parameters = 0 are given in Table 5.7. The p values are derived based on the partial SS (type II). Parameter estimates for all three predictor variables were highly significant. The estimated regression model for predicting the amount of sales ($) is 44270 + 3.80691X1 − 44560X2 + 84.88208X3. The magnitude and sign of the regression parameters provide very useful information relevant to the problem. For example, after controlling for personal disposable income and the advertising ratio, a one-unit ($1000) increase in advertising cost increases product sales, on average, by $3806.91. The information provided in Table 5.8 could be used to find the 95% confidence interval estimates for the parameter. With the 95% level of confidence, we can conclude that on average every $1000 spent on advertising has the potential of increasing the product sales ranging from $3276 to $4337.

Table 5.7 also provides additional statistics such as type I sequential SS, standardized regression coefficient estimates (STB), and type II partial correlations (PCII). Type I SS statistics are influenced by the order of variables entered in the model; therefore, this statistic is useful in testing parameter significance in higher order polynomial models or testing for the significance of parameters in models where the order of entry is important in deciding the significance. The STB and PCII statistics are useful in determining the relative importance of the variables used in the model in terms of accounting for the variation in the response variable. In this example, X3 is the main variable responsible for the variation in the response. VIF estimates larger than 10 usually provide an indication of multicollinearity. Variables X1 and X3 have larger VIF estimates and very large simple correlations, which confirm our observation in the VIF plot that these variables are involved with multicollinearity. However, the impact of multicollinearity on parameter estimates and their significance levels was minimal, as both parameters are highly significant, have the correct signs (positive), and have sensible estimates. The impact of multicollinearity on parameter estimates is usually severe in small samples.

Table 5.9 shows a partial list of observations including the ID, predictor variables used in the model, observed response, predicted response, standard error for predicted values and residuals, and the standardized residual (Student). When the response or predictor variables are missing any values, the entire observation will be excluded from the model fit; however, for any observation, if the response value is missing, but all predictor variables are available, the model estimates the predicted value and reports it in this table, but the residual will be missing for those observations.

5.10.8 Checking for Regression Model Violations

5.10.8.1 Autocorrelation

Figure 5.14 shows the trend plot of the residual over the observation sequence. If the data are time series data, we can examine the residual plot for a cyclic pattern when a sequence of positive residuals follows negative residuals. This cyclical pattern might be due to the presence of first-order autocorrelation where the *i*th residual is correlated with the lag1 residual. The Durbin–Watson (DW) *d* statistic measures the degree of first-order autocorrelation. An estimate of the DW statistic and the significance of first-order autocorrelation are estimated using the PROC REG and are displayed on the title statement. The observed value of the first-order autocorrelation

Table 5.7 Macro REGDIAG: Regression Model Parameter Estimates and Their Significance Levels

Parameter Estimates

Variable	*Label*	*Degrees of Freedom*	*Parameter Estimate*	*Standard Error*	*t Value*	*Pr > |t|*	*Type I Sum of Squares*	*Standardized Estimate*	*Squared Partial Correlative Type II*	*Variance Inflation*
Intercept	Intercept	1	44270	5735.80843	7.72	<.0001	7.076001E13	0	—	0
X1	Advertising ($1000s)	1	3.80691	0.26755	14.23	<.0001	1.748697E13	0.30286	0.64383	27.89067
X2	Advertising ratio[a]	1	–44560	3081.37220	–14.46	<.0001	76680246286	–0.05842	0.65122	1.00485
X3	Personal disposable income ($billion)	1	84.88208	2.59940	32.65	<.0001	3.101658E11	0.69538	0.90495	27.91625

[a] Advertising cost ratio (competing product/own product).

Table 5.8 Macro REGDIAG: Regression Model Parameter Estimates and Their Confidence Interval Estimates

Name of Former Variable	*Label of Former Variable*	*Parameter Estimate*	*Standard Error*	*95% Upper Confidence Interval*	*95% Lower Confidence Interval*
Intercept	Intercept	44269.6137	5735.80843	55634.3829	32904.8446
X1	Advertising ($1000)	3.80691217	0.26755054	4.33702927	3.27679508
X2	Advertising ratio[a]	–44559.685	3081.3722	–38454.341	–50665.029
X3	Personal disposable income ($billion)	84.8820753	2.59939633	90.032446	79.7317047

[a] Advertising cost ratio (competing product/own product).

Table 5.9 Macro REGDIAG: Predicted Values, Residuals, and Standard Errors

ID	*Month*	*X1*	*X2*	*X3*	*Y*	*Predicted Value*	*Standard Error Mean Predicted*	*Residual*	*Standard Error Residual*	*Student Residual*
1	1	4644	1.67	759	51341	51960	4649	–618.734	16409	–0.0377
2	2	16861	0.93	892	143240	142732	3259	507.7364	16741	0.0303
3	3	16734	1.94	1183	120937	121944	3756	–1007	16636	–0.0605
4	4	25616	0.67	1507	234265	239850	3430	–5585	16707	–0.334
5	5	18793	1.58	1394	194812	163734	3190	31078	16754	1.855
—[a]	—	—	—	—	—	—	—	—	—	—
96	96	107665	0.9	9490	1224193	1219568	3258	4625	16741	0.276
97	97	114443	2.15	10022	1219715	1234829	4528	–15114	16443	–0.919

[a] Partial list.

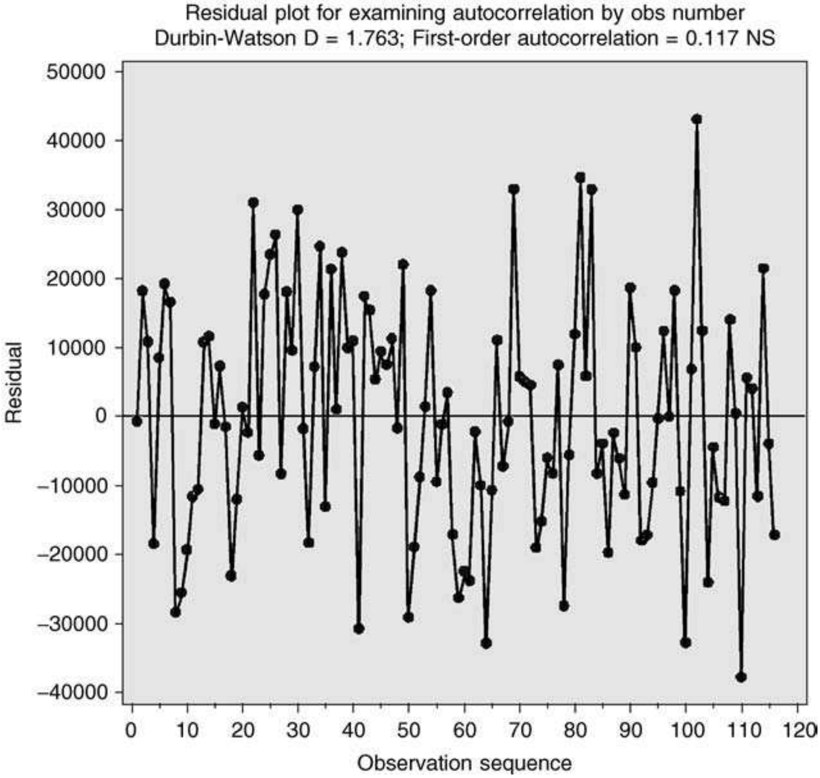


Figure 5.14 Regression diagnostic plot using SAS macro REGDIAG: checking for first-order autocorrelation trend in the residual for "sales".

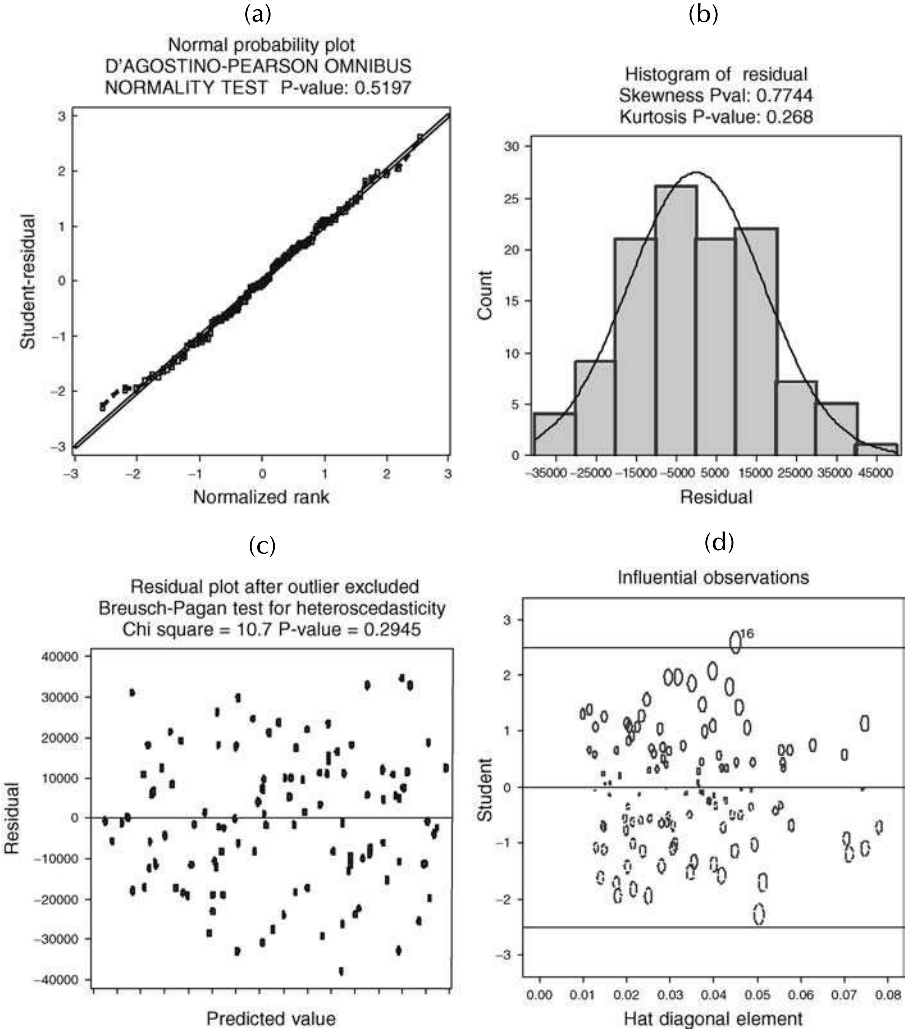
is very small and not significant based on an approximate test (Figure 5.14). Because the dataset used in the case study was simulated, checking for autocorrelation using the DW test is inappropriate.

5.10.8.2 Significant Outlier/Influential Observations

Observations used in the modeling are identified as residual if the absolute Student value exceeds 2.5. Also, observations are identified as influential if the DFFITS statistic value exceeds 1.5. Table 5.10 shows one observation (ID #16) as an outlier (Student value exceeds 2.5). This outlier also showed up in the outlier detection plot (Figure 5.15). Because this outlier has a small DFFITS value and the Student value falls on the borderline, we can conclude that the impact of this outlier on the regression model was minimal.

Table 5.10 Macro REGDIAG: List of Influential Outliers

ID	*Year*	*X1*	*X2*	*X3*	*Y*	*Residual*	*Student*	*DFFITS*	*Outlier*
16	16	20781	1.64	2071	269210	43116.05	2.58691	0.57644	*



(a)

Normal probability plot
D'AGOSTINO-PEARSON OMNIBUS
NORMALITY TEST P-value: 0.5197

(b)

Histogram of residual
Skewness Pval: 0.7744
Kurtosis P-value: 0.268

(c)

Residual plot after outlier excluded
Breusch-Pagan test for heteroscedasticity
Chi square = 10.7 P-value = 0.2945

(d)

Influential observations

Figure 5.15 Regression diagnostic plot using SAS macro REGDIAG: checking for MLR model assumptions. (a) Normal probability plot; (b) histogram of residual; (c) residual plot checking for heteroscedasticity; (d) outlier detection plot for response variable "sales".

5.10.8.3 Checking for Heteroscedasticity in Residuals

The results of the Breusch–Pagan test and the random pattern of the residuals in the residual plot (Figure 5.15) both confirm that the residuals have equal variance.

5.10.8.4 Checking for Normality of the Residuals

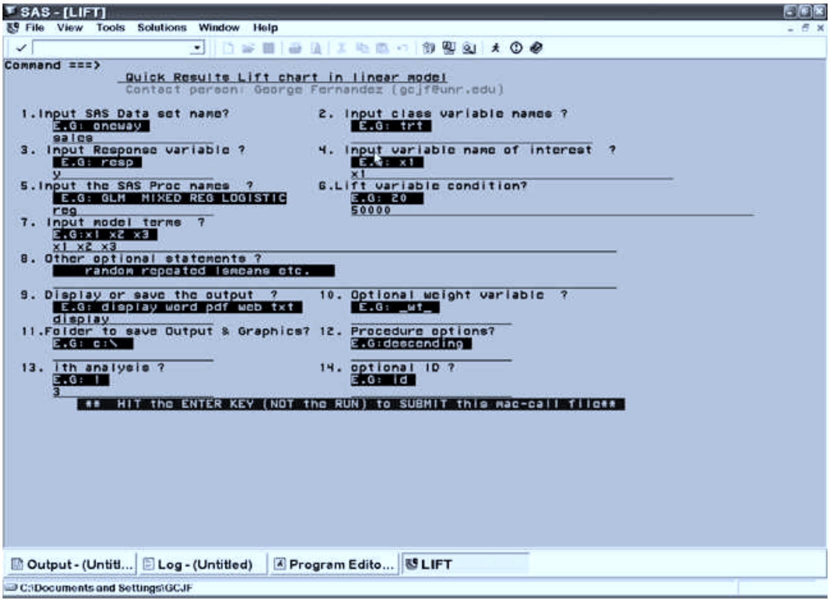
The residuals appeared to have normal distribution based on the *p* values for testing the hypothesis that the skewness and the kurtosis values equal zero and by the d'Agostino–Pearson Omnibus normality test (Table 5.11). This is further confirmed by the symmetrical distribution of the residual histogram and the normally distributed pattern in the normal probability plot (Figure 5.15).

5.10.9 Performing If–Then Analysis and Producing the Lift Chart

The next objective after finalizing the regression model is to perform an if–then analysis and then to create a lift chart showing what happens to the total sales if management spent $50,000 annually on advertising. Open the LIFT.sas macro-call file in the SAS PROGRAM EDITOR window and click RUN to open the LIFT macro-call window (Figure 5.16). Input the SAS dataset (sales) and response (Y) variable names, model statement (X1, X2, X3), variable name of interest (X1), fixed value ($50,000), and other appropriate macro-input values by following the suggestions given in the help file (see Section 5.6.2). Submit the lift macro to get a lift chart (Figure 5.17) showing the differences in the predicted values between the original full model and the new reduced model advertising costs have been kept at

Table 5.11 Macro REGDIAG: Checking for Normality of Error Distributions

Skewness	*p Value for Skewness*	*Kurtosis Statistic*	*p Value for Kurtosis*	*Chi-Square Value for K^2*	*p Value d'Agostino –Pearson Omnibus K^2 Normality Test*
0.061751	0.77439	2.52587	0.26802	1.30899	0.51970


SAS - [LIFT]
File View Tools Solutions Window Help
Output - (Untitl...
Log - (Untitled)
Program Edito...
LIFT
C:\Documents and Settings\GCJF

```
Command ===>
              Quick Results Lift chart in linear model
              Contact person: George Fernandez (gcjf@unr.edu)

 1.Input SAS Data set name?            2. Input class variable names ?
   E.G: oneway                            E.G: trt
   sales
 3. Input Response variable ?          4. Input variable name of interest  ?
   E.G: resp                              E.G: x1
   y                                      x1
 5.Input the SAS Proc names  ?         6.Lift variable condition?
   E.G: GLM  MIXED REG LOGISTIC           E.G: 20
   reg                                    50000
 7. Input model terms  ?
   E.G:x1 x2 x3
   x1 x2 x3
 8. Other optional statements ?
       random repeated lsmeans etc.

 9. Display or save the output  ?      10. Optional weight variable  ?
   E.G: display word pdf web txt          E.G: _wt_
   display
11.Folder to save Output & Graphics?   12. Procedure options?
   E.G: c:\                               E.G:descending

13. ith analysis ?                     14. optional ID ?
   E.G: 1                                 E.G: id
   3
     ** HIT the ENTER KEY (NOT the RUN) to SUBMIT this mac-call file**
```

Figure 5.16 Screen copy of LIFT macro-call window showing the macro-call parameters required for generating a lift chart in linear regression.

$50,000/year continuously. On the lift chart display, the sum of total differences ($8,289,616) between the predicted values of the original model and the predicted value for the reduced model is also displayed. In addition to the lift chart, the if–then analysis outputs a table showing the predicted values of the original, the reduced models, and their differences. A partial list of the if–then analysis output is presented in Table 5.12.

5.10.10 Predicting the Response Scores for a New Dataset

After finalizing the regression model and saving the regression parameter estimates in the SAS dataset "regest", new expected responses can be predicted to any new SAS data containing all the required predictor variables.

Open the RSCORE.sas macro-call file in the SAS PROGRAM EDITOR window and click RUN to open the RSCORE macro-call window (Figure 5.18). Input the name of the new SAS dataset (new), regression parameter estimate dataset (regest), response variable name (optional), model statement (X1 X2 X3), and other appropriate macro-input values by following the suggestions given in the help file (see Section 5.7.2). Submit the RSCORE macro to get an output table showing the predicted scores for the new dataset. If the new data include observed response values, then

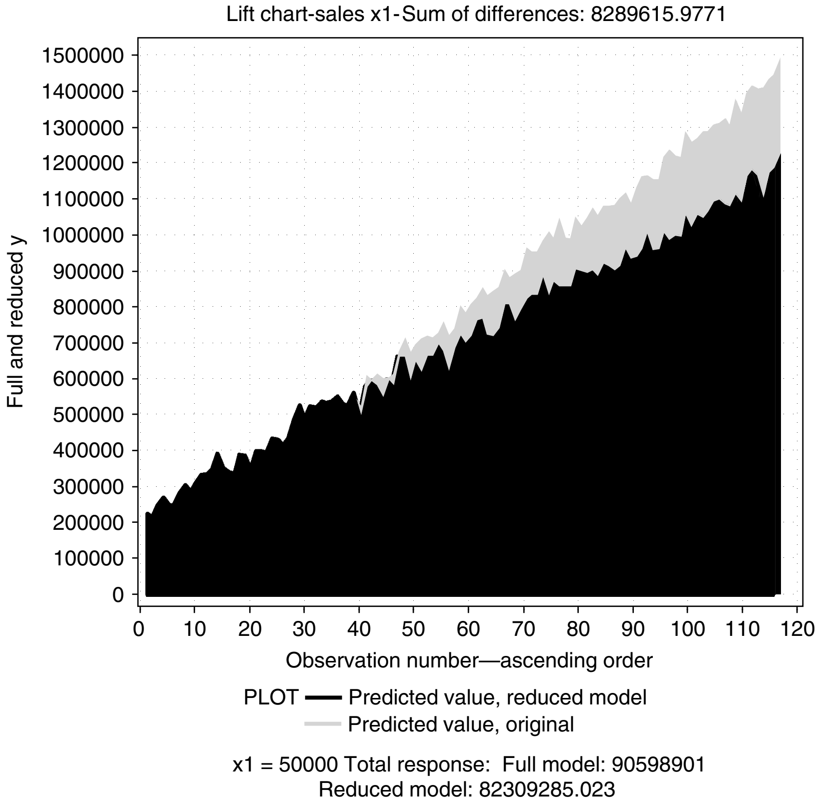


Figure 5.17 Lift chart generated by using SAS macro LIFT: differences in predicted values between the original and the reduced model with fixed level of advertising expenses.

the RSCORE macro also performs model validation and produces additional plots showing the success of the validation.

The predicted sales amounts for a new input dataset containing varying levels of advertising (X1), advertising ratio (X2), and fixed level of personal disposable income (X3) are shown in Table 5.13.

5.11 Case Study 2: Modeling Multiple Linear Regression with Categorical Variables

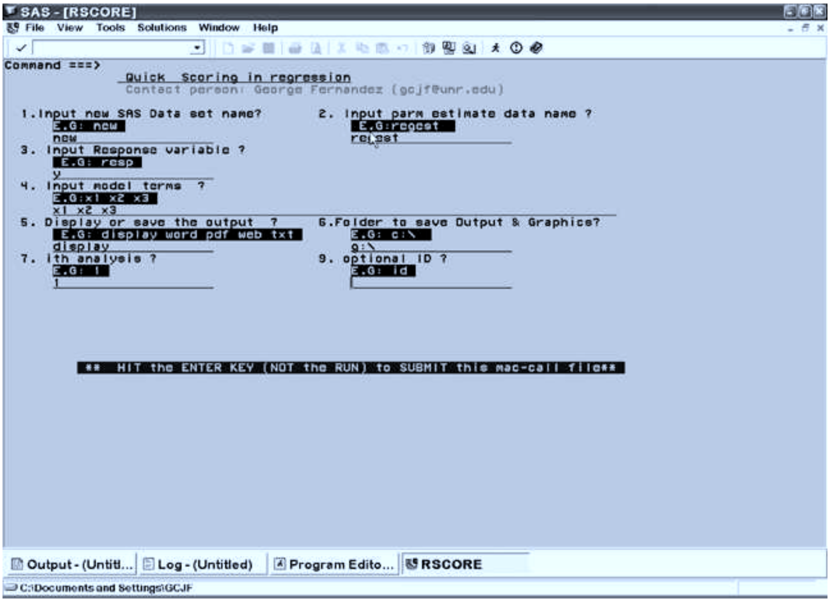
5.11.1 Study Objectives

1. **Data exploration using diagnostic plots:** For each predictor variable, check for linear and nonlinear effects and interactions between categorical and predictors in the scatterplots. Also, check

Table 5.12 Macro REGDIAG: Output From If–Then Analysis

ID	*Predicted Value, Original*	*X1*	*Predicted Value, Reduced Model*	*Fixed X1 Value*	*Differences (Original-Reduced) Model*
35	529793.22	48210	536607.59	50000	–6814.37
36	547198.40	48797	551778.12	50000	–4579.72
42	540740.48	52818	530012.61	50000	10727.88
38	526083.16	49933	526338.22	50000	–255.06
39	548175.69	46138	562877.98	50000	–14702.29
41	520534.15	53520	507133.82	50000	13400.33
—[a]	—	—	—	—	—
111	1433029.57	117895	1174559.27	50000	258470.30
113	1444985.02	117501	1188014.64	50000	256970.38
116	1484121.27	118161	1224638.33	50000	259482.94

[a] Partial list.



```
Command ===>
              Quick  Scoring in regression
              Contact person: George Fernandez (gcjf@unr.edu)

 1.Input new SAS Data set name?        2. Input parm estimate data name ?
   E.G: new                               E.G:regest
   new                                    regest
 3. Input Response variable ?
   E.G: resp
   y
 4. Input model terms  ?
   E.G:x1 x2 x3
   x1 x2 x3
 5. Display or save the output  ?      6.Folder to save Output & Graphics?
   E.G: display word pdf web txt          E.G: c:\
   display                                g:\
 7. ith analysis ?                     9. optional ID ?
   E.G: 1                                 E.G: id
   1


     **  HIT the ENTER KEY (NOT the RUN) to SUBMIT this mac-call file**
```

Figure 5.18 Screen copy of RSCORE macro-call window showing the macro-call parameters required for estimating predicted scores from an independent dataset.

Table 5.13 Macro RSCORE: Scoring New Dataset

Advertising ($1000)	*Advertising Ratio*[a]	*Personal Disposable Income*	*Predicted Score*
118161	0.3	11821	1484121.27
119000	0.3	11820	1487230.39
117500	0.3	11820	1481520.02
119000	1	11820	1456038.61
117500	1	11820	1450328.24
119000	1.3	11820	1442670.70
117500	1.3	11820	1436960.33

[a] Advertising cost ratio (competing product/own product).

the data for very extreme influential observations and, if desirable, fit the regression model after excluding these extreme cases.

2. **Regression model fitting and prediction:** Fit the regression model, perform hypothesis testing on overall regression and on each parameter estimate, estimate confidence intervals for parameter estimates, predict scores, and estimate their confidence intervals.
3. **Checking for any violations of regression assumptions:** Perform statistical tests and graphical analysis to detect influential outliers, heteroscedasticity in residuals, and departure from normality.
4. **Save "_score_" and "regest" datasets for future use:** These two datasets are created and saved as temporary SAS datasets in the work folder. The "_score_" dataset contains observed responses, predicted scores (including those for observations with missing response values), residuals, and confidence interval estimates for the predicted values. This dataset could be used as the base for developing scorecards for each observation. Also, in SAS version 8.0 and later, the parameter estimates dataset called "regest" created by the SAS ODS output option can be used (after slight modification) in the RSCORE macro for scoring different datasets containing the same variables.
5. **If–then analysis and lift charts:** Perform an if–then analysis and construct a lift chart to estimate the differences in the predicted response when one of the continuous or binary predictor variables is fixed at a given value.

5.11.2 Data Descriptions

5.11.2.1 Background Information

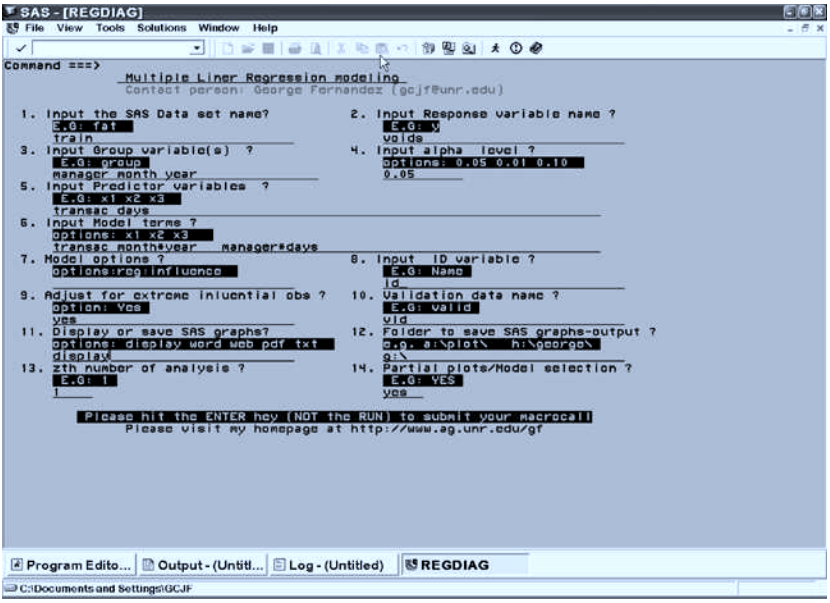
A small convenience store owner suspects that one of the store managers (Mr. X) has stolen a small amount of money ($50 to $150 per day) from the cash register by registering false voids over a 3-year period. The owner caught Mr. X twice in the act of making fraudulent voids. To prove his claim in court, the owner hired a data mining specialist to investigate the 3-year daily sales/transactions records. The average amount of sales for the store varied from $1000 to $2000 per day. A total void amount greater than $400 for any given day was treated as a genuine error. Descriptions of the data and the data mining methods used are presented in the following table:

Data name (3-year daily sales/transactions data)	Permanent SAS dataset "fraud" located in the library "gf". Two temporary datasets, TRAIN and VALID, are randomly selected using the RANSPLT macro.
Response variables/predictor variables	voids: total dollar amount of voids/day (response variable) transac: Total number of daily transactions days: *i*th day of business during the 3-year period year: 1998, 1999, 2000 month: 12 months, from January to December manager: 0, all other managers; 1, Mr. X (indicator variable)
Number of observations	Total dataset: 896 Training: 598 Validation: 298
Source	Actual daily sales data from a small convenience store

Open the REGDIAG.sas macro-call file in the SAS PROGRAM EDITOR window and click RUN to open the REGDIAG macro-call window (Figure 5.19). Use the suggestions provided in the help file (see Section 5.5.2) to input the appropriate macro input values and fit the regression models with indicator variables. Only selected output and graphics produced by the REGDIAG macro are shown here.

5.11.2.2 Exploratory Analysis/Diagnostic Plots

Input dataset name (TRAIN), response variable (voids), group variables (manager), continuous predictor variables (transac and days), model terms (transac, days, manager, and manager*days), and other appropriate macro-input values. To perform data exploration and to create regression diagnostic plots input YES in macro input field #14. Submit the REGDIAG macro to obtain the following regression diagnostic plots:



```
Command ===>
                 Multiple Liner Regression modeling
                 Contact person: George Fernandez (gcjf@unr.edu)

 1. Input the SAS Data set name?          2. Input Response variable name ?
    E.G: fat                                 E.G: y
    train                                    voids
 3. Input Group variable(s) ?             4. Input alpha  level ?
    E.G: group                               options: 0.05 0.01 0.10
    manager month year                       0.05
 5. Input Predictor variables ?
    E.G: x1 x2 x3
    transac days
 6. Input Model terms ?
    options: x1 x2 x3
    transac month*year   manager*days
 7. Model options ?                       8. Input  ID variable ?
    options:reg:influence                    E.G: Name
                                             id
 9. Adjust for extreme inluential obs ?  10. Validation data name ?
    option: Yes                              E.G: valid
    yes                                      vid
11. Display or save SAS graphs?          12. Folder to save SAS graphs-output ?
    options: display word web pdf txt        e.g. a:\plot\   h:\george\
    display                                  g:\
13. zth number of analysis ?             14. Partial plots/Model selection ?
    E.G: 1                                   E.G: YES
    1                                        yes

      Please hit the ENTER key (NOT the RUN) to submit your macrocall
           Please visit my homepage at http://www.ag.unr.edu/gf
```

Figure 5.19 Screen copy of REGDIAG macro-call window showing the macro-call parameters required for performing multiple linear regression (MLR) with categorical variables.

- Simple scatterplots by manager for each predictor variable
- Box plot of voids by manager
- Partial interaction plots for continuous predictors

The relationships between transac and voids by manager are presented in Figure 5.20. The amount of voids/day and the total number of transactions varied from 0 to 350 and 0 to 250, respectively. Approximately 3% of the void amount falls outside $200/day, and a majority of these extreme observations are associated with Mr. X (manager = 1). A significant positive linear trend was observed between the voids and transactions for both manager groups, indicating that larger voids are associated with more transactions. Overall, Mr. X's average void amount is larger than the other managers' (manage = 0) void amount.

A positive linear trend is observed between voids and the *i*th days of transaction for Mr. X and no trend is observed for the other manager group (Figure 5.21). This differential trend indicates the presence of significant interaction for the manager × days term. Figure 5.22 shows the box-plot display of overall variation for voids by two-manager group. The median and third-quartile void amounts for Mr. X are higher than

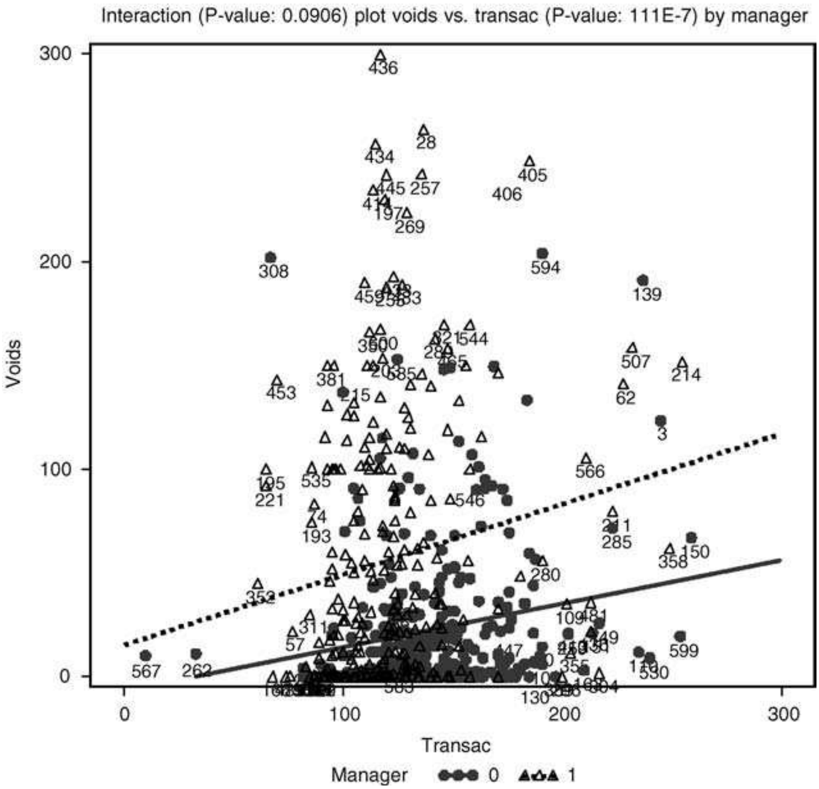


Figure 5.20 Regression diagnostic plot using SAS macro REGDIAG: regression plot between "voids" and "transaction" by "manager".

for the other manager group. The void dollar amount that is approximately greater than $75 is identified as an outlier for the other manager group, whereas a void dollar amount greater than $225 is identified as an outlier for Mr. X. The interaction effect between the transac × days on voids after accounting for all other predictors is displayed in the partial interaction plot (Figure 5.23). The interaction effect is not significant at the 5% level.

These regression diagnostic plots clearly reveal that a differential trend could be observed for the two manager groups, supporting the convenience store manager's hypothesis. This finding could be further confirmed by fitting a multiple regression model and treating the manager as the categorical variable after excluding the extreme outliers. The results of the regression analysis are presented next.

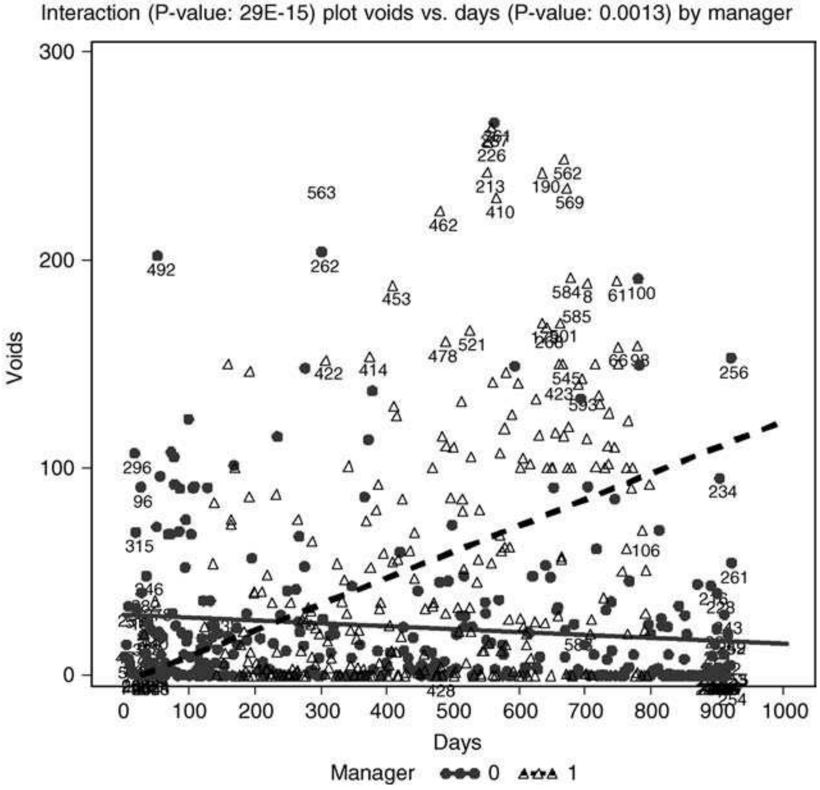


Figure 5.21 Regression diagnostic plot using SAS macro REGDIAG: regression plot between "voids" and "days" by "manager".

5.11.2.3 Fitting Regression Model and Validation

Input dataset name (TRAIN), validation dataset (VALID) response variable (voids), group variables (manager), continuous predictor variables (transac and days) model terms (transac, days, manager, manager*days), and other appropriate macro-input values in the macro-call window REGDIAG. To skip outputting regression diagnostic plots, leave macro input field #14 blank. To exclude extreme outliers and influential observations, input YES in macro input field #15. Submit the REGDIAG macro to exclude outliers, fit regression with indicator variables, estimate predicted scores, check the regression model assumptions, and validate the model using the independent VALID dataset.

Extreme outliers (Student, >4) and influential observations (DFFITS, >1.5) are identified (Table 5.14) and excluded from the regression model

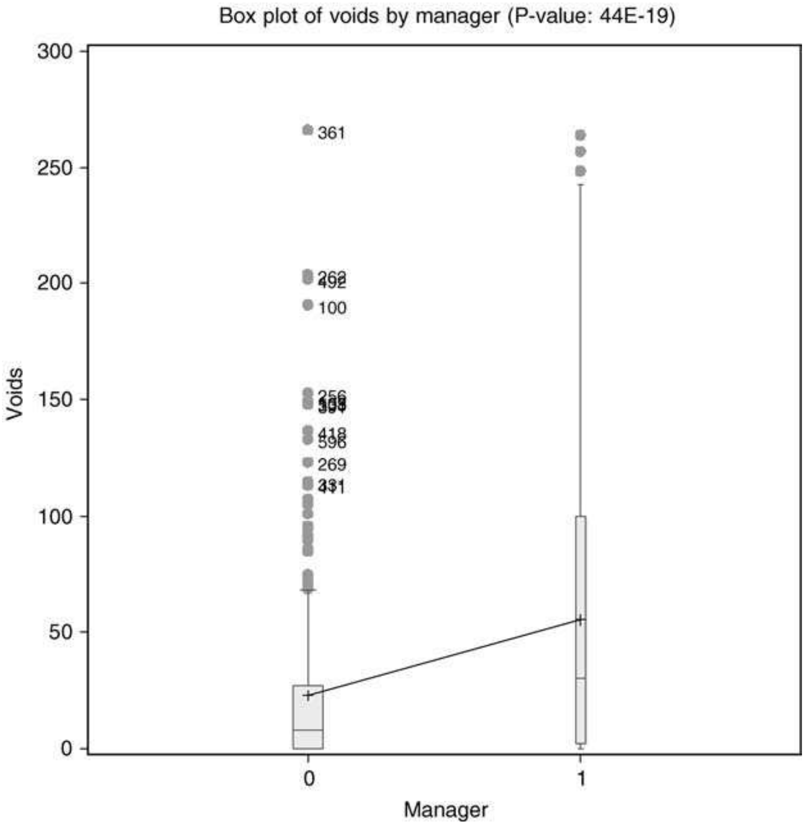


Figure 5.22 Regression diagnostic plot using SAS macro REGDIAG: box plot of "voids" by "manager".

fitting. The number and descriptions of categorical variable levels fitted in the model are presented in Table 5.15. Examine the class level information presented in this table to verify that all categorical variables used in the model are coded correctly. The overall model fit is highly significant (*p* value, <0.0001; Table 5.16). Verify that the total and the model degrees of freedom are correct for the training dataset. The R^2 estimate indicates that about 23% of the variation in the voids could be attributed to the specified model (Table 5.17). The RMSE value reported in Table 5.17 is the smallest estimate among the many regression models tried, indicating that this model is the best.

Figure 5.24 illustrates the total and unexplained variation in voids after accounting for the regression model. The ordered and centered response variable vs. the ordered observation sequence displays the total variability in the voids. This total variation plot shows a right-skewed trend with

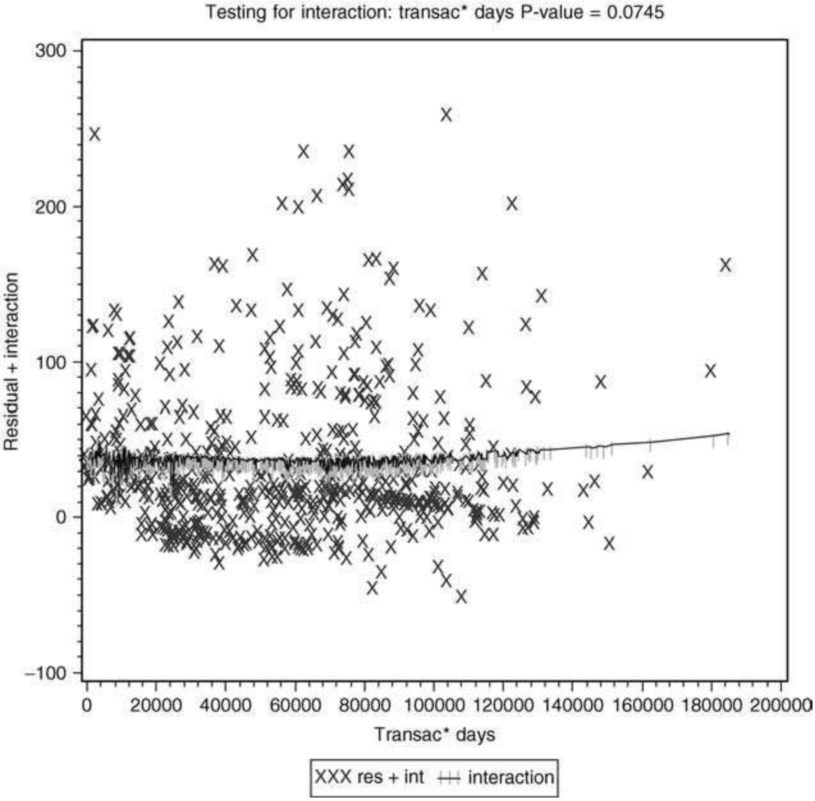


Figure 5.23 Regression diagnostic plot using SAS macro REGDIAG: testing for significant interaction between "transaction" and "days".

Table 5.14 Macro REGDIAG: List of Influential Outliers Excluded

ID	*Manager*	*Transaction*	*Days*	*DFFITS*	*R Student*
146	1	125	150	0.61611	5.00296
290	0	173	295	0.35443	4.77228
543	0	186	563	0.36400	4.07533
633	1	104	656	0.38209	4.24605
661	1	150	685	0.47330	4.59364
668	1	174	692	0.50952	4.20830
758	0	156	788	0.41122	4.60275

Table 5.15 Macro REGDIAG: Class Level Information

Class	*Levels*	*Values*
Manager	2	0, 1

Table 5.16 Macro REGDIAG: Overall Model Fit

Source	*Degrees of Freedom*	*Sum of Squares (SS)*	*Mean Square*	*F Value*	*Pr > F*
Model	4	380478.011	95119.503	44.01	<.0001
Error	586	1266590.125	2161.417	—	—
Corrected total	590	1647068.137	—	—	—

Table 5.17 Macro REGDIAG: Model Fit Statistics

R^2	*Coefficient Variation (CV)*	*Root-Mean-Square Error (RMSE)*	*Voids Mean*
0.231003	124.6285	46.49104	37.30369

very sharp edges at the right end. The unexplained variability in voids is illustrated by the residual distribution. The residual variation shows an unequal variance pattern because the residuals in the higher end are relatively larger than the residual in the lower end, thus illustrating that the equal variance and normally distributed error assumptions are violated. The differences between the total and residual variability shows that only a small portion of the variation in the voids was accounted for by the regression model. The predictive ability of the regression model is given by the estimate for $R^2_{\text{(prediction)}}$ and is displayed in the title statement. A very small difference (4%) between the model R^2 *and* $R^2_{\text{(prediction)}}$ indicates that the impact of the outliers on model prediction is minimal.

All terms included in the final regression models are highly significant based on the partial SS (type III) (Table 5.18). The observed positive trend in voids is significant for transac. The significant interaction for days*manager confirms the differential slope estimates for days by manager. The regression model parameters, their standard errors, and the significance levels produced by the solution option in PROC GLM are presented in

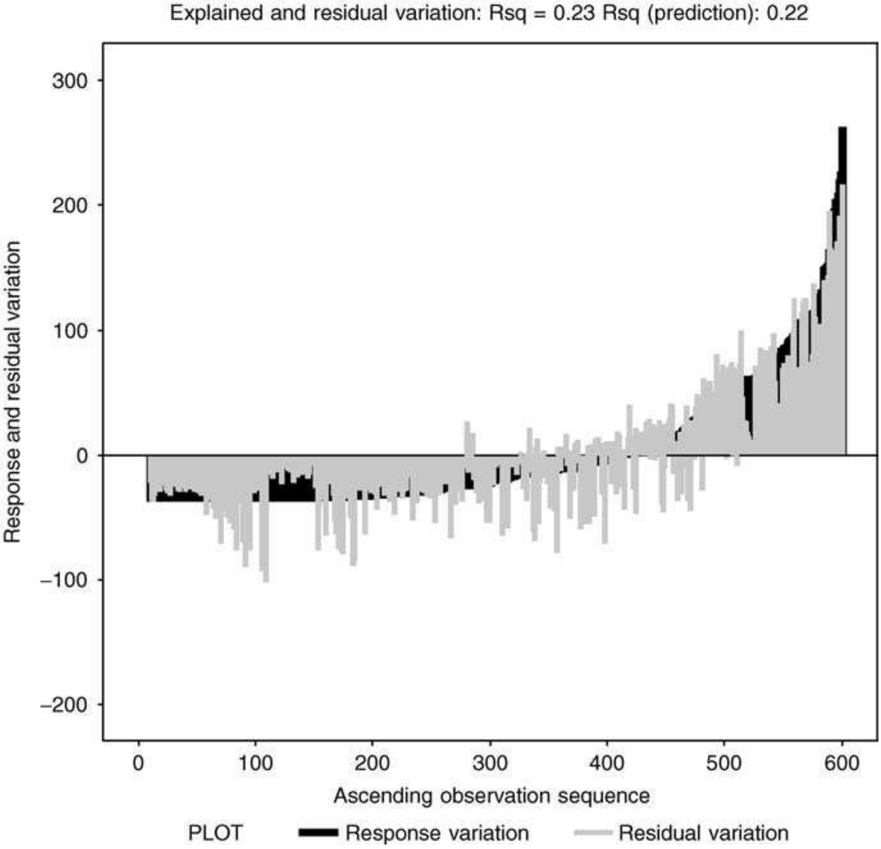


Figure 5.24 Assessing the MLR model fit using SAS macro REGDIAG: plot showing the amount of explained variation (R^2) in "voids".

Table 5.18 Macro REGDIAG: Significance of Model Effects

Source	*Degrees of Freedom*	*Type III Sum of Squares (SS)*	*Mean Square*	*F Value*	*Pr > F*
transac	1	42456.7671	42456.7671	19.64	<.0001
days	1	113345.3946	113345.3946	52.44	<.0001
manager	1	19635.1425	19635.1425	9.08	0.0027
days* manager	1	145128.0467	145128.0467	67.14	<.0001

Table 5.19 Macro REGDIAG: Parameter Estimates and Their Significance

Parameter	*Estimate*	*Standard Error*	*t Value*	*Pr > |t|*
Intercept	–39.74797271 B[a]	10.97681793	–3.62	0.0003
transac	0.27548794	0.06215819	4.43	<.0001
days	0.13356213 B	0.01513612	8.82	<.0001
manager 0	27.13630425 B	9.00332467	3.01	0.0027
manager 1	0.00000000 B	—	—	—
days*manager 0	–0.14130982 B	0.01724511	–8.19	<.0001
days*manager 1	0.00000000 B	—	—	—

Derived Regression Model:

Manager 1: –39.74797271 + 0.27548794 transac + 0.13356213 days

Manager 0: (–39.74797271 + 27.13630425) + 0.27548794 transac + (0.13356213 – 0.14130982) days

[a] B – biased

Table 5.19. These regression estimates could be used to construct the regression equation for both managers as outlined in Table 5.19. The estimated regression model confirms that after adjusting for variations in the number of transactions, the rate of increase in daily void amount for Mr. X is 0.133, whereas for other managers the rate was almost zero (–0.01). These results clearly show that there is a very small chance (<0.0001) that the observed differences in the daily voids between Mr. X and other managers are due just to chance. These regression models could be used to predict the expected void amount for both manager groups.

The estimated predicted scores and their confidence and prediction intervals are given in Tables 5.20 and 5.21, respectively. These predicted scores and the confidence interval estimates could be used to build scorecards for observations in the dataset.

5.11.2.4 Checking for Regression Model Violations

- **Autocorrelation:** Figure 5.25 shows the trend plot of the residual over the observation sequence. No cyclical pattern is evident in the residual plot. Estimates of the Durbin-Watson statistic and the first-order autocorrelation displayed on the title statement indicate that the influence of autocorrelation is small; the approximate significant test shows that the autocorrelation is not significant. This significant test for autocorrelation is highly sensitive when the sample size is large. Because no cyclical pattern is observed in the

Table 5.20 Macro REGDIAG: Partial List of Predicted Scores

ID	*Transactions*	*Days*	*Manager*	*Month*	Year	*Voids*	*Predicted*	*Residual*
2	134	3	0	January	1998	9.00	29.703	–20.703
105	126	108	0	April	1998	91.00	23.498	67.502
235	125	239	0	September	1998	29.10	–6.742	35.842
62	191	64	0	March	1998	6.00	54.754	–48.754
771	115	801	0	April	2000	20.00	0.744	19.256
236	118	240	0	September	1998	0.00	–8.851	8.851
—[a]	—	—	—	—	—	—	—	—
414	137	426	0	March	1999	12.00	18.367	–6.367
346	93	355	1	January	1999	4.00	39.486	–35.486
234	126	238	1	September	1998	0.00	14.723	–14.723
701	129	727	0	February	2000	0.50	30.682	–30.182

[a] Partial list.

Table 5.21 Macro REGDIAG: Partial List of Confidence Interval (CI) and Prediction Interval (PI)

	95% Confidence Interval		*95% Prediction Interval*	
ID	*Lower*	*Upper*	*Lower*	*Upper*
2	6.442	52.963	–62.787	122.193
105	3.253	43.743	–68.280	115.276
235	–27.936	14.453	–98.734	85.251
62	32.387	77.122	–37.515	147.024
771	–23.534	25.022	–92.007	93.495

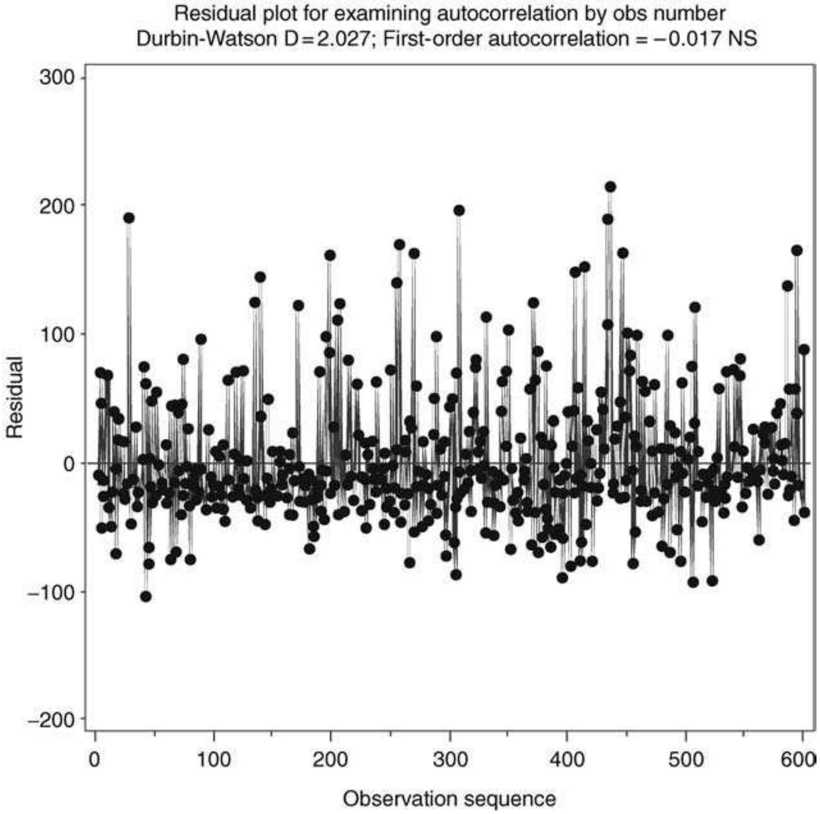


Figure 5.25 Regression diagnostic plot using SAS macro REGDIAG: checking for first-order autocorrelation trend in the residual for "voids".

residual plot and the autocorrelation estimate is very small, we can conclude that the autocorrelation is not significant.

- **Significant outlier/influential observations:** Table 5.22 lists several observations as outliers (the Student value exceeds 2.5). These outliers also showed up in the outlier detection plot (Figure 5.26). Because all of these outliers have a small DFFITS value, we can conclude that the impact of this outlier on the regression model estimate was minimum.
- **Checking for heteroscedasticity in residuals:** The results of the Breusch–Pagan test and the fan-shaped pattern of the residuals in the residual plot (Figure 5.26) both confirm that the residuals have unequal variance. Remedial measures, such as Box–Cox transformation or heterogeneity regression estimates, are recommended to

Table 5.22 Macro REGDIAG: List of Outliers

ID	*Trans-actions*	*Days*	*Voids*	*Residual*	*Student*	*DFFITS*	*Outlier*
230	118	234	115.00	112.904	2.54910	0.13450	*
645	185	669	248.50	113.665	2.57263	0.31528	*
649	114	673	234.50	120.374	2.71158	0.24465	*
505	112	525	166.13	122.467	2.77639	0.12028	*
888	125	922	153.00	128.344	2.89300	0.24059	*
368	100	378	137.00	128.182	2.90668	0.17994	*
271	146	276	148.00	132.092	2.96556	0.13801	*
574	149	594	149.00	131.630	2.98418	0.13726	*
753	169	783	149.50	131.940	2.99926	0.19670	*
397	120	409	188.00	137.430	3.10464	0.16129	**
614	120	636	242.00	137.164	3.12403	0.23700	**
546	119	566	230.00	140.080	3.14918	0.19772	**
532	136	552	242.25	147.835	3.32934	0.21116	**
751	237	781	190.99	153.354	3.50910	0.39615	***
296	191	301	204.00	163.933	3.73538	0.25869	***
464	129	481	223.50	167.550	3.76830	0.18117	***
533	115	553	257.00	168.717	3.80004	0.23183	***
539	137	559	264.00	169.043	3.80117	0.24216	***
664	117	688	300.00	184.460	4.16745	0.36280	***
50	67	52	202.00	197.924	4.48596	0.53484	***

* Outlier.
** Highly significant.
*** Very highly significant outlier.

combat heteroscedasticity problems associated with the hypothesis testing and interval estimates.

- **Checking for normality of the residuals:** The residuals appear to have a strong right-skewed distribution based on the *p* values for testing the hypothesis that the skewness and the kurtosis values equal zero and by the d'Agostino–Pearson omnibus normality test (Figure 5.26). This is further confirmed by the right-skewed distribution of the residual histogram and curved upward trend in the normal probability plot (Figure 5.26). The presence of extreme outliers and heteroscedasticity could be the main cause for the right-skewed error distribution, and the remedial measures suggested to combat outliers and heteroscedasticity in this chapter could reduce the impact of right-skewed residuals.

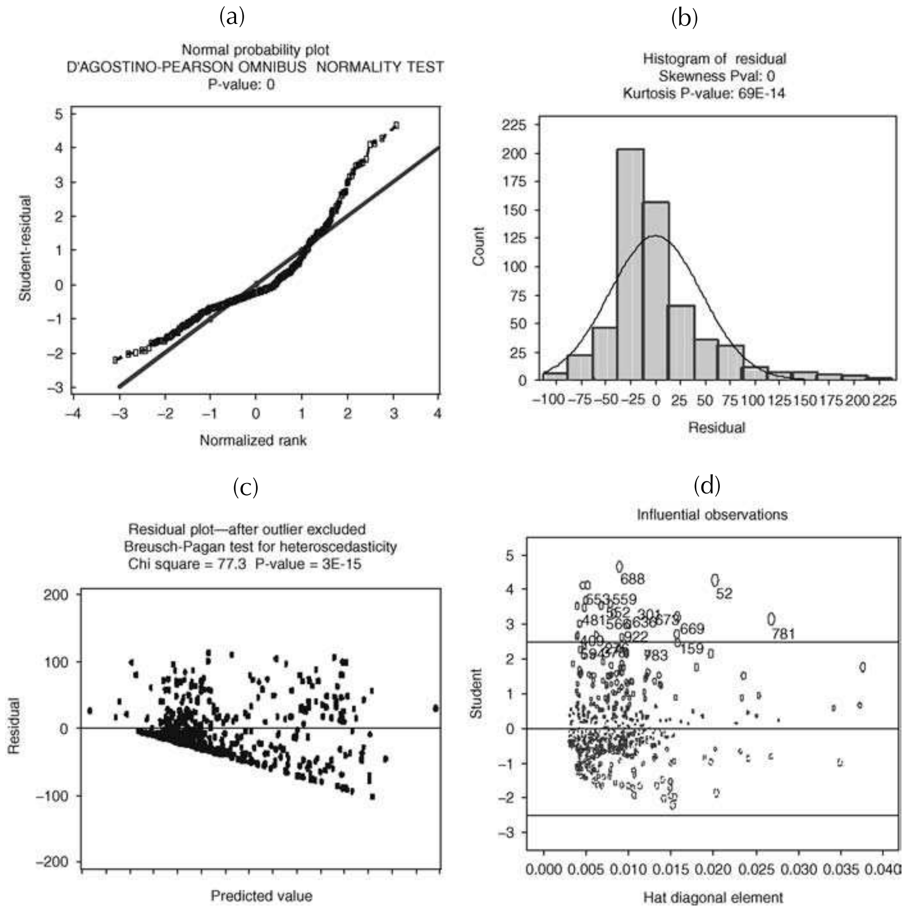


Figure 5.26 Regression diagnostic plot using SAS macro REGDIAG: checking for MLR model assumptions. (a) Normal probability plot; (b) histogram of residual; (c) residual plot checking for heteroscedasticity; (d) outlier detection plot for response variable "voids".

5.11.2.5 Model Validation

To validate the obtained regression model, the regression parameter estimates obtained from the training dataset (n = 591) are used to estimate the predicted voids for the validation dataset (N = 291). Both training and validation regression models and the R^2 values appear to be similar (Figure 5.27). The residuals from the training and validation datasets show a similar distribution pattern in the residual plot (Figure 5.28). Thus, we could conclude that the regression model obtained from the training data provides valid estimates to predict the void amount in this investigation.

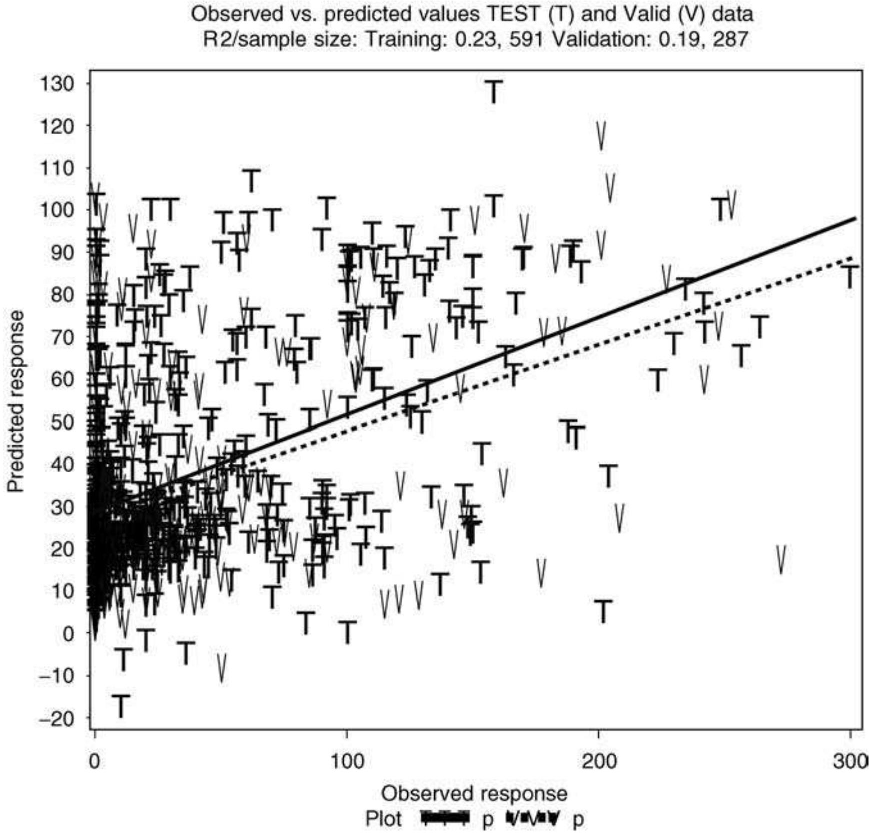


Figure 5.27 Regression model validation using SAS macro REGDIAG: comparing fitted lines between the training and the validation data for "voids".

5.11.2.6 *Performing If–Then Analysis and Producing the LIFT Chart*

The next objective after finalizing the regression model is to perform an if–then analysis and then create a lift chart showing what happens to the total void amount when Mr. X is replaced by any other manager during the 3-year period. Open the LIFT.sas macro-call file in the SAS PROGRAM EDITOR window and click RUN to open the LIFT macro-call window (Figure 5.29). Input the SAS dataset name (fraud) and response variable name (voids), model statement (transac days manager manager*days), variable name of interest (manager), fixed level (manager = others (0)), and any other appropriate macro-input values by following the suggestions given in the help file (see Section 5.6.2). Submit the LIFT macro to obtain a lift chart (Figure 5.30) showing the

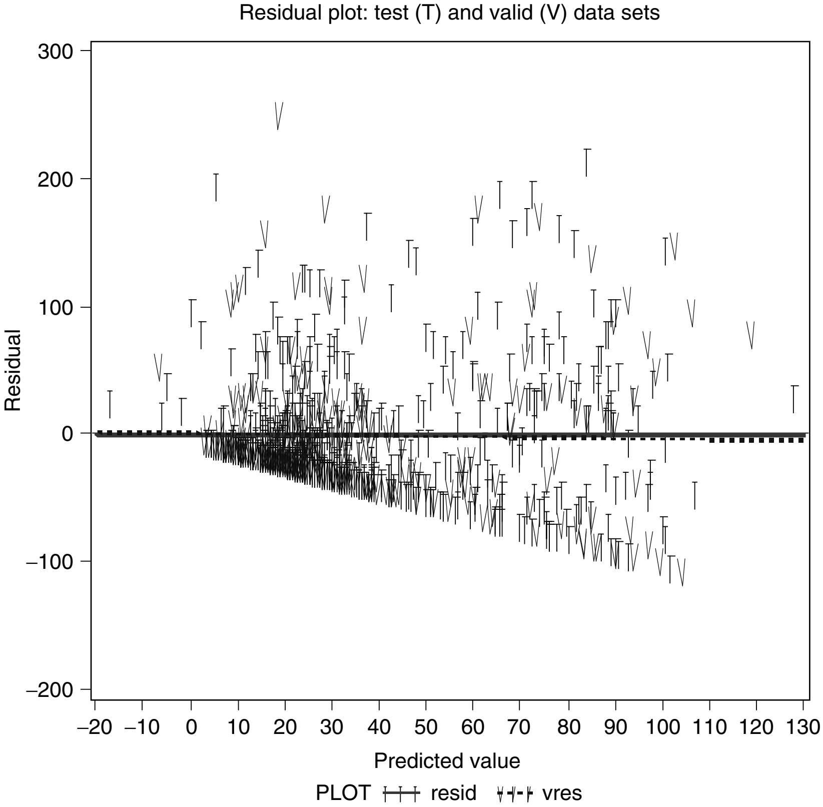
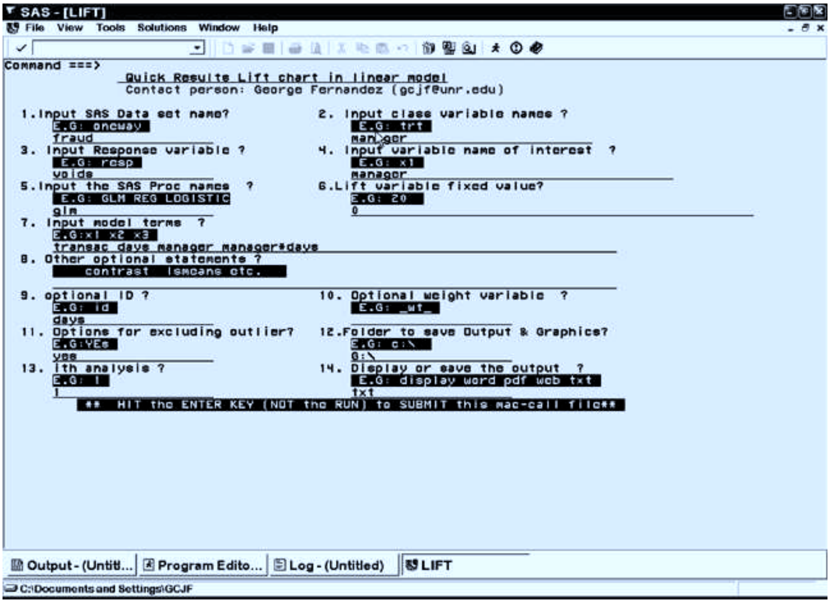


Figure 5.28 Regression model validation using SAS macro REGDIAG: comparing distribution of residuals between the training and the validation data for "voids".

differences in the predicted values between the original full model and the new reduced model, where the manager variable was fixed at 0, meaning other managers. On the lift chart display, the sum of differences ($16,543) between the predicted values of the original model and the predicted value for the reduced model is also displayed. In addition to the lift chart, the if–then analysis output table shows, the predicted values of the original and reduced models, and the differences in the predicted values. A partial list of the if–then analysis output is presented in Table 5.23. We can conclude that this convenience store lost about $16,543 due to fraudulent voids during the 3-year period.



SAS - [LIFT]
File View Tools Solutions Window Help
Output - (Untitl... Program Edito... Log - (Untitled) LIFT
C:\Documents and Settings\GCJF

```
Command ===>
              Quick Results Lift chart in linear model
              Contact person: George Fernandez (gcjf@unr.edu)

 1.Input SAS Data set name?           2. Input class variable names ?
   E.G: oneway                           E.G: trt
   fraud                                 manager
 3. Input Response variable ?         4. Input variable name of interest  ?
   E.G: resp                             E.G: x1
   voids                                 manager
 5.Input the SAS Proc names  ?        6.Lift variable fixed value?
   E.G: GLM REG LOGISTIC                 E.G: 20
   glm                                   0
 7. Input model terms  ?
   E.G:x1 x2 x3
   transac days manager manager*days
 8. Other optional statements ?
      contrast  lsmeans etc.

 9. optional ID ?                     10. Optional weight variable  ?
   E.G: id                               E.G: _wt_
   days
 11. Options for excluding outlier?   12.Folder to save Output & Graphics?
   E.G:YEs                               E.G: c:\
   yes                                   G:\
 13. ith analysis ?                   14. Display or save the output  ?
   E.G: 1                                E.G: display word pdf web txt
   1                                     txt
     ** HIT the ENTER KEY (NOT the RUN) to SUBMIT this mac-call file**
```

Figure 5.29 Screen copy of LIFT macro-call window showing the macro-call parameters required for performing lift chart in linear regression with categorical variables.

5.12 Case Study 3: Modeling Binary Logistic Regression

5.12.1 Study Objectives

1. **Data exploration using diagnostic plots:** Check for the significance of each continuous predictor variable and multicollinearity among the predictors by studying simple logit and delta logit trends. Perform variable selection using the Forward selection. Also, check the data for very high influential observations and, if desirable, fit the regression model after excluding these extreme cases.
2. **Logistic regression model fitting and prediction:** Fit the logistic regression model**,** perform hypothesis testing on overall regression and on each parameter estimate, estimate confidence intervals for parameter estimates and odds ratios, predict probability scores and estimate their confidence intervals, and produce classification tables and false-positive and false-negative estimates.

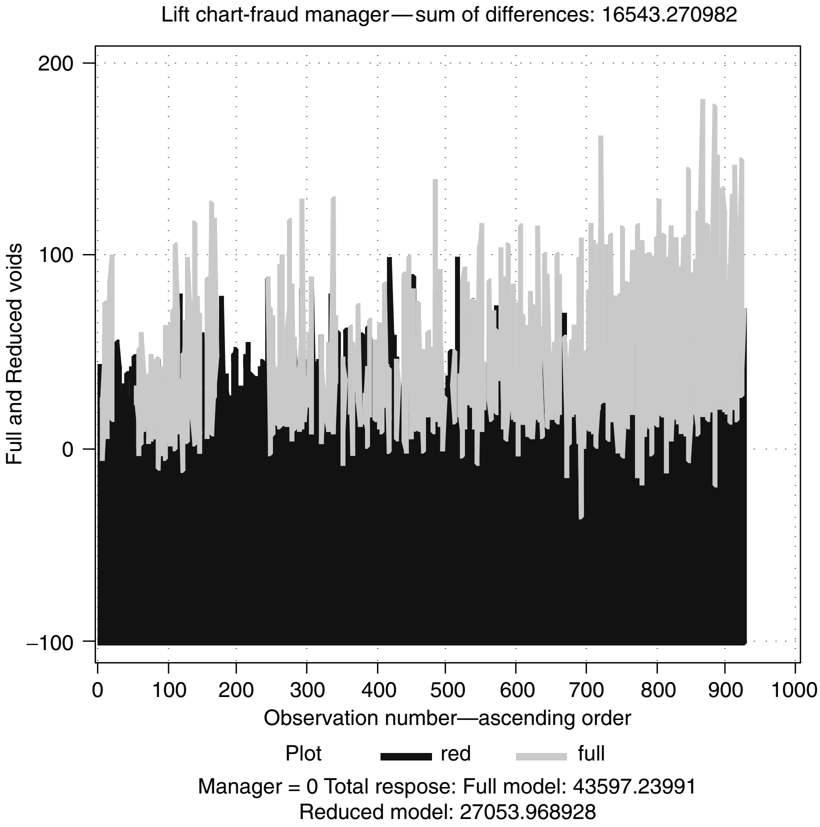


Figure 5.30 Lift chart generated by using SAS macro LIFT: differences in predicted "voids" between the original (full) and the reduced (red) model with the same level of "manager".

3. **Checking for any violations of logistic regression assumptions:** Perform statistical tests to detect overdispersion and graphical analysis to detect influential outliers.
4. **Computing goodness-fit tests and measures for evaluating the fit:** Obtain Hosmer and Lemeshaw goodness fit statistics and model adequacy estimates, generalized and adjusted generalized R^2, Brier scores, and ROC curves.
5. **Save "_score_" and "estim" datasets for future use:** These two datasets are created and saved as temporary *SAS* datasets in the work folder. The "_score_" dataset contains the observed variables, predicted probability scores, and confidence interval estimates. This

Table 5.23 Macro LIFT: Partial List of Differences Between Original and Reduced (If–Then) Model Predicted Scores

Manager	*ID*	*Predicted, Original*	*Predicted, Reduced*	*If–Then*	*Difference in Predicted Values*
0	393	51.749	51.749	0	0.0000
1	398	71.673	39.387	0	32.2860
1	401	50.975	18.280	0	32.6947
1	402	33.705	0.874	0	32.8309
0	412	53.306	53.306	0	0.0000
1	416	51.398	16.660	0	34.7382
1	418	105.838	70.827	0	35.0107
0	419	72.088	72.088	0	0.0000
1	427	45.611	9.374	0	36.2368
—[a]	—	—	—	—	—
1	490	70.785	25.965	0	44.8196
1	735	98.233	20.035	0	78.1972
1	715	98.913	23.440	0	75.4725

Note: Sum of the difference between the original and the reduced model score = $16,543.270.

[a] Partial list.

dataset could be used as the basis for developing the scorecards for each observation. Also, the parameter estimate dataset called "estim" could be used in the LSCORE macro for scoring different datasets containing the same predictor variables.

6. **If–then analysis and lift charts:** Perform an if–then analysis and construct a lift chart to estimate the differences in predicted probabilities when one of the continuous or binary predictor variables is fixed at a given value.

5.12.2 Data Descriptions

5.12.2.1 Background Information

To predict the probability of going bankrupt for financial institutions, a business analyst collected four financial indicators from 200 institutions 2 years prior to bankruptcy. He also recorded the same financial indicators from 200 financially sound firms. To develop a model predicting the probability of going bankrupt, the analysis used a logistic regression. The descriptions of the data and the data mining methods used are presented here.

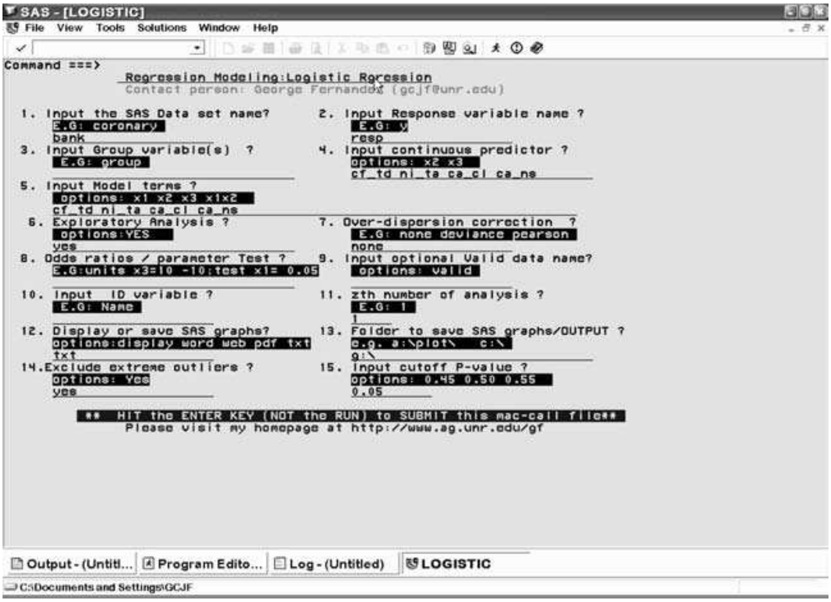
Dataset name	"bank": Permanent SAS dataset "bank" located in the library "gf". Two temporary datasets, TRAIN and VALID, are randomly selected using the RANSPLT macro.
Response variable	resp: 0, financially sound; 1, bankrupt
Predictor variables	CF_TD: cash flow/total debt NI_TA: net income/total assets CA_CL: current assets/current liabilities CA_NS: current assets/net sales
Number of observations	Total dataset: 400 Training: 360 Validation: 40
Source	Simulated data

Open the LOGISTIC.sas macro-call file in the SAS PROGRAM EDITOR window and click RUN to open the LOGISTIC macro-call window (Figure 5.31). Use the suggestions provided in the help file (see Section 5.8.2) to input the appropriate macro input values and fit the binary logistic regression (BLR). Only selected output and graphics produced by the LOGISTIC macro are shown here.

5.12.2.2 Exploratory Analysis/Diagnostic Plots

Input dataset name (bank), binary response variable (resp), continuous predictor variables (CF_TD, NI_TA, CA_CL, and CA_NS) model terms (CF_TD, NI_TA, CA_CL, and CA_NS) and other appropriate macro input values. To perform data exploration and to create regression diagnostic plots input YES in macro input field #6. Once the LOGISTIC macro has been submitted, overlay plots of simple and delta logits for each predictor variable and outputs from forward selection model selection are produced:

Overlay plots of simple and partial delta logit plots for full (four-variable) and reduced (three-variable, after excluding the nonsignificant NI_TA ratio) models are presented in Figures 5.32 to 5.35. The CF_TD variable showed a significant negative trend in both simple and delta logit plots. The significance of CF_TD improved in the reduced three-variable



SAS - [LOGISTIC]
File View Tools Solutions Window Help
Output - (Untitl... Program Edito... Log - (Untitled) LOGISTIC
C:\Documents and Settings\GCJF

```
Command ===>
              Regression Modeling:Logistic Rgression
              Contact person: George Fernandez (gcjf@unr.edu)

 1. Input the SAS Data set name?          2. Input Response variable name ?
    E.G: coronary                            E.G: y
    bank                                     resp
 3. Input Group variable(s)  ?            4. Input continuous predictor ?
    E.G: group                               options: x2 x3
                                             cf_td ni_ta ca_cl ca_ns
 5. Input Model terms ?
    options: x1 x2 x3 x1x2
    cf_td ni_ta ca_cl ca_ns
  6. Exploratory Analysis ?               7. Over-dispersion correction  ?
    options:YES                              E.G: none deviance pearson
    yes                                      none
 8. Odds ratios / parameter Test ?        9. Input optional Valid data name?
    E.G:units x3=10 -10;test x1= 0.05        options: valid

 10. Input  ID variable ?                11. zth number of analysis ?
    E.G: Name                                E.G: 1
                                             1
 12. Display or save SAS graphs?         13. Folder to save SAS graphs/OUTPUT ?
    options:display word web pdf txt         e.g. a:\plot\   c:\
    txt                                      g:\
 14.Exclude extreme outliers ?           15. Input cutoff P-value ?
    options: Yes                             options: 0.45 0.50 0.55
    yes                                      0.05

     ** HIT the ENTER KEY (NOT the RUN) to SUBMIT this mac-call file**
              Please visit my homepage at http://www.ag.unr.edu/gf
```

Figure 5.31 Screen copy of LOGISTIC macro-call window showing the macro-call parameters required for performing binary logistic regression (BLR).

model. A moderate degree of multicollinearity was evident in the full model, as the delta logit data points were clustered near the mean of CF_TD. When the nonsignificant variable NI_TA was dropped, the degree of multicollinearity was reduced and the significance level of CD_TD improved.

The NI_TA variable showed a negative trend in both simple and delta logit plots, but the significance of NI_TA was not significant in the four-variable model (Figure 5.33). A moderate degree of multicollinearity was evident in the full model, because the delta logit data points were clustered near the mean of NI_TA. When this variable NI_TA was dropped, the degree of multicollinearity was reduced and the significance level of the other three predictors improved.

The CA_CL variable showed a significant negative trend in both simple and delta logit plots (Figure 5.34). The significance of CA_CL did not change when the nonsignificant NI_TA was dropped from the full model. No evidence of multicollinearity was observed in the full model, because the delta logit data points spread along the partial logit regression line of CF_TD.

The CA_NS variable showed a significant positive trend in both simple and delta logit plots (Figure 5.35). The positive trend improved from the simple logit to the partial delta curve. The significance of CA_NS did not

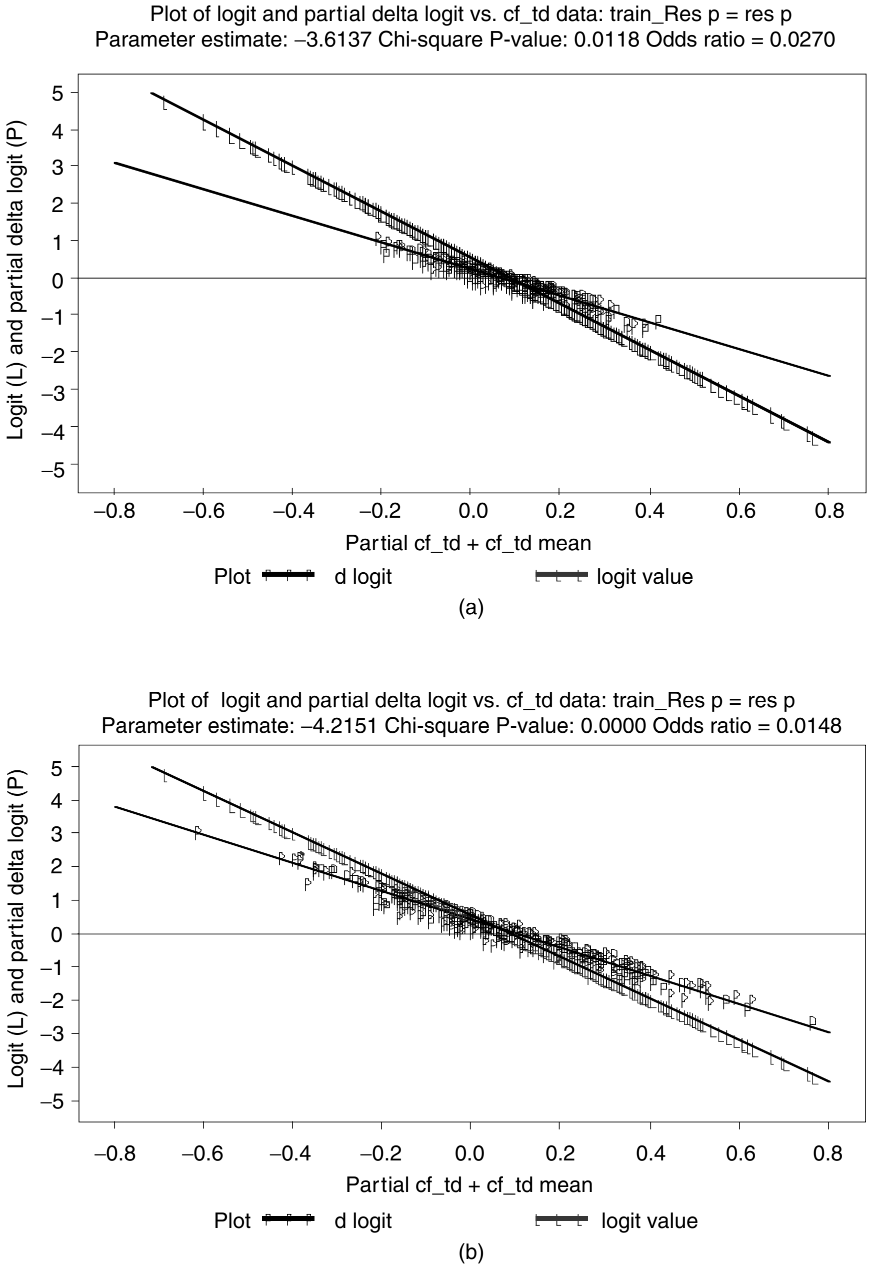


Figure 5.32 Logistic regression diagnostic plot: (a) Full model including all five predictors; (b) reduced model (after removing the nonsignificant variable X2 using SAS macro LOGISTIC. An overlay plot of simple logit and partial delta logit plots for the CF_TD ratio is shown.

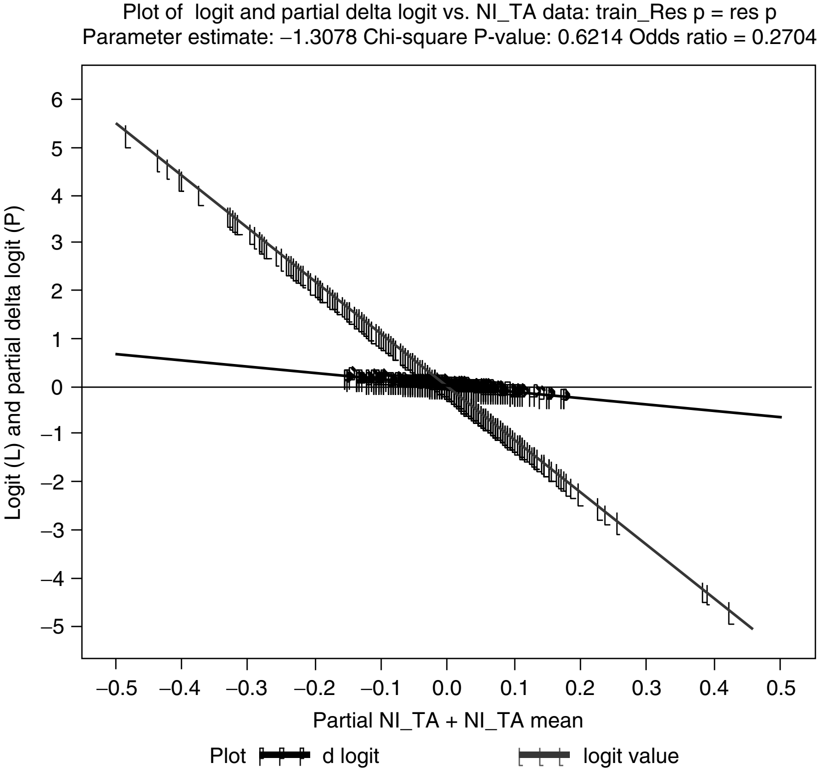


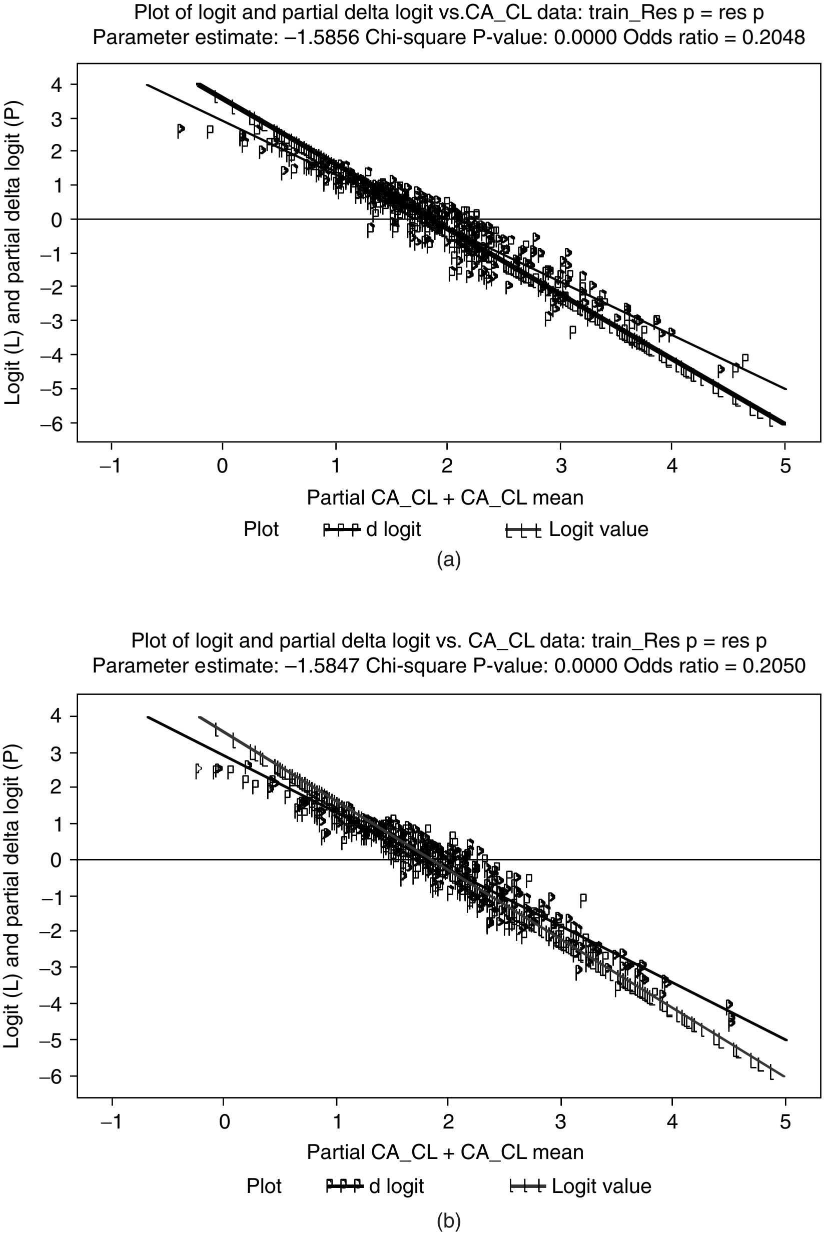
Figure 5.33 Logistic regression diagnostic plot using SAS macro LOGISTIC for full model including all five predictors. An overlay plot of simple logit and partial delta logit plots for the NI_TA ratio is shown.

change when the nonsignificant NI_TA was dropped from the full model. No evidence of multicollinearity was observed in the full model, because the delta logit data points spread along the partial logit regression line of CF_TD.

The summary statistics for the variable selection using the forward selection method are given in Table 5.24. The order of the variables selected and their significance levels based on this sequential fitting show that CA_CL is the most important variable in the single variable model. Variable NI_TA was not included, as it was not significant.

5.12.2.3 Fitting Binary Logistic Regression

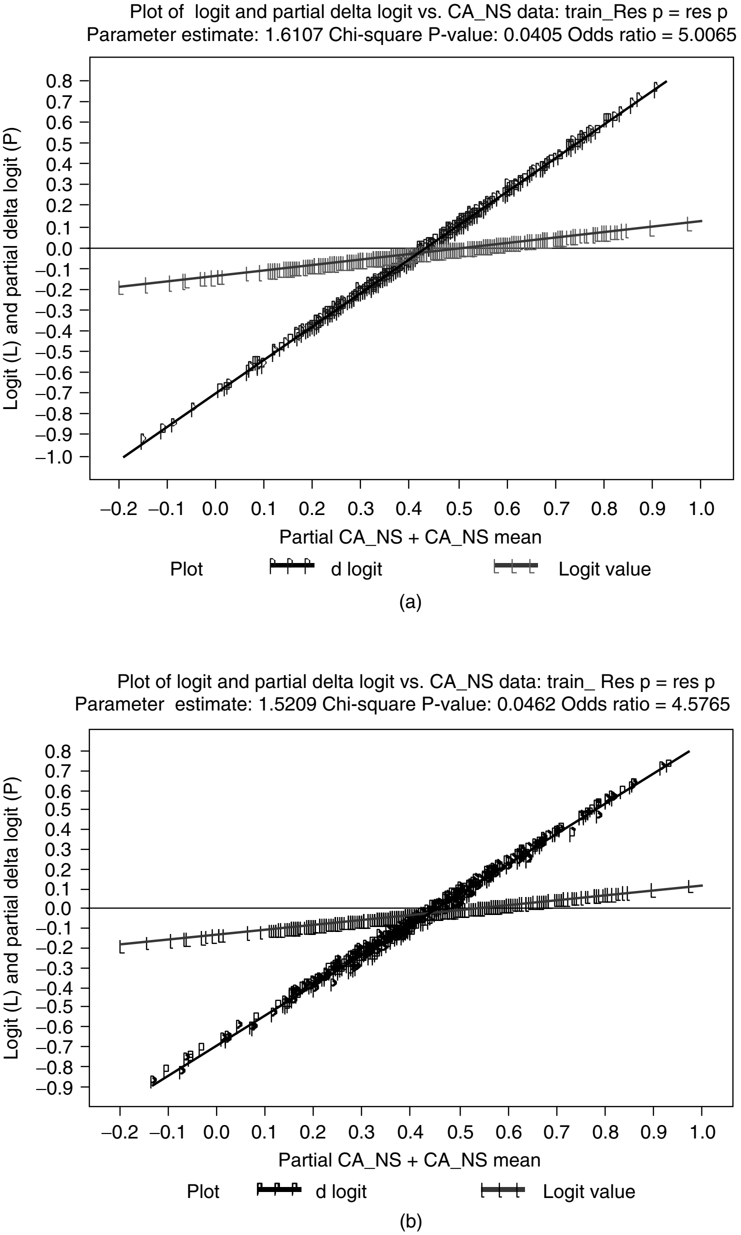
Open the LOGISTIC.sas macro-call file in the SAS PROGRAM EDITOR window and click RUN to open the LOGISTIC macro-call window (Figure



(a)

(b)

Figure 5.34 Logistic regression diagnostic plot: (a) Full model including all five predictors; (b) reduced model (after removing the nonsignificant variable NI_TA using SAS macro LOGISTIC. An overlay plot of simple logit and partial delta logit plots for CA_CL ratio is shown.



(a)

(b)

Figure 5.35 Logistic regression diagnostic plot: (a) Full model including all five predictors; (b) reduced model (after removing the nonsignificant variable NI_TA using SAS macro LOGISTIC. An overlay plot of simple logit and partial delta logit plots for CA_NS ratio is shown.

Table 5.24 Macro LOGISTIC: Summary Statistics from Forward Selection Method

Step	*Effect Entered*	*Degrees of Freedom*	*Number In*	*Score Chi-Square*	*Pr > Chi-Square*	*Variable Label*
1	CA_CL	1	1	144.2088	<.0001	CA_CL
2	CF_TD	1	2	38.5009	<.0001	CF_TD
3	CA_NS	1	3	4.6174	0.0316	CA_NS

5.31). Input training dataset name (TRAIN), validation dataset name (VALID), binary response variable (resp), three significant continuous predictor variables (CF_TD, CA_CL, and CA_NS) and model terms (CF_TD, CA_CL, and CA_NS). To skip data exploration and create regression diagnostic plots leave macro input field #6 blank. To exclude extreme outliers, input YES in macro input field #14. Input other appropriate macro parameters and submit the LOGISTIC macro to fit the binary logistic regression. Only selected output and graphics produced by the LOGISTIC macro are shown here.

The characteristics of the training dataset, number of observations in the dataset, type of model fit, and the frequency of event and non-event are given in Table 5.25. The LOGISTIC macro models the probability of the event. To model the probability of the non-event, recode the non-event to 1 in the data before running the macro.

Table 5.26 shows the model convergence status and statistics for testing the overall model significance. The AIC and SBC statistics for the intercept and covariates are useful for selecting the best model from various logistic models. Lower values are desirable. The –2 log L criteria are the –2 log likelihood statistics for models with only an intercept and the full model.

Table 5.25 Macro LOGISTIC: Model Information

Dataset	WORK.TRAIN
Bankrupt	1 = yes = 178
	0 = no = 182
Response variable	resp
Number of response levels	2
Number of observations	360
Model	Binary logit
Optimization technique	Fisher's scoring

Note: Probability modeled is resp = 1.

Table 5.26 Macro LOGISTIC: Model Fit Statistics

Criterion	*Intercept Only*	*Intercept and Covariates*
AIC	501.022	303.736
SBC	504.908	319.280
–2 log L	499.022	295.736

Table 5.27 Macro LOGISTIC: Testing Global Null Hypothesis that β = 0

Test	*Chi-Square*	*Degrees of Freedom*	*Pr > Chi-Square*
Likelihood ratio	203.2857	3	<.0001
Score	156.3423	3	<.0001
Wald	90.3887	3	<.0001

The results of testing the global null hypothesis that all regression coefficients are equal to zero are given in Table 5.27. All three tests indicate that the overall logistic model is highly significant and at least one of the parameter estimates is significantly different from zero. The parameter estimates, their standard error using maximum-likelihood methods, and significance level for each variable using the Wald chi-square are given in Table 5.28. The estimated logistic regression model for estimating the log (odds of bankruptcy) = 2.6335 – 4.2151CF_TD – 1.5847CA_CL + 1.5209CA_NS. All three regression parameter estimates are statistically significant. These parameter estimate the change in the log odds of bankruptcy when one unit changes in a given predictor variable while holding all other variables constant. For any observation in the data, the probability of the event could be predicted by the BLR parameter estimates.

Table 5.28 Macro LOGISTIC: Analysis of Maximum-Likelihood Estimates and Their Significance Levels, Final Model

Parameter	*Degrees of Freedom*	*Estimate*	*Standard Error*	*Wald Chi-Square*	*Pr > Chi-Square*
Intercept	1	2.6335	0.4704	31.3406	<.0001
CF_TD	1	–4.2151	0.7797	29.2290	<.0001
CA_CL	1	–1.5847	0.2302	47.4103	<.0001
CA_NS	1	1.5209	0.7630	3.9734	0.0462

By default, the LOGISTIC macro outputs the odds ratio estimate for one-unit increases in all predictor variables while holding all other variables constant. These odds ratios and their confidence intervals are computed by exponentiating the parameter estimates and their confidence intervals. Table 5.29 shows the odds ratio estimates and the profile likelihood confidence interval estimates. The interpretation of the odds ratio is valid for a linear continuous or binary predictor when one-unit change in the predictor variable is relevant. If interaction terms are included in the model, the interpretation of the odds ratio for a given variable is not valid.

An odds ratio value equal to 1 implies no change in the odds of the event when the predictor variable is increased by one unit. The chance of the bankruptcy increases multiplicatively when the odds ratio of a variable is greater than 1 and vice versa. For example, one unit increase in the CA_NS ratio while holding all other predictors constant multiplies the odds of becoming bankrupt by 1.03 to 20.41 times. Similarly a one-unit increase in the CF_TD ratio while holding all other predictors constant multiplies the odds of bankruptcy by 0.003 to 0.068 times, thus reducing the chance to a very low level (Table 5.29). Because an increase in one unit in the CF_TD ratio is unrealistic, we could estimate customized odds ratio estimates and their confidence intervals in the LOGISTIC macro by specifying the values in macro input option #8. The customized odds ratio

Table 5.29 Macro LOGISTIC: Profile Likelihood Confidence Interval for Adjusted Odds Ratios

Effect	*Point Estimate*	*95% Wald Confidence Limits*	
		Lower	*Upper*
CF_TD	0.015	0.003	0.068
CA_CL	0.205	0.131	0.322
CA_NS	4.576	1.026	20.417

Table 5.30 Macro LOGISTIC: Profile Likelihood Interval for Customized Odds Ratios

Effect	*Unit*	*Estimate*	*95% Confidence Limits*	
			Lower	*Upper*
CF_TD	0.1000	0.656	0.559	0.760
CF_TD	–0.1000	1.524	1.316	1.788

estimates and their profile confidence interval estimates for a ±0.1-unit change in CF_TD are given in Table 5.30. While holding other predictor variables at a constant level, an increase of 0.1 unit in the CF_TD reduces the odds of bankruptcy by 0.56 to 0.76 times, and a decrease of 0.1 unit in the CF_TD increases the odds of bankruptcy by 1.32 to 1.78 times. The chance of bankruptcy increases 31 to 78% and decreases 24 to 45% when the CF_TD ratio decreases or increases 0.1 units, respectively.

Formal tests results for checking for model adequacy are presented in Table 5.31. The *p* value for the Hosmer and Lemeshow goodness-of-fit test using the Pearson chi-square test is not significant at a 5% level. The estimated R^2 and the max-rescaled R^2 statistics computed from the –2 log likelihood estimates are moderately high. A smaller Brier score closer to 0 indicates that the predictive power of the logistic regression model is high. Neither the deviance nor Pearson goodness-of-fit statistics are significant, indicating that the variance of the binary response variable does not exceed the expected nominal variance. Thus, overdispersion is not a problem in this logistic regression model.

The biased adjusted classification table created by the CTABLE option in the SAS LOGISTIC procedure is presented in Table 5.32. In computing the biased adjusted classification table, SAS uses an approximate jackknifing method using the observed sample proportion as the prior probability estimates. The classification table uses the estimated logistic model based on cross-validation techniques to classify each observation either as events or non-events at different probability cutpoints. Each observation is classified as an event if its estimated probability is greater than or equal to a given probability cutpoint; otherwise, the observation is classified as a non-event. The classification table provides several measures of predictive accuracy at various probability cutpoints.

Table 5.31 Macro LOGISTIC: Model Adequacy Tests

Hosmer and Lemeshow Goodness-of-Fit Test

Chi-Square	*Degrees of Freedom*	*Pr > Chi-Square*	*R^2*	*Max-Rescaled R^2*	*Brier Score*
6.9200	8	0.5453	0.4315	0.5753	0.13121

Deviance and Pearson Goodness-of-Fit Statistics

Criterion	*DF*	*Value*	*Value/DF*	*Pr > Chi-Square*
Deviance	356	295.7358	0.8307	0.9912
Pearson	356	338.4269	0.9506	0.7404

Table 5.32 Macro LOGISTIC: Classification Table

	Correct		*Incorrect*		*Percentages*[a]				
Probability Cutpoint	*Event*	*Non-Event*	*Event*	*Non-Event*	*Correct*	*Sensitivity*	*Specificity*	*False Positive*	*False Negative*
0.050	178	60	122	0	66.1	100.0	33.0	40.7	0.0
0.100	175	81	101	3	71.1	98.3	44.5	36.6	3.6
0.150	175	97	85	3	75.6	98.3	53.3	32.7	3.0
0.200	174	103	79	4	76.9	97.8	56.6	31.2	3.7
0.250	174	115	67	4	80.3	97.8	63.2	27.8	3.4
0.300	171	119	63	7	80.6	96.1	65.4	26.9	5.6
0.350	169	127	55	9	82.2	94.9	69.8	24.6	6.6
0.400	162	132	50	16	81.7	91.0	72.5	23.6	10.8
0.450	157	136	46	21	81.4	88.2	74.7	22.7	13.4
0.500	148	142	40	30	80.6	83.1	78.0	21.3	17.4
0.550	143	146	36	35	80.3	80.3	80.2	20.1	19.3
0.600	134	151	31	44	79.2	75.3	83.0	18.8	22.6
0.650	121	155	27	57	76.7	68.0	85.2	18.2	26.9
0.700	114	159	23	64	75.8	64.0	87.4	16.8	28.7
0.750	98	165	17	80	73.1	55.1	90.7	14.8	32.7
0.800	88	169	13	90	71.4	49.4	92.9	12.9	34.7
0.850	64	170	12	114	65.0	36.0	93.4	15.8	40.1
0.900	45	173	9	133	60.6	25.3	95.1	16.7	43.5
0.950	20	180	2	158	55.6	11.2	98.9	9.1	46.7
1.000	0	182	0	178	50.6	0.0	100.0	—	49.4

[a] *Correct* is the overall percentage of correct classification (total frequency of correct classification/sample size). *Sensitivity* is a measure of accuracy of classifying events or true positives (number of correctly classified events/total number of events). *Specificity* is a measure of accuracy of classifying non-events or true negatives (number of correctly classified non-events/total number of non-events). *False positive* is a measure of error in classification of non-events (number of falsely classified non-events as events/total number of events). *False negative* is a measure of error in classification of events (number of falsely classified events as non-events/total number of non-events).

For each probability cutpoint, the correct and incorrect columns provide the frequency of events and non-events correctly and incorrectly classified. For example, at the 0.5 probability cutpoint, the LOGISTIC procedure correctly classifies 148 events and 142 non-events. It misclassifies 40 events and 30 non-events. The next five columns provide predictive accuracy of the model at various cutpoint probabilities.

The ROC (receiver operating characteristic) curve presented in Figure 5.36 top provides the measure of predicted accuracy of the fitted logistic regression model. Because the ROC curve rises quickly, the fitted model has a relatively high predictive accuracy, which is confirmed by the large area under the ROC curve (0.88) that is displayed in the figure title. The *C* statistic estimated by the LOGISTIC procedure is equal to the area under the ROC curve. Figure 5.36 bottom shows the trend between the percentages of correct classification, false positives, and false negatives and the different cutpoint probability values. If the objective is to find the cutpoint probability that gives overall predictive accuracy, select the cutpoint probability where the false-positive and false-negative curves intersect and the overall classification percentage is high. Approximately at the cutpoint probability equal to 0.575, both false-positive and false-negative estimates are low.

The estimated predicted probability scores and their confidence intervals are given in Table 5.33. These predicted scores and the confidence interval estimates could be used to build scorecards for each financial institution to help identify financially troubled firms.

5.12.2.4 Significant Outlier/Influential Observations

Table 5.34 lists several observations as influential outliers based on the larger DIFDEV statistic (DIFDEV >4). The DIFDEV statistic measures the change in the model deviance statistic when these observations are excluded individually. These outliers also show up in the outlier detection plot (Figure 5.37), which illustrates the leverage of each observation (hat-value). The diameter of the bubbles in the outlier detection plot is proportional to the *cbar* statistic, an influential statistic that quantifies the change in the parameter confidence interval estimates when each observation is excluded. Because some of the outliers have relatively larger *cbar* statistics, to investigate the impact of these outliers this analysis should be repeated by selecting macro input option #14 to exclude extreme influential points.

5.12.2.5 Model Validation

To validate the obtained logistic regression model estimates, the regression parameter estimates obtained from the training dataset (n = 360) are used to predict the bankruptcy for the validation dataset (N = 40), and the results are presented in Table 5.35. The Brier score for this validation dataset is relatively small: 0.111, similar to the Brier score obtained from the training data. The misclassification percentage is 22% (9 out of 40 misclassified). Even though the validation results are satisfactory, the impact of the extreme outliers on model validation could be verified by refitting the model by excluding the outliers from the analysis.

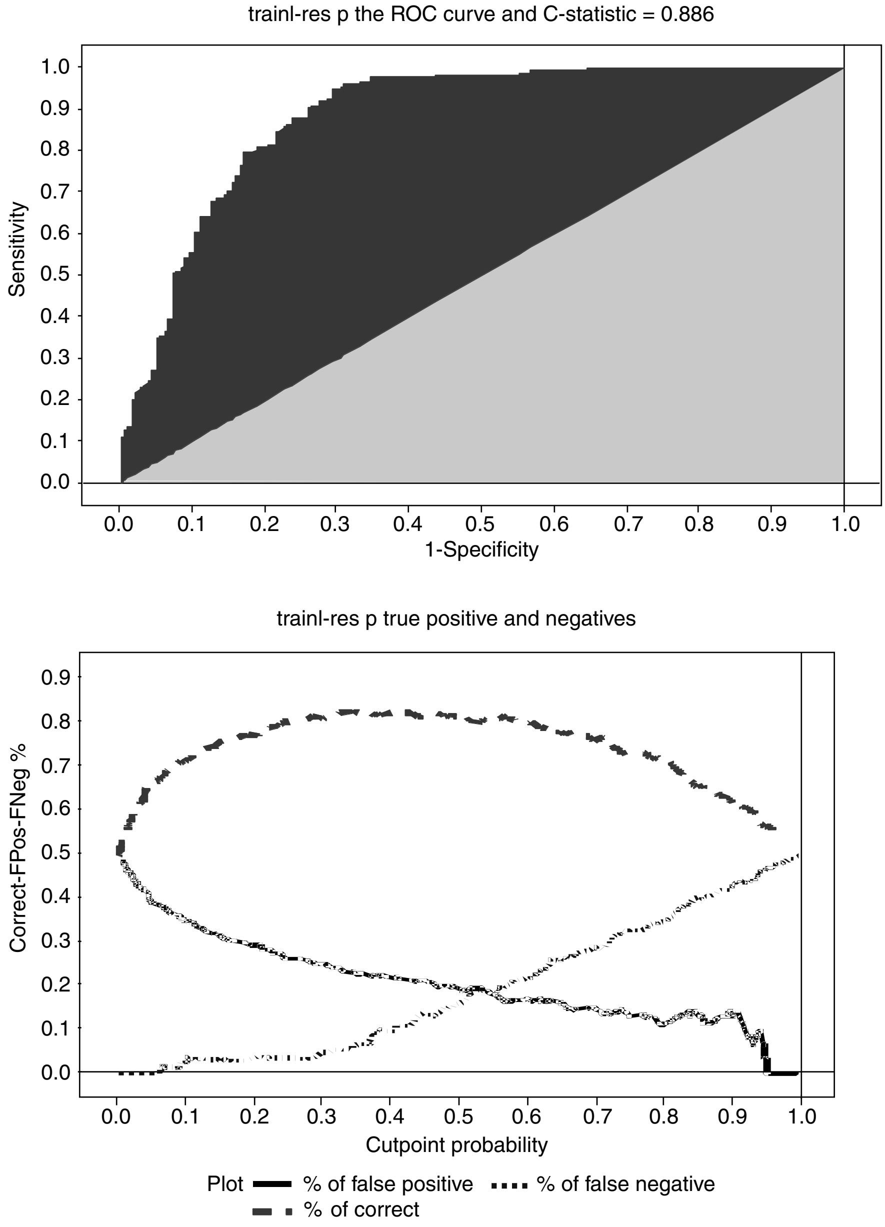


Figure 5.36 Assessing the binary logistic regression (BLR) fit: (top) Receiver operating characteristic (ROC) curve; (bottom) overlay plot showing false positive and false negative percentages vs. different cutpoint probabilities using the SAS macro LOGISTIC.

Table 5.33 Macro LOGISTIC: Partial List of Predicted Probability and Their Confidence Intervals

ID	*Response*	*Predicted Probability*	*CF_TD*	*CA_CL*	*CA_NS*	*p*	*95% Lower Confidence Interval*	*95% Upper Confidence Interval*
1	1	1	–0.15	1.4621	0.6501	0.87416	0.79788	0.92438
2	1	0	0.1079	2.1367	0.7673	0.48993	0.35223	0.62918
3	1	0	0.2651	1.7747	0.8058	0.48227	0.31599	0.65257
—[a]	—	—	—	—	—	—	—	—
360	0	0	0.7063	3.763	0.4697	0.00371	0.00108	0.01273

[a] Partial list.

Table 5.34 Macro LOGISTIC – Partial List of Influential Observations

ID	*CF_TD*	*CA_CL*	*CA_NS*	*Response*	*p*	*Residual Deviance*	*hat*	*cbar*	*Delta Deviance*
3	0.2651	1.7747	0.8058	1	0.48227	1.20769	0.031968	0.03545	1.49396
16	0.3388	0.944	0.5467	1	0.63199	0.95800	0.028605	0.01715	0.93490
24	0.5626	2.2275	0.7125	1	0.10116	2.14059	0.018864	0.17084	4.75297
27	0.0939	1.8677	0.8118	1	0.62541	0.96887	0.024001	0.01473	0.95344
30	–0.224	1.008	–0.141	1	0.85398	0.56186	0.033159	0.00586	0.32156
34	–0.055	1.5323	0.0156	1	0.61280	0.98966	0.031343	0.02044	0.99987
51	–0.211	1.1697	0.0081	1	0.84299	0.58447	0.023468	0.00448	0.34608
52	0.223	1.1789	0.0944	1	0.49220	1.19068	0.024293	0.02569	1.44341
63	0.6359	2.0578	0.3875	1	0.06189	2.35900	0.011672	0.17902	5.74391
—[a]	—	—	—	—	—	—	—	—	—
328	–0.117	2.5411	0.3428	0	0.40674	–1.02188	0.024652	0.01733	1.06157
333	0.031	1.4944	0.0656	0	0.55835	–1.27847	0.023863	0.03091	1.66540
341	–12E-5	0.6978	0.5047	0	0.90853	–2.18710	0.007201	0.07204	4.85544
342	–0.122	1.0481	0.4573	0	0.89858	–2.13939	0.005974	0.05325	4.63023
349	0.0696	0.2605	0.2751	0	0.91258	–2.20772	0.010247	0.10807	4.98210
356	–0.011	0.8628	0.7445	0	0.92013	–2.24825	0.010344	0.12041	5.17503
358	0.6106	0.6698	0.1735	0	0.32352	–0.88413	0.061081	0.03111	0.81280

[a] Partial listing (26 outliers are not shown).

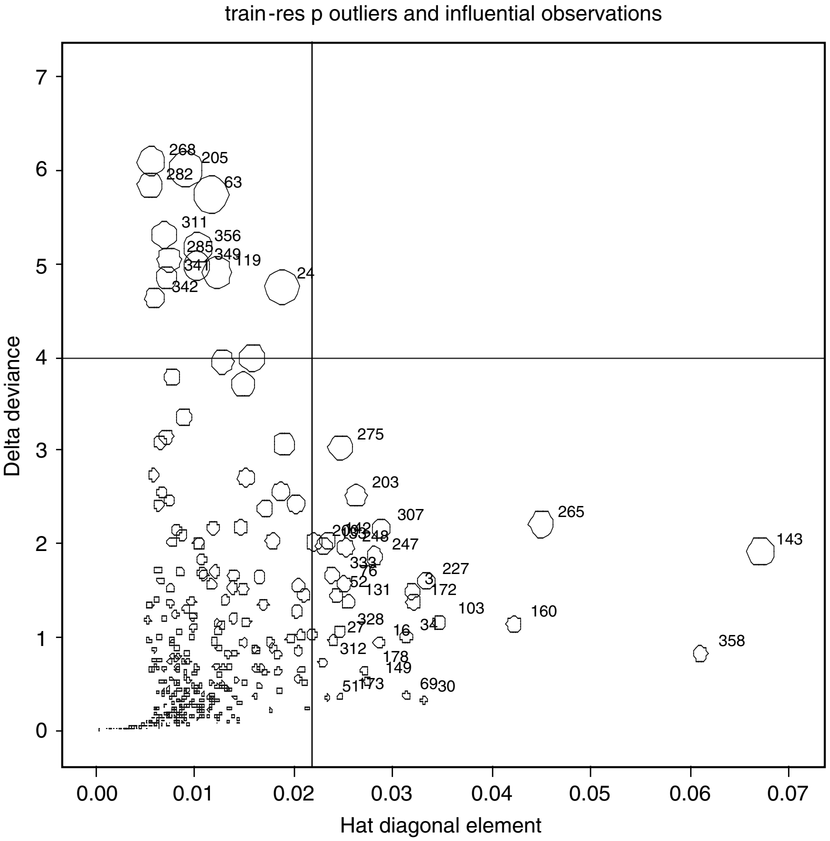


Figure 5.37 Diagnostic plot: checking for influential outliers in binary logistic regression (BLR) using the SAS macro LOGISTIC.

Table 5.35 Macro LOGISTIC: Classification Table of Observed vs. Predicted Probability at the Cut-Off *p* Value 0.5, Validation Data

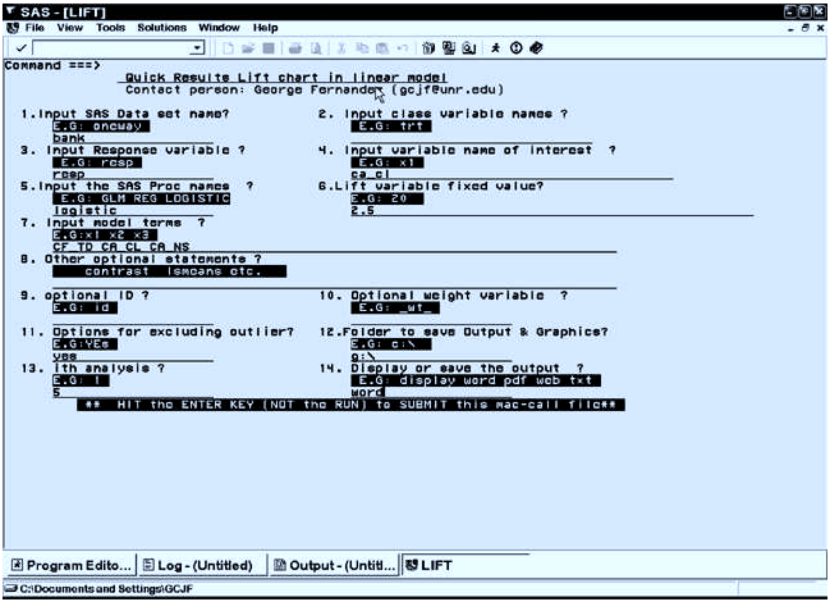
	Predicted Probability[a]		
Frequency[a]	*0*	*1*	*Total*
0	14	4	18
1	5	17	22
Total	19	21	40

Note: N = 40; Brier score = 0.13047.

[a] Bankrupt: yes = 1; no = 0.

5.12.2.6 Performing If–Then Analysis and Producing the LIFT Chart

After finalizing the regression model, the next objective is to perform an if–then analysis and create a lift chart showing what happens to the predictive probability if the CA_CL ratio is held constant at 2.5. Open the LIFT.sas macro-call file in the SAS PROGRAM EDITOR window and click RUN to open the LIFT macro-call window (Figure 5.38). Input the SAS dataset (TRAIN) and response (resp) variable name, model statement (CF_TD, CA_CL, and CA_NS), the variable name of interest (CA_CL), fixed level (2.5), and other appropriate macro input values by following the suggestions given in the help file (see Section 5.8.2). Submit the lift macro to get a lift chart (Figure 5.39) showing the differences in the predicted probability scores between the original full model and the new reduced model where the CA_CL ratio was fixed at 2.5. On the lift chart display, the mean difference (0.19) between the predicted probability values of the original model and the mean predicted probability value for the reduced model are also displayed. In addition to the lift chart, the if–then analysis output table showing, the predicted probability values of the original and the reduced models and the differences in the predicted probability values are also produced. A partial list of the if–then analysis


SAS - [LIFT]
File View Tools Solutions Window Help
Program Edito...
Log - (Untitled)
Output - (Untitl...
LIFT
C:\Documents and Settings\GCJF

```
Command ===>
              Quick Results Lift chart in linear model
              Contact person: George Fernandez (gcjf@unr.edu)

 1.Input SAS Data set name?          2. Input class variable names ?
   E.G: oneway                          E.G: trt
   bank
 3. Input Response variable ?        4. Input variable name of interest  ?
   E.G: resp                            E.G: x1
   resp                                 ca_cl
 5.Input the SAS Proc names  ?       6.Lift variable fixed value?
   E.G: GLM REG LOGISTIC                E.G: 20
   logistic                             2.5
 7. Input model terms  ?
   E.G:x1 x2 x3
   CF_TD CA_CL CA_NS
 8. Other optional statements ?
   contrast  lsmeans etc.

 9. optional ID ?                    10. Optional weight variable  ?
   E.G: id                              E.G: _wt_

 11. Options for excluding outlier?  12.Folder to save Output & Graphics?
   E.G:YEs                              E.G: c:\
   yes                                  g:\
 13. ith analysis ?                  14. Display or save the output  ?
   E.G: 1                               E.G: display word pdf web txt
   5                                    word
     ** HIT the ENTER KEY (NOT the RUN) to SUBMIT this mac-call file**
```

Figure 5.38 Screen copy of LIFT macro-call window showing the macro-call parameters required for performing lift chart in logistic regression.

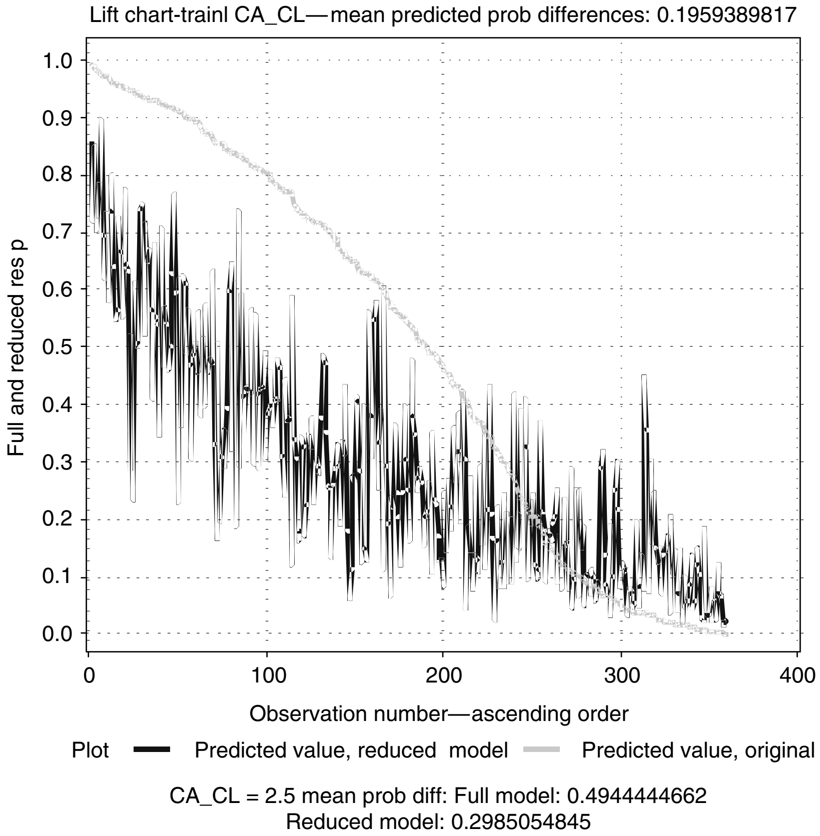


Figure 5.39 Lift chart generated by using SAS macro LIFT: differences in predicted probability for "bankruptcy" between the original and the reduced model if the CA_CL ratio was held constant at 2.5.

output is presented in Table 5.36. Thus, we could conclude that the average probability of bankruptcy drops by 0.19 if the CA_CL ratio is held constant at 2.5 without any change in other financial indicators.

5.13 Summary

The methods of performing supervised predictive models in predicting continuous and binary response variables using user-friendly SAS macro applications are covered in this chapter. Graphical methods to perform diagnostic and exploratory analysis, model assessment and validation, and detecting violation of model assumptions and to produce lift charts are also discussed. Steps involved in using the user-friendly SAS macro appli-

Table 5.36 Macro LIFT Difference in Mean Predicted Probability between Original and Reduced Model (= 0.195)

CA_CL	*ID*	*Predicted Value, Original*	*Predicted Value, Reduced*	*newvalue*	*Difference in Probability Values*
0.5271	5	0.99268	0.85607	2.5	0.13661
0.2874	103	0.98856	0.72162	2.5	0.26694
0.7975	191	0.98835	0.85098	2.5	0.13736
0.4815	177	0.98332	0.70644	2.5	0.27688
0.7829	145	0.98240	0.78602	2.5	0.19638
0.6905	53	0.97923	0.72825	2.5	0.25098
1.5107	41	0.97654	0.89668	2.5	0.07986
0.7595	111	0.97285	0.69438	2.5	0.27847
0.5678	94	0.97230	0.62154	2.5	0.35076
0.0961	92	0.94818	0.28849	2.5	0.65969
0.7318	17	0.94728	0.52162	2.5	0.42567
–0.056	231	0.94633	0.23489	2.5	0.71144
4.8772	335	0.00323	0.12283	2.5	–0.11960
4.5875	360	0.00282	0.07167	2.5	–0.06885

cations REGDIAG for performing regression modeling and validation, LOGISTIC for performing binary logistic regression and validation, LIFT for generating lift charts and if–then analyses, and RSCORE (regression) and LSCORE (logistic) for predicting scores from new datasets are also presented.

References

1. Neter, J., Kutner, M.H., Nachtsheim, C.J., and Wasserman, W., *Applied Linear Regression Models*, Irwin, Homewood, IL, 1996, chaps. 1–5.
2. Montgomery, D.C. and Peck, E.A., *Introduction to Linear Regression Analysis*, 2nd ed., John Wiley & Sons, New York, 1992, chaps. 1–4.
3. Freund, R.J. and Littell, R.C., *SAS System for Regression*, 1st ed., SAS Institute, Inc., Cary, NC, 1986.
4. SAS Institute, Inc., *Regression with Quantitative and Qualitative Variables*, SAS online documentation, SAS Institute, Inc., Cary, NC (http://v8doc.sas.com/sashtml/stat/chap55/sect52.htm; accessed June 2002).

5. Fernandez, G.C.J., Detection of model specification, outlier and multicollinearity in multiple regressions using partial regression/residual plots, in *Proc. SAS Users Group Int. Conf.*, San Diego, CA, 1997, pp. 1246–1251.
6. Mallows, C.L., Augmented partial residual plots, *Technometrics*, 28, 313–319, 1986.
7. Sall, J., Leverage plots for general linear hypothesis, *Am. Statistician*, 44, 308–315, 1990.
8. Stine, R.A., Graphical interpretation of variance inflation factors, *Am. Statistician*, 49, 53–56, 1995.
9. SAS Institute, Inc., *Model Selection Methods: The REG Procedure*, SAS online documentation, SAS Institute, Inc., Cary, NC (http://v8doc.sas.com/sashtml/stat/chap55/sect29.htm-regmsm; accessed June 2002).
10. Neter, J., Kutner, M.H., Nachtsheim, C.J., and Wasserman, W., *Applied Linear Regression Models*, Irwin, Homewood, IL, 1996, chap. 8.
11. Neter, J., Kutner, M.H., Nachtsheim, C.J., and Wasserman, W., *Applied Linear Regression Models*, Irwin, Homewood, IL, 1996, chap. 3.
12. SAS Institute, Inc., *SAS/ETS Users Guide: The AUTOREG Procedure*, SAS online documentation, SAS Institute, Inc., Cary, NC (http://v8doc. sas.com/sashtml/ets/chap8/index.htm; accessed June 2002).
13. Fernandez, G.C.J., *Regression Analysis Using SAS Macros* (http://www.ag.unr.edu/gf/apst705SASlab.htm; accessed June 2002).
14. SAS Institute, Inc., *The REG Procedure: Influential Diagnostics*, SAS online documentation, SAS Institute, Inc., Cary, NC (http://v8doc.sas.com/sashtml/stat/chap55/sect38.htm; accessed June 2002).
15. Neter, J., Kutner, M.H., Nachtsheim, C.J., and Wasserman, W., *Applied Linear Regression Models*, Irwin, Homewood, IL, 1996, chap. 10.
16. SAS Institute, Inc., *The MIXED Procedure: RANDOM Statement*, SAS online documentation, SAS Institute, Cary, NC (http://v8doc.sas.com/sashtml/stat/chap41/sect19.htm-mixedrandom; accessed June 2002).
17. d'Agostino, R.B., Belanger, A., and d'Agostino, R.B., Jr., A suggestion for using powerful and informative tests of normality, *Am. Statistician*, 44, 316–321, 1990.
18. SAS Institute, Inc., *GENMOD Procedure: Overview*, SAS online documentation, SAS Institute, Cary, NC (http://v8doc.sas.com/sashtml/stat/chap29/sect1.htm; accessed June 2002).
19. Zelterman, D., *Models for Discrete Data*, Clarendon Press, Oxford, 1999, chap. 3.
20. SAS Institute, Inc., *Logistic Regression Examples Using the SAS Systems Version 6*, 1st ed., SAS Institute, Inc., Cary, NC.
21. SAS Institute, Inc., *The LOGISTIC Procedure Overview*, SAS online documentation, SAS Institute, Cary, NC (http://v8doc.sas.com/sashtml/stat/chap26/index.htm; accessed June 2002).
22. SAS Institute, Inc., *The LOGISTIC Procedure Odds Ratio Estimation*, SAS online documentation, SAS Institute, Cary, NC (http://v8doc.sas.com/sashtml/stat/chap26/index.htm; accessed June 2002).

23. SAS Institute, Inc., *The LOGISTIC Procedure Confidence Intervals for Parameters*, SAS online documentation, SAS Institute, Cary, NC (http://v8doc.sas.com/sashtml/stat/chap26/index.htm; accessed June 2002).
24. SAS Institute, Inc., *The LOGISTIC Procedure Model Fitting*, SAS online documentation, SAS Institute, Cary, NC (http://v8doc.sas.com/sashtml/stat/chap26/index.htm; accessed June 2002).
25. SAS Institute, Inc., The LOGISTIC Procedure Score Statistics and Tests, SAS online documentation, SAS Institute, Cary, NC (http://v8doc.sas.com/sashtml/stat/chap26/index.htm; accessed June 2002).
26. SAS Institute, Inc., *The LOGISTIC Procedure The Hosmer-Lemeshow Goodness-of-Fit Test*, SAS online documentation (http://v8doc.sas.com/sashtml/stat/chap26/index.htm; accessed June 2002).
27. SAS Institute, Inc., *Logistic Regression Examples Using the SAS Systems Version 6*, 1st ed., SAS Institute, Inc., Cary, NC, 1995, chap. 7.
28. SAS Institute, Inc., *The LOGISTIC Procedure Generalized Coefficient of Determination*, SAS online documentation, SAS Institute, Cary, NC (http://v8doc.sas.com/sashtml/stat/chap26/index.htm; accessed June 2002).
29. SAS Institute, Inc., *The LOGISTIC Procedure Classification Table*, SAS online documentation, SAS Institute, Cary, NC (http://v8doc.sas.com/sashtml/stat/chap26/index.htm; accessed June 2002).
30. SAS Institute, Inc., *The LOGISTIC Procedure Receiver Operating Characteristic Curves*, SAS online documentation, SAS Institute, Cary, NC (http://v8doc.sas.com/sashtml/stat/chap26/index.htm; accessed June 2002).
31. SAS Institute, Inc., *Logistic Regression Examples Using the SAS Systems Version 6,* 1st ed., SAS Institute, Cary, NC, 1995, chap. 6.
32. SAS Institute, Inc., *Logistic Regression Examples Using the SAS Systems Version 6*, 1st ed., SAS Institute, Cary, NC, 1995, chap. 8.
33. SAS Institute, Inc., *The LOGISTIC Procedure Over Dispersion*, SAS online documentation, SAS Institute, Cary, NC (http://v8doc.sas.com/sashtml/stat/chap26/index.htm; accessed June 2002).
34. SAS Institute, Inc., *The REG Procedure Overview*, SAS online documentation, SAS Institute, Cary, NC (http://v8doc.sas.com/sashtml/stat/chap26/index.htm; accessed June 2002).
35. SAS Institute, Inc., *The GLM Procedure Overview*, SAS online documentation, SAS Institute, Cary, NC (http://v8doc.sas.com/sashtml/stat/chap26/index.htm; accessed June 2002).
36. SAS Institute, Inc., PROC LOGISTIC Syntax, SAS online documentation, SAS Institute, Cary, NC (http://v8doc.sas.com/sashtml/stat/chap39/sect3.htm; accessed June 2002).

Chapter 6

Supervised Learning Methods: Classification

6.1 Introduction

The goal of supervised classification models is to fit a model or decision tree that will correctly associate the input variables with the categorical group levels. Classification models use categorical response to approximate the probability of class membership as a function of the input variables. Decision tree analysis is one of the new tools used in data mining to predict the membership of a group response by means of a decision tree. In general, classification models allow analysts to determine the form of the relationship between the response and the predictor variables and further investigate which predictor variables are associated with categorical response. Two supervised classification model techniques, discriminant analysis (DA) and chi-squared automatic interaction detection methods (CHAID) are discussed in this chapter.

Discriminant analysis, a multivariate statistical technique, classifies subjects (cases, observations) to groups or categories. The main purpose of discriminant analysis is to predict membership in two or more mutually exclusive groups from a set of predictors, when the groups have no natural ordering.

Classification trees are used to predict the membership of cases or objects in the levels of a categorical response from their measurements on one or more predictor variables and then to output a decision tree. The flexibility of classification trees makes them a very attractive supervised classification method. When stringent distributional assumptions of traditional methods are met, the traditional methods may be preferable. But, as an exploratory technique, or as a technique of last resort when traditional methods fail, the potentials of classification trees are unsurpassed.

A brief non-mathematical description and application of these supervised classification methods are provided in this chapter. Readers are encouraged to refer to the following statistics books for mathematical accounts of DA: Sharma[1] and Johnson and Wichern;[2] for CHAID, Breiman et al.[3]

6.2 Discriminant Analysis

Discriminant analysis, a multivariate statistical technique, is commonly used to build a predictive or descriptive model of group discrimination based on observed predictor variables and to classify each observation into one of the groups. In DA, multiple quantitative attributes are used to discriminate single mutually exclusive classification variables. DA is different from cluster analysis in that prior knowledge of the group membership is required for DA.

Discriminant analysis is a very powerful tool for selecting the predictor variables that allow discrimination between different groups and for classifying cases into different groups with a better than chance accuracy. The common objectives of DA are:

- To investigate differences between groups of observations
- To discriminate groups effectively
- To identify important discriminating variables
- To perform hypothesis testing on the differences between the expected groupings
- To classify new observations into pre-existing groups

Based on the objectives of the analysis, DA can be classified into three types: stepwise discriminant analysis (SDA), canonical discriminant analysis (CDA), and discriminant function analysis (DFA). SDA is an exploratory component in the discriminant step, where the important discriminating variables are selected in a stepwise fashion. When group membership is known in advance and the purpose of the analysis is to describe and highlight the major group differences, CDA is appropriate. In DFA the focus is on classifying cases into predefined groups or to predict group membership on the basis of predictor variable measures. The statistical theory, methods, and the computation aspects of SDA,[4–6] CDA,[4,5,7] and DFA[4,5,8] are presented in detail in the literature.

6.3 Stepwise Discriminant Analysis

Stepwise discriminant analysis is an exploratory component in DA commonly used to identify a subset of predictor variables that maximize discrimination among group members. The predictor variables within each class are assumed to have a multivariate normal distribution with a common covariance matrix. The SDA analysis is a useful exploratory component to either CDA or DFA. In SDA, continuous predictor variables are chosen to enter or leave the model based on the significance level of an *F* test or squared partial correlation from an analysis of covariance, where the variable is already chosen, the group response acts as predictors, and the variable under consideration is considered as the response variable. A moderate significance level, in the range of 0.10 to 0.25, often performs better than the use of a much larger or a much smaller significance level in the model selection.[6]

Three types of SDA methods are available in the SAS systems:[6]

1. *Forward selection* begins as a null model. At each step, the predictor variable is selected that contributes most to the discriminatory power of the model, as measured by Wilks' lambda (the likelihood ratio criterion). Once a variable is selected, it cannot be removed from the model. When none of the unselected variables meets the entry criterion, the forward selection process stops.
2. *Backward elimination* begins as a full model. At each step, the predictor variable that contributes least to the discriminatory power of the model as measured by Wilks' lambda is removed. Once a variable is excluded, it cannot be entered into the model. When all remaining variables meet the criterion to stay in the model, the backward elimination process stops.
3. *Stepwise selection* begins, like forward selection, as a null model. At each step, the significance of the previously entered predictor variables is compared. The variable that contributes least to the discriminatory power of the model is removed; otherwise, the variable that is not in the model that contributes most to the discriminatory power is entered. When all variables in the model meet the criterion to stay and none of the other variables meets the criterion to enter, the stepwise selection process stops.

Using stepwise selection methods to evaluate the relative importance of predictor variables or selecting predictor variables for finding the best fitted DA is not always guaranteed to give the best results. Limitations in model selection methods include:

- Only one variable can be entered into the model at each step. Thus, all possible combinations of predictors are not evaluated.
- The selection process does not take into account the relationships between variables that have not yet been selected; thus, some important variables could be excluded in the selection process.
- Wilks' lambda also may not be the best criterion for evaluating the discriminatory power in a specific application.

These problems inherent with stepwise methodologies can be serious, especially in small samples. Reducing the number of predictor variables to a manageable size is a useful strategy in the exploratory analysis stage. It is important to exclude highly correlated redundant predictor variables if they do not have discriminatory power; therefore, using SDA as a stand-alone DA method is not recommended. However, SDA can be a treated as a valuable exploratory tool in excluding redundant variables for building a successful discrimination model.

6.4 Canonical Discriminant Analysis

In CDA, linear combinations of the quantitative predictor variables that summarize between-group variation and provide maximal discrimination between the groups are extracted. The methodology used in CDA is very similar to a one-way multivariate ANOVA (MANOVA). In MANOVA, the goal is to test for equality of the mean vector across class levels, while in CDA, the significance of continuous predictor variables in discriminating categorical response is investigated. Thus, the CDA can be conceptualized as the inverse of multivariate MANOVA and all assumptions for MANOVA apply to CDA.

6.4.1 Canonical Discriminant Analysis Assumptions

- **Multivariate normal distribution:** Multivariate normal distribution is a requirement for performing hypothesis testing in CDA. It is assumed that within-group predictor the variables have a multivariate normal distribution. The validity of multivariate normality assumptions can be performed by first standardizing the predictors within each group level and testing for significant Mardia's multivariate kurtosis.[9] Significant departure from multivariate normality can be examined visually in a quantile–quantile (Q–Q) plot.[9] The results of CDA could be misleading if multivariate assumptions are severely violated. When some of the predictor variables are heavily skewed, performing a

Box–Cox type transformation can reduce the severity of the problem.[9]

- **Homogeneity of variance–covariance:** It is assumed that the variance–covariance matrices of predictors are homogeneous across groups. Minor departure from homogeneity of variance–covariance is not that important; however, before accepting the results of CDA, it is probably a good idea to verify the equality of within-group variance–covariance matrices. Problems with the heterogeneous variance–covariance can also be confounded by the lack of multivariate normality. The Box–Cox transformation recommended for reducing the impact of departure from multivariate normality also indirectly corrects the unequal variance–covariance problems. The SAS DISCRIM procedure provides a method for testing the homogeneity of within-error variance–covariance based on the unbiased likelihood ratio statistic adjusted by the Bartlett correction.
- **Multicollinearity among the predictors:** Another assumption of CDA is that the predictor variables that are used to discriminate group members are not highly correlated. If any one of the predictors is totally dependent with the other variables, then the matrix is said to be ill conditioned. The results of CDA become unstable and unpredictable when severe multicollinearity is present.

6.4.2 Key Concepts and Terminology in Canonical Discriminant Analysis

6.4.2.1 Canonical Discriminant Function

A canonical discriminant function (CDF) is a latent variable that is created as a linear combination of predictor variables, such that $CDF_1 = a + b_1x_1 + b_2x_2 + \ldots + b_nx_n$, where the b_i are raw discriminant coefficients, the x_i are predictor variables, and a is a constant. The product of the unstandardized coefficients with the observations yields the discriminant scores. The group centroid is the mean value for the discriminant scores for a given group level. The CDF is considered analogous to multiple regressions, but the b_i are discriminant coefficients that maximize the distance between the means of the group members. The first CDF always provides the most overall discrimination between groups, the second CDF provides second most, and so on. The process of extracting the CDF can be repeated until the number of CDFs equals the number of original variables or the number of classes minus one, whichever is smaller. Moreover, the CDFs are independent or orthogonal, like principal components; that is, their contributions to the discrimination between groups do not overlap. The CDF score is the value resulting from applying a CDF to the data for a given observation.

6.4.2.2 Squared Canonical Correlation (R_c^2)

The squared canonical correlation (R_c^2) is the measure of multiple correlations between the discriminating group and the CDF. The first canonical correlation is at least as large as the multiple correlations between the groups and any of the original variables. The second canonical correlation is obtained by finding the linear combination uncorrelated with the first canonical variable that has the highest possible multiple correlations with the groups. If the original variables have high within-group correlations, the first canonical correlation can be large even if all the multiple correlations are small. In other words, the first canonical variable can show substantial differences between the classes, even if none of the original variables does.

6.4.2.3 MANOVA Test of Group Mean Differences

The F test of Wilks' lambda shows which variables' contributions are significant. Wilks' lambda is also used in discriminant analysis to test the significance of the CDFs as a whole. One can test the number of CDFs that add significantly to the discrimination between groups. Only those found to be statistically significant should be used for interpretation; nonsignificant functions (roots) should be ignored. The variables should have an approximate multivariate normal distribution within each class with a common covariance matrix in order for the probability levels to be valid.

6.4.2.4 Structure Coefficients

The structure coefficients provide the correlations of each predictor variable with each CDF. These simple Pearson correlations are called *structure coefficients*, or *discriminant loadings*. The raw discriminant function coefficients denote the unique (partial) contribution of each variable to the discriminant functions, while the structure coefficients denote the simple correlations between the variables and the functions. To assign meaningful names to the discriminant functions, the structure coefficients should be used. But, to learn each predictor variable's unique contribution to the discriminant function, use the discriminant function coefficients (weights).

6.4.2.5 Bi-Plot Display of Canonical Discriminant Analysis

The bi-plot display of CDA[10] is a visualization technique for investigating the interrelationships between the group members and the canonical function scores in CDA. The term *bi-plot* means a plot of two dimensions with the observation and variable spaces plotted simultaneously. In CDF, relationships between CDF scores and CDF structure loadings associated with any two CDFs can be illustrated in a bi-plot display.[11] The success of CDF in discriminating the group members can be visually verified in the bi-plot.

6.5 Discriminant Function Analysis

Another major purpose of DA is to classify observation into pre-defined groups. In a dataset containing multiple quantitative predictor variables and a classification variable, discriminant functions derived by the DFA can be used to classify each observation into one of the groups. The discriminant function, also known as a *classification criterion*, is determined by a measure of the generalized squared distance The classification criterion can be determined based on either the individual within-group covariance matrices (yielding a quadratic function) or the pooled covariance matrix (yielding a linear function). This classification criterion also takes into account the prior probabilities of the groups. The calibration information can be stored in a special SAS dataset and applied to other datasets. The derived classification criterion from the training dataset can be applied to a validation dataset simultaneously.

6.5.1 Key Concepts and Terminology in Discriminant Function Analysis

- **Parametric DFA:** Using a generalized squared distance measure, a parametric method can be used to develop a classification criterion when the distribution within each group is multivariate normal. The classification criterion can be derived based on either the individual within-group covariance matrices or the pooled covariance matrix that also takes into account the prior probabilities of the classes. Each observation is placed in the class from which it has the smallest generalized squared distance. The posterior probability of an observation belonging to each class can also be computed.
- **Nonparametric DFA:** When the multivariate normality assumption within each group is not met, nonparametric DFA methods can be used to estimate the group-specific densities. Either a kernel or the *k*-nearest-neighbor method can be used to generate a nonparametric density estimate in each group and to produce a classification criterion.

The performance of a discriminant criterion can be evaluated by estimating the probabilities of misclassification of new observations in the validation data.

- **Classification criterion:** The classification functions can be used to determine which group each case most likely belongs to. There are as many classification criterions as there are groups. Each classification criterion allows computation of classification probability scores for each member in each group. The Mahalanobis distance estimates are used to determine proximity in computing the classification criterion. We can use the classification criterion to directly compute classification probability scores for new observations.
- **Mahalanobis distance:** The Mahalanobis distance or the pairwise square distance is the distance between an observation and the centroid for each group in *p*-dimensional space defined by *p* variables and their covariance. Thus, the smaller the Mahalanobis distance, the closer the observation is to the group centroid and the more likely it is to be assigned to that group. In SAS, only the pooled covariance matrix can be used to calculate the Mahalanobis distances in the *k*-nearest-neighbor method. With parametric or kernel DFA methods, either the individual within-group covariance matrices or the pooled covariance matrix can be used to calculate the Mahalanobis distances.[8]
- **Prior probabilities:** Sometimes we know ahead of time that there are more observations in one group than in any other; thus, we should adjust our prediction according to the prior probability. Prior probabilities are the likelihood of belonging to a particular group given that no information about the observation is available. You can specify different prior probabilities that will then be used to adjust the classification of cases and the computation of posterior probabilities accordingly. The prior probabilities can be set to be proportional to the sizes of the groups in the training data or can be set equal to each other in each group.[8] The specification of different prior probabilities can greatly affect the accuracy of the prediction.
- **Posterior probabilities:** The posterior probability is a probability based on knowledge of the values of predictor variables that the respective case belongs to a particular group. The probability that an observation belongs to a particular group is approximately proportional to the Mahalanobis distance from that group centroid. The Mahalanobis distance is one of the main components in estimating the group-specific probability densities. Once we have

computed the posterior probability scores for an observation, we can decide how to classify the observation. In general, we classify a given observation to a predefined group for which it has the highest posterior probability scores. With the estimated group-specific probability densities and their associated prior probabilities, the posterior probability estimates of group membership for each class can be estimated.

- **Classification table:** The classification table is used to assess the performance of DA and is simply a table in which the rows are the observed categories of the group response and the columns are the predicted categories of the response. When prediction is perfect, all observations will lie on the diagonal. The percentage of observations on the diagonal is the percentage of correct classifications.
- **Classification of error-rate estimates:** A classification criterion can be evaluated by its performance in the classification of future observations. Two types of error-rate estimates are commonly used to evaluate the derived classification criterion based on posterior probability values estimated by the training sample. The posterior probability estimates provide good estimates of the error rate when the posterior probabilities are accurate; when a parametric classification criterion (linear or quadratic discriminant function) is derived from a nonnormal population, the resulting posterior probability error-rate estimators may not be appropriate. The overall classification error rate is estimated through a weighted average of the individual group-specific error-rate estimates, where the prior probabilities are used as the weights.

When no independent test data sets are available, the same dataset can be used both to define and to evaluate the classification criterion. The resulting error-count estimate has an optimistic bias and is called an *apparent error rate.*[12] To reduce the bias, the data can be split into two sets, one set for deriving the discriminant function and the other set for estimating the error rate. The error-count estimate is calculated by applying the classification criterion derived from the training sample to a test set and then counting the number of misclassified observations. The group-specific error-count estimate is the proportion of misclassified observations in the group. When the test set is independent of the training sample, the estimate is unbiased; however, it can have a large variance, especially if the test set is small. Such a split-sample method has the unfortunate effect of reducing the effective sample size.

To reduce both the bias and the variance of the estimator, posterior probability estimates can be computed based on cross validation. Cross validation treats (n – 1) out of n observations as a training set. It determines the classification criterion based on these (n – 1) observations and then applies them to classify the one observation left out. This is done for each of the n training observations. The misclassification rate for each group is the proportion of sample observations in that group that are misclassified. This method achieves a nearly unbiased estimate but with a relatively large variance.[13]

To reduce the variance in an error-count estimate, smoothed error-rate estimates are suggested.[14] Instead of summing terms that are either zero or one, as in the error-count estimator, the smoothed estimator uses the posterior probability scores, which are a continuum of values between 0 and 1 in the terms that are summed. The resulting estimator has a smaller variance than the error-count estimate. Two types (unstratified and stratified) of smoothed posterior probability error-rate estimates are provided by the POSTERR option in the PROC DISCRIM statement. The stratified posterior probability error-rate estimates take into account the relative sizes of the groups, whereas the unstratified error-rate estimate does not.[5]

6.6 Applications of Discriminant Analysis

Discriminant analysis can be used to develop classification criteria for grouping bank customers, online shoppers, diabetes types, college admissions, etc. Given categorical (two or more levels) outcomes, DA can estimate the probability that a new cable television customer will upgrade to digital cable or to cable modem. Other DA applications include:

- Investigate differences among groups of college freshmen that drop out in one year, transfer to a new school, or complete the degree.
- Discriminate types of cancer cells effectively.
- Select the important financial indicators for discriminating fast-growing, stable, or diminishing stocks.
- Classify a new credit card customer into pre-existing groups based on paying credit card bills.

It is very clear from these applications that DA can be a useful tool for decision-making endeavors.

6.7 Classification Tree Based on CHAID

Classification tree analysis is a segmentation technique designed to split a population into two or more categories based on available attributes (e.g., gender, employment status, race, age). The results are often presented in the form of a tree. Classification tree analysis can be characterized as a hierarchical, highly flexible technique for predicting membership of cases in the classes of a categorical response using one or more predictor variables. The predictor variables used in developing a classification tree can be categorical, continuous, or any mix of the two types of predictors. The study and use of classification trees are widespread in many diverse fields such as medicine, business, biology, and social sciences. Classification tree analysis can sometimes be quite complex; however, graphical procedures can be developed to help simplify interpretation even for complex trees. The goal of a classification tree analysis is to obtain the most accurate prediction possible.

In a number of ways, classification trees are different from traditional statistical methods (DA, logistic regression) for predicting class membership on a categorical response. When stringent statistical assumptions of traditional methods are met, the traditional methods may be preferable for classification problems, but, as an exploratory technique or when traditional methods fail, classification trees are preferred. Classification methods employ a hierarchy of predictions, sometimes being applied to particular cases to sort the cases into predicted classes. Traditional methods use simultaneous techniques to make one and only one class membership prediction for each and every case. Tree-based methods have several attractive properties when compared to traditional methods. They provide a simple rule for classification or prediction of observations, they handle interactions among variables in a straightforward way, they can easily handle a large number of predictor variables, and they do not require assumptions about the distribution of the data. However, tree-based methods do not conform to the usual hypothesis-testing framework. For finding models that predict well, there is no substitute for a thorough understanding of the nature of the relationships between the predictor and response variables. A brief account of non-mathematical descriptions and applications of classification tree methods are provided in this section. Additional details of decision tree methods are discussed elsewhere.[15,16]

6.7.1 Key Concepts and Terminology in Classification Tree Methods

6.7.1.1 Construction of Classification Trees

A decision tree partitions data into smaller segments called *terminal nodes* or *leaves* that are homogeneous with respect to a target variable. Partitions are defined in terms of input variables, thereby defining a predictive relationship between the inputs and the target. This partitioning goes on until the subsets cannot be partitioned any further using one of many user-defined stopping criterion. By creating homogeneous groups, analysts can predict with greater certainty how individuals in each group will behave.

Classification trees most commonly used are univariate and can be constructed based on splitting a single ordinal-scale predictor variable after performing transformation that preserves the order of values on the ordinal variable. Thus, classification trees based on univariate splits can be computed without concern for whether a unit change on a continuous predictor represents a unit change on the dimension underlying the values on the predictor variable. In short, assumptions regarding the level of measurement of predictor variables are less stringent in classification tree analysis.

Classification tree analyses are not limited to univariate splits on the predictor variables. When continuous predictors are measured on at least an interval scale, linear combination splits, similar to the splits for linear DA, can be computed for classification trees. However, the linear combination splits computed for classification trees do differ in important ways from the linear combination splits computed for DA. In linear DA, the number of linear discriminant functions that can be extracted is the lesser of the number of predictor variables or the number of classes on the group variable minus one. The recursive approach implemented for classification tree module does not face this limitation.

6.7.1.2 Chi-Square Automatic Interaction Detection Method

The chi-square automatic interaction detection (CHAID) method is a classification tree method to study the relationship between a group response variable and a series of predictor variables. CHAID modeling selects a set of predictors and their interactions that optimally predict the response measure. The developed classification tree (or data partitioning tree) shows how major subsets are formed from the predictor variables by differentially predicting a criterion or response variable. The SAS TREEDISC macro uses the CHAID type of analysis.[15]

The decision tree in TREEDISC is constructed by partitioning the dataset into two or more subsets of observations based on the categories of one of the predictor variables. After the dataset is partitioned according to the chosen predictor variable, each subset is considered for further partitioning using the same criterion that was applied to the entire dataset. Each subset is partitioned without regard to any other subset. This process is repeated for each subset until some stopping criterion is met. This recursive partitioning forms a tree structure.[15]

6.7.1.3 Decision Tree Diagram

The decision tree diagram consists of a tree trunk that progressively splits into smaller and smaller branches. The *root* of the tree is the entire dataset. The subsets and sub-subsets form the *branches* of the tree. Subsets that meet a stopping criterion and thus are not partitioned are *leaves*, or the *terminal node*. The number of subsets in a partition can range from two up to the number of categories of the predictor variable. Note that the tree can be pictured in an orientation from top to bottom or left to right or right to left and the results are identical. Different orientations of the same tree are sometimes useful to highlight different portions of the results.

6.7.1.4 Classification Criterion

The predictor variable used to form a partition is chosen to be the variable that is most significantly associated with the response variable according to a chi-squared test of independence in a contingency table. The main stopping criterion used by the TREEDISC macro is the *p* value from this chi-squared test. A small *p* value indicates that the observed association between the predictor and the dependent variable is unlikely to have occurred solely as the result of sampling variability. If a predictor has more than two categories, then there may be a very large number of ways to partition the dataset based on the categories. A combinatorial search algorithm is used to find a partition that has a small *p* value for the chi-squared test. The *p* values for each chi-squared test are adjusted for the multiplicity of partitions.

Predictors can be nominal, ordinal, or ordinal with a floating category.[15] For a nominal predictor, the categories are not ordered and therefore can be combined in any way to form a partition. For an ordinal predictor, the categories are ordered, and only categories that are adjacent in the order can be combined when forming a partition.

6.7.1.5 Assessing the Decision Trees

A general issue that arises when applying tree classification is that the final trees can become very large. In practice, when the input data are complex and, for example, contain many different categories for classification problems and many possible predictors for performing the classification, then the resulting trees can become very large and interpretation becomes difficult.

6.8 Applications of CHAID

- In database marketing, decision trees can be used to segment groups of customers and develop customer profiles to help marketers produce targeted promotions that achieve higher response rates. Also, because the CHAID algorithm will often effectively yield many multi-way frequency tables (e.g., when classifying a categorical response variable with many categories, based on categorical predictors with many classes), it has been particularly popular in marketing research, in the context of market segmentation studies.
- A credit card issuer may use a decision tree to define a rule of the form: "If the monthly mortgage-to-income ratio is less than 30%, the months posted late are less than 1, and salary is greater than $36,000, then issue a gold card."

6.9 Discriminant Analysis Using SAS Macro DISCRIM

The DISCRIM macro is a powerful SAS application for performing complete discriminant analysis. Options are available for obtaining various exploratory and diagnostic graphs and for performing different types of discriminant analyses. SAS procedures STEPDISC and DISCRIM are the main tools used in the DISCRIM macro.[6–8] In addition to these SAS procedures, GPLOT and BOXPLOT procedures and IML modules are also utilized in the DISCRIM macro. The advantages of using the DISCRIM macro over the PROC DISCRIM are:

- Exploratory bivariate plots to check for group discrimination in a simple scatterplot between two predictor variables are generated.
- Plots for checking multivariate normality and influential observations within each group are also generated.
- Test statistics and *p* values for testing equality in variance and covariance matrices within each group level are automatically produced.
- In the case of CDA, box plots of canonical discriminant functions by groups and biplot display of canonical discriminant function scores of observations and the structure loadings for the predictors are generated. When fitting DFA, box plots of the *i*th-level posterior probability by groups are produced.
- Options are available for validating the discriminant model obtained from a training dataset using an independent validation dataset by comparing classification errors.

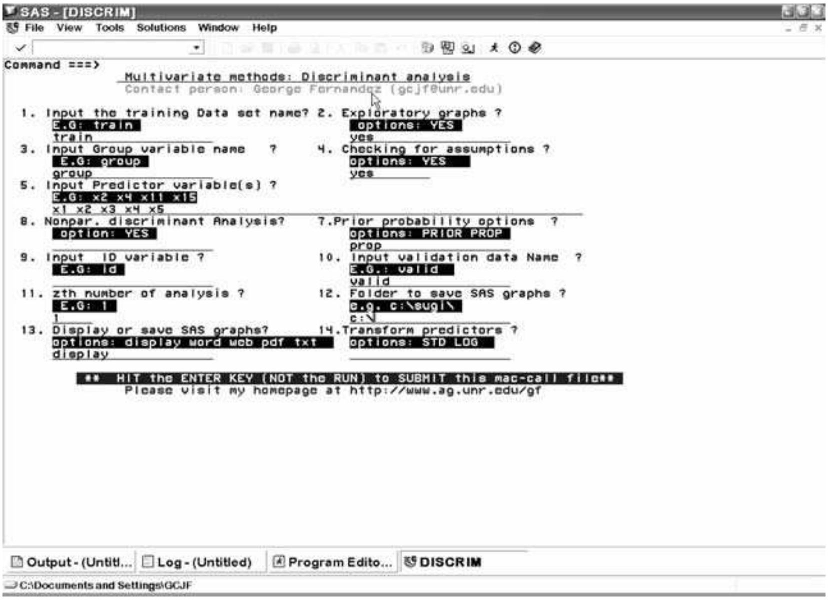
- Options for saving the output tables and graphics in WORD, HTML, PDF, and TXT formats are available.

Software requirements for using the DISCRIM macro are:

- SAS/BASE, SAS/STAT, and SAS/GRAPH must be licensed and installed at the site. SAS/IML is required to check for multivariate normality.
- SAS version 8.0 and above is recommended for full utilization.
- An active Internet connection is required for downloading the DISCRIM macro from the book website if the companion CD-ROM is not available.

6.9.1 Steps Involved in Running the DISCRIM Macro

1. Create an SAS dataset (permanent or temporary) containing at least one categorical group response (target) variable and many continuous and/or ordinal predictor (input) variables. (Disabling the SAS ENHANCED EDITOR is highly recommended in the latest SAS versions; open only the PROGRAM EDITOR window.)
2. *If the companion CD-ROM is not available*, first verify that the Internet connection is active. Open the DISCRIM.sas macro-call file in the SAS PROGRAM EDITOR window. Instructions are given in the Appendix regarding downloading the macro-call and sample data files from the book website. *If the companion CD-ROM is available*, the DISCRIM.sas macro-call file can be found in the mac-call folder in the CD-ROM. Open the DISCRIM.sas macro-call file in the SAS PROGRAM EDITOR window. Click the RUN icon to submit the macro-call file DISCRIM.sas to open the macro-call window called DISCRIM (Figure 6.1).
3. Input the appropriate parameters in the macro-call window by following the instructions provided in the DISCRIM macro help file in Section 6.9.2. Users can choose whether to include exploratory graphs and variable selections by stepwise methods or skip the exploratory methods. After inputting all the required macro parameters, be sure that the cursor is in the last input field and the RESULTS VIEWER window is closed, then hit the ENTER key (not the RUN icon) to submit the macro.
4. Examine the LOG window (only in DISPLAY mode) for any macro execution errors. If any errors are reported in the LOG window, activate the PROGRAM EDITOR window, resubmit the DISCRIM.sas macro-call file, check the macro input values, and correct any input errors.



SAS - [DISCRIM]
File View Tools Solutions Window Help
Output - (Untitl... Log - (Untitled) Program Edito... DISCRIM
C:\Documents and Settings\GCJF

```
Command ===>
             Multivariate methods: Discriminant analysis
             Contact person: George Fernandez (gcjf@unr.edu)

 1. Input the training Data set name? 2. Exploratory graphs ?
    E.G: train                           options: YES
    train                                yes
 3. Input Group variable name   ?     4. Checking for assumptions ?
    E.G: group                           options: YES
    group                                yes
 5. Input Predictor variable(s) ?
    E.G: x2 x4 x11 x15
    x1 x2 x3 x4 x5
 8. Nonpar. discriminant Analysis?    7.Prior probability options  ?
    option: YES                          options: PRIOR PROP
                                         prop
 9. Input  ID variable ?              10. Input validation data Name  ?
    E.G: id                              E.G.: valid
                                         valid
11. zth number of analysis ?          12. Folder to save SAS graphs ?
    E.G: 1                               e.g. c:\sugi\
    1                                    c:\
13. Display or save SAS graphs?       14.Transform predictors ?
    options: display word web pdf txt    options: STD LOG
    display

   **  HIT the ENTER KEY (NOT the RUN) to SUBMIT this mac-call file**
             Please visit my homepage at http://www.ag.unr.edu/gf
```

Figure 6.1 Screen copy of the DISCRIM macro-call window showing the macro-call parameters required for performing data exploration in discriminant analysis.

5. If no errors are found in the LOG window, activate the PROGRAM EDITOR window, resubmit the DISCRIM.sas macro-call file, and change the macro input value from DISPLAY to any other desirable format (see Section 6.9.2.). The SAS output files from complete discriminant analysis and all the exploratory and descriptive graphs can be saved as user-specified-format files in the user-specified folder.

6.9.2 *Help File for SAS Macro DISCRIM*

1. **Macro-call parameter:** Input SAS dataset name (required parameter).
   Descriptions and explanation: Include the name of the temporary (member_name) or permanent (libname.member_name) SAS dataset on which the discriminant analysis will be performed. This dataset should contain at least a categorical group variable and two continuous predictor variables.
   Options/examples:
   Permanent SAS dataset: gf.diabetic1 (LIBNAME: gf; SAS dataset name: diabetic1)
   Temporary SAS dataset: diabetic1 (SAS dataset name)

2. **Macro-call parameter:** Exploratory analysis (optional parameter).
   Descriptions and explanation: This macro-call parameter is used to select the type of analysis (exploratory graphics and variable selection or CDA and DFA).
   Options/examples:

   Yes: Only the scatterplot matrix of all predictor variables by group response are produced. Variable selection by forward selection, backward elimination, and stepwise selection methods are performed. Discriminant analysis (CDA or DFA) is not performed.

   Blank: If the macro input field is left blank, exploratory analysis and variable selection are not performed. Only CDA and parametric or nonparametric DFA are performed.
3. **Macro-call parameter:** Input categorical group response variable name (required parameter).
   Descriptions and explanation: Input the categorical group response name from the SAS dataset to model as the target variable. Code the group levels to numeric values (1, 2, 3, etc.) to automatically generate box plots of posterior probability density estimates by group.
   Option/example:

   group (name of a categorical response)
4. **Macro-call parameter:** Check for assumptions? (Optional statement)
   Descriptions and explanation: To check for multivariate normality and check for the presence of any extreme multivariate outliers/influential observations, input YES. If this field is left blank, this step will be omitted.
   Options/examples:

   Yes: Statistical estimates for multivariate skewness, multivariate kurtosis, and their statistical significance are produced. Also produced are Q–Q plots for checking multivariate normality and multivariate outlier detection plots.

   Blank: If the macro input field is left blank, no statistical estimates for checking for multivariate normality or detecting outliers are performed.
5. **Macro-call parameter:** Input the predictor variable names (required statement).
   Descriptions and explanation: Input the continuous variable names from the dataset to be used in the discriminant analysis as predictors.
   Options/examples:

   X1 X2 X3 X4 X5 mpg murder (names of continuous predictor variables)

6. **Macro-call parameter:** Nonparametric discriminant analysis (optional statement).
   Descriptions and explanation: Select the type of discriminant (parametric or nonparametric) analysis.
   Options/examples:
      Yes: Canonical discriminant analysis and parametric discriminant function analysis will *not* be performed; instead, nonparametric discriminant analysis based on the *k*th nearest neighbor and kernel density methods will be performed. The probability density in the nearest neighbor (k = 2 to 4) nonparametric discriminant analysis method will be estimated using the Mahalanobis distance based on the pooled covariance matrix. Posterior probability estimates in kernel density nonparametric discriminant analysis methods will be computed using these kernel = normal r = 0.5 PROC DISCRIM options. (For details about parametric and nonparametric discriminant analysis options, see SAS online manuals on PROC DISCRIM.[8])
      Blank: Canonical discriminant analysis and parametric discriminant function analysis will be performed assuming all the predictor variables within each group level have multivariate normal distribution.
7. **Macro-call parameter:** Prior probability options (required statement).
   Descriptions and explanation: Input the prior probability option required for computing posterior probability and classification error estimates.
   Options/examples: To set the prior probabilities of group membership.
      Equal: To set the prior probabilities equal.
      Prop: To set the prior probabilities proportional to the sample sizes.
   (For details about prior probability options, see SAS online manuals on PROC DISCRIM.[8])
8. **Macro-call parameter:** Input ID variable (optional statement).
   Descriptions and explanation: If a unique ID variable can be used to identify each member in the dataset, input that variable name here to be used as the ID variable so that any outlier/influential observations can be detected. If no ID variable is available in the dataset, leave this field blank. This macro can create an ID variable based on the observation number from the database.
   Option/example:
      ID NUM

9. **Macro-call parameter:** Input validation dataset name (optional parameter).
   Descriptions and explanation: If you would like to validate the discriminant model obtained from a training dataset by using an independent validation dataset, input the name of the SAS validation dataset.
   Options/examples:
   Permanent SAS dataset: gf.diabetic2 (LIBNAME: gf; SAS dataset name: diabetic2)
   Temporary SAS dataset: diabetic2 (SAS dataset name).
10. **Macro-call parameter:** *z*th number of analysis (required statement).
    Descriptions and explanation: SAS output files will be saved by forming a file name from the original SAS dataset name and the counter value provided in this field. For example, if the original SAS dataset name is train and the counter number included is 1, then the SAS output files will be saved as "train1" in the user-specified folder. By changing the counter value, users can avoid replacing previous SAS output files with new outputs.
11. **Macro-call parameter:** Folder to save SAS graphics and output files? (Optional statement).
    Descriptions and explanation: To save the SAS graphics files in an EMF format suitable for inclusion in PowerPoint presentations, specify output format as TXT in version 8.0 or later. In pre-8.0 versions, all graphic format files will be saved in a user-specified folder. Similarly, output files in WORD, HTML, PDF, and TXT formats will be saved in the user-specified folder. If this macro field is left blank, the graphics and output files will be saved in the default folder.
    Option/example:
    c:\output\ — folder named "OUTPUT"
12. **Macro-call parameter:** Display or save SAS output and graphs (required statement).
    Descriptions and explanation: Option for displaying all output files in the OUTPUT window or saving as a specific format in the folder specified in macro input option #11.
    Options/examples:
    Possible values
    DISPLAY: Output will be displayed in the OUTPUT window, all SAS graphics will be displayed in the GRAPHICS window, and system messages will be displayed in the LOG window.
    WORD: If MS WORD is installed in the computer, the output and all SAS graphics will be saved together in the user-specified folder as a single RTF format file (version

8.0 and later) and will be displayed in the SAS VIEWER window. SAS output files will be saved as a text file in pre-8.0 versions, and all graphics files (CGM) will be saved separately in a user-specified folder (macro input option #11).

WEB: Output and graphics are saved in the user-specified folder and are viewed in the results VIEWER window as a single HTML file (version 8.0 and later) or as a text file in pre-8.0 versions. All graphics files (GIF) will be saved separately in a user-specified folder (macro input option #11).

PDF: If Adobe Acrobat Reader is installed in the computer, output and graphics are saved in the user-specified folder and are viewed in the results VIEWER window as a single PDF file for version 8.2 and later only. For pre-8.2 versions, all graphics files (PNG) will be saved separately in a user-specified folder as a text file (macro input option #11).

TXT: Output will be saved as a TXT file in all SAS versions. No output will be displayed in the OUTPUT window. All graphic files will be saved in the EMF format in version 8.0 or later or in the CGM format in pre-8.0 versions in the user-specified folder (macro input option #11).

13. **Macro-call parameter:** Transforming predictor variables (optional statement).

Descriptions and explanation: Performing a log scale or *z* (0, mean; 1, standard deviation) transformation on all the predictor variables reduces the impact of between-group unequal variance–covariance problems or differential scale of measurement.

Options/examples:

Blank: No transformation is performed; the original predictor variables will be used in discriminant analysis.

LOG: All predictor variables (non-zero values) will be transformed to natural log scale using the SAS LOG function. All types of discriminant analysis will be performed on log-transformed predictor variables.

STD: All predictor variables will be standardized to 0 mean and unit standard deviation using the SAS PROC STANDARD. All types of discriminant analysis will be performed on standardized predictor variables.

Note: System messages are deleted from the LOG window if DISPLAY is not selected as the input.

6.10 Decision Tree Using SAS Macro CHAID

The CHAID macro is a powerful SAS application for performing classification models based on decision trees. Options are available for creating decision tree diagrams and classification plots and for validating the decision tree models using independent validation datasets. This CHAID macro is an enhanced version of the SAS TREEDISC macro. The CHAID macro requires the SAS/IML product. To draw the tree, the SAS/OR product is required, and release 8.0 or later is recommended. In addition to these SAS procedures, GCHART is also utilized in the CHAID macro to obtain classification charts.

The CHAID macro generates an SAS dataset that describes a decision tree computed from a training dataset to predict a specified categorical response variable from one or more predictor variables. The tree can be listed, drawn, or used to generate code for an SAS DATA step to classify observations. Also, the classification results between the observed and the predicted levels for the training and the independent validation data can be viewed in a donut chart display.

The advantages of using the CHAID macro over the TREEDISC macro are:

- The CHAID macro has a user-friendly front end; SAS programming or SAS macro experience is not required to run the CHAID macro.
- Options for creating donut charts illustrating the differences between the observed and the predicted group memberships are available.
- Option for validating the fitted CHAID model obtained from a training dataset using an independent validation dataset by comparing classification errors are available.
- Options for saving the output tables and graphics in WORD, HTML, PDF, and TXT formats are available.

Software requirements for using the CHAID macro:

- SAS/CORE, SAS/BASE, SAS/IML, SAS/GRAPH, and optional SAS/OR must be licensed and installed at the site. Also, the XMACRO.SAS file (supplied by SAS) should be available for execution while running the CHAID macro.
- SAS version 8.0 or above is recommended for full utilization.
- An active Internet connection is required for downloading the CHAID macro and XMACRO.SAS file from the book website if the companion CD-ROM is not available.

6.10.1 *Steps Involved in Running the CHAID Macro*

1. Create an SAS dataset (permanent or temporary) containing at least one categorical response (target) variable and many continuous and/or categorical predictor (input) variables. (Disabling the SAS ENHANCED EDITOR is highly recommended in the latest SAS versions; open only the PROGRAM EDITOR window during execution of this macro.)
2. *If the companion CD-ROM is not available*, first verify that the Internet connection is active. Open the CHAID.sas macro-call file in the SAS PROGRAM EDITOR window. Instructions are given in the Appendix regarding downloading the macro-call and sample data files from the book website. *If the companion CD-ROM is available*, the CHAID.sas macro-call file can be found in the mac-call folder in the CD-ROM. Open the CHAID.sas macro-call file in the SAS PROGRAM EDITOR window. Click the RUN icon to submit the macro-call file CHAID.sas to open the macro-call window called CHAID (Figure 6.20).
3. Input the appropriate parameters in the macro-call window by following the instructions provided in the CHAID macro help file in Section 6.10.2. After inputting all the required macro parameters, be sure that the cursor is in the last input field and the RESULTS VIEWER window is closed, then hit the ENTER key (not the RUN icon) to submit the macro.
4. Examine the LOG window (only in DISPLAY mode) for any macro execution errors. If any errors are reported in the LOG window, activate the PROGRAM EDITOR window, resubmit the CHAID.sas macro-call file, check the macro input values, and correct if you see any input errors.
5. If no errors are found in the LOG window, activate the PROGRAM EDITOR window, resubmit the CHAID.sas macro-call file, and change the macro input value from DISPLAY to any other desirable format (see Section 6.10.2). The SAS output files from complete CHAID modeling and diagnostic graphs can be saved as user-specified-format files in the user-specified folder.

6.10.2 *Help File for SAS Macro CHAID*

1. **Macro-call parameter:** Input SAS dataset name (required parameter).
   Descriptions and explanation: Include the name of the temporary (member_name) or permanent (libname.member_name) SAS dataset on which the CHAID analysis will be performed.

Options/examples:

Permanent SAS dataset: gf.diabetic1 (LIBNAME: gf; SAS dataset name: diabetic1)

Temporary SAS dataset: diabetic1 (SAS dataset name)

2. **Macro-call parameter:** Input group response variable name (required parameter).

   Descriptions and explanation: Input the categorical response variable name from the SAS dataset that is to model as the target variable.

   Options/examples:

   group (name of the categorical response group)

3. **Macro-call parameter:** Input nominal predictor variables? (optional statement)

   Descriptions and explanation: Include categorical variables from the SAS dataset as predictors in CHAID modeling.

   Options/examples:

   TSTPLGP1 FASTPLGP

   Blank: Categorical predictors are not used.

4. **Macro-call parameter:** Input ordinal predictor variable names (optional statement).

   Descriptions and explanation: Include continuous variables from the SAS dataset as predictors in CHAID modeling.

   Options/examples:

   X1 X2 X3

5. **Macro-call parameter:** Input validation dataset name (optional parameter).

   Descriptions and explanation: If you would like to validate the CHAID model obtained from a training dataset by using an independent validation dataset, input the name of the SAS validation dataset. This macro estimates classification error for the validation dataset using the model estimates derived from the training data. Input the name of the temporary (member name) or permanent (libname.member_name) SAS dataset to be treated as the validation data.

   Options/examples:

   Permanent SAS dataset: gf.diabetic2 (LIBNAME: gf; SAS dataset name: diabetic2)

   Temporary SAS dataset: diabetic2 (SAS dataset name)

6. **Macro-call parameter:** Specify the folder containing the required XMACRO.sas file (optional statement).

   Descriptions and explanation: The XMACRO.sas file provided by SAS is required to run the CHAID macro. Input full path and filename of the XMACRO.SAS if running from the companion CD-ROM.

Options/examples:

> **BLANK** (running the macro from the book website): Leave this field blank; the CHAID macro can access the XMACRO.sas file directly from the book website.
>
> **E:\macro\xmacro.sas** (running the macro from the companion CD-ROM)

Otherwise, specify the location of the folder containing the XMACRO.sas file. A copy of the XMACRO.sas file is saved along with the SAS macros in the CD-ROM.

7. **Macro-call parameter:** Folder to save SAS graphics and output files (optional statement).

   Descriptions and explanation: To save the SAS graphics files in the EMF format suitable for inclusion in PowerPoint presentations, specify the output format as TXT in version 8.0 or later. In pre-8.0 versions, all graphic format files will be saved in a user-specified folder. Similarly, output files in the WORD, HTML, PDF, and TXT formats will be saved in the user-specified folder. If this macro field is left blank, the graphics and output files will be saved in the default folder.

   Option/example:

   > c:\output\ — folder named "OUTPUT"

8. **Macro-call parameter:** *z*th number of analysis (required statement).

   Descriptions and explanation: SAS output files will be saved by forming a file name from the original SAS dataset name and the counter value provided in this field. For example, if the original SAS dataset name is "train" and the counter number included is 1, the SAS output files will be saved as "train1" in the user-specified folder. By changing the counter value, users can avoid replacing previous SAS output files with new outputs.

9. **Macro-call parameter:** Display or save SAS output and graphs (required statement).

   Descriptions and explanation: Option for displaying all output files in the OUTPUT window or saving them as a specific format in a folder specified in macro input option #7.

   Options/examples: See Section 6.9.2 for an explanation of these formats.

   Possible values

   > DISPLAY
   > WORD
   > WEB
   > PDF
   > TXT

   Note: System messages are deleted from the LOG window if DISPLAY is not selected as the input.

6.11 Case Study 1: CDA and Parametric DFA

The objective of this study is to discriminate three clinical diabetes groups (normal, overt diabetic, and chemical diabetic) in simulated data using blood plasma and insulin measures**.** The simulated data satisfy the multivariate normality assumption and were generated using the group means and group variance–covariance estimates of real clinical diabetes data reported elsewhere.[11,17] The real diabetes data were used as the validation dataset to check the validity of the discriminant functions derived from the simulated data.

6.11.1 Study Objectives

1. **Data exploration using diagnostic plots:** Used to check for the discriminative potential of predictor variables two at a time in simple scatterplots. These plots are useful in examining the characteristics (scale of measurements, range of variability, extreme values) of predictor variables.
2. **Variable selection using stepwise selection methods:** Used to perform backward elimination, stepwise, and forward selection methods to identify predictor variables that have significant discriminating potentials.
3. **Checking for any violations of discriminant analysis assumptions:** Used to perform statistical tests and graphical analyses to verify that no multivariate influential outliers are present and the data satisfy multivariate normality assumption. The validity of canonical discriminant and parametric discriminant analyses results depends on satisfying the multivariate normality assumption.
4. **Parametric discriminant analyses:** Used to perform canonical discriminant analysis to investigate the characteristics of significant discriminating predictor variables, perform parametric discriminant function analysis to develop classification functions, to assign observations into predefined group levels, and to measure the success of discrimination by comparing the classification error rates.
5. **Save "plotp", "stat", and "out2" datasets for future use:** These three temporary SAS datasets are created and saved in the work folder when running the DISCRIM macro. The "plotp" dataset contains the observed predictor variables, group response value, canonical discriminant function scores, posterior probability scores, and new classification results. This canonical discriminant function and the posterior probability scores for each observation in the dataset could be used as the base for developing the scorecards. The temporary SAS data called "stat" contains the

canonical and discriminant function analysis parameter estimates. If an independent validation dataset is included, the classification results for the validation dataset are saved as a temporary SAS data called “out2”, which can be used to develop scorecards for new members.

6. **Validation:** Validate the derived discriminant functions by applying these classification criteria to an independent validation dataset (real diabetes dataset) and examine the success of classification.

6.11.2 Data Descriptions

Dataset names	Training (simulated dataset): permanent SAS dataset “diabetic1” located in the library “gf” Validation (real dataset): permanent SAS dataset “diabetic2” located in the library “gf”
Group response variables	Three clinical diabetic groups: 1, normal; 2, overt diabetic; 3, chemical diabetic. Use numeric values for group levels to generate box plots of posterior probability density estimates.
Predictor variables (X)	X1: relative weight X2: fasting plasma glucose level X3: test plasma glucose X4: plasma insulin during test X5: steady-state plasma glucose level
Number of observations	Training data: 141 Validation data: 145
Source	Training data: simulated data that satisfy multivariate normality assumption Validation data: real diabetes data[11,17]

Open the DISCRIM.sas macro-call file in the SAS PROGRAM EDITOR window and click RUN to open the DISCRIM macro-call window (Figure 6.1). Input the appropriate macro-input values by following the suggestions given in the help file (see Section 6.9.2).

6.11.3 Exploratory Analysis/Diagnostic Plots

Input the dataset name, group variable, predictor variable names, and prior probability option. Input YES in macro input field #2 to perform data exploration and to create diagnostic plots. Submit the DISCRIM macro to produce discriminant diagnostic plots and stepwise variable selection output.

6.11.4 Data Exploration and Checking

Examining the group discrimination based on simple scatterplots between any two discrimination variables is the first step in data exploring. An example of simple two-dimensional scatterplots showing the discrimination of three diabetes groups is presented in Figure 6.2. These scatterplots are useful in examining the range of variation and degree of linear associations between any two predictor variables. The scatterplot presented in Figure 6.2 reveals that a strong correlation exists between fasting plasma glucose level (X2) and test plasma glucose (X3). These two attributes appeared to discriminate the diabetes group 3 from the other two to a certain degree. Discrimination between the normal and the overt diabetes group is not very distinct.

6.11.5 Variable Selection Methods

Variable selection based on backward elimination, stepwise selection, and forward selections is performed automatically when YES for data exploration is input in the DISCRIM macro-call window. These variable selection methods are especially useful when it is necessary to screen a large number of predictor variables. Information on the number of observations, number of group levels, discriminating variables, and threshold significance level for eliminating nonsignificant predictor variables in the backward elimination method is presented in Table 6.1. About 50% of the cases in the training dataset appear to be in the normal group. Among the clinically diagnosed diabetes group, 25% of them are considered overt type, and 23% are in the chemical diabetes group (Table 6.2).

The results of variable selection based on backward elimination are summarized in Table 6.3. Backward elimination starts with the full model, and the overall significance in discriminating the diabetes groups is highly significant based on the *p* value for Wilks' lambda (<0.0001) and the *p* value for the average squared canonical correlations (<0.0001). In step 1, the nonsignificant (*p* value, 0.28) steady-state plasma glucose (X5) is eliminated from the model, and the resulting four-variable model

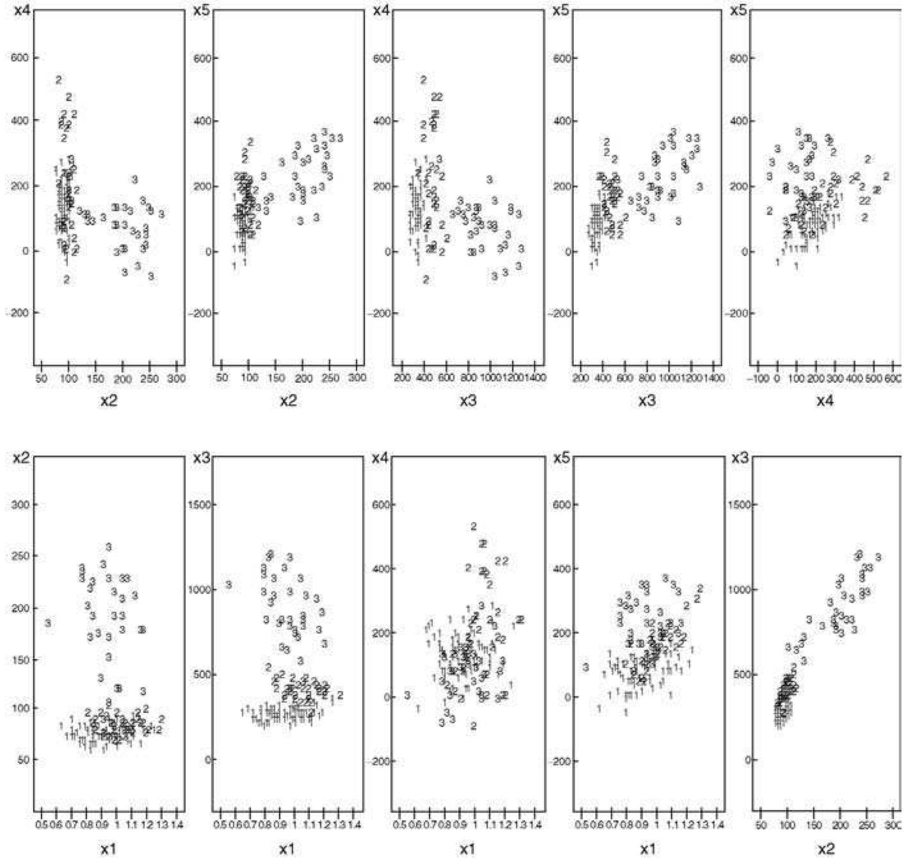


Figure 6.2 Bivariate exploratory plots generated using the SAS macro DISCRIM: group discrimination of three types of diabetic groups (data = diabetic1) in simple scatterplots.

Table 6.1 Data Exploration Using SAS Macro DISCRIM: Variable Selection Using Backward Elimination Method

Observations	141
Class levels	3
Variables in the analysis	5
Variables that will be included	0
Significance level to stay	0.15

Table 6.2 Data Exploration Using SAS Macro DISCRIM: Group Level Information

Group	Variable Name	Frequency	Weight	Proportion
1	_1	72	72.0000	0.510638
2	_2	36	36.0000	0.255319
3	_3	33	33.0000	0.234043

Table 6.3 Data Exploration Using SAS Macro DISCRIM: Backward Elimination Summary

Step	Variable Removed	Label	Partial R^2	F Value	Pr > F	Wilks' Lambda	Pr < Lambda	ASCC	PR> ASCC
0	—	—	—	—	—	0.08032600	<.0001	0.611	<.0001
1	X5	Steady-state plasma glucose	0.0187	1.28	0.2820	0.08185813	<.0001	0.605	<.0001

Note: ASCC = average squared canonical correlation.

is equally as good as the full model. The backward elimination method is stopped here because no other variable can be dropped based on the *p*-value stay criterion (0.15). In the backward elimination method, once a variable is removed from the discriminant model, it cannot be re-entered.

Information on the number of observations, number of group levels, discriminating variables and the threshold significance level for entering (0.15) and staying (0.15) in the stepwise selection method is presented in Table 6.4. The significance of predictor variables in discriminating the

Table 6.4 Data Exploration Using SAS MACRO DISCRIM: Variable Selection Using Stepwise Selection Method

Observations	141
Class levels	3
Variables in the analysis	5
Variables that will be included	0
Significance level to enter	0.15
Significance level to stay	0.15

three clinical diabetes groups is evaluated in a stepwise fashion. At each step, the significance of already entered predictor variables is evaluated based on the significance for staying criterion (*p* value, 0.15), and the significance of newly entering variables is evaluated based on the significance for entering criterion (*p* value. 0.05). The stepwise selection procedure stops when no variables can be removed or entered. The summary results of the stepwise selection method are presented in Table 6.5. The results of the stepwise selection methods are in agreement with the backward elimination method because both methods choose variables X1 to X4 as the significant predictors.

Information on the number of observations, number of group levels, discriminating variables, and the threshold significance level for entering (0.15) in the forward selection method is presented in Table 6.6. The significance of predictor variables in discriminating the three clinical diabetes groups is evaluated one at a time. At each step, the significance of entering variables is evaluated based on the significance for entering criterion (*p* value, 0.15). The forward selection procedure stops when no variables can be entered by the entering *p* value criterion (0.15). In the forward selection method, once a variable is entered into the discriminant model, it cannot be removed. The summary results of the forward selection method are presented in Table 6.7. The results of the forward selection method are in agreement with both the backward elimination and the stepwise selection methods because all three methods choose variables X1 to X4 as the significant predictors.

6.11.6 Discriminant Analysis and Checking For Multivariate Normality

Open the DISCRIM macro-call window, and input the dataset name, group variable, predictor variable names, and prior probability option. Leave macro input field #2 blank to perform CDA and parametric DFA; input YES in macro input field #4 to perform the multivariate normality check. Submit the DISCRIM macro to get the multivariate normality check, CDA output, graphics, and parametric DFA output and graphics.

6.11.6.1 Checking for Multivariate Normality

The correct choice for selecting parametric vs. nonparametric discriminant analysis is dependent on the assumption of multivariate normality within each group. The diabetes data within each clinical group are assumed to have a multivariate normal distribution. This multivariate normality assumption can be checked by estimating multivariate skewness, kurtosis,

Table 6.5 Data Exploration Using SAS MACRO DISCRIM: Stepwise Selection Summary

Step	*Number In*	*Entered*	*Removed*	*Label*	*Partial R^2*	*F Value*	*Pr > F*	*Wilks' Lambda*	*Pr < Lambda*	*Average Squared Canonical Correlation (ASCC)*	*Pr > ASCC*
1	1	X3	—	Test plasma glucose	0.8700	461.68	<.0001	0.13002152	<.0001	0.43498924	<.0001
2	2	X2	—	Fasting plasma glucose	0.1943	16.52	<.0001	0.10475877	<.0001	0.53015812	<.0001
3	3	X1	—	Relative weight	0.1689	13.82	<.0001	0.08706366	<.0001	0.58491556	<.0001
4	4	X4	—	Plasma insulin during test	0.0598	4.29	0.0156	0.08185813	<.0001	0.60571448	<.0001

Table 6.6 Data Exploration Using SAS MACRO DISCRIM: Variable Selection Using Forward Selection Method

Observations	141
Class levels	3
Variables in the analysis	5
Variables that will be included	0
Significance level to enter	0.15

Table 6.7 Data Exploration Using SAS Macro DISCRIM: Forward Selection Summary

Step	*Number In*	*Entered*	*Label*	*Partial R^2*	*F Value*	*Pr > F*	*Wilks' Lambda*	*Pr < Lambda*	*Average Squared Canonical Correlation (ASCC)*	*Pr > ASCC*
1	1	X3	Test plasma glucose	0.8700	461.68	<.0001	0.13002152	<.0001	0.43498924	<.0001
2	2	X2	Fasting plasma glucose	0.1943	16.52	<.0001	0.10475877	<.0001	0.53015812	<.0001
3	3	X1	Relative weight	0.1689	13.82	<.0001	0.08706366	<.0001	0.58491556	<.0001
4	4	X4	Plasma insulin during test	0.0598	4.29	0.0156	0.08185813	<.0001	0.60571448	<.0001

and testing for their significance levels. The quantile–quantile (Q–Q) plot of expected and observed distributions[9] of multi-attribute residuals after adjusting for the group means can be used to graphically examine for multivariate normality. The estimated multivariate skewness (0.772; *p* value, 0.578) and multivariate kurtosis (23.847; *p* value, 0.895) (Figure 6.3) clearly support the hypothesis that, after adjusting for the group differences, these four multi-attributes have a joint multivariate normal distribution. A nonsignificant departure from the 45° angle reference line in the Q–Q plot (Figure 6.3) also supports this finding. Thus, parametric discriminant analysis can be considered to be the appropriate technique for discriminating the three clinical groups based on these four attributes (X1 to X4).

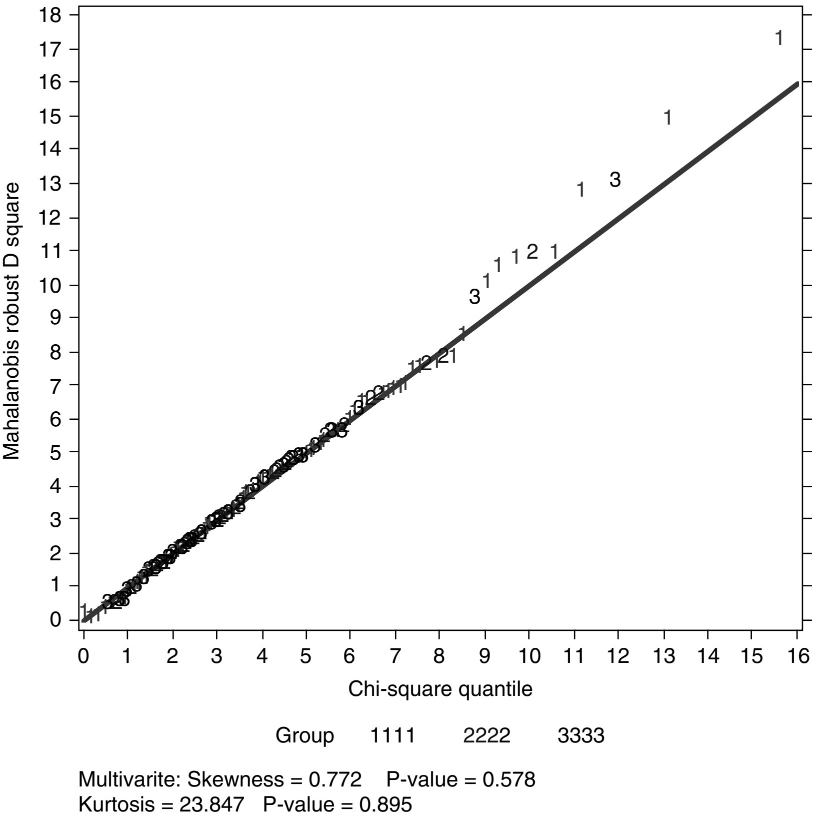


Figure 6.3 Checking for multivariate normality (data = diabetic1) in a Q–Q plot using the SAS macro DISCRIM.

6.11.6.2 Checking for the Presence of Multivariate Outliers

Multivariate outliers can be detected in a plot between the differences of robust (Mahalanobis distance – chi-squared quantile) vs. chi-squared quantile values.[9] No observations are identified as influential observations because the differences between the robust Mahalanobis distance and the chi-squared quantile value are not larger than 2 and fall inside the critical region (Figure 6.4). This can be expected, as the training dataset used is a multivariate, normally distributed, simulated dataset.

6.11.7 Canonical Discriminant Analysis

The main objective of CDA is to extract a set of linear combinations of the quantitative variables that best reveal the differences among the groups. The class level information, group frequency, and prior probability values

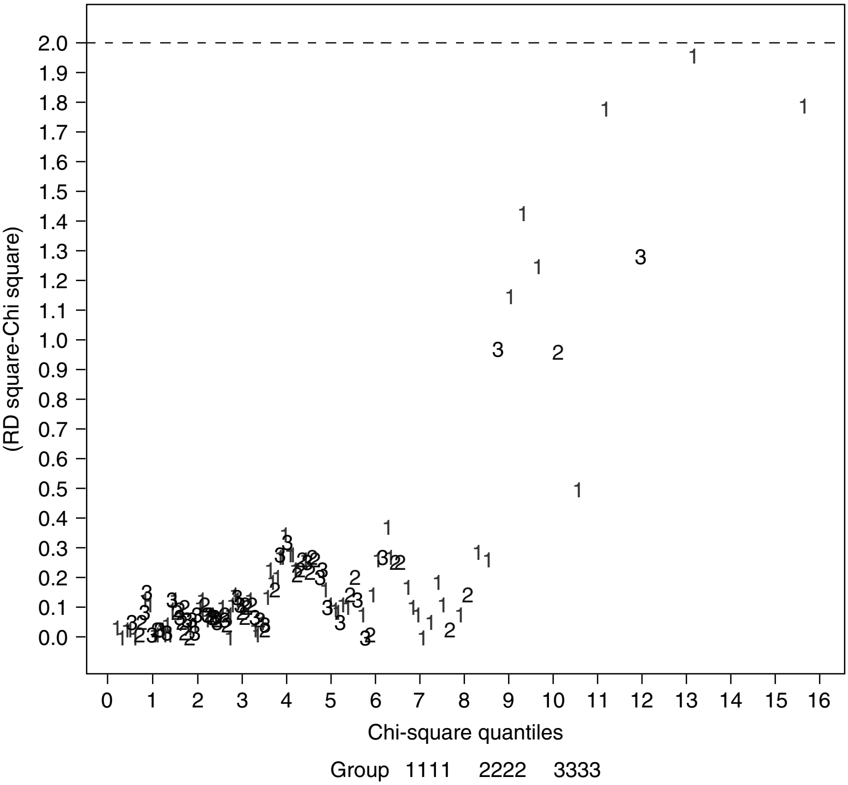


Figure 6.4 Diagnostic plot for detecting multivariate influential observations (data = diabetic1) generated using the SAS macro DISCRIM.

Table 6.8 Canonical Discriminant Analysis Using SAS Macro DISCRIM: Group Level Information

Group	Variable Name	Frequency	Weight	Proportion	Prior Probability
1	_1	72	72.0000	0.510638	0.510638
2	_2	36	36.0000	0.255319	0.255319
3	_3	33	33.0000	0.234043	0.234043

Observations: 141; degrees of freedom: 140.

for training dataset "diabetes1" are presented in Table 6.8. The descriptive statistical measures, sample size, mean, variance, standard deviation for the whole dataset (Table 6.9), and the within-each-group levels are presented in Table 6.10. Very large differences in the means and variances are observed for X2, X3, and X4, among the three levels of diabetes groups. Therefore, in CDA, it is customary to standardize the multi-attributes so that the canonical variables have means that are equal to 0 and the pooled within-class variances are equal to 1.

The univariate ANOVA results indicate that highly significant group differences exist for all the predictor variables (Table 6.11). The total-, between-, and within-group variability in predictor variables is expressed in standard deviation terms. The R^2 statistic describes the amount of variability in each predictor variable accounted for by the group differences. The $R^2/(1 - R^2)$ column expresses the ratio between accounted and

Table 6.9 Canonical Discriminant Analysis Using SAS Macro DISCRIM: Total Sample Statistics

Variable	Label	N	Sum	Mean	Variance	Standard Deviation
X1	Relative weight	141	135.19103	0.95880	0.02034	0.1426
X2	Fasting plasma glucose	141	16565	117.48094	2531	50.3132
X3	Test plasma glucose	141	75170	533.12341	70856	266.1886
X4	Plasma insulin during test	141	25895	183.65149	12431	111.4925

Table 6.10 Canonical Discriminant Analysis Using SAS Macro DISCRIM: Within-Group Statistics

Variable	*Label*	*N*	*Sum*	*Mean*	*Variance*	*Standard Deviation*
Group = 1						
X1	Relative weight	72	66.50201	0.92364	0.01938	0.1392
X2	Fasting plasma glucose	72	6487	90.09403	58.58411	7.6540
X3	Test plasma glucose	72	25463	353.65613	1038	32.2236
X4	Plasma insulin during test	72	12767	177.32023	4736	68.8154
Group = 2						
X1	Relative weight	36	37.52763	1.04243	0.01513	0.1230
X2	Fasting plasma glucose	36	3551	98.64561	80.37419	8.9652
X3	Test plasma glucose	36	17749	493.02776	2866	53.5392
X4	Plasma insulin during test	36	9123	253.40526	25895	160.9207
Group = 3						
X1	Relative weight	33	31.16139	0.94428	0.01856	0.1363
X2	Fasting plasma glucose	33	6527	197.78183	2121	46.0502
X3	Test plasma glucose	33	31958	968.42910	34867	186.7276
X4	Plasma insulin during test	33	4005	121.37012	5989	77.3915

Table 6.11 Canonical Discriminant Analysis Using SAS Macro DISCRIM: Univariate Test Statistics

Variable	*Label*	*Total Standard Deviation*	*Pooled Standard Deviation*	*Between Standard Deviation*	*R^2*	*$R^2/(1 - R^2)$*	*F Value*	*Pr > F*
X1	Relative weight	0.1426	0.1346	0.0608	0.1221	0.1391	9.60	0.0001
X2	Fasting plasma glucose	50.3132	23.2865	54.5354	0.7888	3.7359	257.78	<.0001
X3	Test plasma glucose	266.1886	96.6766	303.0008	0.8700	6.6910	461.68	<.0001
X4	Plasma insulin during test	111.4925	101.9459	57.0604	0.1759	0.2134	14.72	<.0001

Note: F statistics: numerator df = 2; denominator df = 138; and R^2 weighted by variance = 0.767.

unaccounted for variation in the univariate ANOVA model. By comparing the R^2 and the $R^2/(1 - R^2)$ statistics for each significant predictor variable, we can conclude that the test plasma glucose (X3) has the highest amount of significant discriminative potential, while the relative weight has the least amount of discriminative power in differentiating the three diabetes clinical groups. The relatively large average R^2 weighted by variances (Table 6.11) indicates that the four predictor variables have high discriminatory power in classifying the three clinical diabetes groups. This is further confirmed by the highly significant MANOVA test results for all four criteria (Table 6.12). If the MANOVA assumptions, multivariate normality, and equal variance–covariance are not met, the validity of the MONOVA test is questionable. Transforming the predictor variables into a log scale might alleviate the problems caused by the unequal variance–covariance between the groups.

In CDA, canonical variables that have the highest possible multiple correlations with the groups are extracted. The unstandardized coefficients used in computing the raw canonical variables are called the *canonical coefficients* or *canonical weights*. The standardized discriminant function coefficients indicate the partial contribution of each variable to the discriminant functions, controlling for other attributes entered in the equation. The total standardized discriminant functions given in Table 6.13 indicate that the predictor variable, test plasma glucose (X3), contributed significantly to the first canonical variable (CAN1). The fasting plasma glucose (X2) and test plasma glucose (X3) contributed equally in a negative way to the second canonical variable (CAN2).

Table 6.12 Canonical Discriminant Analysis Using SAS Macro DISCRIM: Multivariate ANOVA Statistics

Statistic	*Value*	*F Value*	*Numerator Degrees of Freedom*	*Denominator Degrees of Freedom*	*Pr > F*
Wilks′ lambda	0.08185813	84.21	8	270	<.0001
Pillai's trace	1.21142895	52.23	8	272	<.0001
Hotelling–Lawley trace	7.63338735	128.25	8	190.55	<.0001
Roy's greatest root	7.13094774	242.45	4	136	<.0001

Table 6.13 Canonical discriminant analysis using SAS MACRO: DISCRIM- Total sample standardized canonical coefficients

Variable	*Label*	*CAN1*	*CAN2*
X1	Relative weight	0.249225407	0.407515174
X2	Fasting plasma glucose	0.412500088	2.540795333
X3	Test plasma glucose	3.223110774	2.387172078
X4	Plasma insulin during test	0.040224046	0.319253611

These canonical variables are independent or orthogonal to each other; that is, their contributions to the discrimination between groups will not overlap. This maximal multiple correlation between the first canonical variable and the group variables is called the first canonical correlation. The second canonical correlation is obtained by finding the linear combination uncorrelated with the CAN1 that has the highest possible multiple correlations with the groups. In CDA, the process of extracting canonical variables is repeated until the maximum number of canonical variables has been extracted, equal to the number of groups minus one or the number of variables in the analysis, whichever is smaller.

The correlation between the CAN1 and the clinical group is very high (>0.9), and about 87% of the variation in the first canonical variable can be attributed to the differences among the three clinical groups (Table 6.14). The first eigenvalue measures the variability in the CAN1 and accounts for 93% of the variability among the three group members in four predictor variables. The correlation between the CAN2 and the clinical group is moderate (0.5), and about 37% of the variation in the second canonical variable can be attributed to the differences among the three clinical groups (Table 6.14). The second eigenvalue measures the variability in the second canonical variable and accounts for the remaining 6% of the variability among the three group members in the four predictor variables. Both canonical variables are statistically highly significant based on the Wilks' lambda test (Table 6.15); however, the statistical validity might be questionable if the multivariate normality or the equal variance–covariance assumptions are violated.

The total structure coefficients or loadings measure the simple correlations between the predictor variables and the canonical variable. These structure loadings are commonly used when interpreting the meaning of the canonical variable because the structure loadings appear to be more stable, and they allow for interpretation of the canonical variable in a manner that is analogous to factor analysis. The structure loadings for the first two canonical variables are presented in Table 6.16. The first and the second canonical functions account for 93% and 6% of the variation,

Table 6.14 Canonical Discriminant Analysis Using SAS Macro DISCRIM: Canonical Correlations

					*Eigenvalues of Inv(E)*H = $CANR^2/(1 - CANR^2)$*			
		Adjusted	*Approximate*	*Squared*				
	Canonical	*Canonical*	*Standard*	*Canonical*				
	Correlation	*Correlation*	*Error*	*Correlation*	*Eigenvalue*	*Difference*	*Proportion*	*Cumulative*
1	0.936490	0.934954	0.010394	0.877013	7.1309	6.6285	0.9342	0.9342
2	0.578287	0.572064	0.056252	0.334416	0.5024		0.0658	1.0000

Table 6.15 Canonical Discriminant Analysis Using SAS Macro DISCRIM: Testing the Canonical Correlations in the Current Row (All That Follow Are Zero)

	Likelihood Ratio	*Approximate F Value*	*Numerator Degrees of Freedom*	*Denominator Degrees of Freedom*	*Pr > F*
1	0.08185813	84.21	8	270	<.0001
2	0.66558416	22.78	3	136	<.0001

Table 6.16 Canonical Discriminant Analysis Using SAS Macro DISCRIM: Total Canonical Structure Loadings

Variable	*Label*	*CAN1*	*CAN2*
X1	Relative weight	0.049197	0.599026
X2	Fasting plasma glucose	0.926528	–0.327968
X3	Test plasma glucose	0.994192	–0.096635
X4	Plasma insulin during test	–0.229076	0.623110

respectively, in the discriminating variables (Table 6.14). Two variables, fasting plasma glucose and test plasma glucose, have very large positive loadings on CAN. The predictor variables' relative weight (X1) and plasma insulin during test (X4) have moderate size loadings on CAN2; therefore, the CAN1 can be the plasma glucose factor. The standardized means of three clinical groups on CAN1 and CAN2 are presented in Table 6.17. The mean CAN1 and CAN2 for the normal group is relatively lower than for the other two diabetic groups. The mean CAN1 for the chemical diabetic group is very high, and the overt group has a large mean for CAN2. Thus, in general, these two canonical variables successfully discriminate the three diabetic groups.

Table 6.17 Canonical Discriminant Analysis Using SAS Macro DISCRIM: Group Means on Canonical Variables

Group	*CAN1*	*CAN2*
1	–2.017022608	–0.429657716
2	–0.158292155	1.196872445
3	4.573458950	–0.368244013

For each observation in the training dataset, we can compute standardized canonical variable scores. The box plot of the CAN1 score by group is very useful in visualizing the group differences based on the CAN1. The CAN1 score effectively discriminates the three groups (Figure 6.5). The chemical diabetic group (3) has a relatively larger variation for the CAN1 score than the other two groups.

These standardized canonical variable scores and the structure loadings can be used in two-dimensional bi-plots to aid visual interpretation of the group differences. Interrelationships among the four multi-attributes and discriminations of the three groups are presented in Figure 6.6. The first canonical variable that has the largest loadings on X2 and X3 discriminated the normal (1), overt (2), and chemical diabetic groups (3) effectively. The CAN2, which has a moderate-size loading on X1 and X4, discriminated the normal group (1) and the overt group (2), but CAN2 is not effective

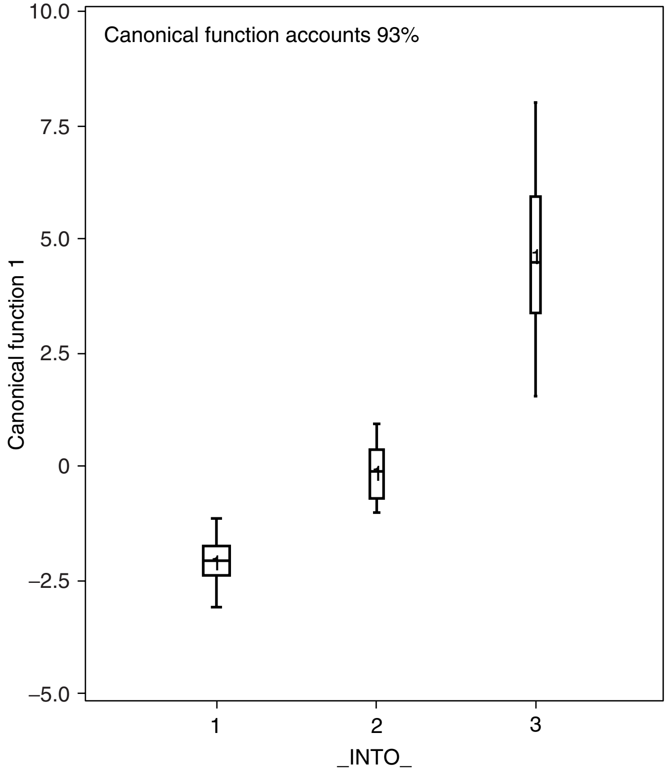


Figure 6.5 Group discrimination using canonical discriminant function1 generated using the SAS macro DISCRIM.

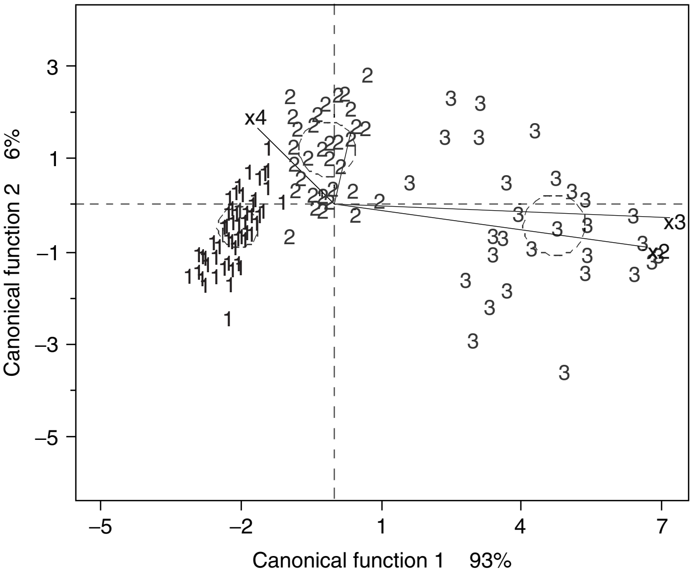


Figure 6.6 Bi-plot display of canonical discriminant functions and structure loadings generated using the SAS macro DISCRIM.

in separating the chemical diabetic group (3). The narrow angle between the X2 and X3 variable vector in the same direction indicates that the plasma glucose variables are positively highly correlated. The correlations between X1 and X4 are moderate in size and act in the opposite direction from the plasma glucose variables.

The two canonical variables extracted from the CDA effectively discriminated the three clinical diabetic groups. The difference between the normal and the chemical group is distinct. The discrimination between the normal group (1) and the overt group (2) is effective when both the CAN1 and CAN2 are used simultaneously; therefore, the CDA can be considered as an effective descriptive tool in discriminating groups based on continuous predictor variables. If the variance–covariance between the groups is assumed to be equal and the predictor variables have joint multivariate normal distributions within each group, then the group differences can be tested statistically for group differences.

6.11.8 Discriminant Function Analysis

Discriminant function analysis is commonly used for classifying observations in predefined groups based on knowledge of their quantitative attributes. Because the distribution within each diabetes group is found to be multivariate normal (Figure 6.5), a parametric method can be used to develop a discriminant function using a measure of their generalized squared distance. The discriminant function, also known as a classification criterion, is estimated by measuring the generalized squared distance.[8] The generalized square distance squares among the three group means are presented in Table 6.18. The generalized squared distances between the normal group and the diabetes groups are farther apart than the distances between the two diabetes groups.

The classification criterion in DFA is based on the measure of squared distance and the prior probability estimates. Either the individual within-group covariance matrices (a quadratic function) or the pooled covariance matrix (a linear function) can be used in deriving the squared distance. Because the chi-square value for testing the heterogeneity of between variance–covariance is significant at the 0.1 level (Table 6.19), the within-covariance matrices are used in computing the quadratic discriminant function.

When computing the classification criterion in this analysis, the prior probability proportional to the group frequency was used. The posterior probability of an observation belonging to each class is estimated and each observation is classified in the group from which it has the smallest generalized squared distance or larger posterior probability. The DISCRIM macro also outputs a table of the *i*th group posterior probability estimates for all observations in the training dataset. Table 6.20 provides a partial list of the *i*th group posterior probability estimates for some of the selected observations in the table. These posterior probability values are very useful estimates as they can be successfully used in developing scorecards and ranking the observations in the dataset.

The performance of a discriminant criterion in classification is evaluated by estimating the probabilities of misclassification or error rates. These

Table 6.18 Discriminant Function Analysis Using SAS Macro DISCRIM: Generalized Square Distance Measures between Group Levels

	To Group		
From Group	*1*	*2*	*3*
1	0	12.70950	16.67706
2	19.02148	0	7.21354
3	493.63117	131.23375	0

Table 6.19 Parametric Discriminant Analysis Using SAS Macro DISCRIM: Chi-Square Test for Heterogeneity of Variance–Covariance among Groups

Chi-Square	*Degrees of Freedom*	*Pr > Chi-Square*
313.500033	20	<.0001

Table 6.20 Parametric Discriminant Function Analysis Using SAS Macro DISCRIM: Posterior Probability Estimates by Groups in Cross Validation Using the Quadratic Discriminant Function

			Posterior Probability Estimates		
Observation	*From Group*	*Classified into Group*	*1*	*2*	*3*
1	1	1	0.9997	0.0003	0.0000
2	1	1	1.0000	0.0000	0.0000
3	1	1	1.0000	0.0000	0.0000
4	1	1	0.9968	0.0031	0.0001
5	1	1	0.9997	0.0003	0.0000
6	1	1	0.9995	0.0005	0.0000
7	1	1	1.0000	0.0000	0.0000
—[a]	—	—	—	—	—
141	3	3	0.0000	0.0000	1.0000
142	3	3	0.0000	0.0000	1.0000
143	3	3	0.0000	0.0000	1.0000
144	3	3	0.0000	0.0000	1.0000
145	3	3	0.0000	0.0000	1.0000

[a] Partial list.

error-rate estimates include error-count estimates and error-rate estimates. When the training dataset is a valid SAS dataset, the error rate can also be estimated by cross validation. In cross validation, (n – 1) out of n observations in the training sample are used. The DFA estimates the discriminant functions based on these (n – 1) observations and then applies them to classify the one observation left out. This cross validation is performed for each of the n training observations. The misclassification rate for each group is estimated from the proportion of sample observations in that group that are misclassified.

Table 6.21 lists the misclassified observations based on the posterior probability estimates computed by the quadratic discriminant function by

Table 6.21 Quadratic Discriminant Function Analysis Based on Cross Validation Using SAS Macro DISCRIM: Posterior Probability of Group Membership for Misclassified Cases

			Posterior Probability Estimates		
Observation	*From Group*	*Classified into Group*	*1*	*2*	*3*
5	1	2	0.1242	0.8731	0.0027
29	1	2	0.0540	0.9392	0.0068
61	1	2	0.1954	0.8008	0.0039
74	2	1	0.9698	0.0297	0.0006
120	3	2	0.0000	0.9991	0.0009

cross validation. Three cases that belong to the normal group (1) are classified into the overt group (2) because their posterior probability estimates are larger than 0.8 for the overt group. Furthermore, one case that belongs to the overt group (2) is classified into the normal group (1), and one case is switched from the chemical group (3) to the overt group (2).

Classification results based on the quadratic discriminant functions and error rates based on cross validation are presented in Table 6.22. The misclassification rates in groups 1, 2, and 3 are 4.1, 2.7, and 3.03%, respectively. The overall discrimination is quite satisfactory because the overall error rate is very low, at 3.5%. The overall error rate is estimated through a weighted average of the individual group-specific error-rate estimates, where the prior probabilities are used as the weights. The posterior probability estimates based on cross validation reduce both the bias and the variance of classification function. The resulting overall error estimates are intended to have both a low variance from using the posterior probability estimate and a low bias from cross validation.

A box-plot display of the *i*th posterior probability estimates by group is a powerful graphical tool to summarize the classification results, to detect false positives and negatives, and to investigate the variation in the *i*th posterior probability estimates. Figure 6.7 illustrates the variation in the posterior probability estimates for the normal group (1). The posterior probability estimates of a majority of the cases that belong to the normal group are larger than 0.9. Three observations (5, 29, 61) are identified as false negatives. One observation (74) that belongs to the overt group (2) is identified as false positive. Figure 6.8 displays the variation in the posterior probability estimates for the overt diabetes group (2). The posterior probability estimates of a majority of the cases that belong to the overt group (2) are larger than 0.85. One observation (74) is identified

Table 6.22 Parametric Discriminant Function Analysis Using SAS Macro DISCRIM: Classification Table and Error-Count Estimates by Groups in Cross Validation Using Quadratic Discriminant Functions

	To Group			
From Group	*1*	*2*	*3*	*Total*
1	69[a]	3	0	72
	95.83[b]	4.17	0.00	100.00
2	1	35	0	36
	2.78	97.22	0.00	100.00
3	0	1	32	33
	0.00	3.03	96.97	100.00
Total	70	39	32	141
	49.65	27.66	22.70	100.00
Priors	0.51064	0.25532	0.23404	
	Error-Count Estimates for Group			
Rate	0.0417	0.0278	0.0303	0.0355
Priors	0.5106	0.2553	0.2340	—

[a] Frequency.

[b] Percentage.

as a false negative. Three observations (5, 29, 61) that belong to the normal group (1) and one observation (120) that belongs to the chemical group (3) are identified as false positives. The variability in the posterior probability estimates for the chemical diabetes group (3) is illustrated in Figure 6.9. The posterior probability estimates of a majority of the cases that belong to the chemical group (3) are larger than 0.95. One observation (120) is identified as a false negative, while no observations are identified as false positives.

Although classification function by cross validation achieves a nearly unbiased overall error estimate, it has a relatively large variance. To reduce the variance in an error-count estimate, a smoothed error-rate estimate is suggested.[14] Instead of summing values that are either 0 or 1 as in the error-count estimation, the smoothed estimator uses a continuum of values between 0 and 1 in the terms that are summed. The resulting estimator has a smaller variance than the error-count estimate. The posterior probability error-rate estimates are smoothed error-rate estimates. The posterior probability error-rate estimates for each group are based on the posterior probabilities of the observations classified into that same group. The posterior probability estimates provide good estimates of the error rate when the posterior probabilities are accurate.

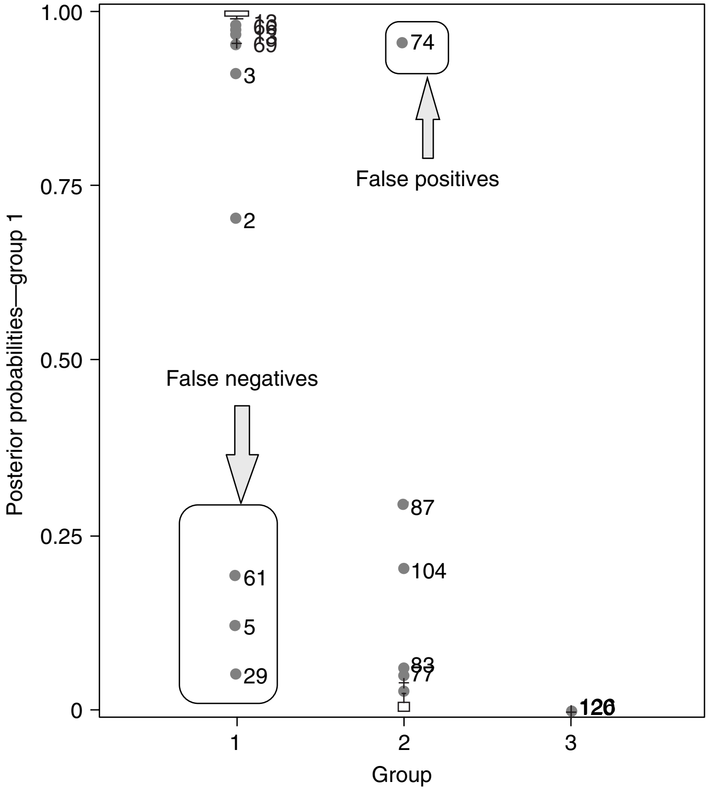


Figure 6.7 Box-plot display of posterior probability estimates for group level = 1 derived from parametric discriminant function analysis by cross validation. This plot is generated using the SAS macro DISCRIM.

When a parametric classification criterion (linear or quadratic discriminant function) is derived from a non-normal population, the resulting posterior probability error-rate estimators may not be appropriate.

The smoothed posterior probability error-rate estimates based on cross-validation quadratic discriminant functions are presented in Table 6.23. The overall error rate for stratified and unstratified estimates are equal because the prior probability option proportional to group frequency is used. The overall discrimination is quite satisfactory, as the overall error rate using the smoothed posterior probability error rate is very low, at 2.1%.

The classification function derived from the training dataset can be validated by classifying the observations in an independent validation dataset. If the classification error rate obtained for the validation data is

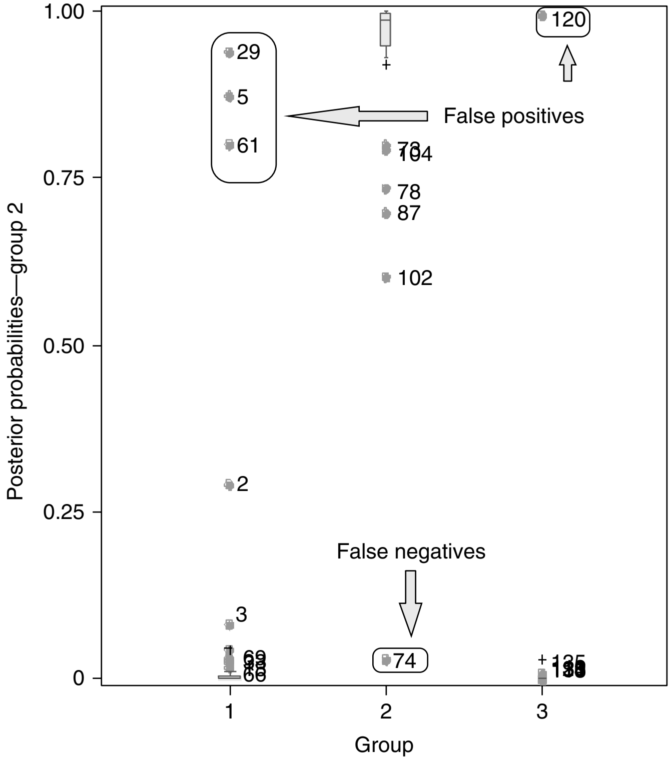


Figure 6.8 Box-plot display of posterior probability estimates for group level = 2 derived from parametric discriminant function analysis by cross validation. This plot is generated using the SAS macro DISCRIM.

small and similar to the classification error rate for the training data, then we can conclude that the derived classification function has good discriminative potential. Classification results for the validation dataset based on quadratic discriminant functions are presented in Table 6.24. The misclassification rates in groups 1, 2, and 3 are 1.3, 5.5, and 9.0%, respectively. The overall discrimination in the validation dataset is quite good, as the weighted error rate is very low, at 4.2%. A total of six observations in the validation dataset are misclassified (see Table 6.25). For example, observation number 75 is classified into the overt group (2) from the normal group (1). Because the training and validation datasets give comparable classification results based on parametric quadratic functions, we can conclude that blood plasma and insulin measures are quite effective in grouping the clinical diabetes.

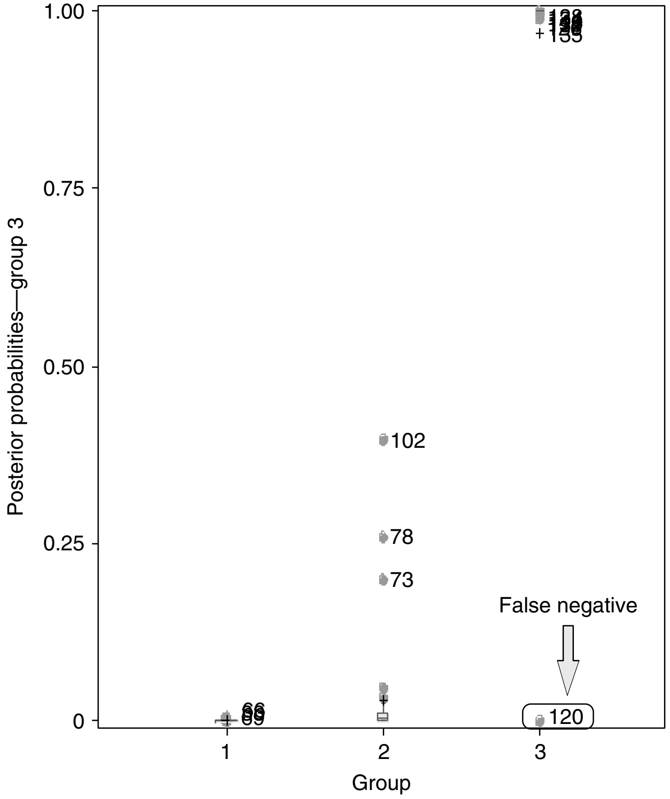


Figure 6.9 Box-plot display of posterior probability estimates for group level = 3 derived from parametric discriminant function analysis by cross validation. This plot is generated using the SAS macro DISCRIM.

6.12 Case Study 2: Nonparametric DFA

When the assumption regarding multivariate normality within each group level is not met, nonparametric DFA is considered to be a suitable classification technique. The diabetes dataset[11,17] used in Section 6.11 for validation, is used as the training dataset to derive the classification function using blood plasma and insulin measures. Because the multivariate normality assumption was not met for this diabetes dataset, it is a suitable candidate for performing nonparametric DFA. The simulated diabetes dataset used in Section 6.11 for training is used as the validation dataset.

Table 6.23 Parametric Discriminant Function Analysis Using SAS Macro DISCRIM: Classification Table and Posterior Probability Error-Rate Estimates by Groups in Cross Validation Using Quadratic Discriminant Functions

	To Group		
From Group	*1*	*2*	*3*
1	69[a]	3	0
	0.9902[b]	0.8710	—
2	1	35	0
	0.9698	0.9473	—
3	0	1	32
	—	0.9991	0.9997
Total	70	39	32
	0.9899	0.9428	0.9997
Priors	0.51064	0.25532	0.23404

	Posterior Probability Error Rate Estimates for Group			
Estimate	*1*	*2*	*3*	*Total*
Stratified	0.0376	–0.0213	0.0306	0.0209
Unstratified	0.0376	–0.0213	0.0306	0.0209
Priors	0.5106	0.2553	0.2340	—

[a] Frequency.

[b] Percentage.

6.12.1 Study Objectives

The objectives are to discriminate three clinical diabetic groups (normal, overt diabetic, and chemical diabetic) using blood plasma and insulin measures.

- **Checking for any violations of discriminant analysis assumptions:** Perform statistical tests and graphical analysis to detect multivariate influential outliers and departure from multivariate normality.
- **Nonparametric discriminant analyses:** Because this dataset significantly violates the multivariate normality assumption, nonparametric discriminant function analyses can be performed based on the nearest neighbor and kernel density methods. Classification functions are developed based on nonparametric posterior probability density estimates and observations are assigned into predefined group levels. This measures the success of discrimination by comparing the classification error rates.

Table 6.24 Parametric Discriminant Function Analysis Using SAS Macro DISCRIM: Classification Table and Error Count Estimates by Groups for Validation Data Using Quadratic Discriminant Functions

	To Group			
From Group	*1*	*2*	*3*	*Total*
1	75[a]	1	0	76
	98.68[b]	1.32	0.00	100.00
2	2	34	0	36
	5.56	94.44	0.00	100.00
3	0	3	30	33
	0.00	9.09	90.91	100.00
Total	77	38	30	145
	53.10	26.21	20.69	100.00
Priors	0.51064	0.25532	0.23404	
	Error-Count Estimates for Group			
	1	*2*	*3*	*Total*
Rate	0.0132	0.0556	0.0909	0.0422
Priors	0.5106	0.2553	0.2340	—

[a] Frequency.
[b] Percentage.

Table 6.25 Parametric Discriminant Function Analysis Using SAS Macro DISCRIM: Misclassified Observations in Validation Data Using Quadratic Discriminant Functions

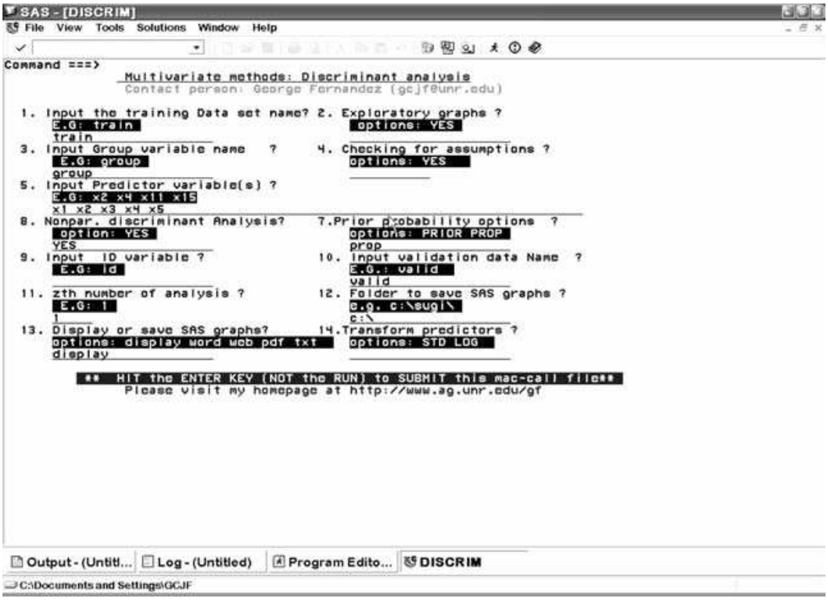
Observation	*X1*	*X2*	*X3*	*X4*	*From Group*	*Into Group*
75	1.11	93	393	490	1	2
83	1.08	94	426	213	2	1
110	0.94	88	423	212	2	1
131	1.07	124	538	460	3	2
134	0.81	123	557	130	3	2
136	1.01	120	636	314	3	2

- **Saving "plotp" and "out2" datasets for future use:** Running the DISCRIM macro creates these two temporary SAS datasets and saves them in the work folder. The "plotp" dataset contains the observed predictor variables, group response values, posterior probability scores, and new classification results. This posterior probability score for each observation in the dataset can be used as the base for developing the scorecards and ranking the patients. If an independent validation dataset is included, the classification results for the validation dataset are saved in a temporary SAS dataset called "out2", which can be used to develop scorecards for new patients.
- **Validation:** This step validates the derived discriminant functions obtained from the training data by applying these classification criteria to the independent simulated dataset and verifying the success of classification.

6.12.2 Data Descriptions

Dataset names	Training: Permanent SAS dataset "diabetic2"[11,17] located in the library "gf" Validation: Permanent SAS dataset "diabetic1" (simulated) located in the library "gf"
Group response variables	Three clinical diabetic groups: 1, normal; 2, overt diabetic; 3, chemical diabetic. Use numeric values for group levels to generate box plots of posterior probability density estimates.
Predictor variables	X1: relative weight X2: fasting plasma glucose level X3: test plasma glucose X4: plasma insulin during test X5: steady-state plasma glucose level
Number of observations	Training data: 145 Validation data: 141
Source	Training data: real[11,17] Validation data: simulated

Open the DISCRIM.sas macro-call file in the SAS PROGRAM EDITOR window and click RUN to open the DISCRIM macro-call window (Figure



```
Command ===>
                 Multivariate methods: Discriminant analysis
                 Contact person: George Fernandez (gcjf@unr.edu)

  1. Input the training Data set name? 2. Exploratory graphs ?
     E.G: train                            options: YES
     train
  3. Input Group variable name   ?       4. Checking for assumptions ?
     E.G: group                            options: YES
     group
  5. Input Predictor variable(s) ?
     E.G: x2 x4 x11 x15
     x1 x2 x3 x4 x5
  8. Nonpar. discriminant Analysis?      7.Prior probability options  ?
     option: YES                           options: PRIOR PROP
     YES                                   prop
  9. Input  ID variable ?                10. Input validation data Name  ?
     E.G: id                               E.G.: valid
                                           valid
 11. zth number of analysis ?            12. Folder to save SAS graphs ?
     E.G: 1                                e.g. c:\sugi\
     1                                     c:\
 13. Display or save SAS graphs?         14.Transform predictors ?
     options: display word web pdf txt     options: STD LOG
     display

      ** HIT the ENTER KEY (NOT the RUN) to SUBMIT this mac-call file**
         Please visit my homepage at http://www.ag.unr.edu/gf
```

Figure 6.10 Screen copy of DISCRIM macro-call window showing the macro-call parameters required for performing nonparametric discriminant analysis.

6.10). Input the appropriate macro-input values by following the suggestions given in the help file (see Section 6.9.2).

6.12.3 Exploratory Analysis/Diagnostic Plots

Input dataset name, group variable, predictor variable names, and prior probability option. Input YES in macro input field #2 to perform data exploration and create diagnostic plots. Submit the DISCRIM macro and discriminant diagnostic plots to obtain stepwise variable selection output.

6.12.4 Data Exploration and Checking

A simple two-dimensional scatterplot matrix showing the discrimination of three diabetes groups is presented in Figure 6.11. These scatterplots are useful in examining the range of variation in the predictor variables and the degree of correlations between any two predictor variables. The scatterplot presented in Figure 6.11 revealed that a strong correlation existed between fasting plasma glucose level (X2) and test plasma glucose (X3). These two attributes appeared to discriminate diabetes group 3 from the other two groups to a certain degree. Discrimination between the normal and the overt diabetes group is not very distinct. The results of

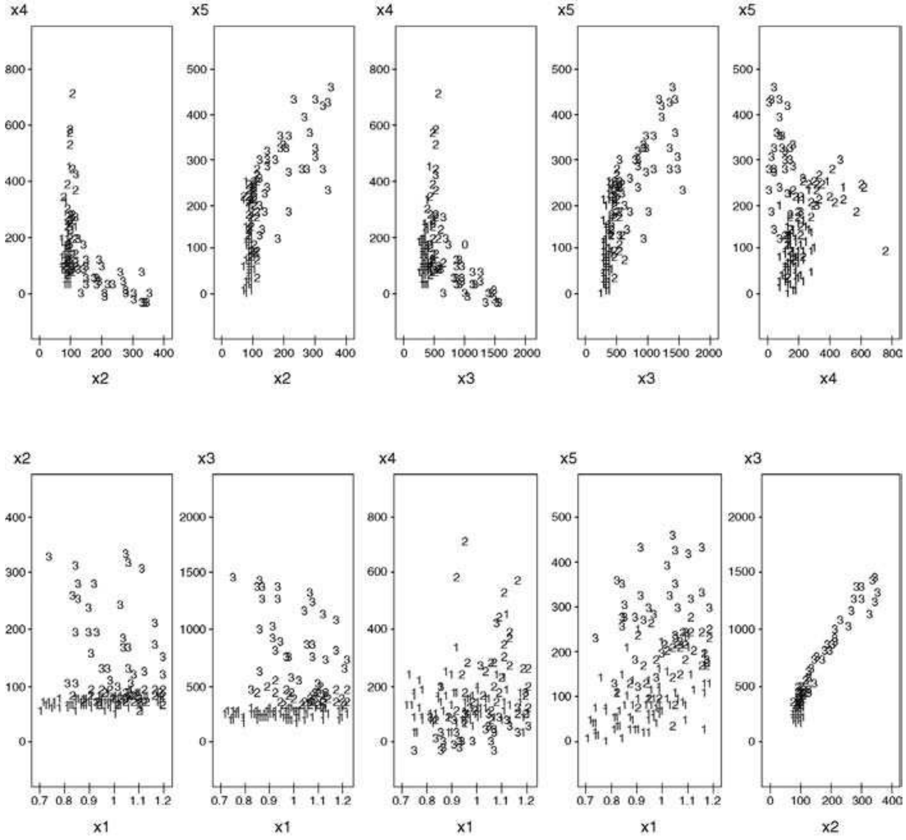


Figure 6.11 Bivariate exploratory plots generated using the SAS macro DISCRIM: group discrimination of three types of diabetic groups (data = diabetic2) in simple scatterplots.

variable selection methods indicate that variable X5 can be dropped from the significant predictor variable list. The details of variable selection results are not discussed here because these results are similar to Case Study 1 in this chapter.

6.12.5 Discriminant Analysis and Checking for Multivariate Normality

Open the DISCRIM.sas macro-call file in the SAS PROGRAM EDITOR window and click RUN to open the DISCRIM macro-call window (Figure 6.10). Input the appropriate macro-input values by following the suggestions given in the help file (see Section 6.9.2). Input the dataset name, group variable, predictor variable names, and prior probability option. Leave macro input option #2 blank and input YES in macro input option #6 to perform nonparametric DFA. Also input YES in macro input option #4 to perform a multivariate normality check. Submit the DISCRIM macro to obtain the multivariate normality check and the nonparametric DFA output and graphics.

6.12.6 Checking for Multivariate Normality

This multivariate normality assumption can be checked by estimating multivariate skewness and kurtosis and testing for their significance levels. The quantile–quantile (Q–Q) plot of expected and observed distributions[9] of multi-attribute residuals after adjusting for the group means can be used to graphically examine multivariate normality. The estimated multivariate skewness (3.945; *p* value, 0.000) and multivariate kurtosis (40.642; *p* value, 0.000) clearly support the hypothesis that, after adjusting for the group differences, these four multi-attributes do not have a joint multivariate normal distribution (Figure 6.12). A significant departure from the 45° angle reference line in the Q–Q plot (Figure 6.12) also supports this finding. Thus, nonparametric discriminant analysis must be considered to be the appropriate technique for discriminating the three clinical groups based on these four attributes (X1 to X4).

6.12.7 Checking for the Presence of Multivariate Outliers

Multivariate outliers can be detected in a plot between the differences of robust (Mahalanobis distance–chi-squared quantile) vs. chi-squared quantile values.[9] Eight observations are identified as influential observations (Table 6.26) because the differences between the robust Mahalanobis distance and chi-squared quantile values are larger than 2 and fall outside the critical region (Figure 6.13).

When the distribution within each group is assumed not to have a multivariate normal distribution, nonparametric DFA methods can be used to estimate the group-specific densities. Nonparametric discriminant methods are based on nonparametric estimates of group-specific probability densities. Either a kernel method or the *k*-nearest-neighbor method can be used to generate a nonparametric density estimate in each group and to produce a classification criterion.

The class level information and the prior probability estimate used in performing the nonparametric DFA are given in Table 6.27. By default, the DISCRIM macro performs three (*k* = 2, 3, and 4) nearest neighbor (NN) and one kernel density (KD) (unequal bandwidth kernel density) nonparametric DFA. We can compare the classification summary and the

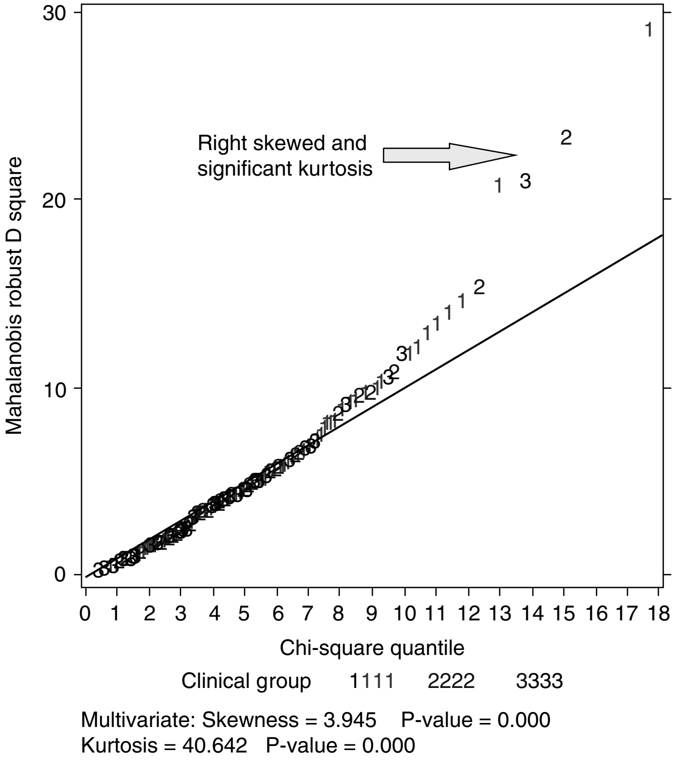


Figure 6.12 Checking for multivariate normality in Q–Q plot (data = diabetic1) generated using the SAS macro DISCRIM.

Table 6.26 Detecting Multivariate Outliers and Influential Observations Using SAS Macro DISCRIM

ID	*Robust Distance Square Statistic (RDSQ)*	*Chi-Square*	*Difference*
82	29.2183	17.6296	11.5887
86	23.4200	15.0041	8.4159
69	20.8619	12.9202	7.9417
131	21.0878	13.7552	7.3326
111	15.4618	12.2891	3.1728
26	14.7255	11.7799	2.9456
76	14.0995	11.3522	2.7474
31	13.5649	10.9827	2.5823

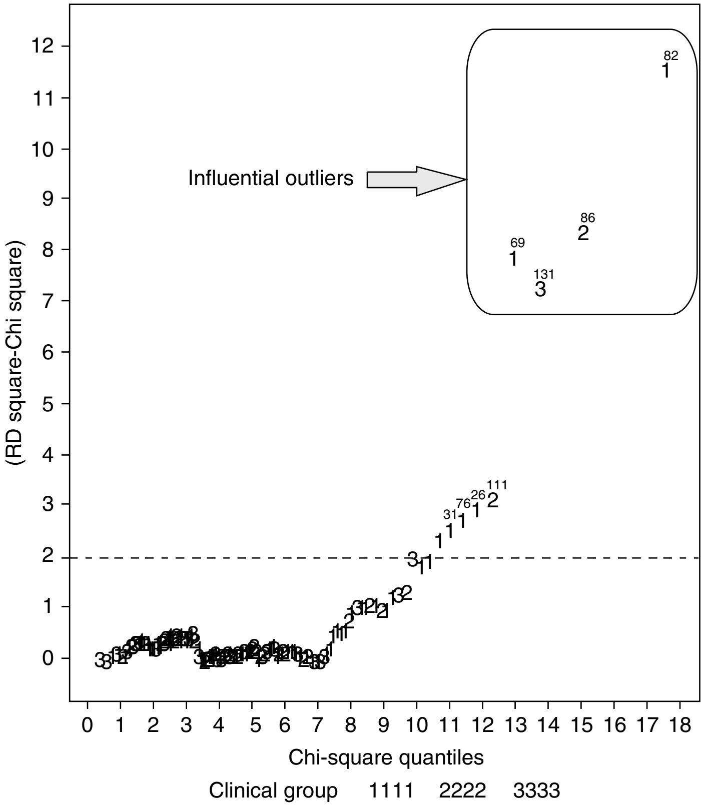


Figure 6.13 Diagnostic plot for detecting multivariate influential observations (data = diabetic2) generated using the SAS macro DISCRIM.

Table 6.27 Nonparametric Discriminant Function Analysis Using SAS Macro DISCRIM: Class Level Information

Group	*Variable Name*	*Frequency*	*Weight*	*Proportion*	*Prior Probability*
1	_1	76	76.0000	0.524138	0.524138
2	_2	36	36.0000	0.248276	0.248276
3	_3	33	33.0000	0.227586	0.227586

misclassification rates of these four different nonparametric DFA methods and can pick one that gives the smallest classification error in the cross validation.

Among the three NN DFAs, classification results based on NN (k = 2) nonparametric DFA give the smallest classification error. The classification

Table 6.28 Nearest Neighbor (k = 2) Nonparametric Discriminant Function Analysis Using SAS Macro DISCRIM: Classification Summary Using Cross Validation

	Number of Observations and Percent Classified into Group			
From Group	*1*	*2*	*3*	*Total*
1	75[a]	1	0	76
	98.68[b]	1.32	0.00	100.00
2	0	36	0	36
	0.00	100.00	0.00	100.00
3	0	4	29	33
	0.00	12.12	87.88	100.00
Total	75	41	29	145
	51.72	28.28	20.00	100.00
Priors	0.52414	0.24828	0.22759	—
	Error Count Estimates for Group			
	1	*2*	*3*	*Total*
Rate	0.0132	0.0000	0.1212	0.0345
Priors	0.5241	0.2483	0.2276	—

[a] Frequency.

[b] Percentage.

summary and the error rates for NN (k = 2) are presented in Table 6.28. When the k-nearest-neighbor method is used, the Mahalanobis distances are estimated based on the pooled covariance matrix. The misclassification rates in groups 1, 2, and 3 are 1.3, 0, and 12.0%, respectively. The overall discrimination is quite satisfactory, as the overall error rate is very low, at 3.45%. The posterior probability estimates based on cross validation reduces both the bias and the variance of classification function. The resulting overall error estimates are intended to have both a low variance from using the posterior probability estimate and a low bias from cross validation.

Figure 6.14 illustrates the variation in the posterior probability estimates for the normal group (1). The posterior probability estimates of a majority of the cases that belong to the normal group are larger than 0.95. One observation (69) is identified as a false negative, while no other observation is identified as false positive. Few intra-group variations for the posterior probability estimates were observed. Figure 6.15 displays the variation in the posterior probability estimates for the overt diabetes group (2). A relatively large variability for the posterior probability estimates is observed for the overt diabetes group (2) and ranges from 0.5 to 1. No observation is identified as a false negative. However, five observations, one belonging to the normal

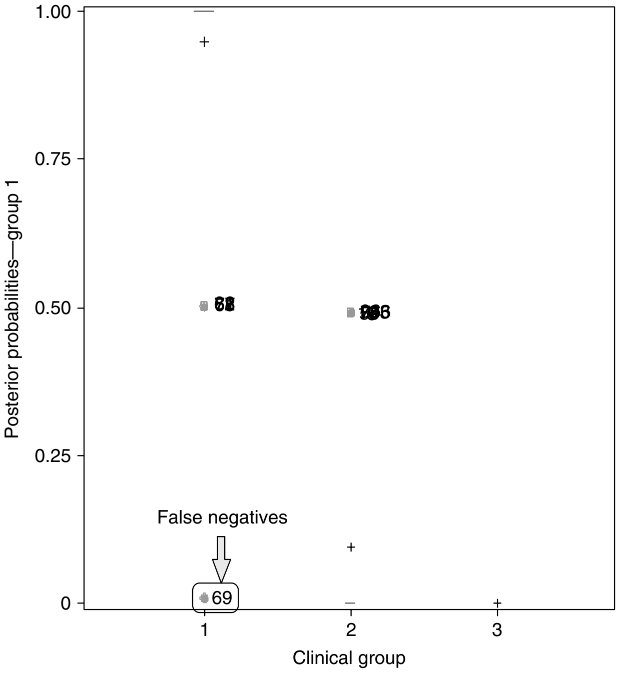


Figure 6.14 Box-plot display of posterior probability estimates for group level = 1 derived from nearest neighbor (k = 2) nonparametric discriminant function analysis by cross validation. This plot is generated using the SAS macro DISCRIM.

group (1) and four observations belonging to the chemical group (3), are identified as false positives. The variability in the posterior probability estimates for the chemical diabetes group (3) is illustrated in Figure 6.16. The posterior probability estimates for a majority of the cases that belong to the chemical group (3) are larger than 0.95. One observation is identified as false negative, but no observations are identified as false positives.

The DISCRIM macro also outputs a table of the *i*th group posterior probability estimates for all observations in the training dataset. Table 6.29 provides a partial list of the *i*th group posterior probability estimates for some of the selected observations in the table. These posterior probability values are very useful estimates because these estimates can be used to develop scorecards and rank the observations in the dataset.

The posterior probability error-rate estimates for each group are based on the posterior probabilities of the observations classified into that same group. The posterior probability estimates provide good estimates of the error rate when the posterior probabilities are accurate. The smoothed

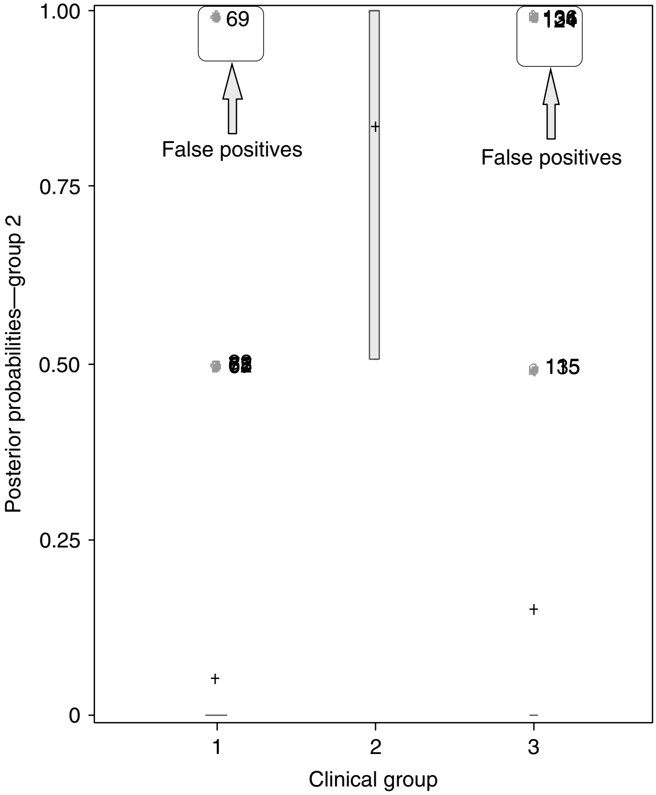


Figure 6.15 Box-plot display of posterior probability estimates for group level = 2 derived from nearest neighbor (*k* = 2) nonparametric discriminant function analysis by cross validation. This plot is generated using the SAS macro DISCRIM.

posterior probability error-rate estimates based on cross-validation quadratic discriminant functions are presented in Table 6.30. The overall error rates for both stratified and unstratified estimates are equal because the group proportion was used as the prior probability estimate. The overall discrimination is quite satisfactory, as the overall error rate using the smoothed posterior probability error rate is relatively low, at 6.8%.

If the classification error rate obtained for the validation data is small and similar to the classification error rate for the training data, then we can conclude that the derived classification function has good discriminative potential. Classification results for the validation dataset based on NN (*k* = 2) classification functions are presented in Table 6.31. The misclassification rates in groups 1, 2, and 3 are 4.1, 25, and 15.1%, respectively.

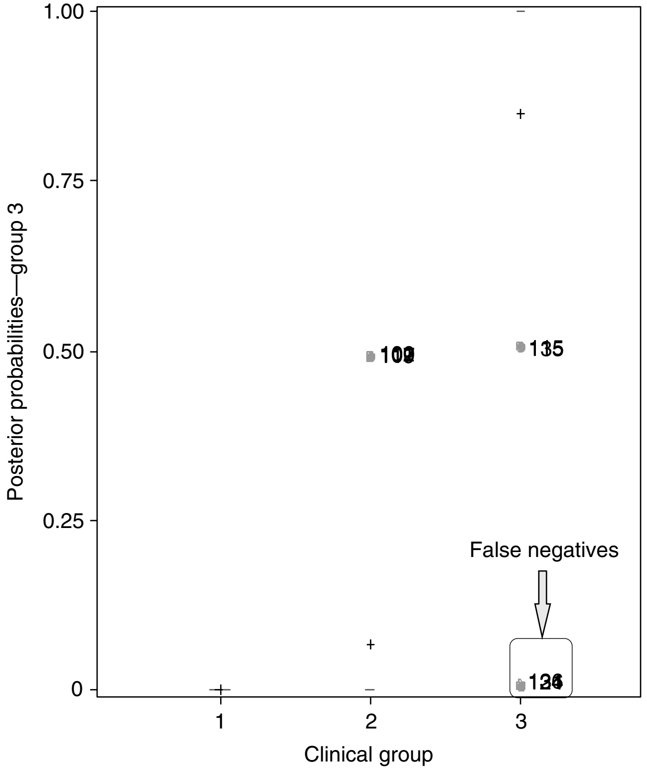


Figure 6.16 Box-plot display of posterior probability estimates for group level = 3 derived from nearest neighbor (*k* = 2) nonparametric discriminant function analysis by cross validation. This plot is generated using the SAS macro DISCRIM.

The overall discrimination in the validation dataset is moderately good, as the weighted error rate is 11.2%. A total of 17 observations in the validation dataset are misclassified. Table 6.32 shows a partial list of probability density estimates and the classification information for all the observations in the validation dataset. The misclassification error rate estimated for the validation dataset is relatively higher than the error rate obtained from the training data. We can conclude that the classification criterion derived using NN (*k* = 2) performed poorly in validating the independent validation dataset. The presence of multivariate influential observations in the training dataset might be one of the contributing factors for this poor performance in validation. Using larger *k* values in NN DFA might do a better job of classifying validation dataset.

The DISCRIM also performs nonparametric discriminant analysis based on nonparametric kernel density estimates (KD) with unequal bandwidth.

Table 6.29 Nearest Neighbor (*k* = 2) Nonparametric Discriminant Function Analysis Using SAS Macro DISCRIM: Posterior Probability Estimates by Group Levels in Cross-Validation

			Posterior Probability of Membership in Group		
Observation	*From Group*	*Classified into Group*	*1*	*2*	*3*
1	1	1	0.9999	0.0001	0.0000
2	1	2[a]	0.1223	0.8777	0.0001
3	1	1	0.7947	0.2053	0.0000
4	1	1	0.9018	0.0982	0.0000
5	1	2[a]	0.4356	0.5643	0.0001
6	1	1	0.8738	0.1262	0.0000
7	1	1	0.9762	0.0238	0.0000
8	1	1	0.9082	0.0918	0.0000
—[b]	—	—	—	—	—
137	3	1*	0.9401	0.0448	0.0151
138	3	3	0.0000	0.3121	0.6879
139	3	3	0.0000	0.0047	0.9953
140	3	3	0.0000	0.0000	1.0000
141	3	3	0.0000	0.0011	0.9988

[a] Misclassified observation.
[b] Partial list.

The kernel method in the DISCRIM macro uses normal kernels in the density estimation. In the KD method, the Mahalanobis distances based on either the individual within-group covariance matrices or the pooled covariance matrix can be used. The classification of observations in the training data is based on the estimated group specific densities from the training dataset. From these estimated densities, the posterior probabilities of group membership are estimated.

The classification summary using KD (normal, unequal bandwidth) nonparametric DFA and the error rates using cross-validation are presented in Table 6.33. The misclassification rates in groups 1, 2, and 3 are 7.8, 16.6, and 9.0%, respectively. Thus, an overall success rate of correct discrimination is about 90%, as the overall error rate is about 10.3%, slightly lower than the overall error rate for the (*k* = 2) NN method.

Figure 6.17 illustrates the variation in the posterior probability estimates for the normal group (1). The posterior probability estimates for a majority of the cases that belong to the normal group are larger than 0.95. Seven observations belonging to the normal group are identified

Table 6.30 Nearest Neighbor (*k* = 2) Nonparametric Discriminant Function Analysis Using SAS Macro DISCRIM: Classification Summary and Posterior Probability Error Rate in Cross-Validation

	To Group		
From Group	*1*	*2*	*3*
1	75[a]	1	0
	0.9603[b]	1.0000	—
2	0	36	0
	—	0.8357	—
3	0	4	29
	—	1.0000	0.9660
Total	75	41	29
	0.9603	0.8557	0.9660
Priors	0.52414	0.24828	0.22759

	Posterior Probability Error Rate Estimates for Group			
Estimate	*1*	*2*	*3*	*Total*
Stratified	0.0524	0.0254	0.1510	0.0681
Unstratified	0.0524	0.0254	0.1510	0.0681
Priors	0.5241	0.2483	0.2276	—

[a] Frequency.
[b] Percentage.

as false negatives. Two observations belonging to the overt group (2) are identified as false positives. Very small amounts of intra-group variation for the normal group posterior probability estimates were observed. Figure 6.18 displays the variation in the posterior probability estimates for the overt diabetes group (2). A relatively large variability for the posterior probability estimates is observed for the overt diabetes group, ranging from 0.75 to 1. Six observations belonging to the overt group are identified as false negatives. Five observations belonging to the normal group (1) and two observations belonging to the chemical group (3) are identified as false positives. The variability in the posterior probability estimates for the chemical group (3) is illustrated in Figure 6.19. The posterior probability estimates for a majority of the cases that belong to the chemical group are larger than 0.95. Two observations from the chemical group (3) are identified as false negatives, while four observations belonging to the overt group (2) are identified as false positives.

The posterior probability error-rate estimates for each group are based on the posterior probabilities of the observations classified into that same

Table 6.31 Classification Summary and Error Rates for Validation Dataset Using SAS Macro DISCRIM: Nearest Neighbors (k = 2) Method

	Number of Observations and Percent Classified into Group			
From Group	*1*	*2*	*3*	*Total*
1	69[a]	3	0	72
	95.83[b]	4.17	0.00	100.00
2	9	27	0	36
	25.00	75.00	0.00	100.00
3	2	3	28	33
	6.06	9.09	84.85	100.00
Total	80	33	28	141
	56.74	23.40	19.86	100.00
Priors	0.52414	0.24828	0.22759	—

	Error Count Estimates for Group			
	1	*2*	*3*	*Total*
Rate	0.0417	0.2500	0.1515	0.1184
Priors	0.5241	0.2483	0.2276	—

[a] Frequency.
[b] Percentage.

Table 6.32 Classification Results in Validation Dataset by Nearest Neighbor (k = 2) Method Using SAS Macro DISCRIM

Observation	*X1*	*X2*	*X3*	*X4*	*Group*	*Into*
1	0.81	80	356	124	1	1
2	0.95	97	289	117	1	1
3	0.94	105	319	143	1	1
4	1.04	90	356	199	1	1
—[a]	—	—	—	—	—	—
142	0.91	180	923	77	3	3
143	0.9	213	1025	29	3	3
144	1.11	328	1246	124	3	3
145	0.74	346	1568	15	3	3

[a] Partial list.

Table 6.33 Unequal Bandwidth Kernel Density Discriminant Function Analysis Using SAS Macro DISCRIM: Classification Summary Using Cross-Validation Results

	Number of Observations and Percent Classified into Group			
From Group	*1*	*2*	*3*	*Total*
1	70[a]	5	1	76
	92.11[b]	6.58	1.32	100.00
2	2	30	4	36
	5.56	83.33	11.11	100.00
3	0	3	30	33
	0.00	9.09	90.91	100.00
Total	72	38	35	145
	49.66	26.21	24.14	100.00
Priors	0.52414	0.24828	0.22759	—

	Error Count Estimates for Group			
	1	*2*	*3*	*Total*
Rate	0.0789	0.1667	0.0909	0.1034
Priors	0.5241	0.2483	0.2276	—

[a] Frequency.
[b] Percentage.

group. The smoothed posterior probability error-rate estimates based on cross-validation discriminant function analysis are presented in Table 6.34. The overall error rates for both stratified and unstratified estimates are equal because the group proportion was used as the prior probability estimate. The overall discrimination is quite satisfactory, as the overall error rate using the smoothed posterior probability error rate is relatively low, at 4.7%.

If the classification error rate obtained for the validation data is small and similar to the classification error rate for the training data, then we can conclude that the derived classification function has good discriminative potential. Classification results for the validation dataset based on KD (normal, unequal bandwidth) nonparametric DFA classification functions are presented in Table 6.35. The misclassification rates in groups 1, 2, and 3 are 4.1, 25, and 15.1%, respectively.

The overall discrimination in the validation dataset is moderately good since the weighted error rate is 11.8%. A total of 17 observations in the validation dataset are misclassified. Table 6.36 shows a partial

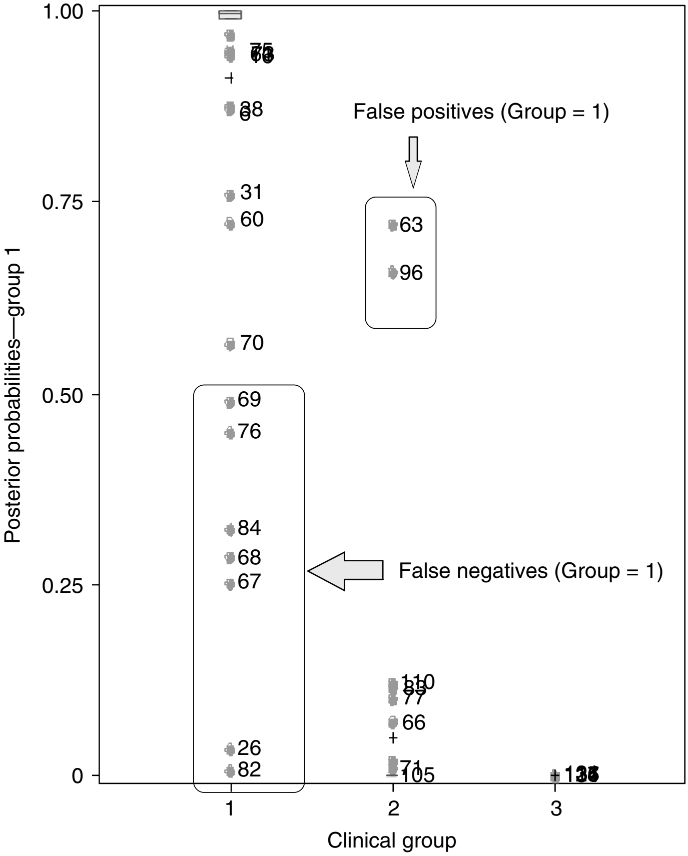


Figure 6.17 Box-plot display of posterior probability estimates for group level = 1 derived from nonparametric unequal bandwidth kernel density discriminant function analysis by cross validation. This plot is generated using the SAS macro DISCRIM.

list of probability density estimates and the classification information for all the observations in the validation dataset. The misclassification error rate estimated for the validation dataset is relatively higher than the error rate obtained from the training data. We can conclude that the classification criterion derived using KD (normal, unequal bandwidth) performed poorly in validating the independent validation dataset. The presence of multivariate influential observations in the training dataset might be one of the contributing factors for this poor performance in validation. Using other types of density options might do a better job in classifying the validation dataset.

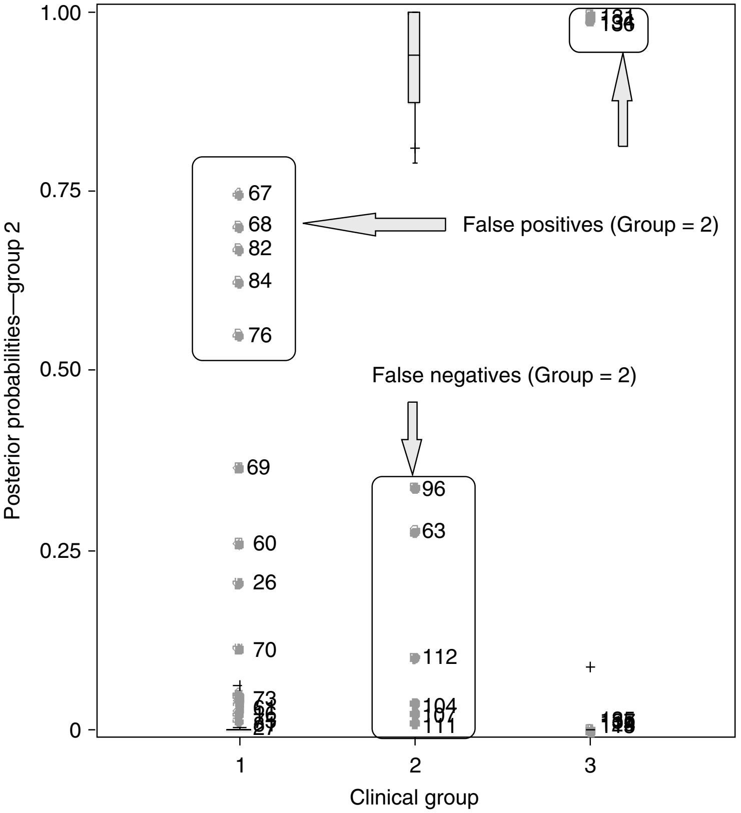


Figure 6.18 Box-plot display of posterior probability estimates for group level = 2 derived from nonparametric unequal bandwidth kernel density discriminant function analysis by cross validation. This plot is generated using the SAS macro DISCRIM.

6.13 Case Study 3: Classification Tree Using CHAID

Chi-squared automatic interaction detection (CHAID) analysis is a powerful classification tree method suitable for classifying observations into predetermined groups based on easy-to-follow decision rules. Unlike the discriminant analysis, both categorical and continuous variables can be used as predictors. Multivariate normality or between-group equal variance–covariance assumptions are not required to run the CHAID analysis. In Case Study 3, the features of CHAID analysis are revealed by fitting a classification model using the diabetes dataset described in Section 6.12. By comparing the classification summary

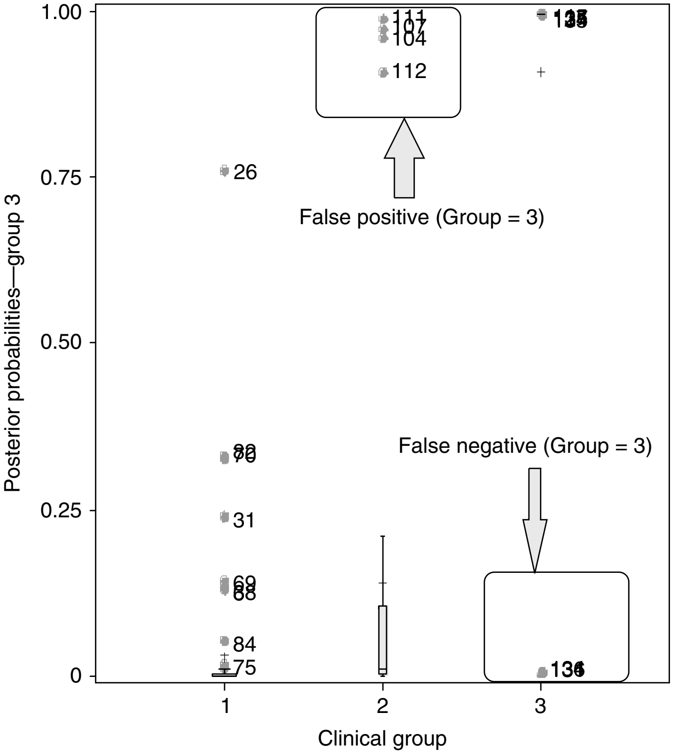


Figure 6.19 Box-plot display of posterior probability estimates for group level = 3 derived from nonparametric unequal bandwidth kernel density discriminant function analysis by cross validation. This plot is generated using the SAS macro DISCRIM.

and misclassification error rates, the similarities and the differences between CHAID and DFA are presented here. The diabetes dataset[11,17] used in Section 6.12 is used as the training dataset to derive the decision tree using blood plasma and insulin measures. The simulated diabetes dataset used in Section 6.12 is used as the validation dataset.

6.13.1 Study Objectives

To discriminate three clinical diabetic groups (normal, overt diabetic, and chemical diabetic) using blood plasma and insulin measures by developing a decision tree model.

Table 6.34 Unequal Bandwidth Kernel Density Discriminant Function Analysis Using SAS Macro DISCRIM: Classification Summary and Posterior Probability Error Rates in Cross-Validation

	To Group			
From Group	*1*	*2*	*3*	
1	70[a]	5	1	
	0.9709[b]	0.6586	0.7597	
2	2	30	4	
	0.6905	0.9469	0.9694	
3	0	3	30	
	.	0.9964	0.9999	
Total	72	38	35	
	0.9632	0.9128	0.9896	
Priors	0.52414	0.24828	0.22759	
	Posterior Probability Error Rate Estimates for Group			
Estimate	*1*	*2*	*3*	*Total*
Stratified	0.0875	0.0364	–0.0496	0.0437
Unstratified	0.0875	0.0364	–0.0496	0.0437
Priors	0.5241	0.2483	0.2276	—

[a] Frequency.

[b] Percentage.

Table 6.35 Classification Summary and Error Rate for Validation Dataset Using SAS Macro DISCRIM: Unequal Bandwidth Kernel Density Discriminant Function Analysis

	Number of Observations and Percent Classified into Group			
From Group	*1*	*2*	*3*	*Total*
1	69[a]	3	0	72
	95.83[b]	4.17	0.00	100.00
2	9	27	0	36
	25.00	75.00	0.00	100.00
3	2	3	28	33
	6.06	9.09	84.85	100.00
Total	80	33	28	141
	56.74	23.40	19.86	100.00
Priors	0.52414	0.24828	0.22759	—
	Error Count Estimates for Group			
	1	*2*	*3*	*Total*
Rate	0.0417	0.2500	0.1515	0.1184
Priors	0.5241	0.2483	0.2276	—

[a] Frequency.

[b] Percentage.

Table 6.36 Misclassified in Observations in Validation Dataset by the Unequal Bandwidth Kernel Density Discriminant Function Analysis Using SAS Macro DISCRIM

Observation	*X1*	*X2*	*X3*	*X4*	*X5*	*Group*	*Into*
2	1.0017	76.127	378.69	296.34	206.07	1	2
29	1.2519	94.588	439.34	260.81	187	1	2
61	1.1095	101.87	423.92	324.94	207.82	1	2
80	0.9916	98.293	445.43	-43.83	162.16	2	1
82	0.8679	92.868	498.67	59.006	113.38	2	1
83	0.9162	87.208	450.96	125.4	155.03	2	1
93	0.8956	112.02	580.41	40.301	90.217	2	1
98	0.9194	89.872	459.24	137.31	103.32	2	1
102	0.9492	117.89	510.63	233.84	122.3	2	1
104	1.0744	93.793	435.95	110.52	187.27	2	1
105	1.1961	114.51	492.52	44.762	232.42	2	1
108	0.9343	106.97	508.96	185	87.276	2	1
120	1.0439	103.84	547.63	322.49	259.23	3	2
121	0.9947	203.72	816.96	176.81	201.23	3	1
126	0.9436	119.83	732.16	162.04	173.31	3	2
135	1.026	133.56	665.97	149.9	193.46	3	2
137	0.9905	228.35	849.61	92.44	145.71	3	1

- **Developing classification tree functions:** Develop classification functions based on CHAID analysis and assign observations into predefined group levels, and measure the success of classification by comparing the classification error rates.
- **Constructing the decision tree:** Construct easy-to-follow decision trees using the decision rules generated in the CHAID analysis.
- **Validation:** Validates the derived classification functions obtained from the training data by applying these classification criteria to the independent validation dataset and verifying the success of classification.

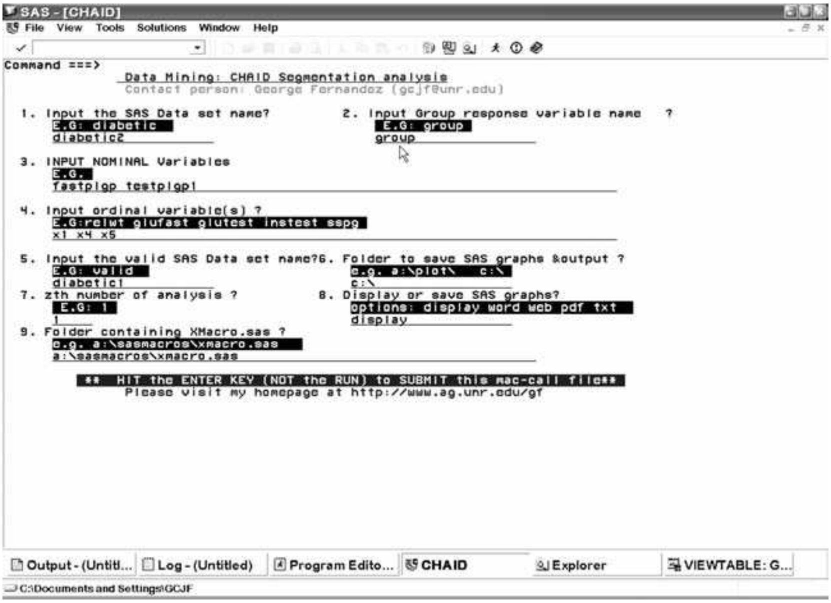
6.13.2 Data Descriptions

Dataset names	Training: Permanent SAS dataset "diabetic2"[11,17] located in the library "gf" Validation: Permanent SAS dataset "diabetic1" (simulated) located in the library "gf"
Group response variables	Three clinical diabetic groups: 1, normal; 2, overt diabetic; 3, chemical diabetic
Predictor variables (continuous)	X1: relative weight X4: plasma insulin during test X5: steady-state plasma glucose level
Predictor variables (nominal)	fastplgp: fasting plasma glucose values ($L < 100$; $M = 100$–200; $H > 200$) tstplgp1: test plasma glucose values ($L < 600$; $M = 600$–1200; $H > 1200$)
Number of observations	Training data: 145 Validation data: 141
Source	Training data: real[11,17] Validation data: simulated

Open the CHAID.sas macro-call file in the SAS PROGRAM EDITOR window and click RUN to open the CHAID macro-call window (Figure 6.20). Input the appropriate macro input values by following the suggestions given in the help file (see Section 6.10.2). Input the dataset name, group response variable, and categorical and continuous predictor variable names. Also, input the validation dataset name to validate the classification tree. Specify the full path and the filename, including the file extension of XMACRO.sas, when running the macro from the companion CD-ROM; otherwise, leave this field blank. The CHAID macro can read the XMACRO.sas file from the book website. Submit the CHAID macro and classification summary plots to produce decision tree diagrams.

The CHAID analysis macro generates very small amounts of output and graphics; therefore, for data exploration and preliminary analysis, users should fit classification models using LOGISTIC and DISCRIM models before trying the CHAID analysis.

If the CHAID macro runs successfully, the classification summary for the training data will be displayed in a donut chart. Classification results based on the CHAID analysis are presented in Figure 6.21. The misclassification rates in groups 1, 2, and 3 are 6, 25, and 0%, respectively. The overall error rate is not computed directly; however, an overall error rate


SAS - [CHAID]
File View Tools Solutions Window Help
Output - (Untitl... Log - (Untitled) Program Edito... CHAID Explorer VIEWTABLE: G...
C:\Documents and Settings\GCJF

```
Command ===>
            Data Mining: CHAID Segmentation analysis
            Contact person: George Fernandez (gcjf@unr.edu)

  1. Input the SAS Data set name?            2. Input Group response variable name   ?
     E.G: diabetic                              E.G: group
     diabetic2                                  group

  3. INPUT NOMINAL Variables
     E.G.
     fastplgp testplgp1

  4. Input ordinal variable(s) ?
     E.G:relwt glufast glutest instest sspg
     x1 x4 x5

  5. Input the valid SAS Data set name?6. Folder to save SAS graphs &output ?
     E.G: valid                                 e.g. a:\plot\   c:\
     diabetic1                                  c:\
  7. zth number of analysis ?              8. Display or save SAS graphs?
     E.G: 1                                     options: display word web pdf txt
     1                                          display
  9. Folder containing XMacro.sas ?
     e.g. a:\sasmacros\xmacro.sas
     a:\sasmacros\xmacro.sas

      **  HIT the ENTER KEY (NOT the RUN) to SUBMIT this mac-call file**
            Please visit my homepage at http://www.ag.unr.edu/gf
```

Figure 6.20 Screen copy of CHAID macro-call window showing the macro-call parameters required for performing chi-squared automatic interaction detection analysis.

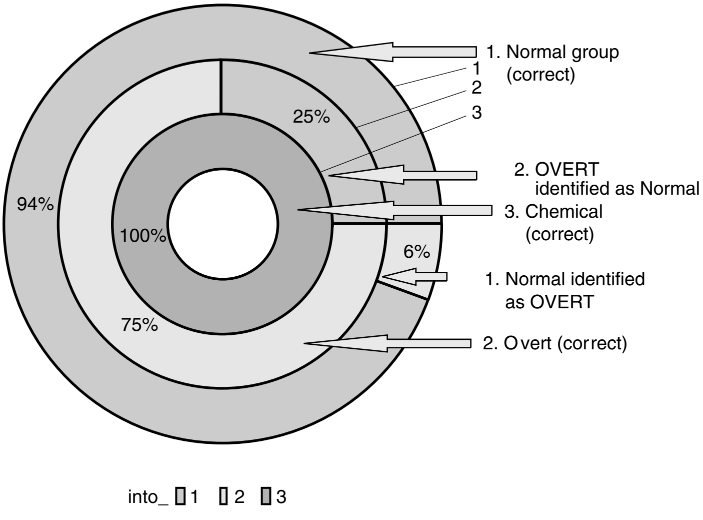


Figure 6.21 Donut chart showing the training data classification summary display generated by using the SAS macro CHAID.

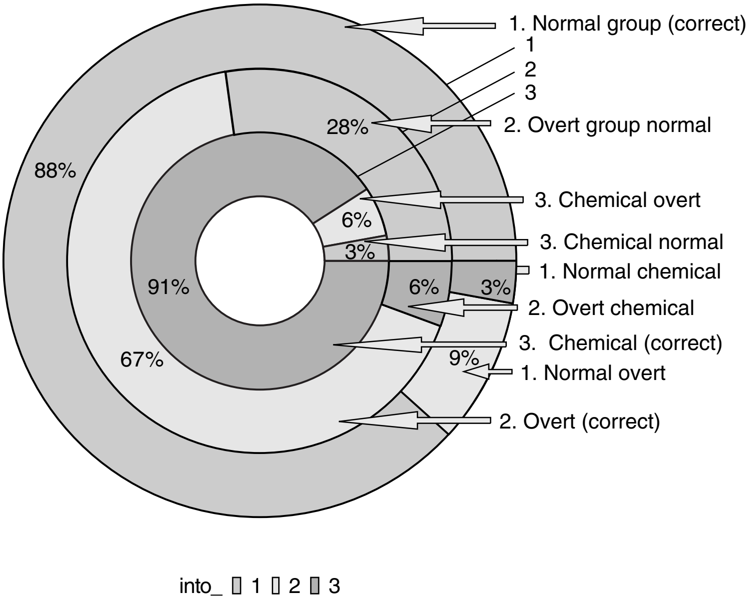


Figure 6.22 Donut chart showing the validation data classification summary display generated by using the SAS macro CHAID.

can be computed by an average error rate weighted by the group frequency. In case of a validation dataset, relatively more classification errors were observed (Figure 6.22). The misclassification rates in groups 1, 2, and 3 are 12, 34, and 9%, respectively. By comparing the classification errors obtained from the CHAID and the DISCRIM macro, we can conclude that the discriminatory power of the DISCRIM macro is superior to the CHAID macro in correctly classifying the three diabetes groups.

The CHAID macro generates a decision tree and produces a graphics file using the NETDRAW procedure available in the SAS OR module. Therefore, to generate the decision tree, the SAS OR module should be installed in the computer. The decision tree generated by the CHAID macro is given in Figure 6.23. When the decision tree graphics are too large to fit within a page, SAS splits the graphics file and outputs them in multiple pages. These split graphics files are joined back together and produced as one single graphic in Figure 6.23. When the graphics files are joined, the resolution of the graphics file is reduced and, as a result, interpretation of the decision tree becomes unclear.

One way to solve this problem is to generate the decision tree manually using the decision rule generated by the CHAID macro. The SAS OR

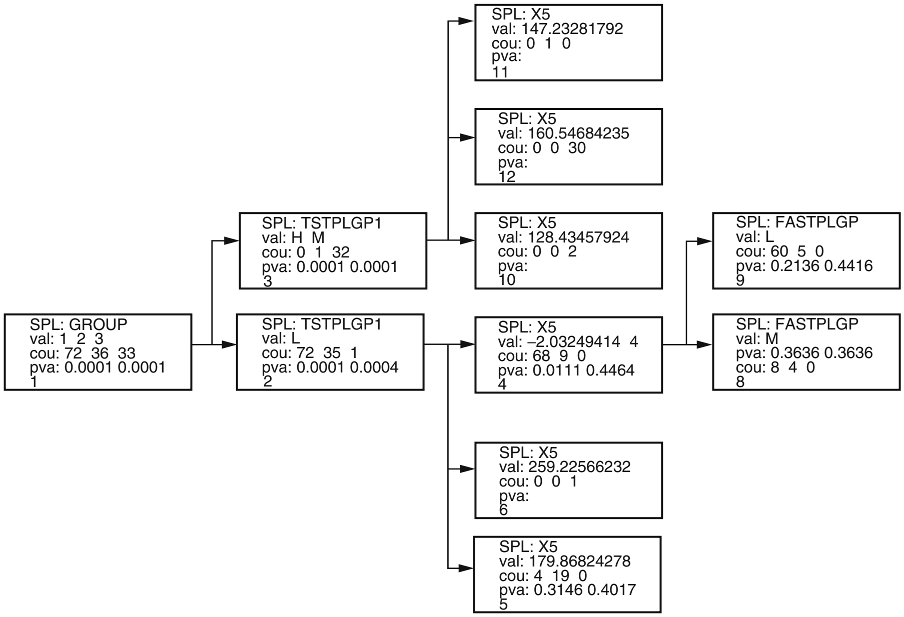


Figure 6.23 Decision tree diagram generated by the SAS macro CHAID using the net draw procedure available in the SAS OR module.

module is not required to generate the decision rules of classification; a copy of the decision rule is presented in Table 6.37. Based on these rules, a decision tree can be constructed manually using any suitable draw program. The decision tree generated by the Corel Presentation DRAW program is presented in Figure 6.24.

In the decision tree construction, all observations in the training dataset are grouped together in the tree trunk node. In the first step, the nominal variable test plasma glucose (tstplgp1) is identified as the most significant predictor in separating observations in the three clinical groups. In the first split, all except one chemical diabetes cases are correctly identified based on the following decision rule:

> Chemical group (terminal/leaf node): tstplgp1 ≥ 600 (with one misclassification)
> Normal and overt groups: tstplgp1 < 600 (with one misclassification)

In the second split, the steady-state plasma glucose level is selected as the important predictor to discriminate the normal group (1) from the overt group (2) by the following decision rule:

Table 6.37 Decision Rules Generated by the CHAID Analysis Using the SAS Macro CHAID: CHAID Analysis of Dependent Variable (DV) GROUP

GROUP values: 1 2 3
DV counts: 72 36 33 Best *p* values: 0.0001 0.0001

TSTPLGP1 value: *L*
DV counts: 72 35 1 Best *p* values: 0.0001 0.0004
X5 values:
–2.03249414 4.3937652377 38.559882904 44.528199324 44.535653211 44.753642423
52.398897375 52.528357771 53.172972712 54.88631257 56.557238318 68.89928
7348 81.149310923 82.399675811, etc.

DV counts: 68 9 0 Best *p* values: 0.0111 0.4464
FASTPLGP value: *M*
DV counts: 8 4 0 Best *p* values: 0.3636 0.3636
FASTPLGP value: *L*
DV counts: 60 5 0 Best *p* values: 0.2136 0.4416

X5 values:
179.86824278 187.00215516 187.27177317 188.26867046 196.45058388 198.86202311
206.06732876 206.21903124 207.82210852 208.4789315 210.51382161 218.196
15889 223.67582281 225.12568308, etc.

DV counts: 4 19 0 Best *p* values: 0.3146 0.4017
X5 value: 259.22566232
DV counts: 0 0 1
X5 values:
263.35227522 271.4201007 272.46053531 274.01080629 328.18269653 351.90019824
374.56331699
DV counts: 0 7 0

TSTPLGP1 value(s): *H M*
DV counts: 0 1 32 Best *p* values: 0.0001 0.0001
X5 values: 128.43457924 145.70590001
DV counts: 0 0 2
X5 value: 147.23281792
DV counts: 0 1 0
X5 values:
160.54684235 173.30593671 193.46150047 201.22972662 206.06366955 211.14031345
211.70428831 216.74283674 231.27195626 234.62979942 237.96478751 240.75
302441 245.17 271.21569933, etc.
DV counts: 0 0 30

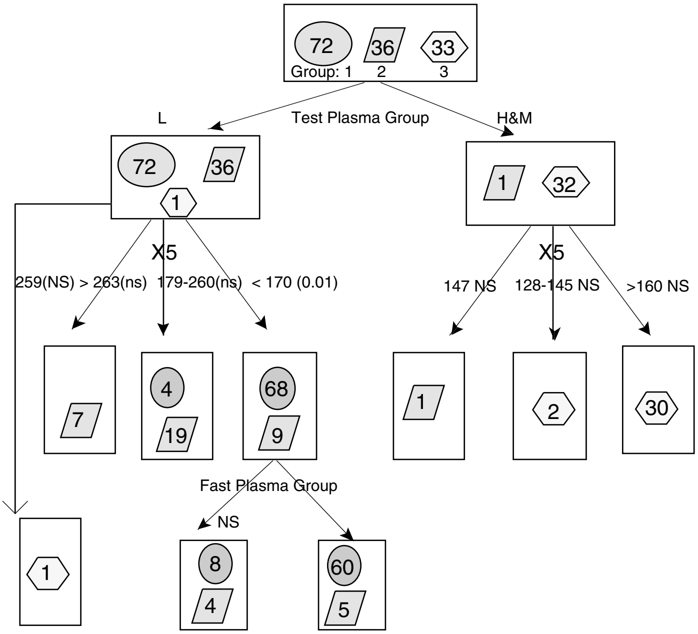


Figure 6.24 Decision tree diagram generated manually using the decision tree information generated by using the SAS macro CHAID.

Normal group: X5 < 170 (69 out of 72 in the normal group plus 9 overt cases misclassified as normal)
Overt group: X5 > 170 (26 out of 35 overt group plus four normal cases misclassified as overt)

The subsequent splits are statistically not significant; thus, we could stop at this step and interpret the decision tree. Therefore, using two predictor variables and easy-to-follow decision rules, 126 out of the 141 cases can be correctly classified. Thus, the SAS CHAID macro provides a simple but very valuable classification tool to classify categorical responses with acceptable predictive accuracy.

6.14 Summary

The methods for performing supervised classification models and for grouping categorical group response variables using the user-friendly SAS macro applications are covered in this chapter. Graphical methods to perform diagnostic and exploratory analysis, classification and discrimination, decision tree analysis, model assessment, and validation are presented using a clinical diabetes dataset. Steps involved in using the user-friendly SAS macro applications DISCRIM, for performing parametric and nonparametric discriminant analysis, and CHAID, for performing CHAID analysis and generating decision trees, are also presented.

References

1. Sharma, S., *Applied Multivariate Techniques*, John Wiley & Sons, New York, 1996, chaps. 8, 9.
2. Johnson, R.A. and Wichern, D.W., *Applied Multivariate Statistical Analysis*, 5th ed., Prentice-Hall, Englewood Cliffs, NJ, 2002, chap. 11.
3. Breiman, L., Friedman, J.H., Olshen, R.A., and Stone, C.J., *Classification and Regression Trees*, Wadsworth, Belmont, CA, 1984.
4. SAS Institute, Inc., *SAS/STAT Users Guide Version 8*, SAS Institute, Inc., Cary, NC, 1999.
5. Khattree, R. and Naik, D.N., *Multivariate Data Reduction and Discrimination with SAS Software*, 1st ed., SAS Institute, Inc., Cary, NC, 2000, chap. 5.
6. SAS Institute, Inc., *The STEPDISC Procedure: An Overview*, SAS online documentation, SAS Institute, Inc., Cary, NC (http://v8doc.sas.com/sashtml/stat/chap60/sect1.htm; accessed July 2002).
7. SAS Institute, Inc., *The CANDISC Procedure: An Overview*, SAS online documentation, SAS Institute, Inc., Cary, NC (http://v8doc.sas.com/sashtml/stat/chap21/sect1.htm; accessed July 2002).
8. SAS Institute, Inc., *The DISCRIM Procedure: An Overview*, SAS online documentation, SAS Institute, Inc., Cary, NC (http://v8doc.sas.com/sashtml/stat/chap25/sect1.htm; accessed May 2002).
9. Khattree, R. and Naik, D.N., *Applied Multivariate Statistics with SAS Software*, SAS Institute, Inc., Cary, NC, 1995, chap. 1.
10. Gabriel, K.R., Bi-plot display of multivariate matrices for inspection of data and diagnosis, in *Interpreting Multivariate Data,* V. Barnett, Ed., Wiley, London, 1981.
11. SAS Institute, Inc., *SAS Systems for Statistical Graphics*, 1st ed., SAS Institute, Inc., Cary, NC, 1991, chap. 9.
12. Lachenbruch, P.A. and Mickey, M.A., Estimation of error rates in discriminant analysis, *Technometrics*, 10, 1–10, 1968.
13. Hora, S.C. and Wilcox, J.B., Estimation of error rates in several population discriminant analyses, *J. Mark. Res.*, 19, 57–61, 1982.
14. Glick, N., Additive estimators for probabilities of correct classification, *Pattern Recognition*, 10, 211–222, 1978.
15. Berry, M.J.A. and Linoff, G.S., *Data Mining Techniques: For Marketing, Sales, and Customer Support*, John Wiley & Sons, New York, 1997, chap. 12.

16. SAS Institute, Inc., *The TREEDISC Macro for CHAID Analysis* (http://www.stat.lsu.edu/faculty/moser/exst7037/treedisc.html).
17. Reaven, G.M. and Miller, R.G., An attempt to define the nature of chemical diabetes using a multidimensional analysis, *Diabetologia*, 16, 17–24, 1979.

Suggested Reading

Eherler, D. and Lehmann, T., *Responder Profiling with CHAID and Dependency Analysis* (http://www.luc.ac.be/iteo/articles/lehmann.pdf).

Huberty, C.J., *Applied Discriminant Analysis*, Wiley Series in Probability and Mathematical Statistics, John Wiley & Sons, New York, 1994.

Kim, H. and Loh, W.H., *Classification Trees with Unbiased Multiway Splits* (http://www.stat.wisc.edu/p/stat/ftp/pub/loh/treeprogs/cruise/cruise.pdf).

McLachlan, G.J., *Discriminant Analysis and Statistical Pattern Recognition*, John Wiley & Sons, New York, 1992.

Robert, M., Brown, R.M., and Balakrisnama, S.B., *Scenic Beauty Estimation Using Linear Discriminant Analysis* (http://www.isip.msstate.edu/publications/courses/ece_4773/projects/1997/group_scenic/paper/paper.pdf).

Zhang, M.Q., *Discriminant Analysis and its Application in DNA Sequence Motif Recognition* (http://argon.cshl.org/reprints/briefing.pdf).

Chapter 7

Emerging Technologies in Data Mining

7.1 Introduction

Information technology (IT) plays a major role in this fast-changing corporate finance world, where enterprise goals change abruptly. During these uncertain times, decision makers in the corporate world count on their information technology departments to deliver technologies that drive superior enterprise performance. The successes of an organization's business strategy depend on utilizing the right information at the right time. As information is collected and enriched into actionable business intelligence, the challenge becomes making this intelligence readily available to the right people in the appropriate form.

Data warehousing (DW), neural net (NN) applications, and market basket association (MBA) analysis are some of the emerging technologies applied in data mining that can be effectively used to deliver the information. With DW, business enterprises can collect data from any source within or outside the organization, reorganize the data, and provide dynamic storage for efficient utilization. NN, or parallel distributed processing, as it is sometimes called, is an information-processing paradigm that closely resembles the densely interconnected, parallel structure of the mammalian brain. NN techniques include collections of predicting and classification models that emulate biological nervous systems and draw on the analogies of adaptive biological learning. MBA is a computer algorithm that examines many transactions in order to determine which items are most frequently purchased together and provides this valuable information to retail store management for better marketing.

The purpose of this chapter is to introduce briefly the concept of these three emerging technologies and to provide some information regarding the capabilities of the SAS software for performing these analyses. For additional information on data warehousing, see Janes et al.;[1] on neural net applications, Ripley[2] and Bishop;[3] and on market basket association analysis, Berry and Linoff.[4]

7.2 Data Warehousing

Business enterprises of all kinds now computerize all their business activities and their abilities to manage their valuable data resources. Databases 100 gigabytes in size are now common, and terabyte (1000-gigabyte) databases are now feasible in enterprises. Data warehousing techniques enable the forward-thinking business to collect, save, maintain, and retrieve data in a more productive way.

A successful data warehousing operation should have the potential to integrate data from wherever its location and whatever its format. It should provide the business analyst with the ability to quickly and effectively extract data tables, resolve data quality problems, and integrate data from different sources. If the quality of data is questionable, then business users and decision makers cannot trust the results. In order to fully utilize data sources, data warehousing should allow maximum use of current hardware investments, as well as provide options for growth as storage needs expand. Data warehousing systems should not limit customer choices but instead should provide a flexible architecture that accommodates platform-independent storage and distributed processing options.

Data quality is a critical factor for the success of data warehousing projects. If the data are of inferior quality, then the business analysts who query the database and the decision makers who receive the information cannot trust the results. High-quality individual records are necessary to ensure that the data are accurate, updated, and consistently represented in the data warehousing.

7.2.1 Key Concepts in Data Warehousing Features

7.2.1.1 Data Import

Data warehousing should have the potential to manage data tables, parallel storage from scaleable performance data (SPD) servers, multidimensional databases, and hierarchical and relational databases such as DB2, Oracle, SQL Server, and to combine any of these storage structures to satisfy unique business requirements. By utilizing the latest parallel processing and data server capabilities, data warehousing should deliver a fully integrated and seamless way to access large volumes of data. Multidimensional databases (MDDB) are another storage option that are especially useful when providing business users with multiple views of their data through drill-down capabilities. MDDBs are specialized storage facilities where data are pulled from a data warehouse or another data source for storage in a matrix-like format for fast and easy access to multidimensional data views.[5]

7.2.1.2 Extraction, Transformation, and Loading (ETL)

The ETL process consists of all the steps necessary to extract data from their various locations; transform raw operational data into consistent, high-quality business data; and load the data into a data warehouse. Easy and timely access to data, regardless of the data sources or platforms, is the first and most critical step in creating enterprise intelligence. The following are some of the desirable features of the ETL process for efficient data warehousing:

- Complete access to all relevant organizational data residing on diverse platforms and servers in a variety of formats
- Improved performance and reduced network traffic due to the ability to pass database queries
- Cleansing, transforming, analyzing, and presenting data from diverse data sources in accordance with established business rules
- Providing a powerful transformation mechanism that handles everything from validation and scrubbing to integration and structuring to ensure that data in the warehouse conform to established business rules

7.2.1.3 Metadata Creation

Metadata contain information about data stored in data warehousing; thus, metadata provide complete information on the data element, including the source, transformation and summarization, a complete list of dimensions, time frame, and any other pertinent information.[6]

7.2.1.4 Online Analytical Processing (OLAP)

To make sense of changes in the ever-competitive business world, flexibility is required to look at the information from all angles and in different dimensions. OLAP gives business analysts the power to provide solutions to multidimensional business problems quickly and easily. The OLAP technology provides fast, efficient access to summarized data and allows complete control over global views of a business. OLAP technology can be applied to sales and marketing analysis, financial reporting, quality tracking, profitability analysis, and manpower and pricing applications.[7]

Decision makers, regardless of computing expertise, can view business scenarios from a number of perspectives. Using OLAP, analysts can produce data tables, charts, and maps to advance multidimensional reports that might include data visualization and geographical analysis. Analysts can also drill down across data views and take advantage of hot-spotting and traffic-lighting capabilities to identify business trends and long-term developments.[8] To focus on key business issues, OLAP can be used to pinpoint critical success factors and key performance indicators. OLAP technology allows virtual visits to any part of a business, anywhere in the world at any time, to ask complex and multifaceted business questions and then have the answers delivered within seconds. Information can be made more accessible to customers, business partners, and the public to improve business performance and reinforce brand loyalties.

7.3 Artificial Neural Network Methods

The recent explosion of artificial neural net (NN) technology has led data miners to explore a variety of computer engineering applications that did not originate based on traditional statistical theory. Borrowing the concept from the human brain, neural systems fit models by learning in repeated trials to achieve the best prediction. In other words, NN learns from examples. The network is composed of a large number of highly interconnected processing elements (neurons) working in parallel to solve a specific problem. In NN systems, the input, output, and intermediate variables act as nodes that are interconnected by weighted network paths of a network diagram. The input layer contains a unit for each input layer. The output layer represents the target. The hidden layer contains hidden units (neurons) that are the intermediate transformed inputs. The connections in the network path represent the unknown parameter coefficients that are estimated by fitting the model to the data.[3,9]

Many NN applications use the supervised learning approach. For supervised learning, training data that include both the input and the target variables must be provided. After successful training, data can be tested to the NN (that is, input data without the target value), and the NN will compute an output value that approximates the response. If trained successfully, NNs may exhibit generalization beyond the training data and predict correct results for new cases in the validation dataset. However, for successful training, a large amount of training data and lengthy computer training time are essential.

Neural network modeling can be used for both prediction and classification. NN models enable the construction of train and validate multiplayer feed-forward network models for modeling large data and complex interactions with many predictor variables. NN models usually contain more parameters than a typical statistical model, the results are not easily interpreted, and no explicit rationale is given for the prediction. All variables are treated as numeric and all nominal variables are coded as binary. Categorical variables must be encoded into numbers before being given to the network. Relatively more training time is needed to fit the NN models.

The NN models are considered flexible multivariate function estimators. Technically speaking, they are multistage parametric nonlinear regression models and classification models. The most common type of NN model used for supervised prediction is the *multilayer perceptron*, which is the feed-forward NN that uses sigmoid hyperbolic functions.[10] For the mathematical aspects of NN, see Bishop[3] and Hastie et al.[11] For an example of fitting a neural network model using the SAS *Enterprise Miner*, see Johnson and Wichern.[12]

Considerable overlap exists between NN and statistics fields.[13] Feed-forward nets with no hidden layer (including functional-link neural nets and higher order neural nets) are basically generalized linear models. Probabilistic neural nets are identical to kernel discriminant analysis.[9] Kohonen nets for adaptive vector analysis are very similar to *k*-means cluster analysis,[9] and Hebbian learning is closely related to principal component analysis.[9] It is sometimes claimed that neural networks, unlike statistical models, require no distributional assumptions. In fact, neural networks involve exactly the same sort of distributional assumptions as statistical models[3] but statisticians study the consequences and importance of these assumptions while many neural net workers ignore them. Many methods are available in statistical literature that can be used for flexible nonlinear modeling. These methods include polynomial regression, *k*-nearest neighbor regression, kernel regression, and discriminant analysis.

7.4 Market Basket Association Analysis

The objective of market basket association analysis (MBA) is to find out what products and services customers purchase together. Knowing what products people purchase as a group can be very helpful to any business. A retail store could use this information to display products frequently sold together in the same aisle. A web-based Internet merchant could use MBA to determine the layout of their online catalogs. Banks and telephone companies could use the MBA results to determine what new products to offer their prior customers. Once an association rule that customers who buy one product are likely to buy another is known, it is possible for a company to market the products together or to make the purchasers of one product the target prospects for another. This is the purpose of market basket analysis — to improve the effectiveness of marketing and sales tactics using customer data already available to the company. For a non-technical account of MBA and its applications, refer to Berry and Linoff.[14] For a mathematical discussion on association rules used in MBA, refer to Hastie et al.[15] For an example of performing MBA analysis using the SAS *Enterprise Miner*, see SAS Institute.[16]

The strength of market basket analysis is that customers' sales data can provide valuable information regarding what products consumers would logically buy together. This is a good example of data-driven marketing. Market basket analysis offers several advantages over other types of data mining. First of all, it is undirected.

It is not necessary to choose a product on which to focus in order to run a basket analysis. Instead, all products are considered, and the data mining software reveals which products are most important to the analysis. In addition, the results of basket analysis are clear, simple, and understandable association rules that can be utilized immediately for better business advantage.

7.4.1 Benefits of MBA

- **Impulse buying:** Knowing which products sell together can be very useful to any business. The most obvious effect is the increase in sales that a retail store can achieve by reorganizing its products so that things that sell together are found together. This facilitates impulse buying and helps ensure that customers who intend to buy a product do not forget to buy it due to not having seen it.
- **Customer satisfaction:** In addition, MBA has the side effect of improving customer satisfaction. Once they have found one of the items they want, customers do not have to search the store for the other items they want to buy. Their other purchases are already located conveniently close together. Internet merchants get the same benefit by conveniently organizing their website so that items that sell together are found together.
- **Actionable:** Unlike most promotions, advertising based on MBA findings is almost sure to pay off; the business has the data to back it up before even beginning the advertising program. This is an example of the best kind of MBA result.
- **Product bundling:** For companies that do not have a physical store, such as mail-order companies, Internet businesses, and catalog merchants, MBA can be more useful for developing promotions than reorganizing product placement. By offering promotions such that buyers of one item get discounts on another they have been found likely to buy, sales of both items may be increased. In addition, basket analysis can be useful for direct marketers for reducing the number of mailings or calls that need to be made. By calling only customers who have shown themselves likely to want a product, the cost of marketing can be reduced while the response rate is increased.
- **Stock inventory:** It can be useful for operations purposes to know which products sell together in order to stock inventory. Running out of one item can affect sales of associated items; perhaps the reorder point of a product should be based on the inventory levels of several products, rather than just one.

7.4.2 *Limitations of MBA*

Though useful and productive, MBA does have a few limitations. It is necessary to have a large number of real transactions to get meaningful data, but the accuracy of the data is compromised if all of the products do not occur with similar frequency. Second, MBA can sometimes present results that are actually due to the success of previous marketing campaigns. Third, association rules sometimes generated by market basket analysis can be trivial and inexplicable and may not always be useful. A trivial rule is one that would be obvious to anyone with some familiarity with the industry at hand. Inexplicable rules are not obvious and do not lend themselves to immediate marketing use. An inexplicable rule is not necessarily useless, but its business value is not obvious and it does not lend itself to immediate use for cross selling.[14]

7.5 SAS Software: The Leader in Data Mining

SAS Institute,[17] the industry leader in analytical and decision support solutions, offers a comprehensive data mining solution that allows exploration of large quantities of data to discover relationships and patterns that can lead to proactive decision making. SAS software provides the industry's most powerful, easy-to-use, metadata-driven warehouse management and ETL capabilities, with the added value of integrated data quality assessment and monitoring to ensure that the consolidated information is consistent and accurate. Data mining is very effective when it is a part of an integrated enterprise knowledge delivery strategy. SAS Warehouse Administrator (a component of the SAS data warehousing solution), OLAP, NN, and MBA are integrated seamlessly with the SAS *Enterprise Miner* software.[18,19]

7.6 Summary

This chapter briefly introduces the three emerging technologies in data mining, data warehousing, artificial neural net applications, and market basket analysis. SAS Institute, the industry leader in analytical and decision support solutions, offers the powerful *Enterprise Miner* software to perform complete data mining solutions. The SAS data mining solution provides business technologists and quantitative experts the necessary tools to obtain enterprise knowledge for helping their organizations achieve a competitive advantage. SAS macros for performing these three emerging technologies are not included in this book because the *Enterprise Miner* software is required to perform these analyses.

References

1. Janes, H., Dixon, S., and Lewis, T., A data warehouse using the SAS systems, in *Proc. 21st Annu. SAS Users Conf.*, SAS Institute, Cary, NC, 1996, pp. 808–811.
2. Ripley, B.D., *Pattern Recognition and Neural Networks,* Cambridge University Press, Cambridge, U.K., 1996.
3. Bishop, C.M., *Neural Networks for Pattern Recognition*, Oxford University Press, London, 1995.
4. Berry, M.J.A. and Linoff, G.S., *Data Mining Techniques: For Marketing, Sales, and Customer Support*, John Wiley & Sons, New York, 1997, chap. 8.
5. SAS Institute, Inc., *SAS Data Warehousing: A Complete Perspective for Managing Enterprise Data*, SAS Institute, Inc., Cary, NC (http://www.sas.com/technologies/data_warehouse/47395_0102.pdf).
6. Hair, J.E., Anderson, R.E., Tatham, R.L., and Black, W.C., *Multivariate Data Analysis*, 5th ed., Prentice-Hall, Englewood Cliffs, NJ, 1998, chap. 12.
7. Berry, M.J.A. and Linoff, G.S., *Data Mining Techniques: For Marketing, Sales, and Customer Support*, John Wiley & Sons, New York, 1997, chap. 15.
8. SAS Institute, Inc. *Online Analytical Processing (OLAP)*, SAS Institute, Inc., Cary, NC (http://www.sas.com/technologies/olap/).
9. Sarle, W.S., Ed., *Neural Network FAQ: Introduction* (part 1 of 7), periodic posting to the Usenet newsgroup at comp.ai.neural-nets (ftp://ftp.sas.com/pub/neural/FAQ.html).
10. SAS Institute, Inc., *Neural Network Modeling Course Notes*, SAS Institute, Inc., Cary, NC, 2000.
11. Hastie, T., Tibshirani, R., and Friedman, J.J., *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*, Springer Series in Statistics, Springer-Verlag, New York, 2001, chap. 11.
12. Johnson, R.A. and Wichern, D.W., *Applied Multivariate Statistical Analysis*, 5th ed., Prentice-Hall, Englewood Cliffs, NJ, 2002, chap. 11.
13. Sarle, W.S., Neural networks and statistical models, in *Proc. 19th Annu. SAS Users Group Int. Conf.*, SAS Institute, Inc., Cary, NC, 1994, pp. 1538–1550.
14. Berry, M.J.A. and Linoff, G.S., *Data Mining Techniques: For Marketing, Sales, and Customer Support*, John Wiley & Sons, New York, 1997, chap. 8.
15. Hastie, T., Tibshirani, R., and Friedman, J.J., *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*, Springer Series in Statistics, Springer-Verlag, New York, 2001, chap. 14.
16. SAS Institute, *Data Mining Using Enterprise Miner Software: A Case Study Approach*, 1st ed., SAS Institute, Inc., Cary, NC, 2000.
17. SAS Institute, Inc., *The Power To Know,* SAS Institute, Inc., Cary, NC (http://www.sas.com).
18. SAS Institute, Inc., *The Enterprise Miner,* SAS Institute, Inc., Cary, NC (http://www.sas.com/technologies/analytics/datamining/miner/index.html).
19. SAS Institute, Inc., *SAS Enterprise Miner Product Review,* SAS Institute, Inc., Cary, NC (http://www.sas.com/technologies/analytics/datamining/miner/miner_review.pdf).

Further Reading

Brauer, B., *Data Quality: Spinning Straw into Gold*, SAS Institute, Inc., Cary, NC (http://www.sas.com/rnd/warehousing/papers/quality0401.pdf).

Fadalla, A. and Lin, C.H., An analysis of the applications of neural networks in finance, *Interfaces*, 31(4), 112–122, 2001.

Fedenczuk, L.L., *To Neural or Not To Neural? This Is the Question*, SUGI 27 (http://www.bc.edu/bc_org/tvp/research/SAS/To_Neural_or_Not.pdf).

Lajiness, M.S, *A Practical Introduction to the Power of Enterprise Miner*, SUGI 27 (http://www.bc.edu/bc_org/tvp/research/SAS/pract.pdf).

McNelis, P.D. and Nickelsburg, J.J., *Neural Networks and Genetic Algorithms as Tools for Forecasting Demand in Consumer Durables (Automobiles)*, SAS Institute, Inc., Cary, NC (http://www2.sas.com/proceedings/sugi27/p245-27.pdf).

Moorman, M., *Data Warehousing Design Issues for ERP Systems*, SAS Institute, Inc., Cary, NC (http://www.sas.com/rnd/warehousing/papers/erpdesign.pdf).

Sarma, K.S., *Using SAS Enterprise Miner for Forecasting*, SUGI 26 (http://www.bc.edu/bc_org/tvp/research/SAS/Using_SAS_EM_for_Forecast.pdf).

Thomas S., Gruca, T.S., Klemz, B.R., and Petersen, E.A.F., *Mining Sales Data Using a Neural Network Model of Market Response* (http://www.acm.org/sigkdd/ explorations/issue1-1/application.pdf).

Wilson, R.L. and Sharda, R., Bankruptcy prediction using neural networks, *Decision Support Syst.*, 11, 545–557, 1994.

Appendix: Instructions for Using the SAS Macros

Prerequisites for Using the SAS Macros

Read all the instructions given in this Appendix first.

SAS Software Requirements

SAS/CORE, SAS/BASE, SAS/STAT, and SAS/GRAPH must be licensed and installed at the site. SAS/IML is required to run the CHAID macro and to check for multivariate normality in the FACTOR, DISJCLUS, and DISCRIM macros. SAS/ACCESS (PC-file types) is required to convert PC files (Excel, Access, Dbase, etc.) to SAS datasets in the EXCELSAS macro. SAS/QC is required to produce control charts in the UNIVAR macro.

SAS version 8.0 and above is recommended for full utilization; some of the enhanced features may not work in SAS version 6.12.

Internet Requirements

If the companion CD-ROM has not been purchased, a working Internet connection is required for downloading the macro-call files and the sample datasets from the book website. A working Internet connection is also required every time the SAS macros are run because the macro-call files must have access to the SAS macro files from the book website while executing the SAS macros. If the companion CD-ROM is available, a working Internet connection is not required because both the macro-call and the macro files and sample datasets are available on the companion CD-ROM.

System Requirements

The SAS system for Microsoft Windows (98, Me, NT, XP) is required to run these macros. Experienced SAS programmers can simply modify and customize the SAS macro-call and SAS macro files available on the companion CD-ROM for use on other platforms (Apple, Unix, and all other mainframe computers).

SAS Experience

No experience in SAS macros or SAS graphics is necessary to run these macros, but a working knowledge of SAS for Windows and creating temporary and permanent SAS datasets is helpful.

Instructions for Downloading Macro-Call Files

1. Visit the book website at http://www.ag.unr.edu/gf/dm.html.
2. Click the download link to be directed to the password-protected macro-call download page.
3. Input the following username (lower case only) and password to go to the download page:
   Username: Please refer to the book for the **Username***
   Password: Please refer to book for the **Password***
4. Click the download link, download the zipped file "dm.zip", and save it in a folder on the PC.
5. Unzip the "dm.zip" file on the PC using any unzip program. The zipped file contains two folders, "sasdata" and "mac-call", and one text file, "README.TXT". The "sasdata" folder contains the Excel data files and the permanent SAS data files used in the book. In the mac-call folder are 13 macro-call files corresponding to the 13 data mining macros described in the book. Do *not* change or modify the contents of these macro-call files. Read the "README.TXT" file for the version number and any update information.
6. Visit the book website at least once a month for any news about update information.

Instructions for Running the SAS Macros

Running these macros using the sample data included in the "sasdata" folder before trying these macros on your own data is highly recommended.

*Please check the hard copy of the book for Username and Password for downloading the macros.

Also, disable the SAS ENHANCED EDITOR window in version 8.2 temporarily by clicking TOOLS→OPTIONS→PREFERENCES→EDIT and unchecking the ENHANCED EDITOR box. Disabling the ENHANCED EDITOR will ensure smooth and less complicated execution of these macros.

Option 1: Downloadable Macros

1. Verify an active Internet connection by browsing the book website to see if you can access it.
2. Create a temporary SAS dataset using one of the sample permanent datasets. For example, to create a temporary dataset called "train" from the permanent dataset "sales" saved in the d:\sasdata\ folder, type the following statements in the SAS PROGRAM EDITOR window:

   ```
   LIBNAME GF d:\ sasdata ;
   /* Assign a libname GF to the sasdata folder
   containing the sample data files*/
   DATA train;
   SET GF.sales;
   RUN;
   ```

3. Click the RUN button to create a temporary dataset called "train" from the permanent dataset "sales" saved in the "sasdata" folder. For example, to run multiple linear regression, click the program window, open the file "REGDIAG.sas" in the program editor (do *not* make any changes to the macro-call file), and click the RUN icon to open the cyan-color macro-call window REGDIAG. Check the LOG window and make sure the macro-call file accessed the corresponding macro from the book website without any problems. Following the instructions given in the specific help file for the REGDIAG macro (Chapter 5), input the necessary macro-input values. When the cursor blinks at the last macro field, hit the ENTER key (not the RUN icon) to execute the macro. To check for any macro execution errors in the LOG window, always run with the DISPLAY option first. Ignore any warnings related to font substitution, as these font specifications are system specific. If no macro-execution errors are reported, then save the output and graphics by changing DISPLAY to the desired file formats (WORD, WEB, PDF, and TXT).
4. Read the specific chapter and macro-help files for specific details.