The LEGACY *of* FISCHER BLACK

Edited by

BRUCE N. LEHMANN

THE LEGACY OF

FISCHER BLACK

This page intentionally left blank

THE LEGACY OF

FISCHER BLACK

Edited by
Bruce N. Lehmann

OXFORD
UNIVERSITY PRESS

2005

OXFORD
UNIVERSITY PRESS

Oxford New York
Auckland Bangkok Buenos Aires Cape Town Chennai
Dar es Salaam Delhi Hong Kong Istanbul Karachi Kolkata
Kuala Lumpur Madrid Melbourne Mexico City Mumbai Nairobi
São Paulo Shanghai Taipei Tokyo Toronto

Published by Oxford University Press, Inc.
198 Madison Avenue, New York, New York 10016

www.oup.com

Library of Congress Cataloging-in-Publication Data
Lehmann, Bruce Neal, 1955–
The legacy of Fischer Black / edited by Bruce N. Lehmann.
p. cm.
ISBN 0-19-516836-4
1. Finance—Congresses. I. Title.
HG63 .L44 2004
332—dc22 2003018743

1 3 5 7 9 8 6 4 2

Printed in the United States of America
on acid-free paper

Preface

Between September 29 and October 1, 1996, the Berkeley Program in Finance held a conference called "On Finance: A Conference in Honor of Fischer Black" at the Breakers in Santa Barbara, California. Organized by David Modest, then a finance professor at the University of California, Berkeley, and presently a Managing Director and Chief Risk Officer at Azimuth Trust, the conference brought together a number of luminaries, including a number of past and future Nobel Laureates. Past presidents of the American Finance Association made up more than half of the paper presenters and discussants, and virtually all of those who have not (yet) received this honor have edited the profession's major journals or presided over its two other principal organizations, the Western Finance Association and the Society for Financial Studies. The assembled multitude was most assuredly not a collection of lightweights.

From the outset, a conference volume was planned to honor Fischer. (We refer to Fischer Black throughout by first name, rather than by the more usual—and less personal—last name.) Eight of the eleven talks given at the conference are chapters in this volume, which is rounded out by my attempted synthesis of Fischer's many contributions to the economics of asset pricing, Myron Scholes's thoughtful reflections on credit and risk management, and Darrell Duffie's appreciation of the Nobel prize winning research of Black, Merton, and Scholes. For a variety of reasons, the task of editing this volume and shepherding it through the production process eventually fell to me. Some tasks proved to be more pleasant than others, and the preparation of this introduction ranks among the most enjoyable.

My own essay, "Fischer Black on Valuation: The CAPM in General Equilibrium," was not presented at the conference. Its main theme is the many dimensions in which the two main features of capital market equilibrium in frictionless markets—that assets in positive supply must be held in equilibrium while buyers must balance sellers in the case of zero net supply assets such as options, futures, and other derivative securities—constitute the central insight of modern asset-pricing theory. Systematically viewing Fischer's research on capital asset pricing through this lens provides an organizing principle for discussing his sweeping insights. However, it provides so much more, the basis for an explication of the economics of asset pricing itself. While my essay may well have fallen short in providing the kind of thorough and comprehensive exposition that this perspective invites,

writing it has crystallized so much of the simplicity and elegance of modern finance, at least for me.

Stewart C. Myers, Gordon Y. Billard Professor of Finance at the Sloan School of Management at MIT, is uniquely positioned to survey Fischer's contributions to corporate finance. He was one of Fischer's colleagues during his time at MIT and is one of the leading scholars in corporate finance, one possessed of the breadth to appreciate Fischer's many and diverse contributions to the field and of a sufficiently practical bent to recognize the way in which they can and should inform corporate financial practice. His chapter (chapter 2) is replete with illustrations of the pervasiveness of Fischer's influence in this field: the Capital Asset Pricing Model (CAPM) and the measurement of the opportunity cost of capital, financial option pricing theory and the valuation of both the liabilities of business firms and the real options embedded in their projects, and his work on such diverse topics as dividend policy, asset/liability management in pension funds, and the proper interpretation and construction of earnings.

Myron S. Scholes, Frank E. Buck Professor of Finance, Emeritus, at Stanford University, Chairman of Oak Hill Platinum Partners, and 1997 Nobel Laureate in Economic Science, worked with Fischer on several of their most influential research projects and, in particular, on the celebrated Black-Scholes option pricing model. His chapter is devoted to a most topical subject: the interaction among major liquidity crises and credit and risk management considerations such as that associated with the Asian financial crisis in 1998. Myron is uniquely positioned to analyze these circumstances because he was a principal in Long-Term Capital Management (LTCM), the hedge find that went under at the tail end of the crisis. Myron traces LTCM's problems and those of similarly situated financial services firms to the dynamics of the market for liquidity for illiquid assets. Such firms earned extraordinary returns prior to the Asian financial crisis by being substantial suppliers of liquidity in such markets and by hedging their positions for "normal" fluctuations in the underlying sources of risk in their positions. These liquidity suppliers experienced large losses during the crisis and their leverage discipline required them to liquidate large positions rapidly. Hence, these firms became large demanders of liquidity at a time in which there were no natural new sources of liquidity. Moreover, market liquidity premiums rose substantially on many illiquid assets at the same time, resulting in extremely high correlation across their positions that reduced or eliminated the benefits of diversification across assets and countries and dramatically impaired the performance of their hedges. Myron notes that one consequence of the crisis has been a dramatic increase in liquidity premiums across many markets engendered by the reduction in the number of liquidity suppliers such as LTCM and predicts the rise of new institutional arrangements for the contingent supply of liquidity during times of unusual market stress.

Robert B. Litterman is Managing Director and Director of Quantitative Resources at Goldman Sachs & Co. He worked extensively with Fischer

throughout his time at Goldman Sachs, culminating in the development of the Black–Litterman Global Asset Allocation Model, a tool still widely applied in the asset allocation process at Goldman Sachs, which is referred to briefly in section 4 of my essay. Bob modestly describes his chapter as "a tutorial on portfolio risk management." It is instead a sweeping reinterpretation of Modern Portfolio Theory in the new language of risk budgeting. As such, it represents a substantial original contribution to an important literature, one possessed of the sort of mixture of analytical rigor and practical usefulness that Fischer so often championed. Evidence for this view may be found in its widespread citation in texts, practitioner journals, and the concept releases of regulatory organizations.

Stephen A. Ross, Franco Modigliani Professor of Financial Economics at the Sloan School of Management at MIT, addresses a subject of great interest to Fischer: the economics of the industrial organization of the market for portfolio management services. In his unpublished paper "The Future for Financial Services" (October 1982), Fischer set forth his views on the role of securities firms, mutual funds, and sophisticated individuals in a world with costly information, management, and marketing.[1] Steve explores the role of the market for mutual fund managers in a world in which it is hard to distinguish lucky managers from skillful ones and unlucky managers from unskillful ones, a problem that is exacerbated by the option-like structure of management compensation. The resulting industrial organization is one in which the skillful, the lucky, and even some of the unskillful and unlucky survive in equilibrium due to a combination of the rewards for ex post performance and ex ante diversification.

Mark Rubinstein, Paul Stephens Professor of Applied Investment Analysis at the Haas School of Business at the University of California at Berkeley, and Jens Jackwerth, Professor of Finance at the University of Konstanz, study methods for drawing inferences regarding the probability beliefs and risk aversion of investors implicit in option prices. One can think of it as the kind of exploration of general equilibrium of the sort favored by Fischer, one rich in the quantitative detail that only option data can provide regarding implicit state prices. While they find it relatively easy to infer implicit state prices from option data in a robust manner, Jackwerth and Rubinstein find modeling two aspects of the distribution of state prices more challenging. First, there are many, economically quite different implied time series models for the S&P 500 that are roughly equally compatible with its volatility smile. Second, there is no representative investor with constant relative risk aversion compatible with observed option prices. As the authors suggest, the latter finding suggests a role for state-dependent preferences for reasons touched on briefly in section 6 of my essay.

Michael Brennan, Emeritus Professor of Finance at the University of California, Los Angeles, who held the Irwin and Goldyne Hearsh Chair in Banking and Finance until his retirement in 2002, and Professor of Finance at the London Business School, Avanidhar Subrahmanyam, Professor of Finance at the Anderson Graduate School of Management at the University

of California, Los Angeles, and Tarun Chordia, Associate Professor of Finance and Caldwell Research Fellow at the Goizueta Business School at Emory University, attempt to distinguish between two general explanations of cross-sectional variation in expected returns. One such explanation is risk-based, for which their baseline model is Connor and Korajczyk's implementation of the Arbitrage Pricing Theory (APT). Against this model, they array a comprehensive list of security attributes that have been found to be correlated with expected returns in prior empirical work. These characteristics come in several flavors, with some thought by many financial economists to be probable determinants of risk exposures, some to be measures of the liquidity of the market for immediacy in individual assets, and others to be anomalies that help explain deviations from the CAPM but that cannot be sorted easily into economically plausible categories. While some of my more specific thoughts on their methods and results may be found in my own commentary, prepared at David Modest's request, in the chapter that follows theirs, their general finding that liquidity measures help account for risk-adjusted expected returns while anomaly variables do not suggests that asset pricing theorists should think hard about the role of liquidity in security price determination. This suggestion is compatible with Fischer's view of the role of market microstructure as expressed, for example, in his 1985 American Finance Association Presidential Address (Black 1986, 1985).

The chapter by Scott F. Richard, Managing Director of Miller, Anderson, and Sherrerd and Morgan Stanley Asset Management, represents the kind of applied research that Fischer thought should be the main preoccupation of students of financial economics. It indirectly provides a measure of the pervasive influence of Fischer and others on the practice of finance; fixed income analysts, traders, and investment professionals now routinely use the language of stochastic differential equations to describe and analyze interest rate dynamics. On the surface, the chapter's main contribution is the extension of the one-factor Black–Karasinski term structure model to a two-factor model, but this pat description does not do justice to the economic insight gained in the passage from one to two factors. The model retains much of the simplicity of the Black–Karasinski model while building in the second, negatively correlated factor for which empirical term structure dynamics and those in the associated derivative asset markets scream, making it a valuable tool for pricing and hedging interest rate contingent claims. What is noteworthy about the analysis is the way in which it employs just the amount of technique required to incorporate simple intuitions about the second factor, a strategy of which Fischer would have doubtless approved.

The chapter by Douglas T. Breeden, Dean and William W. Priest Professor of Finance at the Fuqua School of Business at Duke Univesity and the Chairman and Co-founder of Breeden Associates, Inc., is also the kind of applied research of which Fischer approved. The subject of the chapter is one of the most nettlesome analytical problems in applied finance: mod-

eling the extremely puzzling mortgage refinancing and prepayment behavior of households. Breeden weaves theory and an array of empirical observations on both mortgage-backed security returns and broker forecasts of their prospective behavior into a compelling analytic description. What is appealing about the analysis is the way that it characterizes the sources of risk and return in this market in a simple and clear way unencumbered by formal technique that would have added equations, but not insight, into the behavior of the mortgage market.

Hans Stoll, Anne Marie and Thomas B. Walker Professor of Finance at the Owen Graduate School of Management at Vanderbilt University, and Roger Huang, Kenneth R. Meyer Professor of Global Investment Management at the Mendoza College of Business at Notre Dame University, analyze the link between implicit measures of the revenues earned by suppliers of immediacy such as New York Stock Exchange specialists, proprietary traders at securities firms, and dealers and the execution costs paid by demanders of immediacy among public investors. Data on these revenues and costs come from entirely unrelated sources and so it is quite surprising that Hans and Roger find them to be commensurate. It is also fortuitous because they can use these data to make indirect inferences about the costs and benefits of limit and market orders, a subject of great interest to Fischer in the last half decade of his life. They find that securities firms earn more per share than the revenue on an average trade, which implies that public limit orders—that is, the limit orders placed by liquidity-demanding public investors as opposed to suppliers of immediacy—earn less, the price paid for exposing their orders to the risk of getting "picked off" by informed investors. They also provide crude evidence that the costs paid by public limit order traders are comparable to those of market orders, suggesting that public investors equalize their marginal costs of trading across the two order types.

Darrell Duffie, James I. Miller Professor of Finance at the Graduate School of Business at Stanford University, provides the final chapter of this collection, an appreciation of the contributions to economics made by Fischer Black, Robert C. Merton, and Myron S. Scholes occasioned by the receipt of the 1997 Alfred Nobel Memorial Prize in Economic Science by Merton and Scholes. Had he not died prematurely from cancer, there is no question that Fischer would have shared the prize with them. Treating the research of all three simultaneously confers what amounts to a pedagogical benefit: the ability to explicate the way in which their separate and joint contributions to our understanding of both the equilibrium and arbitrage-free valuation of the cash flows from different assets have permeated the rapid theoretical advances made in financial economics over the last three decades. Darrell has been a keen student of and contributor to many of the developments that followed from their pathbreaking work, which makes him the perfect person to place it in its proper context.

Fischer Black was a remarkable social scientist, one whose contributions range from the lofty perch of highbrow theory to the trenches of practical

application. The chapters that follow span the same range, representing the contributions of a remarkable array of financial economists who embody in different ways Fischer's ideal of insight from economic theory that both guides and is rooted in his kind of detailed observation of relevant aspects of actual financial markets. As such, their efforts constitute a living tribute to this ideal, a reflection of the honor that the many conference participants sought to bestow on him. Fischer never struck me as one for whom such tributes would have mattered much; I am not even sure how he would have viewed the Nobel Prize he would have doubtless shared had he not died prematurely. Nevertheless, I hope that readers find this volume to be both a fitting tribute and a stimulus to further research. After all, the advancement of economic science remained a constant goal throughout Fischer's remarkable career in the many and disparate venues in which he plied his trade.

NOTES

1. As summarized in his abstract, Fischer predicted that "Most individuals will own securities through internally managed funds which will play an active role in the firms whose shares they own. Only sophisticated individuals will own or trade stocks and bonds of firms other than mutual funds. Securities firms will sell to mutual funds and sophisticated individuals." Many pension funds are, but most mutual funds are not, active shareholders. Through my rose-colored glasses, hedge funds today fill the niche Fischer predicted for sophisticated individuals. Parenthetically, I recall Fischer saying that managers should be paid in accord with the quality of their reasoning regarding the strategies they employed and not for their ex post performance.

Contents

1. Fischer Black on Valuation: The CAPM in General Equilibrium, 3
Bruce N. Lehmann

2. Fischer Black's Contributions to Corporate Finance, 33
Stewart C. Myers

3. Crisis and Risk Management, 48
Myron S. Scholes

4. Hot Spots and Hedges, 55
Robert Litterman

5. Markets for Agents: Fund Management, 96
Stephen A. Ross

6. Recovering Probabilities and Risk Aversion from Options Prices and Realized Returns, 125
Mark Rubinstein and Jens Jackwerth

7. Cross-Sectional Determinants of Expected Returns, 161
Michael Brennan, Tarun Chordia, and Avanidhar Subrahmanyam

8. On Cross-Sectional Determinants of Expected Returns, 187
Bruce N. Lehmann

9. Exploring a Two-Factor, Markovian, Lognormal Model of the Term Structure of Interest Rates, 197
Scott F. Richard

10. Convexity and Empirical Option Costs of Mortgage Securities, 212
Douglas T. Breeden

11. The Supply and Demand of Immediacy:
Evidence from the NYSE, 256
Roger D. Huang and Hans R. Stoll

12. Black, Merton, and Scholes—Their Central Contributions
to Economics, 286
Darrell Duffie

Index, 299

THE LEGACY OF

FISCHER BLACK

This page intentionally left blank

1

Fischer Black on Valuation: The CAPM in General Equilibrium

Bruce N. Lehmann

1. THE SPAN OF FISCHER BLACK'S CONTRIBUTIONS TO FINANCIAL ECONOMICS

Sitting before me is a copy of *A Bibliography of Finance* by Richard Brealey and Helen Edwards, which lists 12,037 papers on financial economics published in approximately 120 journals sorted into 40 subject areas. It lists 26 papers written by Fischer Black, covering 14 of those categories. Several papers could have easily been sorted into eight additional subject areas, bringing Black's coverage up to 22 categories. Papers by Fischer published after 1989, those omitted from the bibliography because they were published as book chapters, and his two books increase the coverage to 29 subject areas. Perusal of my incomplete copies of his two self-published series of papers entitled *Fischer Black on Markets* and *Fischer Black on Options* adds three more categories for a grand total of 32. The eight uncovered subject areas are (1) social responsibility, (2) new securities issues, (3) splits and stock dividends, (4) leasing and project finance, (5) working capital, (6) mergers and corporate restructuring, (7) regulated industries, and (8) other speculative markets. I am sure I have overlooked work by Fischer devoted to some of these topics as well, particularly the last three areas.

This number—that is, 32 out of 40 subject areas—is an indicator of the difficulty of my task here: writing a cogent and reasonably thorough synthesis of Fischer Black's research on finance as an introduction to the compendium of chapters that follow. In a world in which economists in general and financial economists in particular specialize in increasingly narrow subfields, Fischer's intellectual legacy stands as one of great breadth, covering virtually all of modern finance and reaching into areas such as general equilibrium theory and macroeconomics. Couple this observation with the fact that Fischer advanced more than the academic enterprise through his lifelong participation in real-world financial markets—culminating in his twelve year stint at Goldman Sachs—and this is a tall order indeed.

Stewart Myers' insightful review of Fischer's contributions to corporate finance (chapter 2) in this volume simplifies the problem before me

considerably. As noted in the preface, Fischer's many theoretical and empirical contributions to research on capital markets had intended and unintended effects on the theory and practice of corporate finance. Stew discusses their full range, making it unnecessary and, indeed, probably undesirable for me to address this body of work.

Accordingly, I shall complement Stew's chapter by focusing on Fischer's main professional preoccupation: the economics of the determination of the prices of risky assets in real-world capital markets. While exchanges with him were sometimes opaque, as is often the case with nonlinear thinkers, Fischer held internally consistent views about the appropriate way to use economic reasoning in the analysis of financial markets. He thought the essence of equilibrium in the market for risky assets—that all assets in positive supply must be held by someone while investors with long positions in zero net supply assets must be balanced by investors with short positions—provided special insight into valuation. There are a variety of settings in which simple application of these necessary equilibrium conditions produces alternative versions of the Capital Asset Pricing Model (CAPM).

My goal is to catalog a number of these economic environments in order to explicate some of the insights obtained by Fischer from the various CAPMs. The sections that follow proceed through the following list: a simple general equilibrium version of the Sharpe–Lintner CAPM, the zero beta CAPM, a few international CAPMs, the CAPM and option pricing, and variations on the intertemporal CAPM (ICAPM) in general equilibrium. In some places, I aim at little more than concise and clear explanation. At others, I stray from this perhaps more appropriate path and attempt to provide more of a synthesis of the underlying economics and its implications for drawing inferences about the nature of financial market equilibrium. At all times, my goal is to explicate Fischer's views, uncorrupted by my own. Irrespective of my success or failure in any of these dimensions, I have used prose only, with no mathematics, at the potential cost of either oversimplifying the relevant theory or glossing over the more troublesome analytical challenges.

This essay is not a memoir since I was certainly not one of Fischer's close colleagues. Nor is it a piece of intellectual history, for I am no historian of science. Rather, I view it as a model of Fischer's evolving thoughts on valuation, one that is empirically accurate at the outset (almost by construction) but that becomes more speculative as it progresses. This model helped me to organize my thinking about his research and to develop a renewed appreciation for its many insights. I hope it produces at least a fraction of the aesthetic satisfaction for readers that its preparation has given me.

2. THE SHARPE–LINTNER–MOSSIN–TREYNOR CAPM IN GENERAL EQUILIBRIUM: BLACK (1972A)

Perhaps counterintuitively, I think the best place to begin is with a little-known and even less cited paper by Fischer entitled "Equilibrium in the

Creation of Investment Goods under Uncertainty" (Black 1972a) that was published in the Jensen volume that housed the far more famous Black, Jensen, and Scholes (1972) paper. The importance I attach to it is perhaps supported indirectly by the following observation: it was the only self-cited paper in Black (1995). The model in question is quite a simple one: a stripped-down general equilibrium model in which wheat can either be eaten now or planted now to be eaten (or planted) next period. Nevertheless, it is a full equilibrium model, perhaps the simplest one in which consumption, production, the riskless rate of interest, and the prices of all risky assets are determined by the interplay of supply and demand. Fischer remained preoccupied with general equilibrium versions of the CAPM throughout his life.

The driving force of the model is the reward for planting wheat. All wheat seed has the same value in consumption, but yields vary randomly across types of wheat. The expected crop yield for all types is well above a bushel of output per bushel planted, while the standard deviation is substantially below the mean. This device makes it possible to assume that crop yields are jointly normally distributed with otherwise arbitrary means, variances, and covariances for valuation purposes while simultaneously truncating actual crop production a bit above zero so that the resulting equilibrium can respect positivity of consumption and output. Wheat yields are taken to be independently distributed over time, removing any hedging demands that might arise if the distribution of wheat output—that is, the investment opportunity set—were stochastic. Accordingly, the investment side of the savings decision is atemporal.

Investor preferences are defined over initial consumption and (random) end-of-period wealth, thus converting a potential multiperiod consumption–investment program into a two-period problem with a single period of uncertainty. Each investor's utility function is otherwise unrestricted except for the assumption of sufficient risk aversion to ensure that no investor borrows enough to risk bankruptcy, making all borrowing and lending risk-free in the absence of moral hazard. The consumer choice problem is closed by endowing each investor with initial quantities of some or all of the types of wheat and with the same beliefs regarding their random yields. Note that there is no need for rational expectations at this juncture.

The last element of the model is the structure of its financial markets. As is commonplace in capital market theory, markets are presumed to be frictionless. That is, there are no taxes, transaction costs, or other impediments to trade in wheat seed. The assets that trade in this market are the wheat seed itself or shares in crop production. The latter interpretation makes it easier to think of the model as empirically relevant.

The structure of the equilibrium devolves from some basic facts about this economy. Set the price of consumption today and end-of-period wealth to 1, which is just a normalization since only relative prices are determined in equilibrium. Since all wheat is identical for consumption purposes, any type of wheat that is planted must have a price greater than or equal to

unity. The returns of such wheat types are (approximately) jointly normally distributed, so investors judge wheat portfolios solely on the basis of their means and variances. Hence, they will choose to hold mean–variance-efficient wheat portfolios—those portfolios with returns that have minimum variance for each given level of mean returns.

It is easy to see that the resulting equilibrium is the CAPM. The mean–variance-efficient set is a hyperbola in mean–standard-deviation space. For a given riskless rate, there is a unique mean–variance-efficient portfolio—the so-called tangency portfolio—on the line passing through the riskless rate that is tangent to this hyperbola at this point, the so-called Capital Market Line. This tangency portfolio must be the market portfolio of risky wheat seed since the riskless asset is in zero net supply and the market portfolio must be held in equilibrium. That is, prices must adjust until the value-weighted portfolio of wheat seed shares is this particular mean–variance-efficient portfolio. The equilibrium interest rate that clears the money market might be negative, but the addition of riskless storage to the model eliminates this possibility. Note that the market portfolio will lie on the Capital Market Line between the tangency portfolio and the riskless rate if the riskless asset—that is, riskless storage—is in positive supply.

Capital market theory is extraordinarily successful in this world because the optimal portfolio of each investor is a combination of riskless borrowing or lending and a position in the wheat market portfolio. The reason is simple and beautiful: the marginal contribution of any type of wheat seed to the expected excess return of the market portfolio exactly equals its marginal contribution to market portfolio risk. As Fischer put it: ". . . The optimal portfolio for any investor could depend on expected excess returns and covariances among excess returns on all available assets. In equilibrium, though, the optimal portfolio for any investor is a mix of the market portfolio with borrowing or lending. The expected returns and covariances cancel one another, so they do not affect the investor's optimal holdings" (Black 1990a, p. 904).

Perhaps more importantly, all investors agree on the fundamentals of risk and return in this world. They define, measure, and manage asset risk as measured by systematic or nondiversifiable risk exposure—that is, by beta. They understand that only systematic risk should be compensated in equilibrium—no investor need bear unsystematic or diversifiable risk and, hence, should not be rewarded for doing so. They even agree on how to evaluate the performance of a managed portfolio. As long as the market portfolio return is measured sufficiently well, systematic risk exposure estimates will be reasonably precise since covariances among long-lived asset returns have modest sampling error.

There are two insights from this simple general equilibrium model that stayed with Fischer throughout his career. The first one was the extraordinary simplicity of the CAPM equilibrium. The market portfolio must be held in equilibrium by someone—in this case, by the investor with average

risk aversion—while zero net supply assets must not be held by this investor. His subsequent models shared this feature, resulting either in proofs as brief as that sketched two paragraphs earlier or in no proofs at all.

A second, more subtle feature of such equilibria became clearer to him over time. Partial equilibrium models and general equilibrium pure exchange models treat the supply of assets and their stochastic payoffs as fixed, and we often fall into the habit of treating the corresponding properties of observed assets as fixed as well. However, the assets that are supplied in equilibrium presumably reflect some balance between investor characteristics such as wealth and risk aversion and aspects of production possibilities such as the productivity of alternative risky technologies. A full equilibrium model must account for both existing asset returns and the assets that exist.

3. THE ZERO BETA CAPM: BLACK (1972B, 1993A), BLACK, JENSEN, AND SCHOLES (1972), BLACK AND SCHOLES (1974A, B)

Central to the Sharpe–Lintner CAPM was the assumption of rational homogeneous beliefs about presumed normally distributed asset returns on the part of investors postulated to maximize the expected utility of end-of-period wealth who could freely borrow or lend at the riskless interest rate. Each of these assumptions was potentially problematic, with some areas of apparently greater concern at the time than others. Fischer thought that the results of Lintner (1969) largely eliminated any important role for heterogeneous expectations (as opposed to asymmetric information), although it is worth noting at the cost of potentially obscuring the link between the ex ante beliefs of investors and the corresponding ex post moments with which the theory might be tested. He also came to argue that "Differences in beliefs must ultimately derive from differences in information" (Black 1986, p. 531), for reasons discussed below.

The assumption of multivariate normality for asset returns might have seemed strained for returns on an annual (or other discrete interval) basis, but Fischer found it natural to think of the CAPM as applying at each instant of time. The reformulation of the model in continuous time can mitigate this concern. Without altering the assumption of time-consuming roundabout production, such a model must be based on information flows about crop yield prospects and, in conjunction with the stochastic process generating wheat yields, must be such that the returns of wheat seed claims are (approximately) jointly normally and independently distributed.

This circumstance will arise when information about crop prospects arrives sufficiently smoothly in the market during the year so that wheat seed values on nearby dates are approximately normally distributed. Discrete chunks of information—say, for example, the harvest itself or periodic crop reports by the U.S. Department of Agriculture that contain unexpected information—can be accommodated as long as they arrive predictably in sufficiently normal increments. Finally, any discrete or smooth information

arrival process cannot create hedging demands without disturbing the CAPM equilibrium, so the parameters of these processes must be sufficiently stable or independently distributed over time as well. The latter circumstance will arise, for example, when the distribution of wheat output at the end of year t depends on random parameters drawn independently from some distribution at the end of year t–1.

Stochastic information flows of this sort take care of potential sources of time variation in risk and return on the supply side but not those that arise on the demand side if investors care about more than consumption and wealth. One minor complication involves the objects of choice over the crop year. It is easiest to retain the dating conventions of the original model since the omission of intermediate consumption is mostly a detail. Investors derive utility from consumption at the beginning and wealth at the end of the crop year and, hence, care about the implications for end-of-year wealth of within-year changes and, specifically, of wealth fluctuations from instant to instant.

More substantively, Fischer came to believe that state-dependent preferences were economically important, but I do not think that they seemed so at the time. In particular, Fama (1970) showed that investors acted as if their utility was derived from current consumption and the next period's wealth in the presence of constant investment opportunity sets such as that described above and in the absence of taste shocks to lifetime utility. In an era before the widespread search for utility-based explanations for consumption smoothing, taste shocks and, hence, potentially state-dependent preferences probably seemed like second-order concerns.

This leaves the risk-free borrowing and lending assumption as the primary concern and the progenitor of Fischer's zero beta version of the CAPM. Like all of his theoretical work, Fischer was partially motivated by empirical evidence. The evidence in question was the Black, Jensen, and Scholes (1972) and Black and Scholes (1974b) observation that the empirical Security Market Line was much flatter than predicted, with the intercept systematically above the risk-free rate, the value predicted by the original Sharpe–Lintner CAPM. The zero beta CAPM could neatly explain both regularities without fundamentally altering the CAPM equilibrium and pricing by the Security Market Line.

The zero beta CAPM dispenses with the assumption that investors can borrow and lend freely at the risk-free rate but retains the remaining structure, including the presumption that investors can freely sell short unlimited quantities of risky assets. The normality of wheat yields still leads investors to choose mean–variance-efficient portfolios. Since it must be held in equilibrium, the market portfolio must be mean–variance efficient. Since short sales must be in zero net supply, the market portfolio must be the tangency portfolio.

However, the intercept of the line tangent to the mean–standard-deviation frontier is no longer the risk-free rate. Rather, it is the common expected return of all assets with returns uncorrelated with those of the market

portfolio, the equality arising from the market portfolio's mean–variance efficiency. There is a unique minimum-variance portfolio with this property, eventually termed the beta portfolio by Fischer and the zero beta portfolio by virtually everyone else. Since investors choose only mean–variance-efficient portfolios, investors who wish to take more risk than that of the market portfolio will fund greater than 100% investment in the market portfolio by selling the zero beta portfolio short. Since short sales are in zero net supply, the short positions of these investors will be balanced in equilibrium by long positions in the zero beta portfolio taken by investors who seek less risk exposure than that of the market portfolio.

Fischer systematically viewed empirical evidence on asset prices, especially that on cross-sectional variation in expected returns, through the lens of the Sharpe–Lintner and zero beta CAPMs, and he had a variety of reasons for doing so. As the theorist in him evolved, a constant was his belief that capital market equilibrium should be modeled as if financial markets were complete, as though investors could hedge against any contingency confronting them by trading in existing financial assets. If financial markets behave like complete ones, equilibrium prices behave as if they were set to satisfy the asset demands of a single representative investor. The optimal portfolio of this investor must be the market portfolio of all risky assets, and Fischer's continued faith in the CAPM presumably reflected, in part, a feeling that this hypothetical representative investor behaved as though the CAPM is true.

A second reason was Black's empirical sense that a more sophisticated model is not needed to extract the signal in the high-volatility noise of long-lived asset returns. He suggested that this volatility could arise from rational forecast revisions and variation in expected returns or from noisy valuation due to irrational fluctuations in investor psychology. Irrespective of its source, this noise makes it hard to make empirically reliable statements about the overall level of risk premiums. For example, the two standard error confidence interval for the mean excess return of the S&P 500 remained a bit higher than 10% per year over the last seven decades, while that of long-term U.S. Treasury bonds exceeded 4% per year, quite large numbers compared to the corresponding mean excess returns themselves of 8.6% and 1.4%, respectively. Of course, cross-sectional differences are measured somewhat, but only somewhat, better, but the CAPM does a tolerably good, if imperfect, job of accounting for such regularities.

Black remained suspicious of the comparative empirical success of Arbitrage Pricing Theory (APT) models in accounting for cross-sectional regularities for these reasons. These models are based on the hypothesis that factors that account for most of the covariation within a large asset menu must, in a sense that can be made precise, account for cross-sectional variation in their expected returns. They predict neither the signs nor the magnitudes of the associated factor risk premiums, leaving them to be determined empirically as part of the fitting of the model. Hence, its "comparative empirical success" may simply reflect the ease of finding factor risk pre-

mium estimates that generate relatively good fits in samples from economies that generate noisy equilibrium risk premiums. Black feared that the unobservability of the market portfolio caused similar problems for the CAPM, but he thought it likely that true market portfolio returns were highly correlated with those of any reasonable proxy, to some extent mitigating concerns regarding the comparative empirical success of the CAPM.

Fischer, an early and ardent critic of apparent data mining in finance, was even more suspicious of the so-called anomalies that have plagued the CAPM. The variables found to help explain cross-sectional variation in average returns were not suggested by a priori theory but rather by empirical exploration guided at best by loose theory. In fact, the variables found in the literature were part of Wall Street folklore, suggesting that they were selected by data miners in the real world long before we had the benefit of computerized databases. Moreover, the small-firm effect, the first major anomaly not readily attributed to statistical problems, largely disappeared from U.S. equity markets after its discovery, lending credence to the data-mining interpretation of its prior significance (although the evidence is also compatible with a market microstructure explanation).

In fact, Fischer long thought that the major empirical question regarding the cross section of expected returns was whether the Sharpe–Lintner or zero beta CAPM better described markets. Three possible explanations were well-expressed early on in Black and Scholes (1974a, p. 405):

> We really don't know whether to believe the theory or the data. It may be that the extra returns on low risk portfolios in the postwar period were simply due to chance. Or it may be that the use of a more comprehensive market portfolio would give different results. Or that the results were due to restrictions on borrowing that were effective for a time, but have now lost their effectiveness.

He later changed this last opinion to "These restrictions have probably tightened in recent decades. . . . Many countries outside the United States seem to have similar restrictions" (Black 1993a, p. 10).

That said, Fischer favored another interpretation. He did not believe the zero beta CAPM because many investors have access to direct or indirect riskless borrowing opportunities. This view also jibed with his belief that financial markets should be modeled as though they were complete, in which case a risk-free asset could be constructed from existing assets. One possibility was that both investors and business firms have an irrational aversion to borrowing, the former unwilling to borrow to buy low-beta stocks and the latter reluctant to take on debt to supply borrowing implicitly to investors by creating leveraged equity. Since they are generally subject to fewer borrowing restrictions than individuals and, in addition, benefit from debt tax shields, Fischer thought corporations should use the Sharpe–Lintner model to value projects. In the wheat seed model, this practice corresponds to forming corporations that borrow to buy low-beta wheat seed for planting.

It is worth noting, in passing, that Fischer thought actual corporate dividend policy difficult to interpret in rational terms for similar reasons. In

partial-equilibrium versions of the CAPM, investors require higher expected returns to hold securities with high dividend yields as compensation for the corresponding increase in tax liabilities, an argument that has long seemed tenuous to many researchers. For example, competition among investors with low or zero marginal tax rates probably should eliminate any dividend yield premiums, and pension, endowment, and insurance company portfolios are arguably marginal investors in most listed securities in the United States.

More importantly, Black and Scholes (1974b) argued that there should be no dividend yield effects on expected returns that corporations can exploit to lower their cost of capital. Firms can supply a range of dividend policies to satisfy any investor dividend clienteles, and competition among firms should ensure that dividend yields should have no effect on expected returns. To be sure, they also showed that any dividend yield effects on expected returns that might arise from the impact of taxes on investor portfolios were empirically ambiguous at best, a finding consistent with such supply effects, and the empirical evidence remains unclear at present. What cannot be explained is the presence of substantial taxable dividend payments in actual capital markets. Fischer ultimately concluded that taxable investors simply like dividends and cannot see the associated decline in their after-tax returns in the high-volatility noise of long-lived asset returns.

4. INTERNATIONAL ASSET PRICING: BLACK (1974, 1978, 1990A), BLACK AND LITTERMAN (1991, 1992)

Although it is out of sequence with respect to its development temporally, I think it best to consider some of Fischer's contributions to international asset pricing at this point. His thoughts on the subject typically involved simple CAPM reasoning. His inferences were generally based on precisely which substantive cross-country differences in the circumstances of investors should impinge on their optimal portfolios and on how these portfolios were priced in equilibrium when the world market portfolio must be held and other assets were in zero net supply. Since other properties of financial market equilibrium are needed in the discussion of derivative asset pricing that follows, it makes sense to review this work first.

Market imperfections are central to much of international economics, and therein lies the source of a gap, small at times and gaping at others, between international finance as practiced by financial economists and by international economists more broadly defined. Tariff and subsidy schemes, capital controls, and constraints on foreign ownership constitute one class of impediments to free trade. A more subtle barrier involves information asymmetries—domestic investors may possess knowledge or insight into domestic asset valuation not available to foreigners. National boundaries can engender substantial cross-border differences in the behavior of investors in such a world.

Black (1974) investigated equilibrium asset prices under fixed exchange rates in the presence of such impediments and, to the best of my knowledge, provided the first equilibrium analysis of such issues under uncertainty. He retained the assumptions of the Sharpe–Lintner CAPM and added one other: taxable investors were subject to a given tax rate when investing in other countries where the tax may be thought of as a "representative trade barrier." The short sale assumption, in particular, had a large impact on the resulting optimal portfolios of investors.

The issues involved can be illustrated in a variant of the wheat seed economy. Imagine that we enrich the wheat economy sketched above with domestic and foreign investors. Since exchange rates are fixed, we might as well normalize them to unity. As in the original model, normalize the price of consumption today and end-of-period wealth to 1 as well. To keep matters simple, suppose that domestic and foreign investors are identical in both endowments and risk aversion. Permitting greater heterogeneity in these parameters changes both riskless rates and makes it desirable for all investors to have a position in the world zero beta portfolio but results in an otherwise qualitatively similar equilibrium.

Both domestic and foreign investors are endowed with different types of wheat, and wheat seed can be freely shifted from consumption to production. However, it cannot be shifted for production across countries, a condition necessary to ensure positive tax revenues. Without it, investors could avoid taxation simply by trading seed until domestic production and consumption came only from domestic claims.

Investors differ in the taxes they pay on claims to wheat seed planted in the two countries. Neither investor pays taxes on claims to wheat seed planted in their own countries. Foreign investors pay a given tax rate on claims to seed planted domestically and, for simplicity, domestic investors pay the same tax rate on claims to seed planted in the foreign country. Short and long positions are treated symmetrically, implying that short sales incur negative taxes (i.e., tax subsidies). The uses of tax revenues are ignored in the model; they can be thought of as being redistributed to investors in a lump sum fashion so as not to disturb the equilibrium.

The tax system gives domestic (foreign) investors a comparative advantage at owning domestic (foreign) assets and foreign (domestic) investors a comparative advantage in short sales of domestic (foreign) assets. Hence, all investors naturally want to hold more assets in their country of origin than would otherwise be optimal and will fund this position with short sales of securities in the other country. The world market portfolio must be held in equilibrium, so the short positions of foreign (domestic) investors in domestic (foreign) securities must be balanced by long positions in excess of the market portfolio weights by domestic (foreign) investors. Note that investors will not hold only assets in their own country; they must balance the tax benefits of domestic investment against the diversification benefits of foreign investment.

This symmetry implies that all investors find it optimal to form their portfolios from a menu of three mutual funds. The first is the world market portfolio. The two others are minimum-variance portfolios with the given tax rate from the perspective of domestic and foreign investors, respectively. The weights of these portfolios sum to that of the world minimum-variance portfolio, so that both the country of origin and the covariance structure of returns determine the weight of each asset in each portfolio. If the correlations among returns are sufficiently weak, these minimum-variance portfolios will concentrate in foreign and domestic assets, respectively. In any event, it is a simple matter to find the equilibrium savings rates and the allocation of wheat seed types to consumption or investment given this collection of asset demands.

Note the simple economics of this equilibrium. The global CAPM describes the equilibrium in the absence of tax distortions, the world market portfolio being the one that all investors would hold in the absence of impediments to free trade. The tax distortions cause investors to engage in zero net supply trading to get the best risk/return/tax tradeoff. These distorted optimal portfolio choices cause wheat seed claim prices to change from those in the zero tax equilibrium. Note also that the equilibrium could just as easily describe that within a country or region with a common currency and different tax jurisdictions. Explicating the difference between closed and open economy financial market equilibria requires a precise notion of what barriers are encountered at national borders.

The capital market equilibrium in this model belies much of the intuition regarding the impact of impediments to trade in international capital markets. Instead of leading to market segmentation, the absence of short sales restrictions in conjunction with the tax barrier leads investors in one country to swap their deviations from the world market portfolio for those of the other country. That is, these investors swap portfolios to reduce their tax liabilities, a transaction reminiscent of swaps created for this purpose in the early 1980s. It may well be that sizable short sales are not feasible even for large institutional investors but the tax distortions are likely to be small at modest levels of taxation, probably making the swap feasible. It is perhaps noteworthy that actual international withholding taxes have historically been on the order of 10–15%.

The presence of these short and long positions in the portfolios of domestic and foreign investors provides another explanation for the tendency of low-beta stocks to plot above the Security Market Line and for high-beta stocks to lie below it. In this version of the model with cross-country withholding tax equalization, Security Market Line deviations are the product of one-half of the tax rate and one minus the beta of each asset with respect to the world market portfolio, thus generating the same regularity. Hence, the model can be tested during fixed exchange rate regimes by seeing if the world zero beta rate exceeds the world risk-free rate, ignoring the problem of measurement of the world market portfolio. The difference

would be zero if such taxation is ineffective and the world Sharpe–Lintner CAPM is true.

That said, Fischer did not really believe this explanation for the apparently flat Security Market Line. As he noted in Black (1974, p. 350), "The trouble with taking such restrictions at face value is that they may not be effective. There are many ways to get around such restrictions, and there are often types of investments not subject to the restrictions." I, along with many others, have listened to both Black and Scholes as they spun portfolio strategies designed to eliminate or substantially reduce taxation within arbitrary tax codes. For example, tax revenues fall to zero if there is frictionless free trade in wheat seed for planting in this model. Similarly, the tax-reducing properties of short sales in this model should serve as a cautionary reminder of the many ways in which financial assets can be used to avoid or reduce apparent barriers. That is, the nature of the supply side is critical for understanding the general equilibrium effects of taxation.

In fact, there is little difference between closed and open economy finance when the only real barrier at a national border is an exchange rate. This viewpoint colors the interpretation of a variety of empirical regularities, and Black (1978) enumerated several of them. Fischer thought that accounting flows such as the current and capital account balances obscure the underlying economics of international investment. Ceteris paribus, increased domestic investment by foreigners implies increased foreign investment by domestic investors. Financial asset flows should be unrelated to rates of return since asset prices are forward-looking, and risk-adjusted expected returns should be equalized across countries. In contrast, there can be substantial cross-country differences in the productivity of physical capital, and physical capital should flow toward high-productivity countries without earning abnormal returns. Accordingly, there is nothing surprising about a negative balance of trade associated with a positive capital account balance arising from the returns of prior investments.

It is instructive to compare these observations on financial and physical capital flows across countries to those within countries and regions. International economists think it noteworthy when countries have large current or capital account balances. Yet we are not surprised when the citizens of exclusive, wealthy communities buy financial assets in the rest of the world, thereby having imports grossly exceed exports. We seldom think their purchases and sales of securities presage rising or falling "foreign" expected returns. We are also not puzzled to find physical capital being installed outside the wealthy community. Similarly, we do not think that coal-mining communities with large current account surpluses and capital account deficits are somehow doing well. That is why "The balance of payments is an even more mysterious concept. We do not even know how to define it meaningfully, let alone measure it. For every proposed definition it is uncertain whether we would rather have a surplus, a deficit, or neither" (Black 1978, p. 25).

Fischer also thought it self-evident that an international version of the CAPM must hold in the absence of effective market imperfections. The gains from international diversification are huge, and investors should care about the properties of returns in their currencies, not about their countries of origin. Of course, foreign investors might fear the possible informational advantages of domestic investors. However, competition among informed domestic investors should produce (approximate) informational efficiency in sufficiently well-developed financial markets, permitting foreign investors to (mostly) free ride on rather than fear the putative informational advantages of domestic investors. Consequently, Fischer, along with a few previous authors and many subsequent ones, thought the big puzzle is "Why don't investors hold more assets in other countries?" (Black 1978, p. 29).

What features does an international CAPM have in the presence of flexible exchange rates? Once again, a variant of the wheat seed economy helps illuminate the relevant issues. However, instead of taxes or other market imperfections, suppose the prices of consumption today and wealth tomorrow are normalized to unity for the domestic investor while those of the foreign investor are the exchange rate today and tomorrow, respectively. The latter statement is more than a definition and represents the presumption of purchasing power parity in the goods market.

What does the resulting equilibrium look like? This question really cannot be answered without explaining the role of exchange rates in this world. If they simply reflect different nominal prices of consumption and wealth, real returns will be independent of exchange rate movements: any price inflation (deflation) in the foreign country will be offset by exchange rate depreciation (appreciation) as long as purchasing power parity holds. Of course, purchasing power parity is violated empirically, but not to the extent necessary to create nontrivial hedging demands. That is, the variability of rates of return on both long-lived assets and exchange rates so exceeds inflation variability as to render long-lived foreign asset returns in domestic currency terms roughly equal to real returns. Fischer seldom took seriously the notion that money illusion had real consequences.

In the absence of important effects attributable to exchange rates as the relative prices of units of account, the alternative is to assume that real exchange rates fluctuate. Real exchange rate changes must come from cross-country variation in relative prices. In principle, relative price variation can arise from both supply or demand sources, but Fischer (and most of the literature) preferred to think in terms of consumption differences across countries. That is, he defined "a group of investors with the same price index as a 'country' . . . a 'currency' for each country as having a known end-of-period value using the corresponding price index. An 'exchange rate' will be the relative price of two currencies" (Black 1990a, p. 900).

Hence, the model must have two or more consumption goods to permit nontrivial cross-country differences in consumption bundles and price indices. Accordingly, suppose we enrich the wheat economy sketched

above by endowing investors with different types of both wheat and cows. Cows can be consumed as beef today or as beef (and, I guess, veal, as long as the relative price of veal and beef is fixed) tomorrow. Cow production is also subject to (truncated) normal random shocks with the same sort of stochastic structure specified for wheat yields above, although it may be best to think of them as taste shocks for different types of beef. Tastes for wheat and beef differ between countries and, hence, domestic and foreign investors have the requisite differing price indices.

Suppose the domestic price index now and in an instant are normalized to unity while the corresponding numbers for the foreign investor are the exchange rate now and in the next instant. Uncertainty regarding future exchange rates makes riskless domestic bonds risky from the perspective of foreign investors and riskless foreign bonds risky from the perspective of domestic investors. However, the uncertainty involves the percentage change in exchange rates from the perspective of domestic investors and its inverse from the perspective of foreign investors. In addition, the inverse of the percentage exchange rate change has a multiplicative effect on risky domestic assets from the perspective of foreign investors, and the level has a multiplicative effect on risky foreign assets from the perspective of domestic investors. These nonlinearities complicate the analysis of international capital market equilibrium.

However, these instantaneous exchange rate effects are linear when exchange rate changes follow a diffusion process, as is the case with Brownian motion information flows. The inverse of the instantaneous percentage change in the exchange rate is the instantaneous volatility of the percentage exchange rate change minus the instantaneous percentage exchange rate change. Hence, the instantaneous return differential between domestic and foreign bonds in domestic currency terms is the nominal interest rate differential minus the percentage change in the foreign exchange rate. The corresponding differential in foreign currency terms is the same minus the instant variance of the percentage change in the exchange rate. Similarly, domestic asset returns in foreign currency terms are the returns in domestic currency minus the percentage exchange rate appreciation plus a constant, the instantaneous percentage exchange rate change variance minus the instantaneous covariance between domestic asset returns and the percentage change in the exchange rate. The appropriate converse is true for foreign asset returns in domestic currency terms.

This implicit linearity simplifies the resulting international capital market equilibrium considerably. The wealth constraints of domestic and foreign investors are identical up to additive constants and weighted averages of the percentage change in the exchange rate. Accordingly, all investors will find it optimal to hold shares in the world wheat and cattle market portfolio combined with international borrowing and lending. International bonds partially hedge the world market portfolio as long as their returns are correlated, and, hence, investors will find it optimal to issue positive quantities of both domestic and foreign bonds. The intrinsic symmetry of

exchange rate risk—that appreciation to a domestic investor is depreciation to a foreign investor plus its volatility—implies (after some nontrivial manipulations) that domestic and foreign investors will hold identical domestic and foreign bond portfolios coupled with domestic lending. This is well-known from Stulz (1981), a paper taken from the only dissertation research supervised by Fischer, and other papers cited in the Adler and Dumas (1983) review article.

What is less well-known is that the assumption that both foreign and domestic riskless bonds are in zero net supply sharply constrains optimal hedging in equilibrium. Consider the case in which all investors have the same degree of risk aversion so that they hold the same risky asset portfolio in proportion to their share in world wealth. Since each investor has the same riskless bond position to hedge risky asset returns from the optimal portfolio result, there is a fixed aggregate demand for the riskless bonds of all countries to hedge risky asset returns. Gross lending by domestic investors must be offset by the borrowing of both foreign and domestic investors in this bond portfolio—that is, gross domestic lending exceeds net domestic lending. Coupled with the zero net supply constraint for domestic and foreign bonds, this condition implies that the gross lending of all investors is the same fraction of their corresponding share of world wealth. The same basic result obtains when investors have different risk aversions since the gross lending in each bond in excess of the aggregate demand for hedging purposes must sum to zero.

The resulting hedge ratio takes a remarkably simple form not anticipated in the voluminous literature on exchange rate hedging. The fraction of foreign investments hedged is the ratio of two quantities: the mean excess return of the world market portfolio less its variance and the mean excess return of the world market portfolio minus one half of the average exchange rate variance. Moreover, one minus the fraction hedged gives the average risk tolerance of investors. Note the complete absence of the usual inputs for hedging calculations: exchange rate means, variances, and correlations with other asset returns. These variables do appear in the marginal conditions of optimizing investors, but the mean and covariance terms net out in equilibrium because of both the zero net supply constraint for all bonds and the symmetry of all investors. In equilibrium, exchange rates will adjust until real exchange rate risk takes this form.

Exchange rate hedging is a free lunch, and thus all investors prefer flexible exchange rates in this model. When exchange rates follow diffusion processes, the difference between the instantaneous return differential between domestic and foreign bonds in domestic and foreign currency terms is not zero but rather is the instantaneous variance of the percentage change in the exchange rate. An investor in one country whose portfolio risk is reduced by exchange rate hedging is counterbalanced by one in another country who experiences an increase in portfolio expected return from this source. The Sharpe ratios of both portfolios rise because investors in the two countries implicitly divide up the Jensen's inequality term,

which is known as Siegel's paradox in international finance. This gain from trade does not arise with fixed exchange rates such as those implicitly prevailing within any region with a common unit of account.

The model also has an interesting empirical implication for risk premiums. One of the problems plaguing the empirical asset pricing literature is the extraordinary imprecision with which the mean excess returns of long-lived assets are estimated. In the Sharpe–Lintner CAPM, the risk premium of the market portfolio can be inferred from the risk aversion of investors and the fraction of their wealth in the riskless asset. In this model, investors' risk aversion is implicitly embedded in the fraction of the portfolio that is hedged, a potential observable.

The practical implications of this observation are more important still. Quantitative portfolio management has long suffered from the noise in expected return estimation. For example, sample mean–variance-efficient portfolios based on simulated returns typically have extreme positions in some assets. There is a general tendency to take large long positions in assets with large sample, but not population, mean returns and to place large negative weights on those with negative sampling error in average returns. Practitioners of quantitative portfolio management have several methods for dealing with this problem that generally fall into one of three categories: the imposition of artificial constraints to produce reasonably well-diversified portfolios, the introduction of benchmarks from which expected tracking error is comparatively precisely estimated, or covariance-based estimation of risk premiums based on false models. I must confess a preference for the third approach.

Goldman Sachs Asset Management has produced a family of models for global asset allocation based on Black and Litterman (1991,1992) that fit into the third category. They use this model to produce estimates of the expected excess returns of uninformed investors. This device permits the model to support active management based on investor beliefs about expected asset returns and exchange rates expressed as deviations from the postulated beliefs implicit in this global CAPM. The result is a set of asset allocation strategies with well-behaved portfolio weights based on economic theory.

From index funds to the Black–Scholes model to the Black–Litterman model, Fischer systematically believed in theory as the basis for practical financial products. Now the international CAPM is patently false since actual investor portfolios differ substantially from the optimal ones, if only because of the extraordinary home country bias in actual portfolios. Nevertheless, the question is not whether the theory is false but rather whether the errors in the model's expected returns are larger or more misleading (from the perspective of portfolio formation) than those produced by other methods. My guess is that they are not in this case.

Three aspects of Fischer's work on universal hedging strike me as noteworthy. First, the (international) CAPM is at its center, with market clearing in asset markets describing the real equilibrium. Second, he teased yet another unexpected and noteworthy implication from the zero net supply

nature of financial assets in this familiar setting that had escaped the notice of many scholars interested in exchange rate hedging. Finally, the resulting model was both abstract theory and the basis for practical asset allocation tools.

5. THE BLACK–SCHOLES MODEL: BLACK (1989A), BLACK AND SCHOLES (1973)

In the wheat economy, the only assets that trade are claims to wheat seed itself, not securities that permit hedging against different realizations from the multivariate normal yields. If the parameters of this economy were sufficiently similar to those in the real world, wheat seed returns would exhibit high volatility. Investors with sufficiently diverse endowments and risk aversions would perceive substantial gains from trade from opening markets in derivative assets such as options on different wheat seed types. Moral hazard poses no barrier to writing such contracts if wheat seed claim prices and crop output are observable. In the absence of costs of writing, enforcing, and trading such contracts, the profit motive for introducing options and other derivative assets would lead investors to complete the market by issuing such claims.

There are two natural questions engendered by the addition of zero net supply financial assets such as options. First, does their introduction destroy the CAPM and pricing by the Security Market Line? After all, wheat seed returns may be normally distributed, but the distributions of payoffs on derivative assets such as options are decidedly nonnormal. Second, what prices would derivative securities command in equilibrium? While this might seem like a strange way to pose these questions given the role of the riskless hedge in the Black–Scholes model, it strikes me as the correct formulation because Black and Scholes discovered the model by finding the right way to apply the CAPM to option valuation when the underlying assets follow diffusion processes.

In fact, the introduction of sufficiently many financial contracts to complete the market changes the resulting equilibrium but in ways that make CAPM pricing *more* plausible. What is lost is the most troublesome prediction of the CAPM: that all investors hold the same portfolio of risky assets and differ only in their position in either riskless bonds or the zero beta portfolio. What remains is the pricing of types of wheat seed by the Security Market Line.

It is worth emphasizing why the CAPM need not hold despite the fact that risky wheat seed returns remain normally distributed. The CAPM obtained in the original wheat economy because all investors found it optimal to hold mean–variance-efficient portfolios precisely for this reason. In the present economy, investors also have access to derivative assets, many of which have skewed returns, and, hence, investors will generally no longer judge portfolios solely on the basis of their means and variances. Optimal portfolios of wheat seed claims and derivatives written on them

will typically vary across investors due to their differences in risk aversion and endowments. That is, the asset menu no longer supports portfolio separation.

CAPM pricing of wheat seed shares remains valid following the introduction of derivative assets for two reasons: the completeness of the wheat seed market and the dependence of investor utility solely on initial consumption and end-of-period wealth. The former guarantees that the asset demands of a representative investor describe financial market equilibrium, although not aggregate market excess demand curves. The latter implies that the only random quantity that concerns this investor is terminal wealth. Now the market portfolio has normally distributed returns, and the representative investor holds only this portfolio since the derivative assets are in zero net supply. The result is CAPM pricing for risky wheat seed with risk premiums based on the risk aversion of this investor. Note that the risk premiums and betas in the wheat economy with and without derivative assets will generally differ.

However, CAPM pricing of derivatives such as options need not obtain. Recall that the stochastic information flows introduced earlier came in two flavors: frequent smooth information arrivals well-approximated by diffusion processes and large but normally distributed shocks that occur at predictable times. Derivative assets that are written and mature between these periodic jumps have values that depend on underlying assets with returns that follow diffusion processes, while those outstanding when such discrete changes occur do not.

One can solve the valuation problem for options and other derivative assets in the former setting when the parameters of the underlying diffusion are constant using the riskless hedge discovered by Merton (see Darrell Duffie's discussion in section 5 of chapter 12) but Black and Scholes (1973) took the CAPM route. As Fischer reported in Black (1989a), there were two basic stages in Black and Scholes' development of the option pricing model. The first was Fischer's finding that option values did not depend on the expected return of any asset when the underlying asset price followed a diffusion process and the CAPM held continuously. The second was their realization that the absence of expected returns from the determinants of option values meant that they could apply what has come to be known as risk-neutral valuation. Since the expected return of the underlying asset did not appear in the option valuation formula, they could simply set it to the risk-free rate without affecting the valuation relation. Since the underlying asset would have a zero beta when its expected return was the riskless rate, the beta of the option would be zero in this case, making the appropriate discount rate for the option the risk-free rate as well. Discount the expected terminal payoff of a truncated lognormal random payoff at the risk-free rate and you have the Black–Scholes formula.

Fischer often said that he believed the CAPM and not the hedging derivation of the Black–Scholes model, and it is worth contemplating some of the implications of this view. Suppose we modify the information arrival

process in two dimensions: (1) discrete chunks of information can arrive randomly as well as periodically, a modest generalization of the earlier model, and (2) the price dynamics of the wheat market portfolio are a diffusion process with drift and diffusion coefficients that depend only on its market value irrespective of the jumps that arise in wheat seed claim prices. The latter represents a restriction on both information arrival and the asset menu: all discontinuities in this process reflect seed-type-specific information, and the number of assets is sufficiently large to make perfect diversification of jumps across wheat seed claims possible.

The CAPM pricing of derivative assets obtains in this market even if insufficiently many contracts have been issued to complete the market. This result does not rely on the assumption that payoffs on derivative assets are perfectly correlated with—that is, lie in the span of—those of existing risky wheat seed claims. Rather, it arises because all investors view jump risk as idiosyncratic in these circumstances, diversifiable risk that should not command a risk premium by construction. This supply-side assumption about the idiosyncratic nature of large information arrivals yields a market portfolio with returns that follow a diffusion process. The resulting linearity of the marginal conditions of the representative investor in the market portfolio implies that the expected excess returns of derivative assets are linear in their (generally time-varying) betas.

This variant of the wheat economy combines elements of the Sharpe–Lintner CAPM and the APT. The CAPM says that zero beta assets should earn the risk-free rate. The premise that jumps in the prices of wheat seed claims represent diversifiable, idiosyncratic risk is an APT-style assumption; the idea that they command no risk premiums—that is, that they have zero betas—corresponds to pricing within an equilibrium version of the APT. Note that CAPM pricing holds irrespective of the stochastic process of jumps in this model. For example, Merton's (1976) option valuation model arises when jumps follow a lognormal distribution with parameters that are fixed over the life of the contract.

6. STATE DEPENDENCE, BUSINESS CYCLES, AND ASSET PRICING: BLACK (1986, 1987, 1990B, 1995), BLACK AND SCHOLES (1973)

The wheat economy CAPM has gone through several changes in these pages but remains a quite restrictive model, an observation that might appear to gainsay my suggestion that it guided much of Fischer's subsequent thinking. Two sorts of assumptions were of particular concern to him: (1) the omission of any sort of state dependence in household preferences, and (2) the simple stochastic constant returns to scale production technology. The question at hand is whether the CAPM intuition survives enriching the model in these dimensions.

On the demand side, Fischer came to believe that many sorts of state dependencies in preferences were economically important. In particular, he thought households were quite willing to substitute consumption over

time in the short run but engage in consumption smoothing over intermediate and long horizons, resulting in a consumption function depending on both current and lagged wealth. He also thought that any reasonable wealth construct included human capital, making fluctuations in the value of human capital, the volatility of which he thought comparable to that of physical capital, another potential source of state dependence in the indirect utility of wealth. Taste shocks fit into this category as well.

On the supply side, he came to believe that economically important frictions in the allocation of capital were the dominant cause of business cycles. Capital flows freely in this model since seed not allocated to consumption can be planted with no barriers. In his later work, Fischer emphasized the long gestation lags associated with roundabout production. In particular, both physical and human capital investments are often made on the basis of expectations about supply and demand conditions that will prevail far in the future, well before their actual marginal value products are determined. Both individual assets and the market as a whole can experience prolonged periods of unusually high or low expected returns if capital in all its forms can only be moved between sectors with nontrivial adjustment costs.

Suppose, for example, there are two inputs into the production process, land and wheat seed, and that the productivity of each type of wheat seed on different plots of land is subject to imperfectly correlated random shocks. In addition, suppose a given seed type can only be planted in a given wheat field if special preparations are made after the prior harvest—that is, six months before the next planting—and that wheat seed productivity is subject to permanent and transitory shocks. Output will be high next year if there is a good match between wheat types and field preparation after the productivity shocks are realized, and output will be low if there is a poor match. These random fluctuations in the quality of the match between needs and resources are made persistent by costs associated with reallocating resources such as costly capital or, in this model, wheat seed mobility. Mix in longer gestation lags and shocks on the demand side when wheat types represent differentiated products to consumers, not perfect substitutes as in this model, and you have Fischer Black's theory of business cycles.

Fischer meant for this sort of model to be quite different from the simple wheat economy. The pithy reference to longer gestation lags conceals the belief that "production takes too many years and too many distinct inputs to be summarized by an aggregate function that depends only on the current values of a few inputs" (Black 1995, p. 31). Roundabout production includes roundabout production of the production process itself because "Only a little of what we do each year is to produce final goods and services. Most of what we do is to turn this year's capital into next year's capital" (Black 1995, p. 24). He thought "of the world as having *billions* of relevant sectors" (Black 1995, p. 51) across which there are nontrivial costs of reallocating resources, thus generating persistent business cycles.

Then there is the asset menu. Fischer viewed capital markets and the macroeconomy through the rose-colored glasses of the Arrow–Debreu

model with complete markets in state-contingent claims. As in the previous section, such claims would include derivative asset contracts written on wheat seed claims. However, market completeness might require a menu of contingent claims with payoffs dependent on the consumption indices that enter investor preferences, human capital values, and any taste shocks. This assumption would appear to require considerable suspension of disbelief; unlike wheat seed derivatives, one cannot blithely state that moral hazard, observability, and enforcement are not issues for such claims.

CAPM pricing of wheat seed shares does not obtain in this model because only half of the sufficient conditions are met in this version of the model. Market completeness once again ensures the existence of a representative investor who holds only the conditionally normally distributed wheat market portfolio since derivative assets are in zero net supply. However, this representative investor will generally care about many state variables in addition to end-of-period wealth.

The CAPM might no longer govern pricing, but its close relative, Merton's ICAPM, describes the equilibrium in the economic setting of the previous section. Parenthetically, Fischer preferred the ICAPM to the consumption CAPM because the latter requires time- and state-separable preferences while the ICAPM does not (Black 1995, p. 149). What is essential for the ICAPM is that the intertemporal marginal rate of substitution of the representative investor is linear in the innovations to both market portfolio returns and the state variables. If the innovations to the state variables follow (possibly state-dependent) diffusion processes, portfolio separation obtains and expected returns are linear in the usual market beta plus the covariances between asset returns and state variable innovations. In addition to the market portfolio and the riskless asset or zero beta portfolio, all investors will select from a menu of hedge portfolios that have returns that are maximally correlated with state variable innovations. Jumps that do not result in dependence between asset prices and future marginal utilities can be accommodated within the ICAPM as well.

At this level of generality, this version of the ICAPM seems to be of little empirical relevance since it would appear to take an implausibly large number of both contingent claims to complete the market and state variables to describe the random shocks to the indirect utility of the representative investor. Fischer thought not because:

> Introducing tastes does not make the theory into a black box, though, since tastes are largely observable. We can estimate them with surveys, introspection, and detailed marketing and engineering data . . . Tastes and technology . . . are observable to roughly the same degree. Estimating supply curves is as difficult as estimating demand curves. It's easy to see what and how much a person or a household or a community will buy at given prices as it is to see how much a plant or an office or a machine will make at given prices . . . In principle, we can use marketing and engineering data to identify supply and demand. In practice, we will find it very costly to gather this information and interpret it. (Black 1995, pp. 84, 126)

Perhaps he also thought that we empirical types are unwilling to do this much work!

Like nature, capitalism abhors a vacuum, and an unfilled or underserved product niche is that worst of all vacuums, a missed profit opportunity. By the same token, a flood of similar products can overfill a niche, generating economic losses for the firms involved. Firms make the physical capital investment decisions, and firms and workers make the human capital decisions that determine how well-positioned firms are to satisfy the future product demand. Future demand will depend on taste shocks, the time paths of both physical and human wealth, and other sources of state dependence in preferences. Hence, the market value of the firm measures the extent to which the past planning of firms and workers matched subsequent consumer product demand.

On this view, markets may not be as incomplete as they might appear at first blush. The range of possible values of the common stocks of business firms might come close to spanning the uncertainties confronting firms and workers. If such spanning obtains, options written on equities can help complete the market. Several tiers of debt securities with differing priority in bankruptcy can help serve this function as well. This argument might fail for nontraded assets such as human capital or unlisted firms, but Fischer did not think so because:

> Human and physical capital take relatively stable shares . . . Thus we can use the market prices for firms that have traded securities to suggest market values for human capital. We assume that the return on total human capital is equal to the return on total physical capital. We do the same for the physical capital of firms that have no traded securities. (Black 1995, p. 34)

Home mortgages, again possibly with several tiers, and credit card debt can be viewed as contingent claims that help span individual specific risks.

To Fischer, it simply was not plausible that an economically important state of the world would not be reflected in a market value somewhere. Investors can write financial contracts that ensure against any outcome for which the configuration of returns is different. As long as each economically important state is distinctly value-relevant for some security, the usual gains from trade argument implies that they have the incentive to do so. If, in addition, information flows are sufficiently smooth, such contracts would only need to span small uncertainties since diffusion processes have continuous sample paths. The combination of value-relevant states and smooth information flows makes for more plausibly complete markets.

If anything, obtaining the ICAPM with a small number of state variables is easier still. Many of the risks confronting workers and firms are in zero net supply. For example, an industry portfolio return is unaffected by which firms prove to be winners and losers when there are several firms with competing technologies in a market with given demand. Fischer did not think that all sectoral demand and supply shifts were in zero net supply since he viewed booms and busts as nothing more than the simultaneous

occurrence of predominantly good or bad matches across numerous sectors. Nevertheless, the relatively small number of factors needed to account for covariation among asset returns suggests that a small number of state variables are empirically relevant if the ICAPM is true.

Fischer thought that the identification of the relevant state variables is another matter entirely. He remained suspicious of efforts to identify factors empirically by either factor analysis as in the APT or through cross-sectional regression of returns on security characteristics that are putatively correlated with risk exposures. He viewed direct measurement by studying the details of tastes and technology as both prohibitively expensive and potentially subject to rapid depreciation in the rich dynamics predicted by his general equilibrium world.

If correct, this seemingly small measurement problem has nearly fatal consequences for our ability to discover the economics of financial market equilibrium. If financial markets are in equilibrium, they do not permit arbitrage opportunities. If market prices are arbitrage-free, there is a set of strictly positive state prices implicit in market prices. If the market is complete, these state prices are unique and there is a representative investor whose asset demands characterize the equilibrium. This representative investor sets expected marginal utilities proportional to state prices state by state.

Since the representative investor's preferences are state-dependent, it is trivial to construct a utility function that supports this equilibrium. Take arbitrary concave preferences, probability beliefs across states, and a rate of time preference. Calculate the implied expected intertemporal marginal rate of substitution at the aggregate consumption of goods and services in each state of nature. Compute the ratio of the state price to the expected intertemporal marginal rate of substitution state by state. Set representative investor preferences next period equal to the product of this ratio and the original (arbitrary) concave utility. The asset demands of this investor will support equilibrium prices and quantities. Other constructions are possible; for example, state-dependent risk-neutral preferences—that is, those that are linear in wealth—will suffice as well.

In the absence of a priori identification of the relevant state variables and how they impinge on investor preferences, the hypothesis that financial markets are in equilibrium is entirely vacuous as long as market prices are arbitrage-free. Since beliefs are arbitrary, a rational expectations equilibrium is both easily constructed and similarly vacuous. That is, there are an infinite number of observationally equivalent equilibria. One cannot speak of a complete markets model of this form as making testable predictions that can be confronted with data to see if it is true or false.

This sort of reasoning is implicit in Fischer's work, particularly Black (1995). On this view, any positive theory of risk premiums is really a model of specific representative investor preferences. Any rejection can be interpreted as simply implying that the preference specification is wrong, not that financial markets are out of equilibrium or are incomplete. If one is

willing, as was Fischer, to treat market completeness as the maintained hypothesis, empirical efforts to model risk premiums should be viewed as explorations of functional forms in applied demand analysis.

Instead, he favored a different mixture of applied theory and empirical work:

> [I]t's better to "estimate" a model than to test it. I take "calibration" to be a form of estimation, so I'm sympathetic with it, so long as we don't take seriously the structure of a model we calibrate. Best of all, though, is to "explore" a model. This means creating many specific examples of a general model, where each one explains a single stylized fact or perhaps a few features of the world. It means using some of these examples to elucidate microeconomic evidence . . . It means changing examples quickly when they don't fit the facts. It means avoiding formal testing, or even estimation, of heavily restricted models. (Black 1995, pp. 4, 5)

Black (1988, 1990b) presented two related explorations of general equilibrium. These examples—explicit in Black (1990b) and implicit in Black (1988)—were constructed to show that a standard stochastic growth model with a representative investor, one without directly state-dependent preferences, could reproduce three empirical regularities: consumption smoothing, the equity premium puzzle, and wealth volatility that is both high and declining in the level of wealth. High wealth volatility implies that it is hard to measure both the average level and time variation of expected returns, a feature Fischer used in his account of the October 1987 stock market crash.

What characteristics must such a model possess? Representative investor preferences are those routinely used in applied stochastic growth theory; they are additively separable over time with constant relative risk aversion utility of (instantaneous) consumption. Consumption smoothing arises when consumption changes less than one-for-one with wealth. The second and third regularities obtain when wealth dynamics are those used in the CAPM pricing of derivatives in the preceding section: a diffusion process with drift and diffusion coefficients that depend only on wealth. The parameters of this stochastic process can be chosen to match its observed moments. Equilibrium requires determination of two prices—the risk premium for wealth and the risk-free rate—that make the representative investor happy to hold risky wealth and to have no position in (zero net supply) riskless bonds.

Two natural relative risk aversion measures govern the amount of consumption smoothing in equilibrium. The first is direct risk aversion: the coefficient of relative risk aversion of instantaneous utility or, equivalently, one minus the consumption elasticity of instantaneous utility. The second is indirect risk aversion: the product of wealth and the ratio of the second derivative of discounted expected utility with respect to wealth to the corresponding first derivative or, equivalently, one minus the wealth elasticity of discounted expected utility. In equilibrium, the ratio of direct to indirect risk aversion is equal to both the wealth elasticity of consumption

and the ratio of the standard deviation of consumption to that of wealth. Hence, consumption smoothing arises when this ratio is less than one.

These risk aversion measures determine the equilibrium properties of expected returns. The market price of risk—the ratio of the instantaneous expected excess return to the instantaneous variance of wealth—is given by indirect risk aversion. A more delicate calculation reveals that the wealth elasticity of the instantaneous expected excess market portfolio return is negative when direct risk aversion exceeds one—that is, more risk-averse than log utility—and indirect risk aversion is below unity. The same result shows that the instantaneous variance of wealth is declining in wealth. Thus, the model qualitatively reproduces the regularities outlined above.

It also yields roughly the right orders of magnitude for reasonable parameter values. If the Sharpe ratio of stock market wealth is roughly the same as that of overall market wealth, the average risk premium is equal to the model value when indirect risk aversion is 5.6. Using Fischer's estimate of one-third for the ratio of the standard deviation of consumption to that of wealth, the ratio of direct to indirect risk aversion is roughly three, implying a direct coefficient of relative risk aversion of 16.7. Both risk aversion estimates seem a bit high, but they are both very imprecisely estimated and of the right order of magnitude, unlike those found in much of the equity premium puzzle literature. A modest amount of the right kind of state dependence in instantaneous utility or the introduction of an appropriate supply side would easily reduce these numbers.

Moreover, the model makes for a simple distorted beliefs story about the October 1987 stock market crash. Expectations regarding volatility should probably be treated as rational since volatility is estimated quite precisely. However, nothing in the model required that investors possess rational expectations about either expected returns or mean reversion. Accordingly, suppose that beliefs about the indirect risk aversion of the representative investor fluctuate randomly, the role of randomness being to avoid adding hedging demands that would change the equilibrium of the model. Such beliefs are plausible due to the imprecision with which expected returns and mean reversion are measured.

The stock market rose considerably in the first three quarters of 1987 without an associated increase in fundamentals. This was also a period during which many institutional investors adopted portfolio insurance strategies, making their trading much more sensitive to changes in the values of their portfolios. Just before the crash, I remember tactical asset allocators saying that their models were screaming for them to get out of equities in September and early October, a signal that prompted some to add 'earnings momentum' to their models to justify staying in the stock market.

Suppose investors systematically misjudged the extent to which the representative investor changed in this dimension. That is, suppose investors underestimated the increased mean reversion in expected returns that could be expected from wealth fluctuations. On this hypothesis, the

behavior of returns during the week ending on October 16 forced investors to realize that mean reversion from this source in expected returns was much higher than they had expected. Fischer thought that this change in expectations explained the global decline in equity prices.

Note that such an explanation is always consistent with the hypothesis that financial markets were in such an equilibrium that changed in this fashion. As in the earlier construction of the representative investor, compute the ratio of state price to the expected intertemporal marginal rate of substitution state by state under rational expectations. A representative investor with beliefs proportional to the product of rational probabilities and this ratio will support any equilibrium prices and quantities. That is, this distorted beliefs interpretation is vacuous, too, as it is always available when market prices are arbitrage-free.

This story of changing expectations is one in which markets are grossly inefficient since it supposes that a revision of beliefs regarding mean reversion in expected returns caused a 20% market correction. Said correction occurred on a day on which $40 billion of stock and stock index futures changed hands in the U.S., and this observation raises a major concern regarding general equilibrium models of financial market equilibrium. Throughout this chapter, financial markets have been an abstract place in which asset prices are determined. In contrast, the microstructure of actual securities markets is rich, and understanding the information flows within them was a subject of great interest to Fischer. While I will not discuss his market microstructure research so as not to further lengthen an overly long essay, I do want to briefly address the role of trading in general equilibrium models. If the most disconcerting prediction of the Sharpe–Lintner CAPM is that all investors hold the same portfolio of risky assets, the most counterfactual implication of general equilibrium models under rational expectations concerns the volume of trade.

That there might be a tenuous connection between the microstructure of financial markets and general equilibrium models is nowhere more obvious than in the predictions of the latter for the volume of trade. In the initial Arrow–Debreu formulation, all trading takes place at time zero, and no investor feels a need for markets to reopen thereafter. In the sequential markets version of the model, investors need only a modest amount of trading to rebalance their portfolios as the economy evolves stochastically. Hence, complete market models of financial market equilibrium in the absence of information asymmetries predict little or no trade, an observation grossly at variance with the volume of trade in the real world.

How do information flows and investor trading interact in a general equilibrium model of the sort favored by Fischer? More precisely, how should one refine general equilibrium models of financial markets to account for the volume of trade? As in so much of his work, Fischer discussed the relevant economics in words as though the implicit theorems and proofs are obvious. The discussion that follows reflects an effort to make them a bit more explicit.

Consider first the sources of short-run variation in indirect marginal utilities. Fischer's thoughts can be gleaned from his beliefs regarding consumption smoothing: "At intervals of a quarter or longer, aggregate consumption shows a preference for smoothing. At much shorter intervals, though, people seem very willing to shift consumption through time. They exhibit 'local substitution'" (Black 1995, pp. 22–23). If consumption on nearby dates should be treated as nearly perfect substitutes, marginal utilities should be expected to be similar on nearby dates as well so long as state variable innovations don't shift marginal utilities much in the short run. I will take the 'local substitution' assumption to mean that most investors have few or no noninformation-related motives for trade most of the time.

Trading must be in zero net supply. That is, there must be a buyer for every seller and a seller for every buyer. In this setting, prospective traders know two things: they are trading on what they believe to be private information, and potential counterparties are (nearly all) similarly situated. Suppose all investors know that they all have rational expectations and common prior beliefs, but not common information, and that markets are complete. In this case, all investors can infer the relative valuations of potential counterparties from their willingness to trade. Hence, prices must adjust until all investors agree on the price. In equilibrium, prices will reveal the private information of all investors with no trading.

This very general no-trade theorem leaves only three potential reasons for trade. One is that the presumption of solely information-related trading motives is wrong but it is unlikely that the volume of observed trade can be explained solely in life cycle or risk-sharing terms. The common priors assumption is a bit more delicate. Fischer thought that "differences of opinion will not exist" (Black 1986, p. 531) in complete markets, perhaps on the hypothesis that heterogeneous priors are much like private information in equilibrium or because Bayesian updating can yield common posterior beliefs in large samples. This leaves the rational expectations assumption as the likely source of trade in a complete market model.

Enter noise trading:

> Noise trading provides the essential missing ingredient. Noise trading is trading on noise as if it were information. Perhaps they think the noise they are trading on is information. Or perhaps they just like to trade . . . the information traders can never be sure that they are trading on information rather than noise. What if the information they have has already been reflected in prices? Trading on that kind of information would be just like trading on noise. (Black 1986, pp. 531–532)

Prospective traders still know that all investors are trading on what they believe to be private information but can balance the fear that the expectations of potential counterparties are rational against the hope that they are not. Risk aversion and fear of their own fallibility will restrict the willingness of putatively informed investors to trade, permitting prices to persistently deviate from fundamental value.

Hence, we must add descriptors of the beliefs of the representative investor to account for the volume of trade in a full-blown general equilibrium model. In the absence of theoretical restrictions on beliefs, we can always construct a representative investor with rational or irrational beliefs that support any equilibrium. However, as in the case of preferences and technology, Fischer thought of "expectations as observable, at least in principle. We can see expectations in the career choices people make and in the investments firms make. We can even find out about expectations by asking people what they think will happen in the future" (Black 1995, p. 99). If we are willing to condition on a theoretical model, we can use observed investor risk exposures to infer expectations. That is, we can measure expectations from the choices made by investors if we are willing to take seriously the notion that they solve the maximum problems that underlie our models. Fischer always believed in conditioning measurement on theory.

7. CONCLUSION

Throughout this essay, I have tried to sketch a picture of a mind that viewed financial markets from a general equilibrium perch. The view was not static, involving a progression from the simple CAPM to an ICAPM with state dependencies arising from the potentially rich dynamics of a business cycle model driven by sectoral demand and supply dynamics arising from costly resource reallocation. The common theme was the simplicity of equilibria in which all investors or a representative stand-in held the market portfolio and not derivative assets. Fischer was a master at fleshing out the implications of the hypothesis that the latter are in zero net supply.

I have also stressed the sense in which this model does not produce falsifiable, testable hypotheses that can be compared with evidence. It is distressing to discover that it is a *theorem* that economics is vacuous but, viewed in isolation, the hypothesis that asset markets are in equilibrium is just so much empty rhetoric. But we do not view markets in isolation; we implicitly and explicitly bring a mixture of intuition, conjecture, and knowledge to our research. As Fischer put it:

> I think more about a group of stylized facts that summarize the world as I see it. These are my observations, derived from everyday experience, from reading newspapers and magazines, and from studying technical publications . . . I think we can explain all of these puzzles using sample models derived from the full general equilibrium model with general von Neumann–Morgenstern utility. What's important, I think, is to explain the stylized facts in a deeper sense. We want to understand the underlying economics. (Black 1995, p. 7)

On this view, our enterprise involves the exploration of a region of model space centered around the stylized facts and economic intuitions that we find compelling. Students of financial markets differ in precisely what economics and evidence they find compelling. These differences arise largely

because we do not have direct and credible measurements of the determinants of preferences, technology, beliefs, and the extent to which investors make choices and interact in markets according to our models. In the absence of such knowledge, the model constitutes a language for describing and organizing our thinking about financial market equilibrium, not for making predictions about it. This view is reflected in Fischer's notions of the appropriate standards of evidence in economic discourse:

> My approach to evidence is similar to McCloskey's . . . and Summers'. . . . My job, I believe, is to persuade others that my conclusions are sound. I will use an array of devices to do this: theory, stylized facts, time-series data, surveys, appeals to introspection and so on. (Black 1995, p. 83)

In my view, Fischer was a sophist, a label I do not mean to be a slander. In the dialogue *Gorgias*, the sophist Gorgias described rhetoric as the art of persuasion or the manufacturer of conviction. As evidenced by the previous quotation, Fischer thought that economic reasoning was sophistry of just this sort. Irrespective of whether one agreed with his reasoning, he assembled an impressive collection of "devices," as can be seen by perusing Black (1995), especially his thorough and critical reading of so much of the literature.

In this respect, he made it part of the way down the path he set out on, fleshing out the theoretical and empirical details of a general equilibrium model of the sort he thought described the world. However, he only made it part of the way, to some extent because of his untimely passing but also because he did not "think we are ready to create a model with intermediate scope that explains in numerical detail many kinds of evidence at once" (Black 1995, p. 4). He remained an optimist about the eventual success of this endeavor to the end:

> No doubt the reader's glasses differ from mine. But we are all looking at the same world, and the technology for making glasses is constantly improving. Someday it will all be clear. (Black 1995, p. xi)

I can think of no more fitting conclusion.

REFERENCES

Adler, Michael and Bernard Dumas, 1983, "International Portfolio Choice and Corporation Finance: A Synthesis," *Journal of Finance* 38 (June), pp. 925–984.

Black, Fischer, 1972a, "Capital Market Equilibrium with Restricted Borrowing," *Journal of Business* 45, pp. 444–454.

Black, Fischer, 1972b, "Equilibrium in the Creation of Investment Goods under Uncertainty," in Michael C. Jensen (ed.), *Studies in the Theory of Capital Markets* (New York: Praeger).

Black, Fischer, 1974, "International Capital Market Equilibrium with Investment Barriers," *Journal of Financial Economics* 1 (December), pp. 337–352.

Black, Fischer, 1978, "The Ins and Outs of Foreign Investment," *Financial Analysts Journal* 34 (May/June), pp. 25–32.

Black, Fischer, 1986, "Noise," *Journal of Finance* 41 (July), pp. 529–543.
Black, Fischer, 1988, "An Equilibrium Model of the Crash," in Stanley Fischer (ed.), *NBER Macroeconomics Annual 1988* (Cambridge, Mass.: MIT Press).
Black, Fischer, 1989a, "How We Came Up with the Option Formula," *Journal of Portfolio Management* 15 (Winter), pp. 4–8.
Black, Fischer, 1989b, "Universal Hedging: Optimizing Currency Risk and Reward in International Equity Portfolios," *Financial Analysts Journal* 45 (July/August), pp. 16–22.
Black, Fischer, 1990b, "Mean Reversion and Consumption Smoothing," *Review of Financial Studies* 3, pp. 107–114.
Black, Fischer, 1990a, "Equilibrium Exchange Rate Hedging," *Journal of Finance* 45 (July), pp. 899–907.
Black, Fischer, 1993a, "Beta and Return," *Journal of Portfolio Management* 20 (Fall), pp. 8–18.
Black, Fischer, 1993b, "Estimating Expected Return," *Financial Analysts Journal* 49 (September/October), pp. 36–38.
Black, Fischer, 1995, *Exploring General Equilibrium* (Cambridge, Mass.: MIT Press).
Black, Fischer, Jensen, and Myron Scholes, 1972, "The Capital Asset Pricing Model: Some Empirical Tests" in Michael C. Jensen (ed.), *Studies in the Theory of Capital Markets* (New York: Praeger).
Black, Fischer and Robert Litterman, 1991, "Asset Allocation: Combining Investor Views with Market Equilibrium," *Journal of Fixed Income* 1 (September), pp. 7–18.
Black, Fischer and Robert Litterman, 1992, "Global Portfolio Optimization," *Financial Analysts Journal* 48 (September/October), pp. 28–43.
Black, Fischer and Myron S. Scholes, 1973, "The Pricing of Options and Corporate Liabilities," *Journal of Political Economy* 81, pp. 637–654.
Black, Fischer and Myron Scholes, 1974a, "From Theory to a New Financial Product," *Journal of Finance* 29 (May), pp. 399–412.
Black, Fischer and Myron S. Scholes, 1974b, "The Effects of Dividend Yield and Dividend Policy on Common Stock Prices and Returns," *Journal of Financial Economics* 1 (May), pp. 1–22.
Brealey, Richard and Helen Edwards, (eds.), 1991, *A Bibliography of Finance* (Cambridge, Mass.: MIT Press).
Fama, Eugene F., 1970, "Multiperiod Consumption–Investment Decisions," *American Economic Review* 60 (March), pp. 163–174.
Jensen, Michael. C., (ed.), 1972, *Studies in the Theory of Capital Markets* (New York: Praeger).
Lintner, John, 1969, "The Aggregation of Investors' Diverse Judgements and Preferences in Perfectly Competitive Security Markets," *Journal of Financial and Quantitative Analysis* 4 (December), pp. 347–400.
McCloskey, D. N., 1985, *The Rhetoric of Economics* (Madison: University of Wisconsin Press).
McCloskey, D. N., 1990, *If You're So Smart: The Narrative of Economic Expertise* (Chicago: University of Chicago Press).
Merton, Robert C., 1976, "Option Pricing when the Underlying Stock Returns are Discontinuous," *Journal of Financial Economics* 5, pp. 125–144.
Stulz, René, 1981, "A Model of International Asset Pricing," *Journal of Financial Economics* 9, pp. 383–406.
Summers, Lawrence H., 1991, "The Scientific Illusion in Empirical Macroeconomics," *Scandinavian Journal of Economics* 93, pp. 129–148.

2

Fischer Black's Contributions to Corporate Finance

Stewart C. Myers

Fischer Black's impact on corporate finance is insufficiently noticed. He did not specialize in that subject, and the fame of "Black–Scholes" has drawn attention from his broader contributions.

But Fischer was co-author of the most influential early test of the Capital Asset Pricing Model (CAPM), now probably the most widely used tool for estimating the opportunity cost of capital and valuing risky real assets. The CAPM is not essential to the now-accepted framework of corporate finance—any linear asset pricing model works—but it made that framework accessible and, at least in principle, implementable.

The Black–Scholes option pricing model provided the tools to value the real options embedded in almost all corporate investments. It also revealed the true structure of corporate liabilities. And there are great Black papers on dividend policy, the normative theory of corporate investment, the meaning of accounting income, and on taxes and pension management.

Fischer had a gift for finding interesting ideas in unexplored or neglected territory. He rarely added twigs to existing branches of the literature. Therefore, it pays to take an interest in what interested him. I think some of his more provocative statements:

> [D]ividends that remain taxable will gradually vanish. (Black 1990, p. 5)
>
> The objective of accounting is to use a set of rules that makes the price-earnings ratio as constant as possible. (Black 1980, p. 6)
>
> [W]e might define an efficient market as one in which price . . . is more than half of value and less than twice value. (Black 1986, p. 533)
>
> In the real world of research, conventional tests of [statistical] significance seem almost worthless. (Black 1993, p. 9)

were intended in part to jolt researchers to question conventional methodology or to work on something new. So in this review I will emphasize some things we don't know about corporate finance and also known problems that are often conveniently forgotten. The review is not a complete survey of Fischer's work or of those parts of his work that may touch on corporate finance.

This review starts with the CAPM and its role in the standard framework of corporate finance. Used normatively, this framework requires a way of valuing real assets, and for that task, the CAPM, combined with discounted cash flow (DCF), seems ideal. But of course there are problems, and Fischer's work on real investment decisions suggests an alternative approach, which I describe in section 2.

Section 3 considers real options and the applicability of valuation methods developed for traded options. The last section covers financing, dividend policy, accounting, and tax issues.

1. THE CAPM AND THE STANDARD FRAMEWORK FOR CORPORATE FINANCE

The standard framework for corporate finance starts with a market-value balance sheet.

$PV = \Sigma PV_i$	$D = \Sigma D_i$
PVGO	E
V	V

where $PV = \Sigma PV_i$ is the sum of the values of the firm's projects—"projects" referring to its real assets and existing operations; PVGO is the present value of future investment opportunities; $D = \Sigma D_i$ is the market value of outstanding securities, excepting equity; E is the market value of outstanding common stock; and V is the value of the firm. Note that all entries are current market values. For example, PV_i is the market value of project i if it were separately traded. The assumed objective is maximization of E, the current value of shareholders' wealth.

This *balance sheet model* of corporate finance is more than definitions or an accounting identity. It implicitly assumes that capital markets are tolerably perfect, efficient, and complete. This assumption supports the objective of market-value maximization. It also supports *value additivity*, the assertion that project values add up. The balance sheet says that the value of the firm can be calculated by breaking its assets and operations into discrete projects, valuing each one separately, and summing.

This framework is used normatively as well as positively. It governs how much of corporate finance is taught. But finance teachers and textbook writers can't just describe the firm as a value-maximizing bundle of projects. Their audiences expect analytical methods and advice.

Think of teaching finance in 1970. By then it was clear that market-value maximization made sense. The value additivity principle had been demonstrated for perfect and (sufficiently) complete markets.[1] Modigliani and Miller's (MM's) papers had shown that financing and dividend policies

ought to have only second-order effects: value added should come mostly from the left side of the balance sheet.[2]

Thus, we had the $64,000 question, "How is project value calculated?" The presumptive answer was discounted cash flow:

$$PV_i = \sum_{t=1}^{H} \frac{C_{it}}{(1+r_i)^t} \tag{1}$$

where C_{it} is the expected after-tax cash flow for project i in period t, extending to a horizon date H; and r_i is the opportunity cost of capital (that is, the expected rate of return offered by securities or portfolios with risks matching project i.[3]

But this DCF formula immediately raises three further questions:

1. What assumptions about the expected cash flows C_{it} are necessary for the formula to work?
2. How can the expected cash flows be estimated?
3. What determines r_i, the opportunity cost of capital? Is it really likely to be independent of the timing of cash flows?

The CAPM offered a possible answer to question 3:

$$r_i = r_f + \beta_i (r_m - r_f) \tag{2}$$

where r_f is the risk-free interest rate, $r_m - r_f$ the expected market risk premium, and β_i project i's beta. The formula as written assumes these parameters are constant.

The CAPM was potentially an enormous advance in valuing real assets. Beta could, in principle, be estimated from returns on stocks with risks similar to the project under consideration. Estimating the expected return on the market was harder, but not as tough as estimating expected returns on individual securities. Moreover, betas "added up." The CAPM made it easy to see why present values add and why "portfolio effects" have no place in corporate investment decisions if shareholders have access to perfect and complete capital markets.[4]

But could the CAPM be trusted? By 1970, its theoretical importance was well-understood but its empirical relevance unknown. Thus, the stage was set for Black, Jensen, and Scholes (1972). That paper's econometric setup, carefully designed to avoid bias and reduce measurement errors, inspired confidence. The paper also showed that betas mattered. There was a significant positive relationship between average returns and beta from 1931 to 1965, though returns on low-beta portfolios were too high—higher than predicted by the CAPM—and returns on high-beta portfolios were too low.[5]

In other words, the relationship between return and beta was too flat—but that could be accommodated in generalized versions of the CAPM in which beta plays its customary role. For example, Black (1972) showed how the CAPM changes when there is no truly risk-free asset or when investors face restrictions on, or extra costs of, borrowing. He offered this

generalized model as a possible explanation of the too flat slope of the empirical Security Market Line. By the 1990s, he regarded this model as the *most likely* explanation.[6]

The Black–Jensen–Scholes paper, with contemporaneous work by Miller and Scholes (1972) and others,[7] gave researchers, teachers, and ultimately financial managers confidence in the CAPM and therefore a relatively easy way of thinking about risk and the cost of capital.[8] The CAPM fit easily into the balance sheet model of corporate finance and eventually made that model seem no more than common sense.

Today, that early confidence in the CAPM may seem misplaced. The positive relationship between beta and average return seems to have disappeared since 1965, and other factors, such as size and market-to-book ratios, seem much more powerful in explaining differences in average returns.

Black (1993a) gives a critical review of much of this work and a strong defense of beta and the CAPM.

With Fischer Black on its side, the CAPM can't be completely dead. The CAPM won't disappear until a replacement is found—a replacement with support from theory as well as empirical tests. In the meantime, those involved in applied corporate finance will often end up using the CAPM because it is consistent with the balance sheet model and most of the time seems to give sensible answers. When better asset-pricing models are finally fit for use, the CAPM will depart, but the balance sheet model will stay. That model seems second nature today because the CAPM helped us understand it.

But Fischer would give a final, contrarian twist to this story. The balance sheet model may seem second nature to most, but he didn't fully trust it. For example, he offers "irrational pricing" as a partial explanation for the CAPM's empirical shortcomings (1993a, p. 10). The balance sheet model assumes rational investors. Fischer also disagrees with MM's leverage irrelevance proposition, arguing that corporations should borrow to take advantage of the too flat Security Market Line, thereby substituting corporate for personal leverage. I wish he were here to debate this.[9]

2. FISCHER BLACK ON CORPORATE INVESTMENT DECISIONS

Financial managers, consultants, and textbook writers now seem comfortable with DCF, but a devil's advocate can easily paint Eq. (2) as simpleminded. Here are three common problems:

1. Estimates of expected cash flows several years away are vaporous, biased, or maybe both.
2. Even if the CAPM is OK, some betas can't be estimated with acceptable accuracy, and there is no assurance that future betas approximate past betas. Estimates of the market risk premium are frequently controversial.

3. In order to use a constant (not time-varying) discount rate, detrended cash flows must follow a geometric random walk with constant proportional variance.[10] This implies lognormally distributed cash flows and rules out negative future cash flows if the expected cash flow is positive.[11] (Managers would love to find projects with guaranteed positive cash returns. Unfortunately. . . .[12]) Mean reversion, as might be predicted from competitive responses to unexpectedly high or low cash flow realizations, is likewise ruled out.[13]

Black (1988) presents "a simple discounting rule" that may alleviate some of these problems. His idea can be put as follows. Suppose that a cash flow to be received one period hence is linearly related to the return on the market portfolio:

$$C_{i1} = a + b(1 + r_{m1}) + e_1 \tag{3}$$

where a is a constant and e_1 is independent noise with a mean of zero. The present value of $b(1 + r_{m1})$, which corresponds to b dollars invested in the market, is just b. The present value of the diversifiable error term e_1 is zero. The present value of the constant a is found by discounting at the risk-free rate.[14] Therefore, the value of a cash flow to be received at date 1 is

$$\begin{aligned} \mathrm{PV}_0(C_{i1}) &= \frac{a}{1+r_f} + b \\ &= \frac{a + b(1+r_f)}{1+r_f} \end{aligned} \tag{4}$$

In other words, the certainty equivalent of C_{i1} is $a + b(1 + r_f)$, the expectation of the cash flow conditional on $r_{m1} = r_f$. Black's discounting rule replaces the unconditional expected cash flow with a "conservative" forecast that assumes that investors in the market end up earning no risk premium. This seems natural: in practice, one rarely observes financial managers varying discount rates project by project, but they do dial in different degrees of conservatism in forecasting projects' cash flows.

The discounting rule is easily extended to two or more periods. Suppose the cash flow is to be received at date 2:

$$C_{i1} = a + b(1 + r_{m1})(1 + \mathrm{r}_{m2}) + e_2$$

where e_2 captures the noise between dates 0 and 2. The present value of $b(1 + r_{m1})(1 + r_{m2})$, which corresponds to b dollars invested in the market for two periods, is again b, and

$$\begin{aligned} \mathrm{PV}_0(C_{i2}) &= \frac{a}{(1+r_f)^2} + b \\ &= \frac{a + b(1+r_f)^2}{(1+r_f)^2} \end{aligned} \tag{4a}$$

The extension to a long-lived stream of cash flows, as in the DCF formula given as Eq. (1), is straightforward.

Could this rule help the financial manager? It depends on whether the manager can come up with a conditional forecast of future cash flows. Once such a forecast is in hand, the manager does not need to know the expected market risk premium $r_m - r_f$. The manager does not need to know beta or the opportunity cost of capital. Time-varying or uncertain betas or risk premiums cause no problems. The manager does not even have to worry whether the Security Market Line is too flat. The discount rate for the conditional (certainty equivalent) cash flow forecast is just the spot Treasury rate for the date at which the cash flow is to be received.

But if the manager *starts* with an unconditional forecast, Black's discounting rule is much less help. Such a forecast could, of course, be written down from the unconditional expectation $a + b(1 + r_{m1})$ to its certainty equivalent $a + b(1 + r_f)$, but this writedown would require knowledge of b (equivalent to knowing beta) and of the difference between r_m and r_f (the expected market risk premium). In this case, why not use the conventional DCF formula directly? The only disadvantage of doing so is relying explicitly on the CAPM. (Once the conditional, certainty equivalent cash flow is obtained, the CAPM is discarded. However, the choice of the market portfolio as the traded valuation benchmark may implicitly assume that model. More on this below.)

Fischer claimed that conditional forecasts are, for most people, easier than unconditional ones. I think he's right if the right questions are posed to the forecaster.

Suppose the question is put this way: "Assume that next year's return to investors in the stock market is only the one-year Treasury rate. Then what is next year's expected cash flow for the proposed expansion of our refinery?" This question is almost impossible to respond to directly because the manager is given no way of translating the assumed market return into assumptions about the business conditions relevant to refining.

Thus, Fischer's discounting rule seems to call for a two-step forecast. First construct scenarios for the business variables corresponding to the macroeconomic conditions implied by a market return equal to the risk-free rate. Then ask the manager to forecast cash flow for these scenarios.[15] If everything is done consistently, the result should be the conditional forecast Fischer calls for.

Better still, replace the market with a stock or portfolio that's closer to the project under consideration. For example, the refinery expansion project's cash flows could be forecasted on the assumption that an oil industry portfolio will earn only the risk-free rate. Black's reasoning works for any benchmark security or portfolio that's priced in an efficient and competitive market.

Fischer's discounting rule does not require the CAPM, only that the cash flow can be linked to the traded benchmark by an equation such as (3). Suppose that asset prices conform to a two-factor arbitrage pricing model; for the benchmark portfolio B,

$$r_B = r_f + b_{B1}(r_{\text{factor 1}} - r_f) + b_{B2}(r_{\text{factor 2}} - r_f) \tag{5}$$

where the b's are the benchmark's "betas" on the factor returns and $r_{\text{factor 1}} - r_f$ and $r_{\text{factor 2}} - r_f$ the expected factor risk premiums. Fischer's discounting rule requires that the project cash flow depend linearly on the benchmark's return and that no factor shows up in the noise variable e. For the first cash flow at date 1,

$$C_{i1} = a + b(1 + r_{B1}) + e_1 \tag{3a}$$

This works only if the project cash flows also depend on factors 1 and 2 and have the same relative factor weights as the benchmark portfolio. Otherwise, the "surprises" in r_B will show up in the noise term e, and we can no longer assert that the present value of e is zero, even if it has a zero mean. If the present value of e is not zero, then Fischer's rule does not follow.

In short, you can't apply Fischer's discounting rule without choosing an asset-pricing model. You need to specify a model to know what factors determine expected returns and to choose a benchmark. But if a benchmark can be found, and if cash flow forecasts can be made conditional on the benchmark return equal to the risk-free rate, then knowing the parameters of the asset-pricing model is no longer necessary.

Could Fischer's valuation procedure substantially improve capital investment practice? I think it's worth trying. At least the attempt might wake up casual users of DCF plus CAPM to the assumptions they have been making.

3. REAL OPTIONS

Fischer's discounting rule works only for project cash flows that can be expressed as linear functions of security or portfolio returns. In other words, it does not work for real options or for assets with option-like characteristics. In this respect, it is no better than conventional DCF. To keep things simple, I will set the rule aside and concentrate on the deficiencies of DCF when options are important.

How important are real options in corporate investment decisions? Judging from practice, where explicit real options analyses are very rare, one might conclude that DCF solves, say, 95% of the investment valuation problem, leaving option-pricing methods as a possible 5% refinement. In fact, options are at the heart of the valuation problem in all but the most pedestrian corporate investments. If I am right in this, the option-pricing methods first developed by Black and Scholes (1973) are a first-order contribution to corporate finance.

It's hard to think of an investment project that does *not* include important real options:

1. When a new project is undertaken, no one knows how long it will last. There is no predetermined economic life. Successful projects are extended, failures cut short. If one views each project as

potentially long-lived, then there is a put option to abandon. The exercise price is the value of the project's assets in their next-best use.[16] This abandonment value put is encountered in *all* projects, except for a few with contractually determined lives. Myers and Majd (1990) have shown the put's importance numerically.

2. Some investment decisions are "go or no go," now or never. But when delay is possible, the firm holds a call option to invest. The call is not exercised unless the project's net present value (NPV) is sufficiently far in the money to justify cutting the call's life short. The decision rule, "Invest if NPV is positive," is no longer right.[17]
3. The design of production facilities has to trade off specialization versus flexibility. Flexibility generally costs more, either in investment or production costs, but keeps the facilities useful if the intended product doesn't sell.[18]
4. Most "strategic" investments involve outlays today undertaken to open up further investment opportunities (i.e., options to invest) tomorrow. Thus, a company may enter a new market not to earn immediate high returns but to acquire technology or an established base of customers, or to "get down the learning curve" to lower costs more than later-entering competitors. These advantages then give the option for follow-on investment. There need be no certainty of positive NPV for these investments, only the possibility. In fact, the more uncertainty the better, other things equal, because options on volatile assets are always worth more than options on safe ones.[19]
5. Investments in R&D, though neglected in the finance literature, are similar to strategic investments—made not in expectation of immediate profit but in hopes of generating follow-on investments with positive NPV.

My point is that real options are nearly ubiquitous. They account for PVGO, the present value of growth opportunities in the balance sheet model, and they are embodied in, or attached to, virtually every real asset or investment project. Thanks to Black and Scholes (1973), and the financial engineering techniques built on the Black–Scholes theory, these real options can be valued. I think this is Fischer's biggest single contribution to corporate finance.

But why are practical valuations of real options so scarce? The most common types of real options have been identified (and listed above) and solution techniques laid out. Two comprehensive books on the analysis of real options have been published.[20]

The problem is not that option-pricing methods are untested or that they require unusual or arbitrary assumptions. The methods are routinely used in financial markets worldwide. The assumptions required to apply the methods to real options are no more stringent than the assumptions required to apply DCF.[21]

I attribute the lag in applications to two things. First is simply lack of understanding. Most financial managers are not comfortable with option-pricing methods. They do not fully trust the valuation methods or grasp the meaning of assumptions and inputs. Thus, the numerical packages usually required look like black boxes. Moreover, the language of options is not widely understood. Remember, DCF is not just a tool for valuing projects but a way of *talking about* projects; that is, a framework for assembling information and debating projects' prospects. Unless real options can be talked about, calculations of real option values will not be trusted.

The second, deeper problem is the *fuzziness* of many real options. Their terms are not contractual but part and parcel of the business, so they have to be identified and modeled before inputs can be estimated and valuations calculated. In some cases, this is not a big hurdle. The abandonment put, for example, is easily recognized, and its exercise price is the "salvage" or terminal value used in conventional DCF. But in other cases—for example, strategic investments or outlays for R&D—the real option may be easy to see intuitively but very hard to write down. The option may be too complex, or its boundaries may not be crisply defined.

That may be discouraging for quantitative, normative finance. If the object is positive—that is, to explain corporate investment behavior—option-pricing theory clearly helps. Managers who have never heard of Black and Scholes respond to real options by judgment and common sense. For example, they make strategic investments and commit to R&D even when conventional DCF would advise otherwise, so financial economists may be able to understand financial managers' actions even though we can't tell them what to do.

4. FINANCING, DIVIDEND POLICY, PENSIONS, AND ACCOUNTING

So far, I have concentrated on the biggest general issue in corporate finance, the valuation of real assets and options. But Fischer Black had important things to say on many other topics. I will briefly cover four of them, beginning with financing and the structure of corporate liabilities.[22]

A. Financing

Black and Scholes (1973) is doubly famous. The paper showed how to price options and explained the structure of corporate liabilities.

Common shares are call options, options to take (or retain) the firm's assets by paying off its debt. By put–call parity, we can also say that the value of debt is marked down by the value of a default put: stockholders can put the assets of the firm to its creditors and walk away without further liability. The exercise price of the put is the face value of the debt.

From this insight, it was clear how the *relative* prices[23] of the firm's liabilities are determined. Effects that might have seemed odd (for example, equity value goes up when the volatility of asset returns increases) were obvious. The conflicts of interest between stockholders and creditors were

easy to see.[24] Thus, the Black–Scholes paper accelerated the development of agency theories of capital structure.

The value of a debt guarantee equals the value of shareholders' default put. An extensive literature on debt guarantees, especially deposit insurance, therefore started up shortly after Black–Scholes. The moral hazard problems created by deposit insurance were worked out in theory in the 1970s.[25] They were worked out in practice in the savings and loan industry's subsequent debacle.

Of course, we now use Black–Scholes methods routinely to price all sorts of corporate securities—convertibles, warrants, stock options, debt issues subject to call, and so on. Black and Cox's paper (1976) investigating how bond indenture provisions affected default puts and bond values was especially influential.

B. Dividend Policy

Fischer's main contribution to dividend policy was to remind us that we don't understand it. In "The Dividend Puzzle," he concluded (1976b, p. 8):

> What should the individual investor do about dividends in his portfolio? We don't know.
>
> What should the corporation do about dividend policy? We don't know.

By 1990 (Black 1990, p. 5), he had much more definite opinions:

> Why do firms pay dividends? I think investors simply like dividends. They believe that dividends enhance stock value . . . , and they regard dividends as a more ready source of wealth because they feel uncomfortable spending out of capital.

In the same work, the information content of dividends now is dismissed and the tax disadvantages of dividends emphasized:

> Changing dividends seems a poor way to tell about a firm's prospects.
>
> I think dividends that remain taxable will gradually vanish.

There is no model or analysis to back up these statements, but Fischer's line of thought can be seen pretty clearly. First is the admissability of irrational investors or managers. It's not that people are generally stupid. *Noise* makes it hard to know when our behavior is rational and when it is not; see Black (1986). Second is the recognition of taxes as a first-order effect. The tax burden on cash distributed as dividends (vs. share repurchases) couldn't be clearer. What offsetting benefits do cash dividends provide? Black says investors "like" dividends. He may be right. But I think dividends won't be fully understood until we have a formal, general agency theory of corporate finance.

C. Pensions

Speaking of tax effects, how about tax arbitrage? Black (1980b) and Black and Dewhurst (1981) identify an arbitrage opportunity in corporate pension

funding. Suppose a corporation has a defined benefit pension obligation that it *can* fund. That is, the Internal Revenue Service will allow the firm to contribute, say, $100 million more to its pension fund. So the firm borrows $100 million, puts this amount in the pension, and invests in corporate bonds similar to the corporation's own debt. The new debt liability and pension asset exactly offset, except that interest paid on the corporation's new debt is tax-deductible and the interest earned in the pension fund is tax-free. For long-lived pension liabilities and assets, the present value gained is roughly equal to the marginal corporate tax rate times the amount invested, at today's tax rate about $35 million in this example.[26]

So the rule for blue chip, tax-paying corporations ought to be, "Borrow and fund the maximum the IRS will allow. Put the pension assets in corporate debt." That's Fischer's advice. He's right: you can't argue with arbitrage. I don't understand why his advice isn't more widely followed.[27]

D. Accounting

I close with one of Fischer's most important and interesting insights, namely the correct definition of accounting earnings.

Economists are trained to think of earnings as economic income, that is, cash flow plus change in value. For stocks, income is obviously dividends plus capital gains or losses. Thus, we tend to assume that accounting income equals economic income. Accountants may say they are shooting for economic income, and in some special cases they may even try to measure it. But as Black (1980a, 1993b) shows, economic income can't be the true, general objective of accounting rules. If it were, the rules would attempt to measure *changes* in the value of the firm or its assets, and they clearly do not.

What then is accounting income supposed to measure? As far as I know, the accounting literature does not say. Fischer's answer is the only one that makes sense:

> all the users of accounting reports ". . . want the same kind of earnings figure. They all want earnings to be a measure of value, not a measure of the change in value." (Black 1980a, p. 3)

Therefore,

> "The objective of accounting is to use a set of rules that makes the price–earnings ratio as constant as possible." (Black 1980a, p. 6)

This is a testable statement. It implies that price–earnings ratios vary less cross-sectionally than, say, market-to-book ratios, and that price–earnings ratios are more stable over time than price, earnings, or other common measures such as free cash flow. Fischer's results (in Black 1993b) seem to confirm this hypothesis.

Accounting is the language of practical finance, and accounting numbers are one of the most important sources of information about firms. Yet financial economists have not paid much attention to the "should" and

"why" of accounting. Academic accountants research the "why" but rarely the "should."

As Fischer said (1993b, p. 1):

> I am fascinated by the "should" of accounting rules. We must pick some rules. How should we go about it?

5. NO CONCLUSIONS

What were Fischer's contributions to corporate finance? We don't know yet. He left us with too many open questions and unabsorbed ideas. It's wrong to presume to wrap up Fischer's research. So I offer no conclusions.

NOTES

This paper was prepared for the Berkeley Program in Finance conference "On Finance: In Honor of Fischer Black" and was also published as Myers (1996). I thank John Cox and Milton Harris for helpful comments.

1. Myers (1968).
2. Modigliani and Miller (1958) and Miller and Modigliani (1961). The latter paper also distinguished PVGO from assets in place.
3. In practice, the opportunity cost of capital would be adjusted (decreased) to reflect the value of interest tax shields on debt supported by the project. The simplest device for this adjustment is the weighted average cost of capital. See Brealey and Myers (2003, ch. 19). Fischer Black did not consider tax-adjusted discount rates, so I bypass this issue in this chapter.
4. The irrelevance of portfolio effects was *not* obvious pre-CAPM. The analogy between portfolio selection for investors in securities and a firm's choice of investment projects led John Lintner, the CAPM's co-inventor, to comment that "the problem of determining the best capital budget of any given size is formally identical to the solution of a security portfolio analysis" (1965, p. 65). Several adaptations of portfolio selection techniques to corporate investment decisions were published, including Van Horne (1966) and Weingartner (1966).
5. A companion paper, Black and Scholes (1974), showed that the too flat slope was not attributable to the higher dividend yields of low-beta stocks.
6. Black (1993a).
7. Much of this work appeared in Jensen (1972).
8. The CAPM also made it easy to see the risk–return tradeoffs implied by MM's proposition II relating the expected rate of return on equity to financial leverage. See Hamada (1969).
9. If the Security Market Line has been too flat, then corporations have had a strong incentive to borrow more, and the supply of corporate debt should have expanded to the point where the effects of any restrictions on investors' borrowing are eliminated and leverage irrelevance holds at the margin. Why has this not happened? Black (1993a, p. 17) suggests that corporate managers may be irrationally averse to leverage, perhaps "carry[ing] over the investor psychology that makes individuals reluctant to borrow."
10. In order for r_i to be constant (independent of t), beta must be constant from periods 1 to t. Thus, the covariance of the unexpected change in $PV(C_t)$ with the market return must be constant, and $PV(C_t)$ must follow a geometric random walk. Thus, $PV_t(C_t)$ is lognormal. Since $PV_t(C_t) = C_t$, the cash flow C_t must be lognormal, too. See Myers and Turnbull (1977), Fama (1977, 1996) and Treynor and Black (1976).
11. See Fama (1996).

12. DCF can accommodate negative cash flows if fixed costs are split out and valued separately, so that PV = PV(revenues less variable costs) – PV(fixed costs). Revenues net of variable costs are more likely to be lognormal and strictly positive. However, this approach requires two discount rates. Fixed costs are relatively safe and deserve a low discount rate. Net revenues are also safer than overall project cash flows because discounting fixed costs separately eliminates operating leverage.

13. The geometric random walk is a natural first description of asset values in an information-efficient market. But there is no economic reason why a time series of (detrended) cash flows should behave in the same way.

14. Actually, the after-tax risk-free rate (i.e., $r_f(1 - T_c)$, where T_c is the marginal corporate rate. See Myers and Ruback (1992). Here I have left out taxes for simplicity.

15. Construction of the scenarios would start with values for macroeconomic variables consistent with a stock market rate of return equal to the risk-free rate. The macroeconomic variables would in turn imply conditional forecasts for the relevant industry and company variables. Of course, there would be a large number of possible scenarios consistent with $r_f = r_m$ because various combinations of macro economic variables could yield that market performance. A small number of easy-to-interpret scenarios would have to be chosen. For consistency, each would have to fit in equations such as (3) without bias; that is, with the expectation of the noise term e equal to zero.

16. The same project could be modeled as short-lived, but including a call option to reinvest. The exercise price is the same.

17. See McDonald and Siegel (1986), Ingersoll and Ross (1992), and Ross (1995).

18. See Triantis and Hodder (1990).

19. See Myers (1984) and Brealey and Myers (2003, ch. 22).

20. Dixit and Pindyck (1996) and Trigeorgis (1996).

21. Real option applications require identification of an underlying asset, usually valued by DCF. For example, the present value of a project with uncertain life is the sum of (1) the DCF value of the project assuming it will last a very long time and (2) the abandonment put. If financial markets are sufficiently complete to justify value maximization as the firm's objective and DCF valuation of the underlying asset, then they are also complete with respect to options contingent on that asset's future value. In other words, if investment in the project does not expand investors' opportunity set, acquisition of an option on the project will not expand it either. See also Mason and Merton (1985, pp. 38–39).

22. This subsection was not included in the original version of the paper at the conference honoring Fischer Black. Milton Harris noted my oversight in his comments at the conference. I added the subsection to fill an obvious hole in the paper.

23. Nothing in option-pricing theory upsets the Modigliani–Miller theorem that the total value of the firm does not depend on the nature of securities issued against its assets.

24. Jensen and Meckling (1976) stressed the temptation to increase asset risk once debt was issued. Myers (1977) relied on option valuation theory to show the underinvestment problem, that is, shareholders' reluctance to invest when the firm has risky debt outstanding.

25. Merton (1978) is an early example.

26. The pension contribution is tax-deductible regardless of whether the pension is funded now or later. This tax shield has nothing to do with pension funding policy. However, the tax arbitrage argument does ignore the possibility of default on pension obligations.

27. Putting the pension assets in corporate debt guarantees arbitrage but may not be necessary for it. Suppose, for example, that the CAPM is strictly right. Then pension assets put in the stock market would generate the same present value of tax savings. However, in this case the pension assets and liabilities would not be well-hedged.

REFERENCES

Black, F. (1972), "Capital Market Equilibrium with Restricted Borrowing." *Journal of Business* 45 (July), 444–455.

Black, F. (1976a), "The Accountant's Job." *Financial Analysts Journal* 32 (September/October), 18.
Black, F. (1976b), "The Dividend Puzzle." *Journal of Portfolio Management* 2 (Winter), 5–8.
Black, F. (1980a), "The Magic in Earnings: Economic Earnings Versus Accounting Earnings." *Financial Analysts Journal* 36 (November/December), 3–8.
Black, F. (1980b), "The Tax Consequences of Long-Run Pension Policy." *Financial Analysts Journal* 36 (July/August), 21–28.
Black, F. (1986), "Noise." *Journal of Finance* 41 (July), 529–543.
Black, F. (1988), "A Simple Discounting Rule." *Financial Management* 17 (Summer), 7–11.
Black, F. (1990), "Why Firms Pay Dividends." *Financial Analysts Journal* 46 (May/June), 5.
Black, F. (1993a), "Beta and Return." *Journal of Portfolio Management* 20 (Fall), 8–18.
Black, F. (1993b), "Choosing Accounting Rules." *Accounting Horizons* 7 (December), 1–17.
Black, F. (1993c), "Estimating Expected Return." *Financial Analysts Journal* 49 (September/October), 36–38
Black, F. and Cox, J. C. (1976), "Valuing Corporate Securities: Some Effects of Bond Indenture Provisions." *Journal of Finance* 31 (May), 351–368.
Black, F. and Dewhurst, M. P. (1981), "A New Investment Strategy for Pension Funds." *Journal of Portfolio Management* 7 (Summer), 26–34.
Black, F., Jensen, M. C., and Scholes, M. (1972), "The Capital Asset Pricing Model: Some Empirical Tests." In Michael C. Jensen, ed., *Studies in the Theory of Capital Markets* (New York: Praeger), 79–121.
Black, F. and Scholes, M. (1973), "The Pricing of Options and Corporate Liabilities." *Journal of Political Economy* 81 (May/June), 637–654.
Black, F. and Scholes, M. (1974) "The Effects of Dividend Yield and Dividend Policy on Common Stock Prices and Returns." *Journal of Financial Economics* 1 (May), 1–22.
Brealey, R. A. and Myers, S. C. (2003), *Principles of Corporate Finance*, Seventh Edition (New York: McGraw-Hill).
Dixit, A. K. and Pindyck, R. S. (1994), *Investment under Uncertainty* (Princeton, N.J.: Princeton University Press).
Fama, E. F. (1977), "Risk-Adjusted Discount Rates and Capital Budgeting under Uncertainty." *Journal of Financial Economics* 5 (August), 3–24.
Fama, E. F. (1996), "Discounting under Uncertainty," *Journal of Business* 69 (October), 415–428.
Hamada, R. S. (1969), "Portfolio Analysis, Market Equilibrium and Corporation Finance." *Journal of Finance* 24 (March), 13–31.
Ingersoll, J. E. and Ross, S. (1992), "Waiting to Invest: Investment and Uncertainty," *Journal of Business* 65 (January), 1–30.
Jensen, M. C., ed. (1972), *Studies in the Theory of Capital Markets* (New York: Praeger).
Jensen, M. C. and Meckling, W. H. (1976), "Theory of the Firm: Managerial Behavior, Agency Costs and Ownership Structure." *Journal of Financial Economics* (October), 305–360.
Lintner, J. (1965), "Optimal Dividends and Corporate Growth under Uncertainty." *Quarterly Journal of Economics* 77 (February), 59–95.
Mason, S. P. and Merton, R. C. (1985), "The Role of Contingent Claims Analysis

in Corporate Finance." In E. I. Altman and M. G. Subramangam, eds., *Recent Advances in Corporate Finance* (Homewood, Ill.: Irwin).
McDonald, R. and Siegel, D. (1986), "The Value of Waiting to Invest." *Quarterly Journal of Economics* 101 (November), 707–728.
Merton, R. C. (1974), "On the Pricing of Corporate Debt: The Risk Structure of Interest Rates." *Journal of Finance* 29 (May), 449–470.
Merton, R. C. (1978), "On the Cost of Deposit Insurance When There Are Surveillance Costs." *Journal of Business* 51 (July), 439–452.
Miller, M. H. and Modigliani, F. (1961), "Dividend Policy, Growth and the Valuation of Shares." *Journal of Business* 34 (October), 411–433.
Miller, M. H. and Scholes, M. (1972), "Rates of Return in Relation to Risk: A Reexamination of Some Recent Findings." In Michael C. Jensen, ed., *Studies in the Theory of Capital Markets* (New York: Praeger), 47–78.
Modigliani, F. and Miller, M. H. (1958), "The Cost of Capital, Corporation Finance and the Theory of Investment." *American Economic Review* 48 (June), 261–297.
Myers, S. C. (1968), "Procedures for Capital Budgeting under Uncertainty." *Industrial Management Review* 9 (Spring), 1–20.
Myers, S. C. (1977), "Determinants of Corporate Borrowing." *Journal of Financial Economics* 5 (November), 147–175.
Myers, S. C. (1984), "Finance Theory and Financial Strategy." *Interfaces* 14 (January–February), 126–137.
Myers, S. C. (1996), "Fischer Black's Contributions to Corporate Finance." *Financial Management* 25 (Winter), 95–103.
Myers, S. C. and Majd, S. (1990), "Abandonment Value and Project Life." In F. Fabozzi, ed., *Advances in Futures and Options Research*, volume 4 (Greenwich, Conn.: JAI Press).
Myers, S. C. and Ruback, R. (1992), "Discounting Rules for Risky Assets." Working paper, MIT.
Myers, S. C. and Turnbull, S. M. (1977), "Capital Budgeting and the Capital Asset Pricing Model: Good News and Bad News." *Journal of Finance* 32 (May), 321–332.
Ross, S. A. (1995), "NPV and the Investment Timing Option." *Financial Management* 24 (Autumn), 96–101.
Treynor, J. L. and Black, F. (1976), "Corporate Investment Decisions." In Stewart C. Myers, ed., *Modern Developments in Financial Management* (New York: Praeger), 310–327.
Triantis, A. J. and Hodder, J. E. (1990), "Valuing Flexibility as a Complex Option." *The Journal of Finance* 45 (June), 549–565.
Trigeorgis, L. (1996), *Real Options, Managerial Flexibility and Strategy in Resource Allocation* (Cambridge, Mass.: MIT Press).
Van Horne, J. (1966), "Capital Budgeting Decisions Involving Combinations of Risky Investments." *Management Science* 13 (October), B84–92.
Weingartner, H. M. (1966), "Capital Budgeting of Interrelated Projects." *Management Science* 12 (March), 485–516.

3

Crisis and Risk Management

Myron S. Scholes

From theory, alternative investments require a premium return because they are less liquid than market investments. This liquidity premium varies considerably over time as a function of preferences, leverage technology, the developments in financial technology, and changes in institutional arrangements. The dynamics of the liquidity premium depend on institutional reactions to financial crises.

During 1997–1998, we saw the movement of a financial crisis around the world. It started in Southeast Asia, moved through Latin and South America, and then visited Russia and returned again to South America. The financial crisis also infected Europe and the United States, especially during August through October of 1998.

The increase in volatility (particularly in the equity markets) and the flight to liquidity around the world led the firm with which I was associated, Long-Term Capital Management (LTCM), to experience an extraordinary reduction in its capital base. This reduction in capital culminated in a form of negotiated bankruptcy. A consortium of 14 institutions with outstanding claims against LTCM infused new equity capital into LTCM and took over the firm and the management of its assets. They hired LTCM's former employees to manage the portfolio under their direct supervision and with sufficient incentives to undertake the task efficiently.

Although the Federal Reserve Bank (FRB) facilitated the takeover, it did not bail out LTCM. Many debtor entities found it in their self-interest not to post the collateral that was owed to LTCM, and other creditor entities claimed to be ahead of others to secure earlier payoffs. Without the FRB acting quickly to mitigate these holdup activities, LTCM would have had to file for bankruptcy, for some a more efficient outcome but for others a far more costly outcome for society. If there was a bailout, it failed: LTCM has been effectively liquidated.

Because of LTCM, the press and others have taken the opportunity to criticize financial modeling and, in particular, the value of option pricing models. In truth, mathematical models and option pricing models played only a minor role, if any, in LTCM's failure. At LTCM, models were used to hedge local risks. LTCM was in the business of supplying liquidity at levels that were determined by its traders. In 1998, LTCM had large positions, concentrated in less-liquid assets. As a result of the financial crisis,

LTCM, too, was forced to switch from being a large supplier to a large demander of liquidity at a cost that eliminated its capital.

Although the Russian default, the LTCM bankruptcy, and the financial difficulties of other financial service firms are the most visible manifestations of the crisis of the late summer and fall of 1998, to this day we observe much greater volatility and lack of liquidity in many debt-related and equity-related financial markets. For example, during the summer of 1999, three- to five-year long-dated volatility on the Standard and Poor's (S&P) 500 index was quoted in the 25–30% range, average volatility levels on the S&P index that have not been seen before. To be consistent with market expectations, the realized quarterly volatility on an annualized basis of the S&P 500 would have to average 30% over the next five years, and even higher levels starting one year from now, since the current quoted one-year volatility is far less than 30%. In my view, this is extremely unlikely even given the evolving nature of the stocks that make up the index. To put this in perspective, the quarterly realized volatility of the S&P 500 has averaged well below 15% over the last ten years and has never averaged more than 25% in any five-year period.

In addition, credit spreads and mortgage spreads have widened dramatically. Although early in 1999 spreads narrowed somewhat, during the summer of 1999 they widened to even higher levels than those of August–September of 1998. For these spreads to be default premiums, the market must expect large numbers of defaults and defaults with little chance of recovery.

Moreover, during August 1999, the ten-year on-the-run swap spread was as high as 112 basis points over Treasuries, more than 15 basis points greater than at the height of the September 1998 crisis. These spread levels are extraordinary in that swap spreads (in basis points) were generally in the high 20s to the low 30s from 1992 to mid-1998 and never reached this level even in 1990 when banks, including Citicorp and Bank of America, were experiencing extreme difficulties.

It is hard to believe that these spread levels are attributable only to expectations of defaults in the credit market. Take the off-the-run swap spread as an example. The London Interbank Offered Rate (LIBOR) is set for a time frame, say 3 months, by averaging the quoted borrowing rates on a truncated set of the then 16 top-rated banks in the world and does not depend on the survivorship of any particular bank. That is, if a bank were to become risky because its own prospects had diminished, it would be excluded from the computation of the next LIBOR index. Thus, for swap spreads to be entirely credit spreads, the market must perceive that the entire worldwide banking sector is to experience difficult times. What is even more amazing is that this perception would have to be true not for this coming year but for nine years starting one year from now. Currently, 1-year LIBOR is quoted at only 25–35 basis points over a general-collateral-reverse repurchase agreement (reverse REPO). That is, to borrow Treasury bonds to sell to someone else in the market and to return similar bonds to the lender,

the bond borrower would receive about 30 basis points below LIBOR. Thus, for the swap spread to be a credit spread, LIBOR must increase dramatically relative to REPO, on average, during the nine years starting one year from now.

If these spreads are not entirely credit-related, they must be liquidity spreads. At different times, the market demands more liquidity and will pay for it. During the last couple of years, the number of liquidity providers diminished. Many financial institutions that previously devoted part of their capital to earning returns by supplying liquidity to the market withdrew from doing so or would only commit capital at much higher expected premiums. To provide liquidity, an investor must have a longer horizon than the average market participant. Interestingly, because the liquidity premium is generally small relative to the expected return on alternative investments, liquidity providers are generally leveraged investors that must hedge other factor exposures. For them, risk management, particularly during a crisis, when both credit risk and liquidity risk premiums balloon, is of crucial importance.

1. RISK MANAGEMENT

Understanding risk management technology provides insights into the dynamics of liquidity premiums in asset returns. The risk management practice at large financial institutions such as Citicorp or Merrill Lynch affects the supply of liquidity and therefore the required liquidity premium. As liquidity premiums change, credit spreads and other spreads increase in the debt and equity markets around the world.

For a financial institution, a conventional balance sheet does not provide adequate information to insiders or to outsiders such as investors or creditors as to the risk of the entity. Balance sheet leverage is a reduced-form static measure of risk; it provides no forecast of the firm's profit and loss as economic factors unfold in the economy.

A risk management system is an exposure-accounting system and a control system. An exposure-accounting system is a dynamic system that gives managers an opportunity to assess the effects of changes in economic factors such as interest rate movements, yield curve shifts and reshaping, currency and commodity price moves, stock price movements, etc., on the economic profit and loss of the entity. It determines the firm's need for capital to support its positions.

During the last five or so years, value-at-risk (VAR) has become an accepted standard in the financial industry, and it forms the basis for determining a bank's regulatory capital for market risk. Many financial entities use VAR as a dynamic risk measure. VAR is often disclosed to investors. This approach to exposure accounting assumes that the future movements in risk factors are similar to past movements. That is, the variances and correlation matrix among factor exposures affecting profit and loss do not change over time. They are assumed to be stationary and normally distrib-

uted. The VAR measure is a probabilistic measure of loss potential, measured over a specified holding period and to a specified level of statistical confidence. For example, the VAR might be computed to be $100 million for a two-week period with 99% probability. Loosely put, there is about a 1% chance that a loss greater than $100 million would be sustained in the next two weeks.

Correlation patterns and variances, however, are not stationary, especially when market prices move dramatically. Factors that might exhibit low levels of correlation or association most of the time appear to be highly correlated in volatile times. When the values of nearly all asset classes are moving in lockstep, diversification is not helpful in reducing risk. The actual realized correlation patterns appear to be close to one. In these times, the volatility of profits and losses will be far greater than VAR would predict. Liquidity and risk premiums change dramatically as well, resulting in far greater measured asset volatility.

In periods of extreme market stress, such as globally in 1987, 1990 in Japan, 1991 in Europe, 1992 in Sweden, 1994 in the United States, 1995 in Mexico, and 1997–1999 in Asia, the Americas, Europe, and the United States, many statistically uncorrelated activities using historical data exhibited high degrees of association. For example, in 1998 the spreads over Treasuries widened on U.S. AAA bonds, AAA commercial mortgage pools, credit instruments, country risks, and swap contracts. Moreover, volatilities on stocks and bonds increased to levels that had not been observed in decades.

For example, on August 21, 1998, one week after Russia defaulted on its debt, swap spreads, the difference between AA bank risk and U.S. Government bonds in the 10-year sector, shot up from 60 basis points to 80 basis points in one day. This 20 basis point change was a 10 standard deviation move in the swap spread. After this date, the volatility of the swap spread increased from 8/10 of a basis point a day to 8 basis points a day and remained high throughout 1999.

To protect against extreme shocks such as these, many financial entities impose stress-loss limits on their portfolios. These stress limits attempt to protect against extreme shocks in individual risk factors as well as groups of risk factors. Their intent is to capture more extreme moves, the so-called "tail exposures." These stress limits might preclude the entity from concentrating in any one strategy or project, or from maintaining a position even though additional or continued investment had expected positive present value when using conventional present-value analysis to decide its worth.

Before the financial crisis in August 1998, most financial institutions were well within the guidelines for capital adequacy specified by the Bank for International Settlements (BIS) on standard measures such as VAR, leverage, or Tier I or Tier II capital. Then, in August, investors rushed to more liquid securities, increasing the demand and price of liquidity around the world. Investors liquidated large portfolios of assets in Asia and Latin and

South America by selling into a market with high transaction costs. Many leveraged investors were forced to liquidate holdings to cover margin requirements.

Maybe part of the blame for the flight to liquidity lies with the International Monetary Fund (IMF). Investors believed that the IMF had given implicit guarantees to protect their investments against country-specific risks in the underdeveloped and less-developed regions of the world. But when Russia defaulted on its debt obligations, market participants realized that the implicit guarantees were no longer in place.

In an unfolding crisis, most market participants respond by liquidating their most liquid investments first to reduce exposures and to reduce leverage. Transaction costs, including spreads, tend to be smaller in these markets. Since it is not possible to know the extent of the unfolding crisis, holding and not selling the less liquid instruments is similar to buying an option to hold a position. More liquid markets tend to be large and can handle large trading volumes relatively quickly. But, after the liquidation, the remaining portfolio is most likely unhedged and more illiquid. Without new inflows of liquidity, the portfolio becomes even more costly to unwind and manage.

There has been little modeling of the stress-loss liquidity component of risk management and its implication for the price of liquidity. Financial institutions use stress-loss limits and capital cushions to mitigate crisis risk. They have moved from a static risk measure (leverage) to a dynamic risk measure (VAR) with a static overlay (a stress-loss cushion) to provide an extra capital reserve in the event of a stress loss. A static risk measure, however, is not time-consistent. In a dynamic world, a dynamic policy is required that describes what actions to take as the cushion deteriorates or after it has been breached.

As is commonly known, as the adjustment gap between the stop-loss (demanding liquidity) and the price at which you reacquire the position (providing liquidity) becomes small enough, the strategy is equivalent to replicating an option in the Black–Scholes world. Thus, a dynamic stop-loss policy values an option.

A put option provides the equivalent of a dynamic liquidity cushion. A put-protected position self-liquidates as money is lost and markets become more illiquid. The cost of this protection is the value of liquidity. In reality, put options replace the role of the static stress cushion.

Conceptually, to value risk or to price reserves for its position, an entity must value the options it is not buying to protect itself in the event that it has an increased demand for liquidity. Since the stress limit is not priced, this tends to create the wrong capital allocation incentives within financial entities.

If an entity buys options, it protects itself in a crisis and against negative jumps in asset values. If, however, it establishes its own reserves, they must increase as position values fall, thereby forcing a dynamic adjustment to reserves. The cushion, so to speak, must be dynamic. The entity, how-

ever, by dynamically hedging on its own account, cannot protect itself entirely. Gaps or jumps (unless of specific forms) cannot be hedged by employing internal dynamic adjustments. But this dynamic cushion is superior to the static risk cushions. Many financial products have two-way markets. Financial entities enter into contracts with customers and with other institutions. They tend to be long and short contracts with customers and other dealers. Because its exposures tend to net, the net risk position is quite low. This activity is called a matched book or agency business. The gross number of positions, however, becomes quite large. In addition, to reduce credit risk, many dealers and sophisticated entities post collateral to each other on price moves in the amount of the payment that would have to be made to a counterpart on a forced liquidation.

For many of its proprietary products, however, financial entities need to hedge risks by using the bond or equity markets. In a market crisis, the greatest losses most likely occur in this hedged-book business. In August of 1998, those who were receiving in swaps and hedging by shorting government issues or selling long-dated options and hedging by buying equity forward contracts suffered the greatest loss as spreads widened dramatically. The hedged books suffered loss because of changes in the economic fundamentals and because of an unanticipated jump in the demand for liquidity. Again in the summer of 1999, as corporations and other entities were issuing bonds or hedging an anticipated increase in interest rates, the demand for liquidity increased with a decrease in institutional supply as these institutions also demanded liquidity. Stress-loss cushions were violated, and many financial entities reduced the size of their hedged-book positions at significant liquidation costs. Because the stress-loss cushions are static, entities have an ill-defined policy on when to supply liquidity and in what amounts. As a result, banks and financial entities are not the natural suppliers of liquidity and add to the volatility in financial crises.

2. CONCLUSION

In recent years, regulators have encouraged financial entities to use portfolio theory to produce dynamic measures of risk. VAR, the product of portfolio theory, is used for short-run day-to-day profit and loss risk exposures. Now is the time to encourage the BIS and other regulatory bodies to support studies on stress test and concentration methodologies. Planning for crises is more important than VAR analysis, and such new methodologies are the correct response to recent crises in the financial industry.

The financial industry will become more creative in supplying or finding a source of supply of "liquidity" options and contingent capital to supply liquidity in times of stress. As the reinsurance market for excess loss has developed, similar markets could develop and add value in financial markets. This becomes an important role for alternative investments. The financial industry's use of the stop-loss technology produces volatility in liquidity premiums in many financial instruments. It takes time, however,

to develop new products and to educate potential new entrants into the market to replace them. More dynamic cushions will reduce the fluctuations in the price of liquidity, and markets will become less prone to a financial crisis. The marketplace will find alternative providers and ways to supply liquidity.

From time to time, it is argued that financial quantitative modeling has failed because, even with the increasing number of measurement techniques, their use has not prevented financial crises or financial failures. Financial crises are prevalent throughout time and across countries. Although this might seem somewhat discouraging and a slam against financial modeling, it is not. This is so because better risk measurement models reduce costs and, as a result, financial firms develop new products and activities that make their constituents better off. Most likely, these new developments increase risk levels once again. As costs fall, economics predicts that agents move to the envelope once again.

4

Hot Spots and Hedges

Robert Litterman

1.

This chapter is a tutorial on portfolio risk management. It describes how to identify the primary sources of risk in complex portfolios—the "hot spots"—and how to identify the trades that will reduce those risks—the "hedges." While this topic, identifying and reducing risk, would seem to be at the heart of risk management, many of the ideas presented here are not well-understood. Unfortunately, too much of the recent focus in risk management has been directed elsewhere—for example, toward understanding and incorporating the risks of derivative securities. While derivatives can create complex and highly nonlinear payoffs, they are often not the primary source of risk in a portfolio. Also, while risk managers have focused much attention recently on developing firmwide systems to aggregate information from many traders or portfolio managers, they have generally not yet concentrated on how to use that information to obtain a comprehensive understanding of enterprise risk. More generally, while risk managers have also recently developed a greater appreciation of the statistical nature of risk and the need to use statistical measures such as value-at-risk (VaR)[1] to quantify, monitor, and set limits on risk-taking activities, the tools required to understand and incorporate these statistical measures in portfolio risk management have lagged behind.[2] This most important area of risk management—the development of portfolio analytics based on the statistical understanding of risk—has not received as much attention as it deserves. We hope this chapter helps to address this concern by describing a number of portfolio analysis tools that we have developed at Goldman Sachs in recent years to better understand, manage, and monitor the risks in our clients' investment portfolios as well as our own trading positions. We focus on the concepts of risk management and how they apply in realistic trading and portfolio management contexts rather than on details of implementation.

The plan of the chapter is as follows. In section 2, we discuss volatility and VAR, the two main statistical measures of portfolio risk. In section 3, we contrast them with the accounting measures that currently are commonly used for understanding portfolio risk—measures such as position reports, duration analysis, and stress tests. In section 4, we consider what

additional information is needed for managing, as opposed to monitoring, risk. The key idea here is the need for a decomposition of risk. In introducing this decomposition, we make extensive use of a simple triangle analogy to illustrate how the risks of assets add up to create the risk of the portfolio and the role played by the correlations of their returns. We also describe the Trade Risk Profile, a graph of portfolio risk as a function of trade or position size. In section 5, we explain a number of additional portfolio analysis tools that we have found useful in managing risk at Goldman Sachs. These include the Best Hedges report, which helps locate potential portfolio hedges; the Market Exposure report, which characterizes the exposure to certain predefined risk factors; the Best Replicating Portfolio report, which provides a simplified representation of complex portfolios; and the Implied Views Analysis, which helps make clear whether the views that a portfolio actually represents are the same as those it is intended to represent. Section 5 ends with a discussion of how these analyses are related to one another and to the construction of optimal portfolios. Section 6 provides a brief summary.

Portfolio Risk Analysis

This chapter builds on a framework for risk analysis that should be familiar to most investment managers. The framework is generally referred to as "Modern Portfolio Theory" and was first described mathematically by Harry Markowitz in the early 1950s (Markowitz 1952) as "mean–variance" analysis and later extended by William Sharpe in the mid 1960s (Sharpe 1964) as the basis for the Capital Asset Pricing Model. Professors Markowitz and Sharpe were awarded Nobel Prizes in 1990 for their work. This is also the framework adopted by Black in his 1989 extension of the Sharpe model (Black 1989) and by Black and Litterman (1991) in our application to asset allocation.

Much of what is described in this chapter is implicit, if not explicit, in the original work of Markowitz. The basic approach is to recognize that an investor faces a trade-off between risk and return and to develop the implications of that trade-off. Markowitz suggested the use of volatility as a measure of risk. The volatility of a portfolio is most often obtained by combining measures of exposures in a set of asset classes with a covariance matrix that provides numerical estimates of the volatilities and correlations of the returns of those asset classes.[3]

Even in his original paper, Markowitz recognized that volatility is an oversimplification of the concept of risk. Moreover, the use of exposures and a covariance matrix is itself just a "linear approximation" to measuring the volatility of the portfolio. The approximation itself may be poor when the exposures change with market moves—that is, when the portfolio has embedded nonlinearities such as those that arise in options and other derivative securities. Such nonlinearities are much more commonly found in portfolios now than they were when the theory was developed.

Thus, the traditional portfolio theory has important limitations that have to be recognized and that have limited its use in risk management today.

This lack of common application is unfortunate, however. Many risk managers today seem to have lost sight of the fact that the key benefit of a simple approach, such as the linear approximation implicit in traditional portfolio analysis, is the powerful insight it can provide in contexts where it is valid. With very few exceptions, portfolios will have locally linear exposures about which the application of portfolio risk analysis tools can provide useful information. Moreover, tools such as stress tests, developed for understanding nonlinearities in specific securities, are generally not well-suited for understanding portfolio effects—for example, the impacts of the volatilities of different assets and their correlations on portfolio risk.

Thus, while we emphasize that it would be extremely dangerous to suggest that you can manage risk using only linear approximations, we also feel that the insights afforded by the types of portfolio analytics originally developed by Markowitz, and extended here, do apply broadly and should be a part of each risk manager's information set. Moreover, as shown in this chapter, when nonlinear risks are important, extensions of the linear portfolio analytics can be developed.

2. MEASURES OF PORTFOLIO RISK

Risk is inherently a statistical concept. In the context of investments, risk refers to the degree of uncertainty in the distribution of gains and losses. A complete picture of risk requires a complete description of this distribution. In practice, however, investors focus on summary measures of the degree of dispersion in this distribution. Two of the most common measures of risk, which we focus on in this chapter, are volatility and value-at-risk.

Volatility and VaR have different strengths when used to characterize the risk in a portfolio. Volatility is the usual measure of the uncertainty of investment portfolios. Volatility—that is, a measure of one standard deviation in the return distribution—provides an estimate of the size of a "typical" return over some period, usually a year. VaR, on the other hand, which has become quite popular for characterizing the risk of trading positions, focuses on a point in the distribution of gains and losses. The VAR is the size of a loss that occurs with a specified probability over a particular period of time. If certain conditions are met—for example, the time period is the same, the return distribution is normal, and the specified probability is approximately one-sixth—then the VaR is the same as the volatility. More commonly, however, the VaR applies over a shorter period of time, such as a day; the probability is much smaller—for example, one in a hundred—and the return distribution is nonnormal. In such cases, the VaR is generally many multiples of the volatility and, rather than representing a typical return, is intended to represent a loss that, while very unlikely, will occur every so often.

At Goldman Sachs, we use many approaches to measure and characterize risk. We rely heavily on both the daily volatility of positions and a daily once-per-year VaR. Given the nonnormality[4] of daily returns that we find in most financial markets, we use as a rule of thumb the assumption that four-standard-deviation events in financial markets happen approximately once per year.[5] Given this assumption, the daily once-per-year VaR for portfolios whose payoffs are linear is approximately four standard deviations. Of course, when positions include nonlinear responses, this rule of thumb breaks down. Depending on the nature of the nonlinearity of the instruments being held, the once-per-year VaR of the distribution of returns can be virtually any multiple of its volatility, smaller or larger than four standard deviations.

As a measure of risk, volatility has the advantage that, because it represents a typical event, it is easy for the trader or the risk manager to validate through observations of a small number of events. For example, if we observe three or four days in a row with returns greater than a measured one standard deviation, we quickly start to question our measure. A few months of daily observations would generally be adequate to validate or raise suspicion about a volatility measure. On the other hand, volatility has the disadvantage that the risk manager is not particularly concerned with determining the typical gain or loss; his main concern is what to be prepared for in a rare event. VaR has the advantage that it tries to quantify this concern precisely. Although it has the disadvantage that it may appear to be more difficult to validate, such is the nature of any estimate of a low-probability event; this limitation can't really be avoided. The important point to remember is that the VaR differs from a simple multiple of volatility only when there are nonlinearities, and when this is the case, VaR is the more relevant measure.

To illustrate this difference between VaR and volatility measures of risk, consider what happens when you buy an out-of-the-money call option. This security is like a lottery ticket; most likely it will expire worthless, but there is a chance that it will become highly valuable. The volatility of the payoff of this security is relatively high, but as with a lottery ticket, you might consider this volatility as "upside potential" rather than risk. Thus, at least at the time of purchase, the risk is not well-represented by this volatility. The most that can be lost is the premium, which for an out-of-the-money option will be small relative to the volatility. VaR, on the other hand, captures the fact that the largest possible loss is the premium, and the VaR accurately reflects the true risk. If the market rallies and the call option goes in the money, then the risk increases and the VaR, which now may be several multiples of the volatility, will accurately reflect this fact.

In figures 4.1 and 4.2, we show two histograms that illustrate the difference between the type of distribution generated by a security with a linear payoff and the distribution generated by a nonlinear payoff. In the former case, the daily once-per-year VaR is approximately four daily standard deviations. In the latter case, the VaR can be a much smaller or larger

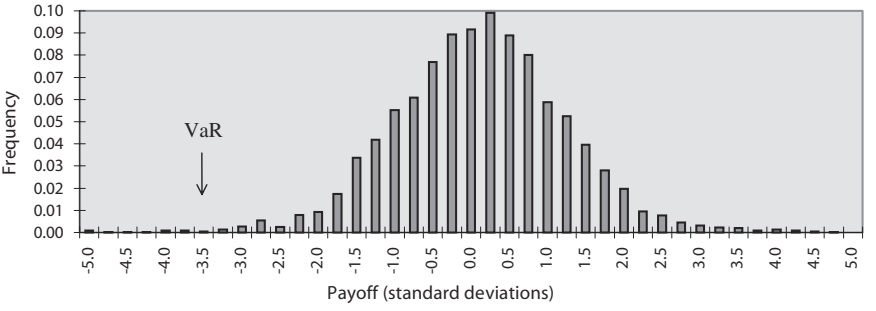


Figure 4.1. Linear Payoff Outcomes

multiple; in the example shown here—the distribution of outcomes from being long an out-of-the-money option—the VaR is just over one standard deviation.

3. UNDERSTANDING RISK

There are many contexts in which the sources of risk in a portfolio are not obvious. A simple example is a portfolio of individual stocks being managed against a market index. A more complicated example would be a global portfolio of long and short positions in bonds, futures, and options being managed by a fixed income trader. Another example is the assets of a pension fund relative to its strategic benchmark. A final example is the "enterprise" portfolio of assets and liabilities of a bank, insurance company, or other financial institution. These are all examples of contexts where the types of analysis that we describe here may usefully be applied.

To clarify the ideas in this chapter, we will pursue three examples in detail. The first is a global equity portfolio managed relative to a market-weighted equity benchmark. This is an example in which the basic risks are locally linear, but understanding the risk is complicated by the interactions of different volatilities and correlations. The second example will be a set of Eurodollar futures and options positions representative of a proprietary

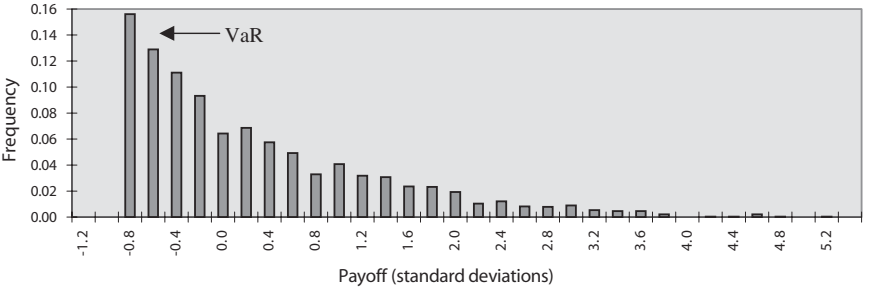


Figure 4.2. Nonlinear Payoff Outcomes: Call Option Payoff Function

trader. In this case, there are fewer risk factors, but nonlinearities are significant. The third example is an actual set of positions held by Goldman Sachs in its European fixed income businesses one day in the summer of 1996. This example illustrates how we put the ideas in this chapter to use in the risk management of our own trading positions. The latter two examples were generated prior to the conversion to the Euro, making the formerly independent currencies separate sources of risk.

In trying to understand the sources of risk embedded in portfolios such as these, it is often not very helpful to simply look at the positions. Generally, there will be too many individual securities, and the volatilities and correlations that create portfolio risk will not be easily discerned. Consider our first example, the global equity portfolio. Even when the number of positions is reduced—for example, by aggregating the individual positions into total exposures in asset class buckets—the relationships among the asset classes may not be clear.[6] The asset class positions are shown as deviations from the benchmark in table 4.1.[7]

Such a table is often used to summarize the risks in a portfolio. Unfortunately, simply viewing the positions—in this case deviations from a benchmark—is insufficient to reveal the important sources of risk, as there are still too many different volatilities and correlations that need to be taken into account. In the second example, that of a proprietary trading portfolio, the problem is not the number of positions—just the four positions as shown in table 4.2—but rather the nonlinearities that play a prominent role, making the risk analysis more complex.

Historically, risk managers have often tried to treat risk as if it were an "accounting" problem. Positions were somehow converted into risk equivalents and added together. For example, in fixed income markets, participants have for many years scaled their positions into units of a common duration. In this way, each position is converted into a basis—for example, a number of "10-year duration equivalents"—which should have approximately equal sensitivity to the main source of fixed income risk, a parallel movement in interest rates. Whether the unit is 10-year equivalents, position size itself, or some other fraction or "haircut" applied to a position, the validity of any such accounting approach is questionable and thus has

Table 4.1. Global equity portfolio deviations from benchmark weights.[a]

United States	–7.0	Italy	2.0
United Kingdom	–6.0	Germany	2.0
Japan	4.5	Brazil	2.0
Canada	4.0	Thailand	2.0
France	–3.5	Australia	–1.5
Switzerland	2.5	South Africa	–1.0

[a]Annualized tracking error 1.82%.

Table 4.2. Proprietary trading portfolio Eurodollar positions.[a]

Contract	Type	Strike	Position (# of Contracts)
September '96	Future		–4,500
	Call Option	94.00	8,000
December '96	Future		2,200
	Call Option	94.00	–6,600

[a]Valuation date June 14, 1996.

very limited value. Nonetheless, at Goldman Sachs we continue to create "risk" reports for fixed income positions based on such an accounting approach. Table 4.3 is an example of a report for the aggregate of our European fixed income positions.[8] When you try to understand the sources of risk in such a report, however, the need for something better becomes clear.

An example of an accounting approach to managing the risk of the proprietary derivative portfolio appears in table 4.4, where we show what are often called the "Greek letter" exposures—the sensitivities to various market risk factors. In this case, the delta is the net interest rate sensitivity measured in thousands of dollars per basis point; the vega is the net volatility sensitivity in thousands of dollars per basis point change in volatility; and the gamma is the change in delta with respect to a 1 bp change in interest rates—that is, the degree of nonlinearity in the position. While these bits of information are important to understanding and managing the position, they do not provide an adequate basis for risk management.[9] In this case, the directions of the various risks are revealed: There is a small short interest rate bias, which arises from the September contract delta; a flattening bias because the September contract delta is short while the December contract delta is long; a net short volatility position arising from the short position in December calls; a volatility spread position; and a net long gamma position because of the much larger gamma arising from the September contract. However, it is certainly not clear from this table alone what is the magnitude of the risks or even which are the major risks of the positions.

Over the past several years, the accounting approach to risk management has been largely supplanted by the use of "stress" tests. Stress tests are the output of an exercise in which positions are revalued in scenarios where the market risk factors sustain large moves. There is no doubt that the use of stress tests is an improvement over a situation of not knowing what might happen in such circumstances. However, as we discuss, there are important limitations in stress testing that need to be recognized. In figures 4.3 through 4.6, we show examples of stress tests for the proprietary portfolio of Eurodollar futures and options. In this example, the nonlinearities of the payoffs from the options figure prominently and are revealed in the stress tests.

Table 4.3. Europe Fixed Income U.S. 10-year equivalent summary.[a]

Asset	Total	Ger	Ndl	Fra	Bgm	UK	US	Ita	Spn	Dnk	Sw	Fin	Nor	Swz	Ecu	Jap	Hkg	SAf
Currency	–14.5	14.3	0	4.0	–35.5	0		–27	–2.5	0	0	0.3	13.3	–4.0	35.8	0	–12.5	0
< 1 year	95.5	23.0	0	6.3	0	6.3	0.5	38.3	2.5	0	0	18.8	0	0	0	0	0	0
1 to 5 year	116.3	–4.3	3.8	40.5	–2.5	49.8	16.8	44.8	4.3	–47.5	11.0	1.3	0	0	–1.8	0	0	0
5 to 12 year	–329.3	–224.3	2.0	–100.5	2.3	–11.0	6.3	100.3	–117.3	32.8	–12.8	–16.0	0	0	10.5	–2.5	0	1.0
12 to 30 year	318.0	226.8	0.5	11.5	0	84.0	–12.0	7.8	0	0	0	0	0	0	0	0	0	0
Swaps	–18.3	–18.3	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Total FI	182.5	3.0	6.3	–42.3	–0.3	129.3	11.3	191.0	–110.5	–14.8	–1.8	4.0	0	0	8.8	–2.5	0	1.0

[a]Currency positions are shown in U.S. dollars, while other positions are shown in the dollar market value of the position. However, we have also scaled all the numbers by an arbitrary constant so as not to reveal the actual absolute sizes of the positions that were taken.

Table 4.4. Proprietary trading portfolio Eurodollar positions Greek letter exposures.[a]

Contract	Type	Strike	# of Contracts	Delta	Vega	Gamma
September '96	Future		–4,500	112.50	0.00	0.00
	Call Option	94.00	8,000	–131.25	2.15	1.45
December '96	Future		2,200	–55.00	0.00	0.00
	Call Option	94.00	–6,600	70.38	–2.80	–0.75
Portfolio Total				–3.38	–0.625	0.70

[a]Thousands of dollars per basis point.

We begin by showing three different stress tests that highlight different dimensions of risk. In each case, the two horizontal axes represent separate risk factors, while the vertical axis shows the profit or loss, in millions of dollars, in that scenario. Figure 4.3 shows the profits and losses of the portfolio as a function of a parallel shift in interest rates on one axis and a parallel shift in the volatilities of the option positions on the other. In this chart, which focuses on the two basic risk factors affecting these securities, the outcomes look quite good. There is considerable upside and very little potential for loss. However, as shown in figures 4.4 and 4.5, there are other risks. Figure 4.4 shows the profits and losses as a function of separate interest rate shifts in each of the Eurodollar contracts, the September contract on one horizontal axis and the December contract on the other. Figure 4.5

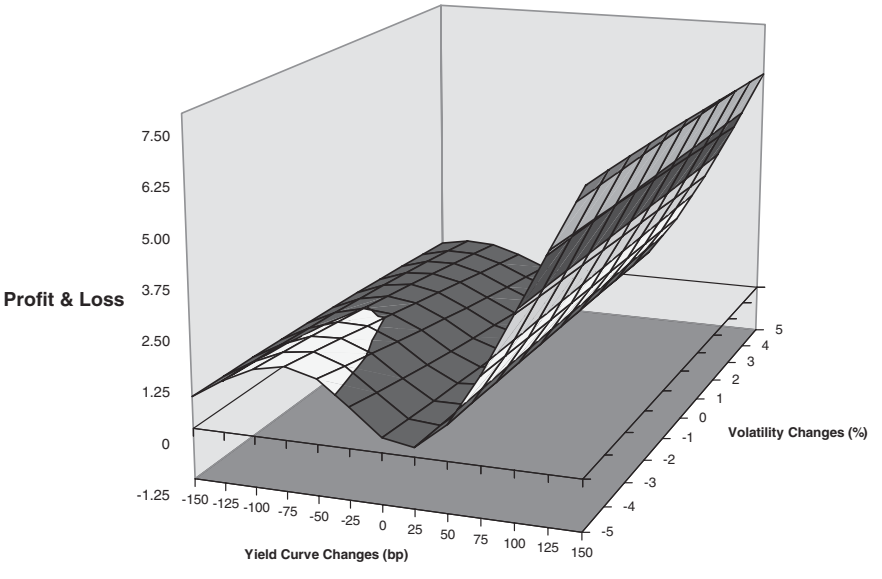


Figure 4.3. Proprietary Portfolio Stress Test Results: Parallel Yield and Volatility Changes

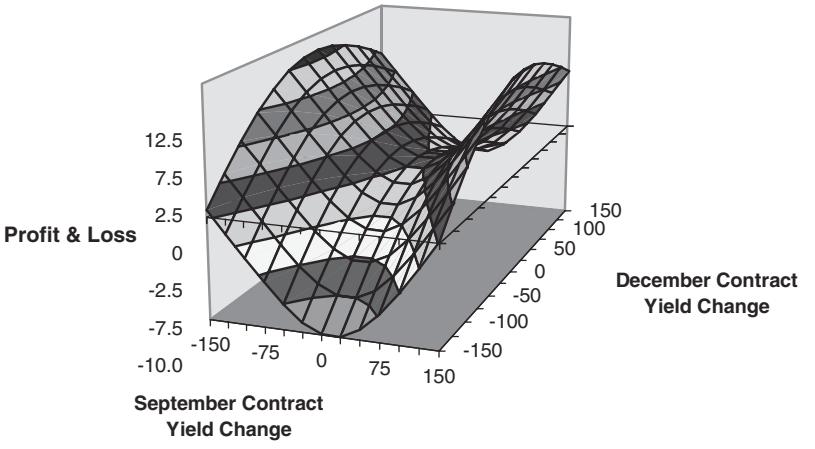


Figure 4.4. Proprietary Portfolio Stress Test Results: Yield Changes for September and December

shows the dimension of risk that these positions are designed to highlight—namely, profits and losses as a function of separate volatility shifts for the September and December contracts.

Stress tests clearly provide useful information about the profile of gains and losses. What are their important limitations? First, it is not always clear which dimensions of risk need to be considered. This problem is obvious in comparing figures 4.3–4.5. Also, stress tests do not reveal the relative probabilities of different events. A position with negative gamma—one that loses money in large moves in either direction—will look very bad in extreme scenarios but will generally look very attractive when only local moves are considered. The positions shown here, which have positive

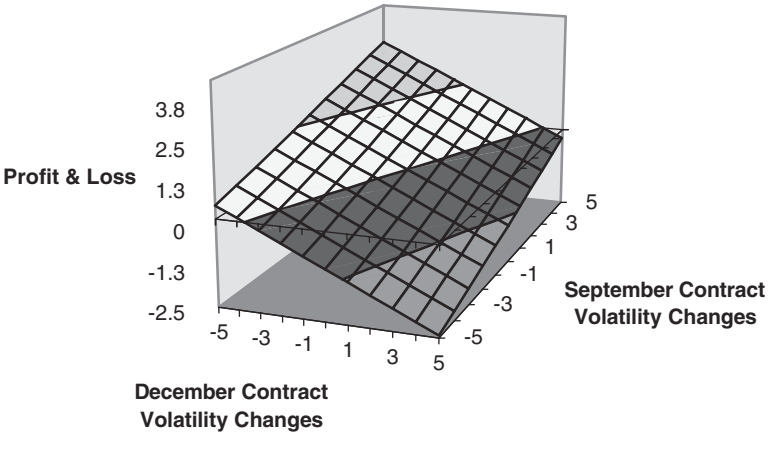


Figure 4.5. Proprietary Portfolio Stress Test Results: Volatility Changes for September and December

gamma, have the opposite property. In any case, the shrewd trader or salesperson can tailor his positions to look attractive relative to any particular set of scenarios or, given the opportunity, can find a set of scenarios for which a particular set of positions looks attractive. For example, the only differences between figure 4.3 and figure 4.6 are the scale of the axes and the direction of the view. Figure 4.6, which covers the most likely outcomes for a given day, does not look nearly as attractive as figure 4.3, which covers a much wider region and focuses attention on the positive outcomes.

Moreover, in complex portfolios, there are many sets of scenarios to look at; in fact, it may be virtually impossible to know which among many risk factors need to be considered. Furthermore, even if an exhaustive set of scenarios is considered, how does the trader or risk manager know how to take into account the risk reduction resulting from diversification of the risk factors? Thus, while stress testing is a useful tool, it often leaves large gaps in the understanding of risk.

Today, such approaches are quickly being supplemented by measures such as volatility and VAR, which explicitly recognize that risk is a statistical concept.[10] These statistical approaches make assumptions about the distribution of the outcomes of risk factors, map the points in that distribution into different valuations for the portfolio, and then measure the uncertainty in the distribution of portfolio gains and losses. There are a number of different approaches to computing these measures; they differ primarily in the nature of the assumptions about the distribution of outcomes. For example, one approach is to use the historical distribution; another is to use a parameterized distribution estimated from historical or market data. In addition, approaches may differ in the degree of accuracy achieved through different degrees of aggregation or—especially with respect to derivatives—in the degree of accuracy achieved through repricing versus forming linear or nonlinear approximations to the payoff functions of the derivatives.

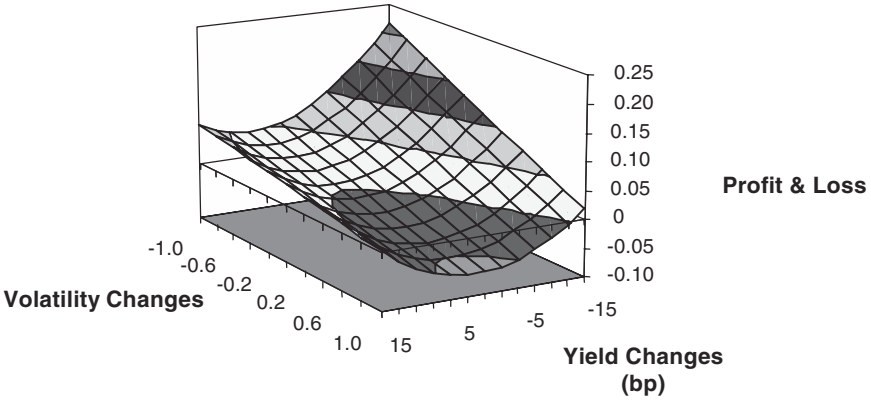


Figure 4.6. Proprietary Trading Stress Test Results: Local Changes in Yield and Volatility

Different statistical approaches have different strengths that may make them more or less appropriate in different contexts. We will illustrate two approaches in our examples—historical simulations and the use of a covariance matrix—but it is not the purpose of this chapter to try to defend one or another statistical approach. Much research is continuing in this area, and, while we hope and expect that practitioners and researchers will eventually develop a common understanding of what constitutes best practice, we anticipate that one component of that understanding will be the recognition that no one approach is adequate for all purposes and that it is best to monitor risk using several different approaches. In addition to the two approaches illustrated in this chapter, we also use Monte Carlo simulations extensively in risk management exercises at Goldman Sachs.

In the context of the global equity portfolio that we are using as an example, we will construct a covariance matrix of returns of the asset classes from which we can form an estimate of the tracking error—the volatility of the portfolio of deviations from benchmark weights.[11] As applied to the positions in the equity portfolio shown in table 4.1, the volatility of the portfolio has an annualized tracking error of 1.82% relative to the benchmark.[12] Since the risks in this portfolio are linear, the use of a VaR does not add much, except to allow comparability with other VaR analyses. In order to do such a comparison, we divide the annualized volatility by the square root of 252 (business days in a year) to get a daily volatility of 11.5 basis points, and then we appeal to our four-standard-deviation rule of thumb to come up with a once-per-year daily VaR of 0.46%. As we shall see later, the positions themselves (which are ordered in table 4.1 according to the absolute size of the deviation) reveal little about the actual sources of risk in the portfolio.

In the context of the proprietary trading portfolio, we use historical price and volatility changes rather than a covariance matrix to calculate the volatility and VaR of the positions.[13] In this case, the daily volatility is $58,000 and the once-per-year daily VaR is $127,000. The fact that the VaR is a smaller multiple of the volatility (only 2.2 rather than our rule of thumb of 4) reflects the nonlinearity, in particular the positive gamma, of the positions. The use of historical data has two distinct advantages in this context. First, it does not require any assumptions about the distribution that might smooth out the large jumps or other unusual moves that could create the biggest risks. Second, a very practical benefit is that the historical analysis provides a list of the "worst days" losses of this portfolio. That is, rather than simply claiming that a portfolio of positions is risky, the analysis provides a list of examples of days on which the portfolio would have lost such and such amounts. In table 4.5, we show the five best and five worst days for the proprietary portfolio.[14]

The results in table 4.5 show clearly that the portfolio gains from large moves in rates in either direction—the benefit of positive gamma—and that the risk is dominated by the short volatility contributed by the December option and the volatility spread contributed by the September option.

Table 4.5. Best five and worst five days: analysis of historical simulation of the proprietary trading portfolio.[a]

5 Best Days	Profits & Losses	Sept. Future Profit	Basis Points Change in Yield	Sept. Option Profit	% Change in Implied Vol.	Dec. Future Profit	Basis Points Change in Yield	Dec. Option Profit	Change in Implied Vol.
09/04/92	0.43	–3.36	–28	4.65	9.6	1.57	–30	–2.43	3.8
12/20/91	0.43	–2.94	–27	3.92	–8.0	1.22	–26	–1.77	–1.2
07/02/92	0.42	–3.51	–27	4.73	–1.7	1.78	–31	–2.58	–4.6
07/06/95	0.34	–2.85	–18	3.71	–8.5	1.40	–25	–1.92	–4.9
03/08/96	0.30	4.92	29	–3.60	11.5	–2.86	44	1.84	20.4
5 Worst Days									
04/28/94	–0.15	1.67	9	–1.72	2.9	–1.02	15	0.92	2.9
02/09/93	–0.14	0.64	3	–0.74	–0.5	–0.44	6	0.39	0.5
12/07/92	–0.13	–0.56	–3	0.71	1.5	0.54	–6	–0.82	1.3
05/31/95	–0.13	0.09	1	–0.14	–2.9	–0.00	1	–0.07	–0.6
03/02/94	–0.12	0.06	1	–0.01	0.5	–0.08	1	–0.08	2.2

[a]Profits and losses in millions of U.S. dollars.

Nonetheless, the downside is significantly limited relative to the upside. A histogram of the profits and losses, shown in figure 4.7, reflects this as well, although the distribution of outcomes is perhaps surprisingly symmetric, given the appearance of the stress test in figure 4.3. The symmetry reflects the fact that even for this highly nonlinear portfolio, the payoffs are close to linear for most daily moves.

4. RISK MANAGEMENT VERSUS RISK MONITORING

Volatility and VaR characterize, in slightly different ways, the degree of dispersion in the distribution of gains and losses, and therefore they are useful for monitoring risk. They do not, however, provide much guidance for risk management. To manage risk, you have to understand what the sources of risk are in the portfolio and what trades will provide effective ways to reduce risk. Thus, risk management requires additional analysis—in particular, a decomposition of risk, an ability to find potential hedges, and an ability to find simple representations for complex portfolios.

The risk of positions is not additive. This is fortunate for investors because it reduces overall risk. However, it is unfortunate for risk managers because it complicates their job. Since the returns of different assets are more or less correlated, the risk of a portfolio of positions in different assets is always less than the sum of the individual risks. This reduction in risk is the benefit of diversification, and, in a well-balanced portfolio, the risk reduction can be significant. Because of this nonadditivity of risk, it is not easy to create a simple decomposition of risk. However, as we show below, although the total risks of the portfolio is not the sum of the risks of individual positions, it is in fact the sum of the marginal impacts on portfolio risk from small percentage increases in each of the portfolio positions. Such a marginal analysis thus provides the basis for a risk decomposition.

Before we discuss this risk decomposition, however, consider figure 4.8. This is a useful diagram for understanding how, in a simple linear context, the volatilities of two positions combine to form the volatility of the portfolio. The diagram illustrates that the volatilities of the two positions combine to form portfolio volatility in the same way that the lengths of two

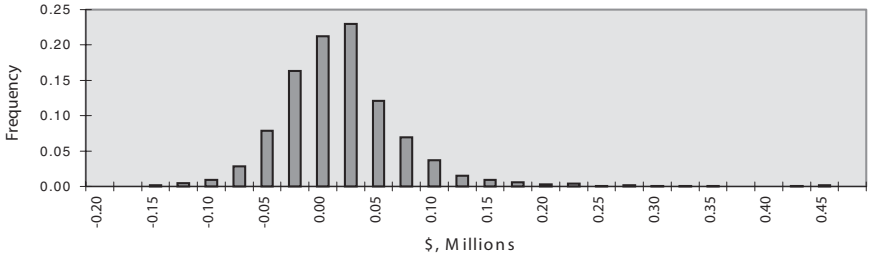


Figure 4.7. Profits and Losses from Proprietary Trading Portfolio

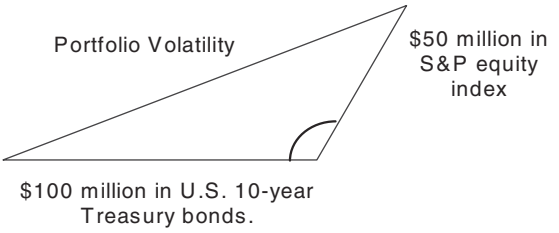


Figure 4.8. Combining Volatilities. The volatilities of two positions sum to create portfolio volatility in a manner analogous to the way in which the lengths of two sides of a triangle create the length of the third side. Correlation plays the role of the angle between the two sides

sides of a triangle combine to form the length of the third side. In the case of portfolio risk, the correlation between the returns of the two assets plays the same role as the angle between the two sides plays in the case of the triangle. Correlations range between –1 and +1 and map into angles ranging from 0 to 180 degrees. The case of no correlation corresponds to a 90-degree angle. Positive correlations correspond to angles between 90 and 180 degrees, and negative correlations correspond to angles between 0 and 90 degrees.[15]

In figure 4.8, the lower side represents the volatility of a $100 million position in U.S. 10-year Treasury bonds. The shorter side represents a $50 million (equity) position in the S&P 500 Index. The annualized return volatilities of the bonds and stocks are 10.2% and 13.1%, respectively, leading to position volatilities of $10.2 and $6.5 million. The correlation between the two assets is 0.47, and the volatility of the total portfolio is $14.5 million, only 86% of the sum of the two volatilities. We describe this as a 14% reduction in risk resulting from diversification. It is obvious in this example that risks are not additive and that the decomposition of portfolio risk requires attention to this fact.

In figure 4.9, we modify the previous diagram and its interpretation slightly to illustrate the common case where we are interested in the impact on portfolio risk of a trade that is small relative to the size of the portfolio. The same diagram applies, except that now we use one side of the triangle to represent the tracking error of an investment of $1 billion in our example equity portfolio, while the second side represents a $10 million trade out of Japan into various other asset classes.[16] The combined port-

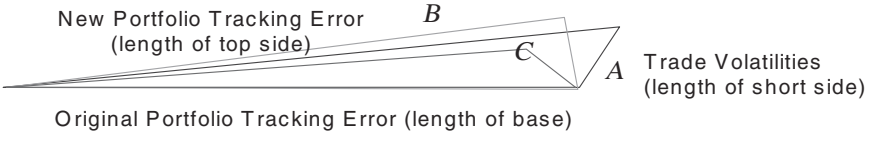


Figure 4.9. Portfolio Volatility Example

folio risk is represented by the length of the third side. As is clear in the diagram, as long as the risk in the trade is small relative to the risk of the original portfolio, the correlation of the returns of the trade with those of the original portfolio is much more important than the volatility of that position itself in determining the impact on portfolio risk.

In this example, we assume a sale of Japanese equity and look at the impact of several different potential purchases: Brazilian, Thai, and U.S. equities. The sale of Japanese equity represents a reduction in the largest single deviation of the portfolio from its benchmark. Japanese equity is also a fairly volatile asset, so the result of such a sale in which the proceeds are invested in cash is not surprising—the tracking error of the portfolio is reduced from 1.82% to 1.70%. If the proceeds are used to purchase Brazilian equity, which represents an increase in the deviation from the benchmark in a highly volatile asset, then without taking into account the portfolio effects, it is not clear which impact will dominate: the reduction in the Japanese exposure or the increase in the Brazilian exposure. In fact, the higher volatility of the Brazilian exposure leads to an increase in tracking error, up to 1.98%. Alternatively, a sale of Japanese equity coupled with a purchase of Thai equity, which in the original portfolio had the same exposure as Brazil, leads to a slight reduction in overall tracking error, down to 1.81%. Finally, the largest reduction in risk would be generated by investing the proceeds of the Japanese equity sale in U.S. equity, which represents a significant underweight position in the original portfolio. Such a trade is very negatively correlated with the portfolio risk and reduces the tracking error to 1.67%.

In figure 4.9, in each case the triangle represents the effect of selling Japanese equity and buying equity of another country. Triangle A represents increasing exposure to Brazil, triangle B represents increasing exposure to Thailand, and triangle C represents reducing the underweight position in the United States. In each case, the volatility of the trade, represented by the length of the short side of the triangle, is not that different; but what clearly matters in determining the effect on the portfolio volatility (represented by the length of the third side) is the correlation of that trade's returns with the original portfolio tracking error, represented by the angle between the short side and the base of the triangle.

Another interesting analysis results from starting with an existing portfolio and considering how its risk will be affected by adding different amounts of one particular trade. We call this simple graph the "Trade Risk Profile" and use it to better understand the effects of correlations on portfolio risk. As an example of this type of analysis, we start from the triangle diagram in figure 4.9. Once the original portfolio and the particular trade are chosen (in this case we focus on the trade out of Japanese equity into U.S. equity), the length of the base of the triangle is fixed, as well as the angle between it and the side representing the particular trade. The relationship we are interested in is the one between the length of the second side, representing the size of the position in the trade, and the length of

the third side, representing the effect on portfolio volatility. This relationship always has the same shape, which is shown in figure 4.10. Of particular interest will be both the size of the second side of the triangle (which we show in figure 4.10 on the horizontal axis) at the point where the length of the third side (which we show in figure 4.10 on the vertical axis) is minimized and the length of the third side at that minimum.

The triangle diagrams can also be very helpful in understanding how correlations affect the ability of hedges to reduce portfolio risk. In particular, many traders and portfolio managers may not be aware of how sensitive risk reduction is to the degree of correlation between the returns of the positions being hedged and the hedging instruments. As we will show, for example, a correlation of 0.8, which is a high degree of correlation, creates a risk reduction potential of only 40%. The risk reduction potential rises to 56% with a correlation of 0.9, 69% for a correlation of 0.95, 86% for a correlation of 0.99, and 96% for a correlation of 0.999. In figure 4.11, we illustrate what we mean by a risk reduction potential. In figure 4.12, we show the relationship between correlation, the angle between the two sides of a triangle, and the risk reduction potential.

Figure 4.11 shows a risk triangle with the base representing a long position of $100 million worth of Italian lira exposure. We are interested in what fraction of the risk, which in this case is a daily volatility of $408,000, we can hedge. The second side of the triangle represents the risk of a position in deutsche marks that, from the point of view of a U.S. dollar-based investor at the time this report was being written, had a correlation of 0.24 with the lira. The correlation of –0.24 between lire and short deutsche marks implies an angle of 76 degrees between the two sides of the triangle. The risk-minimizing hedge is represented by a point on the second side that minimizes the distance to the opposite side of the base—that is, that minimizes the length of the third side of the triangle. That risk-minimizing position in this case corresponds to a short deutsche mark position of $23 million. The resulting portfolio has a daily volatility of $396,210. This

Figure 4.10. Trade Risk Profile

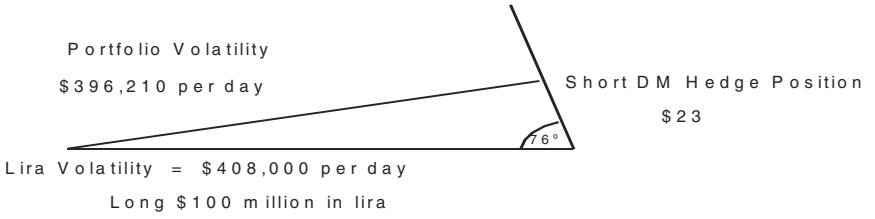


Figure 4.11. Risk Reduction Potential

represents only a 2.9% reduction in risk. Clearly, in order to create a meaningful reduction in portfolio risk, we need to find hedging instruments that have a much higher degree of correlation with the risk of the portfolio.

This relationship between correlation and risk reduction potential appears graphically in figure 4.12. Here we show the different correlations, the angles to which they correspond, and the percentage risk reduction potential. We highlight the point corresponding to a correlation of 0.5, which leads to an angle of 60 degrees and creates a risk reduction potential of only 13.4%. As correlations increase, the angle gets smaller and the potential for risk reduction increases, but only quite slowly.

We next use the triangle diagram to illustrate how we will decompose risk. In figure 4.13, the length of the third side of the triangle represents the portfolio volatility that we wish to decompose. In this example, the two assets are currency positions of a dollar-based investor: long $100 million in Swiss francs and short $100 million in deutsche marks. If we were to focus on the volatilities of the two assets alone, we would attribute just over one-half of the risk to francs and a little under one-half to deutsche marks (the annualized volatilities of these currencies are 12.9% and 11.6%, respectively). Such an attribution would be misleading, however. The risk of the joint position is only 5.0%, and the dominant risk in the portfolio is clearly

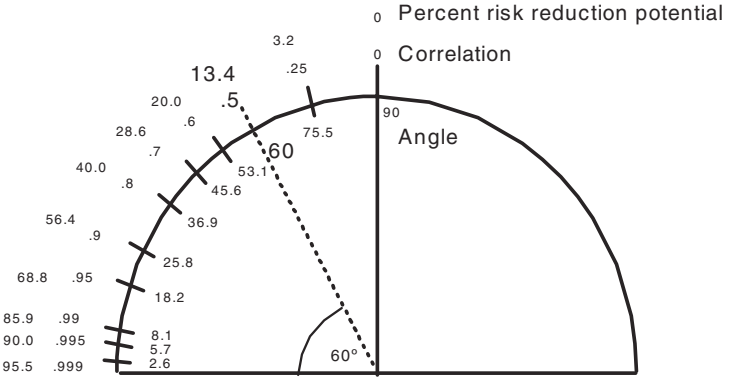


Figure 4.12. Correlation and Risk Reduction Potential

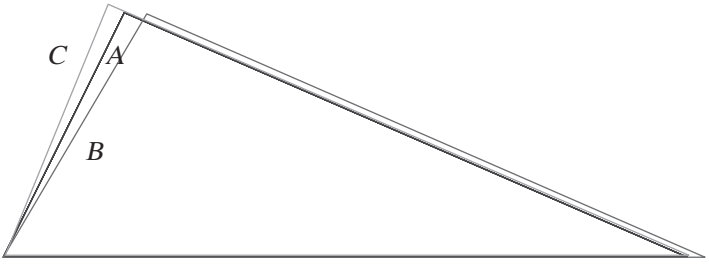


Figure 4.13. Decomposing Portfolio Volatility

the long position in Swiss francs; the returns of the deutsche mark position are highly (–0.92) negatively correlated with those of the franc and are offsetting much of its risk. Rather than focus on the volatilities, which correspond to the lengths of the two sides of the triangle and do not sum to create portfolio volatility, we will concentrate on what happens to the volatility of the portfolio as we make small percentage increases in each of the positions. This is illustrated in figure 4.13 as changes in the length of the third side as we make small percentage increases in the lengths of each of the two other sides of the triangle. What we see is that the small increase in the length of the franc side creates an increase in the length of the third side, whereas an increase in the length of the deutsche mark side reduces the length of the third side. For a 1% increase in the sizes of the positions, the percentage changes in risk are an increase of 1.34% and a reduction of 0.08%, respectively, for the Swiss franc and the deutsche mark. These impacts represent a useful decomposition of the portfolio volatility—the franc position is creating the portfolio risk, and the deutsche mark position is reducing it. As we show below, this idea of looking at the marginal impact on risk of percentage changes in position sizes generalizes as a risk decomposition to the case of VaR, as well.[17]

In figure 4.13, triangle A represents a portfolio of long $100 million of Swiss francs, the base side, and short $100 million of deutsche marks, the upper right side. The effects of marginal increases in the size of the positions are shown in triangle B for francs and in triangle C for deutsche marks. At the margin, an increase in the franc position increases portfolio risk (represented by the length of the third side), whereas an increase in the deutsche mark position has no discernible effect (in fact, it is a very slight reduction).

Finally, in figure 4.14, we show what happens when we increase the size of the deutsche mark position to $200 million. Now the marginal impact of increasing the length of the side representing deutsche marks increases the length of the side representing portfolio risk, whereas increasing the length of the side representing Swiss francs has the opposite effect. The revised position has a risk of 12.4%, and the percentage change in risk with respect to a 1% increase in the size of the positions is a decrease of 0.69% for the Swiss franc and an increase of 1.73% for the deutsche mark. In other

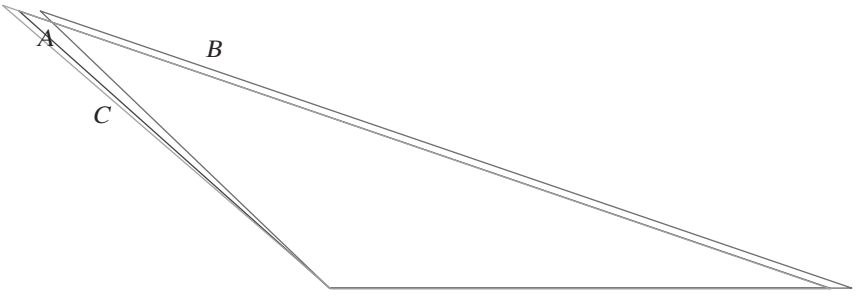


Figure 4.14. Decomposing Risk: Increased DM Position

words, the increased size of the deutsche mark position has made it the dominant risk in the portfolio.

Using the same letters as before to represent the marginal impacts, in figure 4.14 we see that having increased the size of the deutsche mark position in the portfolio, the marginal changes in each of the two positions have different effects: Now an increase in francs (triangle B) reduces risk, while an increase in deutsche marks (triangle C) increases risk.

Both volatility and VAR have a feature in common that allows us to form this useful marginal risk decomposition: They are both linear in position size. In other words, if all positions are increased by a common factor, then risk, as measured by both volatility and VAR, is also increased by that same factor.[18]

Because of this linear nature of the two risk measures, marginal impacts provide a decomposition of risk. Scaling all positions by a common factor increases the volatility or VAR by that same factor. In other words, $R(kx) = k*R(x)$, where k is a factor greater than zero, x is the n-dimensional vector of positions, and R is the risk function. Taking the derivative with respect to k of both sides of this equation reveals that

$$R_1(x) * x_1 + R_2(x) * x_2 + \ldots R_n(x) * x_n = R(x) \quad (1)$$

where $R_i(x)$ is the partial derivative of R with respect to the ith position and x_i is the ith position. Loosely speaking, the product $R_i(x) * x_i$, which is the marginal rate of change in risk per unit change in the position (at the current position size) times the position size itself, can be thought of as the rate of change in risk with respect to a small percentage change in the size of the position. Thus, the risk itself, R, is the sum of the rates of change in risk with respect to percentage changes in each of the positions. This is a useful decomposition because it highlights the positions to which the risk of the portfolio is most sensitive.

We define

$$R_i(x) * x_i * 100 \,/\, R(x) \quad (2)$$

to be the percentage contribution to portfolio risk of the ith position.

There is one important limitation to this risk decomposition—it is a marginal analysis. If we find, for example, that in the decomposition a particular position accounts for half the risk, that implies that a small percentage increase in that position will increase risk as much as the sum of a similar percentage increase in all other positions. It does not imply, however, that eliminating that position entirely will reduce risk by half. As might be expected, as the size of the position of a contributor to risk is reduced, the marginal contribution of that position to risk will be reduced as well.

Not all positions need provide a positive contribution to risk. In fact, as was clear in the risk triangle diagrams (figures 4.8, 4.9, 4.11, 4.13, and 4.14), any trade whose returns are negatively correlated with the returns of the portfolio (so that the angle between the first and second sides of the triangle is less than 90 degrees) will, at least at the margin, reduce risk. The interpretation of a negative contribution to risk is very straightforward—increasing the size of such a position will, at the margin, reduce risk.[19] Eventually, as the size of such a position is increased, it will contribute enough to the returns of the portfolio that its returns are no longer negatively correlated with the returns of the portfolio but rather are uncorrelated. At that point, a marginal increase in the position does not affect the risk of the portfolio.

Indeed, this point is a very interesting position size. The risk contribution is zero, and the position is such that it is the risk-minimizing position for that asset class, holding all other positions at their current size.[20] We call such a position a "best hedge."

Using the notation above:

$$R_i(x) = 0 \text{ implies that } x_i \text{ is the best hedge position, holding all other positions fixed.} \quad (3)$$

Let us return to the "Trade Risk Profile" graph (see figure 4.10) and consider the relationship between it and the contribution to portfolio risk of an individual position. In figure 4.15, we again show an analysis of the risk for our global equity portfolio as a function of the amount moved from Japanese equity into U.S. equity. We also show, for the same-sized trade, the contribution of the trade (treated as a separate position) to portfolio risk. The risk contribution of a trade is really just a simple generalization of the risk contribution of any other asset class. In either case, we look at the change in risk as a function of a small percentage increase in the size of the position. As the chart shows, the trade will (almost) always have two positions where it contributes zero to portfolio risk. The first is where the trade size is zero; the second is where the risk of the portfolio is minimized with respect to the size of the trade—that is, the point where the trade represents a "best hedge" position. At all points between these two, the trade will have a negative impact on portfolio volatility, and outside this range it will have a positive impact—increasingly so as the size of the trade gets big relative to the other risks in the portfolio. (The only time there will not be two positions of zero contribution is when the trade is uncorrelated with

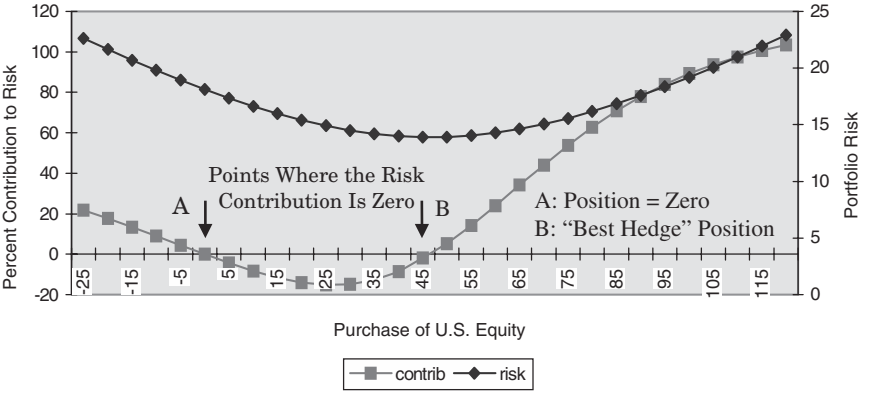


Figure 4.15. Trade Risk Profile and Contribution to Portfolio Risk

the portfolio, in which case the only zero contribution position is the null position.)

Earlier, we analyzed the impact of small changes in position or trade size on the portfolio risk to create a portfolio risk decomposition. We can generalize this concept still further and consider the decomposition of risk in a large portfolio into that contributed by smaller components of the portfolio. For example, we find it useful to decompose the firmwide risk of Goldman Sachs into that contributed by its divisions, the divisional risk into that contributed by business units, and so on. The contribution to firmwide risk can vary considerably from the absolute level of risk. The latter clearly is a function of the correlation of the risks of different businesses. If the dominant risk faced by the firm is long global interest rates,

Table 4.6. Firmwide risk decomposition.

	VaR[a]	Percentage Contribution to Firmwide Risk
Firmwide Total	11.8	100.0
Fixed Income	6.7	46.0
Asia	2.8	5.8
Emerging Markets	1.8	4.1
Europe	5.6	32.4
Europe Governments	4.8	24.7
Gilts	1.1	2.9
Trader #1	0.9	1.9
Trader #2	0.5	1.1
North America	2.3	3.7

[a]We have scaled the numbers to arbitrary units in order to avoid revealing the actual level of risk taken on this date.

then a large short in one particular business may have significant risk in itself but contribute negatively to firmwide risk. In table 4.6, we show a simple example of this decomposition. On this particular day, the VaR for North America Fixed Income was over 40% of that of Europe Fixed Income, while the contribution to firmwide risk was less than 12%. Mathematically, this decomposition follows immediately by considering the components of the overall portfolio in the same way that we consider trades—that is, we take the derivative of risk with respect to percentage changes in their position sizes. Similarly, a pension fund or mutual fund may want to decompose its overall portfolio risk into the amounts contributed by each of its components. Those components could represent geographic regions, countries, asset types, or portfolio managers.

5. PORTFOLIO ANALYSIS TOOLS

In this section, we describe a number of portfolio analysis tools that we have developed and found useful in managing risk at Goldman Sachs. We show how they apply to our example portfolios and discuss how we try to use them.

"Hot Spots" Reveal Which Positions in the Portfolio Have the Biggest Impact on Risk

Since the marginal impacts on risk sum to the risk of the portfolio, it is natural to express the risk decomposition in terms of the percentages of risk accounted for by each position or asset class. When we do this for a portfolio—and sort the resulting decomposition according to the asset classes and countries that create the largest contributions to risk—we have what we at Goldman Sachs call a "Hot Spots" report. We show reports for our example portfolios in tables 4.7–4.10. In our Hot Spots reports, we generally show two numbers for each asset class. The first is the percentage of marginal risk accounted for by the position in that asset class. The second is the position size itself. The dark shading represents all positions whose marginal contribution is more than 5% of the total, and the light shading indicates all positions whose marginal contribution is to reduce risk by more than 5% of the total.

The Hot Spots report for the equity portfolio (table 4.7) clearly highlights which positions merit the risk manager's attention. The combination of Japan and Brazil contributes more than half of the total portfolio's risk decomposition. This is surprising because their deviations from benchmark weights do not stand out as obvious candidates for attention. In contrast, the two countries with the biggest absolute deviations, the United States and the United Kingdom, contribute a total of less than 20% to the risk decomposition. Canada, which has a large absolute deviation from the benchmark, contributes virtually no risk. Given the size of the underweight position in the United States, the overweight position in Canada is close to a best hedge position.

Table 4.7. Hot Spots Report for the global equity portfolio.[a]

Country	Percentage Contribution to Risk	Position
Japan	30.7	4.5
Brazil	23.6	2.0
United States	13.0	–7.0
Thailand	10.4	2.0
Italy	6.9	2.0
United Kingdom	6.6	–6.0
Germany	3.7	2.0
France	3.3	–3.5
Switzerland	2.7	2.5
Canada	0.2	4.0
South Africa	–0.8	–1.0
Australia	–2.1	–1.5

[a]Annualized Tracking Error 1.82%.

In the case of the proprietary portfolio, the risk decomposition can be applied to the four individual securities or to the two different contract expiration dates.[21] In the first case, the analysis shows that the December option contributes the dominant risk in the portfolio. The short position in the December Eurodollar option contributes 131% of the risk. Clearly, as seen in table 4.5 (the best and worst days analysis), the risk in this portfolio, limited as it is, is largely that implied volatility on the December contract will rise. Of course, when volatility in the December contract rises, it is likely that volatility in the September contract will rise also, offsetting some of the losses; and indeed, the September option shows a negative contribution to risk of 15% (table 4.8).

Table 4.8. Hot Spots Report for the proprietary trading portfolio.[a]

Contract	Type	Percent Contribution to Risk	Position
September		–5	
	Future	9	–45
	Option	–15	80
December		105	
	Future	–25	22
	Option	131	–66

[a]Daily once-per-year VAR $127,000.

Our example consists of a portfolio long gamma and short vega. If we flip the signs of all the positions, then we have a portfolio with the classic profile that risk managers hate. It makes small amounts of money most of the time and, on rare occasion, blows up with large losses. The five "best" days become the five "worst" days. If we compute the risk decomposition for this portfolio (table 4.9), the dominant risk, now short gamma, now comes from the September option, which contributes 290% of the risk. The December option reduces risk by 130%, and the futures offset the additional component.

We constructed these examples in part to be easy to understand. But because of this simplicity, they do not adequately illustrate the main benefit of a risk decomposition, which is to highlight, in large, complex portfolios, the positions that require attention. To illustrate this type of benefit, we show in table 4.10 the percentage contributions from the Hot Spots report for the Europe Fixed Income portfolio. Although the positive contributors to risk are not necessarily long positions, we can see from table 4.3 that in this example that is generally the case.

Once we have identified the hot spots in a portfolio, the next step in reducing risk is to decide how to change those positions. A complete solution to this problem requires a process of portfolio optimization. However, because of the existence of transaction costs, traders and portfolio managers often update portfolios incrementally. As discussed above, it is frequently useful in this context to visualize how portfolio risk, either VAR or volatility, varies as a function of the position size in a given asset or trade.

For the equity portfolio, we found that the hot spots included Japan and Brazil. We first focus on the Trade Risk Profile for a trade out of Japanese equity into U.S. equity, which was shown in figures 4.10 and 4.15. This trade risk analysis is not particularly surprising. The risk rises as the Japanese equity position increases away from the benchmark weight—at first slowly, but then with an increasing slope as this becomes the dominant risk in the portfolio. The risk-minimizing position is at a point where the Japanese equity position is close to the benchmark weight.

Table 4.9. Hot Spots Report for the proprietary trading portfolio with all positions reversed.[a]

Contract	Type	Percent Contribution to Risk	Position
September		273	
	Future	–17	45
	Option	290	–80
December		–173	
	Future	–43	–22
	Option	–130	66

[a]Daily once-per-year VaR $322,000.

Table 4.10. Hot Spots Report for the Europe Fixed Income Portfolio.

	Total	Italy	U.K.	Spain	Germany	Belgium	France	Denmark
Total	100	78	25	–12	9	5	–4	–1
10-year future	–67	9	–16	0	–62	0	1	0
5-year bond	25	19	17	0	–22	1	6	1
7-year bond	60	7	11	0	38	0	–2	5
10-year bond	8	24	6	–12	4	0	–11	0
30-year bond	34	3	0	0	31	0	0	0
5-year future	26	0	0	0	9	0	0	0
2-year bond	–8	0	–10	0	1	0	1	0
20-year bond	15	0	14	0	0	0	1	0
Currency	–3	–3	0	0	–2	5	0	0
3-year bond	–4	2	1	0	1	–1	0	–7
1-year LIBOR	13	11	0	0	1	0	0	0
4-year bond	–3	0	2	0	–5	0	0	0
2nd Euro contract	4	3	1	0	0	0	1	0
Other	0	2	0	0	–2	0	1	0

Consider the appearance of the Trade Risk Profile for two limiting cases: first, the case where only a single trade contributes all of the risk; and second, the case where a particular trade contributes a small and dwindling percentage of the risk. In the first case, the Trade Risk Profile appears as in figure 4.16. Risk is minimized at the zero position point and increases linearly in trade size in either direction. The second case is illustrated in figure 4.17. Here, the portfolio risk is a flat line. By assumption, it is not sensitive to the trade size (although, of course, if there is any sensitivity at all, then for some trade size that is large enough, the portfolio risk will start to be affected). For an asset that contributes a dominating risk to the port-

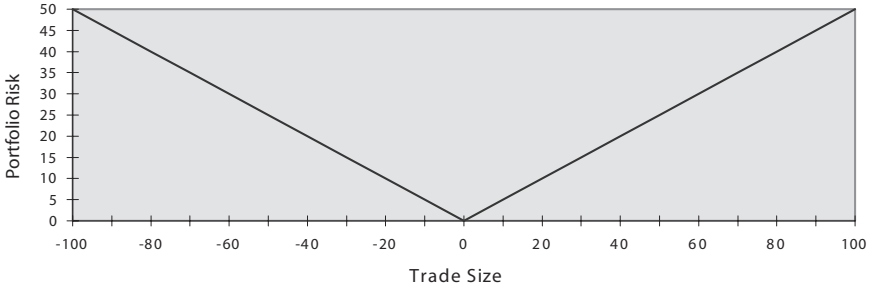


Figure 4.16. Trade Risk Profile: Dominant Risk

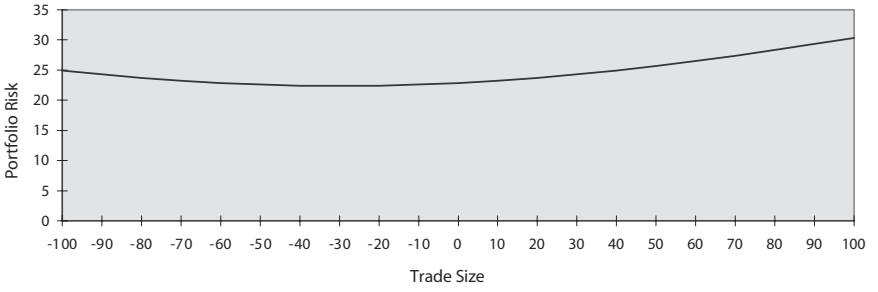


Figure 4.17. Trade Risk Profile: Small Risk Case

folio, the Trade Risk Profile for trades of that asset will look like that of the first limiting case, except that the minimum position will not be zero but will be the point where the trade minimizes the risk. The more interesting Trade Risk Profiles are the ones for trades of assets that are not dominant risks in the portfolio. We focus, for example, on trades of Canadian equity in the context of our global equity portfolio. Perhaps the most common misperception among portfolio managers is that they can minimize their risk from a given asset class by holding a position that is at, or close to, its benchmark weight. As you can see in figure 4.18, the minimum risk position for Canadian equity,[22] 3.8% above the benchmark weight, is close to the current position in the portfolio. As noted above, this is because the Canadian equity is acting as a hedge to the risk contributed by the underweight position in U.S. equity.

The Trade Risk Profile graph is an important one to keep in mind when thinking about portfolio risk. As noted above, it always has the same basic shape. The two aspects of the graph that change are the location of the minimum, which is the trade size that leads to the best hedge position, and the drop between the level of risk at the current position and that at the minimum, which reflects how much risk can be reduced by hedging using that particular asset or trade.

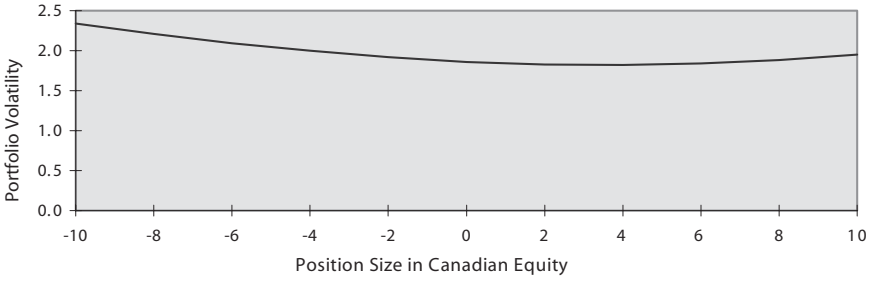


Figure 4.18. Position Risk Profile

The Best Hedges Report

A portfolio manager or trader should have a sense of what the Trade Risk Profile looks like for each of his assets. The slope of the graph at the current position reveals how sensitive risk is to that asset; whether the current position is to the left or the right of the minimum reveals whether purchases increase or decrease risk, and knowing how much higher than the minimum height the current position is reveals how much risk reduction is available through purchases or sales of this position. We have found that a useful report, which we call "Best Hedges," can be formed by measuring the purchase or sale of each individual asset required to reach the risk-minimizing position and then calculating and sorting on the percentage risk reduction available through that purchase or sale. Table 4.11 shows an example of this report for the equity portfolio. This report highlights the importance of the overweight position in Japan in terms of its risk reduction potential far more clearly than do the percentage deviations from the benchmark weights. The current 4.5% overweight in Japan, for example, is small relative to the 7% and 6% underweight positions in the United States and the United Kingdom. But the risk reduction potential is much larger. Conversely, despite a 4% overweight position in Canada, there is virtually no risk reduction potential. This is essentially the risk-minimizing position.

Another example of a Best Hedges report is given in table 4.12 for the Europe Fixed Income Portfolio. This report clearly shows that the biggest opportunity for risk reduction is through selling Italian bonds. Other alternatives are to sell U.K. bonds, Belgian bonds, German bonds, or ECU bonds,

Table 4.11. Best Hedges Report for the global equity portfolio.

Country	Volatility at the Best Hedge	Percentage Reduction in Volatility	Current Position	Trade Required to Reach the Best Hedge Position
Japan	1.48	18.62	4.5	–4.93
Brazil	1.66	8.87	2.0	–1.50
Thailand	1.71	5.95	2.0	–2.30
Italy	1.75	3.73	2.0	–2.18
United States	1.75	3.72	–7.0	3.80
Germany	1.79	1.88	2.0	–2.06
Australia	1.80	1.28	–1.5	–1.89
United Kingdom	1.80	1.22	–6.0	2.10
Switzerland	1.81	0.75	2.5	–1.45
France	1.81	0.57	–3.5	1.18
South Africa	1.82	0.22	–1.0	–0.65
Canada	1.82	0.02	4.0	–0.11

Table 4.12. Best Hedges Report for the Europe Fixed Income Portfolio.[a]

Hedge Asset	VaR at the Best Hedge	Percentage Reduction in VaR	Trade Required to Reach the Best Hedge Position
Italian 30-year bond	3.37	39.42	–138.75
Italian 7-year bond	3.45	37.91	–258.25
Italian 10-year future	3.57	35.81	–180.00
Italian 2-year swap	3.66	34.12	–575.00
U.K. 5-year bond	4.49	19.26	–298.50
U.K. 7-year bond	4.49	19.25	–247.50
Belgian 2-year bond	4.53	18.38	–847.25
U.K. 10-year future	4.55	18.08	–179.25
U.K. 10-year bond	4.56	17.88	–180.50
German 2-year bond	4.64	16.45	–756.75
ECU 7-year bond	4.68	15.73	–263.25

[a]Current VaR is 5.55.

in decreasing order of effectiveness. Notice that selling Belgian bonds is a portfolio hedge despite the fact that, as shown in table 4.3 above, the current position in Belgian bonds is already a net short position.

The Relationship Between Best Hedge and "Market Exposure"

The risk-minimizing, or "Best Hedge," position is also important in that it represents a key reference point for risk management purposes. Kurt Winkelmann and I (Litterman and Winkelmann 1996)[23] have discussed how portfolio managers could use the regression coefficient of portfolio returns on those of a market portfolio to characterize their exposure to "the market." As reported in that paper, at Goldman Sachs we find it convenient in fixed income traders' portfolios to define the market as the 10-year benchmark bond and to characterize the amount that a trader is long as the amount that he would have to sell to make his portfolio have zero market exposure—that is, to be uncorrelated with the market. In cases such as this, when the market is defined as one asset, that market exposure is exactly the distance between the current position in that asset and the risk-minimizing position. At positions larger than the risk-minimizing position, the portfolio's returns are positively correlated with the asset returns; at positions less than this, the portfolio's returns are negatively correlated with those of the asset. At the risk-minimizing position, the portfolio is uncorrelated with the asset—a position that represents a market-neutral position relative to that asset. More generally, the same idea applies in an investment context where we typically do not represent the market with a single asset, but we can think of buying or selling amounts of an index or

assets correlated with the index in order to manage the investment portfolio's market exposure.

Table 4.13 shows the Market Exposure report for the Europe Fixed Income Portfolio expressed in terms of German 10-year bond equivalents. While the basic exposures in most countries correspond to the duration-based measure of positions shown in table 4.3, notice that there are some significant differences in such holdings as Belgium and the ECU, where significant currency positions contribute to Market Exposure but not to duration. Another exception is in the U.K. 10-year sector, where a basis trade long bonds and short futures nets out to a small short position on a duration basis and a small long position on a market exposure basis. Finally, notice that the long Finland position in the under-one-year sector contributes significantly less market exposure than duration because of the relatively low correlation between Finnish short rate changes and 10-year yield changes in Germany.

Best Replicating Portfolios

A natural generalization of the Best Hedge is what we call the "Best Replicating Portfolio." The general idea is that in most portfolios there is probably no one asset that hedges very effectively. For a variety of reasons, we may want to find some small combination of asset positions that does provide an effective hedge—that is, a replicating portfolio. One benefit is simply to aid in understanding. For example, as we have seen, even when the hundreds of positions in our Europe Fixed Income Portfolio are aggregated, it is difficult to comprehend all that is going on. Here's where a small Best Replicating Portfolio can help. In addition, if a more effective hedge is desired, the Best Replicating Portfolio is the optimal sale of a small number of assets. In our risk system, we have an algorithm designed to quickly identify and compute the best three-, five-, and ten-asset replicating portfolios. We also allow traders to choose a set of assets for a replicating portfolio, in which case we simply solve for the weights.

Table 4.14 shows the three-, five-, and ten-asset Best Replicating Portfolios for the Europe Fixed Income Portfolio that we have been following. In addition to the choice of assets and optimal weights, the table shows the percentage contribution to risk of the replicating portfolio for each asset and the percentage of risk explained—that is, 100 times one minus the ratio of residual risk to the original risk of the portfolio.

The Hot Spots and Best Replicating Portfolio reports provide similar information—a picture of where the risks in a portfolio are coming from. They do differ, however, and each has strengths and weaknesses. In general, Hot Spots is more disaggregated, has more information, but unfortunately is more difficult to interpret. The Best Replicating Portfolio is especially useful in the context of a large, complex portfolio. These differences are highlighted by comparing the two reports for the fixed income portfolio.

The three-asset Best Replicating Portfolio in table 4.14 clearly summarizes the main risks of the fixed income positions in a concise, easy-to-

Table 4.13. Market Exposure Summary for the Europe Fixed Income Portfolio.[a]

Asset	Total	Ger	Ndl	Fra	Bgm	U.K.	U.S.	Ita	Spn	Dnk	Swe	Fin	Nor	Swz	Ecu	Jap	Hkg	SAf
Currency	–7.3	–6.8	0	–1.5	16.0	0		–0.8	0.8	0	0	0	–4.8	1.5	–11.5	0	0	0
< 1 year	55.8	8.5	0	3.3	0	4.0	0.3	34.5	0.8	0	0	4.5	0	0	0	0	0	0
1 to 5 year	132.0	0.5	4.5	35.8	–2.3	56.3	18.3	52.8	2.3	–44.0	8.8	0.8	0	0	–1.5	0	0	0
5 to 12 year	–253.5	–200.5	2.0	–75.0	2.0	3.0	6.5	120.0	–126.3	34.3	–14.8	–14.5	0	0	9.3	–0.5	0	1.0
12 to 30 year	313.3	228.8	0.5	7.5	0	80.8	–12.0	7.8	0	0	0	0	0	0	0	0	0	0
Swaps	–20.3	–20.3	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Total FI	220.0	10.3	7.0	–30.0	15.8	144.0	13.3	213.8	–122.8	–9.8	–6.0	–9.3	-4.8	1.5	–4.0	–0.5	0	1.0

[a]Expressed in 10-year German equivalents.

Table 4.14. Best Replicating Portfolio Report for the Europe Fixed Income Portfolio.

Best Three-Asset Portfolio, 59% of Risk Explained

Asset	Replicating Portfolio Weight	Percentage Contribution to Risk
Italian 10-year bond	226.25	92
U.K. 10-year future	109.25	24
Spanish 10-year bond	–144.25	–16

Best Five-Asset Portfolio, 63% of Risk Explained

Asset	Replicating Portfolio Weight	Percentage Contribution to Risk
Italian 10-year bond	178.75	70
Italian floating rate bonds	298.00	18
Spanish 10-year bond	–151.25	–16
U.K. 10-year future	69.00	15
German 7-year bond	99.75	13

Best Ten-Asset Portfolio, 82% of Risk Explained

Asset	Replicating Portfolio Weight	Percentage Contribution to Risk
German 10-year future	–54.25	–68
German 7-year bond	40.75	47
Italian 10-year bond	109.75	38
German 30-year bond	139.50	32
Italian 5-year bond	126.50	29
U.K. 10-year future	70.50	13
Spanish 10-year bond	–143.00	–13
U.K. 5-year bond	103.50	12
Italian floating rate bond	213.50	12
Danish 3-year bond	–76.75	–3

comprehend format. The portfolio is basically long $226 million in Italian and $109 million in U.K. 10-year bonds and short $144 million in Spanish 10-year bonds. If more detail is needed, one can look at the five- and ten-asset Best Replicating Portfolios. Thus, the information in these reports is similar to that in the Hot Spots reports and in some respects is easier to understand. The Hot Spots report can often be confusing. For example, the one in table 4.10 indicates that the United Kingdom is a source of 25% of the risk, but there are several U.K. asset classes contributing and others hedging the overall risk. A careful look at the positions in table 4.3 shows

that the contributors to risk are long positions while the short positions are hedges; but in general the risks may not be so easily understood by looking at the Hot Spots report alone.

On the other hand, the Hot Spots report does reveal the role that a particular asset class is playing, which unfortunately the Best Replicating Portfolio might not. In this example, the U.K. 10-year bond future in the Best Replicating Portfolio has a positive weight and is serving as a proxy for the long position in U.K. bonds. In the actual portfolio, however, the position in the U.K. 10-year bond future is short and hedges the overall long bond position in the portfolio, as can be seen in the Hot Spots report. Nonetheless, the two reports are complementary and together should provide a clear picture of what is going on in the portfolio.

Implied Views

In marketing asset allocation services, we have found that in a large number of cases the optimal portfolios generated when we put the client's views into our optimizer are very different from the client's current portfolio. In trying to explain these differences, we have found it useful to compare the client's stated views with those that would have been required to generate the client's current portfolio from the optimizer. We call those views for which the current portfolio is optimal the "implied views" of the portfolio. The comparison of actual views with the implied views of a portfolio is a useful way to understand how and why portfolios can be improved.

We think of the creation of implied views as the process of reverse engineering a portfolio. Instead of taking views as input and creating an optimal portfolio, we take the portfolio as input and create the set of implied views. Because every portfolio generates a set of implied views and the analysis requires no further input from the portfolio manager or trader, it can be useful to incorporate this exercise into the risk management routine.

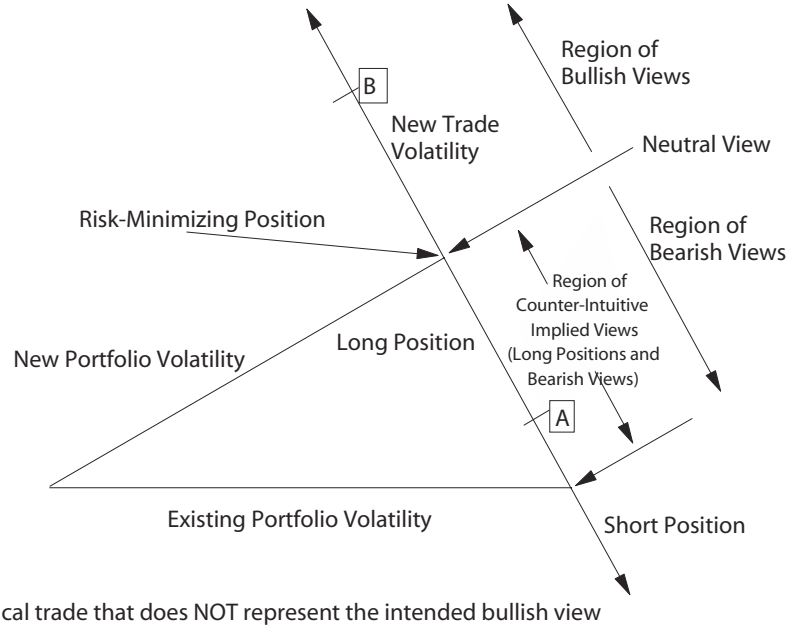
At Goldman Sachs, we do not create our trading positions from the top down. Our positions are not the result of a large portfolio optimization. Rather, at any given time, the positions are the aggregate of many individual decisions by traders operating in many different markets. At the end of each day, however, we do a risk decomposition analysis to highlight the hot spots, and we do an implied views analysis to highlight any inconsistency between the views expressed by our economists and those implicit in our trading positions.

One reason that the implied views of a portfolio are interesting stems from the above-mentioned misperception that many traders and portfolio managers have—specifically, that their "neutral" position in an asset is either no position or, for managers with a benchmark, close to their benchmark weight. Of course, this is true if a particular asset is the only one in a portfolio. But, in general, the neutral position for any given asset is a function of all of the other positions in the portfolio. The implication of taking other positions into account is that a portfolio manager has to understand the volatilities and correlations of the returns of all of his assets in order to

know what view a particular position represents. A long position (or overweight the benchmark) can represent a bearish view; similarly, a short position can represent a bullish view. It all depends upon what else is going on in the portfolio.

How can an overweight position represent a bearish view? This can occur anytime the returns of a position are negatively correlated with the returns of the rest of the portfolio. As illustrated in figure 4.19, whenever there is a position with such a negative correlation, there will be a region of long position sizes, between 0 and the risk-minimizing or "best hedge" position, for which the implied views will be bearish and therefore counterintuitive. If the investor's actual view is bullish—which is generally the case when the investor is long (or overweight)—then there is an opportunity to increase the expected return and reduce the risk by increasing the size of the position. In figure 4.19, this would be true for all positions between points A and B.

To develop the intuition that underlies the implied views of a portfolio, note that, for a portfolio to be optimal, the return per unit of contribution to portfolio risk must be the same for all assets; otherwise, a portfolio could be improved by moving out of an asset with a lower return per unit of portfolio risk into an asset with a higher return per unit of portfolio risk. Of course, this implies that you must understand the overall risks in a



A: Typical trade that does NOT represent the intended bullish view
B: The trade that has equal risk to A and higher expected return, and DOES represent a bullish view

Figure 4.19. Counter-Intuitive Implied Views Represent Opportunities. They arise often whenever a new trade is negatively correlated with an existing portfolio

portfolio in order to understand what views the portfolio represents. In the region of counterintuitive views, a decrease in position size increases portfolio risk. In this situation, there is an implied negative return (more accurately an implied negative expected excess return relative to the risk-free rate) despite the long or overweight position, in order to justify not increasing the position and reducing the risk.

Most traders and portfolio managers find the analysis above counterintuitive. No doubt it is strange to regard a long position as representing a bearish view. But don't think that this situation shows up only in some unusual set of circumstances in leveraged portfolios. It is actually an everyday occurrence in investment portfolios. An example is provided by the Canadian position in our global equity portfolio. Although the 4% position is 0.11% larger than the risk-minimizing position, any overweight position less than 3.8% would have represented a bearish view. Most investment portfolios are dominated by one risk: They are long or short the market (or over- or underweight relative to their benchmark). In any portfolio in which there is such a dominant risk, whether or not a position in some other asset represents a bullish or bearish view is simply a function of the correlation of that position's returns with the dominant risk—it is not particularly sensitive to the size of the position in that asset. For instance, suppose the dominant risk is a long position in the U.S. equity market. Suppose further that the investor is bullish on oil prices and takes a long position in oil via purchases of the equities of oil companies or directly through commodity derivatives. Assume also that, as is almost always the case, the returns on oil are negatively correlated with the returns on the equity market. As long as the long position in oil is small relative to the overall size of the portfolio, it will have a negative correlation with the portfolio and will therefore, at the margin, reduce its overall risk. Consequently, such a position represents a bearish view despite being long.

You do not even have to look as far afield as commodities to find these types of situations. A global equity manager who manages relative to a market benchmark will generally have lots of equity risks that correlate negatively with the dominant risk of the portfolio. Throughout the early 1990s, the international equity portfolios managed for U.S. investors often had risk dominated by an underweight position in Japan. Whenever you had such a situation, the overweight positions in every equity market that correlated positively with Japanese equity were risk-reducing and therefore represented bearish views. Such situations were common in these global equity portfolios, and they generally went unrecognized by the portfolio managers.

When these types of counterintuitive implied views arise, they generally represent investment opportunities. In the example of the investor bullish on oil prices, the investor has an opportunity to increase his expected returns and reduce his risk by increasing the size of his oil position. Of course, this opportunity is not unlimited. At some point, the size of the oil position will grow larger than the risk-minimizing, or best hedge, position.

At that point, the position starts to represent a bullish view, and the increase in expected return has to be balanced against the increase in risk.

In table 4.15, we show the implied views for the Europe Fixed Income Portfolio. Notice that the overall long position in Europe is reflected in the implied views of rate declines. Not surprisingly, the largest source of risk, long Italian bonds, is reflected in the strongest implied view. On the other hand, this is not a simple straightforward mapping. Notice also that the significant short positions in Spain and France do not imply increases in yields—just a spread widening relative to Italy.

Two Different Types of Marginal Analyses

A key message of this chapter is that it is the contribution to portfolio risk that matters, not the volatility of a position in isolation. Both the Hot Spots report and the implied views are driven by this marginal contribution to portfolio risk. Yet there is an important distinction between how marginal risk drives the contribution to risk and how it creates the implied views. Implied views are a function of the increase in risk generated by the next marginal *unit* of investment. In other words, they are a function of the slope of the Trade Risk Profile for that asset at the current position. Contribution to risk, on the other hand, is a function of the impact on portfolio risk of a marginal *percentage* increase in the position size. There can be a big difference, especially when the minimum risk is far from the origin. In such a case, having no position may represent a strong view, but clearly it can't be a contributor to risk. For example, the 10-year U.S. Treasury bond's returns may be highly correlated with the returns of a current portfolio in which the risk is dominated by a long position in 30-year U.S. Treasuries, even though there is no current position of 10-year bonds in the portfolio. In this case, it makes sense to consider the contribution to risk of the 10-year bonds as zero, but it also makes sense that the implied view on 10-year bonds, the returns of which are highly correlated with those of the 30-year bonds, is strongly bullish. A neutral view on 10-year bonds, for example, is inconsistent with having no position. This is because it would suggest using 10-year bonds as a hedge, which would generate a substantial risk reduction with no change in expected return.

Optimal Portfolios

Portfolio optimization is a relatively simple computational exercise but a rare procedure in the world of traders and portfolio managers. The practical difficulties that arise in optimizations—difficulties of systematically formulating views in a quantitative manner, of estimating a covariance matrix, of rebalancing portfolios, and of understanding the counterintuitive results that often come out of the complex interactions of views, correlations, constraints, and measuring risk relative to benchmarks—leave many portfolio managers scratching their heads. Our experience in working with portfolio managers suggests that the answer is not to give up but rather to start with a thorough grounding in understanding the sources of risk in

Table 4.15. One-month horizon Implied Views for the Europe Fixed Income Portfolio.[a]

Asset	Ger	Ndl	Fra	Bgm	U.K.	U.S.	Ita	Spn	Dnk	Swe	Fin	Nor	Swz	Ecu	Jap	Hkg	SAf
Currency	.89	.86	.66	.87	–.43		–.51	1.00	.86	.30	.67	.76	.97	–.51	.27	0	.91
2-year	–14	–6	–18	–13	–17	–14	–53	–5	–11	–20	–5						
5-year	–14	–11	–14	–12	–18	–17	–50	–8	–15	–20					–1		
10-year	–11	–10	–10		–17	–16	-47	–9	–13	–19	–8			–13	0		5
30-year	–12	–10	–10		–15	–13	–40							–12			

[a]Percentage change in U.S. dollar exchange rate for currencies. Yield changes are in basis points. In all cases except the U.K., a positive number for the exchange rate represents appreciation of the dollar. In the U.K. case, where the exchange rate is quoted in dollars per pound, a positive number represents sterling appreciation.

the portfolio, the correlations of each asset with the dominant risks, and the other portfolio effects. The Hot Spots report, the Best Hedges report, Trade Risk Profiles, Best Replicating Portfolios, Market Exposure, and the Implied Views of the portfolio are all useful in this context. Once these tools are mastered, portfolio optimization—or, at the very least, improved portfolio management—follows naturally. We discussed earlier a situation where a bullish view in oil might have led to a surprisingly large position because of the negative correlation between that asset and the dominant equity risk in the portfolio. More generally, the art of successful portfolio management is not only to be able to identify opportunities but also to balance them against the risks that they create in the context of the overall portfolio.

6. CONCLUSION

Risk management has received more attention in recent years as the complexity of financial markets has increased. Unfortunately, so far that focus has been applied almost exclusively to derivatives and the risks of individual securities. Investors and traders need to recognize—and the practice of risk managers needs to reflect—that what matters is the marginal impact on the risk of the portfolio, not the risk of individual securities. With that in mind, this chapter has highlighted some of the tools and analyses that we have developed along these lines at Goldman Sachs to help improve the profitability of our traders and the investment performance of our clients.

ACKNOWLEDGMENTS Very few of the ideas expressed in this chapter are original. This is basically a compilation of insights obtained over the years from my colleagues and our clients. I am especially indebted to the late Fischer Black for his intellectual guidance over the past decade. Much of the content of this chapter was developed when we jointly prepared a course on portfolio theory for the fixed income professionals at Goldman Sachs in 1993. Also, the ideas here have benefited greatly from the intellectual stimulation and thoughtful criticism of my colleagues Armen Avanessians, Arnout Eikeboom, Jacob Goldfield, Bob Litzenberger, Mark Painting, Ken Perry, Scott Pinkus, Barbara Smigelski, Susan Waters, Nick Weir, Anthony Williams, and Kurt Winkelmann; and my former colleagues Marta Amieva, Doug Greenig, Chi-fu Huang, Ugo Loser, Tom Macirowski, Ravi Narula, Scott Richard, Jon Ruff, Jose Scheinkman, Ken Singleton, Suresh Wadhwani, and Larry Weiss. In addition, support and insights were provided from people too numerous to mention individually, including the management of Goldman Sachs, the derivatives and market research specialists, the other members of the risk analysis team, the traders whom we both serve and monitor, and the many Goldman Sachs clients and risk management professionals at other firms with whom I've had the pleasure of interacting.

NOTES

1. Value-at-risk is a measure of a point in the distribution of possible outcomes. It has two parameters: a horizon and a probability. For example, a common regulatory definition of VaR is the amount of capital that you should expect to lose no more than once in a

hundred two-week intervals, given your current positions. At Goldman Sachs, we commonly focus on an amount of capital that we should expect to lose no more than once per year in a given day. We think of this not as a "worst case" but rather as a regularly occurring event with which we should be comfortable.

2. Despite widespread acceptance of VAR as a statistical measure of risk, many risk managers still think it has been oversold—they feel it represents an attempt to generate one number from a "black box" analysis that can summarize risk, but they argue it provides little understanding of risk or guidance in risk management. Of course, no one number can provide much information by itself. But, as discussed in this chapter, because statistical measures such as VAR can be used appropriately as the basis for understanding the sources and magnitudes of risk in a portfolio, they should in fact provide the cornerstone of modern risk management.

3. Of course, in actually trying to implement such an approach, the risk manager will encounter many interesting and difficult statistical issues that must be addressed but that are beyond the scope of this chapter.

4. Note that an assumption of nonnormality, in particular fat tails, is natural in financial markets. The usual justification of the assumption of a normal distribution—that many independent sources of uncertainty sum to create the total uncertainty—is clearly not valid. On those days when important information is observed, most market participants will react to the same information. In some special situations, as, for example, was the case in the crash of October 1987, that information may be in large part the market price movement itself. (The crash was approximately a 25–standard-deviation event relative to previous daily price moves.)

5. Currency devaluations, especially during periods when central banks are attempting to fix or stabilize exchange rates, are a notable exception to this rule of thumb. In such situations, the probability of a devaluation is uncertain, and when it occurs, it is generally much larger than four standard deviations relative to the daily volatility of the preceding period.

6. This aggregation can be handled in various ways. Here, we simply assume that the individual equity positions are small and sufficiently representative of the index that the dollar value in each asset class is a fair representation of the index exposure. An alternative would be to compute the "beta" of the individual equity positions with respect to the asset class benchmark in order to get a more precise measure of exposure. The choice of how finely to disaggregate the data is a compromise between accuracy of risk measurement, which comes from disaggregation, and the clarity that comes from aggregation. In addition, lack of availability of data or computing resources may limit the degree to which the positions can be analyzed on a disaggregated basis.

7. For simplicity, we assume that half of the currency exposures created by the deviations from the benchmark weights are hedged. At this level of hedging, the currency exposures do not create significant risk relative to the equity exposures, and in this chapter we report the sum of the two.

8. This report shows positions aggregated into the sectors of the fixed income markets of different countries. This actually represents the second level of aggregation: from positions into asset classes, and then from asset classes into sectors. In addition, we aggregate into risk factors, such as a parallel move in interest rates, a steepening of a yield curve, etc.

9. Traders often add other Greek letters, such as the theta (the time decay of the portfolio) and the rho (the sensitivity to financing rates). We do not report them here. Theta is better viewed as a cost, not a risk factor, since it is a known quantity, and rho is not a significant risk in this context.

10. In the course of defending the use of statistical approaches to risk management, we do not mean to suggest that other approaches are invalid or obsolete. In particular, in this chapter we will say nothing about issues such as looking for illiquid or aged items, two particular classes of securities that need to be monitored closely by a risk manager but do not get highlighted by our statistical measures.

11. We estimated the covariance matrix used here, and volatilities and correlations quoted later in this chapter, using daily excess return data from February 1, 1988, through

April 12, 1996. We gave all observations equal weight in the estimation. When applying the covariance approach to trading positions that have a short expected holding period, we typically use an exponential decay function that puts more weight on more recent observations. Such an approach has drawbacks, though. There is a delicate trade-off between the benefit of capturing current market conditions by downweighting older data and the increased noise in the estimates that arises from having fewer observations. This is a particularly acute concern when the covariance matrix is to be used for understanding sources of risk and constructing hedges, not simply for measuring risk. For example, no matter how short the expected holding period, we would never put the majority of the weight on data less than one month old, an unfortunate and dangerous feature of one widely used approach. Our default decay rate for trading positions is 20% per month. For analyses of client portfolios, which typically have a much longer expected holding period, we typically let the data decay much more slowly.

12. Because we are focusing on tracking error, we don't need to specify benchmark weights. We have, however, set underweighted positions to be no greater in absolute value than the market capitalization weight.

13. An advantage of using a covariance matrix is that it can be estimated from a relatively short period of data and therefore can capture the current state of volatilities and correlations. In the context of a portfolio with nonlinearities, however, a portfolio manager is likely to be especially concerned about large, rare events and may prefer to use a long history of actual price and volatility changes rather than to rely on the assumption that current conditions, which may reflect recent low volatility, will persist. On the other hand, the historical simulation analysis does give up two important advantages of the covariance approach: first, the ability to track time-varying volatilities and correlations and their impact on risk; and second, the ability to search the "event space" thoroughly rather than simply being based on the limited set of actual historical events.

14. The yield and implied volatility changes for the particular contracts in the portfolio are not actual yield and implied volatility changes for historical contracts; rather, they are linearly interpolated values with weights that produce "constant-maturity" yield and volatility changes. While not perfect, this avoids some of the instability of using yield and volatility changes on "rolling" contracts.

15. Note the similarity. The formula for the length of a side of a triangle, A, given the lengths of the two other sides, B and C, and the angle between them, θ, is given by: $A^2 = B^2 + C^2 - 2BC\cos(\theta)$. The formula for the volatility of a portfolio, A, given the volatilities of its two constituent assets, B and C, and the correlation between their returns, cor(), is given by: $A^2 = B^2 + C^2 + 2BC\ \mathrm{cor}()$. Clearly, the analogy works when we let the lengths of the two sides of the triangle correspond to the volatilities of the two assets and the angle between them, θ equal the arccos of the negative of the correlation of the returns of the two assets. I discovered the usefulness of this analogy while preparing a continuing education class for fixed income professionals at Goldman Sachs, where it was quite well-received. While none of my colleagues were aware of the use of this analogy, if it has been previously used (which I suspect is likely), I apologize for the lack of a reference.

16. At the risk of complicating the example, we have recognized that the purchase of the new asset is generally funded from the sale of another asset.

17. This marginal decomposition of risk contrasts with the common practice of looking at the incremental impact on risk of a given position. The incremental view asks the question, "What happens to the portfolio risk if each position individually is removed in its entirety from the portfolio?" Unfortunately, such an analysis generally does not help to identify the sources of risk in a portfolio. Consider, for example, that, in this simple two-asset case, removing either position increases the risk of the portfolio by a large multiple. This information is certainly not as useful for understanding the risks in the current portfolio as is knowing the impacts of marginal changes in each.

18. Not all measures of risk have this convenient property. For example, one that does not is the "probability of shortfall." This measure, which is defined as the probability of a loss greater than some specified amount over some period of time, does not exhibit the lin-

earity that allows the decomposition that we develop in this chapter. This measure may be useful if one's disutility jumps at a shortfall point, but it is not well-suited for general risk management. For example, when applied to an out-of-the-money call option that expires during the period, the probability of shortfall is zero as long as the premium is less than the shortfall, and then jumps to nearly one as soon as the premium becomes greater than the shortfall—a discontinuity that is unlikely to be meaningful.

19. We can measure the risk contribution of a "position" or a "trade." From a mathematical point of view, there is no useful distinction in this context between a single position or a combination of positions that might be considered a "trade."

20. Using the notation above, if the risk contribution is 0, then $R_i(x) * x_i = 0$. In this case, either $x_i = 0$ or $R_i(x) = 0$. The latter condition, that the derivative of risk with respect to changes in the ith position is zero, implies that the ith position is risk-minimizing.

21. The computation of the risk decomposition is, as described above, a derivative with respect to percentage changes in position size. However, if we take a particular day to represent the VAR—for example, the fifth worst out of the five years of data—then the result will be very sensitive to what happened on that particular day. To avoid this problem, we apply a smoothing filter to a set of worst days centered on the fifth worst. The filter we use is simply a seven-day triangular window with weights {1, 2, 3, 4, 3, 2, 1}.

22. Earlier, we looked at the impact of trades such as selling Japanese equity and purchasing Canadian equity. Here we look at the risk of the portfolio as a function of the position size in Canadian equity, holding all other risk positions constant—or, equivalently, of exchanging cash for Canadian equity.

23. Robert Litterman and Kurt Winkelmann, *Managing Market Exposure*, Goldman Sachs & Co., Risk Management Series, January 1996.

REFERENCES

Black, Fischer, "Universal Hedging: Optimizing Currency Risk and Reward in International Equity Portfolios," *Financial Analysts Journal*, Vol. 45, July/August 1989, pp. 16–22.

Black, Fischer, and Robert Litterman, *Global Asset Allocation with Equities, Bonds, and Currencies*, Goldman Sachs & Co., Fixed Income Research, October 1991.

Litterman, Robert, and Kurt Winkelmann, *Managing Market Exposure*, Goldman Sachs & Co., Risk Management Series, January 1996.

Markowitz, Harry, "Portfolio Selection," *Journal of Finance*, Vol. 7, March 1952, pp. 77–91.

Sharpe, William F., "Capital Asset Prices: A Theory of Market Equilibrium under Conditions of Risk," *Journal of Finance*, Vol. 19, No. 3, September 1964, pp. 425–442.

5

Markets for Agents: Fund Management

Stephen A. Ross

During the period from the 1970s through the 1990s, the financial markets of the United States and of the world became institutionalized. By some estimates, over this period of time, the ownership of assets has reversed from one in which individuals held approximately 75% of the assets and institutions held 25% to one in which institutional ownership is now approximately 75%. While this has occasioned much discussion in the popular press, it is probably fair to say that the response in the academic community has been one of indifference.

In large part, the academic response stems from a mind-set in which equilibrium in the financial markets is determined by the interactions of many individuals operating with their own information, which is partially communicated through prices in efficient markets and under the discipline of no arbitrage. This paradigm is exceptionally useful, but it brings with it a strong predisposition to think of institutions as somewhat irrelevant. After all, aren't financial institutions only a reflection of the constituencies they serve, and, as such, aren't they merely more efficient facilitators of the original concept of equilibrium?

This is, of course, a quite reasonable first approximation to the impact of institutions that manage large blocks of funds, such as pension funds and insurance companies and fund management companies, but it also somewhat overstates the case. In particular, we know that institutions have access to information that individuals do not, and as long as acquiring such information is costly with attendant economies of scale and scope, and as long as prices do not fully reveal the information of market participants, this may well be a primary reason for the existence of such institutions. Furthermore, insofar as institutions act as agents for primary participants in the market, this will bring with it all of the attendant problems of agency, including incentive problems, moral hazard issues, and adverse-selection problems.

Not surprisingly, Fischer Black had an enduring interest in the role of economic institutions in general and the money management business in particular. One of his earliest papers (Black and Treynor 1973) directly addressed the question of how money managers could combine portfolio theory with security analysis to obtain 'alphas', and he continued this theme

of using Modern Portfolio Theory to improve portfolio management in his work on global asset allocation (Black and Litterman 1991, 1992). Nevertheless, despite his defining contributions to modern finance, numerous conversations convinced me that Fischer's faith in the explanatory power of economic equilibrium made him more of a skeptic than an agnostic on the value of quantitative active portfolio management. Fischer's view that the rational expectation equilibrium was 'noisy'—perhaps by a factor of two—fits nicely with the explanation offered in this chapter for the rise of money managers who would probably add little value in a world of perfectly efficient markets.

This chapter takes the view that institutions may well be important and concerns itself with modeling their impact on the market in a context where they are informed and offer their services to less informed individuals. We will examine this situation in a two-period model in which prices can convey information. The model we use is the noisy rational expectations model first introduced by Grossman (1976), refined by Hellwig (1980), and subsequently modified to analyze a wide variety of financial problems involving information in a series of chapters by Admati and Pfleiderer (1986, 1987, 1990). Bhattacharya and Pfleiderer (1985) have examined the issue of delegating portfolio management, but their concern was less with the equilibrium market for agents. A more closely related paper by Huberman and Kandel (1993) is also directly concerned with the incentives for money managers, but it does not address the sorting of agents by ability in an equilibrium in which abilities are continuous and the reward to reputation is endogenous.

In section 1 of this chapter, we examine a noisy rational expectations model and point out that the equilibrium is unstable to individuals offering to manage funds for others whose precision of information is lower than their own. This provides a rationale for the formation of mutual funds, and we examine the equilibrium when this occurs and agents endogenously emerge. The new equilibrium is similar in form to the original one in which participants trade directly, but it has some differences. One feature of the new equilibrium is that the market risk premium will reflect the attitudes of the agents who are choosing funds and their precisions as opposed to those of the original participants. Another is that it is profitable to be a fund manager with a high precision of information. Since fund management is assumed to be costless—we do not consider the costs of information acquisition—shirking is not an issue, and the agency problem is simply solved by paying managers fees that depend only on their precisions and not on performance.

The model developed in section 1 is of interest in its own right, but for the purposes of this chapter our primary interest is in the behavior of a market in which institutions and agents have yet to be identified. Section 2 considers such a situation one period prior to the equilibrium of section 1. At this time, individuals are assumed to know their own precisions, but, while they may have probabilistic assessments, they do not know the pre-

cisions of others. This is unlike the second stage, at which precisions are assumed to be common knowledge. As a consequence, agents and potential agents are interested in taking actions that will signal that they have high precisions, and the market will reach an equilibrium in which they are disciplined so that they cannot credibly forge their credentials.

1. THE NOISY RATIONAL EXPECTATIONS EQUILIBRIUM MODEL

The model we will explore has the following schematic. There are two periods. In the second period, principals have some information about how well-informed agents are, and they use this information to allocate their investments across principals. Agents compete for principals by the terms on which they offer their services. We will assume that a revelation principle is at work or just note that, in our model, at the second stage agents are either indifferent between doing what the principals would want and any other policy either because they are paid a fixed sum independent of their actions or because the incentives are perfectly aligned.

In the first stage, the actions and consequences of the actions taken by agents provide information for the second-stage assessments of their abilities. In the first stage, then, the terms under which they are employed by principals must take account not only of first-stage rewards and incentives but also of the second-stage implications for the agents' reputations. This latter is what makes the first-stage control issues of paramount importance.

We begin at the second stage and work backward. The basic model is Hellwig's (1980) noisy rational expectations model with an infinity of traders (see also Pfleiderer (1984) and Admati (1985)). Hellwig's model is an equilibrium model in which investors receive signals about future values and trade on the basis of this information and the information revealed to them from the equilibrium price. It is useful for our purposes since our rationale for the agency relation to exist at all is differentially informed participants in markets and since we are also looking for the equilibrium implications of markets with institutional agents.

The market has only a single risky asset and a riskless asset, and, just to simplify the algebra, we will assume that the rate of interest is zero. Participants in the market have constant absolute risk aversion utility functions

$$U(w) = -e^{-Aw}$$

and the coefficient of absolute risk aversion, A, is the same for all. It is well-known that the demand for the risky asset is given by

$$x = \frac{E - p}{A\sigma_v^2}$$

where x is the demand for units of the risky asset, E denotes the expected end-of-period payoff on one unit of the risky asset, p is the risky asset's current price in terms of the riskless asset numeraire, and σ_v is the

standard deviation of the random terminal payoff, v, on one unit of the risky asset.

Indexing agents by i ($i = 1, \ldots, n$), market clearing requires that

$$\frac{1}{n}\sum x^i + z = \frac{1}{n}\sum \frac{E^i - p}{A\sigma_i^2} + z = 0$$

where z is a per capita random demand for the risky asset that provides the "noise" that—in the words of the Green Hornet—"clouds men's minds" and prevents the equilibrium price from being fully revealing of the economy-wide information.

Information comes in the form of signals about the future value, v, of the risky asset. Each participant, i, observes a signal, $v + s^i$, where s^i is a mean zero random variable. All random variables are assumed to be normally distributed and independent of one another. Summarizing,

$$v \sim N(\bar{v}, \sigma_v^2)$$
$$z \sim N(\bar{z}, \sigma_z^2)$$

and

$$s^j \sim N(0, \sigma_s^2)$$

An individual's demand for the risky asset depends on his individual assessment of its expected payoff and its risk. Thus,

$$E^j = E\{v \mid v + s^j, p\}$$

and

$$\sigma_{jv}^2 = \mathrm{Var}\{v \mid v + s^j, p\} \quad .$$

We will let μ denote the measure of the participants as $n \to \infty$ and assume that the measure is sufficiently nonatomic that we can eliminate individual effects on the equilibrium. Under this condition, there is an equilibrium price linear in v and z

$$p = a_0 + a_v v + a_z z$$

where we now index participants by r and have

$$a_0 = \frac{1}{C}\left[\bar{v}\sigma_z^2 A + \sigma_v^2 \bar{z} AH\right]$$

and

$$a_v = \frac{1}{C}\left[\sigma_v^2 H\sigma_z^2 + \sigma_v^2 AH^2\right]$$

and

$$a_z = \frac{1}{C}\left[\sigma_v^2\sigma_z^2 + \sigma_v^2 AH\right]$$

where

$$C = A\sigma_z^2 + \sigma_v^2 H\sigma_z^2 + \sigma_v^2 AH^2$$

and

$$A = \int \frac{1}{A_r} d\mu_r, \quad H = \int \frac{1}{A_r\sigma_r^2} d\mu_r$$

The demand by any individual for the risky asset is a function of his signal precision and the particular signal he has received:

$$E[v|v+s^r, p] = \frac{1}{D}\left[\frac{\sigma_z^2}{H^2}\sigma_r^2\bar{v} + \frac{\sigma_z^2}{H^2}\sigma_v^2(v+s^r) + \sigma_v^2\sigma_r^2\left(\frac{p-a_0-a_z\bar{z}}{a_v}\right)\right]$$

and

$$\mathrm{Var}[v|v+s^r, p] = \frac{1}{D}\frac{\sigma_z^2}{H^2}\sigma_r^2\sigma_v^2$$

where

$$D = \frac{\sigma_z^2}{H^2}\sigma_v^2 + \frac{\sigma_z^2}{H^2}\sigma_r^2 + \sigma_v^2\sigma_r^2$$

Institutional Instability of the Equilibrium

Without any loss of generality, we need only consider variables as deviations from their unconditional means. The equilibrium condition is now split into two conditions, the original ones above, with all variables mean zero, and the additional condition where all are set at their unconditional means

$$\frac{1}{n}\sum E[x^i] + \bar{z} = \frac{1}{n}\sum \frac{E[E^i] - (a_0 + a_v\bar{v} + a_z\bar{z})}{A\sigma_{iv}^2} + \bar{z} = 0$$

In what follows, then, all variables will have mean zero.

Notice that the uninformed trader can only condition on the equilibrium price, p, and his demand will therefore be proportional to p,

$$x = \frac{\beta}{A}p$$

where β is the limit of the informed demand given above as the precision of the signal goes to zero (see appendix).

There are a number of difficulties with this and nearly all versions of noisy rational expectations models. For one thing, the source of the noise is unclear, and in reality its only real function is to serve as a deus ex machina to prevent the equilibrium from being fully revealing. In the limiting state where the noise disappears, individuals supposedly use or contribute their individual signals to the overall price, which is fully revealing of the terminal value, but since in the final equilibrium the price is a suffi-

cient statistic, no individual has any incentive to use his private information. This paradox is well-known and is the price we pay for wanting to achieve relevant results in a world in which individuals have limited computational capacities that we cannot yet model convincingly.

Less well-known is that the equilibrium itself is unstable on its own grounds in the sense that individuals have incentives to form institutions such as funds and by so doing can improve their lot. Suppose, for example, that we consider two individuals, i and j, with signal standard deviations $\sigma_i < \sigma_j$. In this case, i and j can strike a deal that makes them both better off. Suppose that i agrees to manage j's money for a fixed fee of b. Since this is the final stage, i is unconcerned about reputation effects, and since the fee is fixed, i is indifferent to following any policy, and we will assume that i simply does what j would do if j had access to the same information signal that i does. In other words, we will assume that i uses his signal s_i to maximize

$$E[-e^{-A_j w} \mid v + s^i, p]$$

The following lemma verifies that this is an improvement for participant j.

LEMMA 1

If $\sigma_i < \sigma_j$, then

$$E[\max E[-e^{-A_j w} \mid v + s^i, p]] > E[\max E[-e^{-A_j w} \mid v + s^j, p]]$$

Proof:

The result is quite general and is independent of the form of the utility function as long as there exist some actions that will alter utility. Since $\sigma_i < \sigma_j$, the signal s^i is a refinement of s^j, and for any utility function, U, we have

$$\begin{aligned} \max E[U(w) \mid v + s^j, p] &= \max E[E[U(w) \mid v + s^i, p] \mid v + s^j, p] \\ &< E[\max E[U(w) \mid v + s^i, p] \mid v + s^j, p] \end{aligned}$$

The inequality is strict because the inner expectation will vary with the signal. Taking unconditional expectations produces the desired result.

□

Since the manager will improve j's expected utility, there is a fee that i could charge j and still leave j better off than by self-managing. Not only, then, is the original equilibrium not Pareto-efficient, but in a world where the precisions of signals are common knowledge, the mechanism of one party managing assets for another hints that the equilibrium is not robust to the creation of financial institutions that manage funds.

The Market for Fund Management

The next step in our analysis is to develop an equilibrium in a world in which some participants select themselves out to be fund managers. It might be thought that this could degenerate into a situation where only the single best

manager survives and manages all of the assets. Upon reflection, though, this would only occur if principals had to put all of their funds with a single agent and agents could serve many principals. This would not be an equilibrium if principals could divide their funds among agents since a single fund manager would not offer diversification against the risk attendant in relying on his one signal. This opens up the possibility that inferior managers could survive by offering fund management that diversified the principals' positions and did so at lowered fees to compensate for the lowered precision of their signals. Alternatively, if principals are only able to hire a single agent and agents can only serve a single principal, then we would expect some sort of pairing equilibrium in which the wealthiest principals with the worst information paid the most for the best agents and so on.

To make this precise, we will assume that all of the participants are identical in all respects except that they differ in the precision of their information signals. Since they all have constant absolute coefficients of risk aversion, A, and identical wealth, w_0, setting up funds can be a bit tricky. In particular, since the constant coefficient of absolute risk aversion investor puts a fixed amount in risky assets independent of wealth, we have to determine how a fund will operate when it may have many investors and may only have a portion of each of their portfolios. We will adopt the following rules. We assume that a participant may set up a fund, at which point they are said to be an "agent" or a "manager", and we will refer to participants who invest in funds as "principals" or "clients." Funds are open-ended and do not trade as separate entities, and funds charge a fixed fee, θ, as a percentage of assets given by the clients. Later, we will examine other pay arrangements. The funds invest according to the following rule. First, the fund manager, the agent, determines the optimal amount of the risky asset for an individual with risk aversion A and the signal he receives. Call this optimal amount x. The fund then multiplies this by (W^f/w_0), where W^f is the amount invested in the fund and w_0 is the initial wealth of each investor, to obtain the amount of units of the risky asset held in the fund, $(W^f/w_0)x$. This is equivalent to managing the funds on a separate account basis, where each account with an amount w_j holds $(w_j/w_0)x$ units of the fund. Principals are free to invest in many agents—this could be thought of as either an individual diversifying across funds or a company hiring many managers or traders—and agents are free to attract funds from many principals, which really fits the fund model best. We will assume that these are mutually exclusive activities and that fund managers cannot also invest in other funds.

We begin the analysis by considering the problem of a client who is deciding on a portfolio of funds. Although a client doesn't see a manager's signals, he knows that each fund in which he holds ω_j units (as a proportion of initial wealth) will return an investment in the risky asset of

$$\omega_j x^j = \omega_j \frac{E\left[v|v+s^j,p\right]-\mathrm{p}}{A\,\mathrm{Var}\left[v|v+s^j,p\right]}$$

The client's end-of-period wealth is thus given by

$$w = \left[\sum \omega_j x^j\right](v-p) - \left[\sum w_0 \omega_j \theta_j\right] + w_0$$

In what follows, we will assume that everyone is a potential agent and that the decision to become an agent is a function only of the precision or, more precisely, the market's perception of the agent's precision. At this stage, perceptions are reality because precisions are common knowledge. Agents are said to be informed if they receive a signal that is useful for predicting the terminal price, and they are said to be uninformed if their precision is zero, which is equivalent to not receiving a signal.

With this structure, we will assume that principals allocate to the funds before the market has opened and that the equilibrium price, p, has been observed. A principal will thus choose a portfolio of funds that maximizes unconditional expected utility

$$E[-e^{-Aw}]$$

The following theorem solves for the expected utility for any choice of a fund allocation. This theorem is of some interest in its own right, but it is primarily useful for establishing the first-order conditions for an optimal allocation across funds.

THEOREM 1

Conditional on any set of precisions, $<\sigma_j^2>$, the expected utility takes the form

$$E\left[-e^{-Aw} \middle| \left\langle \sigma_j^2 \right\rangle\right] = -\left[\phi - \psi\Omega\right]^{-\frac{1}{2}} e^{Aw_0 \sum \omega_j \theta_j}$$

where ϕ and ψ are as defined in Lemma A2 and

$$\Omega = \sum \left(\frac{\omega_j^2}{\sigma_j^2} \right)$$

Proof:

See appendix.

□

Deriving the first-order conditions from the form for the expected utility given in Theorem 1, we can obtain the demand for any fund j.

THEOREM 2

The demand for fund j is given by

$$\omega_j = 1 - \frac{k}{h_j} - c\frac{\theta_j}{h_j}$$

where

$$q = \frac{a_z^2 \sigma_z^2 \sigma_v^2 \beta^2}{\Psi} > 0$$

and

$$k = q\left(\sum \omega_j - 1\right)$$

and

$$c = A\bar{w}_0 \frac{1}{\Psi}[\phi - \Omega\Psi] > 0$$

Proof:

See appendix.

□

Theorem 2 will be the workhorse of our analysis. It shows that for any given fee θ_j charged by agent j, the direct impact of the fee on demand is proportional to the agent's precision, h_j. Notice, too, that as the precision goes to infinity, the amount invested in the fund goes to one and the fee becomes irrelevant. In a general equilibrium, the other terms in the demand function, c and k, are also functions of the fees, but the exercise of looking at the impact of a rise in precision is not a movement across equilibria but rather a cross-sectional movement within a given equilibrium moving across different agents. In this way, the equilibrium price remains unchanged and these terms may be kept constant.

Viable Agents and Equilibrium Analysis

Theorem 2 suggests a simple definition of a viable agent.

DEFINITION 1

An agent and, equivalently, the agent's fund is said to be viable at a given fee if there is positive demand for the fund at that fee.

Viability is important since there is no mechanism for shorting funds. This, in turn, means that only funds held in positive amounts will contribute to the equilibrium. From Theorem 2, viability requires that the precision be sufficiently high. A necessary condition for viability is that the fund be viable at a zero fee, which is equivalent to requiring that

$$h_j > k$$

Among viable agents, there are some simple orderings. Comparing the demand for fund i with that for fund j, we see that if the funds charge the same fee, then the one with the higher precision will have the greater demand,

$$\frac{\omega_i}{\omega_j} = \frac{1 - \frac{(k + c\theta)}{h_i}}{1 - \frac{(k + c\theta)}{h_j}}$$

Lemma A3 in the appendix verifies that $k + c\theta > 0$, which for two viable agents with both demands positive implies that $\omega_i > \omega_j$ as $h_i > h_j$.

Examining the structure of c and k in the appendix, it can be seen that in the special case where there is only a single agent, the demand is solved by setting the purchase at a single unit. This is as we would expect since the agent is more informed than the principal and the principal would optimize by simply putting one unit into the risky asset. In general, though, the solution of the demand equations is a fixed point problem. Not only is the demand for any given fund functionally dependent on the total number of units of all funds purchased, but it also depends through c and k on Φ, Ψ, and Ω, which are functions of weighted sums of the individual holdings. One example that is easily solved is where all of the agents have identical precisions, h, and charge the same fee, θ. In this case, it follows immediately from Theorem 2 that, reflecting the symmetry of the problem, the optimum holding of each fund is identical and can be solved as the root of a quadratic equation. Even in this case, though, analyzing the equilibrium is complex.

The demand equations are particularly easy to solve and very illuminating when all fees are set at zero. In this base case, the system of demand equations is of the form

$$\omega_j = 1 - \frac{q}{h_j}\left(\sum \omega_j - 1\right)$$

This implies that the total units of funds purchased is

$$\sum \omega_j = n\frac{\left(1+\frac{q}{h}\right)}{\left(1+n\frac{q}{h}\right)} = n\frac{(h+q)}{(h+nq)}$$

where n is the number of funds and h is the harmonic mean of the precisions,

$$h \equiv \left(\frac{1}{n}\sum \frac{1}{h_j}\right)^{-1}$$

Hence,

$$\begin{aligned}\omega_j &= 1 + \frac{q}{h_j}\left[\frac{1-n}{1+\frac{nq}{h}}\right] \\ &= 1 + \left(\frac{h}{h_j}\right)\left[\frac{q(1-n)}{h+qn}\right]\end{aligned}$$

From this it can easily be seen that agents whose precision is too low will not be viable. Any agent for whom

$$h_j < h\left[\frac{q(n-1)}{h+qn}\right]$$

is not viable. In particular, then, no uninformed participant with zero precision can be a viable agent unless all participants have zero precision.

As the number of agents grows large, the cutoff for viability approaches the harmonic mean itself. This implies a sort of "rolling up" of the equilibrium since the harmonic mean is taken over only viable agents. Suppose, for example, that all agents have the same precision, h. This is a perfectly acceptable equilibrium situation, and the market equilibrium price, p, will be the same as would obtain in a model without agents in which all of the participants had this precision. If we introduce a worse-informed agent into this situation, then he will obviously not be viable. On the other hand, if a better-informed agent with greater precision is introduced, he will be viable, but he will not attract a gross (atom in a continuum economy) amount, rather only that amount consistent with his precision and the formulas above. What limits the demand for the agent's funds is the need to diversify across signals even though one has a superior signal.

To see this more clearly, suppose that there are two types of agents, high and low, with bounded precisions,

$$0 < h_L < h_H < \infty$$

Since the coefficient, q, is a function of the average variance in agents' forecasts (see appendix), and since the noise term prevents the price from ever becoming fully revealing, q is bounded from above and below as the economy grows by adding agents (i.e., as n increases). This implies that the coefficient of relative precision that determines viability converges to 1, and

$$\begin{aligned}\omega_j &= 1+\frac{q}{h_j}\left[\frac{1-n}{1+\frac{nq}{h}}\right]\\ &\to 1-\frac{h}{h_j}\end{aligned}$$

It follows, then, that if there are two types of agents and the proportion represented by the high-precision agents is bounded away from zero, then the harmonic mean precision, h, will be strictly bounded above the precision of the low agents, $h > h_L$. Thus, as the economy grows, the low-precision agents will cease to be viable. This is what is meant by 'rolling up' the equilibrium. The only way to prevent this is for high-precision agents to become a vanishingly small proportion of the economy.

The Market Risk Premium

As we look at markets with different structures of agents, although the coefficients can change, the form of the equilibrium price is unaltered; it remains a linear combination of the true value, v, and the per capita noise, z. The investments are still made by agents receiving signals as in the original model, but now what matters is the precision of the fund managers and

not of the clients. Since fund managers have higher precisions than clients, and since the equilibrium will reflect the demand by these funds with higher precision, it is to be expected that the price will be a better signal of the true value than when there are no agents. As the number of agents increases and the equilibrium rolls up, only still viable agents will be used, and agents with lower precision will be discarded. This implies that the price will be a better predictor, which will set a higher hurdle rate for the fund manager to be viable.

One consequence of this is that the conditional variance that agents use in their demands will be lower than is used in an individual market without agents. This implies that the market risk premium declines as the market institutionalizes.

LEMMA 2

The market risk premium declines when the market is institutionalized with agents investing for individual participants.

Proof:

$$\begin{aligned} E[v]-E[p] &= E[z]A\,\mathrm{Var}[v|p] \\ &= (A\bar{z})\left[\frac{a_z^2\sigma_z^2\sigma_v^2}{a_v^2\sigma_v^2+a_z^2\sigma_z^2}\right] \\ &= (A\bar{z})\left[\frac{\sigma_z^2\sigma_v^2}{\left(\frac{a_v}{a_z}\right)^2\sigma_v^2+\sigma_z^2}\right] \\ &= (A\bar{z})\left[\frac{\sigma_z^2\sigma_v^2}{H^2\sigma_v^2+\sigma_z^2}\right] \end{aligned}$$

The market equilibrium is determined by agents with higher precisions than the average of the participants. This means that the harmonic mean of the precisions, AH, rises, and, since A is assumed constant, H must also rise. From the equilibrium conditions above and the definitions of the coefficients of the equilibrium price on v and z, a_v and a_z, respectively, the gross market risk premium is a decreasing function of H.

□

The Equilibrium Fee Structure and the Rewards to Management

We have assumed that agents charge a variable fee. As Admati and Pfleiderer (1990) point out, though, it is natural to consider nonlinear pricing systems and, canonically, a fixed fee or load for the fund. In the case of a load and no variable fee, the analysis becomes relatively straightforward. Conditional on having a surplus from investing in the fund (i.e., if the fee is less than the certainty equivalent of the gain from entering the fund), the demand for the fund will be as given in Theorem 2 with the variable fee set at zero. In other words, demand will be exactly as we have examined in

the no-fee case above. In fact, it is easy to show that in this case there exists a set of fixed fees that extract all of the surplus from the clients, which implies that a fixed load is actually superior in our simple model where all traders are symmetric. Presumably, with clients being differentiated, a variable fee might be part of the equilibrium.

Lastly, we turn to the determination of the equilibrium fee structure. Consider what the structure of the equilibrium might look like. If the agents compete among themselves for the business of the principals, there will generally be an equilibrium fee for their services—perhaps dependent on their perceived precision. Even though the marginal cost of agency has been assumed to be zero, unlike the case of competitors selling identical products, here there is complementarity, and each agent is of value at the margin because of his usefulness in diversifying the noise in the signals of other agents.

Furthermore, as the number of agents grows and as we approach a continuum of agents, it seems reasonable to assume that each agent will act as though he has no influence over the equilibrium itself. We will take this to mean that the agent does not feel that he can influence broad market averages or the equilibrium price. Rather, the agent will be assumed to maximize profits—revenues since we have assumed there are no costs to agency—by considering only the direct effects. The result will be a (Nash) equilibrium in the sense of monopolistic competition.

THEOREM 3

A viable fund manager achieves maximum profits by setting

$$\theta_j = \frac{h_j - k}{2c}$$

which produces profits of

$$m\frac{\left(h_j - k\right)^2}{4ch_j}$$

where m is the number of clients served.

Proof:

See appendix.

□

In Theorem 3, the agent maximizes utility by ignoring his impact on the equilibrium price parameters or the term $\Phi - \Omega\Psi$. In a large economy, the agent will have a negligible effect upon this term since it depends on the investment-weighted precisions and the equilibrium price parameters. Such an assumption would be untenable if an agent were atomic and loomed large in the determination of the equilibrium, but it is quite sensible in our context. For example, it is easy to show that in the case of n

identical agents the impact of a marginal change in the agent's fee or fund holdings on these variables is of order $(1/n)$ and, therefore, that is the order of the effect of the agent's decision. In other words, the direct effect computed in Theorem 3 dominates in large economies.

In the special case where all of the agents have identical precision, h, the equilibrium is the noisy rational expectations equilibrium for a market with participants whose precision is h. In this case, it can be shown that the parameters of individual fund demand, k and c, depend only on the number of total fund units purchased by each of the m identical participants, ω. It can then be shown that k and c remain of first order while the demand for the fund is of order $1/n$, where n is the number of funds. This implies that the profits

$$m\frac{(h-k)^2}{4ch} = m(\omega/n) = \omega(m/n)$$

depend on the proportion of clients to agents, (m/n), which will be large in a typical equilibrium. This verifies the basic intuition that being an agent can be very profitable. In particular, viable agents with high precision will make large profits.

We have yet to analyze viability generally with positive fees or in the monopolistic competition equilibrium, but presumably low-precision agents would have some ability to survive at lowered fee levels. This and other features of this model, such as the equilibrium with nonlinear fees and loads, are of enormous interest in their own right. There is certainly a practical sense that demand depends much more sensitively on precision and perceptions of precision than on fees, and it is particularly interesting to see if the model sustains such a result. Nevertheless, at the least we know that agents with sufficiently low precision (namely, no signals at all) will be forced out of business to find alternative employment. This is important because it puts a floor on the downside for agents' compensation. In the following section, we will use these results to examine the stage 1 decision of an agent.

2. THE FRAMEWORK

In this section, we consider the stage 1 decisions of a participant who understands the reward for having a high perceived precision at stage 2. We will adopt the same noisy rational expectations model structure as for stage 2, but here we will have a different mechanism for determining who invests for whom. We will assume that all participants seek to maximize the expected value of an exponential two-period utility,

$$-E[e^{-Aw_1 - Aw_2}]$$

and we can ignore endowments since with constant absolute risk aversion the demand for units of risky assets is independent of endowments.

In this formulation of preferences, wealth is assumed to be perfectly substitutable across the two periods. This could be easily modified to accommodate, for example, a different coefficient of risk aversion or time preference for period 2 versus period 1, but the additional generality would needlessly complicate the analysis. It is also entirely possible to think of the prices that are determined in the second period as being the terminal values in the first period. This provides an explicit intertemporal link between the two periods. The results that we will develop are entirely consistent with this possibility, but, since it only further complicates matters and brings no real additional insights, we will not explicitly develop this linkage.

From section 1, we know that the reward at stage 2 depends on whether the participant is a viable agent. If his precision is perceived to be high enough, then it will pay him to have or to be employed by an institution that manages funds, and his reward will be

$$m\frac{\left(h_j - k\right)^2}{4ch_j}$$

On the other hand, agents whose precision is below the critical cutoff of viability will leave the fund management business and seek alternative employment. Thus, the terminal wealth for an agent is given by

$$w_2 = \max\left[m\frac{(h-k)^2}{4ch}, w_a\right] \equiv \max\left[R(h), w_a\right]$$

where w_a is the wage in an alternative occupation. We write the payoff to a high-precision agent as $R(h)$ to emphasize its dependence on the precision, h.

The convex call option-like appearance of the second-period payoff as a function of precision has important implications for the first-period analysis. Clearly, agents have a strong stake in enhancing the market's perception of their precision, h. Our interest is in the properties of the stage 1 market equilibrium as participants look forward to the second-stage equilibrium.

We will assume that participants know their own precision—despite the fact that this is rarely the case—and that they consider all other participants as symmetric and assign a prior probability distribution to their precisions that is the same as that for the population as a whole. We will also assume at this stage that a set of participants has been chosen who are managers of the funds of others. These may be employees of fund firms, traders, or any group of young entrants to the financial markets, and we will call them agents at this stage. From this set of agents, some may emerge as second-stage agents as well, depending on whether they are perceived as having sufficiently high precision at the second stage. Since the agents are all symmetric from the perspective of the principals, the equilibrium will also be symmetric unless agents of differing precision can find an a priori way to

sort themselves out at stage 1. For example, if we let the payoff functions for fund managers differ, high-precision agents might choose to adopt payoff functions that would signal their knowledge of their high precision to the clients and they might, at stage 1, be able to sort themselves out.

To be explicit about this, we will assume that there are only two types of agents, uninformed agents, who receive no signals and, therefore, have zero precision, and informed agents, who receive a signal with precision h. Agents manage funds subject to payoff functions that determine their first-period compensation, $f(\cdot)$. The payoff depends on observables, and we will assume that the agent's portfolio choice, x, as well as the current price, p, and the terminal value, v, are observable. Perhaps the simplest way to think of this is that agents at this stage are endowed with some funds that they manage. This implies that

$$\text{payoff} = f(\xi)$$

where

$$\xi = (p, v, x).$$

Notice that, in principle, the payoff for any one agent could depend on the actions of all of the other agents, as in a tournament. This is permitted here as well, but there is no need to make it explicit since with a large number of agents each agent will be assumed to play a Nash game in which the other agents' actions are taken as given. This permits us to fold these strategic considerations into the form of the payoff function.

Given the first-period payoff functions, the actions taken by the agents, and the market results, the clients will make inferences about the precisions of the agents. A variety of different types of equilibria are possible. In general, an equilibrium will be a payoff function in the first period and a mapping from portfolio choices and outcomes into precisions. The payoff function describes what the agent receives as a function of both his first-period choice and first-period outcome, and to be an equilibrium this payoff function must satisfy the condition that, given the precision mapping, agents maximize their expected two-period utility over actions. Second, the mapping must be verified in the sense that it correctly identifies the precisions of the agents. This concept of an equilibrium is the same as that used in signaling theory and has been applied before in financial markets (see Ross 1977, 1978).

We will distinguish two versions of equilibria, one in which the precisions are perfectly revealed and one in which they are only probabilistic: as a formal matter, a fee, $f: R^3 \to R^+$, and the precision mapping, $M: R^{3n} \to S^{+n}$, where n is the number of agents and S^{+n} denotes the n-dimensional positive simplex of probabilities. The elements of R^{3n} are n tuples of triples, $\Upsilon = (\xi^1, \ldots, \xi^n)$, and an element of S^{+n} is an n-vector of probabilities $(\pi_1, \ldots, \pi_n)$, where π_j is the probability that agent j is informed. We will let I and U, respectively, denote the sets of informed and uninformed agents.

We assume that the fee mapping is to the positive orthant because we are making the important assumption that agents have limited liability. We

will also assume that the fee to the uninformed agent is bounded above (i.e., conditional on having no information beyond the market price, the uninformed agent cannot expect unbounded compensation). Since at stage 1 only the agent knows his precision, this compensation could exceed the alternative wage of uninformed participants at the second stage. In equilibrium, we would expect this bound to be the certainty equivalent compensation for an agent whose precision is 0 or h with weights representing the proportion of the population of stage 1 agents who are informed and uninformed. For the present, though, we will simply assume that entry to the market for agents sets this at w_c and require

$$E[-e^{-Af(\xi)} \mid p] \leq -e^{-Aw_c}$$

Equilibrium and Separation

We are now in a position to define a stage 1 equilibrium.

DEFINITION 2

An equilibrium (E) is a set $\{f(\cdot), M(\cdot), p\}$ such that, given the mapping and the fee structure, agents are optimizing, for $j = 1, \ldots, n$,

$$\text{argmax } \pi_j E[-e^{-Af(\xi) - AR(h)} \mid p, v + s^j] + (1 - \pi_j)E[-e^{-Af(\xi) - Aw_a} \mid p]$$

and such that the mapping is verified,

$$\pi_j \equiv M(\Upsilon)_j = \text{Prob } [j \in I \mid p, v, \Upsilon]$$

It should be admitted that this definition of equilibrium is something of a 'fudge'. It assumes that the reward to an agent who is perceived to have a probability π of being informed is distributed as a π chance of receiving $R(h)$ and a $(1 - \pi)$ chance of receiving w_a, the alternative wage. No doubt the payoff to such an agent will be of this form, but it cannot be asserted that the payoff function will be the same as the $R(h)$ derived above. This payoff function is the reward at the second stage for an agent with *known* precision h. We have not worked out the stage 2 equilibrium for a market in which agents are assigned probabilities that they are informed or not. Nevertheless, we can safely assume that there is still a reward for being perceived as informed, and we will write the equilibrium as we have defined it with the understanding that while $R(h)$ is an increasing function of h, it may not take the form we derived in section 1.

No such problem arises when the equilibrium perfectly separates out the informed agents from the uninformed agents.

DEFINITION 3

A separating equilibrium (SE) is a set $\{f(\cdot), M(\cdot), p\}$ such that, given the mapping and the fee structure, (i) clients do at least as well using agents as the uninformed choice they would make on their own, (ii) agents are optimizing, for $j = 1, \ldots, n$

$$\xi^j = \text{argmax } E[-e^{-Af(\xi) - A(\pi_j R(h) + (1-\pi_j)w_a)} \mid p, v + s^j]$$

and (iii) such that the mapping is verified with agents being perfectly sorted,

$$M(\Upsilon)_j \equiv \pi_j = 1 \text{ if } j \in I \wedge \pi_j = 0 \text{ if } j \in U$$

Agents will be perceived as informed if and only if their actions are the actions of informed agents. In what follows, we will let $x(I)$ and $x(U)$ denote the sets of actions taken by the informed and uninformed agents, respectively. Formally, the space of actions and outcomes is then partitioned into two disjoint sets,

$$x(I) \equiv [\xi_j \mid M(\Upsilon)_j \equiv \pi_j = 1]$$
$$x(U) \equiv [\xi_j \mid M(\Upsilon)_j \equiv \pi_j = 0]$$

The conditions that have to be satisfied for a separating equilibrium are twofold. First, the mapping has to separate informed from uninformed in the sense that informed will have no incentive to dissemble and pretend they are uninformed:

$$\max_{\xi \in x(I)} E[-e^{-Af(\xi) - AR(h)} \mid p, v + s^j] \geq \max_{\xi \notin x(I)} E[-e^{-Af(\xi) - Aw_a} \mid p, v + s]$$

Second, the fee schedule and the mapping have to induce uninformed agents to act honestly:

$$\max_{\xi \in x(I)} E[-e^{-Af(\xi) - AR(h)} \mid p] \leq \max_{\xi \notin x(I)} E[-e^{-Af(\xi) - Aw_a} \mid p]$$

The Impossibility of a Separating Equilibrium

Not surprisingly, when the second-stage reward for agency is high and limited liability constrains the available penalties for dissembling, no separating equilibrium exists.

THEOREM 4

If the second-stage reward for being an agent with perceived high precision, $R(h)$, exceeds the sum of the maximum first-stage compensation of untried agents, w_c, and the alternative second-stage wage for participants with low precision, w_a, then there is no separating equilibrium.

Proof:

The difficulty with finding a separating equilibrium when the terminal reward is high lies in satisfying the honesty constraint. From the analysis above, we know that a necessary condition for a separating equilibrium to exist is that

$$\max_{\xi \in x(I)} E[-e^{-Af(\xi) - AR(h)} \mid p] \leq \max_{\xi \notin x(I)} E[-e^{-Af(\xi) - Aw_a} \mid p]$$

The only variable that the agent controls is the allocation, x, so that the requirement that $\xi \in x(I)$ is simply that $x \in x(I)$ (i.e., the acceptance set projected onto the allocation decision), and, similarly, $\xi \in x(U)$ is that $x \in x(U)$. Thus, we can rewrite the honesty constraint as

$$e^{-AR(h)}\max_{x\in x(I)} E[-e^{-Af(\xi)} \mid p] \le e^{-Aw_a} \max_{x\notin x(I)} E[-e^{-Af(\xi)} \mid p]$$

From limited liability,

$$\max_{x\in x(I)} E[-e^{-Af(\xi)} \mid p] \ge -1$$

In addition, from the bound on the compensation of untried agents, we must have

$$\begin{aligned}\max_{x\notin x(I)} E[-e^{-Af(\xi)} \mid p] &\le \max_{x} E[-e^{-Af(\xi)} \mid p] \\ &\le -e^{-Aw_c}\end{aligned}$$

Combining these results, we have that a necessary condition for separation is simply that

$$R(h) \le w_c + w_a$$

□

Given the large ratio of participants to agents and the fact that $R(h)$ rises with precision h, it would not be surprising to expect the reward of an informed agent to exceed the combined first- and second-stage compensation of an uninformed agent. This implies that there will not generally be a separating equilibrium.

The Pooling Equilibria

If an equilibrium exists and is not a separating equilibrium, then it will be said to be a pooling equilibrium. In a pooling equilibrium, there is no mechanism for telling agents apart, and they will all receive the same compensation function, $f(\cdot)$. Of course, after the fact, some of the agents will do well and some will do poorly, and the market will make inferences from their performance. Exactly what occurs is unclear since models of this sort generally have an embarrassing richness of potential equilibria.

Suppose, for example, that the first-stage function for all agents is chosen to be a constant wage, w. Now the only concern of agents is the perception the market will have of them in the second stage. Whatever actions the informed take will be based upon their signals. If the informed act honestly—and they have no reason to do otherwise—then they will choose an amount in the risky asset equal to

$$x^j = x + \frac{1}{A\sigma_j^2}\left(v - p + s^j\right)$$

where x (= $x(p)$) is the demand of uninformed agents acting honestly. If uninformed agents simply chose x, then the market equilibrium at stage 1 would be determined solely by the demand of informed agents with precision h. This is not, however, a stage 1 equilibrium.

Clearly, the market would be able to tell the informed from the uninformed simply by observing who chose x and who did not. Any agent who

varied from x would be considered informed and that would signal their type. But, we have already shown that no such separating equilibrium is possible. Indeed, in this and in all such cases, there is a simple and superior strategy open to the uninformed agents. Uninformed agents would simply mimic the informed ones by using a 'signal' of their own making, y, which would replace $v + s^j$ in the demand equation for the risky asset, where y is drawn from a distribution

$$y \sim N(0, \sigma_v^2 + \sigma_s^2)$$

Doing so will make it impossible for the market to tell them from informed agents on the basis of their actions alone. The practical import of Theorem 4 is that there is no separating equilibrium possible at stage 1, so in all cases the uninformed participants will be unable to distinguish the agents on the basis of their period 1 actions alone. As a consequence, the stage 1 equilibrium will be one in which the market precision is lower than if the informed alone determined the price, and the market risk premium will be higher.

Of course, at stage 2, the market will have observed that some of the agents at stage 1 were successful and others were not. Since informed agents make better decisions, the market will presumably have a posterior distribution across the agents conditional on the outcomes. Informed agents will more likely have $x^j > x$ when $v > p$ and $x^j < x$ when $v < p$. Since the informed demand is strictly monotone (linear) in their signal, no other fee schedule could induce them to reveal more information than simply a constant at which they optimize in the usual honest fashion. In this pooling equilibrium, uninformed agents will mimic by using a false noise signal, y.

It is straightforward to derive the Bayes updated probabilities of an agent being informed or uninformed conditional on observing (x^j, v, p).

THEOREM 5

$$\text{Prob}\left[j \epsilon I \middle| v, p, x^j = c\right] = \frac{\pi}{\pi + (1-\pi)\left(\frac{\sigma_s}{\sigma_y}\right) e^{\frac{\alpha^2 - \beta^2}{2}}}$$

where

$$\pi \equiv \text{Prob}\left[j \epsilon I\right]$$

$$\alpha \equiv \frac{Ah^j (c - x) + p - v}{\sigma_s}$$

$$\beta \equiv \frac{(\sigma_s \alpha + v)}{\sigma_y}$$

Proof:

See appendix.

□

In a pooling equilibrium, participants would use Theorem 5 to update their assessments of each agent. To understand the implications of this formula, observe first that as the standard deviation $\sigma_s \to 0$, the agent becomes perfectly informed but, for any given c, the probability that he is informed approaches zero. The paradox is easily resolved by noting that, as the precision approaches infinity, the perfectly informed agent takes larger positions and the probability that the position is equal to a given c must go to zero. In addition, since the agent takes larger positions as his information becomes more precise, the equilibrium price must converge to the terminal value, v. Conversely, as the agent becomes less informed and $\sigma_s \to \infty$, the probability that he is informed approaches π, the proportion of informed agents, but, as $\sigma_s \to \infty$, all agents are equally uninformed. At less extreme positions, though, when informed agents are not all that precisely informed, the posterior probability is not all that different from the prior probability. This simply verifies that a single observation of performance is not terribly informative about precisions.

A central result of this section is that the equilibrium can be expected to be a pooling equilibrium where all agents are perceived to behave in a similar fashion and uninformed agents mimic informed ones. Together with Theorem 5, we can now expect that in financial markets in equilibrium a substantial percentage and possibly even a majority of agents could be truly uninformed. Having said that, though, at the second stage of our model, clients are not fully informed about the precisions of their agents, and it must be noted that we have not solved for the equilibrium in the second-stage market with unknown precisions. Nevertheless, as long as the reward for being perceived (exactly or in probability) to be of high precision is significantly greater than not being so perceived, the pooling equilibrium of this section will still hold.

CONCLUSION

Financial markets have become increasingly institutionalized, with individual participants holding assets through intermediaries such as mutual funds or pension plans that, in turn, make investment decisions for them. We began our analysis by showing that the noisy rational expectations equilibrium in which individual investors have information of varying degrees of precision is unstable to the formation of institutions in which agents with higher precisions manage assets for those with lower precisions. As the market becomes more institutionalized, with participants of known high precision becoming agents, they will come to dominate the market. Within limits, though, agents of a range of precisions can survive even when their precisions are known to be inferior because they provide a mechanism for diversifying across the information received by agents. Furthermore, markets dominated by institutions with higher precisions than individuals will exhibit lower risk premiums.

Another observation from this analysis is that the rewards from being perceived to have a high-precision signal (i.e., to being a "top" manager) are great. At this stage, we have assumed that the precisions of all agents are common knowledge. We then examined the situation at a prior stage when agents are untried and their precisions are known only by them. As potential agents look forward to a market where it is highly desirable to be perceived to be of high precision, they will have an incentive to engage in a wide variety of "mimicking" behaviors so as to appear to be highly informed even when they are not. The resulting equilibrium is a pooling equilibrium in which there is no definitive way to separate the uninformed from the truly informed. That being the case, all agents will appear to behave similarly, and it is only after the results are in that the market can update its perceptions of the agents. Furthermore, uninformed agents follow mimicking strategies that make the probability that they will appear to be informed significant and not all that different from the probabilities assigned by the market to agents who are truly informed.

Central to the result that it will generally be difficult to separate the informed from the uninformed is what is known in trading firms as the "trader's option." Traders who do well are very well rewarded as the posterior probability of their precision (i.e., the perception of their trading ability) rises, while, if they are poorly perceived, they always have a reasonable opportunity wage available to them. This call option-like feature is an important part of our model. Traders and agents implicitly, if not explicitly, hold call options on the market's perception of their abilities. This underlies their willingness to behave so as to alter the market's perceptions of them and undermines any hope of separating agents on the basis of their actions alone.

Coupled with the limited liability features of compensation, the trader's option also greatly limits the ability of institutions and markets to control agents through incentives. That being the case, the natural market response is to monitor and constrain the space of actions that agents can take. An important area of future research would be to explain the limits that institutions place on their employees and that markets place on institutions as equilibrium responses to the call option-like features that agents have. On a more narrow plane, there is much that remains to be done with the models of this chapter. In particular, we know very little about the equilibrium when participants assign different probabilities to agents' precisions based on their past performance and next to nothing about what occurs when agents are only imprecisely informed about their own abilities.

APPENDIX

LEMMA A1

If $x \sim N(0, \sigma^2)$, then

$$E\left[e^{-(\alpha x^2+\beta x+\gamma)}\right] = \left(1+2\alpha\sigma^2\right)^{-\frac{1}{2}} e^{\frac{\beta^2\sigma^2}{2\left(1+2\alpha\sigma^2\right)}-\gamma}$$

Proof:

Straightforward integration.

□

LEMMA A2

Let v and z be independent mean zero normals with variances σ_v^2 and σ_z^2, respectively. The following is true:

$$E\left[e^{-\left(\alpha_{vv}v^2+\alpha_{vz}vz+\alpha_{vv}z^2+\alpha_v v+\alpha_z z\right)}\right]=\phi^{-\frac{1}{2}}e^{\frac{\Psi}{2\phi}}$$

where

$$\phi = (1 + 2\alpha_{vv}\sigma_v^2)(1 + 2\alpha_{zz}\sigma_z^2) - \alpha_{vz}^2\ \sigma_v^2\sigma_z^2$$

and

$$\psi = \alpha_v^2\sigma_v^2\ (1 + 2\alpha_{zz}\sigma_z^2) + \alpha_z^2\sigma_z^2\ (1 + 2\alpha_{vv}\sigma_v^2) - 2\alpha_v\ \alpha_z\alpha_{vz}\sigma_v^2\sigma_z^2$$

Proof:

Holding z constant, integrate over v using Lemma A1. The result is an exponential quadratic in z, and integrating over z using Lemma A1 verifies the lemma.

□

For the next result, we will use the demeaned model in which the equilibrium price is given by

$$p = a_v v + a_z z$$

and the demand of the uniformed agent is given by

$$x=\frac{\beta}{A}p$$

and we will define

$$g_0=\sum\omega_j,\ g_1=\sum\frac{\omega_j}{\sigma_j^2},\ g_2=\sum\frac{\omega_j}{\sigma_j^2}s^j,\ g_3=\sum\omega_j\theta_j$$

THEOREM A1

Conditional on any set of precisions, $<\sigma_j^2>$,we have the following result:

$$E\left[e^{-Aw}\left|\left\langle\sigma_j^2\right\rangle\right.\right]=\left[\phi-\Psi\Omega\right]^{-\frac{1}{2}}e^{Aw_0g_3}$$

where ϕ and ψ are as defined in Lemma A2 and

$$\Omega=\sum\left(\frac{\omega_j^2}{\sigma_j^2}\right)$$

and

$$\alpha_{vv} = (1 - a_v)(\beta a_v g_0 + (1 - a_v)g_1)$$
$$\alpha_{vz} = \beta a_z (1 - 2a_v)g_0 - 2a_z (1 - a_v)g_1$$
$$\alpha_{zz} = -\beta a_z^2 g_0 + a_z^2 g_1$$
$$\alpha_v = (1 - a_v)$$
$$\alpha_z = - a_z$$

Proof:

The terminal wealth of a principal who invests in n funds, each of which puts x^j in the risky asset, is given by

$$w = \left(\sum \omega_j x^j\right)(v - p) - w_0 \sum \omega_j \theta_j$$

From the expressions for the demand x^j in the text, it is easy to see that, as the precision of agent j's information converges to zero, we obtain

$$x^j = x + \frac{1}{A\sigma_j^2}\left(v - p + s^j\right)$$

Thus,

$$\begin{aligned} w &= \left(\sum \omega_j x^j\right)(v - p) - w_0 \sum \omega_j \theta_j \\ &= \left[\left(\sum \omega_j\right)x + \frac{1}{A}\sum \frac{\omega_j}{\sigma_j^2}(v - p) + \frac{1}{A}\sum \frac{\omega_j}{\sigma_j^2} s^j\right](v - p) - w_o \sum \omega_j \theta_j \\ &= \frac{1}{A}\left[g_0 \beta (a_v v + a_z z) + g_1\left((1 - a_v)v - a_z z\right) + g_2\right]\left((1 - a_v)v - a_z z\right) - w_o \sum \omega_j \end{aligned}$$

This implies that

$$-Aw = -[\alpha_{vv} v^2 + \alpha_{vz} vz + \alpha_{zz} z^2 + (\alpha_v v + \alpha_z z)g_2] - w_0 g_3$$

We can now apply Lemma A2 to obtain

$$E\left[e^{-Aw} \middle| \left\langle s^j \right\rangle, \left\langle \sigma_j^2 \right\rangle\right] = \phi^{-\frac{1}{2}} e^{\frac{\psi}{2\phi} g_2^2} e^{A w_0 g_3}$$

Since the signals $\langle s^j \rangle$ are independent normals,

$$g_2 \sim N\left(0, \sum \frac{\omega_j^2}{\sigma_j^2}\right) = N(0, \Omega)$$

Applying Lemma A2 again, we now have

$$\begin{aligned} E[e^{-Aw} \mid \left\langle \sigma_j^2 \right\rangle] &= \mathrm{E}\left[\phi^{-\frac{1}{2}} e^{\frac{\psi}{2\phi} g_2^2} e^{A w_0 g_3} \middle| \left\langle \sigma_j^2 \right\rangle\right] \\ &= \phi^{-\frac{1}{2}} \left[\phi - \frac{\psi}{\phi}\Omega\right]^{-\frac{1}{2}} e^{A w_0 g_3} \\ &= [\phi - \psi\Omega]^{-\frac{1}{2}} e^{A w_0 g_3} \end{aligned}$$

□

In what follows, we will use, for the precision as the inverse of the variance, the notation

$$h_j = \frac{1}{\sigma_j^2}$$

Some algebra reveals that ψ can be simplified to

$$\psi = \alpha_v^2 \sigma_v^2 + \alpha_z^2 \sigma_z^2$$

and is therefore dependent only on the equilibrium price parameters and not directly on ω_j, the client's allocation to any fund j. Two additional useful, if algebraically tedious, relations are that

$$\beta = \frac{\alpha_v \sigma_v^2 - \psi}{a_z^2 \sigma_z^2 \sigma_v^2}$$

and that

$$\frac{\partial \phi}{\partial \omega_j} = 2\psi h_j + 2\beta\left[\alpha_v \sigma_v^2 - \psi\right](1 - g_0)$$

These observations together with some further algebra permit us to verify the following result.

THEOREM A2

The demand for fund j is given by

$$\omega_j = 1 - \frac{k}{h_j} - c\frac{\theta_j}{h_j}$$

where

$$k = q\left(\sum \omega_j - 1\right)$$

and

$$q = \frac{a_z^2 \sigma_z^2 \sigma_v^2 \beta^2}{\psi} > 0$$

and

$$c = Aw_0 \frac{1}{\psi}[\phi - \Omega\psi] > 0$$

Proof:

Differentiating the expression for expected utility in Theorem A1 and setting the result to zero provides the necessary first-order conditions,

$$\frac{E[e^{-Aw}]}{\partial \omega_j} = -\frac{1}{2}[\phi - \psi\Omega]^{-\frac{3}{2}} \frac{\partial[\phi - \psi\Omega]}{\partial \omega_j} e^{Aw_0 g_3} + [\phi - \psi\Omega]^{-\frac{1}{2}} e^{Aw_0 g_3} Aw_0$$

Since ψ is independent of ω_j and

$$\frac{\partial \Omega}{\partial \omega_j} = 2h_j\omega_j$$

rearranging the first-order conditions produces

$$\omega_j = -Aw_0\frac{1}{\psi}[\phi - \psi\Omega]\frac{\theta_j}{h_j} + \frac{1}{2\psi h_j}\frac{\partial \phi}{\partial \omega_j}$$

Differentiating ϕ with respect to ω_j verifies the desired result. The constant $c > 0$ from Theorem A2 shows that it has the sign of the expectation of a positive function, and $\omega > 0$ implies that $q > 0$.

□

LEMMA A3

If all $\theta_j = \theta$, then

$$k + c\theta = q\left(\sum \omega_j - 1\right) + c\theta > 0$$

Proof:

From the demand equations,

$$\sum \omega_j = n\left[\frac{1 + \frac{q}{h} - \frac{c\theta}{h}}{1 + n\frac{q}{h}}\right]$$

Hence,

$$q\left(\sum \omega_j - 1\right) + c\theta = \frac{q(n-1) + c\theta}{1 + n\frac{q}{n}} > 0$$

□

THEOREM A3

Ignoring the impact of the fee charged by fund j on the total invested in all funds, a viable fund manager achieves maximum profits by setting

$$\theta_j = \frac{h_j - k}{2c}$$

which produces profits of

$$m\frac{\left(h_j - k\right)^2}{4ch_j}$$

where m is the number of clients served.
Proof:

The profits of a viable fund manager are given by

$$m\omega_j\theta_j = m\left(1-\frac{k}{h_j}-c\frac{\theta_j}{h_j}\right)\theta_j$$

where m denotes the number of clients. Differentiating with respect to θ_j and setting the derivative equal to zero produces the desired result.

□

THEOREM A4

$$\text{Prob}\left[j\epsilon I|v,p,x^j=c\right]=\frac{\pi}{\pi+(1-\pi)\left(\frac{\sigma_s}{\sigma_y}\right)e^{\frac{\alpha^2-\beta^2}{2}}}$$

where

$$\pi\equiv\text{Prob}\left[j\epsilon I\right]$$

$$\alpha\equiv\frac{Ah^j(c-x)+p-v}{\sigma_s}$$

$$\beta\equiv\frac{(\sigma_s\alpha+v)}{\sigma_y}$$

Proof:

Letting lowercase "prob" denote probability density,

$$\text{Prob}\left[j\epsilon I|v,p,x^j=c\right]=\frac{\text{prob}\left[v,p,x^j=c|j\epsilon I\right]\pi}{\text{prob}\left[v,p,x^j=c\right]}$$

$$\begin{aligned}\text{prob}\left[v,p,x^j=c|j\epsilon I\right]\pi&=\text{prob}\left[x^j=c|v,p,j\epsilon I\right]\text{prob}\left[v,p|j\epsilon I\right]\\&=\text{Prob}\left[x+\frac{1}{Ah}(v+s-p)\right]\text{prob}\left[v,p\right]\\&=n(\alpha)\frac{Ah}{\sigma_s}\end{aligned}$$

where

$$\alpha\equiv\frac{Ah(c-x)+p-v}{\sigma_s}$$

Similarly,

$$\text{prob}\left[v,p,x^j=c|j\notin I\right]=n(\beta)\left(\frac{Ah}{\sigma_y}\right)\text{prob}\left[v,p\right]$$

where

$$\sigma_y = \left[\sigma_v^2 + \sigma_s^2\right]^{\frac{1}{2}}$$

and

$$\beta \equiv \frac{\sigma_s \alpha + v}{\sigma_y}$$

Hence

$$\begin{aligned}\mathrm{Prob}\left[j \epsilon I | v, p, x^j = c\right] &= \frac{\pi\left(\frac{n(\alpha)}{\sigma_s}\right)}{\pi\left(\frac{n(\alpha)}{\sigma_s}\right) + (1-\pi)\left(\frac{n(\beta)}{\sigma_y}\right)} \\ &= \frac{\pi}{\pi + (1-\pi)\left(\frac{\sigma_s n(\beta)}{\sigma_y n(\alpha)}\right)} \\ &= \frac{\pi}{\pi + (1-\pi)\left(\frac{\sigma_s}{\sigma_y}\right) e^{\frac{\alpha^2-\beta^2}{2}}}\end{aligned}$$

□

ACKNOWLEDGMENTS I am grateful to Bengt Holmström and Alan Schwartz for helpful comments, and, of course, to Fischer Black, who so profoundly changed my thinking about finance as to defy my ability to identify what is truly new from what is merely delayed recognition. Needless to say, any errors are my own doing.

REFERENCES

Admati (1985), "A Noisy Rational Expectations for Multi-Asset Securities Markets," *Journal of Business*, 53, 629–658.

Admati, Anat R., and Paul Pfleiderer (1986), "A Monopolistic Market for Information," *Journal of Economic Theory*, 39, 400–438.

Admati, Anat R., and Paul Pfleiderer (1987), "Viable Allocations of Information in Financial Markets," *Journal of Economic Theory*, 43, 76–115.

Admati, Anat R., and Paul Pfleiderer (1990), "Direct and Indirect Sale of Information," *Econometrica,* 58, 901–928.

Bhattacharya, Sudipto, and Paul Pfleiderer (1985), "Delegated Portfolio Management," *Journal of Economic Theory*, 36, 1–25.

Black, Fischer, and Jack Treynor (1973), "How to Use Security Analysis to Improve Portfolio Selection," *Journal of Business*, 46, 66–86.

Black, Fischer, and Robert Litterman (1991), "Asset Allocation: Combining Investor Views with Market Equilibrium," *Journal of Fixed Income,* 1, 7–18.

Black, Fischer, and Robert Litterman (1992), "Global Portfolio Optimization," *Financial Analysts Journal,* 48, 28–43.

Grossman, Sanford J. (1976), "On the Efficiency of Competitive Stock Markets Where Traders Have Diverse Information," *Journal of Finance*, 31, 573–585.
Hellwig, Martin F. (1980), "On the Aggregation of Information in Competitive Markets," *Journal of Economic Theory*, 22, 477–498.
Huberman, Gur, and Shmuel Kandel (1993), "On the Incentives for Money Managers," *European Economic Review,* 37, 1066–1081.
Pfleiderer, Paul (1984), "The Volume of Trade and the Variability of Prices: A Framework for Analysis in Noisy Rational Expectations Equilibria," Working paper, Stanford University.
Ross, Stephen A. (1977), "The Determination of Financial Structure: The Incentive-Signalling Approach," *The Bell Journal of Economics*, 8, 23–40.
Ross, Stephen A. (1978), "Some Notes on Financial Incentive-Signalling Models, Activity Choice and Risk Preferences," *Journal of Finance*, 33, 777–794.

6

Recovering Probabilities and Risk Aversion from Options Prices and Realized Returns

Mark Rubinstein & Jens Jackwerth

1.

Standard equilibrium models in financial economics are, in their essential nature, ways of determining *state-contingent prices*: the price today of a dollar to be received at only a specific date in the future and given a specific description of the state of the economy at that time. If there are no *riskless arbitrage opportunities*, each of these prices is positive. The sum of the state-contingent prices for dollars received at a single date over all possible states is the current price of a dollar received for sure at that date. This is one divided by the current *riskless return* for that date. Therefore, multiplying the state-contingent prices by this return converts them into a probability measure over the states, which financial economists call *risk-neutral probabilities* (figure 6.1). This chapter is largely about ways of recovering these probabilities from the current riskless return, the currently observed prices of traded assets, and the current prices of traded derivatives on those assets.

The usual way of applying the equilibrium model goes about this differently. It takes as given the subjective probabilities and risk preferences of an "average investor" and uses them to determine the risk-neutral probabilities.

The argument is that, ceteris paribus, a risk-neutral probability will be higher the higher the *subjective probability* of achieving its associated state: the probability measuring the investor's degree of belief that the corresponding state will occur. If the investor were indifferent to risk, then corresponding risk-neutral and subjective probabilities would be equal. However, the investor may value an extra dollar more highly in one state than another. For example, if he were risk averse, he would value an extra dollar more highly in states when, ceteris paribus, his wealth was relatively low. This motivates him to spread his wealth out evenly across states. However, aggregate economic uncertainty prevents this since the aggregate supply of dollars in all states is not the same. As a result, what he is willing to pay today for a dollar received tomorrow depends not only on

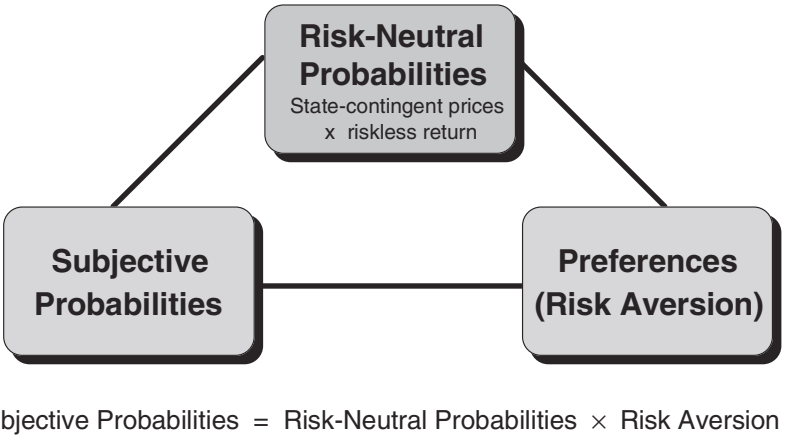


Figure 6.1. Risk-Neutral Probabilities: The Link between Probabilities and Preferences

his subjective probabilities but also on his degree of *risk aversion*. Risk-neutral probabilities, therefore, can be interpreted as subjective probabilities that are adjusted upward (downward) if they correspond to states in which dollars are more (less) highly valued (see figure 6.2).

In the standard approach, given the riskless return and having determined the state-contingent prices in this way, assuming *perfect markets*, traded securities are simply portfolios of state-contingent securities. Therefore, the value of traded securities can be easily calculated, and the model may be tested by comparing these values to quoted market prices. As a practical matter, the standard equilibrium model has been difficult to test

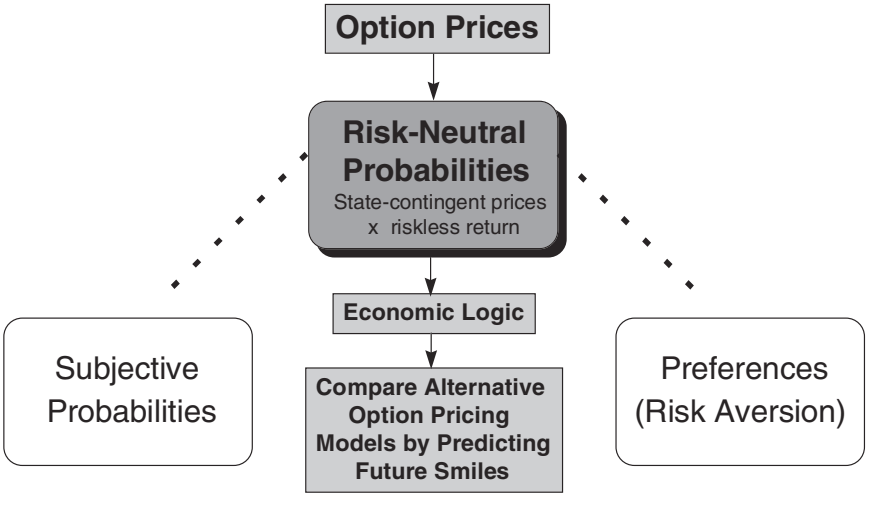


Figure 6.2. Step 1: Recover Risk-Neutral Probabilities From Option Prices

empirically because it has been difficult to identify the relevant subjective probabilities and risk aversion.

The approach of this chapter is to break this Gordian knot and determine the risk-neutral probabilities directly—and only then try to say something about how these probabilities decompose into subjective probabilities and risk aversion.

We take as data the current prices of traded options on a proxy for the *market portfolio* (see figure 6.3): the portfolio of assets that has the same proportionate payoffs across states as aggregate wealth. Our proxy is the portfolio measured by the Standard and Poor's 500 Index of common stocks (S&P 500 Index). Since a highly liquid market has existed for about a decade on a wide variety of different European puts and calls on the S&P 500 Index, it is tempting to take advantage of this comparatively recent development in financial markets. Admittedly, this is an incomplete and probably biased proxy, and some, though not all, of our results may be affected by this.

We begin by discussing methods of recovering risk-neutral probabilities from the concurrent prices of these options (along with the concurrent level of the index and the riskless return). If these prices were set according to the *Black–Scholes formula*, our task would be a simple one (Black and Scholes 1973). In that case, the entire risk-neutral probability distribution could be summarized by its volatility (its mean must equal the riskless return). Unfortunately, since the stock market crash of 1987, the Black–Scholes formula fits the market prices of S&P 500 Index options very poorly, so we need to investigate other methods of recovering these probabilities from market prices. If European options expiring on the target expiration date existed on the Index spanning all possible strike prices from zero to infinity, then (ignoring trading costs) the simultaneously observed prices

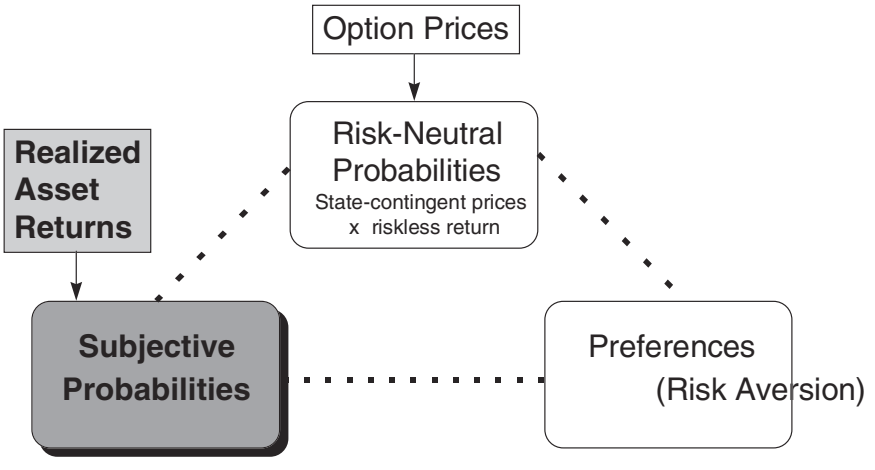


Figure 6.3. Step 2: Use Realized Asset Returns as a Proxy for Subjective Probabilities

of these options would uniquely determine the risk-neutral probability distribution (Breeden and Litzenberger 1978).

Of course, such a complete set of options does not currently exist. In practice, strike prices are set at discrete intervals, and there is a lowest (highest) strike price significantly greater (less) than zero (infinity). This opens the recovery problem to different possible methodologies. We consider a number of possibilities, including quadratic optimization and the method of maximizing smoothness (Jackwerth and Rubinstein 1996). Because of the richness of the market for S&P 500 Index options, the most important properties of the recovered distribution—its extreme leptokurtosis (peakedness) and left skewness—are not sensitive to the particular methodology. While all methods tested result in much more probability in the lower left tail than in the lognormal, because there are few options with strike prices covering that region, the exact distribution of this greater probability in this region is sensitive to the methodology chosen. For example, whether the distribution contains another mode in this region can depend on the recovery methodology.

We might hope to recover even more information from option prices; in particular, the stochastic process followed by the underlying index price. Unfortunately, given the recovered risk-neutral probability distribution for a given expiration date, there are an infinite number of possible stochastic processes that are consistent with these prices. To sort through these processes, we need to make additional assumptions. We design a model that is as close as possible to the standard binomial option-pricing model while allowing options to be valued under an arbitrary prespecified risk-neutral probability distribution applying to a single given expiration date that corresponds to the ending nodes of a recombining binomial tree. We call the resulting stochastic process an *implied binomial tree* (Rubinstein 1994). The strongest assumption we make initially (also a property of standard binomial trees) is that all paths leading to the same ending node have the same risk-neutral probability. Applying this model to postcrash option prices, produces a tree of local (or one move) volatilities with the following general features:

- on a given date prior to expiration, local volatilities are higher the lower the level of the underlying index;
- for a given change from the initial underlying index price, the faster it occurs, the greater the change in the local volatility;
- for index levels near the initial underlying index price, the farther into the future the local volatility, the lower it tends to be.
- One line of research has been to drop the assumption that all paths leading to the same ending node have the same risk-neutral probability. Fortunately, the model can be generalized by adding path-weighting parameters that can be calibrated so that the generalized implied binomial tree now also fits the prices of options that expire on earlier dates. (Jackwerth 1997a)

Stepping away from the purely modeling problems, we ask what fundamental features of the economy could create the recovered risk-neutral distribution and implied binomial tree. We provide four potential explanations. A goal of future research will be to find some way of determining what combination of these explanations actually underlies the observed phenomena. With this in hand, we will have a much deeper understanding than we now have of the economic forces that determine security prices, and we will be able to anticipate the effects on security prices of structural economic changes.

While the recovered risk-neutral probability distribution for a given expiration date is quite robust to our assumptions, this is not true for the implied binomial tree (which requires a much stronger set of assumptions). Fortunately, an implied tree has several empirical implications that are amenable to empirical tests (Jackwerth and Rubinstein, 2001). Most important of these is the prediction of future Black–Scholes *implied volatility smiles* given the corresponding future underlying index price. Since other option-pricing models also can be interpreted as making this kind of forecast, we have an opportunity not only to test the validity of implied binomial trees but also to compare their predictive power to that of other popular option-pricing models or cruder smile prediction techniques used in practice ("naïve trader" models). We find that despite the greater sophistication of "academic" approaches, a very crude rule of thumb used in practice produces the best predictions in our postcrash empirical sample. However, while as expected the Black–Scholes formula does very poorly, a *CEV model* and implied binomial trees only do a little worse than the best naïve trader model.

Relying only on our robust approach to estimate expiration-date risk-neutral distributions, we then try to break these risk-neutral probabilities apart into a product of subjective probabilities and risk aversion (see figure 6.4) (Jackwerth, 2000). We measure subjective probabilities using the traditional technique of historical frequency distributions. In the past, the two key problems with this kind of inference have been, first, estimating the mean of the subjective probability distribution (since the mean of the realized frequency distribution is highly unstable), and, second, the difficulty of ascertaining the shape of the tails. Fortunately, we show that our conclusions about inferred market-wide risk aversion need rely only on information about the shape of the subjective distribution near its mean, wherever that mean may be.

Unfortunately, the logic of the model breaks down, implying for example that in aggregate the market actually prefers risk, or at best has increasing absolute risk aversion. We then consider a number of explanations for this implausible result. The most disturbing of these is that the index options market is highly inefficient. We test this hypothesis by following a postcrash investment strategy where we accumulate profits by rolling over a sequence of out-of-the-money puts and find that this strategy leads to highly excessive risk-adjusted excess returns even if we adopt general risk adjustments that account for the utility benefits of positive skewness and even if we in-

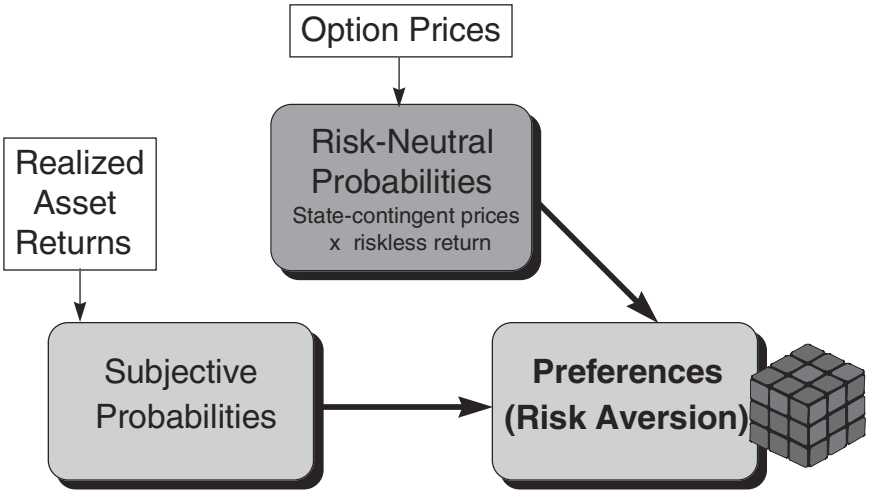


Figure 6.4. Step 3: Recover Risk-Aversion from Risk-Neutral and Subjective Probabilities

ject frequent crashes of the October 1987 magnitude into the historically realized index returns.

The research reported in this chapter summarizes a four-year effort, some published and some still in unfinished working paper form, including Jackwerth (1997a, b, c) and Jackwerth and Rubinstein (1996, 2001). To pursue this in more detail, it will be necessary to look at those papers.

2. THE PROBLEM

The interest in this research arises because the popular approach of explaining option prices—the Black–Scholes formula—fails miserably as an explanation of postcrash U.S. index option prices (as well as postcrash index option prices in several other countries). This anomaly stands out since the formula works much better in explaining the prices of most individual stock and foreign currency options.

Figure 6.5 shows the implied volatility smile for 164-day S&P 500 Index options traded on the Chicago Board Options Exchange on July 1, 1987 at 8:59 A.M. Central Time. If the Black–Scholes formula were true for these options, the smile should be perfectly flat. There can be only one risk-neutral probability distribution for the underlying index behind these options (since all the options are on the same underlying index and are only exercisable on the same date). Black–Scholes assumes that this distribution is lognormal, with its two free parameters, mean and variance, fully determined by the riskless return and implied volatility.

As seen in figure 6.5, the smile is remarkably flat, well within the bounds of realistic trading costs. So in this precrash period, the Black–Scholes formula appears to be doing extremely well, justifying its reputation as the

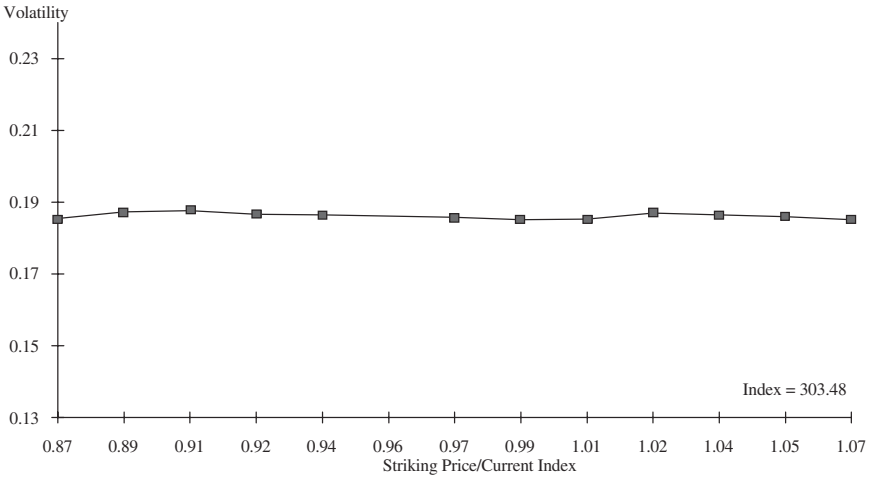


Figure 6.5. Typical 164-Day Pre-Crash Smile Implied Combined Volatilities of S&P 500 Index Options July 1, 1987 (8:59 A.M.)

last word in option pricing.[1] Moreover, this time can be shown to be typical for these options priced before the 1987 stock market crash.

In stark contrast, after the stock market crash, a very steep smile developed in the S&P 500 Index option market, roughly similar to figure 6.6, from mid-1988 to the present. This smile betrays an extreme departure from the predictions of the Black–Scholes formula. One way to place a lower bound on this departure is to select as the implied volatility in the Black–Scholes formula the volatility that minimizes the largest dollar or percentage error of a single option price over the set of all available options. This gives Black–Scholes the full benefit of the doubt. However, even if we do this, one of the options will have a pricing error of about $4.00, or one will have a pricing error of 15%. Such errors are probably well beyond the range that could be created by realistic trading costs. In any event, it is difficult to believe that changes in trading costs could account for the change in the smile across the divide of the stock market crash.

Options with shorter time-to-expiration are more liquid. Had we chosen these, the smile would have been even steeper, implying even greater departure from Black–Scholes predictions than for the 164-day options in figure 6.6.

This pricing deviation from Black–Scholes is striking for several reasons:

- it has existed more or less continuously over a 10-year period;
- it resides in one of the most liquid and active option markets, with a very large open interest;
- it is found in a market that, one might argue on theoretical grounds, is most likely to be the one for which the Black–Scholes formula works best.[2]

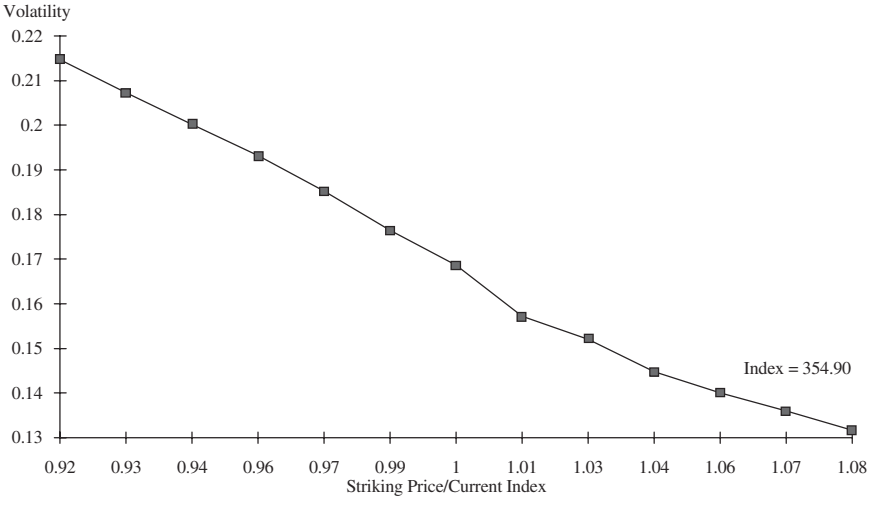


Figure 6.6. Typical 164-Day Pre-Crash Smile Implied Combined Volatilities of S&P 500 Index Options: January 2, 1990 (11:07 A.M.)

This situation just cries out for an alternative way to approach option pricing.

3. RECOVERING RISK-NEUTRAL PROBABILITY DISTRIBUTIONS

One possibility is to let the option prices speak for themselves. In contrast with Black–Scholes, the approach advocated here is nonparametric in the sense that any risk-neutral probability distribution could result. Instead, Black–Scholes begins by assuming that the risk-neutral distribution must be lognormal; the only question remaining is what its volatility is (its mean is anchored to the riskless return).

However, whatever methodology is selected should satisfy the following properties:

- As the number of available options with different strike prices becomes more dense, or spans a larger range, the methodology should result in a recovered distribution that is closer in a useful sense to the unique distribution recovered from a complete set of options.
- If the methodology uses a *prior distribution* as an input, and if the option prices can be explained by this distribution, the *recovered* (*posterior*) *distribution* should be the same.
- If any buy-and-hold arbitrage opportunities exist among the options, the underlying asset, and cash, the methodology should fail to recover any distribution.
- If option prices were determined by the Black–Scholes formula, the recovered distribution should be lognormal.

In table 6.1, we start by making a prior guess of the implied risk-neutral distribution, P_j', over all possible levels of the underlying asset price S_j at expiration, $j = 0, 1, 2, \ldots, n$. Also assumed known are the current bid and ask underlying asset prices, S^b and S^a, the current bid and ask prices of associated call options, C_i^b and C_i^a, with strike prices K_i, $i = 1, 2, 3, \ldots, m$, all with the same time-to-expiration, t, the current annualized return on a riskless zero-coupon bond maturing on the expiration date r, and the current annualized payout return on the underlying asset through the expiration date d.

The problem is to determine from this information the posterior risk-neutral probabilities P_j, which explain the current prices of the options as well as the underlying asset. The first constraints in table 6.1, $\Sigma_j P_j = 1$ and $P_j > 0$, assure that the P_j will indeed be probabilities. The second constraints, $S^b \le S \le S^a$ and $S = d^t \Sigma_j P_j S_j / r^t$, assure that the current value placed on the underlying asset, S, is the discounted expected value of its future possible prices using the posterior risk-neutral probabilities after adjusting for payouts and that this value lies between the market bid and ask prices.

The third constraints, $C_i^b \le C_i \le C_i^a$ and $C_i = \Sigma_j P_j \max[0, S_j - K_i] / r^t$, assure that the current value placed on the calls, C_i, is the discounted expected value of their possible future payoffs using the posterior risk-neutral probabilities and that this value lies between the market bid and ask prices.

Among all the posterior risk-neutral probability distributions that satisfy these constraints, the distribution chosen by this methodology is the one that is "closest" to the prior distribution in the sense of minimizing the average squared distance between these two probability distributions.

While there is some arbitrariness created by the assumed prior distribution and the quadratic measure of closeness, the method does satisfy the previous four properties claimed to be desirable for any technique for recovering risk-neutral probabilities from options.[3]

Table 6.1. Recovering risk-neutral probabilities: Optimization method.

$$\min \Sigma_j (P_j - P_j')^2 \text{ subject to:}$$

$$P_j$$

$$\Sigma_j P_j = 1 \text{ and } P_j \ge 0 \text{ for } j = 0, \ldots, n$$

$$S^b \le S \le S^a, \text{ where } S = (d^t \Sigma_j P_j S_j) / r^t$$

$$C_i^b \le C_i \le C_i^a, \text{ where } C_i = (\Sigma_j P_j \max[0, S_j - K_i])\, r^t \text{ for } i = 1, \ldots, m$$

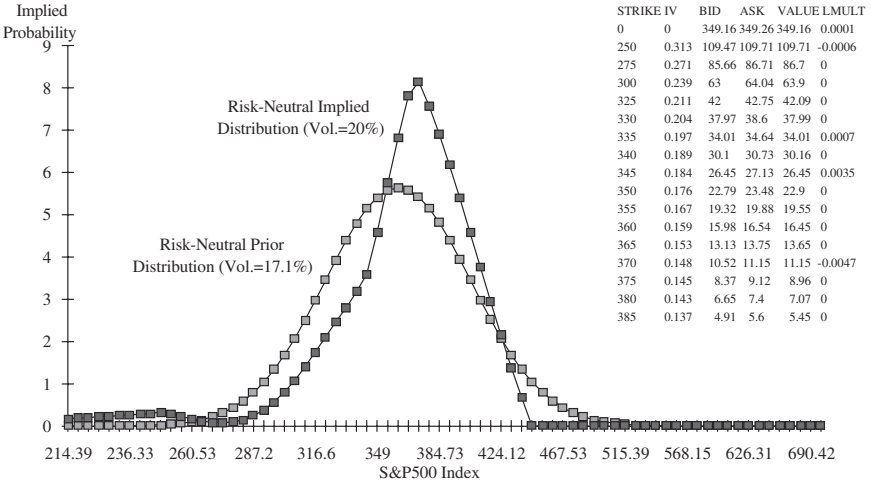
j indexes the ending binomial nodes from lowest to highest.
$P_j \equiv$ implied (posterior) ending nodal risk-neutral probabilities.
$P_j' \equiv$ prespecified (prior) ending nodal lognormal risk-neutral prohahi1ities.
$S_j \equiv$ underlying (ex-payout) asset prices at end of standard binomial tree.
S^b (S^a) $\equiv$ current *observed* bid (ask) underlying asset price.
C_i^b (C_i^a) $\equiv$ current *observed* bid (ask) call option price with striking price K_i.
$d \equiv$ *observed* annualized payout return.
$r \equiv$ *observed* annualized riskless return.
$t \equiv$ time to expiration.

Figure 6.7 shows a typical postcrash distribution recovered by this method. The distribution is based on the simultaneously observed bid and ask prices of 16 164-day European S&P 500 Index options with strike prices ranging from 250 to 385 and a current index level of 349.16 on January 2, 1990. This information closely matches the postcrash smile reported earlier.

The lighter-colored distribution is the one we would expect from Black–Scholes using the at-the-money options to determine the single implied volatility (17.1%) applied to all the options. It is derived by taking logarithms of returns to be a normal distribution. In contrast, the darker-colored distribution is the recovered posterior distribution. Even though this distribution was in a sense prejudiced to come up lognormal (since the prior was lognormal), its shape is markedly different, showing significant left skewness, much higher leptokurtosis, and slight bimodality. Perhaps the key feature is the much larger concentration of probability in the lower left-hand tail.

While we don't present the detailed evidence here, it turns out that these features of the recovered distribution are continuously displayed from about mid-1988 to the present in this market. On the other hand, prior to October 1987, the two distributions are nearly indistinguishable. The crash, then, marks a divide in the pricing of S&P 500 Index options. Evidence now available on smiles for other U.S. index options and for options on foreign stock market indexes is confirmatory: the features observed here for risk-neutral distributions carry over to other equity index options (Gemmill and Kamiyama 1997).

With enough options, the methodology we have used for recovering probabilities becomes insensitive to our choice of prior or our choice of the



STRIKE	IV	BID	ASK	VALUE	LMULT
0	0	349.16	349.26	349.16	0.0001
250	0.313	109.47	109.71	109.71	-0.0006
275	0.271	85.66	86.71	86.7	0
300	0.239	63	64.04	63.9	0
325	0.211	42	42.75	42.09	0
330	0.204	37.97	38.6	37.99	0
335	0.197	34.01	34.64	34.01	0.0007
340	0.189	30.1	30.73	30.16	0
345	0.184	26.45	27.13	26.45	0.0035
350	0.176	22.79	23.48	22.9	0
355	0.167	19.32	19.88	19.55	0
360	0.159	15.98	16.54	16.45	0
365	0.153	13.13	13.75	13.65	0
370	0.148	10.52	11.15	11.15	-0.0047
375	0.145	8.37	9.12	8.96	0
380	0.143	6.65	7.4	7.07	0
385	0.137	4.91	5.6	5.45	0

Figure 6.7. Prior and Implied Risk-Neutral 164-Day Probabilities S&P 500 Index Options: January 2, 1990 (11:00 A.M.)

quadratic measure of closeness. In effect, the recovered distribution becomes driven solely by the constraints.

To test the robustness of the approach with the number and span of options usually available for S&P 500 Index options, we tried alternative optimization criteria besides the quadratic (see table 6.2). Alternative criteria that could replace min $\Sigma_j (P_j - P_j')^2$ include:

- goodness of fit: min $\Sigma_j (P_j - P_j')^2 / P_j'$
- absolute difference: min $\Sigma_j | P_j - P_j |$
- maximum entropy: min $-\Sigma j\ P_j' \log(P_j/P_j')$
- maximum smoothness: min $\Sigma_j (P_j - 1 - 2P_j + P_j + 1)^2$

Each of these has its own rationale. The goodness of fit criterion places greater weight on states with lower probabilities; the absolute difference criterion places less weight on the most extreme differences between priors and posteriors. Perhaps, from a purely theoretical standpoint, the maximum entropy criterion is superior since it selects the posterior that has the highest probability of being correct given the prior. The maximum smoothness criterion, similar to fitting a cubic spline, minimizes the sum of the square of the second derivative $\partial^2 P/\partial S^2$ over the entire probability distribution. The expression $P_{j-1} - 2P_j + P_{j+1}$ is a finite-difference approximation for this second derivative. Note that this last criterion does not rely on a prior.

In practice, although the maximum entropy criterion may be best in theory, it is difficult to apply. In contrast, using the maximum smoothness criterion almost permits the problem to be transformed into solving a set of triangular linear equations and so produces very quick and reliable solutions. In any event, in the region between the lowest and highest strike prices, all the optimization criteria result in almost the same recovered probability distributions. In each case, the distribution is also heavily skewed to the left (postcrash). However, while all approaches agree that the recovered distribution has much more probability in the lower left tail than in the normal (postcrash), they disagree about how that probability is distributed in that tail. For example, one approach may produce slight bimodality while another may not.

Table 6.2. Recovering risk-neutral probabilities: Alternative nonparametric methods.

Basic method
Smile interpolation method
Optimization methods:
• Quadratic: $\Sigma j\ (P_j - P_j')^2$
• Goodness of fit: $\Sigma j\ (P_j - P_j')^2/P_j$
• Absolute difference: $\Sigma j\ \lvert P_j - P_j' \rvert$
• Maximum entropy: $-\Sigma j\ P_j \log (\boldsymbol{P}_j/P_j')$
• Smoothness: $\Sigma j\ (P_{j-1} - 2P_j + P_{j+1})^2$ [no prior]

For those methods that require priors, again it turns out that, at least for S&P 500 Index options, available strike prices are sufficiently dense that the implied risk-neutral distribution is not particularly sensitive to the imposition of a uniform in place of a lognormal prior.

4. RECOVERING RISK-NEUTRAL STOCHASTIC PROCESSES

As we indicated earlier, obtaining a good estimate of the risk-neutral probability distribution at the expiration date is only part of the story. We also want to recover the stochastic process that leads to this distribution. In a discrete version of the Black–Scholes model, this can be described by a recombining binomial tree with constant multiplicative up and down moves, and constant riskless and payout returns. After a sequence of these moves, the probabilities at the end of the tree can be made to approximate closely a risk-neutral lognormal distribution with a prespecified volatility and mean (Cox, Ross, and Rubinstein 1979). However, if the target risk-neutral distribution departs significantly from lognormal, as we have seen for the postcrash index option market, this simple binomial stochastic process must perforce be inconsistent with this.

So one might ask whether there is a way to modify the binomial model that leaves its major advantages—its intuitive simplicity and numerical tractability—intact but at the same time is consistent with the actual recovered risk-neutral distribution. It turns out that this can be done even while retaining the main attractive features of the the binomial approach:

- binomial price moves,
- recombining nodes,
- ending nodal values organized from lowest to highest,
- constant riskless and payout returns, and
- all paths leading to the same ending node having the same risk-neutral probability.

This last feature means that if you stand at a node at the end of the tree and look backward, you will see many paths from the beginning of the tree that lead to that node. Each of these paths has the same probability. This does not mean that all paths in the tree have the same probability but that, *conditional on ending up at a particular terminal node*, the paths have the same probability.

However, in an important way, the modified binomial tree differs from the standard tree: *it does not require constant move sizes*. It allows the local volatility of the underlying asset return to vary with changes in the underlying asset price and time. In addition, it can be shown that, given the ending risk-neutral distribution, the riskless and payout returns, and with the assumptions in table 6.3, there is a unique, consistent binomial tree that, moreover, preserves the property that there are no arbitrage opportunities in the interior of the implied tree (all risk-neutral move probabilities, although they may be different at each node, are nonnegative).

Table 6.3. Implied binomial trees: Assumptions.

Objective: value options for arbitrary risk-neutral expiration date probability distributions

- Underlying asset follows binomial process
- Binomial tree is recombining
- Ending nodal values ordered from lowest to highest
- Riskless (and payout) return constant
- All paths leading to the same ending node have the same risk-neutral probability

New Objective: generalize fifth assumption but retain the simplicity of the recursive solution process

Perhaps the most undesirable feature of this modified binomial approach—even though it is shared with the standard binomial approach—is the assumption that all paths leading to the same ending node have the same risk-neutral probability. Fortunately, it can be shown that this last assumption can be dropped, and the implied tree can be tractably designed to fit simultaneously options on the same underlying asset but with different times-to-expiration as well as different strike prices.

The modified binomial approach can be used to imply the stochastic process for S&P 500 Index options on January 2, 1990 with 164 days-to-expiration. Instead of depicting the resulting process in the usual way as a tree of up and down moves, it is perhaps more instructive to depict the tree in terms of the evolution of implied volatility, as in table 6.4. The volatility shown here is actually the annualized volatility from its associated node to the end of the tree, called the "global volatility." It turns out that

Table 6.4. Annualized global volatility structure for January 2, 1990 (10:00 A.M.) based on S&P 500 Index June call and put options maturing in 164 days.

S&P 500 Index	Days into the Future						
	0	12	32	47	61	76	91
397			9.3	9.9	10.4	11.1	11.9
386			10.8	11.1	11.8	12.4	13.0
376		11.8	12.1	12.5	12.9	13.4	13.9
365		14.3	14.0	14.0	14.1	14.5	14.8
355[a]	20.0	18.8	17.1	16.2	15.8	15.7	15.9
338		26.6	24.1	22.1	20.4	18.8	17.8
318		34.0	31.5	29.5	27.5	25.1	22.5
297		38.6	36.4	34.4	32.6	30.3	28.2
273			38.7	36.7	34.5	32.2	30.2

[a]S&P 500 Index reported at 10:00 A.M. is 354.75.

this will be similar to the Black–Scholes implied volatility for an option that is at-the-money at that node.

As we can see from the tree, the global volatility starts at 20% on January 2, when the options have 164 days-to-expiration. If the index falls 16% to 297 over the next 12 days (so the options now have only 152 days-to-expiration), the volatility almost doubles to 38.6%. This may seem like an excessive increase in volatility, but something like this happened during the 1987 stock market crash. If this same fall were to take 91 days, then the volatility would only rise to 28.6%. On the other hand, if the index rises, the volatility falls. Also note that if the index ends up in 91 days at the same level as it started, at 355, then the volatility will fall to 15.9%. The implied binomial tree shows that one way to make sense out of the downward-sloping smile (or alternatively, the left skewness of the recovered probability distribution) in index options is to suppose that the implied volatility varies inversely with the underlying asset price.

It is important to realize that these predictions concerning volatility are all recovered from the January 2, 1990 prices of S&P 500 Index options. They embody predictions about future option prices and are therefore amenable to an empirical test. For example, if the predictions are accurate, when we move 12 days into the future, say to January 14, 1990, if the index is then at 297, at-the-money options should be priced in the market such that their implied volatility is about 38.6%. Of course, the world is much more complex than our model, but we still might hope that the model gives an unbiased, low-error-variance prediction of future implied volatility, conditional on the future underlying asset price and the time remaining to expiration. One of our tasks will be to check this out and to compare the predictions from this method of constructing implied binomial trees to the predictions from other approaches.

Recovery of the stochastic process through implied binomial trees, as we have seen, strongly suggests that at-the-money implied volatility should vary inversely with the underlying asset price. We can make a quick check of this prediction.

Table 6.5 shows the results of regressions that attempt to use the current at-the-money volatility, σ_t, and the log return over the next 14 trading days, $\log(S_{t+14}/S_t)$, to predict the at-the-money volatility 14 trading days into the future, σ_{t+14}. In general, the option used to calculate σ_t and the option used to calculate σ_{t+14} will not be the same since the option that is at-the-money after 14 days will generally change since the underlying asset price has changed. The time period for the regressions, April 2, 1986 (the first day the S&P 500 Index calls traded as European options) to November 11, 1994, is divided into two subperiods, precrash (April 2, 1986 to September 4, 1987) and postcrash (May 18, 1988 to November 11, 1994).

The regressions over the *precrash period* show that adding the 14-day return to the current volatility does little to improve the prediction of the future volatility. This fits with what we already know about this period. Index option smiles were almost flat, suggesting that the Black–Scholes

Table 6.5. Forecasting future ATM implied volatility.

$\sigma_t \equiv$ current ATM implied volatility

$\sigma_{t+14} \equiv$ future ATM implied volatility (14 days later)

S&P 500 Index Options

Precrash April 12, 1986 to September 14, 1987

$\sigma_{t+14} = a\sigma_t$:	$r^2 = .47$, $a = 1.0024$
$\sigma_{t+14} - \sigma_t = (b/n)\log(S_{t+14}/S_t)$:	$r^2 = .04$, $b = -0.6518$
$\sigma_{t+14} = a\sigma_t + (b/n)\log(S_{t+14}/S_t)$:	$r^2 = .49$, $a = 1.0058$, $b = -0.7327$

Postcrash May 18, 1988 to November 25, 1994

$\sigma_{t+14} = a\sigma_t$:	$r^2 = .82$, $a = 0.9899$
$\sigma_{t+14} - \sigma_t = (b/n\log(S_{t+14}/S_t)$:	$r^2 = .49$, $b = -4.0715$
$\sigma_{t+14} = a\sigma_t$: $(b/n)\log(S_{t+14}/S_t)$:	$r^2 = .91$, $a = 1.0007$, b = –4.0800

formula, based on a constant volatility, worked well during this period. However, over the *postcrash period*, the 14-day return variable improves the prediction considerably.

In both cases, the coefficient a in the regression $\sigma = a\sigma_{t+1} + \varepsilon$, being near one, indicates that σ_t by itself is an unbiased forecast of σ_{t+14}. Surely, this is to be expected. Interestingly, this independent variable did a much better job forecasting the volatility 14 days ahead in the postcrash period. A second series of regressions sees how much of the variance of the forecast error $(\sigma_{t+14} - \sigma_t)$ can be explained by $\log(S_{t+14}/S_t)$. Precrash, this variable was of little assistance in helping explain this error, while, in stark contrast, postcrash this variable was of considerable value, confirming the prediction of implied binomial trees. Indeed, postcrash, σ_t and $\log(S_{t+14}/S_t)$ taken together explain 91% of the variance in σ_{t+14}.

Jump movements, not contemplated by Black–Scholes, could also potentially explain the observed leftskewness of the risk-neutral probability distribution if it were supposed that downward jumps are much more likely than upward jumps. Since 1987, there has been some rough empirical evidence of this in the U.S. stock market. However, these jumps would not also explain the observed negative relation between volatility and index levels.

To recapitulate, so far we have identified a significant departure in postcrash S&P 500 Index option pricing from the Black–Scholes formula. We have shown that this translates into a left-skewed highly leptokurtotic risk-neutral distribution for the future underlying asset price. Using implied binomial trees, this further translates into a stochastic process for which the salient departure from Brownian motion is the inverse relation of implied volatility and the underlying asset price. Finally, we have verified that, as predicted, this inverse relation was only present marginally precrash but was much stronger postcrash.

Perhaps we should ask what might be the economic causes of this departure from Black–Scholes. We are aware of four explanations in the current literature (see table 6.6):

Leverage effect: When stock prices fall, the firm's debt-to-equity ratio in market value terms tends to rise since its denominator falls faster than its numerator. If returns from assets remain the same, the increasing debt-to-equity ratio magnifies the influence of return from assets on stock returns, thereby increasing volatility. Thus, indirectly through automatic changes in the debt-to-equity ratio, a fall in stock prices causes an increase in stock volatility. Not only should this affect smiles of individual stock options, but since index returns are a convex combination of constituent stock returns, a similar smile effect should be observed from index options.

Correlation effect: Suppose that when stock index prices fall, or fall significantly, individual stock returns become more highly correlated. Some empirical evidence supports this. For example, in the 1987 stock market crash, most stock markets around the world fell together. If this occurs, then with the attendant reduced advantage from diversification, volatility will rise.

Wealth effect: Suppose that when stock index prices fall, investors become noticeably less wealthy and because of this more risk averse, so that when the same type of information hits the market, they respond by buying more or selling more than they would have with higher stock prices. In turn, this causes stock prices to be more sensitive to news, and volatility increases.

Risk effect: This reverses the order of causality of the wealth effect. In this case, something exogenous happens to increase stock market risk. Because investors are risk averse, they demand a higher expected return to hold stock. Assuming unchanged expectations, this leads to a reduction of current stock market prices.

It may be that each of these effects has some truth. One way to disentangle them is to compare the smiles for individual equities with the smile

Table 6.6. Economic causes of negative skewness and higher downside implied volatility.

Leverage: as asset price falls ⇒ market value debt–equity ratio rises ⇒ volatility rises
Correlation: equities become more highly correlated in down markets ⇒ volatility rises
Wealth: as market falls ⇒ wealth falls ⇒ investors become more risk averse ⇒ same news leads to greater reaction and trading ⇒ volatility rises
Risk: as volatility rises ⇒ risk premium increases ⇒ market falls
To separate these potential causes, compare the volatility behavior of individual stocks vs. market index.
Jumps: market is more likely to jump down than up ⇒ explains negative skewness but not negative correlation between implied volatility and index level

for indexes. In the United States, postcrash, the S&P 500 Index smile is far more pronounced than the smiles observed for its constituent equities. This suggests that the leverage effect may be quite weak, lending increased weight to the three other possibilities (however, see Toft and Prucyk 1997).

5. EMPIRICAL TESTS OF ALTERNATIVE FORECASTS OF RISK-NEUTRAL PROBABILITIES

Although the implied binomial tree model correctly anticipates the negative relation between volatility and asset price, it is not the only option-pricing model with this implication. This motivates our tests of alternative option-pricing models. Earlier tests of option-pricing models often relied on estimates of volatility from historically realized returns. Unfortunately, we can not separate errors in volatility estimation from option formula errors, and it is very easy to err in volatility estimation. So, in our tests, we will be careful to avoid such a joint hypothesis, and the tests will not depend on historical volatility estimates. Another problem with many tests of option-pricing models is that they often rely on following the outcome of managing a sequentially revised portfolio (usually chosen to replicate an option payoff). This makes these results subject to questionable assumptions about transaction costs and errors in asset price measurement. Again, our tests will not rely on dynamic replication and so will also avoid these complications.

Table 6.7 contains four types of predictions from option-pricing models that can be used, without relying on historical volatility estimation or dynamic replication, to test the validity of these models. The first and simplest simply compares the differential pricing of options across strike prices with model predictions (Rubinstein 1985). Unfortunately, this simple test cannot be used for implied binomial trees because it takes these option prices as data and fits the stochastic process to them.

Another test is to compare the concurrent prices of otherwise identical options but with different times-to-expiration. This can be used to test implied binomial trees (but not the generalized version). One can construct an implied tree from long-term options. Then, this tree can be used to value options that mature earlier.

Table 6.7. Types of Comparisons of option prices.

Volatility-free testing of alternative option-pricing models.
Comparisons among the prices of otherwise identical options but:
with different strike prices: explain the current smile
with different times-to-expiration: predict shorter-term smile from concurrent longer-term smile
observed at different points in time: predict future conditional smile from current smile
with different underlying asset: index option smiles vs. individual option smiles

A more interesting and stronger test, and the one we will report here, is to construct a tree based on current option prices and then to use this tree to predict the future prices of these same options. Having constructed a tree, as the future evolves, we can think of ourselves as moving along a path in the tree. If we stop before the options expire at the then current node, we can now infer a subtree from the original tree that should govern the stochastic movement of the underlying asset from that node forward to the expiration date. Using this inferred subtree, we can then value the options at that node and compare these to the market prices observed at that time. Stated equivalently, we can use the subtree to calculate the predicted implied volatilities of each option (the smile) and compare these to the observed smile in the market.

A final test we have not yet performed is to compare smiles across different underlying assets, including a comparison of smiles for individual equity options with smiles for indexes.

We shall be comparing alternative option-pricing models. The empirical test we shall emphasize uses the current prices of options to parameterize the models. Then the parameterized model is used to predict future option smiles. We then compare the predicted future smile with the actual smile subsequently observed in the market, conditional on knowing the new underlying asset price. We will prefer the model with a predicted smile that is closest (in absolute difference) to the realized smile.

Table 6.8 illustrates this test concretely. It lists options maturing in 189 days that have a strike price to current asset price ratio ranging from .8748 (out-of-the money puts) to 1.0756 (out-of-the-money calls). For example, the option with strike price to current asset price ratio of 1.0039 currently has an implied volatility of .1586. Thirty days from now, the options will have 159 days-to-expiration. At that time, the underlying index rose from 1.000 to 1.024 (up 2.4%). Each option-pricing model will supply a different prediction about the implied volatility of that option at that time. Our task will be to compare those predictions.

Although not reported in detail here, the alternative option-pricing models were also parameterized using current prices for options expiring in 189 days, as well as current prices for options maturing in 91 days. Again, the parameterized models are then used to predict option smiles in 30 days. We have not reported these results because they were little changed from the results where only options maturing in 189 days were used.

We have compared nine alternative approaches to option pricing (see table 6.9). They can be grouped into four categories:

Standard benchmark model
  Black–Scholes model
"Naïve trader" models
  Relative smile model
  Absolute smile model

Table 6.8. Empirical test: Inferring future conditional smiles from current option prices.

	Implied Volatilities		
K/S	159-day options in 30 days	91-day options now	189-day options now
.8748	?	(.2492)	.2278
.8892	?	(.2403)	.2209
.9035	?	(.2294)	.2139
.9179	?	(.2192)	.2049
.9322	?	(.2087)	.1961
.9466	?	(.1953)	.1888
.9609	? Index	(.1840)	.1814 Index
.9752	? = 1.024	(.1720)	.1718 =1.000
.9896	? (.1472)	(.1607)	.1645
1.0039	?	(.1506)	.1586
1.0183	?	(.1398)	.1503
1.0326	?	(.1298)	.1436
1.0756	?	(.1070)	.1244

Models emphasizing a functional relationship between volatility and asset price
- Constant elasticity of variance diffusion: restricted
- Constant elasticity of variance diffusion: unrestricted
- Implied binomial trees
- Displaced diffusion model

Models emphasizing other deviations from Black–Scholes
- Jump diffusion
- Stochastic volatility

The Black–Scholes model is parameterized by setting the volatility parameter in the formula equal to the current at-the-money implied volatility. The prediction of the Black–Scholes model is that in the future all options will have that same implied volatility.

The "naïve trader" models are so named because they are simple rules of thumb commonly used by professionals. The relative smile model predicts that the future implied volatility of an option with strike price K when the underlying asset price is S_1 is the same as the current implied volatility of an option with a strike price equal to $K(S_0/S_1)$. In contrast, the absolute smile model predicts that the future implied volatility of an option with strike price K is the same as the current implied volatility of that option. For this model, it is as if for each option its current implied volatility stays pinned to it.

Table 6.9. Alternative option-pricing models.

Black–Scholes ("flat smile") [$(dS)/S = \mu dt + \sigma dz$]
future implied σs set equal to current at-the-money implied σ
Relative Smile
future implied σs set equal to current implied σs of options with *same K/S*
Absolute Smile
future implied σs set equal to current implied σs of options with *same K*
Constant Elasticity of Variance: Restricted [$(dS/S = \mu dt + \sigma' S^{\rho-1} dz$]
future (σ', ρ) set equal to best fitting current (σ', ρ) ($0 \le \rho \le 1$)
Constant Elasticity of Variance: Unrestricted
future (σ', ρ) set equal to best fitting current (σ', ρ) ($\rho \le 1$)
Implied Binomial Trees
future option prices derived from implied binomial tree fitting current option prices
Displaced Diffusion [$S_t = (\alpha e^y + (1-\alpha) r^t) S_0$]
future (σ, α) [= % in risky asset] set equal to best fitting current (σ, α)
Jump Diffusion [$dS/S = (\alpha - \lambda k)dt + \sigma dz + dq$]
future (σ, λ, k) set equal to best fitting current (σ, λ, k)
Stochastic Volatility [$(dS)/S = \mu dt + v(t)^{1/2} dz_1$, $dv(t) = \kappa(\theta - v(t)) + v(t)^{1/2} dz_2$]
future (σ, $v(0)$, κ, θ, ρ) set equal to best fitting current (σ, $v(0)$, κ, θ, ρ)

The restricted CEV model assumes that the local volatility of the underlying asset is $\sigma' S \rho^{-1}$, where $0 \le \rho \le 1$ and σ' are constants (Cox 1996). This model builds in directly an inverse relation between the local volatility and the underlying asset price S. The closer ρ is to 0, the stronger this relation; and as ρ gets close to 1, the model becomes identical to the Black–Scholes formula. A more general version of this model that allows for an even stronger inverse relation is what we call the unrestricted version since it only requires that $\rho \le 1$.

The displaced diffusion model is also based on the assumption that the volatility is a function of the underlying asset price (Rubinstein 1983). As it was originally developed for individual stock options, the source of this dependence arose from the risk composition of the firm's assets and its financial leverage. Indeed, in contrast to the CEV model, the displaced diffusion model actually permits the volatility to vary in the same direction as the underlying asset price if the asset composition effect is stronger than the leverage effect. But, we can anticipate that, given the observed

empirical inverse relation between asset price and volatility for both individual stocks and the index, the displaced diffusion model is likely to have no advantage over the CEV model in forecasting future implied volatilities (postcrash).

Many academics and professionals believe that diffusion-based option models that only allow the volatility to depend at most on the underlying asset price and time are too restrictive. Therefore, we also want to test models incorporating the two other key generalizations of the Black–Scholes formula, jump asset price movements and volatility, which can depend on other variables. We have therefore included Merton's jump-diffusion model (Merton 1976) and Heston's stochastic volatility model (Heston 1993).

Figure 6.8 illustrates the potential difference in estimated risk-neutral probability distributions of three of the alternative pricing models. It shows, as we have seen before, that the implied distribution is left-skewed, with much greater leptokurtosis than the lognormal. Notice that the unrestricted CEV model with a sufficiently low ρ parameter (about –4) fits the implied distribution reasonably well. However, the methods for inferring the implied stochastic process are different. In the CEV case, the ρ and σ' parameters determining the fit above are held fixed when the CEV formula is reapplied to value options in the future. The only changed inputs in the formula are S and t, whereas, in the implied tree approach, a more elaborate backward-recursive tree construction is used, followed by the inference of the future subtree. In particular, in contrast to the CEV model, the implied tree approach builds in a dependence of the local volatility not only on the underlying asset price but on time as well.

Nonetheless, because of the similarity between the two risk-neutral expiration date distributions and because, as it turns out, the time dependence

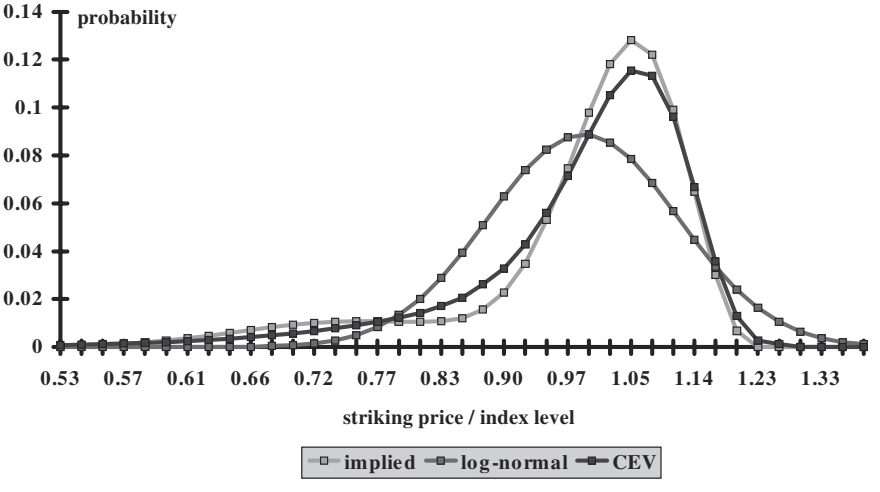


Figure 6.8. Lognormal, Implied and CEV Probability Distributions

of volatility appears slight, we can anticipate that the unrestricted CEV model will give results similar to the implied tree model in forecasting future implied volatility.

Recall that we will be comparing the forecasts of future option implied volatility. Alternatively, we will be using alternative option-pricing models, parameterized using current data, to forecast the prices of currently existing options at a specified future date before their expiration. That forecast will be conditional on knowing the underlying asset price at that future date since we are clearly not trying to forecast that as well. Having forecasted the future option prices using the Black–Scholes formula, we will translate those prices into the metric of implied volatilities, construct the implied volatility smile, and compare these predicted implied smiles across the different pricing models.

Figure 6.9 provides an illustration. The upper sloped line is the current smile, summarizing along with the current underlying asset price (indexed to 1.00) most of the information we need to make our predictions. Note that it is considerably downward-sloping, typifying the smiles for S&P 500 Index options in the postcrash period. The lower sloped line is the observed smile that was later observed at a specified future date when the index had risen to 1.0545. Not surprisingly, it too is downward-sloping since we remain within the postcrash period. The prediction of the Black–Scholes model, based on constant volatility for all strike prices and times, is described by the horizontal line. It simply says that the current at-the-money volatility of about 26% should continue to reign in the future for all the options.

The two other lines illustrate the predictions from our two "naïve trader" models. Since the horizontal axis is the ratio of the strike price to the underlying asset price, the relative smile model simply makes the prediction that the smile, scaled in terms of this ratio, will remain unchanged. So the sloped

Figure 6.9. Comparative Naïve Black–Scholes Predictions

line is at once the current smile and the prediction of the relative model of the future smile. In contrast, the absolute smile model predicts that options with the same striking price will have the same implied volatilities in the future that they have now. In the example, since the index moved up to 1.0545 in the future, the option that currently is at-the-money when the index is 1 will have in the future a strike price to index ratio of about .95. Since the option's implied volatility is currently about 26%, the model predicts it will continue to have that same implied volatility in the future. Thus, the future prediction is graphed by the ordered pair (.95, .26), which is indeed a point along the lower sloped line that lies just above the future smile containing the absolute model's prediction. In general, if the index increases, the absolute model predicts that the smile will fall, while if instead the index had fallen, the model would predict that the smile will rise.

Comparing the three predictions—Black–Scholes, relative, and absolute—it is easy to see in figure 6.9 that the absolute model has worked best since the bottom two lines lie close to one another.

For the precrash period sampling once per day, table 6.10 summarizes the average absolute errors between the realized future smile and the smile prediction from each model. For each model, two smile predictions are made, one 10 trading days in advance and the other 30 trading days in advance. For example, for the 10-day prediction, the Black–Scholes formula makes an average error of about 50 cents, and the median error across all the trading days is 39 cents. The median error for the 30-day prediction is about twice this, at 73 cents.

All the models perform about the same. But this is just what would have been expected since all models nest the Black–Scholes formula as a special

Table 6.10. Future versus current precrash pricing errors for S&P 500 Index options for the period April 2, 1986 to October 16, 1987 (365 observations). All models perform about the same.

	Error in Cents of 10(30)-Day Forecast		
Forecasting Method	Mean	Median	Std. Dev.
Black-Scholes	50	39 (73)	36
Relative smile	51	42 (72)	34
Absolute smile	52	42 (74)	35
CEV: restricted	49	40 (72)	34
CEV: unrestricted	50	40 (72)	34
Implied binomial trees	54	44 (69)	40
Displaced diffusion	50	40 (72)	34
Jump diffusion	49	39 (71)	34
Stochastic volatility	50	41 (72)	34

case and, as far as we can judge, the Black–Scholes formula worked quite well in this period. Even the relative and absolute models are special cases of Black–Scholes since if the current smile were flat, both the relative and absolute models would predict that the future smile would remain unchanged.

For the postcrash period, as expected, the Black–Scholes model works very poorly, with a median absolute error of $1.72 over a 10-day forecast period (see table 6.11). The jump-diffusion model does almost as poorly. Again, given a strongly downward-sloping smile, with the near-symmetric jump, up or down, of that model, we would not expect that it would offer much improvement. Smile patterns where the jump-diffusion model would help are weak smiles that turn up on both ends. Similarly, although the restricted CEV model can explain a downward-sloping smile, it can only explain a much weaker slope, so it also offers little improvement.

However, substantial improvement over Black–Scholes is offered by the relative smile model, the absolute smile model, the unrestricted CEV model, implied binomial trees, and Heston's stochastic volatility model. Of these, the best performing is the absolute model. It is ironic that the simplest predictive rule (apart from Black–Scholes) does the best: every option simply retains whatever Black–Scholes implied volatility it started with. This model is a considerable improvement over Black–Scholes, reducing the median 10-day error to $0.44, about one-fourth of the Black–Scholes error. The absolute smile model is also best over the longer 30-day prediction interval.[4]

Table 6.11. Future versus current postcrash pricing errors for S&P 500 Index options for the period June 1, 1988 to December 31, 1994 (1562 observations).

	Error in Cents of 10(30)-Day Forecast		
Forecasting Method	Mean	Median	Std. Dev.
Black–Scholes	175	172 (171)	58
Relative smile	73	55 (78)	61
Absolute smile	56	44 (63)	43
CEV: restricted	139	136 (145)	54
CEV: unrestricted	74	56 (78)	60
Implied binomial trees	83	67 (86)	62
Displaced diffusion	107	96 (112)	53
Jump diffusion	168	164 (166)	56
Stochastic volatility	75	57 (83)	67

Black–Scholes very poor at 172¢. Absolute smi1e best at 44¢, but relative smile, CEV unrestricted, implied binomial trees, and stochastic volatility not far behind.

We should not conclude from this that academic attempts to improve the Black–Scholes model—such as the CEV model, implied binomial trees, or the stochastic volatility model—have therefore failed. Rather, they do provide worthwhile improvements, cutting the Black–Scholes error to about one-third. But it is true that a "naïve trader" approach such as the absolute smile model, which has no academic foundations, does even better. This throws down a challenge to academic and professional theorists to explain why the absolute model should work so well.

Our fascination with the absolute smile model led us to decompose its remaining $0.44 error. We divided that error into three parts:

- the error in predicting the future at-the-money volatility;
- the error in predicting the implied volatility of other options, conditional on knowing the future at-the-money volatility;
- the error if, in addition, it is assumed that transactions can only take place at the bid–ask prices rather than at their midpoint.

Knowing the 10-day-ahead at-the-money volatility in advance, cuts the forecast error from $0.44 to $0.23, or even further to $0.14 if the error is measured relative to the bid–ask spread (see table 6.12). This suggests that one way to approach future research on this issue is first to explain the changes in at-the-money volatility since that alone can explain about half of the $0.44 error.

The success of the absolute model over the fancier academic models, including implied binomial trees, motivated us to test it directly in a time series analysis. In our previous time series analysis, we compared the implied volatilities of options that were at-the-money at the beginning and at the end of a 14-day trading interval. Table 6.13 compares the implied volatilities of the *same* options at the beginning and end of 14-day trading intervals.

In the precrash period, using the implied volatility at the beginning of the period explains about half of the variance in the implied volatility of the same option at the end of the period and, with a coefficient close to one, provides an almost unbiased forecast. Adding the 14-day logarithmic re-

Table 6.12. Future versus current: Summary.

Best prediction is absolute smile. But relative smile, CEV unrestricted, implied binomial trees, and stochastic volatility are not far behind.
Postcrash 10–day pricing errors for these models are about 1/3 to 1/4 of Black–Scholes or CEV restricted errors.
In general, knowing current short-term option prices in addition to long-term option prices doesn't seem to help.
For absolute smile, postcrash 10-day forecast errors based just on the current long-term option prices are 44¢. This is cut to 23¢ if, in addition, the future ATM option price is assumed known. This error is further cut to 14¢ if, in addition, errors are only measured outside the bid–ask spread.

Table 6.13. Naïve forecast of future implied volatility.

$\sigma_t \equiv$ current ATM implied volatility

$\sigma_{t+14} \equiv$ future implied volatility of *same* option (14 days later)

S& 500 Index Options

Precrash April 2, 1986 to September 4, 1987

$\sigma_{t+14} = a\sigma_t$:	$r^2 = .49$, $a = 1.0018$
$\sigma_{t+14} - \sigma_t = (b/n)\log(S_{t+14}/S_t)$:	$r^2 = .02$, $b = -0.4268$
$\sigma_{t+14} = a\sigma_t + (b/n)\log(S_{t+14}/S_t)$:	$r^2 = .50$, $a = 1.0040$, $b = -0.4813$

Postcrash May 18, 1988 to November 25, 1994

$\sigma_{t+14} = a\sigma_t$:	$r^2 = .91$, $a = 0.9709$
$\sigma_{t+14} - \sigma_t = (b/n)\log(S_{t+14}/S_t)$:	$r^2 = .05$, $b = -1.1705$
$\sigma_{t+14} = a\sigma_t + (b/n)\log(S_{t+14}/S_t)$:	$r^2 = .92$, $a = 0.9727$, b = –0.8960

turn does little to improve this forecast. Again, given how well the Black–Scholes model fits option prices during this period, this should come as no surprise.

In the postcrash period, the beginning implied volatility now explains a much greater percentage of the variance of the ending volatility (91%) and continues to be a nearly unbiased forecast. When we looked at *at-the-money* implied volatility comparisons previously, we found that adding the log 14-day return substantially improved the forecast in the postcrash period. But, if the regressions are recast in terms of predicting 14-day-ahead volatilities of the *same* options, then adding the log 14-day return offers almost no improvement in the forecast. This result is, of course, to be expected from our earlier analysis of comparative option-pricing models.

6. RECOVERING RISK AVERSION

What kind of a market would produce risk-neutral distributions so much at variance with the Black–Scholes predictions?

- One possibility is that postcrash the market dramatically changed the subjective probabilities it attached to the future performance of the S&P 500 Index.
- Another possibility is that the market postcrash became much more averse to downside risk.

If, following a time-honored tradition in financial economics, we measure the consensus market subjective probability distribution by its future realized frequency distribution, the result is the nearly normal curve in figure 6.10. Superimposed is the left-skewed risk-neutral distribution deduced by our techniques from March 15, 1990 S&P 500 Index option prices.

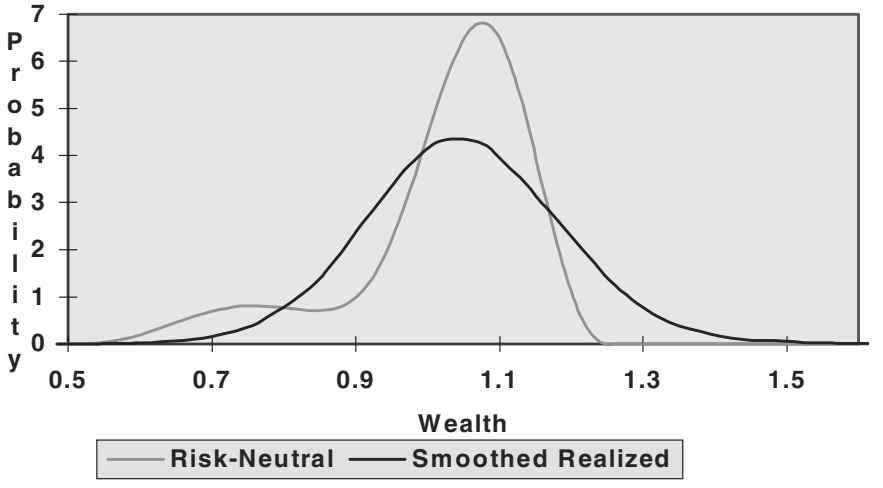


Figure 6.10. Risk Neutral vs. Realized Distributions (March 16, 1990)

The difference between the two distributions is striking. If we have measured the market subjective distribution accurately, then the shape of this distribution has not changed very much pre- and postcrash. So, we must look elsewhere for an explanation of the postcrash risk-neutral distribution, perhaps to changed risk aversion.

But before taking a look at this, we need to discuss an important objection. Using the realized frequency distribution, either drawn from realized prices prior to March 15, 1990 or from realized prices after March 15, 1990, time-honored though it may be, is a highly suspect measure of the subjective distribution that was actually in the minds of investors. In particular, if the market were anticipating an improbable but extreme event (such as a second crash) that had not yet been realized, it would not show up in our estimate of the subjective distribution. At the same time, these events may, despite their infrequency, be very important for understanding the pricing of options, particularly out-of-the-money puts.

Our way around this problem is to draw implications for market-wide risk aversion only from the comparative shapes of the realized and risk-neutral distributions around their means, without needing to consider the more questionable tails of these distributions. Around the means, it seems likely that the realized distribution provides a reasonably reliable approximation of the true subjective distribution in this region. In addition, our earlier analysis also shows that our techniques for estimating risk-neutral distributions from option prices are very robust around the means to alternative methods since available options are dense in this region.

Given the risk-neutral distribution, we can estimate the subjective distribution by imposing popular risk-aversion assumptions. We infer risk aversion using a simple but widely used model of financial equilibrium.

Table 6.14 defines the variables we will be using. Assuming a consensus investor, we maximize his expected utility $\Sigma_j Q_j U(R_j \delta^n)$ subject to a constraint anchoring his present wealth $(\Sigma_j P_j R_j)/(r/\delta)^n$ to 1. Choosing his portfolio of state-contingent claims is equivalent to choosing the returns R_j he will realize in each state j. (δ^n is a correction for R_j that is defined only to be the market portfolio return after payouts, so that $R_j\delta^n$ is the market portfolio's total return.)

As the optimal choice, we have the familiar first-order condition: $U'(R_j\delta^n) = \lambda(P_j/Q_j)/r^n$. Except for λ, this is a state-by-state restriction on the relation of risk aversion, subjective probabilities, and risk-neutral probabilities.

If we assume logarithmic utility so that $U'(R_j\delta^n) = \log(R_j\delta^n)$, then this first-order condition becomes

$$1/(R_j\delta^n) = \lambda(P_j/Q_j)/r^n$$

so that

$$Q_j = \lambda P_j R_j(\delta/r)^n$$

Summing over all j and since $\Sigma_j Q_j = 1$:

$$1 = \lambda(\Sigma_j P_j R_j)(\delta/r)^n$$

Since the investor is constrained so that $\Sigma_j P_j R_j(\delta/r)^n = 1$, then $\lambda = 1$. Substituting this into one of the equations above leads to the very simple decomposition of subjective probabilities

$$Q_j = (P_j/r^n)(R_j\delta^n)$$

so that the subjective probability of a state equals the state-contingent price for that state weighted by the total market return in that state.

Figure 6.11 shows the relation of subjective and risk-neutral probabilities for January 2, 1990 if we derive the subjective probability distribution not from past or future index realizations but from a simple model of financial equilibrium based on logarithmic utility and risk-neutral probabilities estimated from option prices.

Note how close the risk-neutral and subjective distributions are. The

Table 6.14. Notation.

$P_j \equiv$ risk-neutral probability for state $j = 1, \ldots, n$
$Q_j \equiv$ subjective probability for state j
$R_j \equiv$ market portfolio return (ex-payout) for state j
$\delta \equiv$ market portfolio payout return over a single period
$r \equiv$ riskless return over a single period
$U(R_j\delta^n) \equiv$ utility function of representative investor
$\underset{Rj}{\text{Max}}\ \Sigma_j Q_j U(R_j\delta^n) - \lambda\,[\Sigma_j P_j R_j)/r/\delta^n - 1]$
differentiating once: $U'(R_j\delta^n) = \lambda\,(P_j/Q_j)r^n$

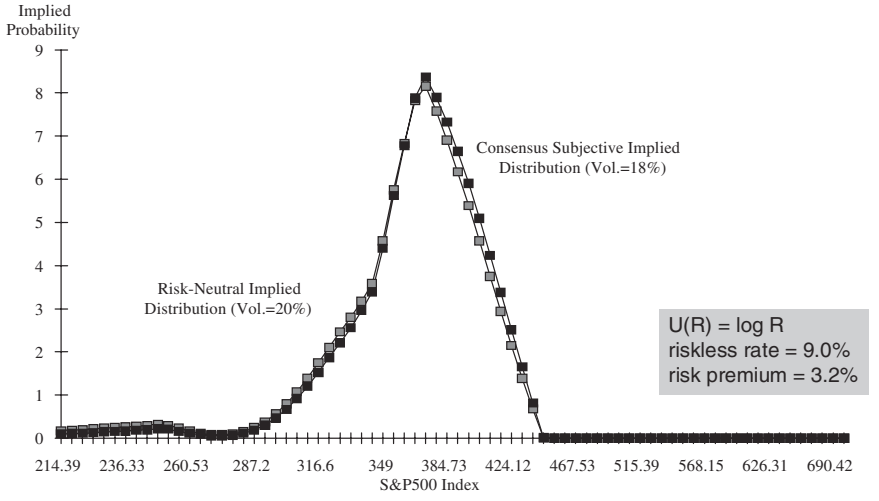


Figure 6.11. Subjective and Risk-Neutral 164-Day Probabilities S&P 500 Index Options: January 2, 1990 (11:00 A.M.)

main difference is that the subjective distribution is shifted to the right with a mean of 12.2%, in contrast to the 9% mean of the risk-neutral distribution. The risk-aversion property of logarithmic utility accounts for this shift. But the shapes of the two distributions are almost the same.

This contrasts sharply with our previous comparison of subjective and risk-neutral distributions, where the risk-neutral distribution was estimated in the same way, but the subjective distribution was estimated from realized index prices. Clearly, a simple model of equilibrium with logarithmic utility does not explain this disparity. So we now ask what consensus utility function could simultaneously rationalize these two distributions.

A trick to this comparison is to differentiate the general first-order condition a second time to obtain another condition that needs to hold in equilibrium:

$$-U''(R_j\delta^n)/U'(R_j\delta^n) = (\delta^{-n})[(Q'_j/Q_j) - (P'_j/P_j)]$$

Q'_j (P'_j) is the change in the subjective (risk-neutral) distribution across the nearby state. For example, $Q'_j = \partial Q_j/\partial S_j$ and is approximated by $(Q_{j+1} - Q_{j-1})/(S_{j+1} - S_{j-1})$. This has the advantage of being a true state-by-state condition, where λ has been eliminated. This permits us to examine only the states near the mean, in which we have the greatest confidence of our estimate of the subjective distribution inferred from realizations. In particular, we can determine the utility function fit in this region without needing to estimate the shape of the tail probabilities, in which we have very little confidence (see table 6.15). This condition also conveniently isolates the measure of absolute risk aversion on its left-hand side.

With our equilibrium result for absolute risk aversion in hand, the risk-neutral and subjective distributions were estimated following the

Table 6.15. Equilibrium preference probability relation.

$$\max_{Rj} \Sigma_j Q_j U(R_j \delta^n) - \lambda \left[\Sigma_j P_j R_j)/(r/\delta)^n - 1\right]$$

differentiating once: $U'(R_j\delta^n) = \lambda (P_j/Q_j) r^n$

differentiating twice: $U''(R_j\delta^n) = (\lambda/\delta^n r^n)[P_j'Q_j - P_jQ_j')/Q_j^2]$

combining: $-U''(R_j\delta^n)/U'(R_j\delta^n) = (\delta^{-n})[Q_j'/Q_j) - (P_j'/P_j)]$

This shows how *absolute risk aversion* for a given state is related to subjective and risk-neutral probabilities for that state, independent of other states. With this, we can examine center states that have the highest probability while neglecting the notoriously unreliable tail estimates.

techniques described in table 6.16 for several nonoverlapping time periods from April 2, 1986 to December 30, 1994.[5]

Unreported tests show that the estimated subjective probabilities are robust to perturbations in all of these assumptions. In particular, assuming a risk premium in the range of 5–10% leaves the results essentially unchanged.

The resulting absolute risk aversion is described in figure 6.12 for each time period. For example, for the single precrash period, April 1986 to September 1987, absolute risk aversion is positive but more or less declining with increasing wealth, and within the range 0 to 5—a plausible result. Unfortunately, in all the postcrash periods the results make no sense. Absolute risk aversion is not only increasing over levels of wealth greater than current wealth but is even negative over the range 0.9 to 1.06 times current wealth. In addition, this bizarre result worsens as we move farther into the future from the 1987 crash.

This result is essentially being driven by the extreme difference between the risk-neutral and measured subjective distributions around the mean. As we saw in figure 6.10, postcrash, on both sides of the mean, the risk-neutral distributions changed much more rapidly than the subjective dis-

Table 6.16. Methodology.

Risk-neutral probability distributions:

- inferred from S&P500 Index options with 135–225 days-to-expiration
- using the maximum smoothness method

Other parameters (S, r, d, t) as observed in the market

Subjective probability distributions:

- bootstrapped from 4-year historical samples
- 25,000 returns matching the option's time-to-expiration are generated and smoothed through a Gaussian kernel
- mean is reset to risk-free rate plus 7% annualized
- volatility is reset to volatility of risk-neutral distribution

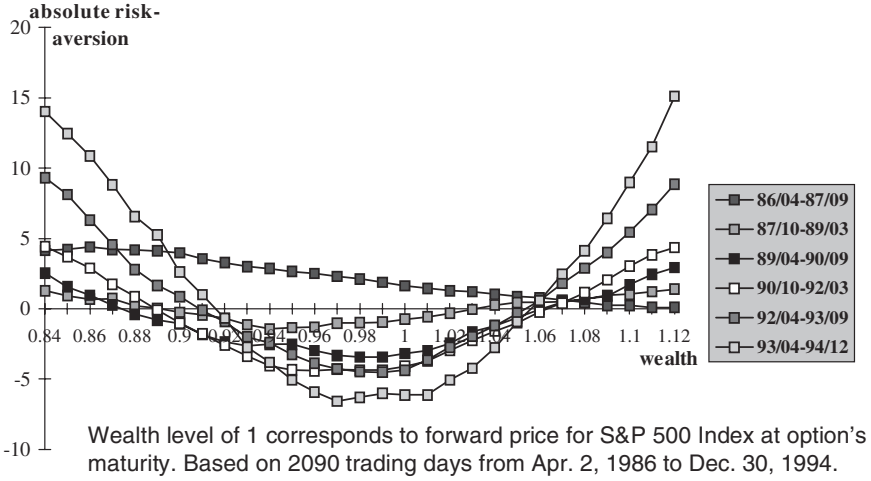


Figure 6.12. Absolute Risk Aversion Across Wealth

tribution. It is simply the case that our equilibrium model can not make sense of this.

For the first-order condition of this model to be a necessary condition of equilibrium, the second-order condition, which requires a negative second derivative of the utility function, must hold. But for absolute risk aversion to be negative, either $U' < 0$ or $U'' > 0$. If, for example, $U' > 0$ but $U'' > 0$, then the first-order condition need not characterize the optimum. In essence, figure 6.12 says that something is seriously wrong somewhere.

If the assumed risk premium is pushed from 7% to as high as 23%, all the lines of absolute risk aversion now fall just above the horizontal axis, so that $U' > 0$ and $U'' < 0$. However, even in this extreme case, the shape of the lines remains about the same. In particular, postcrash, they continue to exhibit increasing absolute risk aversion in the range above current wealth.

So what could be wrong? Something we have assumed must be at fault. One possibility (see table 6.17) is that our use of a representative investor could be a very bad assumption, or the S&P 500 Index could be a very poor approximation of the market portfolio. For that to be true, the market portfolio must be relatively uncorrelated with the returns of the S&P 500 Index. Utility could be a function of other significant variables besides wealth. Perhaps investors prefer risk over a range of their future wealth, such as suggested by prospect theory.

Even though our results only depend on probabilities near the mean, it may still be the case that historically realized returns are not reliable indicators of subjective probabilities.

So far, for the most part, we have ignored trading costs. In particular, although some methods used for estimating risk-neutral probabilities require

Table 6.17. Potential explanations.

- Representative investor is a poor assumption.
- S&P 500 Index is a poor proxy for the market portfolio.
- Utility functions depend on other variables besides wealth.
- More general frameworks for utility functions admitting risk preference (prospect theory).
- The subjective distribution Q is not well-approximated by realizations.
- Trading costs, particularly for deep out-of-the-money put options.
- Mispricing of deep out-of-the-money puts and calls.

that securities valued under these probabilities fall within the bid–ask spread, even these methods do not give full consideration to the role of trading costs in all their varying guises: commissions, bid–ask spread, and market impact. For example, it may be that the relatively high prices of out-of-the-money puts, which drive the postcrash S&P 500 Index smile, are somehow the result of trading costs that we have not considered. To us, given the magnitude of the smile effect and the high absolute dollar prices of these options (since the underlying asset is scaled to a high price), this seems unlikely, but it should not be dismissed without a deeper analysis.

Another problem with looking to trading costs as the solution to the puzzle is that the implied risk-neutral distribution changed markedly from before to after the stock market crash, yet it seems unlikely that trading costs did.

Finally, although it may be heretical to suggest this, the high prices of out-of-the-money puts may be the result of mispricing that a normally efficient market fails to correct. For some reason, not enough capital may be mobilized to sell these types of options.

We consider this last possibility by examining the returns from following a strategy where out-of-the-money six-month S&P 500 Index puts are sold every three months during the postcrash period. Each period, we assume that the number of puts sold equals the number that could be sold with $100 margin under the requirement that the margin for a sold uncovered put is 30% of the index level less the out-of-the-money amount. We compare these realized returns to risk measured by a version of the Capital Asset Pricing Model that considers positive preference toward skewness, an aspect of investor preferences that may be important in the pricing of securities with adjustable asymmetric outcomes, such as options. (Rubinstein 1976, Leland 1999).

Table 6.18 states that replacing

$$\beta \equiv \mathrm{Cov}(r_p, r_m)/\mathrm{Var}\, r_m$$

Table 6.18. Adjusted excess return measure.

- Assume the market portfolio exhibits lognormal returns
- Instead of β, we use $B \equiv \text{Cov}[r_p, -r_m^{-b}]/\text{Cov}[r_m, -r_m^{-b}]$, where B is an adjusted beta measure for the portfolio

  $r_p \equiv$ return of the portfolio

  $r_m \equiv$ return of the market

  $r \equiv$ riskiess return
- In this case: $b = (\ln(E[r_m]) - \ln r)/\text{Var}[\ln r_m]$
- Instead of α, we use $A \equiv (r_p - r) - B(r_m - r)$, where A is an adjusted portfolio expected excess return

is the generalized risk measure (adjusted beta)

$$B \equiv \text{Cov}(r_p, r_m^{-b})/\text{Cov}(r_m, r_m^{-b})$$

where r_p is the return (one plus the rate of return) of an arbitrary portfolio, r_m is the return of the market portfolio, and b is the consensus market relative risk aversion.

Using this measure, the realized excess over risk-adjusted return is

$$A = (r_p - r) - B(r_m - r)$$

which we call the realized adjusted alpha. Based on the formula

$$b = (\ln(E[r_m]) - \ln r)/\text{Var}[\ln r_m]$$

we set $b = 3.63$. But even if b were as high as 10, our results would be essentially unchanged.

Figure 6.13 shows the results of our adjusted alpha and adjusted beta return analysis. The riskless return itself is located at the origin, and the market return is located along the horizontal axis at 1 (adjusted alpha of 0, adjusted beta of 1).

Each line looks at the alpha–beta ordered pairs for strategies using puts of varying degrees of being out-of-the-money. For example, the puts on the upper line were about 5% out-of-the-money at the time of sale.

An important objection to our analysis as it has so far been described is that our strategy of selling out-of-the-money puts may do well in the post-crash periods because the much-feared second crash has not yet occurred, and had it occurred our strategies would have done poorly. To allow for this, we have inserted crashes into the data at varying frequencies. For example, the alpha–beta ordered pairs labeled 4 are constructed from the time series of S&P 500 Index returns by adding crashes of the October 19, 1987 magnitude (down 20% in a single day) at the expected rate of once every four years. That is, each day a number is drawn at random with re-

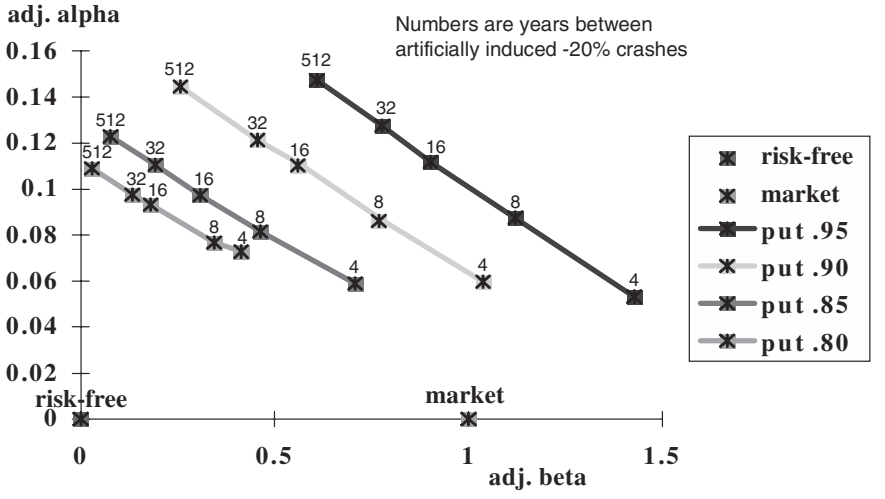


Figure 6.13. Excess Returns Selling Out-of-the-Money Puts

placement from a bowl containing about 999 zeros and 1 one. If 0 is drawn, no crash occurs. If 1 is drawn, then the return for that day is adjusted downward by 20% and future returns continue as before unless yet another crash is drawn. For example, if the return for that day were actually –1%, we assume instead that the return was –21%.

Thus, the exhibit shows that with almost no crashes added to the postcrash historical record (the ordered pairs labeled 512), the adjusted alpha ranges from 11% to 15% per annum. Thus, given the actual outcomes in the postcrash period, the strategy of selling out-of-the-money puts would have beaten the market by a good margin, and the more out-of-the-money the puts, the better.

Perhaps this is not surprising since almost no artificially induced crashs have been added to the realized historical time series. However, at the other extreme, suppose a crash is inserted into the data with a frequency of every four years. This means that about two daily 20% crashes during the postcrash periods were assumed to occur. In that case, the adjusted alpha is about 6–7% per annum. Even in this case, the strategy produces superior returns.

Now you may object that our result does not adequately consider the extreme fear the market may have of downside returns. But we have given some consideration to this because we have been careful to adjust our risk measure for dislike of negatively skewed returns, for a relative risk-aversion level of $b = 3.63$. Also, imposition of commissions plus bid–ask spread transaction costs, with a crash expected every four years, still leaves adjusted alphas in the range of 4–5% per annum.

Table 6.19 provides a summary of this chapter.

Table 6.19. Summary

Motivation: postcrash failure of the Black–Scholes formula in the index options market

Recover risk-neutral probabilities from option prices

- robust methods available
- leptokurtosis and left-skewness

Imply a consistent risk-neutral stochastic process

- simple generalization of standard binomial trees

Empirical tests of alternative smile forecasts

- "naïve trader" model best, several academic models similar

Recover risk aversion from realized returns and option

- strange results, index option market possibly inefficient

NOTES

1. Note, however, that a flat smile is only a necessary condition for the Black–Scholes formula to hold, not also a sufficient condition.

2. It is sometimes argued that while the Black–Scholes formula can be expected to hold for individual equity options since their underlying asset returns should be approximately lognormal, it will not hold for index options whose underlying index would then be a weighted sum of lognormal variables, clearly itself not lognormal. However, we believe this puts the cart before the horse. It seems to us more probable that when "God" created the financial universe, he made the market portfolio lognormal; and man, in his efforts to create exchange arrangements, then created individual equities and other securities he called bonds with returns that are not lognormal. We suspect that empirical analysis would show that the returns of diversified portfolios of stocks are closer to lognormal than their typical constituent components. Moreover, jumps, the Achilles heel of the Black–Scholes formula, are much more likely to prove a problem for a typical individual equity than for an equity index.

3. The fourth property, recovering a lognormal distribution if all available options have the same Black–Scholes implied volatility, is met if the prior distribution is assumed to be lognormal, as assumed in figure 6.7.

4. The paper by Dumas, Fleming, and Whaley (1998) seems to contain a similar result. However, here, rather than emphasize the failure of implied binomial trees, we instead emphasize that implied binomial trees do much better than Black–Scholes and about as well as any competing "academic" model we have tested.

5. The use of a bootstrapping method destroys serial correlation. However, Jackwerth (1997b) indicates that a lognormal distribution (which is the result of destroying serial dependence) provides a reasonable fit to these half-year returns. On another matter, the risk-neutral distribution is strictly a point estimate. Some results concerning the degree to which probabilities can vary around point estimates is contained in Jackwerth (1997c).

REFERENCES

Black, F. and M. Scholes, "The Pricing of Options and Corporate Liabilities," *Journal of Political Economy* 81, No. 3 (May–June 1973), pp. 637–659.

Breeden, D. T. and R. H. Litzenberger, "Prices of State-Contingent Claims Implicit in Option Prices," *Journal of Business* 51, No. 4 (October 1978), pp. 621–651.

Cox, J. C., "The Constant Elasticity of Variance Option Pricing Model," *Journal of Portfolio Management* (Special Issue: A Tribute to Fischer Black, December 1996), pp. 15–17.

Cox, J. C., S. A. Ross, and M. Rubinstein, "Option Pricing: A Simplified Approach," *Journal of Financial Economics* 7, No. 3 (September 1979), pp. 229–263.

Dumas, B., J. Fleming, and R. Whaley, "Implied Volatility Functions: Empirical Tests," *Journal of Finance* 53, No. 6 (1998), pp. 2059–2106.

Gemmill, G. and N. Kamiyama, "International Transmission of Option Volatility and Skewness: When You're Smiling, Does the Whole World Smile?" City University Business School, London, working paper (February 1997).

Heston, S. L. "A Closed-Form Solution for Options with Stochastic Volatility and Applications to Bond and Currency Options," *Review of Financial Studies* 6, No. 2 (Summer 1993), pp. 327–343.

Jackwerth, J. C., "Generalized Binomial Trees," *Journal of Derivatives* 5, No. 2 (1997a), pp. 7–17.

Jackwerth, J. C., "Do We Live in a Lognormal World?" University of Wisconsin at Madison, working paper in progress (February 1997b).

Jackwerth, J. C., "Geometric and Probabilistic Bounds on Option Prices," University of Wisconsin at Madison, working paper in progress (March 1997c).

Jackwerth, J. C., "Recovering Risk Aversion from Option Prices and Realized Returns," *Review of Financial Studies* 13, No. 2 (2000), pp. 433–451.

Jackwerth, J. C. and M. Rubinstein, "Recovering Probabilities from Option Prices," *Journal of Finance* 51, No. 5 (December 1996), pp. 1611–1631.

Jackwerth, J. C. and M. Rubinstein, "Recovering Stochastic Processes from Option Prices," London Business School, working paper (2001).

Leland, H. E., "Beyond Mean-Variance: Risk and Performance Measurement in a Nonsymmetrical World," *Financial Analyst Journal* 55, (1999), pp. 27–36.

Merton, R. C., "Option Pricing when Underlying Stock Returns Are Discontinuous," *Journal of Financial Economics* 3, No. 1 (January–March 1976), pp. 125–144.

Rubinstein, M., "The Valuation of Uncertain Income Streams and the Pricing of Options," *Bell Journal of Economics and Management Science* 7, No. 2 (Autumn 1976), pp. 407–425.

Rubinstein, M., "Displaced Diffusion Option Pricing," *Journal of Finance* 38, No. 1 (March 1983), pp. 213–217.

Rubinstein, M., "Nonparametric Tests of Alternative Option Pricing Models Using All Reported Trades and Quotes on the 30 Most Active CBOE Option Classes from August 23, 1976 through August 31, 1978," *Journal of Finance* 40, No. 2 (June 1985), pp. 455–480.

Rubinstein, M., "Implied Binomial Trees," *Journal of Finance* 49, No. 3 (Presidential Address to the American Finance Association, July 1994), pp. 771–818.

Toft, K. B. and B. Prycyk, "Options on Levered Equity: Theory of Empirical Tests," *Journal of Finance* 52, No. 3 (July 1997), pp. 1151–1180.

7

Cross-Sectional Determinants of Expected Returns

Michael J. Brennan, Tarun Chordia, & Avanidhar Subrahmanyam

1.

In theory, the primary determinant of the cross section of expected returns on securities is risk, although other factors such as taxes, liquidity, agency, and informational considerations can also, in principle, affect the demand for securities and therefore expected returns. The measure of risk that is relevant for pricing is the covariance of the security's return with the pricing kernel, which may be expressed in terms of one or more return factors, depending on the particular equilibrium model. Thus, in the Capital Asset Pricing Model (CAPM), the pricing factor is the return on the market portfolio; in the consumption CAPM, it is the return on the portfolio that is most highly correlated with the growth rate of consumption; in the Epstein and Zin (1989) recursive utility CAPM, it is a linear function of the return on the market and the return on the maximum consumption correlation portfolio; and in the Arbitrage Pricing Theory (APT) it is a linear function of a set of priced factors that span the pricing kernel.

In empirical work, a variety of approaches have been used to specify the factors. First, they may be determined by the theory. This is the case with the simple CAPM, in which the pricing factor is taken as some proxy for the return on the market portfolio as in Black, Jensen, and Scholes (1972) or Fama and MacBeth (1973). Chen, Roll, and Ross (1986) also adopted this approach in their empirical implementation of the Arbitrage Pricing Theory, in which the factors relevant for pricing were taken to be macroeconomic factors that affect stock prices. A second approach is to extract the factors from the covariance structure of asset returns. This approach has been followed by many different authors, including Roll and Ross (1980), Chen (1983), Connor and Korajczyk (1988) and Lehmann and Modest (1988). A third approach is to leave the factors unspecified but to assume that factor loadings can be written as linear functions of a small set of security characteristics: Rosenberg (1974) and Sharpe (1982) are early examples of this approach.[1] Finally, Fama and French (1993) have pioneered a hybrid approach in which the factors are chosen not on the basis of their

contribution to the covariance matrix or an a priori assessment of their economic importance but as returns on portfolios constructed according to particular characteristics of the securities, namely the market-to-book ratio and firm size. Fama and French show that their factors are able to explain the excess returns on a full set of portfolios ranked according to the book-to-market ratio and firm size.

However, work by Roll (1994), Daniel and Titman (1997), and Haugen and Baker (1996) has cast doubt on the ability of a factor model to explain the cross section of expected returns. First, Roll shows in a paper that is closely related to this chapter that relative returns on eight portfolios sorted by size, earnings–price ratio, and book-to-market ratio cannot be explained by their loadings on five Connor–Korajczyk (1988) type factors. Secondly, Daniel and Titman show that portfolios with the same book-to-market and size characteristics but different loadings on the Fama–French factors do not have different returns, as they should if expected returns are determined by the factor loadings. Finally, Haugen and Baker show that dynamically rebalanced portfolios formed on the basis of a large number of firm characteristics and the lagged rewards to those characteristics have a wide dispersion in realized mean returns without a corresponding dispersion in risk. Haugen and Baker include among their firm characteristics the market beta and estimated loadings on macroeconomic variables such as changes in industrial production and inflation. However, these loadings are measured with error, which is not taken into account in their monthly cross-section regressions of returns on the firm characteristics. Despite this, the Haugen and Baker results present a formidable empirical challenge to factor pricing theory.

This chapter complements these earlier studies by investigating further the ability of firm characteristics to explain the cross section of returns after accounting for the effects of APT type factors. First, rather than prespecifying the factor structure as in Daniel and Titman and Fama and French, we follow Roll (1994) in selecting the factors (as in the second approach above) according to their importance in explaining the covariance matrix of returns. Secondly, we consider a broader range of firm characteristics than do these authors, though one that is smaller than that of Haugen and Baker. Thirdly, in most of our tests we employ individual security data rather than portfolios, while avoiding the errors-in-variables problem that arises from putting estimated factor loadings on the right-hand side of a cross-sectional regression in which returns are the dependent variable.

The characteristics that we focus on are firm size, stock price, turnover, bid–ask spread, analyst following, dispersion of analyst opinion, book-to-market ratio, institutional holdings, membership in the S&P 500 Index, dividend yield, and lagged returns. These characteristics are chosen either because of prior empirical evidence of their association with returns, as in the case of size, share price, the book-to-market ratio, and lagged returns, or because there are sound economic reasons for expecting that the variables may affect expected returns. One objective of this study is to determine whether the size and book-to-market anomalies may be accounted

for by related variables whose economic role is understood, such as institutional ownership, liquidity, and dividend yield. While the effects of several of these firm characteristics have been analyzed in earlier studies, this is the first attempt to provide a joint analysis of the marginal effects of the characteristics above on expected returns, given the role of five risk factors that are estimated according to the approach of Connor and Korajczyk (1988), and it is also the first to consider jointly the effect of institutional holdings and index membership on expected returns.

The traditional approach to assessing the importance of risk and other security characteristics for expected returns has been to form portfolios by first sorting securities on the characteristic of interest and then, following Fama and MacBeth (1973), averaging the time series of coefficients from cross-sectional regressions of (excess) returns on the portfolios on the average beta estimates and other portfolio characteristics. Securities are aggregated into portfolios for the estimation because of the problem of errors in the estimates of the betas, first noted in this context by Black, Jensen, and Scholes (1972). A more direct approach to dealing with the errors-in-variables problem is to estimate the security betas simultaneously with the coefficients of the security characteristics in a single set of seemingly unrelated regressions, as in Roll (1994), Pontiff (1995), and Brennan and Subrahmanyam (1996). However, this approach also requires that the securities be aggregated into portfolios before estimation, not to overcome an error-in-variables problem but so that the number of risk parameters to be estimated and the dimension of the error variance matrix will remain manageable.

The use of portfolios rather than individual securities in tests of asset-pricing models has been criticized from opposing perspectives by Roll (1977) and by Lo and MacKinlay (1990). Roll argues that the portfolio formation process, by concealing possibly return-relevant security characteristics within portfolio averages, will make it difficult to reject the null hypothesis of no effect on returns. Lo and MacKinlay make the almost precisely opposite point that if the researcher forms portfolios on the basis of characteristics that prior research has shown to be relevant to expected returns, he will be inclined to reject the null hypothesis too often due to a "data-snooping bias." It is worth emphasizing that the Roll and the Lo and MacKinlay critiques of the portfolio formation approach are complementary rather than competing; portfolio formation may both make some return-irrelevant characteristics appear significant and disguise the empirical relevance of other return-relevant characteristics.

Consistent with the critiques of Roll and of Lo and MacKinlay, we find that the empirical significance of security characteristics for expected returns is dependent on the way in which the portfolios are formed. We therefore turn to a modification of the Fama–MacBeth approach that can be applied to individual securities rather than to portfolios. We find that after risk adjustment using the Connor–Korajczyk (1988) factors, mean returns remain significantly related to several firm characteristics. First, we con-

tinue to find that size is important when share turnover is measured as a proportion of shares outstanding. However, when trading volume is measured by the dollar value of shares traded, size is no longer significant, while trading volume is. This is consistent with size being priced because it is an important component of liquidity. The market-to-book ratio is not significant. We find a significant positive association with analyst following and membership in the S&P 500 Index; it is possible that the latter, which implies a return premium of 0.27% per month for Index membership, is associated with the growth in index portfolio investment strategies during our sample period.[2] On the other hand, we find no significant return effects associated with a measure of the dispersion of analysts' opinions, the proportion of stock owned by institutions, the (reciprocal of) share price, dividend yield, or 3- and 6-month lagged returns. We do find a significant positive association with the 12-month lagged returns, and our most striking finding is of a strong *negative* association between risk-adjusted returns and the proportional bid–ask spread. This association persists, albeit in somewhat weaker form, when the returns are risk-adjusted using the Fama–French (1993) factors[3] or when they are not risk-adjusted at all. This finding of a negative association with the bid–ask spread is contrary to that of Amihud and Mendelson (1986) and to the idea that investors require compensation for illiquidity as measured by the bid–ask spread. However, it is consistent with the empirical findings of Eleswarapu and Reinganum (1993) and Brennan and Subrahmanyam (1996), who report a negative association between returns and the spread using in one case the Capital Asset Pricing Model and firm size and in the other the Fama–French (1993) three-factor model to adjust for risk. Moreover, Petersen and Fialkowski (1994) show that the correlation between the posted bid–ask spread measure that is used in empirical work and the effective spread that investors pay is only of the order of 0.10, so that the posted bid–ask spread is a poor measure of the cost of transacting. However, if the posted spread is not a measure of transaction costs, there remains the issue of what it is measuring and why it has such a powerful (and negative) association with expected returns. We have no satisfactory answer to this question. The analysis of MacKinlay (1995) suggests that the *t*-statistic for the coefficient on this security characteristic is too high to be consistent with an omitted risk factor explanation.[4] One possible explanation, suggested by Harris (1994), is that high posted spreads are associated with more liquidity ceteris paribus because they make it more attractive for traders to post limit orders, which increase market depth. While accepting this point, we find it hard to believe that the (perverse) liquidity effect can be strong enough to explain our findings.

As interesting as the variables for which we find a relation to risk-adjusted expected returns are those for which we find no such relation. In contrast to Stattman (1980), Rosenberg, Reid, and Lanstein (1985), Fama and French (1993), and Roll (1994), we find no relation between expected returns and the book-to-market ratio, suggesting that the effects of this

variable are subsumed by our risk variables and firm characteristics.[5] The dividend yield variable is also insignificant, providing no support for the Brennan (1970) differential taxation hypothesis. The dispersion of analyst earnings forecasts also has no reliable association with expected returns.

The remainder of the chapter is organized as follows. In section 2, we describe the empirical hypotheses we test. In section 3, the data are described, and in section 4 the empirical results are presented. Section 5 concludes.

2. HYPOTHESES

The basic equation that we wish to estimate is

$$E\left[\tilde{R}_j\right] - R_F = c_0 + \sum_{k=1}^{5} \lambda_k \beta_{jk} + \sum_{m=1}^{M} c_m Z_{mj} \tag{1}$$

where R_j is the return on security j, R_F is the risk-free interest rate, β_{jk} is the loading of security j on factor k, λ_k is the risk premium associated with factor k, Z_{mj} is the value of (nonrisk) characteristic m for security j, and c_m is the premium per unit of characteristic m. Our null hypothesis is that expected returns are determined solely by the risk variables, which are the factor loadings, β_{jk}, and that in the presence of the risk variables, the other security characteristics have no marginal explanatory power. We include 14 security characteristics as possible determinants of expected returns.

The five risk factors are taken to be the first five (asymptotic) principal components of stock returns estimated over the sample period. In deciding which firm characteristics to include, attention was given to those that had been found to be important in prior studies as well as those for which there exists a theoretical rationale. Thus, firm size is included because of the widespread evidence of a "small firm effect."[6] Fama and French (1993) implicitly treat firm size as a loading on a risk factor, although the work of Daniel and Titman (1995) suggests that the firm size itself is a stronger determinant of expected returns than is the loading on the size factor. Berk (1995) suggests that firm size may appear as a determinant of expected returns because of imperfect risk adjustment in the empirical analysis. It is therefore important to assess whether size has any residual explanatory power for expected returns once account is taken of the five risk factors[7] and other firm characteristics. Share price has also been found in a number of studies to be (negatively) related to expected return. It has also been hypothesized that the low-price effect reflects the fact that firms with low prices are often in financial distress and that financial institutions may be reluctant to invest in them on account of the prudent man rule.[8] Therefore, we include the reciprocal of share price as a possible determinant of expected returns, as well as including institutional ownership.

Turnover, or trading volume, is included because this variable is associated with liquidity, and the work of Amihud and Mendelson (1986) and Brennan and Subrahmanyam (1996) suggests that expected returns are

affected by liquidity. The bid–ask spread is included for similar reasons. The limited-information model of Merton (1987) suggests that the expected return on securities will depend on how well-known the securities are to investors, and we hypothesize that this will depend on how many brokerage analysts follow the firms.[9] We therefore include a measure of analyst following.

The dispersion of analyst opinion is included as a further possible explanatory variable because the theoretical work of Miller (1977) and Jarrow (1980) suggests that, in the presence of short-sales constraints, disagreement among investors will tend to be associated with overpricing and therefore lower expected returns. The book-to-market ratio is included because this variable has been found to be important empirically by Stattman (1980), Rosenberg, Reid, and Lanstein (1985), and Frankel and Lee (1996) and because Fama and French (1993) found that a factor portfolio formed from securities by sorting on this characteristic helped to explain the cross section of expected returns on portfolios sorted by size and book-to-market ratio. We consider the possible influence of institutional holdings and membership in the S&P 500 Index because the agency model of Brennan (1994) suggests that, insofar as institutional portfolio managers are rewarded on the basis of relative rather than absolute performance, their portfolio decisions will distort the structure of expected returns; it is also possible that institutional investors are superior analysts and that institutional ownership is associated with superior investment performance. Dividend yield is included because Brennan (1970) suggests that differential taxation of dividends and capital gains could make this variable relevant, and the resulting empirical work of Miller and Scholes (1978) and Litzenberger and Ramaswamy (1979) has been inconclusive. Finally, we include lagged return variables because the work of Jegadeesh and Titman (1993) has shown these to be relevant, and by including them we should improve the efficiency of the estimates of the coefficients of the other variables.

3. DATA

The basic data consist of monthly returns and other characteristics for a sample of NYSE securities for the period January 1977 to December 1989. The sample period is limited by the availability of data on analyst following and the bid–ask spread. To be included in the sample for a given month, a security had to satisfy the following criteria: (1) its return in the current month and in 24 of the previous 60 months be available from CRSP, and sufficient data be available to calculate the size, price, turnover, and dividend yield as of the previous month; (2) sufficient data be available on the COMPUSTAT tapes to calculate the book-to-market ratio as of December of the previous year; (3) annual earnings forecasts be available on the I/B/E/S tape for the previous month; (4) institutional holdings be reported in the previous year's December issue of the S&P Security Owner's Stock Guide; (5) the average bid–ask spread be available for the previous year.

This screening process yielded an average of 980 stocks per month. In comparison, Fama and French (1992) had an average annual sample size of 2,267 stocks. The Fama and French sample, however, also included AMEX and NASDAQ stocks.

The data come from several sources: Institutional Brokers' Estimate System (I/B/E/S) (analyst following and earnings forecasts); S&P Security Owner's Stock Guide (institutional holdings); Wells Fargo Investment Advisors and Robert Whaley (S&P Index membership); Hans Stoll and Marc Reinganum (bid–ask spread data); and the CRSP and COMPUSTAT tapes (returns, market capitalization, price, turnover, dividend yield, and book-to-market ratio). For each security, the following variables were calculated each month as follows:

SIZE—the natural logarithm of the market value of the equity of the firm as of the end of the previous month.

BM—the natural logarithm of the ratio of the book value of equity plus deferred taxes to the market value of equity using the end of the previous calendar year market value and the most recent book value available at the end of the previous calendar year.

TO—the natural logarithm of share volume for the previous month expressed as a proportion of the number of shares outstanding.

DVOL—the natural logarithm of the dollar share trading volume for the previous month.

NANAL—the natural logarithm of the number of analysts making annual earnings forecasts as reported on the I/B/E/S tape for the previous month.

DISP—the absolute value of the coefficient of variation of analysts' one-year earnings forecasts as of the previous month.

SPREAD—the natural logarithm of the average bid–ask spread as a proportion of the closing stock price for the previous year (calculated as the average of the beginning and the end-of-year closing bid–ask spread relative to the mean quote).

PINST—the natural logarithm of the proportion of the stock held by institutions as reported in the S&P Security Owner's Stock Guide in December of the previous year.

S&P—a dummy variable that takes on a value of unity if the security is included in the S&P 500 Index at the end of the previous month and zero otherwise.

PRICE—the natural logarithm of the reciprocal of the share price as reported at the end of the previous month.

YLD—the dividend yield as measured by the sum of all dividends paid over the previous 12 months divided by the share price at the end of the previous month.

RET3—the cumulative return from month –3 to month –1.

RET6—the cumulative return from month –6 to month –1.

RET12—the cumulative return from month –12 to month –1.

The lagged return variables are constructed to exclude the immediate previous month's return to avoid any spurious association with the current month return due to thin trading or bid–ask spread effects. Finally, for all of the regressions reported below, all of the firm characteristics variables were expressed as deviations from their cross-sectional means each month; this implies that the expected return for a security with average values of these characteristics will be determined solely by its factor loadings and the factor risk premiums.

Table 7.1 reports the grand time series and cross-sectional means, medians, and standard deviations of the raw security characteristics. Note that the variables in table 7.1 are not in logarithms. Table 7.2 reports the averages of the month-by-month cross-sectional correlations of the variables that we use in our analysis. The largest correlations with SIZE are DVOL (positive), NANAL (positive), SPREAD (negative), S&P (positive), and PRICE (negative); with NANAL, they are SPREAD (negative), PRICE (negative), PINST (positive), and S&P (positive); with SPREAD, they are PRICE (positive) and S&P (negative). The correlations of DVOL with the other variables are similar to those of SIZE. The other correlations are smaller than 0.4 in absolute value.

For the first set of regressions reported below, 25 portfolios were constructed as follows. First, the securities were assigned to one of five equal-

Table 7.1. Summary statistics. The summary statistics represent the grand cross-section, time series average for an average of 980 NYSE stocks over 156 months from January 1977 through December 1989. Each stock had to satisfy the following criteria: (1) its return in the current month and in 24 of the previous 60 months be available from CRSP, and sufficient data be available to calculate the size, price, turnover, and dividend yield as of the previous month; (2) sufficient data be available on the COMPUSTAT tapes to calculate the book-to-market ratio as of December of the previous year; (3) annual earnings forecasts be available on the I/B/E/S tape for the previous month; (4) institutional holdings be reported in the previous year's December issue of the S&P Security Owner's Stock Guide; (5) the average bid–ask spread be available for the previous year.

Variable	Mean	Median	Std. Dev.
Firm size ($bill)	1.302	0.373	3.685
Book-to-market ratio	1.234	0.956	4.167
Share turnover (% per month)	4.81	3.44	5.28
Dollar trading volume ($mill. per month)	64.74	12.39	19.5
Number of analysts	10.06	8.00	8.30
Coefficient of variation of analysts' earnings forecasts	0.199	0.047	2.098
Proportional bid–ask spread (%)	1.31	1.07	1.11
Institutional ownership (%)	29.54	26.23	20.67
Share price ($)	28.97	24.88	21.97
Dividend yield (%)	4.30	3.63	8.09

Table 7.2. Correlation matrix of transformed firm characteristics. This table presents time series averages of monthly cross-sectional correlations between transformed firm characteristics used in pricing regressions. The variables relate to an average of 980 stocks over 156 months from January 1977 through December 1989. RETURN denotes the excess monthly return (i.e., the raw return less the risk-free return). SIZE represents the logarithm of the market capitalization of firms in billions of dollars. BM is the ratio of book value of equity plus deferred taxes to market capitalization. TO is the logarithm of the ratio of the monthly trading volume to the number of shares outstanding. DVOL is the logarithm of the dollar trading volume. NANAL is the logarithm of the number of analysts following a stock. DISP is the absolute value of the coefficient of variation of analyst forecasts. SPREAD is the logarithm of the relative bid–ask spread. PINST is the logarithm of the fraction of firm shares held by institutions. S&P is a dummy variable, which equals one if the stock belongs to the S&P500 Index. PRICE denotes the logarithm of the share price reciprocal. YLD is the dividend yield.

	RETURN	SIZE	BM	TO	DVOL	NANAL	DISP	SPREAD	PINST	S&P	PRICE	YLD
RETURN	1.00	–0.020	0.015	–0.001	–0.019	–0.015	–0.014	0.013	–0.001	–0.004	0.008	0.005
SIZE	–0.020	1.00	–0.266	0.065	0.890	0.798	–0.056	–0.743	0.323	0.606	–0.672	0.125
BM	0.015	–0.266	1.00	0.012	–0.228	–0.182	0.075	0.219	–0.119	–0.081	0.252	0.154
TO	–0.001	0.065	0.012	1.00	0.507	0.186	0.056	–0.081	0.169	0.160	–0.045	–0.171
DVOL	–0.019	0.890	–0.228	0.507	1.00	0.771	–0.021	–0.676	0.356	0.592	–0.600	0.024
NANAL	–0.015	0.798	–0.182	0.186	0.771	1.00	0.023	–0.586	0.402	0.564	–0.483	0.089
DISP	–0.014	–0.056	0.075	0.056	–0.021	0.023	1.00	0.074	–0.008	0.020	0.125	–0.037
SPREAD	0.013	–0.743	0.219	–0.081	–0.676	–0.586	0.074	1.00	–0.341	–0.441	0.746	–0.147
PINST	–0.001	0.323	–0.119	0.169	0.356	0.402	–0.008	–0.341	1.00	0.260	–0.396	–0.039
S&P	–0.004	0.606	–0.081	0.160	0.592	0.564	0.020	–0.441	0.260	1.00	–0.355	0.027
PRICE	0.008	–0.672	0.252	–0.045	–0.600	–0.483	0.125	0.746	–0.396	–0.355	1.00	–0.016
YLD	0.005	0.125	0.154	–0.171	–0.024	0.089	–0.037	–0.147	–0.039	0.027	–0.016	1.00

sized groups based on SIZE; then, within each of the SIZE quintiles, the securities were again sorted on the basis of a second criterion and assigned to one of five portfolios. In this way, 25 portfolios were formed each year so as to maximize the dispersion of firm size and the second sorting criterion. This second criterion was, in turn, YLD, DISP, PINST, NANAL, TO, BM, and SPREAD, all measured prior to the current month. Thus, there are seven sets of 25 portfolios based on the different sorting criteria. For each of the portfolios, the monthly return was calculated as the arithmetic average of the returns of the securities in the portfolio, and the portfolio characteristics, SPREAD, SIZE, etc., were calculated in a similar fashion.

Finally, five factors were estimated by using the asymptotic principal components methodology applied by Connor and Korajczyk (1988)[10] (henceforth C–K factors). These factors were selected in preference to the three Fama and French (1993) factors because Daniel and Titman's (1995) analysis suggests that the F–F factors may not be adequate to capture the cross section of equity returns. We find that the C–K factors explain 95%, 63%, and 39% of the variation in the Fama–French market, size, and book-to-market factors, respectively, and for robustness we repeat part of the analysis with the Fama–French factors. For the purposes of calculating excess returns, the risk-free interest rate is taken as the 1-month risk-free rate from the CRSP bond files.

4. EMPIRICAL RESULTS

The null hypothesis against which we evaluate the influence of security characteristics is the five-factor APT. Thus, assume that returns are generated by a five-factor approximate factor model[11]:

$$\tilde{R}_{jt} = E\left[\tilde{R}_{jt}\right] + \sum_{k=1}^{5} \beta_{jk} \tilde{f}_{kt} + \tilde{e}_{jt} \tag{2}$$

where f_{kt} are mean zero and $E[e_{jt} \mid f_{kt}] = 0$. Then, the exact or equilibrium version of the APT implies that expected returns may be written as

$$E\left[\tilde{R}_{jt}\right] - R_{Ft} = \sum_{k=1}^{5} \lambda_{kt} \beta_{jk} \tag{3}$$

where R_{Ft} is the return on the riskless asset, and λ_{kt} is the risk premium for factor k. Substituting from (3) into (2), the APT implies that realized returns are given by

$$\tilde{R}_{jt} - R_{Ft} = \sum_{k=1}^{5} \beta_{jk} \tilde{F}_{kt} + \tilde{e}_{jt} \tag{4}$$

where $F_{kt} \equiv \lambda_{kt} + f_{kt}$ is the factor plus its associated risk premium.

Then, in order to test whether security characteristics have incremental explanatory power for returns relative to the Connor–Korajczyk (C–K) factors, the following equation is estimated:

$$\tilde{R}_{jt} - R_{Ft} = \beta_0 + \sum_{k=1}^{5} \beta_{jk} \tilde{F}_{kt} + \sum_{m=1}^{14} c_m Z_{mjt} + \tilde{e}_{jt} \tag{5}$$

where Z_{mjt} is the value of characteristic m for portfolio j in month t for each of the seven sets of 25 portfolios. Under the null hypothesis that expected returns depend only on the risk characteristics of the returns, as represented by β_{jk}, the loadings on the C–K factors, the coefficients c_m ($m = 1, \ldots, 14$) will be equal to zero.

A. Fama–MacBeth Portfolio Regressions

For comparability with previous studies, we first employed the following adaptation of the Fama–MacBeth (1973) procedure on the portfolio data. For each month j, the risk-adjusted return on each of the portfolios, R^*_{jt}, was estimated as

$$\tilde{R}^*_{jt} = \tilde{R}_{jt} - R_{Ft} - \sum_{k=1}^{5} \hat{\beta}_{jk} \tilde{F}_{kt} \tag{6}$$

where β_{jk} was estimated by regressing the portfolio excess returns on the C–K factors over the entire sample period from January 1977 through December 1989[12]; to allow for the possible effects of thin trading, the Dimson (1979) procedure with one lag was followed, and the estimate of β_{jk} was obtained by summing the coefficients on the current and lagged factors. Then, the risk-adjusted portfolio return was regressed on the (nonrisk) portfolio characteristics for each month from January 1977 to December 1989 according to

$$\tilde{R}^*_{jt} = c_{0t} + \sum_{m=1}^{14} c_{mt} Z_{mjt} \tilde{e}_{jt} \tag{7}$$

Finally, the estimates of the coefficients, c_{mt} ($m = 0,1, \ldots 14$), were averaged over time and their standard errors calculated. Note that although the factor loadings are measured with error, this error affects only the dependent variable in regression (7), and while the factor loadings will be correlated with the security characteristics, there is no a priori reason to believe that *errors* in the estimated loadings will be correlated with the security characteristics, so the averaged coefficients should be unbiased estimates of the true coefficients. The average coefficients and their associated t-statistics are reported in table 7.3. There is very little evidence in these results that firm characteristics exert any independent effect on security returns. For example, the only variable that is significant in the SIZE/BM sorted portfolio regressions is the institutional ownership variable, PINST, which has a t-statistic of 2.26; however, this variable becomes insignificant in the regressions using data on portfolios sorted on SIZE and institutional ownership. No other firm characteristic is significant in more than one of the regressions.

While it is tempting to interpret these results as supportive of the factor model since there is no strong evidence against the null, there are two rea-

Table 7.3. Fama–MacBeth regression estimates of Equation (7) using portfolio data. Coefficient estimates are time series averages of cross-sectional OLS regressions. The dependent variable is the portfolio excess return adjusted for the C–K factor realizations, and the independent variables are the firm characteristics measured as the deviation from the cross-sectional mean in each period. Quintile portfolios were first formed by sorting on SIZE. Then, each size portfolio was further subdivided into quintile portfolios by sorting (in turn) on YLD, DISP, PINST, NANAL, TO, BM, and SPREAD, yielding a total of 25 portfolios. The sample and the variables are defined in table 7.1. All coefficients are multiplied by 100. *t*-statistics are in parentheses.

	SZ/YLD	SZ/DISP	SZ/PINST	SZ/NANAL	SZ/TO	SZ/BM	SZ/SPRD
Intercept	0.030	–0.031	–0.031	–0.031	–0.030	–0.031	–0.031
	(0.84)	(0.85)	(0.85)	(0.85)	(0.84)	(0.85)	(0.85)
SIZE	–0.329	0.033	–0.290	0.109	0.023	–0.092	–0.171
	(1.09)	(0.12)	(1.23)	(0.31)	(0.08)	(0.33)	(0.47)
BM	0.180	0.024	0.304	0.007	0.533	0.559	0.184
	(0.47)	(0.70)	(0.85)	(0.02)	(1.58)	(0.95)	(0.51)
TO	–0.401	0.104	–0.425	–0.172	–0.165	–0.231	–0.375
	(1.23)	(0.35)	(1.56)	(0.62)	(0.22)	(0.83)	(1.15)
NANAL	0.143	–0.121	0.140	–0.358	–0.510	–0.166	–0.220
	(0.51)	(0.40)	(0.51)	(0.90)	(1.70)	(0.55)	(0.70)
DISP	–0.083	–0.254	0.284	–0.139	0.093	–0.694	0.209
	0.51)	(0.34)	(0.55)	(0.31)	(0.21)	(1.46)	(0.32)
SPREAD	–0.953	–0.548	–0.380	–0.157	0.396	0.259	–0.336
	(1.36)	(0.88)	(0.60)	(0.24)	(0.62)	(0.46)	(0.31)
PINST	–0.172	0.236	0.019	–0.334	0.636	0.448	0.089
	(0.67)	(1.04)	(0.05)	(1.38)	(2.64)	(2.26)	(0.38)
S&P	0.169	0.343	0.793	–0.059	.503	0.734	0.502
	(0.35)	(0.78)	(1.77)	(0.11)	(1.02)	(1.56)	(1.11)
PRICE	0.350	1.12	0.285	–0.562	–0.379	0.301	–0.246
	(0.59)	(1.89)	(0.50)	(0.92)	(0.66)	(0.58)	(0.42)
YLD	–1.76	1.03	–8.53	8.87	1.78	–3.70	–5.51
	(0.13)	(0.15)	(1.29)	(1.33)	(0.27)	(0.48)	(0.81)
RET3	0.600	–2.71	–3.57	0.078	1.99	–2.18	0.213
	(0.20)	(1.05)	(1.53)	(0.03)	(0.82)	(0.91)	(0.08)
RET6	1.95	0.857	2.03	–1.39	–4.37	–0.268	1.71
	(1.00)	(0.42)	(1.03)	(0.72)	(2.20)	(0.13)	(0.82)
RET12	0.477	1.63	1.26	0.834	1.61	1.56	0.849
	(0.40)	(1.62)	(1.14)	(0.77)	(1.37)	(1.50)	(0.73)

sons to suspect these results. First, the Fama–MacBeth technique may lack power against the alternative hypothesis insofar as it fails to take account of any residual correlation in the risk-adjusted portfolio returns. Second, as we shall discuss further below, the portfolio formation procedure may be obscuring significant relations between the variables. The natural estimation procedure to take account of cross-sectional correlation in the residual returns, e_{jt}, is generalized least squares.

B. Generalized Least Squares Regressions

We run pooled cross-section time series regressions of the portfolio returns on the C–K risk factors and portfolio characteristics; in this way we estimate simultaneously the factor loadings and the coefficients of the characteristics. We thus avoid the errors-in-variables problems associated with more traditional Fama and MacBeth (1973) procedures.[13] The estimation proceeds as follows. Define R as the $25T \times 1$ vector of portfolio excess returns, where T is the total number of time series observations, and the vector is ordered by month so that the first 25 observations correspond to the portfolio returns in month 1. Define X as the partitioned matrix $X = [W\ Z]$, where W is a $25T \times 125$ matrix of the C–K factors. The first 25 columns of W consist of T stacked 25×25 diagonal matrices with identical elements F_{1t}, the return on the first factor in month t, $t = 1, \ldots, T$; the second 25 columns consist similarly of the second factor, F_{2t}, and so on up to column 125. Z is a $25T \times 15$ matrix whose first column is a vector of units and whose remaining 14 columns are the vectors of the 14 portfolio attributes (SIZE, PRICE, etc.) whose incremental effect on expected returns we wish to assess. We first estimate the following OLS pooled cross-section time series regression:

$$R = X\beta + \varepsilon \tag{8}$$

where β is a 140×1 vector of coefficients: the first 125 elements are the coefficients of the five C–K factors for the 25 portfolios, ordered by portfolio; the next element is the constant term of the regression; and the last 14 elements are the coefficients of the 14 security characteristics. ε is a $25T \times 1$ vector of errors. In our application, the sample consists of monthly returns from January 1977 to December 1989 so that $T = 156$ and the total number of observations is 3,900. In order to obtain the GLS estimator of β, Ω, the variance–covariance matrix of errors in (8), is estimated assuming that the portfolio return errors are serially independent but allowing for cross-sectional dependence. Then Ω is a $25T \times 25T$ block diagonal matrix, whose typical element is the 25×25 covariance matrix of portfolio return errors. This is estimated using the residuals from (8). The generalized least squares estimate of β is given by

$$\hat{\beta} = X' \hat{\Omega}^{-1} X)\ (X' \hat{\Omega}^{-1} Y) \tag{9}$$

where $\hat{\Omega}$ is the estimate of Ω from the first-stage regressions.

The results of these GLS regressions for each of the seven sets of portfolios are reported in table 7.4.[14] Now we find much greater evidence that

Table 7.4. Generalized least squares regressions. Coefficient estimates of seven pooled time series, cross-section GLS regression estimates of equation (5). The dependent variable is the portfolio excess returns, and the independent variables are the C–K factors and firm characteristics, with the characteristics measured as the deviation from the cross-sectional mean in each period. The column headings describe the sorting criteria. Quintile portfolios were first formed by sorting on SIZE. Then, each size portfolio was further subdivided into quintile portfolios by sorting (in turn) on YLD, DISP, PINST, NANAL, TO, BM, and SPREAD, yielding a total of 25 portfolios. In addition to the variables whose coefficients are reported, the regressions include the Connor–Korajczyk factors as independent variables, whose coefficients are not reported for brevity. RETn denotes the *n*-month lagged returns. The sample and the other variables are defined in table 7.2. All coefficients are multiplied by 100. *t*-statistics are in parentheses.

	(1) SZ/YLD	(2) SZ/DISP	(3) SZ/PINST	(4) SZ/NANAL	(5) SZ/TO	(6) SZ/BM	(7) SZ/SPREAD
Intercept	–0.03	–0.04	0.04	–0.03	–0.05	4.63	5.04
	(0.73)	(1.08)	(0.99)	(0.92)	(1.34)	(3.88)	(3.78)
SIZE	–0.214	–0.057	–0.039	0.054	–0.066	0.205	0.214
	(1.29)	(0.35)	(0.25)	(0.30)	(0.41)	(1.51)	(1.65)
BM	0.670	0.916	0.878	0.104	0.971	0.730	0.228
	(1.95)	(2.74)	(2.24)	(0.30)	(2.79)	(2.21)	(0.71)
TO	–0.469	–0.153	–0.501	–0.422	–0.718	–0.033	–0.006
	(2.79)	(0.93)	(3.07)	(2.54)	(2.35)	(0.30)	(0.05)
NANAL	0.016	–0.030	0.121	–0.392	–0.227	–0.023	0.059
	(0.10)	(0.16)	(0.74)	(1.81)	(1.36)	(0.16)	(0.41)
DISP	–0.043	0.046	0.020	–0.105	–0.105	–0.050	–0.035
	(0.59)	(0.59)	(0.28)	(1.41)	(1.36)	(0.65)	(0.48)
SPREAD	–0.495	–0.268	–0.210	–0.489	0.54	1.02	1.49
	(1.33)	(0.74)	(0.61)	(1.34)	(1.52)	(3.68)	(4.64)
PINST	0.283	0.324	0.114	0.115	0.396	–0.211	0.207
	(1.39)	(1.64)	(0.59)	(0.59)	(2.01)	(0.86)	(1.19)
S&P	–3.85	0. 90	–1.15	1.14	4.35	–0.107	0.352
	(2.23)	(0.54)	(0.72)	(0.72)	(2.60)	(0.07)	(0.22)
PRICE	0.218	0.474	0.714	0.108	–0.527	–0.006	–0.594
	(0.68)	(1.54)	(2.41)	(0.35)	(1.72)	(0.02)	(2.09)
YLD	0.028	–0.658	0.141	–0.07	0.03	–0.938	–2.13
	(0.08)	(1.99)	(0.44)	(0.23)	(0.08)	(1.53)	(3.59)
RET3	0.75	–0.63	0.37	–0.59	–2.11	–0.30	–0.63
	(0.86)	(0.77)	(0.45)	(0.68)	(2.46)	(1.02)	(2.18)
RET6	0.44	1.56	1.32	0.87	0.02	0.15	0.16
	(0.83)	(3.06)	(2.59)	(1.67)	(0.12)	(0.71)	(0.78)
RET12	0.19	0.45	0.43	0.06	0.44	0.43	0.26
	(1.10)	(2.63)	(1.90)	(0.33)	(3.05)	(3.00)	(1.84)

the risk-adjusted returns are affected by the security characteristics. Using a *t*-statistic of two as a criterion of significance, we find that SIZE, NANAL, and DISP are nowhere significant; BM is significant (positive) in four regressions; TO is significant (negative) in four regressions; RET12 is significant (positive) in three regressions; SPREAD is significant (positive) in two regressions; and RET3 is significant (negative) in two regressions. S&P is significant in two regressions but has opposite signs in the two regressions. The remaining variables are significant in only one regression. The intercept terms are insignificantly different from zero except in the cases of the SIZE/BM and SIZE/SPREAD sorted regressions.

The fact that the significance and even the sign of the coefficients of the firm characteristic variables can vary so drastically across the regressions using data on portfolios sorted according to different criteria points to the serious limitations of analyses of this type that employ portfolio returns as the basic unit of analysis. Roll (1977) has pointed out that when portfolios are used to test an asset-pricing model the results can support the model even when it is false because individual asset deviations from the pricing relation can cancel out in the formation of portfolios. In our context, this corresponds to a situation in which the portfolio formation procedure reduces the cross-sectional variation in a particular security characteristic and thus obscures its effect on expected returns.[15] Roll's observation suggests that, neglecting sampling variability, the null hypothesis that a given characteristic has no effect on returns should be rejected if the coefficient of the characteristic is significant for *any* portfolio formation procedure. On this basis, we would conclude that the effects of the variables BM, SPREAD, PINST, PRICE, RET6, and RET12 are positive and that the following variables have a negative effect on expected returns: TO, YLD, and RET3. As noted, S&P has a different sign in different regressions.

While the foregoing discussion suggests that, if a variable appears to be significant in the regression corresponding to *any* sort, then the null hypothesis of no effect should be rejected, Lo and MacKinlay (1990) argue that portfolio formation procedures based on some empirically motivated characteristic can lead to the spurious conclusion that such a characteristic (or, by extension, a characteristic that is correlated with it) has a significant effect on returns. Both the GLS and FM regresssions reported above are vulnerable to such criticism. Therefore, regressions that depend on portfolio formation are hard to interpret—significance can be explained away as spurious, following Lo and MacKinlay, and insignificance can be attributed to the portfolio formation used, following Roll. Despite this, it is notable that almost all analysis of the cross section of security returns has relied on the use of portfolios sorted according to a single criterion.[16] In the balance of this chapter, we turn to *individual security* regressions.

C. Individual Security Fama–MacBeth Regressions

First, we follow Fama and French (1992) in performing Fama–MacBeth style cross sectional regressions of excess returns of individual securities on the

security characteristics, *including the estimated factor loadings*. The factor loadings were estimated as follows. In December of each year, all NYSE stocks were allocated to one of ten size-ranked portfolios. Each size decile was then subdivided into ten portfolios on the basis of individual stock preranking loading on the first factor. The preranking factor loadings were estimated using the prior 24 to 60 monthly returns as available. After assigning stocks to size-factor loading portfolios in December, we calculated the equal-weighted monthly returns for the next 12 months, continuing this process for the entire sample period. Then, following Dimson (1979), for each portfolio, the factor loadings were estimated as the sum of the slopes in the regression of the portfolio returns on the current and the prior month's factor returns,[17] using the full sample of returns. The factor loadings assigned to individual stocks each year were then the loadings of the portfolio to which the stock was assigned that year.

Cross-sectional regressions were then estimated each month with the security excess returns as the dependent variable and the factor loadings and other characteristics as the independent variables. The coefficients were then averaged over time in the standard Fama–MacBeth manner, and the results are reported in table 7.5. The estimated coefficients of the (estimated) factor loadings are all insignificant, but the coefficients of several of the security characteristics are significant. These include the coefficients of SIZE (negative), TO (negative), SPREAD (negative), PINST (positive), S&P (positive), and RET12 (positive). Moreover, the intercept is large and significant, indicating that the model does not properly price securities with average characteristics (recall that the characteristic variables are expressed as deviations from the monthly means).

While these individual security regressions are free of the problems of portfolio formation we have discussed, we have no assurance that the procedure used to assign factor loadings to individual securities leads to errors in the factor-loading estimates that are orthogonal to the other security characteristics. Indeed, we suspect that they will be dependent on the security characteristics since we expect the true factor loadings to be related to the security characteristics, while the factor loadings we have assigned are independent of the individual security characteristics. To address this issue of errors in variables, a second modified Fama–MacBeth approach for individual securities was adopted. First, the risk-adjusted monthly return was estimated for each security using the preranking factor loadings estimated from the previous 24 to 60 months as in equation (4),[18] and then the risk-adjusted returns were regressed on the security characteristics each month as in equation (7). Note that since the security characteristics are measured largely without error, there is no errors-in-variables problem induced by this procedure, despite the fact that we are using individual security data. On the other hand, we are imposing the constraint that the rewards to factor risk be equal to the returns on the underlying factor portfolios. The cross-sectional regression coefficients were averaged, and the results are summarized in table 7.6. SIZE remains significantly negatively

Table 7.5. Fama–MacBeth regression estimates of equation (5) using individual security data. Coefficient estimates are time series averages of cross-sectional OLS regressions. The dependent variable is the security excess returns, and the independent variables are the C–K factors and firm characteristics, measured as the deviation from the cross-sectional mean in each period. Dimson betas with one lag are used. The sample and the variables are defined in tables 7.1 and 7.2. The bold coefficient on size is the result replacing TO by DVOL in the independent variables. All coefficients are multiplied by 100. *t*-statistics are in parentheses.

Intercept	0.828	NANAL	0.050
	(2.44)		(0.72)
β_1	0.063	DISP	–0.147
	(0.15)		(1.30)
β_2	–0.164	SPREAD	–0.219
	(0.37)		(2.87)
β_3	0.771	PINST	0.114
	(1.27)		(2.46)
β_4	0.000	S&P	0.279
	(0.00)		(2.63)
β_5	–0.520	PRICE	0.168
	(0.91)		(1.31)
SIZE	–0.248	YLD	2.24
	(3.50)		(1.14)
	–1.037	RET3	–0.599
	(1.07)		(1.11)
BM	0.070	RET6	–0.162
	(0.89)		(0.36)
TO /	–0.144	RET12	1.36
DVOL	(1.98)		(4.68)

associated with returns; BM now has a significant *negative* effect[19]; TO, PRICE, YLD, and the lagged return variables are all insignificant. However, the additional characteristics NANAL, DISP, and PINST are significant, and the most striking finding is that SPREAD has a highly significant *negative* effect on returns. The factor model does a good job of pricing securities with average characteristics since we observe that the intercept of the regression is small and insignificant.

Although our procedure should ensure that errors in the estimated factor loadings are uncorrelated with the security characteristics, to the extent that they are correlated, the coefficients in the cross-sectional regressions will be biased by an amount that is proportional to the factor realizations. There-

Table 7.6. Fama–MacBeth regression estimates of equation (7) using individual security data. Coefficient estimates are time series averages of cross-sectional OLS regressions. The dependent variable is the excess return adjusted for the C–K factors (Dimson betas with one lag are used). The independent variables are the firm characteristics, measured as the deviation from the cross-sectional mean in each period. The sample and the variables are defined in tables 7.1 and 7.2. The bold coefficient on size is the result replacing TO by DVOL in the independent variables. All coefficients are multiplied by 100. *t*-statistics are in parentheses.

Intercept	–0.028	PINST	0.179
	(0.21)		(3.30)
SIZE	–0.422	S&P	0.184
	(5.95)		(1.66)
	–0.315		
	(2.65)		
BM	–0.256	PRICE	0.069
	(2.51)		(0.43)
TO /	–0.108	YLD	2.33
DVOL	(1.30)		(1.20)
NANAL	0.214	RET3	–0.280
	(2.98)		(0.20)
DISP	–0.287	RET6	–0.929
	(2.53)		(0.85)
SPREAD	–0.534	RET12	0.778
	(5.24)		(1.92)

fore, as a final check on the robustness of our results, the monthly regression coefficients of each of the characteristics were regressed on the C–K factors,

$$c_{mt} = \gamma_{m0} + \sum_{k=1}^{5} \gamma_{mk} \tilde{F}_{kt} + \tilde{u}_{mt} \tag{10}$$

and the intercepts from these regressions, γ_{m0}, which are purged of any influence from the factor realizations, are reported in table 7.7 with their associated *t*-statistics. The results do not change very much.

Thus, we continue to find that SIZE is strongly and negatively related to returns, while RET12 is positively associated; we find no independent effect associated with PRICE or BM once account is taken of the C–K factors. Thus, while Daniel and Titman (1995) continue to find a significant BM effect in the presence of the Fama–French factors, it is not present in our (shorter) sample, once account is taken of the influence of the other

Table 7.7. Risk-adjusted Fama–MacBeth Regression Estimates of equation (7) using individual security data. Coefficient estimates are the intercept terms from equation (10), obtained by regressing the parameter estimates from the monthly cross-section regressions on the C–K factors. The sample and the variables are defined in tables 7.1 and 7.2. The bold coefficient on size is the result replacing TO by DVOL in the independent variables. All coefficients are multiplied by 100. *t*-statistics are in parentheses.

Intercept	–0.015	PINST	0.073
	(0.33)		(1.61)
SIZE	–0.299	S&P	0.273
	(4.91)		(2.65)
	–0.081		
	(0.91)		
BM	–0.048	PRICE	0.132
	(0.53)		(1.63)
TO /	–0.217	YLD	2.89
DVOL	(2.87)		(1.56)
NANAL	0.202	RET3	1.99
	(2.75)		(1.45)
DISP	–0.177	RET6	–2.02
	(1.47)		(1.80)
SPREAD	–0.404	RET12	0.957
	(4.13)		(2.27)

variables. Other new findings include a significant negative relation between TO and returns, which is consistent with a liquidity effect, and a positive effect associated with NANAL, which is inconsistent with the 'neglected firm effect' hypothesis. Interestingly, we find that S&P membership is associated with positive risk-adjusted returns; this is perhaps due to the growth in index funds during our sample period (see Chan and Lakonishok 1993). There is no YLD effect. The most significant finding is the negative coefficient associated with the SPREAD variable, which has a *t*-statistic in excess of 4 even after adjustment by regression on the factors.

To check whether the SPREAD results are being driven by securities with extreme values of the spread, we formed spread-sorted quintiles each year and allowed the intercept in the cross-sectional Fama–MacBeth regression to depend on the spread quintile. Differences in the intercepts across quintiles would be indicative of nonlinearity in the return–spread relation. In fact, we found virtually no differences in the intercepts across the quintiles, suggest-

ing that the spread effect is not being driven by the stocks with extreme values of the bid–ask spread.

The risk-adjusted rewards for SIZE and TO in table 7.7 are of the same sign and similar magnitude. This suggests that it is the sum of these variables that is important; that is, the logarithm of dollar trading volume. To test this hypothesis, TO was replaced by the logarithm of the product of firm size and turnover to yield the variable DVOL, log dollar volume, and the regression results are reported in tables 7.6 and 7.7. Note that all the coefficients are the same except for that of SIZE (shown in bold in tables 7.5 through 7.10), which is no longer significant. This is not surprising given the fact that all our variables are in logarithms and DVOL = TO + SIZE. Considering the risk-adjusted results in table 7.7, we see that DVOL has a significant negative effect on returns, which is consistent with investors requiring a lower rate of return on liquid stocks; moreover, the coefficient of the SIZE variable is now small and insignificant. Thus, our results are consistent with the much analyzed 'size effect' being a liquidity effect.

It is evident that the SPREAD variable is not proxying for liquidity in the sense of Amihud and Mendelson (1986). Nor is it likely that SPREAD is associated with a negative bias in measured returns; in fact, any bias associated with bid–ask bounce is likely to be positive. We are not sure what role this variable plays. The finding of Petersen and Fialkowski (1994) that the quoted spread is only loosely associated with the effective spread that investors pay further raises the issue of how the quoted spread is set. Tinic (1972) presents a statistical model that explains 84% of the variance in spreads of NYSE stocks. The factors that are statistically significant in his regression are price, trading volume, number of institutional investors, an index of concentration of trading on the NYSE, and the fraction of days on which the stock was traded. Since we have already included price, turnover, and proportional institutional ownership as firm characteristics, SPREAD cannot be proxying for these variables in our regressions. Concentration of trading on the NYSE and thin trading[20] hardly seem sufficient to account for the importance of the SPREAD effect on returns that we find.

To further investigate the robustness of the SPREAD effect, the analysis was repeated using the three Fama–French (1993) factors to risk-adjust returns. The results, which are reported in tables 7.8 and 7.9, show that the coefficient of SPREAD remains negative and significant, although there are changes in some of the other coefficients. Thus (from table 7.9), NANAL and S&P are no longer significant, while PINST is now only marginally significant.

As a final check on the robustness of the spread effect, Fama–MacBeth regressions were run using individual security excess returns *without any risk adjustment* as the dependent variable. The results are reported in table 7.10 SPREAD remains strongly significant. Thus, there is no evidence that the SPREAD effect is due to the risk-adjustment procedure.

Table 7.8. Fama–MacBeth regression estimates of equation (9) using individual security data. Coefficient estimates are time series averages of cross-sectional OLS regressions. The dependent variable is the excess return adjusted for the F–F factors (Dimson betas with one lag are used). The independent variables are the firm characteristics, measured as the deviation from the cross-sectional mean in each period. The sample and the variables are defined in tables 7.1 and 7.2. The bold coefficient on size is the result replacing TO by DVOL in the independent variables. All coefficients are multiplied by 100. *t*-statistics are in parentheses.

Intercept	0.835	PINST	0.115
	(2.06)		(2.40)
SIZE	–0.256	S&P	0.280
	(4.13)		(2.63)
	–0.107		
	(1.10)		
BM	0.091	PRICE	–0.170
	(1.15)		(1.29)
TO /	–0.148	YLD	0.980
DVOL	(1.93)		(0.41)
NANAL	0.033	RET3	–0.653
	(0.47)		(1.19)
DISP	–0.132	RET6	–0.113
	(1.26)		(0.24)
SPREAD	–0.213	RET12	1.30
	(2.78)		(4.42)

5. CONCLUSION

This chapter makes several contributions to the growing literature on the determinants of the cross section of security returns. First, we have brought together the attributes approach of Fama and French (1992) and the factors approach of Connor and Korajczyk (1988) to examine the relationship between returns and 14 firm characteristics after accounting for the Connor and Korajczyk factors. Second, we have shown that the use of portfolio data to test asset-pricing models is fraught with difficulty—the results are highly sensitive to the particular portfolio formation criteria that are used. Finally, we have discussed how data on individual securities may be used to test a given asset-pricing model against a specific alternative hypothesis.[21] Using data on NYSE securities for the period 1977–1989, we have found that, after adjustment for risk by the Connor–

Table 7.9. Risk Adjusted Fama–MacBeth regression estimates of equation (9) using individual security data. Coefficient estimates are the intercept terms from equation (10) obtained by regressing the parameter estimates from the monthly cross-section regressions on the Fama–French factors. The sample and the variables are defined in tables 7.1 and 7.2. The bold coefficient on size is the result replacing TO by DVOL in the independent variables. All coefficients are multiplied by 100. As the Durbin–Watson statistic for the dividend yield regression showed significant evidence of serial correlation in residuals, this coefficient is estimated assuming an AR(2) process for the residuals. *t*-statistics are in parentheses.

Intercept	–0.033	PINST	0.099
	(0.56)		(1.99)
SIZE	–0.216	S&P	0.165
	(3.89)		(1.54)
	–0.001		
	(0.02)		
BM	–0.015	PRICE	–0.055
	(0.24)		(0.45)
TO /	–0.215	YLD	2.00
DVOL	(3.11)		(1.03)
NANAL	0.071	RET3	0.476
	(0.98)		(0.92)
DISP	–0.167	RET6	–0.148
	(1.53)		(0.30)
SPREAD	–0.269	RET12	1.29
	(2.91)		(4.22)

Korajczyk factors, security returns are reliably related to the following security characteristics: the dollar volume of share trading, analyst following, S&P membership, spread, and the 12-month lagged return. We find no significant evidence of a relation to firm size, the book-to-market ratio, dispersion of analyst forecasts, institutional ownership, price, dividend yield, or 3- and 6-month lagged returns.

Our most puzzling finding is the very strong negative relation between returns and the measured bid–ask spread. This relation appears to be constant over time and is (log)linear across different levels of the spread. We have no satisfactory explanation for this phenomenon, which clearly warrants further research.

ACKNOWLEDGMENTS We thank Larry Harris, Bob Korajczyk, Jeremy Evnine (formerly of Wells Fargo Investment Advisors), Marc Reinganum, Hans Stoll, and Robert Whaley, who all of whom provided data used in this study. We also thank

Table 7.10. Fama–MacBeth regression estimates of equation (9) using individual security data. Coefficient estimates are time series averages of cross-sectional OLS regressions. The dependent variable is the risk-unadjusted excess return. The independent variables are the firm characteristics, measured as the deviation from the cross-sectional mean in each period. The sample and the variables are defined in tables 7.1 and 7.2. The bold coefficient on size is the result replacing TO by DVOL in the independent variables. All coefficients are multiplied by 100. *t*-statistics are in parentheses.

Intercept	0.835	PINST	0.115
	(2.06)		(2.40)
SIZE	–0.256	S&P	0.280
	(4.13)		(2.63)
	–0.107		
	(1.10)		
BM	–0.091	PRICE	0.170
	(1.15)		(1.29)
TO /	–0.148	YLD	0.980
DVOL	(1.93)		(0.41)
NANAL	0.033	RET3	–0.653
	(0.48)		(1.19)
DISP	–0.132	RET6	–0.113
	(1.26)		(0.24)
SPREAD	–0.213	RET12	1.30
	(2.78)		(4.42)

Matthew Richardson and participants at the 1997 Utah Winter Conference and at Tulane University for comments, and Christoph Schenzler for providing programming assistance. We gratefully acknowledge the contribution of I/B/E/S International, Inc., for providing data on earnings forecasts. We also thank Eugene Fama and Richard Roll for helpful comments on an earlier draft of this chapter. We are responsible for any remaining errors.

NOTES

1. Connor (1995), who discusses the three approaches, refers to the third approach as a "fundamental factor model." He compares the three approaches in terms of their ability to explain the covariance matrix of U.S. stock returns: the model with five macroeconomic factors explains 10.9% of the variance, a (five) statistical factor model explains 39.0%, and his "fundamental factor model" explains 42.6%. Note that this ranking conveys little information about the ability of the different approaches to explain pricing.

2. See Chan and Lakonishok (1993).

3. However, when returns are risk-adjusted using the Fama–French factors, the coefficient of the dividend yield becomes positive and significant.

4. The *t*-statistic is an estimate of the Sharpe ratio for a particular portfolio, and MacKinlay (1995) places bounds on the maximum Sharpe ratio for a risky portfolio.

5. Fama and French (1993) also find that the book-to-market effect disappears when the risk model contains a factor based on the returns on portfolios formed on the basis of the book-to-market ratio.

6. Banz (1981) and Fama and French (1992).

7. Lehmann and Modest (1988) found that their implementation of a five-factor APT was unable to account for the size anomaly.

8. Falkenstein (1996) shows that mutual funds "show an aversion to low-price stocks."

9. The empirical work of Arbel et al. (1983) also suggests the existence of a "neglected firm" effect in expected returns.

10. Connor and Korajczyk (1993) "find evidence for one to six pervasive factors generating returns on the NYSE and AMEX over the period 1967 to 1991." We are grateful to Bob Korajczyk for providing us with updated estimates of the factors.

11. See Connor and Korajczyk (1988), for example, for the definition of an approximate factor model.

12. Fama and French (1992) state that their results are not sensitive to whether the factor loadings are calculated using the full-period postranking betas or 5-year preranking betas. Following these authors, we use the full-period betas in our cross-sectional regressions.

13. Roll (1994) follows a similar procedure with eight portfolios but uses dummy variables to represent three security characteristics.

14. To conserve space, the estimated factor loadings, β_{jk}, are not reported.

15. In addition, the portfolio formation procedure may induce multicollinearity between the average security characteristics in the portfolios, making it even more difficult to detect the influence of particular characteristics on expected returns.

16. Fama and French (1992) is a notable exception.

17. This procedure allows for thin trading that would cause biases in estimated factor loadings.

18. The Dimson procedure with one lag was used.

19. Note that this is despite the fact that our procedure implicitly assumes that the book value is known as of the end of the calendar year, whereas it may be reported only later. Our assumption may impart an upward bias to this coefficient.

20. Tinic (1972) finds that the *less* continuous the trading, the higher the spread.

21. Bossaerts and Hillion (1995) show how to use individual security data to test for the mean–variance efficiency of a portfolio.

REFERENCES

Amihud, Y., and H. Mendelson, 1986, Asset pricing and the bid–ask spread, *Journal of Financial Economics* 17, 223–249.

Arbel, A., S. Carvell, and P. Strebel, 1983, Giraffes, institutions and neglected firms, *Financial Analysts Journal* 39, 57–63.

Banz, R. W.,1981, The relationship between return and market value of common stocks, *Journal of Financial Economics* 9, 3–18.

Berk, J., 1995, A critique of size related anomalies, *Review of Financial Studies* 8, 275–286.

Black, Jensen, and Scholes, 1972, "The Capital Asset Pricing Model: Some Empirical Tests" in Michael C. Jensen (ed.) *Studies in the Theory of Capital Markets* (New York: Praeger).

Bossaerts, P., and P. Hillion, 1995, Testing the mean variance efficiency of well-diversified portfolios in very large cross-sections, *Annales d'Economie et de Statistique* 40, 93–119.

Brennan, M. J., 1970, Taxes, market valuation, and corporate financial policy, *National Tax Journal* 4, 417–427.

Brennan, M. J., 1994, The individual investor, *The Journal of Financial Research* 18, 59–74.

Brennan, M. J., and A. Subrahmanyam, 1996, Market microstructure and asset pricing: On the compensation for illiquidity in stock returns, *Journal of Financial Economics* 41, 341–364.

Chan, L. K. C. and J. Lakonishok, 1993, Are reports of beta's death premature?, *Journal of Portfolio Management* 19, 51–62.

Chen, N., 1983, Some empirical tests of the theory of arbitrage pricing, *Journal of Finance* 38, 1393–1414.

Chen, N., R. Roll, and S. Ross, 1986, Economic forces and the stock market, *Journal of Business* 59, 383–404.

Connor, G., 1995, The three types of factor model, *Financial Analysts Journal* 51, 42–46.

Connor, G., and R. A. Korajczyk, 1988, Risk and return in an equilibrium APT: Application of a new test methodology, *Journal of Financial Economics* 21, 255–290.

Connor, G., and R. A. Korajczyk, 1993, A test for the number of factors in an approximate factor model, *Journal of Finance* 48, 1263–1291.

Daniel, K., and S. Titman, 1997, Evidence on the characteristics of cross sectional variation in stock returns, *Journal of Finance* 52 (March), 1–33.

Dimson, E., 1979, Risk measurement when shares are subject to infrequent trading, *Journal of Financial Economics* 7, 197–226.

Eleswarapu, V. R., and M. Reinganum, 1993, The seasonal behavior of the liquidity premium in asset pricing, *Journal of Financial Economics* 34, 373–386.

Epstein, L., and S. Zin, 1989, Substitution, risk aversion, and the temporal behavior of consumption and asset returns, *Econometrica* 57, 937–969.

Falkenstein, E. G., 1996, Preferences for stock characteristics as revealed by mutual fund holdings, *Journal of Finance* 51, 111–135.

Fama, E. F., and K.R. French, 1992, The cross section of expected stock returns, *Journal of Finance* 47, 427–466.

Fama, E. F., and K. R. French, 1993, Common risk factors in the returns on stocks and bonds, *Journal of Financial Economics*, 33, 3–56.

Fama, E. F., and J. MacBeth, 1973, Risk and return: Some empirical tests, *Journal of Political Economy* 81, 607–636.

Frankel, R. and C. M. C. Lee, 1996, Accounting valuation, market expectation, and the book-to-market effect, *Journal of Finance* 51 (July), 1040.

Harris, L. 1994, Minimum price variations, discrete bid–ask spreads, and quotation sizes, *Review of Financial Studies* 7, 149–178.

Haugen, R., and N. Baker, 1996, Commonality in the determinants of expected stock returns, *Journal of Financial Economics* 41, 401–439.

Jarrow, R. 1980, Heterogeneous expectations, restrictions on short sales, and equilibrium asset prices, *Journal of Finance* 35, 1105–1113.

Jegadeesh, N., and S. Titman, 1993, Returns to buying winners and selling losers: Implications for stock market efficiency, *Journal of Finance* 48, 65–92.

Lehmann, B., and D. A. Modest, 1988, The empirical foundations of the arbitrage pricing theory, *Journal of Financial Economics* 21, 213–254.

Litzenberger, R. H., and K. Ramaswamy, 1979, Dividends, short-selling restrictions, tax-induced investor clienteles and market equilibrium, *Journal of Financial Economics* 7, 163–196.

Lo, A. W., and A. C. MacKinlay, 1990, Data-snooping biases in tests of financial asset pricing models, *Review of Financial Studies* 3, 431–468.

MacKinlay, A. C., 1995, Multifactor models do not explain deviations from the CAPM, *Journal of Financial Economics* 38, 3–28.

Merton, R. C. 1987, A simple model of capital market equilibrium with incomplete information, *Journal of Finance* 42, 483–510.

Miller, E. M. 1977, Risk, uncertainty and divergence of opinion, *Journal of Finance* 32, 1151–1168.

Miller, M. M., and M. S. Scholes, 1978, Dividends and taxes, *Journal of Financial Economics* 6, 333–364.

Petersen, M., and D. Fialkowski, 1994, Posted versus effective spreads: Good prices or bad quotes, *Journal of Financial Economics* 35, 269–292.

Pontiff, J., 1995, Closed-end fund premia and returns: Implications for financial market equilibrium, *Journal of Financial Economics* 37, 341–370.

Roll, R., 1977, A critique of the asset pricing theory's tests: On past and potential testability of theory, *Journal of Financial Economics* 4, 129–176.

Roll, R., 1994, Style return differentials: Illusions, risk premia, or investment opportunities?, in *Handbook of Equity Style Management*, T. B. Coggin and F. Tabotti (eds.), 1995 (New York: Wiley and Sons).

Roll, R., and S. A. Ross, 1980, An empirical investigation of the arbitrage pricing theory, *Journal of Finance* 35, 1073–1103.

Rosenberg, B., 1974, Extra-market components of covariance in security returns, *Journal of Financial and Quantitative Analysis* 44, 29–42.

Rosenberg, B., K. Reid, and R. Lanstein, 1985, Persuasive evidence of market inefficiency, *Journal of Portfolio Management* 11, 9–17.

Sharpe, W. F., 1982, Factors in NYSE security returns 1931–79, *Journal of Portfolio Management* 8, 5–19.

Stattman, D., 1980, Book values and stock returns, *The Chicago MBA: A Journal of Selected Papers* 4, 25–45.

Tinic, Seha M., 1972, "The Economics of Liquidity Services," *Quarterly Journal of Economics* 86 (February), 79–93.

8

On Cross-Sectional Determinants of Expected Returns

Bruce N. Lehmann

1.

It is both a pleasure and a privilege, albeit somewhat daunting ones, to discuss Brennan, Chordia, and Subrahmanyam's chapter "Cross-Sectional Determinants of Expected Returns" in this volume (see chap. 7). It is a pleasure because this is an interesting essay, one that challenges both the reliability of recent empirical evidence on risk/return relations obtained by Fama and French (1992) and others and the interpretation of postulated risk exposures in factor-pricing models. It is an honor because Fischer was one of the most extraordinary students of all areas of financial economics and of most of economics more broadly defined. I am somewhat intimidated, however, for Fischer was justly famous for his voracious reading of an extraordinarily large fraction of the published and unpublished research in finance and economics available at any point in time along with his detailed, insightful commentary on those aspects of any analysis he found either compelling or wanting.[1] Fischer set a very high bar indeed for commentary on papers prepared for a conference in his honor.

One aspect of Fischer's commentary on empirical work on asset-pricing relations seem relevant to the chapter by Brennan et al.: his insistence on rigorously interpreting evidence within the theoretical framework of the Capital Asset Pricing Model (CAPM). Now, I would be forsaking both comparative and absolute advantage were I to try to articulate what Fischer might have said about the Brennan et al. chapter if only because I lack his remarkable ability to provide cogent economic interpretations of empirical evidence in terms of false models such as the CAPM. Nevertheless, my own approach to empirical work is not entirely dissimilar to Fischer's because I try to rigorously interpret evidence by explicating what is being measured to (and, perhaps, past) the point of pedantry.

What do these high-sounding principles have to do with this chapter? Much of the empirical asset-pricing literature of the 1990s has revolved around the extent to which the security attributes that generated anomalies in the CAPM should be thought of as microdeterminants of risk exposures or as fundamental flaws in the fabric of the no-arbitrage or equilibrium

approaches to pricing capital assets in perfect markets. Chapter 7 fills a void in the literature by simultaneously considering a large set of asset characteristics in the context of a single benchmark asset-pricing model, the Connor and Korajczyk (1988) implementation of the Arbitrage Pricing Theory (APT). The inclusion of a laundry list of security attributes holds out the hope of sorting out the wheat from the chaff by harvesting a consistent and coherent description of mispricing in conventional asset-pricing models.

My goal is to suggest dimensions in which Brennan et al. can better achieve these goals. The next section discusses the difficulty in using their empirical specification to distinguish the microdeterminants of risk exposures and mispricing interpretations of the role of security attributes. The third section suggests a simple procedure for simultaneous measurement of individual security risk exposures and characteristic risk premiums, an approach that can mitigate any concerns about sequential estimation of these parameters. The penultimate section addresses two problems associated with the inference in this literature: the extreme fragility of inference about characteristic risk premiums based on grouped portfolios as documented by Brennan et al. and the data-mining concerns engendered by using attributes found largely through specification searches in the literature, a subject of great concern to Fischer. Brief concluding remarks round out my comments.

2. WHAT THESE LINEAR ASSET-PRICING RELATIONS MEASURE

It is profitable to work through the arithmetic of equations (3)–(5) in chapter 7 in order to be precise about what is being measured in these asset-pricing relations. In a vector version of equation (5) in chapter 7, the model for excess returns is

$$R_t - \iota R_{ft} = \iota\beta_0 + BF_t + Z_t c + e_t$$

where ι is an N-vector of ones. Letting I_{t-1} denote information available to investors at time t–1 with $Z_t \in I_{t-1}$, conditional mean returns are given by

$$E[R_t - \iota R_{ft} \mid I_{t-1}] = \iota\beta_0 + B\lambda_t + Z_t c + E[e_t \mid I_{t-1}]; \qquad \lambda_t = E[F_t \mid I_{t-1}]$$

which is satisfied arithmetically given only the existence of the relevant conditional means. Brennan, Chordia, and Subrahmanyam append the standard, though nontrivial, assumption that $E[e_t \mid I_{t-1}] = 0$, a presumption they use to justify their estimation procedures but one not actually necessary for this purpose.

What exactly do the characteristic risk premiums c measure? The authors provide the obvious answer: c reflects the reward for bearing the risks associated with the security attributes Z_t, compensation that should be zero if the five-factor model spans conditional mean excess returns or, more precisely, if Z_t does not help account for any mispricing engendered by this model. Yet there is another interpretation, one that strikes me as equally plausible a priori. Note that a maintained hypothesis in this model is that

the conditional factor loadings B are constant. If these loadings, in fact, vary over time, Z_t might simply be proxying for time variation in conditional risk exposures.[2]

There is a sense in which time variation in conditional loadings is the more natural interpretation. If this market admits no arbitrage opportunities, the mean–variance efficient frontier has finite slope and returns satisfy

$$R_t - \iota R_{ft} = \beta_{pt}\,[R_{pt} - R_{ft}] + \varepsilon_{pt};\; E[\varepsilon_{pt} \mid I_{t-1}] = 0$$

where p is the maximum conditional Sharpe ratio portfolio of these assets. Now there are two sets of portfolio returns implicit in equation (5) of chapter 7: F_t, the excess returns of the Connor–Korajczyk basis portfolios, and c, the unconditional mean returns of zero net investment portfolios of these assets with weights that are linear combinations of the columns of Z_t since $Z_t'\iota = 0$.

Accordingly, actual returns and conditional expected returns also satisfy

$$R_t - \iota R_{ft} = \iota\beta_0 + BF_t + Z_t W_{Zt} Z_t' R_t + u_t; \qquad u_t = Z_t\,[c - W_{Zt} Z_t' R_t] + e_t$$
$$E[R_t - \iota R_{ft} \mid I_{t-1}] = [B\beta_{pFt} + Z_t\beta_{pZt}]\,E[R_{pt} - R_{ft} \mid I_{t-1}] + E[u_t \mid I_{t-1}]$$

where $W_{Zt} \in I_{t-1}$ is a linear combination of the characteristic basis portfolios (see the discussion in the following section) and β_{pFt} and β_{pZt} are the conditional betas of the Connor–Korajczyk and characteristic basis portfolios, respectively. If $E[u_t \mid I_{t-1}] = 0$, B and Z_t provide conditional beta models in the sense that

$$\beta_{pt} = B\beta_{pFt} + Z_t\beta_{pZt}$$

Put differently, a correctly specified (conditional) multifactor asset-pricing model can always be interpreted as a conditional beta model within the corresponding single beta pricing relation.

Parenthetically, ignoring the specification and data-mining issues discussed further below, Fischer would doubtless have tried to interpret any characteristic risk premiums he thought empirically reliable along these lines—that is, within the CAPM assuming that portfolio p was the market portfolio of all risky assets. If he adopted this view, he would have construed the characteristic risk premiums as reflections of the implicit betas of the excess characteristic portfolio returns on the market portfolio.[3] At minimum, he would have demanded consistency in the attribute premiums such as the same signs and similar premium estimates for those characteristics that, loosely speaking, reflected the same economic quantities. In fact, he would probably have doubted the empirical reliability of any premium that he could not rationalize in this fashion, although judging whether implicit betas make economic sense is not an entirely straightforward matter.

Alternatively, suppose the five-factor model did, in fact, price this collection of assets but with time-varying conditional factor loadings B_t as in

$$E[R_t - \iota R_{ft} \mid I_{t-1}] = B_t\lambda_t$$

a circumstance that arises when

$$B_t\lambda_t = B\lambda_t + Z_t c + E[e_t \mid I_{t-1}]$$

since β_0 is zero in this model. Given the unconditional projection

$$B_t - B = Z_t\Pi + V_t; \qquad Z_t'V_t = 0$$

the characteristic risk premiums are given by

$$c = \Pi\lambda_t - [Z_t' Z_t]^{-1}Z_t'E[e_t \mid I_{t-1}]$$

Under the authors' assumption that $E[e_t \mid I_{t-1}] = 0$, one would naturally model c as the time average of $\Pi\lambda_t$ (i.e., $c = \Pi E[\lambda_t]$) if λ_t is time-varying and as the appropriate compensation for bearing time-varying risk exposure when it is time-invariant.

In any event, there is a natural way to mitigate this ambiguity: incorporate a linear model for conditional loadings into the model specification.[4] That is, the excess return model (5) in chapter 7 can be rewritten in the general form

$$R_t - \iota R_{ft} = \iota\beta_0 + [B+Z_t\Pi]F_t + Z_t c + e_t$$

To be sure, the characteristic risk premiums c still might not be purged of the effects of time-varying conditional risk due to the potential correlation between the factor risk premiums and V_t, the unmodeled part of the conditional risk exposures B_t (i.e., $E[V_tF_t] = E[V_t\lambda_t]$ need not be zero). Nevertheless, most researchers would probably conclude that there is some mispricing or omitted risk factor related to the characteristics Z_t if c proved to be nonzero in this revised model.

3. THE MEASUREMENT OF RISK PREMIUMS

One of the admirable qualities of chapter 7 is the authors' insistence on using individual securities to estimate the characteristic risk premiums, as opposed to basing inference solely on results obtained from grouped portfolios. However, the authors share a common misapprehension: that the Fama–MacBeth procedure is necessary because it is not generally possible to estimate risk exposures and premiums simultaneously when the number of securities exceeds the number of time series observations in the absence of restrictions on the residual covariance matrix. My purpose in this section is to show both that inference for risk premiums can proceed in a straightforward way without such restrictions and that the difference between such inferences and those produced by the Fama–MacBeth procedure reflect distinctions in the premiums being estimated.

Before proceeding, it is worth noting that the Fama–MacBeth procedure might just as well be called the Black, Jensen, and Scholes procedure. Black, Jensen, and Scholes (1972) developed this procedure to estimate the zero beta rate, the intercept in the zero beta CAPM pricing relation, via cross-sectional regression. At each time t, the regressand was the vector of ex-

cess returns of portfolios sorted by prior period sample beta less the product of the excess return of a market index and the vector of portfolio beta estimates. The regressors were comprised of a vector of ones less the same portfolio beta estimates. The cross-sectional regression coefficient was the actual return of a portfolio constructed to have an estimated beta of zero. This simple idea—that least squares coefficients estimated from cross-sectional regressions of security returns on asset characteristics are portfolio returns—has provided many insights linking financial econometrics with asset-pricing theory. Fischer was justly proud of this contribution.

Under whatever name, the inference issue is easily explicated in the pooled time series/cross-sectional regression (8) in chapter 7:

$$\mathbf{R} = X\beta + \varepsilon; \qquad \Omega = E[\varepsilon\varepsilon']$$

where $\mathbf{R}$ is the $NT \times 1$ vector of excess individual security returns, X is a matrix comprised of the Connor–Korajczyk excess basis portfolio returns, the pricing intercept, and the security attributes, β is a conformable parameter vector consisting of the factor risk exposures in the first $5N$ elements and the zero beta rate and the characteristic risk premiums in the last 15 elements, and ε is an $NT \times 1$ residual vector. The authors assume that the errors are serially uncorrelated and homoskedastic with potentially nonzero contemporaneous correlation, resulting in $\Omega = \Sigma \otimes I_T$, where Σ is the contemporaneous covariance matrix of ε_t.

This is a set of seemingly unrelated regressions since the elements of X are not identical across securities at each time t due to the presence of the individual asset characteristics. It is natural to estimate β, ε, and Σ by ordinary least squares via

$$\hat{\beta}_{\mathrm{OLS}} = [X'X]^{-1} X'R; \quad \hat{\epsilon} = \mathrm{R} - \hat{\beta}_{\mathrm{OLS}}; \quad \hat{\epsilon} = \{\hat{\epsilon}_1' \ldots \hat{\epsilon}_{\mathrm{T}}'\}; \quad \hat{\Sigma} = \frac{1}{\mathrm{T}} \sum_{t=1}^{T} \hat{\epsilon}_t \hat{\epsilon}_t'$$

for use in the asymptotically efficient generalized least squares estimate:

$$\hat{\beta}_{\mathrm{GLS}} = [X' (\hat{\Sigma}^{-1} \otimes I_T) X]^{-1} X' \hat{\Sigma}^{-1} \otimes I_T)\mathbf{R}; \qquad \mathrm{Var}[\hat{\beta}_{\mathrm{GLS}}] = [X' (\hat{\Sigma}^{-1} \otimes I_T) X]^{-1}$$

where $\mathrm{Var}[\hat{\beta}_{\mathrm{GLS}}]$ is a consistent estimate of the asymptotic covariance matrix of $\hat{\beta}_{\mathrm{GLS}}$. Of course, this estimate is only feasible when $N<T$ and generally performs well only if $N<<T$.

However, there is an alternative to grouping if $N>T$ because inefficient methods can be applied to very large asset menus when their dimension renders efficient approaches infeasible. That is, it is possible to implement any weighted least squares estimator given a nonsingular weighting matrix $\hat{W}$, in this case of the form $\hat{A} \otimes I_T$ with $\hat{A}$ converging in probability to some nonsingular limit A. Feasible estimators include ordinary least squares with $W = I_{NT} = I_N \otimes I_T$ and conventional weighted least squares with $\hat{W} = [\mathrm{Diag}(\hat{\Sigma})]^{-1} \otimes I_T$, where Diag[•] denotes a diagonal matrix comprised of the diagonal elements of its argument. The weighted least squares estimator and a consistent estimate of its covariance matrix are given by[5]

$$\hat{\beta}_{WLS} = [X'\hat{W}X]^{-1}X'\hat{W}\mathbf{R}; \quad \text{Var}[\hat{\beta}_{WLS}] = [X'\hat{W}X]^{-1}[X'\hat{W}(\hat{\Sigma}\otimes I_T)\,\hat{W}X]^{-1}[X'\hat{W}X]^{-1}$$

which uses the same estimate of Σ but does not require that it be invertible. Of course, the finite sample reliability of $\text{Var}[\hat{\beta}_{WLS}]$ need not be good, but it is likely to be no better or worse than that of the Fama–MacBeth estimator.

The justification for the last statement is evident after contemplation of these expressions. Suppose that the factor risk exposures were known and included in the model so that X contained only the security characteristics and a vector of ones, while β included only the zero beta rate and the characteristic risk premiums. In this case, both the Fama–MacBeth risk premium estimator and the standard estimator of its covariance matrix are numerically identical to those of the corresponding weighted least squares estimator. Of course, it is possible that the simultaneous estimation of risk exposures and characteristic risk premiums results in an estimator with small sample properties drastically different from that of the Fama–MacBeth estimator, but this outcome strikes me as an unlikely one given the relative precision with which risk exposures are estimated.

There is a subtle difference in the nature of the risk premiums produced by the Fama–MacBeth and weighted least squares procedures. Holding constant the weighting matrices and the risk exposure and zero beta rate estimates (which would, of course, differ across estimation procedures), the Fama–MacBeth and weighted least squares estimators are given by

$$\hat{c}_{FM} = \frac{1}{T}\sum_{t=1}^{T}\left[Z_t'\hat{A}Z_t\right]^{-1} Z_t'\hat{A}\left[R_t - \iota R_{ft} - \iota\hat{\beta}_0 - BF_t\right]$$

$$\hat{c}_{WLS} = \left(\frac{1}{T}\sum_{t=1}^{T}Z_t'\hat{A}Z_t\right)^{-1}\frac{1}{T}\sum_{t=1}^{T}Z_t'\hat{A}\left[R_t - \iota R_{ft} - \iota\hat{\beta}_0 - BF_t\right]$$

As is readily apparent, these estimators differ with regard to when $Z_t'\,\hat{A}\,Z_t$ is inverted in the estimation procedure.

In either case, the risk premium estimates are the average returns of linear combinations of zero net investment portfolios. The portfolios implicit in the Fama–MacBeth estimator differ in each time period due to the weighting by $[Z_t'\,\hat{A}Z_t]^{-1}$. In contradistinction, those associated with the weighted least squares estimator are time-invariant, being the inverse of the time series mean of $Z_t'\,\hat{A}Z_t$, due to the implicit imposition of the restriction that the characteristic risk premiums are time-invariant as well. Which is superior depends on the source of the characteristic risk premiums in the population.

Finally, modifications along these lines need not materially change the risk premium estimates. In fact, ignoring the common practice of estimating betas in prior sample periods before running cross-sectional regressions, the Fama–MacBeth and weighted least squares procedures (with identical weighting matrices) will produce numerically identical risk exposure estimates if there is a security-specific intercept (i.e., the usual α coefficient) in the regression. Nevertheless, it seems sensible to use joint estimation if it can eliminate any concerns about inference.[6]

4. THE MEASURED RISK PREMIUMS

The main focus of chapter 7 is on the sample characteristic risk premiums and their relevance for interpreting asset-pricing anomalies. What should we make of these estimates even ignoring the changes that might transpire if either the time-varying risk exposure model or the simultaneous estimation procedure was employed? Does this experiment teach us about anomaly risk premiums or the methods we use to investigate them? Chapter 7 has much to say on both scores.

As the authors stress, the results are quite sensitive to the choices of asset menu and econometric procedures. Reading the tables is an exercise in confusion if one thinks that commonly employed methods measure *the* characteristic risk premiums. The risk premium estimates often change sign and significance across both grouping strategies and estimation procedures. Some of this variation is doubtless due to sample size—as the authors note, 13 years is a small sample for measuring unconditional risk premiums, as emphasized by Merton (1980). Similarly, the screening criteria dramatically reduce the number of firms in the sample. However, the main cause of variation is likely to be the implicit experimental design.

What asset menus facilitate the measurement of characteristic risk premiums in this model? The experimental design that sharpens the risk premium estimates is that which facilitates precise estimation of the parameters of the multivariate linear regression model.[7] Assets with small residual variances and modest cross-sectional correlations are clearly desirable. Considerable cross-sectional variation in security characteristics alone is insufficient; it is also necessary that the variation be sufficiently independent across securities as well so that the separate effects of the attributes—that is, their individual risk premiums—can be reliably ascertained.

These observations suggest that portfolio grouping *can* be beneficial. Grouping securities into a smaller number of portfolios reduces residual variance because of the obvious diversification effect. Similarly, ranking on characteristics creates cross-sectional variation in those included in the sort. However, there is no a priori reason to think that sorting on attributes creates portfolios with modest cross-sectional residual correlations. If anything, the difference between the Fama–MacBeth and seemingly unrelated regression results for grouped portfolios suggests the opposite is the case.[8] Moreover, the limitations of grouping are obvious; as more characteristics are added to the sort, the number of securities in each portfolio declines, causing a corresponding reduction in the diversification benefits.

The results reported in tables 7.3 and 7.4 suggest that the potential benefits of grouping are not realized in this application.[9] Put differently, chapter 7 provides further documentation of the tenuousness of inference based on portfolio grouping. I find it unsurprising that inferences about the characteristic risk premiums prove so sensitive to the choice of grouping variables as well as to the use of individual securities since such findings have

cropped up in the literature when researchers have bothered to look. What is surprising to me is the common reliance on grouping procedures when they are so unreliable.

The experimental design issues are somewhat different in the case of the individual security Fama–MacBeth regressions. The correlation matrix in table 7.2 suggests nontrivial linear dependence among a subset of the characteristics across all securities, but it does not provide enough information to indicate whether it is severe enough to cause econometric problems. My guess is that collinearity among asset characteristics is high among those securities with extreme values of the attributes—that is, those with large (in absolute value) magnitudes of the characteristics in the metric of their squares and cross products—that have the most influence on the risk premium estimates. For example, there are many small firms with high yields, large spreads, and low volume and turnover that are not in the S&P 500 Index, much followed by analysts, or owned by institutions.

Finally, Fischer would doubtless have asked: "Where do the characteristics come from?" All are the result of loose economic reasoning, some of it a priori but much of it ex post, and the quarter century search for attributes correlated with pricing errors in the CAPM and other asset-pricing models. One cannot judge by the sign or significance of risk premium estimates since anomalies only remain anomalous as long as they remain statistically important in different asset menus. The only sure way to tell if a surviving security characteristic is a consequence of data mining is to see if it succumbs to the law of large numbers, which probably involves a long wait due to the imprecision with which mean excess stock returns are measured.

The concern regarding data mining makes it difficult to judge the liquidity-based explanation offered by the authors.[10] They observe that the coefficients on the logarithms of size and turnover have the same sign and similar magnitudes, suggesting that their effects are additive and represent a liquidity effect. Maybe so, but the coefficient on analyst's dispersion is similar, too, and one would surely not add it to size. Moreover, this finding might not survive the addition of securities, particularly of small firms, that would arise if the authors dropped some of the characteristics from the experiment and, hence, loosened the screening criteria. Similarly, the passage of the bid/ask spread coefficient from large and negative to small and positive by lagging size, price, and yield by one month suggests that this is a good choice due to the common factor of the closing month-end price. However, the authors would probably have looked for other empirical explanations for the puzzling negative spread coefficient had this not worked, raising concerns that this is a sample, not a population, effect.

5. CONCLUDING REMARKS

Except for concerns about sample size, my comments on chapter 7 can be accurately summarized in terms of hypotheses about potential collinearity between the security attributes and the corresponding risk/return

characteristics. Perhaps security characteristics are collinear with risk exposures but not risk premiums. Perhaps the collinearity among expected returns, risk exposures, and characteristics varies so much across grouping procedures, estimation methods, and weighting matrices that it is hard to know which set of results, if any, to believe. Perhaps they are collinear with expected returns as well, but there is so much collinearity among the attributes that it is hard to tell which attributes have economically significant risk premiums. Perhaps risk premiums also differ across characteristics that loose economic reasoning suggests should reflect similar economic risks.

Some of these concerns can be addressed using different methods, and some are the irremediable difficulties associated with inductive learning. Nevertheless, chapter 7 is a good start on resolving some of the outstanding disputes in the literature regarding the role of security characteristics in asset pricing. Perhaps at the end of this part of the road, the liquidity-based attributes will remain economically and statistically important with the right signs and plausible magnitudes with the other characteristics insignificant in both senses. It will still be difficult to tell if their economic relevance reflects a need to substantially relax the perfect markets assumptions so central to modern asset-pricing theory or if the CAPM (or other asset-pricing model) interpretation Fischer would doubtless have championed made more intuitive economic sense. Nevertheless, such a finding will have further narrowed the domain of the competition between these alternative views.

NOTES

1. Many others have noted the joy associated with their first receipt of unsolicited comments from Fischer Black when they were unknown assistant professors or the great surprise they experienced on visiting him and finding that he had a file of comments on their papers. I offer another such anecdote. A few years ago, David Modest and I prepared a paper entitled "Market Structure and Liquidity on the Tokyo Stock Exchange" for the National Bureau of Economic Research Conference on The Industrial Organization and Regulation of the Securities Industry. While writing about some of the limitations on the inferences that could be drawn from the trade and quote record in the middle of this paper, I wanted to contrast the demand for immediacy in the form of market orders and the supply of liquidity in the form of limit orders. David and I both viewed the precommitment associated with limit orders as a defining characteristic of the supply of immediacy. However, Fischer most assuredly did not, and I must confess that, with malice aforethought, I wrote two offending paragraphs on this topic in a fashion that I imagined would most upset Fischer since the contra-Black view of the supply of immediacy best characterized what I wanted to say at the time. Some two weeks after we sent the paper off to the NBER, certainly well before it was widely circulated (I had sent out fewer than 20 copies of the paper by that time), I received an email message from Fischer that began with "When I start reading a paper like yours, I am flooded with basic questions. What is liquidity exactly? . . . What does it mean to supply liquidity? To demand liquidity?" and followed with a coherent argument that the precommitment associated with limit orders had zero value in equilibrium. While the correctness of his insights on the value of limit orders may be open to dispute, I still find it astonishing that Fischer found two paragraphs written in a fashion designed to irritate him in a narrowly circulated working paper.

2. The sometimes substantial differences between the raw and purged Fama–MacBeth risk premium estimates in tables 7.5 and 7.6 are consistent with this hypothesis.

3. One more comfortable with less restrictive no-arbitrage models would focus more generally on the covariances among the characteristics' portfolios.

4. Lehmann (1992) shows that linear models for conditional factor risk exposures naturally arise when the APT holds both conditionally and unconditionally.

5. I have glossed over some solvable implementation details, most notably the fact that many securities will have missing data for at least part of the sample.

6. The comparison of the Fama–MacBeth and GLS risk premium estimates based on grouped portfolios in chapter 7 shed little light on this question because of the OLS weighting of the former and the GLS weighting of the latter. Perhaps the sometimes substantial differences between the raw and purged risk premium estimates in tables 7.5 and 7.6 reflect potential gains from the use of joint estimation procedures.

7. Warga (1989) discusses these issues in a different multifactor asset-pricing model.

8. Other potential problems include those engendered by sample size.

9. The sorts reported in chapter 7 all involve size and one other variable; my conjecture is that the variation in estimated risk premiums would be greater still if sorting was carried out on other characteristic pairs or, for that matter, triplets.

10. The virtues of including many of the anomalies found in prior research to see if liquidity-based variables can systematically account for prior findings must be balanced against the cost that sampling variation in many variables facilitates the finding of spurious correlations in the sample.

REFERENCES

Connor, Gregory, and Robert A. Korajczyk, 1988, "Risk and Return in an Equilibrium APT: Application of a New Test Methodology," *Journal of Financial Economics* 21 (September), pp. 255–289.

Fama, Eugene F., and Kenneth R. French, 1992, "The Cross-Section of Expected Stock Returns," *Journal of Finance* 47 (June), pp. 427–465.

Lehmann, Bruce N., 1992, "Notes on Dynamic Factor Pricing Models," *Review of Quantitative Finance and Accounting* 2, pp. 69–87.

Merton, Robert C., 1980, "On Estimating the Expected Return on the Market," *Journal of Financial Economics* 8 (December), pp. 323–361.

Warga, Arthur D., 1989, "Cross-sectional Tests of Linear Factor Models: Some Diagnostic Checks on the Experimental Design," *Journal of Business and Economic Statistics* 7 (April), 191–198.

9

Exploring a Two-Factor, Markovian, Lognormal Model of the Term Structure of Interest Rates

Scott F. Richard

The purpose of this chapter is to better understand two-factor, Markovian, lognormal models of the term structure of interest rates. This class of models is useful in valuing and hedging fixed-income securities. As such, we find it a valuable tool in managing fixed-income portfolios.

The one-factor, Markovian, lognormal model of the instantaneous interest rate first appeared in Dothan (1978). The model became useful and popular because of the binomial tree solution invented by Black, Derman, and Toy (1990) (BDT). In the models of both Dothan and BDT, the logarithm of the short rate follows a random walk and does not mean-revert. The lack of mean reversion makes it difficult for the model to price interest rate options such as caps and swaptions. To remedy this problem, Black and Karasinski (1991) extended the binomial tree in BDT to include mean reversion.

The one-factor, Markovian, lognormal model proved extremely useful in practice. It was adopted by many Wall Street firms to price derivatives, mortgage securities, and other interest-rate contingent claims. Over time, however, single-factor models proved inadequate for valuing and hedging simultaneously fixed-income securities, as argued in Canabarro (1995). Wall Street and other financial industry participants are largely replacing single-factor models with two-factor Markovian models or with (non-Markovian) multifactor Heath, Jarrow, and Morton (1992) models.

To my knowledge, there are two papers, Chan (1992) and Tenney (1995), that contain a two-factor, Markovian, lognormal model of the term structure of interest rates. Both papers develop essentially the same model, as we show in an appendix, but with different representations of the two state variables. Both models have been developed into successful commercial applications. Chan's model is used in Salomon Brothers' Yield Book; Tenney's model is implemented by Chalke, Inc., in their insurance company models.[1]

An empirically important feature of these models and the one presented here is that the two state variables are in general (negatively) correlated.

We shall see in section 2 that autonomous, independent state variables restrict the correlation of spot interest rates with futures prices on interest rates to be too high to match empirical observations.[2]

Why has this model proven successful in portfolio management? The answer, I think, is that the model is a very parsimonious representation, which can simultaneously value and hedge many fixed-income securities. In general, valuation and hedging are the two main uses of a term structure model. First, we calibrate the model to the current yield curve, futures prices, and liquid option prices and then value other securities. Second, having calibrated the model, we can determine what combination of bonds, futures, and options will hedge the security.

A calibrated two-factor, Markovian, lognormal model has three desirable characteristics. The first is the shape of the term structure. The model can simultaneously approximate both the term structure of interest rates and the term structure of London Interbank Offered Rate (LIBOR) futures prices. The peak of the hump in the term structure is often at 20 years or more; this peak is difficult to match using a one-factor model. Of course, to match the term structure exactly, we must use time-dependent fudge factors, but these should be small and in some sense average to zero. (Otherwise, the shape of the endogenous term structures will become unrealistic over time.)

Term structure shocks are the second characteristic. An empirical analysis of the covariance matrix of monthly changes in the yields on zero-coupon bonds between 1986 and 1995 reveals that there are two (or three) statistically significant factors determining the changes in the term structure of interest rates.[3] Figures 9.1 and 9.2 show the first two principal components of the correlation matrix.[4] Together, these two factors account for about 97% of the variation in monthly yield curve changes. The first factor is commonly called a yield curve shift and the second a yield curve twist.[5] The shape of the yield curve shift is evidence of mean reversion in interest rates. The factor shocks for the two-factor, Markovian, lognormal model closely resemble those shown in figures 9.1 and 9.2.

The final characteristic is the term structure of volati1ity. Typically, the term structure of cap prices is determined as if implied Black volatility is first rising and then falling. Implied Black volatility for long-dated swaptions is usually less than for caps. The price of a forward cap resetting to at-the-money should not necessarily decline with the starting date. Taken together, these stylized facts are further evidence for mean reversion of interest rates but not for decreasing conditional short-rate volatility.

Other models may have these characteristics as well, but that needs to be verified before we use them in practice.

This chapter is organized as follows. Section 1 contains the general formulation of the one-factor, Markovian, lognormal model and a simple and intuitive numerical solution technique. Section 2 extends this model and numerical solution to two factors. Section 3 contains some illustrative output and a discussion of the model results. The Appendix shows how to map

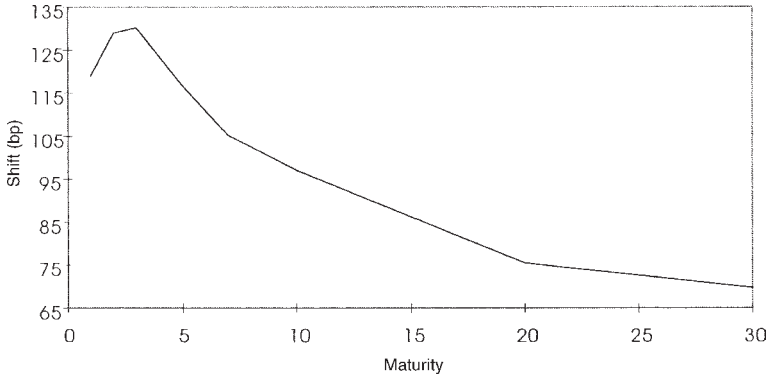


Figure 9.1. First Principal Component of the Correlation Matrix of Log Yield Changes (87.5%)

the models of Chan (1992), Tenney (1995), and Canabarro (1996) into our framework.

1. THE ONE-FACTOR, MARKOVIAN, LOGNORMAL MODEL

In this section, we review the one-factor, Markovian, lognormal model and present a new explicit numerical technique for finding contingent claims prices for the model. An explicit numerical technique is one where we build a grid or lattice to approximate the instantaneous interest rate. Perhaps the best-known explicit technique is the one-dimensional, recombining, binomial tree derived in Black, Derman, and Toy (1990). Their model was extended to accommodate mean reversion in Black and Karasinski (1991). We present an alternative numerical technique, a grid, which is extremely simple to build, intuitively understandable, and appears to have good

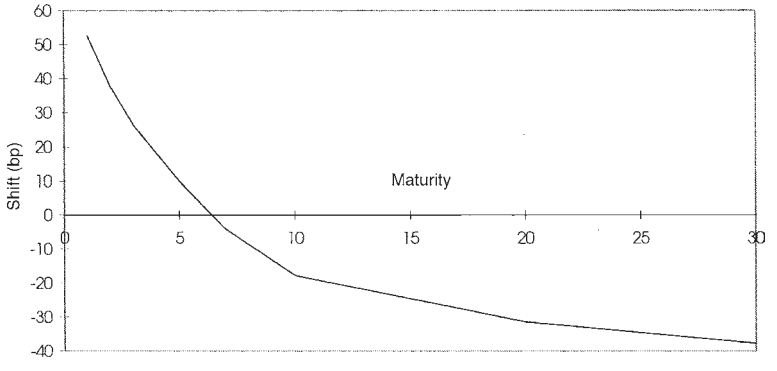


Figure 9.2. Second Principal Component of the Correlation Matrix of Log Yield Changes (9.5%)

numerical properties. Furthermore, the grid can easily be extended to the two-dimensional case.

The risk-adjusted model we consider is the same as the one in Black and Karasinski (1991):

$$dx(t) = m(h - x(t))dt + \sigma dw(t) \tag{1}$$

where $\ln(r(t)) = x(t)$, h, $m > 0$, and $\sigma > 0$ are constants, and $w(t)$ is a standard Wiener process.[6] The solution to equation (1) is

$$x(t) = h + \varphi(t)(x(0) - h) + \varphi(t)\int_0^t \varphi^{-1}(s)\sigma dw(s) \tag{2}$$

where $\varphi(t) = \exp(-mt)$. Asymptotically, $x(\infty)$ is normally distributed with mean h and variance $\sigma^2/(2m)$.

We want to build a discrete-time, binomial grid to approximate $x(t)$ in such a way that the conditional mean and variance of $x(t)$ are correct over the approximating interval. We assume it is currently time zero and divide time into equal intervals of length Δt. (Typically, we choose Δt equal to 1/12 or 1/24 of a year for actual applications.) Over an interval of length Δt, we calculate that the conditional mean of the change in x is

$$E_0\Delta x = (h - x(0))(1 - \beta) \tag{3}$$

where $\beta = \exp(-m\Delta t) < 1$. The conditional variance of the change in x is

$$V_0\Delta x = \sigma^2 \frac{(1-\beta^2)}{2m} \equiv v^2 \tag{4}$$

For small Δt, the conditional mean is approximately equal to $(h - x(0))m\Delta t$, and v^2 is approximately equal to $\sigma^2\Delta t$. It is convenient to define the normalized process

$$z(t) = (x(t) - h)/v \tag{5}$$

so that the conditional mean of the change in z is

$$E_0\Delta z = -z(0)(1 - \beta) \tag{6}$$

and the conditional variance of the change in z is one. Clearly, if we calculate $z(t)$, then we can invert equation (5) to find $x(t)$.

The grid approximating $z(t)$ is symmetric about $z_0 = 0$ the grid points are given by the formula

$$z_n = -z_{-n} = \sum_{i=1}^{n} \beta^{n-2i+1}, n = 0, 1, 2, \ldots \tag{7}$$

It is easily verified that $z_{n+1} > z_n$. Over an interval Δt, z_n either moves up to z_{n+1} or down to z_{n-1}.The probability of moving up is

$$p_n = \frac{\beta^n}{\beta^n + \beta^{-n}}, n = 0, 1, 2, \ldots \tag{8}$$

Naturally, the probability of moving downward is $1–p_n$. It can be verified by direct computation that Δz satisfies equation (6) and has a conditional

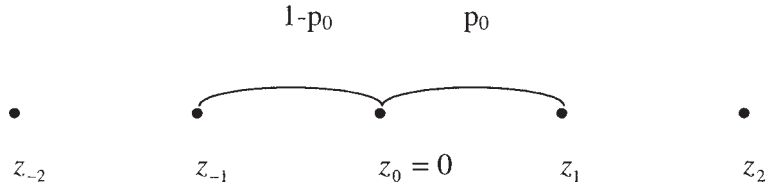


Figure 9.3. A Segment of the One Dimensional Grid with Transition Probabilities

variance of one. Figure 9.3 illustrates a section of the grid and the direction of the transition probabilities.

Intuitively, the grid is centered around the stationary mean of x (or z) rather than the current value. Conversely, the BDT tree is centered around the current value of x. This makes sense to me in that, with mean reversion, the future distribution of x is centered about h, while without mean reversion the future distribution is centered around $x(0)$. Also, the grid depends only on $m\Delta t$, so that it does not need to be recalculated if h or σ changes.

There are some practical considerations in actually building the grid (which can easily be done in a spreadsheet). First, the grid must be truncated at some N (chosen so that all interesting values of x and hence r are spanned). At the boundaries, we have the choice of making the grid either reflecting or absorbing. We will discuss how to do this in turn.

If the grid is to be made reflecting, then at the upper boundary node, z_N, we must redefine p_N as the probability of staying in z_N. In order to match the conditional mean given by equation (6), we calculate that

$$p_N = \frac{\beta^N}{\beta^N + (1-\beta)z_N} \tag{9}$$

Similarly, at the lower boundary node, z_{-N}, we redefine $(1-p_{-N}) = p_N$ to be the probability of staying in z_{-N} The conditional variances at the boundary nodes will be less than one, but this will have no practical effect on contingent claims pricing.

If the grid is to be made absorbing, then at the upper boundary node, z_N, we must redefine $p_N = 1$ as the probability of staying in z_N. Similarly, at the lower boundary node, z_{-N}, we redefine $(1-p_{-N}) = 1$ to be the probability of staying in z_{-N}. We have found that the absorbing boundary has better numerical properties when calculating the Green's functions (which are a proper subset of the Arrow–Debreu prices).

Having constructed the grid, contingent claims pricing proceeds in the obvious way. At each node, we use equation (5) to calculate the short rate $r_n = \exp(h + vz_n)$ and the discount function $d_n = \exp(-r_n\Delta t)$.[7] Denote by $\pi_n(t)$ the value at time $t = K\Delta t$ in node n of any contingent claim. If the cash flow for the claim is not path-dependent, then we can value it through backward induction. We begin at the maturity date, $T = K\Delta t$, of the claim where the cash flow determines $\pi_n(T)$. For example, for a zero-coupon bond ma-

turing at T, we have $\pi_n(T) = 1$, and for a caplet struck at s, we have $\pi_n(T) = \max(0, r_n - s)$. Valuation proceeds through backward induction:

$$\pi_n(t) = d_n(p_n\pi_{n+1}(t + \Delta t) + (1 - p_n)\pi_{n-1}(t + \Delta t)), n = 0, \pm 1, \pm 2, \ldots, \pm N - 1 \quad (10)$$

Figure 9.4 illustrates the backward induction from time $t + \Delta t$ to time t at node z_0.

At the boundaries, the backward induction is

$$\pi_N(t) = d_N(p_N\pi_N(t + \Delta t)(1 - p_N)\pi_{N-1}(t + \Delta t)) \quad (11a)$$

and

$$\pi_{-N}(t) = d_{-N}(p_{-N}\pi_{-N+1}(t + \Delta t) + (1 - p_{-N})\pi_{-N}(t + \Delta t)) \quad (11b)$$

The present value of the contingent claim is calculated by interpolating between the values at the two nodes straddling $r(0)$. The value of a path-dependent claim, such as a mortgage security, is calculated by simulating paths of interest rates from the grid.

In the following section, we extend the one-factor model and solution technique to a two-factor model.

2. THE TWO-FACTOR, MARKOVIAN, LOGNORMAL MODEL

In our model, the instantaneous interest rate is lognormally distributed, mean reverts, has a stationary steady-state distribution, and is described by two stochastic variables. We set

$$\ln(r(t)) = x_1(t) + x_2(t) \quad (12)$$

where x_1 and x_2 are jointly normally distributed with risk-neutral dynamics

$$dx = M(h - x)dt + Bdw(t) \quad (13)$$

In equation (13), h is a 2-vector of constants, M is a diagonal matrix with elements $m_2 > m_1 > 0$, B is a 2-by-2 matrix of constants such that $\Sigma = BB' > 0$, and w is a two-dimensional vector of independent Wiener processes. It is very important to note that x_1 and x_2 are, in general, correlated random variables. If we assume them to be independent, then it is not possible to

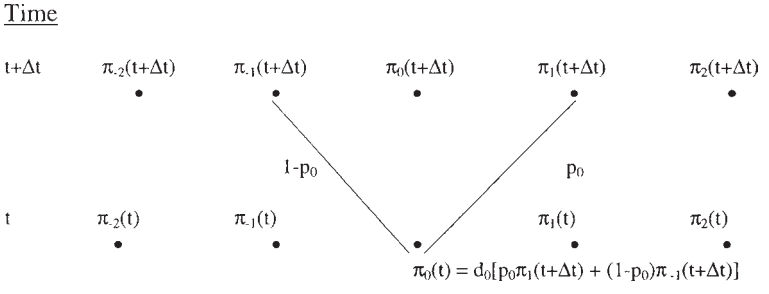


Figure 9.4. Backward Induction in the Grid

guarantee the existence of parameter values that match the covariance of the spot interest rate and futures on the interest rate, as we will see below.

Because M is diagonal, the stochastic differential equations given by (13) are autonomous and can be solved for $x(t)$. (In fact, it is easily shown that any pair of linear equations for $\ln(r)$ can be reduced to autonomous equations.) To solve equation (13), define $\Phi(t)$ to be a diagonal matrix with elements $\varphi_1(t) = \exp(-m_1 t)$ and $\varphi_2(t) = \exp(-m_2 t)$. Then we find that

$$x(t) = h + \Phi(t)(x(0) - h) + \Phi(t)\int_0^t \Phi^{-1}(s)Bdw(s) \tag{14}$$

From equation (14), we calculate the mean of $x(t)$ to be

$$Ex(t) = h + \Phi(t)(x(0) - h) \tag{15}$$

and the covariance matrix to be

$$Vx(t)\begin{bmatrix} v_{11}(t) & v_{12}(t) \\ v_{12}(t) & v_{22}(t) \end{bmatrix} \tag{16}$$

where $v_{11}(t) = (\sigma_{11}/2m_1)\,(1 - \exp(-2m_1 t))$, $v_{12}(t) = (\sigma_{12}/(m_1 + m_2))(1 - \exp(-(m_1 + m_2 t))$, and $v_{22}(t) = (\sigma_{22}/2m_2)(1 - \exp(-2m_2 t))$.

We calculate the moments of $\ln(r(t))$

$$E\ln(r(t)) = \mathbf{1}'h + \mathbf{1}'\Phi(t)(x(0) - h) \tag{17}$$

and

$$V\ln(r(t)) = v_{11}(t) + 2v_{12}(t) + v_{22}(t) \tag{18}$$

Asymptotically $\ln(r)$ is normally distributed with mean

$$E\ln(r(\infty)) = h_1 + h_2 = \eta \tag{19}$$

and variance

$$V\ln(r(\infty)) = \frac{\sigma_{11}}{2m_1} + \frac{2\sigma_{12}}{m_1 + m_2} + \frac{\sigma_{22}}{2m_2} \tag{20}$$

I have no formal proof, but experimentation shows that prices of interest-rate contingent claims depend only on η and not on its allocation to h_1 and h_2.

The dynamics for $\ln(r(t))$ are given by

$$d\ln r(t)) = \mathbf{1}'M(h - x(t))dt + \mathbf{1}'Bdw(t) \tag{21}$$

The instantaneous variance of $r(t)$ is a constant

$$\sigma_r^2 = \sigma_{11} + 2\sigma_{12} + \sigma_{22} \tag{22}$$

Using equation (12), we can eliminate one of the unobserved state variables in equation (21) but not both. We need to introduce a second observed state variable.

Our second state variable is the τ-year futures price of r denoted by $f(t; \tau)$. If we approximate r by 3-month LIBOR, then we might choose τ equal to 5 years and approximate $f(t; \tau)$ by the 5-year futures price for 3-month LIBOR. Each day, we would have to interpolate $f(t; \tau)$ between the two

contracts straddling 5 years until expiration. Under the risk-neutral distribution, we have

$$f(t;\tau) = E[r(t+\tau) \mid r(t)] \tag{23}$$

Calculating the conditional expectation on the right-hand side of equation (23), we find that

$$\ln(f(t;\tau)) = c(\tau) + \mathbf{1}'\Phi(\tau)x(t) \tag{24}$$

where

$$c(\tau) = \tfrac{1}{2}\,(v_{11}(\tau) + 2v_{12}(\tau) + v_{22}(\tau)) + \eta - \mathbf{1}'\Phi(\tau)h \tag{25}$$

Denoting $v(\tau) = v_{11}(\tau) + 2v_{12}(\tau) + v_{22}(\tau)$, we find that $\ln(f(t;\tau))$ is normally distributed with mean

$$E\ln(f(t;\tau)) = \tfrac{1}{2}\,v(\tau) + \eta + \mathbf{1}'\Phi\,(t+\tau)(x(0) - h) \tag{26}$$

and variance

$$V\ln(f(t;\tau)) = \varphi_1^2\,(\tau)v_{11}\,(t) + 2\varphi_1(\tau)\varphi_2\,(\tau)v_{12}\,(t) + \varphi_2^2\,(\tau)v_{22}\,(t) \tag{27}$$

We can study $f(t;\tau)$ asymptotically for fixed τ as t gets large or for fixed t as τ gets large. Asymptotically, as $t\to\infty$, $\ln(f(t;\tau))$ is normally distributed with mean

$$E\ln(f(\infty;\tau)) = \tfrac{1}{2}\,v(\tau) + \eta \tag{28}$$

and variance

$$V\ln\big(f(\infty;\tau)\big) = \varphi_1^2(\tau)\frac{\sigma_{11}}{2m_1} + \varphi_1(\tau)\varphi_2(\tau)\frac{2\sigma_{12}}{m_1+m_2} + \varphi_2^2(\tau)\frac{\sigma_{22}}{2m_2} \tag{29}$$

From equations (19) and (20) and from equations (26) and (27), we calculate that for fixed t

$$Er(\infty) = \exp\left[\eta + \frac{1}{2}\left(\frac{\sigma_{11}}{2m_1} + \frac{2\sigma_{12}}{m_1+m_2} + \frac{\sigma_{22}}{2m_2}\right)\right] = f(t;\infty) \tag{30}$$

(i.e., the value of all futures prices converges to the asymptotic expected instantaneous interest rate or, equivalently, the term structure of futures prices is asymptotically flat). Finally, we turn to the dynamics for $\ln(f(t;\tau))$. From equation (24), we calculate that

$$d\ln(f(t;\tau)) = \mathbf{1}'\Phi\,(\tau)M(h - x(t))dt + \mathbf{1}'\Phi(\tau)Bdw(t) \tag{31}$$

The instantaneous variance of $\ln(f(t;\tau))$ is a constant

$$\sigma_f^2 = \varphi_1^2\,(\tau)\sigma_{11} + 2\varphi_1(\tau)\varphi_2\,(\tau)\sigma_{12} + \varphi_2^2(\tau)\sigma_{22} \tag{32}$$

as is the instantaneous covariance of $\ln(r(t))$ and $\ln(f(t;\tau))$,

$$\sigma_{rf} = \varphi_1(\tau)\sigma_{11} + \varphi_1(\tau)\sigma_{12} + \varphi_2\,(\tau)\sigma_{12} + \varphi_2\,(\tau)\sigma_{22} \tag{33}$$

We now return to the claim that x_1 and x_2 must in general be correlated in order to match the correlation, ρ_{rf}, between the spot rate and the futures

rate. By inverting equations (22), (32), and (33), we find that $\rho_{12} = 0$ if and only if

$$\rho_{rf} = \frac{\varphi_1\varphi_2\sigma_r^2 + \sigma_f^2}{(\varphi_1 + \varphi_2)\sigma_r\sigma_f} \tag{34}$$

Since φ_1 and φ_2 must both be positive and less than one, we have $\rho_{rf} > \sigma_f/\sigma_r$, which is empirically too large a lower bound. For example, if $\tau = 5$ years, then ρ_{rf} is typically about 0.5, but σ_f/σ_r is typically greater than 0.65.

We are finally ready to derive an observable dynamic system for $\ln(r(t))$ and $\ln(f(t;\tau))$. Solving equations (12) and (24) simultaneously for $x(t)$, we get

$$x_1(t) = [\varphi_2(\tau)\ln(r(t)) - \ln(f(t;\tau)) + c(\tau)]/[\varphi_2(\tau) - \varphi_1(\tau)] \tag{35a}$$

and

$$x_2(t) = [-\varphi_1(\tau)\ln(r(t)) + \ln(f(t;\tau)) - c(\tau)]/[\varphi_2(\tau) - \varphi_1(\tau)] \tag{35b}$$

Equations (35) can be substituted into equations (21) and (31) to find a pair of internally consistent stochastic differential equations for $\ln(r(t))$ and $\ln(f(t;\tau))$.

We now turn to building the two-dimensional grid. We will roughly replicate the steps in building the one-dimensional grid. First, we calculate the conditional mean of the change in x using equation (15):

$$E_0\Delta x_i = (h_i - x_i(0))(1 - \beta_i),\ i = 1, 2 \tag{36}$$

where $\beta_i = \exp(-m_i\Delta t)$, $i = 1, 2$. Next, we calculate the conditional variance of the change in x using equation (16):

$$V_0\Delta x(t)\begin{bmatrix} v_{11}(\Delta t) & v_{12}(\Delta t) \\ v_{12}(\Delta t) & v_{22}(\Delta t) \end{bmatrix} \tag{37}$$

Finally, we define the normalized processes

$$z_i(t) = \left(x_i(t) - h_i\right)/\sqrt{v_{ii}(\Delta t)},\ i = 1, 2 \tag{38}$$

The conditional mean of the change in z is

$$E_0\Delta z_i = -z_i(0)(1 - \beta_i),\ i = 1,2 \tag{39}$$

and the conditional covariance matrix is

$$V_0\Delta z = \begin{bmatrix} 1 & \rho' \\ \rho' & 1 \end{bmatrix} \tag{40}$$

where

$$\rho' = \frac{v_{12}(\Delta t)}{\sqrt{v_{11}(\Delta t)v_{22}(\Delta t)}} \tag{41}$$

For small Δt, $\rho' \approx \rho \equiv \sigma_{12}/\sqrt{\sigma_{11}\sigma_{22}}$

The grid approximating the correlated processes $z_1(t)$ and $z_2(t)$ is simply the Cartesian product of the grids approximating each process. The grid points are given by the obvious specialization of equation (6) to β_1 and β_2. The probabilities must be modified to reflect the correlation of the processes. Denote a node of the grid by (n, j), where n is the index for z_1 and j is the index for z_2. Denote the probability of moving from (n, j) to (n', j') by $p_{nj}(n', j')$. Then we have that

$$p_{nj}(n+1, j+1) = (\beta_1^n \beta_2^j + \rho') \,/\, D \tag{42a}$$

$$p_{nj}(n+1, j-1) = (\beta_1^n \bar{\beta}_2^j - \rho') \,/\, D \tag{42b}$$

$$p_{nj}(n-1, j+1) = (\bar{\beta}_1^n \beta_2^j - \rho') \,/\, D \tag{42c}$$

$$p_{nj}(n-1, j-1) = (\bar{\beta}_1^n \bar{\beta}_2^j + \rho') \,/\, D \tag{42d}$$

where $D = (\beta_1^n + \bar{\beta}_1^n)(\beta_2^j + \bar{\beta}_2^j)$. Inspecting equations (42) shows that the probability of each of the processes moving up or down is unaltered from the one-dimensional case, so the conditional means automatically satisfy equation (39) and the conditional variances are each equal to one. Direct calculations will verify that the covariance between Δz_1 and Δz_2 equals ρ'.

Unlike the one-dimensional grid, the probabilities in (42) can become negative. When one of the probabilities as given by (42) is negative, we are at a corner point and must modify the probabilities. To show how this is done, we will assume that $\rho' < 0$, which is the case empirically. When $\rho' < 0$, $(\beta_1^n \beta_2^j + \rho')$ will become negative for large enough positive n and j. Assuming (n, j) is such a corner node, we replace (42) with

$$p_{nj}(n+1, j+1) = 0 \tag{43a}$$

$$p_{nj}(n+1, j-1) = \beta_1^n / (\beta_1^n + \beta_1^{-n}) \tag{43b}$$

$$p_{nj}(n-1, j+1) = \beta_2^j / (\beta_2^j + \beta_2^{-j}) \tag{43c}$$

$$p_{nj}(n-1, j-1) = (\bar{\beta}_1^n \bar{\beta}_2^j - \beta_1^n \beta_2^j) / D > 0 \tag{43d}$$

The choices of probabilities given in (43) will cause the conditional means of the change in z to satisfy equation (39), but the conditional covariance matrix will not be correct. At the boundary nodes of the grid, as opposed to the corners, we continue to make the grid absorbing.

3. DISCUSSION OF SOLUTIONS

In this section, we present some solutions of the model that are indicative of the term structure, cap, and swaption prices at the end of July 1996. The values of the state variables are $r = 5.61\%$ and $f = 7.13\%$, where f is the 5-year Eurodollar futures; the parameters of the model are $m_1 = 6\%$, $m_2 = 60\%$, $\sigma_r = 25\%$, $\sigma_f = 15\%$, $\rho_{rf} = 0.5$, and $\eta = -2.533$. The grid we built is 121 by 91 and has $\Delta t = 1/24$ year. We calculate that $\sigma_1 = 20.86\%$, $\sigma_2 = 24.65\%$, and $\rho_{12} = -0.406$.

The computed term structure is shown in figure 9.5. Figures 9.6 and 9.7, respectively, show the first and second principal components of the co-

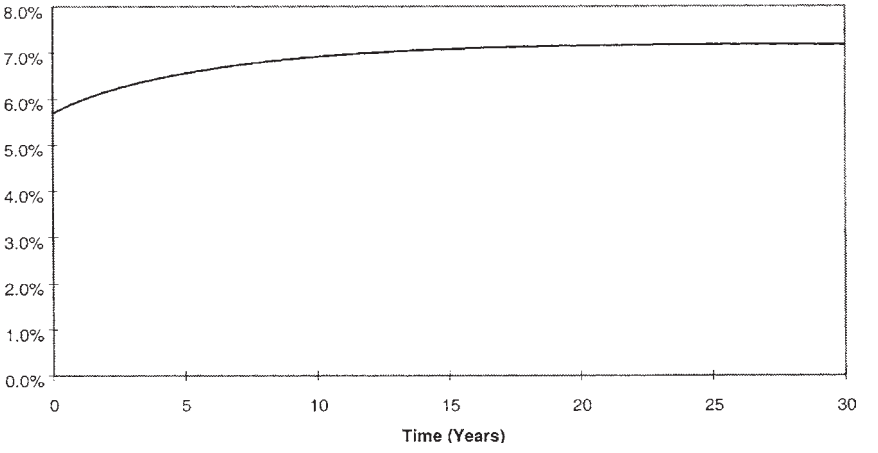


Figure 9.5. The Term Structure of Interest Rates

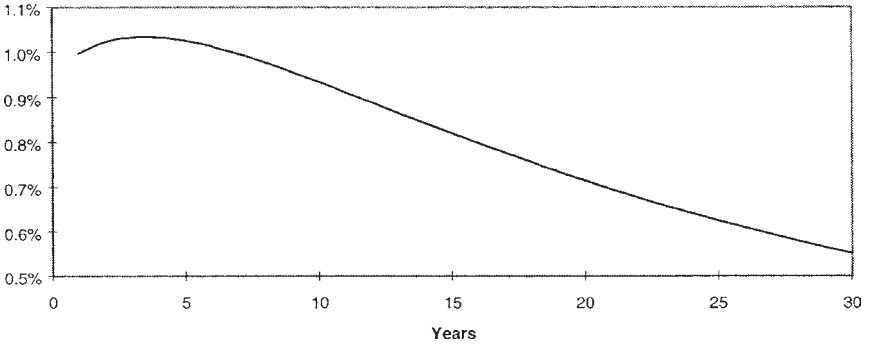


Figure 9.6. First Principal Component of the Covariance Matrix of Log Yield Changes

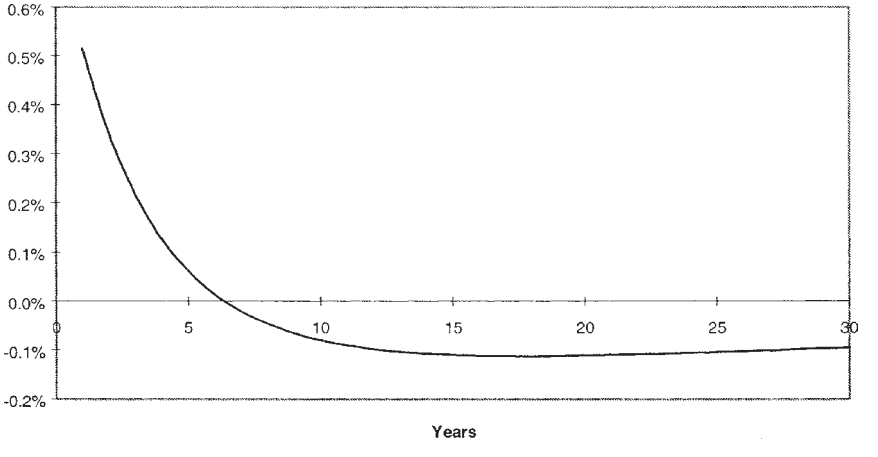


Figure 9.7. Second Principal Component of the Covariance Matrix of Log Yield Changes

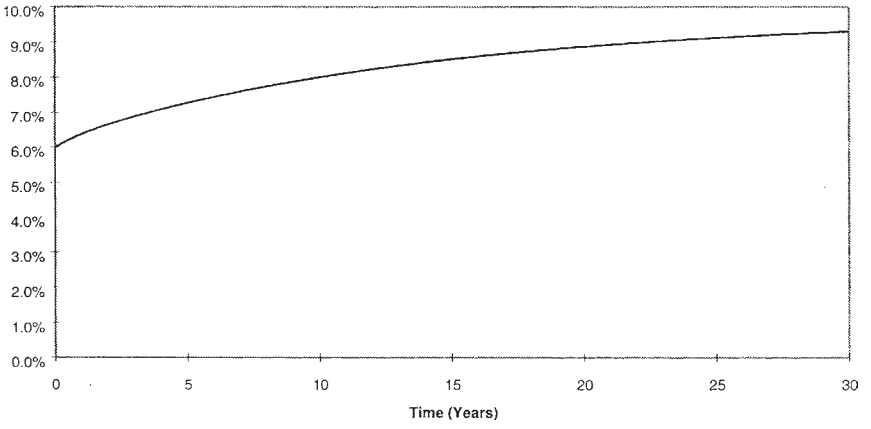


Figure 9.8. The Term Structure of Futures Prices for the Short Rate. The term structure of futures prices for the short term interest rate is shown in Figure 9.8. The five year rate was an input; the rest are calculated.

variance matrix of log yield changes. If actual log yield changes were generated by the model, then the observed empirical principal components would look like figures 9.6 and 9.7. (Of course, we cannot calculate more than two components since we have generated the data from a two-factor model.) The similarity of the model output (figures 9.6 and 9.7) to the actual empirical results (figures 9.1 and 9.2) should be noted.

The term structure of futures prices for the short-term interest rate is shown in figure 9.8. The 5-year rate was an input; the rest are calculated. Notice that the rate is asymptotically approaching 9.46%.

Finally, we turn to the evaluation of options. Table 9.1 shows the price and implied Black volatility of at-the-money caps of various maturities.

Table 9.1. At-the-money strikes, cap prices, and Black volatilities.

Maturity	Stike	Cap Price	Black Volatility
1	5.82%	29.30	28.9%
2	6.01%	85.70	35.9%
3	6.16%	152.93	24.7%
5	6.37%	305.58	23.6%
7	6.53%	463.98	22.0%
10	6.67%	688.10	21.5%

Table 9.2. Swaption prices, yield volatilities, and price volatilities.

Option Maturity	Swap Maturity	Strike	Price	Yield Vol.	Price Vol.
1	7	6.88%	2.19%	16.8%	5.8%
1	10	6.99%	2.69%	16.4%	7.2%
2	7	7.03%	2.89%	16.5%	5.8%
2	10	7.12%	3.54%	16.1%	7.1%
3	7	7.15%	3.29%	16.1%	5.8%
5	10	7.32%	4.33%	15.1%	6.7%

Notice that the implied Black volatilities rise from 1-year caps to 2-year caps and then decline, a pattern similar to market prices.

Swaption prices and volatilities are shown in table 9.2. The implied yield volatilities for longer-dated swaptions are indicative of market quotes. The two-factor, Markovian, lognormal model appears to be a successful representation of the term structure of interest rates. It mimics well the shape of the term structure, the principal components of the change in the term structure, the term structure of futures prices, the term structure of cap volatility, and the term structure of swaption volatility. Increasing experience in using it to buy and hedge securities should lead to the discovery of its shortcomings and hence improvement.

ACKNOWLEDGMENTS Thanks to Ben Gord, Jaidip Singh, and Eduardo Schwartz, the discussant, for useful suggestions and error corrections.

NOTES

1. Canabarro (1996), contains essentially the same model as Chan (1992) and Tenney (1995), but with the correlation between changes in the short rate and changes in the second factor restricted to be zero; a model implemented at Goldman Sachs.

2. There have been a number of papers using two uncorrelated square root interest rate processes. See, for example, Richard (1978), Cox, Ingersoll, and Ross (1985), Chen and Scott (1992), and Longstaff and Schwartz (1992).

3. The original study of factor shifts in the yield curve is Litterman and Scheinkman (1991). They find that the first factor, the yield curve shift, is nearly a parallel shift. If we restrict our data to their time period (1984–1989), we closely replicate their results.

4. The results of factor analysis, as opposed to principal components, are nearly identical.

5. The fact that so much of the variation in yields is described by two principal components by itself guides us to look at two-factor models. Another equally important consideration points us toward a two (or more)-factor model: We have been unable to find any single-factor model that qualitatively has a principal component in the shape shown in figure 9.1.

6. In Black and Karasinski (1991), h is a time function used to match exactly the current term structure of interest rates. We will discuss in note 7 how to accommodate this modification in our solution technique.

7. If you wish to match the current term structure of interest rates exactly, then the constant h is replaced with $h(t)$ for $t = 0, \Delta t, 2\Delta t, \ldots, T$. The algorithm given below can be used to first estimate $h(t)$ by matching the price of zero-coupon bonds before proceeding to contingent claims pricing.

REFERENCES

Black, F. E., E. Derman, and W. Toy (1990): "A One Factor Model of Interest Rates and its Application to Treasury Bond Options," *Financial Analysts Journal*, 46, January/February, 33–39.

Black, F. E., and P. Karasinski (1991): "Bond and Option Pricing when Short Rates Are Lognormal," *Financial Analysts Journal*, 47, July/August. 52–59.

Canabarro, E. (1995): "Where Do One-Factor Interest Rate Models Fail?," *The Journal of Fixed Income*, September, 31–52.

Canabarro, E. (1996): "A Two-Factor Markovian Interest Rate Model, Goldman Sachs and Co., New York, NY.

Chan, Y. K. (1992): "Term Structure as a Second Order Dynamical System, and Pricing of Derivative Securities," Salomon Brothers, Inc., New York, NY.

Chen, R. R., and L. Scott (1992): "Pricing Interest Rate Options in a Two-Factor Cox–Ingersoll–Ross Model of the Term Structure," *Review of Financial Studies*, 5, 613–636.

Cox, J. C., J. E. Ingersoll, and S. A. Ross (1985): "A Theory of the Term Structure of Interest Rates," *Econometrica*, 53, March, 385–407.

Dothan, L. U. (1978): "On the Term Structure of Interest Rates," *The Journal of Financial Economics*, 6, March, 59–69.

Heath, D., R. Jarrow, and A. Morton (1992): "Bond Pricing and the Term Structure of Interest Rates: A New Methodology for Contingent Claims Valuation," *Econometrica*, 60:1, January, 77–105.

Litterman, R., and J. Scheinkman (1991): "Common Factors Affecting Bond Returns," *The Journal of Fixed Income*, June, 54–61.

Longstaff, F., and E. S. Schwartz (1992): "Interest Rate Volatility and the Term Structure: A Two-Factor General Equilibrium Model," *Journal of Finance*, 47, September, 1259–1282.

Richard, S. F. (1978): "An Arbitrage Model of the Term Structure of Interest Rates," *Journal of Financial Economics*, 6, March, 33–57.

Tenney, M. S. (1995): "The Double Mean Reverting Process™," Chalke, Inc., Chantilly, VA.

APPENDIX

We want to show that the models of Chan (1992), Tenney (1995), and Canabarro (1996) can be put into the form of equation (13). With some modification of notation, but not of content, all three models can be written as

$$dy = P(k - y)dt + Sdw \tag{A1}$$

where $y_1(t) = \ln(r(t))$, y_2 is a second (unobservable) state variable, P is a 2-by-2 matrix of constants, k is a 2-vector, and the instantaneous covariance matrix of the state variables is $SS' > 0$; in Canabarro, the covariance matrix is

diagonal, implying that the state variables are uncorrelated. Let $m_2 > m_1 > 0$ be the eigenvalues of P, and let E be the corresponding 2-by-2 matrix of eigenvectors. Denote the diagonal matrix with elements m_1 and m_2 by M so that

$$PE = EM \tag{A2}$$

Defining $x = E^{-1}y$, we see that x satisfies equation (13) with $h = (EM)^{-1}k$ and $B = E^{-1}S$. In fact, we have shown a more general result: Any two-factor, Markovian, lognormal model can be reduced to equation (13).

10

Convexity and Empirical Option Costs of Mortgage Securities

Douglas T. Breeden

The option to prepay is the main focus of researchers in mortgage securities, as it is extremely important for both hedging and valuation. The first-order effect of prepayments is to shorten the lives of mortgages, thereby reducing their effective durations and price elasticities with respect to interest rates.

While researchers at major brokerage houses differ significantly in forecasting the effective durations of various fixed-rate mortgage coupons, on average their forecasts for standard fixed-rate mortgages are closely related to subsequent empirical durations (see Breeden 1994). Their forecasts of effective durations for interest-only strips vary hugely, by contrast, and are highly inaccurate forecasts of subsequent price elasticities. The risks of interest-only strips are extremely difficult to forecast, despite the major talent and resources that investment banks have devoted to research on mortgage securities. Errors in these risk estimates may have contributed to the many well-publicized losses in derivative mortgage securities in 1992–1995.

This chapter examines the second-order effect of prepayments on mortgage risk (i.e., the cost of the "negative convexity" of mortgages) (see Diller 1984). This is the asymmetry in mortgage returns; losses are larger for rate increases than the gains are for corresponding decreases in rates, due to the borrowers' use of the prepayment option adverse to the mortgage investor's interests. For hedgers of interest rate risk (which includes most of the investment banks and major players in the market), the first-order duration risk is routinely hedged, which makes this second-order convexity risk really of the first order with regard to movements in hedged profits.

Mortgage derivatives such as interest-only strips (IOs) and principal-only strips (POs) may have positive or negative convexity and positive or negative skewness of returns, depending upon the level of interest rates. Thus, the option risk adjustment may reverse sign and reflect an option benefit, making the certainty equivalent yield greater than the yield of the base case. The option risk adjustment may also be very large and, indeed, overwhelm the duration risk adjustment.

For example, in 1993, some interest-only strips had estimated option benefits of positive convexity that were estimated to be worth as much as 1,000 basis points, which was a very large portion of their projected risk-adjusted spreads to Treasuries. Also in 1993, other IO strips had option costs of *negative* convexity of 1,000 basis points, as estimated by one broker.

The expected patterns of option costs and option-adjusted durations for IO and PO strips are considerably more complex and interesting than for whole mortgage-backed securities (MBS), as is shown by work of Roll (1986) and Asay and Sears (1988). For this reason, we focus on brokers' forecasts of option costs for IOs, as they illustrate the differences most vividly. We also empirically estimate the option costs using a dynamic hedging strategy and compare these estimates to brokers' forecasts. Following the analysis of IO convexity option costs is a similar analysis for conventional (whole) fixed-rate mortgage-backed securities.

Section 1 gives an overview of the cross-sectional diversity and time series behavior of brokers' forecasts of option costs, option-adjusted durations, and option-adjusted spreads for mortgage securities. Section 2 reviews the theory of pricing and expected patterns of risk for interest- and principal-only strips, as well as for whole mortgages. Section 3 presents the empirical data for IO price behavior and empirical durations and compares the data to the predictions of theory, as well as to the brokers' forecasts.

Section 4 presents estimates for the scale and pattern of empirical option costs for IOs. The estimates are from a dynamic hedging strategy based upon brokers' forecasts of durations. These empirical option costs are then compared to the brokers' forecasts of option costs.

Section 5 briefly shows similar results for standard (whole, principal, and interest) FNMA mortgage-backed securities. Section 6 concludes with a few remarks on the results and future work.

1. MORTGAGE ENVIRONMENT AND BROKERS' FORECASTS

Figure 10.1 shows the roller coaster ride in FNMA 9% IO prices for 1991–1996. Prices dropped by over 50% from 1991 to 1993, as rates dropped sharply and prepayments surged. IO prices then doubled to near their original levels in 1994, as rates increased by 260 basis points. Prices fell again in 1995 by over 30%, and then rose in 1996 by 25% (through July), mirroring moves in interest rates.

Figure 10.2 shows corresponding movements in prices for principal-only strips, which also moved dramatically but opposite to IO strip prices, as expected.

Prices of both IO and PO strips are closely related to mortgage prepayments, which are closely related to interest rates. Figure 10.3 shows the movements in the prepayment rate on the conventional mortgage coupon (with over $1 billion outstanding) with the highest prepayment rate, which is usually a coupon rate 100 to 300 basis points over the current par mortgage rate.

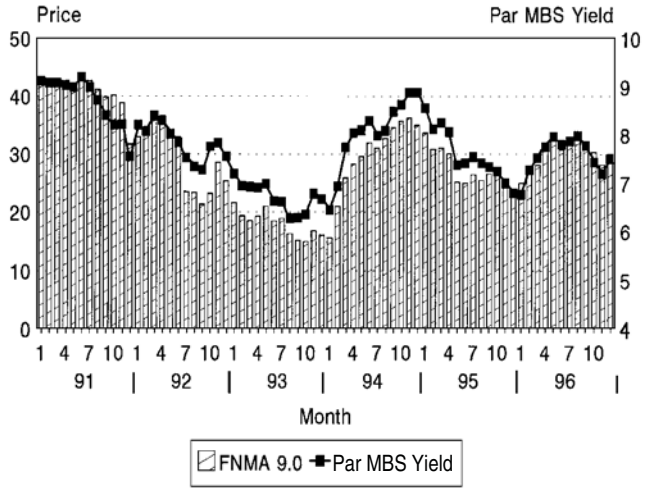


Figure 10.1. FNMA 9% Interest Only Prices: 1991–1996

In this chart, four unprecedented surges in prepayment rates are seen in 1992–1993, as they rise from 30% in 1991 to a peak of 70% annualized paydowns in late 1993. Then follows a huge plunge in prepayments in 1994 to a 15% maximum, responding to a 260 basis point increase in mortgage rates.

In 1995 and 1996, prepayment rates mirrored mortgage rate movements, increasing sharply in 1995 and early 1996 to 40% annualized prepayments before falling back to a 25% pace in mid-1996.

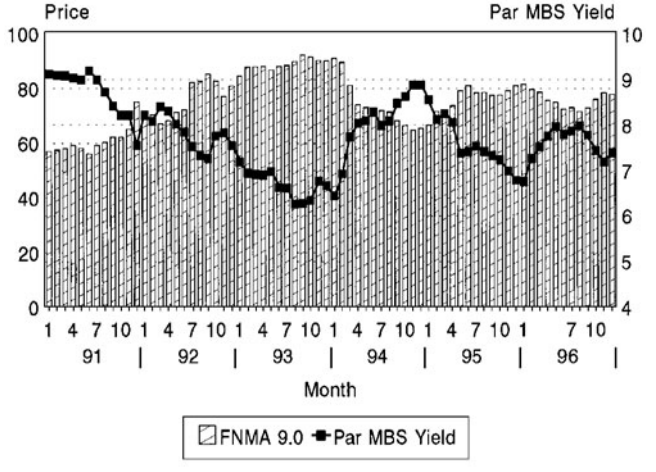


Figure 10.2. FNMA 9% Principal Only Prices: 1991–1996

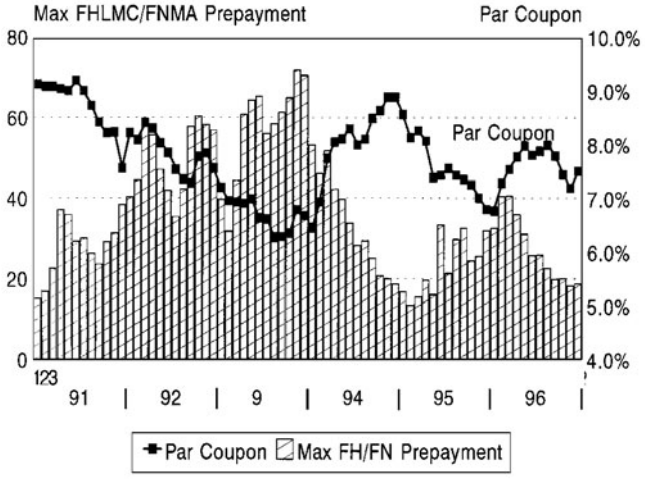


Figure 10.3. Mortgage Prepayments: 1991–1996

Figure 10.4 shows some of the differences in the prepayment response curves of the 1992–1993 epoch, vis-à-vis the 1986–1987 and 1988–1991 epochs. Technological change, program rule changes, consumer learning, and a pronounced media effect have led to great nonstationarities in the mortgage prepayment function. This is what makes the valuation and forecasting of risk problems so difficult for mortgage researchers.

The workhorse model for researchers in mortgages has been estimation of "option-adjusted spreads" (OAS) to Treasury rates or to the London Interbank Offered Rate (LIBOR) and selection of securities for purchase that have wide OAS. Although in recent years there have been criticisms of OAS

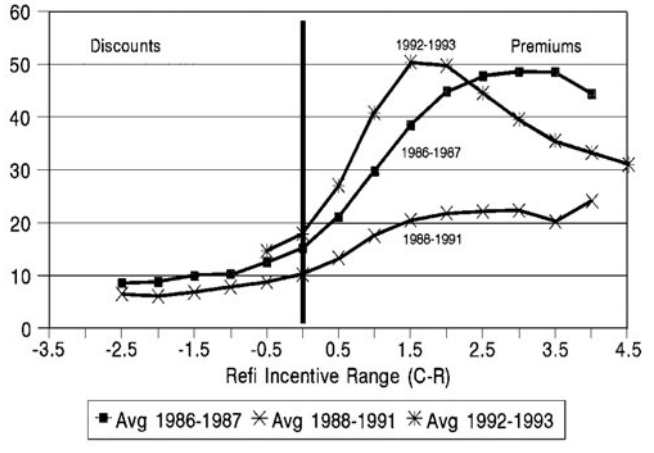


Figure 10.4. Empirical Data for FHLMC and FNMA. Prepayments Sorted by Refinancing Incentive (C-R)

models, as well as some improvements in focusing on projected total returns, OAS models continue to prevail.

A typical OAS analysis starts with a prepayment model to simulate the mortgage's cash flows under a wide variety of interest rate paths. From these cash flows, an internal rate of return is estimated, which is then risk-adjusted based upon the interest rate sensitivity (effective duration or price elasticity) and upon the prepayment option risk (negative or positive convexity).

Figure 10.5 shows the quarterly time series of the median broker's forecasts of option-adjusted spreads for IO strips with coupons of 7–10%. In late 1990, OAS were negative. Then, as prices of IOs fell sharply from 1991 to 1993, OAS surged to as much as 1,500 basis points (over Treasury rates).

Figure 10.6 shows that the brokers' forecasts of OAS were very useful in predicting the subsequent quarter's hedged excess returns on IOs. The relatively low OAS (at least in hindsight) that brokers forecasted for IOs in 1990 and 1991 were followed by very negative hedged returns in 1992 and 1993. In contrast, the very high OAS that brokers forecasted in 1992 and 1993 were followed by outstanding hedged performance of IOs in 1994 and 1995.

Unfortunately, while the OAS valuation results are very good on IOs for the brokers' researchers, the duration estimation and hedge precision results are not so comforting. (These results are opposite from those for whole fixed-rate mortgages, which have good duration estimation but poor OAS correlation with hedged returns, as Breeden (1994) shows.)

Table 10.1 shows the various investment bankers' published forecasts of option-adjusted durations at (or near) the ends of years from 1991 to 1995

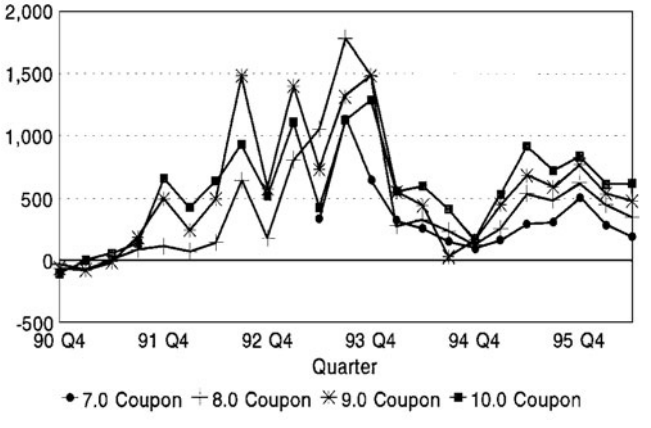


Figure 10.5. Median Broker Forecasts Interest Only Strips: Option Adjusted Spreads. End of Quarter: December 1990–June 1996

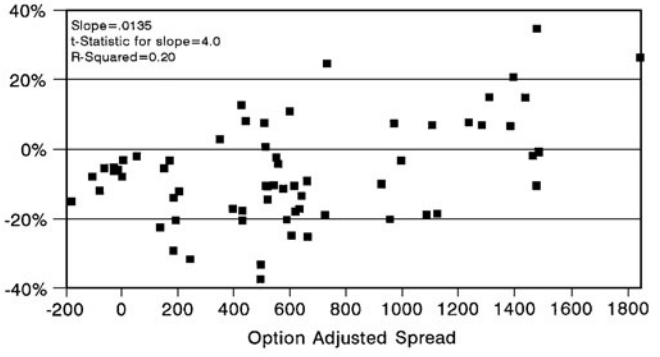


Figure 10.6. Median Broker Forecasts Interest Only Strips: Broker OASs vs. Hedged Returns

and at mid-year 1996. For portfolio managers, the spread in risk estimates is not comforting. In many cases, one broker's estimate of an IO's duration will be a multiple of another broker's forecast, and occasionally they are even of the opposite sign (1991, 1993).

We can see the difficulty that brokers had in forecasting IO durations by examining figure 10.7, which shows the option-adjusted durations forecasted by one broker for 8%, 9%, and 10.5% IOs. Major revisions to the model are apparent in both April 1993 and December 1993, as forecasts of durations of 8% IOs flipped from –15 years to 3 years and back to –12 years, without corresponding interest rate movements.

This illustrates what traders call "model whipsaw," as researchers almost everywhere frequently changed their models as the unprecedented prepayment waves came in. Also, it should be noted that some brokers have other estimates that they publish for the best recommended hedges, which may be based more on empirical durations than on option-adjusted durations from computer models.

Figure 10.8 shows that brokers' estimates of the durations of IOs were significantly smaller in absolute value than were the empirical durations measured. Brokers' duration estimates averaged about –15 years, while empirical durations averaged –25 to –30 years in the 1992–1994 period—a substantial difference. As Figure 10.9 shows, however, brokers' option-adjusted duration forecasts were useful predictors of realized durations (but statistically biased toward zero), as there is a significant correlation of the sizes of the forecasts with the sizes of realized durations.

In this chapter, the focus is on the "option cost" that is subtracted in risk-adjusting the mortgage's projected return for its negative convexity, which is due to the borrower's prepayment option. Even if an OAS approach is not used, all pricing models in mortgages must reflect these option features and, implicitly if not explicitly, adjust for the value of the negative convexity or negative skewness in normal mortgage returns. This chapter

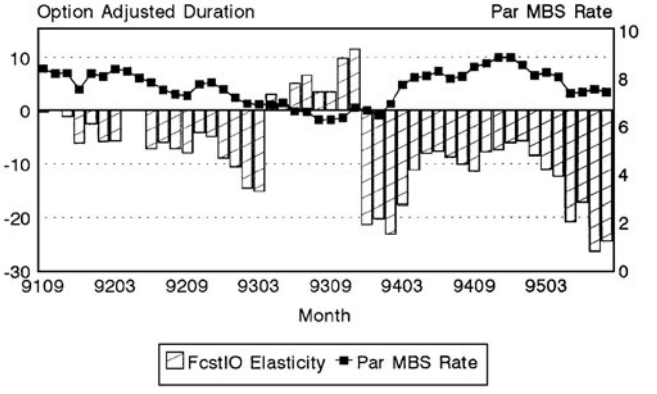
Table 10.1. Broker Forecasts for interest-only strip option-adjusted durations.

	December 31, 1991 (Par Yield = 7.55)					December 31, 1994 (Par Yield = 8.88)				
	#1 Goldman	#2 Prudential	#3 J.P. Morgan	#5 BS	#6 DLJ	#1 Goldman	#3 J.P. Morgan	#4 Salomon	#5 BS	#6 DLJ
6.5						1.3			(1.5)	1.9
7.0						(0.4)	(1.9)		(2.7)	0.8
7.5						(1.7)	(3.8)		(4.3)	(0.8)
8.0	(6.4)	(3.7)	(6.1)			(2.6)	(6.1)		(6.0)	(3.1)
8.5	(9.5)	(8.6)	8.9			(2.9)	(7.7)			(2.7)
9.0	(16.8)	(16.8)	12.1			(4.2)	(8.9)		(6.7)	(6.3)
9.5	(22.0)	(16.9)	12.8			(4.9)	(10.0)		(5.6)	(9.9)
10.0	(24.0)	(12.1)	12.1			(4.5)	(11.8)		(5.8)	(15.4)
10.5		(5.7)	(9.6)				(12.9)		(8.2)	

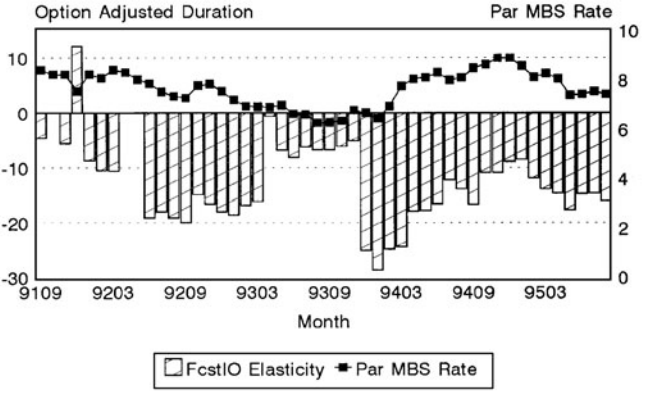
	December 31, 1992 (Par Yield = 7.55)					December 31, 1995 (Par Yield = 6.79)				
	#1 Goldman	#2 Prudential	#3 J.P. Morgan	#5 BS	#6 DLJ	#1 Goldman	#3 J.P. Morgan	#4 Salomon	#5 BS	#6 DLJ
6.5						(6.1)		(17.2)	(16.1)	(5.0)
7.0						(14.9)		(28.1)	(22.0)	(12.2)
7.5						(25.0)		(40.5)	(28.9)	(21.5)
8.0	(4.0)	(8.1)	(9.0)			(33.9)		(37.7)	(34.5)	(40.8)
8.5	(7.4)	(12.4)	(15.0)	(7.3)		(23.4)		(22.3)	(25.1)	(61.2)
9.0	(17.3)	(14.7)	(18.0)	(7.6)		(19.4)		(17.9)	(13.7)	(12.6)
9.5	(22.8)	(12.5)	(17.0)	(8.8)		(14.5)		(15.4)	(9.0)	(10.9)
10.0	(28.3)	(8.6)	(12.0)	(9.0)		(10.7)		(13.2)	(7.4)	(11.8)
10.5		(5.9)	(8.0)	(12.4)				(12.6)	(8.5)	

	December 31, 1993 (Par Yield = 6.67)					June 30, 1996 (Par Yield = 7.80)				
	#1 Goldman	#2 Prudential	#3 J.P. Morgan	#5 BS	#6 DLJ	#1 Goldman	#3 J.P. Morgan	#4 Salomon	#5 BS	#6 DLJ
6.5					(13.3)	1.3			(1.5)	1.9
7.0	(22.4)	(17.0)	8.5		(16.5)	(0.4)		(1.9)	(2.7)	0.8
7.5	(34.6)	(27.8)	9.8		(26.8)	(1.7)		(3.8)	(4.3)	(0.8)
8.0	(41.6)	(32.4)	11.5		(31.3)	(2.6)		(6.1)	(6.0)	(3.1)
8.5	(11.3)	(17.4)	3.8		(26.7)	(2.9)		(7.7)		(2.7)
9.0	(5.8)	(14.1)	(5.1)	(12.4)	(24.2)	(4.2)		(8.9)	(6.7)	(6.3)
9.5	(5.6)	(11.0)	(5.1)	(12.5)	(24.7)	(4.9)		(10.0)	(5.6)	(9.9)
10.0	(4.1)	(6.1)	(0.6)	(12.1)	(24.5)	(4.5)		(11.8)	(5.8)	(15.4)
10.5			4.7	(14.8)				(12.9)	(8.2)	

(a)



(b)



(c)

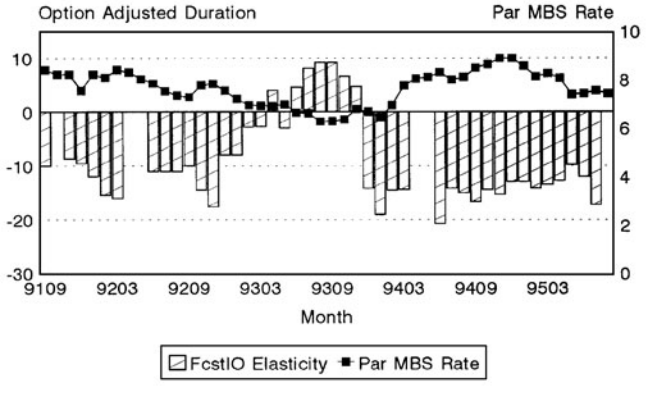


Figure 10.7. (a) J.P. Morgan, 8.0% Interest Only Option Adjusted Duration Monthly September 1991–August 1995. (b) J.P. Morgan, 9.0% Interest Only Option Adjusted Duration Monthly September 1991–August 1995. (c) J.P. Morgan, 10.5% Interest Only Option Adjusted Duration Monthly September 1991–August 1995.

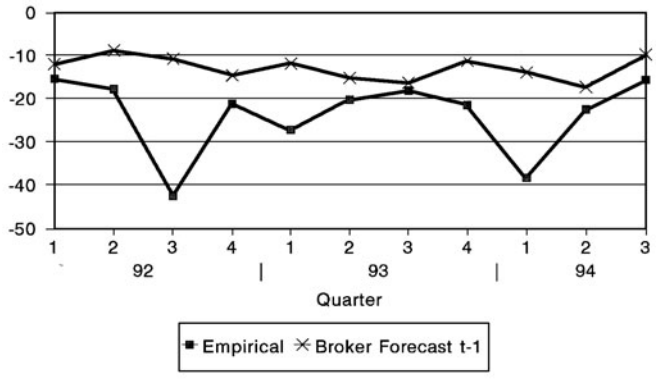


Figure 10.8. FNMA Interest-Only Strips: Average 8%–10% Empirical Durations vs. Median Broker Forecast

addresses whether or not the scale and pattern (both cross-sectional and time series) of option cost estimates make theoretical sense and are validated by the empirical data.

Empirical estimates of option costs ("whipsaw costs") for MBS using a dynamic hedging strategy are presented in Breeden (1991) for 1982–1990 using monthly data. The much more volatile dynamic option hedging costs for stripped securities are not examined. Also, the series of actual brokers' forecasts of durations is not used in those earlier estimates of dynamic option hedging costs as they are here.

Table 10.2 shows brokers' forecasts of option costs and OAS for IO strips at year-end 1991–1995 and mid-year 1996. Positive numbers indicate option costs (due to negative convexity), and negative numbers indicate option benefits (due to positive convexity).

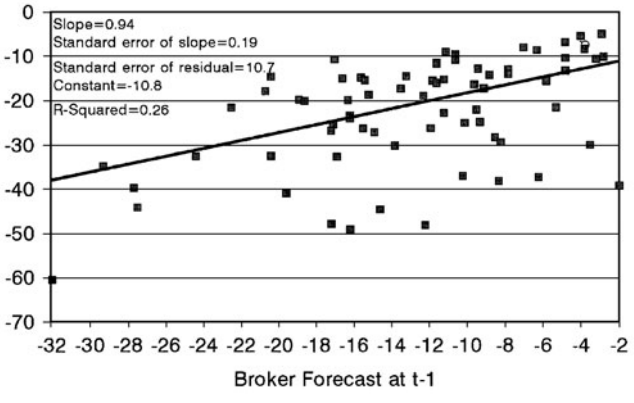


Figure 10.9. FNMA Interest-Only Strips. Empirical Durations vs Median Broker Forecasts

Table 10.2 Broker Forecasts of option costs, and OAS for interest-only strips.

	December 31, 1991 (MBS Par Yield = 7.55%) (FH/FN Max. Prep. = 38%)					
	Estimated Option Cost (Benefit)			Estimated OAS		
	#1 GS	#2 Pru	#3 JPM	#1 GS	#2 Pru	#3 JPM
8.0	(169)	19	766	110	116	35
8.5	(344)	(36)	1,104	469	191	365
9.0	(776)	(139)	301	568	(2)	496
9.5	(933)	(198)	128	691	43	634
10.0	(1,153)	(202)	(366)	661	49	904
10.5		(219)	(671)		226	1,004

	December 31, 1992 (MBS Par Yield = 7.55%) (FH/FN Max. Prep. = 57%)					
	Estimated Option Cost (Benefit)			Estimated OAS		
	#1 GS	#2 Pru	#3 JPM	#1 GS	#2 Pru	#3 JPM
8.0	120	192	840	623	179	(65)
8.5	163	47	551	968	370	500
9.0	280	(41)	(116)	982	295	717
9.5	335	(70)	(970)	931	372	693
10.0	366	(76)	(1,174)	684	454	426
10.5		(123)	(631)		988	128

	December 31, 1993 (MBS Par Yield = 6.67%) (FH/FN Max. Prep. = 70%)					
	Estimated Option Cost (Benefit)			Estimated OAS		
	#1 GS	#2 Pru	#3 JPM	#1 GS	#2 Pru	#3 JPM
8.0	322	93	965	529	1,555	1,400
8.5	(84)	41	390	1,093	2,487	1,233
9.0	(106)	35	(324)	1,482	2,343	760
9.5	(168)	12	(302)	1,420	1,747	769
10.0	(215)	25	(233)	1,286	1,931	808
10.5			210			1,042

continued

Table 10.2 continued

	December 31, 1994 (MBS Par Yield = 8.88%) (FH/FN Max. Prep. = 19%)			
	Estimated Option Cost (Benefit)		Estimated OAS	
	#1 Goldman	#3 J.P. Morgan	#1 Goldman	#3 J.P. Morgan
6.5	46		46	
7.0	70	334	70	(17)
7.5	100	430	100	(16)
8.0	141	560	141	(52)
8.5	127	603	127	27
9.0	182	589	182	98
9.5	195	574	192	184
10.0	153	475	153	140
10.5		271		129

	December 31, 1995 (MBS Par Yield = 6.79% (FH/FN Max. Prep. = 32%)			
	Estimated Option Cost (Benefit)		Estimated OAS	
	#1 Goldman	#4 Salomon	#1 Goldman	#4 Salomon
6.5	178	242	423	297
7.0	384	444	611	415
7.5	616	628	668	481
8.0	682	254	680	583
8.5	401	68	833	683
9.0	244	52	752	504
9.5	171	59	848	661
10.0	121	82	1,088	724
10.5		104		922

	June 30, 1996 (MBS Par Yield = 7.80%) (FH/FN Max. Prep. = 26%)			
	Estimated Option Cost (Benmefit)		Estimated OAS	
	#1 Goldman	#4 Salomon	#1 Goldman	#4 Salomon
6.5	69	108	135	62
7.0	126	139	189	129
7.5	191	189	239	220
8.0	260	260	376	310
8.5	251	291	432	407
9.0	250	205	478	304
9.5	157	98	511	349
10.0	114	43	622	371
10.5		20		462

There are several interesting points to be seen in table 10.2. First, the general scale of the option cost is quite large in absolute value and is of the same order of magnitude as the OAS. Second, brokers differ significantly in their forecasts of IO option costs. For example, as of December 31, 1991, Goldman Sachs forecasted option costs of –169 and –1,153 for 8% and 10% IOs, respectively, while J.P. Morgan forecasted +766 and –366 basis points (bp) of cost for those same coupons. They don't even agree on whether these IOs have positive or negative convexities!

As we'll see in section 2, given the form expected for the option cost function, this is not quite as surprising and implausible as it seems, although it is unusual and causes legitimate concerns among portfolio managers.

Despite their differences, both Goldman Sachs and J.P. Morgan found the 10% IOs to have about 1,000 bp of better convexity than did the 8s. Prudential, however, projected +19 bp and –202 bp of option cost for the 8s and 10s, respectively, for an option cost advantage to the 10s of only 221 bp. Thus, there are wide differences in brokers' views on the general magnitudes and coupon structures of IOs' option costs.

Figure 10.10 shows four brokers' estimates of option costs for IO coupons of 7%, 8%, 9%, and 10%. Scanning these charts, we do see a generally positive correlation of brokers' estimates of option costs and benefits for IOs (8s generally positive, 10s negative in 1991–1993), but there are still very wide differences for the brokers cross-sectionally, as well as through time.

Prepayment and valuation model revisions have occurred at all mortgage research firms during this volatile period, dramatically affecting a research group's option cost estimates. Figure 10.11 shows the option cost estimates of Goldman Sachs, which is acknowledged by most researchers as one of the leaders in mortgage research, particularly on IOs and POs.

Without having been privy to model changes, we can see that the discontinuities in option cost estimates for both 8.5% and 9.5% coupons (and others not shown) clearly indicate a model revision implemented in August 1992. For the 9.5s, an option cost (benefit) estimate of over –1,000 basis points in July 1992 turns into an estimate of option costs of over +200 basis points the next month, with relatively little intervening movement in interest rates. These model revisions were common at many firms, as researchers dealt with nonstationarities, nonlinearities, and prepayment movements never seen before.

Our main point is to show that analysis of these mortgage derivatives is not easy and that there are many interesting questions to examine.

1. From finance theory, what do we expect to be the scale and pattern of durations and option costs for IOs across coupons?
2. How can we empirically estimate durations and option costs, and how do the results of these estimates conform to the theory?
3. Do the forecasts of the brokers' research groups conform to either the theory or the data?

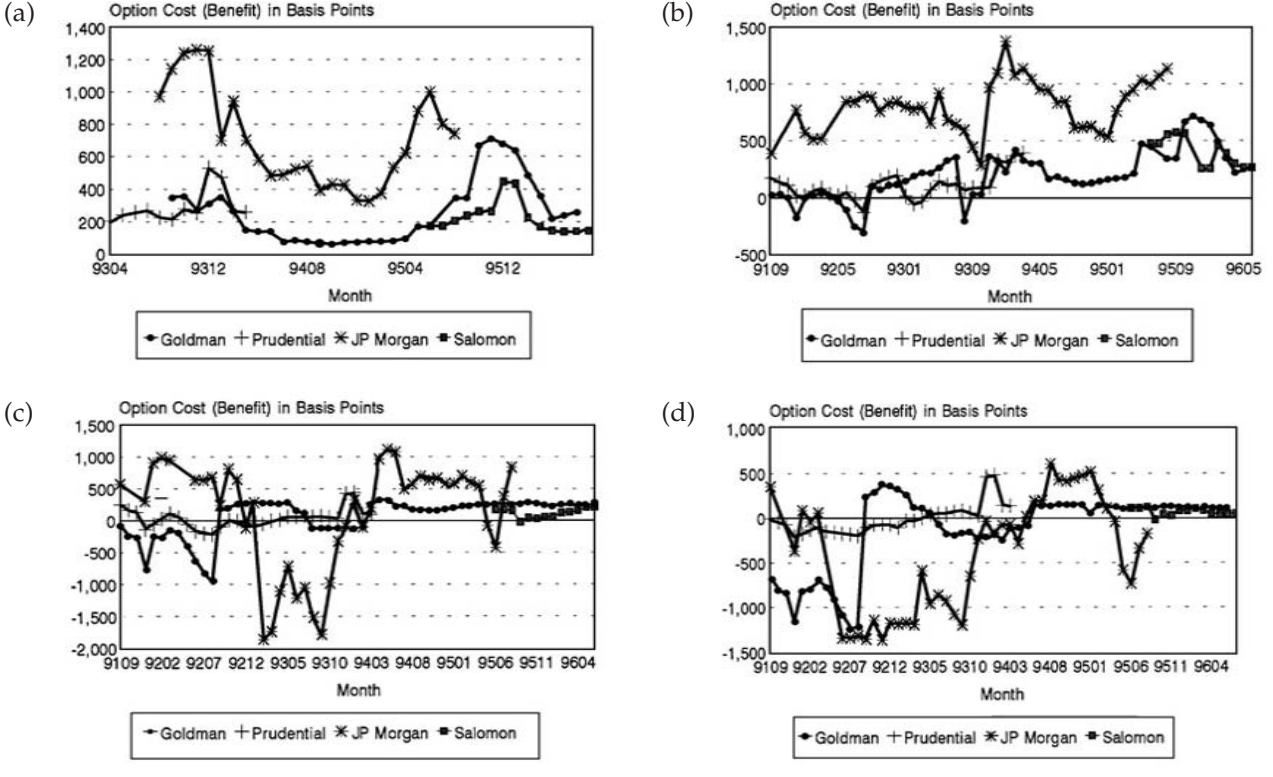
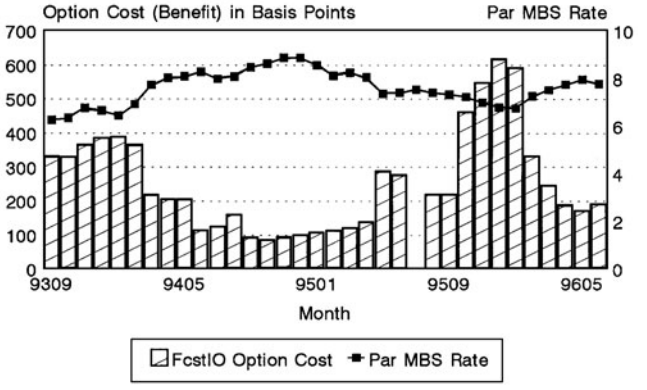
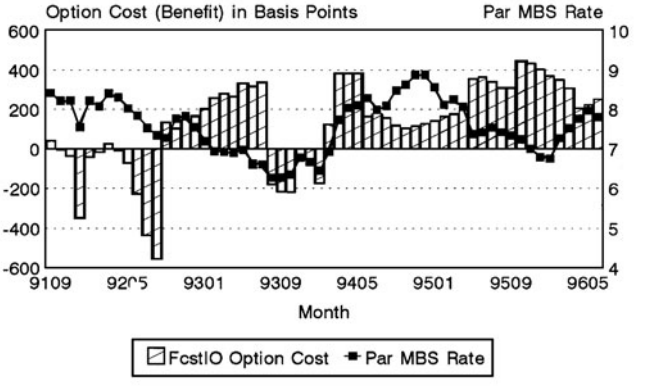


Figure 10.10. (a). 4 Brokers: 7.0% Interest Only Option Costs Monthly April 1991–June 1996. (b) 4 Brokers: 8.0% Interest Only Option Costs Monthly April 1991–June 1996. (c) 4 Brokers: 9.0% Interest Only Option Costs Monthly April 1991–June 1996. (d) 4 Brokers: 10.0% Interest Only Option Costs Monthly April 1991–June 1996

(a)



(b)



(c)

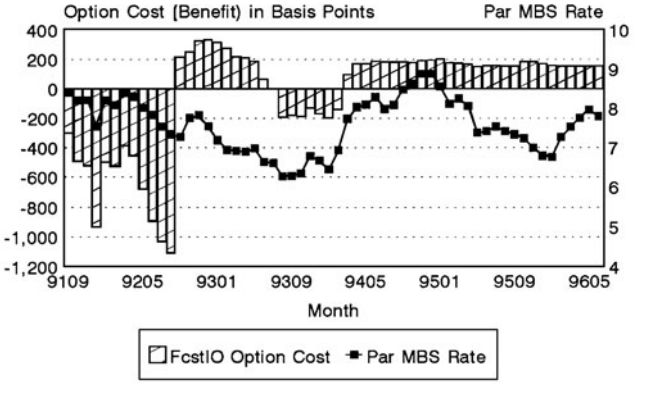


Figure 10.11. (a) Goldman Sachs, 7.5% Interest Only Option Costs Monthly September 1993–June 1996. (b) Goldman Sachs, 8.5% Interest Only Option Costs Monthly September 1993–June 1996. (c) Goldman Sachs, 9.5% Interest Only Option Costs Monthly September 1993–June 1996

2. REVIEW OF THEORETICAL PREDICTIONS OF RISKS IN INTEREST-ONLY STRIPS

The theory on durations and option risks for IOs precedes all the empirical data examined, so if it explains much of what we subsequently observed in this volatile period, it is a nice victory for the theorists and, indeed, shows the power of theory (much as Fischer Black believed and implemented).

Richard Roll produced the major path-breaking work on stripped mortgage-backed securities in 1986 while working at Goldman Sachs. His work was followed by fine work by his colleagues, Michael Asay and Timothy Sears (1988). As all the data examined in this chapter come from the 1988–1996 period, their theoretical work clearly precedes all the empirical results. What is described in subsequent sections is impressive support for their theory.

An illustration of a prepayment function and pricing for an FHLMC or FNMA 9% fixed-rate mortgage and its interest-only and principal-only strips is in table 10.3. For research on mortgage prepayment functions, see Richard and Roll (1989), Hayre (1994), and Patruno (1994). Valuations for the mortgage, the IO, and the PO are normally achieved with a Monte Carlo model, as is common in mortgage research, building on the approach of Black, Derman, and Toy (1990).

Figure 10.12 graphs the prices of these securities for par mortgage rates from 3% to 13% and indicates the option-like payoffs that these investments have at different interest rate levels. At very low interest rates, when prepayments are near their peak level on their S-curve, for example, an IO sells for a very low price but has a favorable asymmetric return pattern (positive convexity). At that low price the IO has little to lose (as prepays are near their peak and unlikely to increase much more with lower rates) but much to gain (as prepays will fall dramatically if rates increase).

Thus, at very low interest rates, the IO has a limited downside in price but a substantial potential upside if rates increase, very much like a put option on bond prices. Corresponding to this, at low rates, the IO has a substantial option benefit of positive convexity, rather than an option cost of negative convexity, as a normal MBS has.

At very high interest rates, the situation is reversed for the IO in that prepayment rates are then extremely low (near their minimum on their S-curve) and IO prices are quite high, but have limited upside for rate increases and a very substantial downside if rates drop. At high rates, the payoff pattern for the IO resembles having written a call option on bond prices in that, if rates increase and bond prices decline, the position has small gains, but if rates decline and bond prices increase, the position has large losses (as the call is in-the-money). Thus, at high rates, the IO has negative convexity and a substantial option cost.

It is important to note that the effective duration of an IO can actually change signs (to be positive, i.e., bond-like) both at very high and very low interest rates. At very low rates, as prepayments peak out quickly, addi-

Table 10.3. FHLMC/FNMA 9% illustration, interest-only and principal-only strip values.

Current-Coupon Mortgage Rate (%)	Coupon – Refi Rate (%)	Assumed Prepay Rate (%)	Principal-Only Value	Interest-Only Value	FHLMC 9% MBS
3.00	6	37.4	95.83	18.21	114.0
4.00	5	35.1	93.16	19.36	112.5
5.00	4	33.4	90.16	20.52	110.6
6.00	3	32.2	86.59	22.03	108.6
7.00	2	30.0	81.51	24.88	106.3
8.00	1	20.5	73.52	30.19	103.7
9.00	0	9.0	63.20	37.00	100.2
10.00	–1	6.3	54.42	41.35	95.77
11.00	–2	5.3	47.32	43.66	90.98
12.00	–3	4.8	41.68	44.50	86.18
13.00	–4	4.4	37.15	44.43	81.58

tional drops in rates might not accelerate prepayments but will benefit the IO by a lower discount rate for its cash flows (a standard positive duration effect). Similarly, at very high rates, when prepayment rates approach their minimum levels, additional rate increases will not benefit the IO much with lower prepays but will decrease the value of the IO as its cash flows are discounted at higher rates. Thus, while IOs usually have a large negative duration, their durations can become slightly negative and even positive at rates that are both very high and very low (as measured by the coupon minus refinancing rate (C – R) for the IO coupon).

Theoretical option-adjusted durations for interest-only and principal-only strips are illustrated in figure 10.13. Note that coupons that are 100 to 200 basis points above the current refinancing rate should have the great-

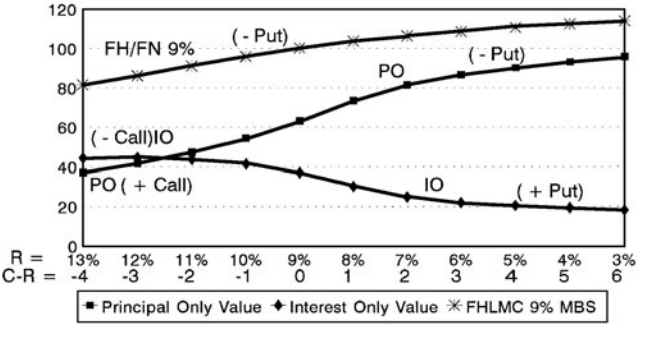


Figure 10.12. FHLMC/FNMA 9% Illustration. Prices of Interest Only and Principal Only Strips

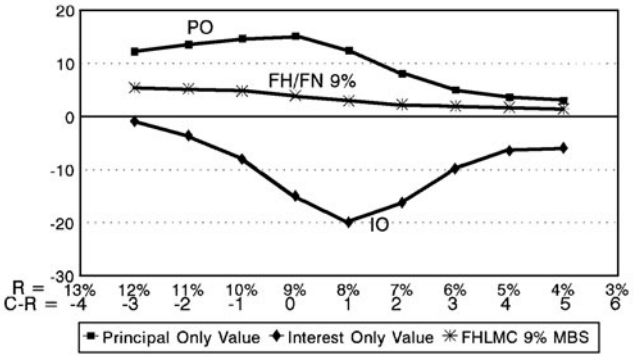


Figure 10.13. FHLMC/FNMA 9% Illustration. Option Adjusted Durations for Interest Only and Principal Only Strips

est negative duration, as they are on the cusp of the prepayment curve and have values most sensitive to interest rates. Both very high premiums (e.g., C – R = 4%) and discounts (C – R = –1% or –2%) have much lower durations (in absolute value), as their prepayments are at relatively flat segments of the prepayment curve—either near maximum prepayment levels or near minimum prepayment levels. Note that the IO duration curve is approximately V-shaped (we examine brokers' forecasts and empirical estimates for them in section 3).

Theoretical option costs for IOs and POs are illustrated in figure 10.14. Note that IOs on discount mortgages have negative convexity and a projected positive option cost. In contrast, IOs on premium mortgages have positive convexity and therefore projected negative option costs, or option benefits. The crossover point from positive to negative option cost here is at a premium of 100 basis points, but bear in mind that this is related more to the prepayment function than it is to interest rates (and figure 10.4 showed the substantial shifts in the prepayment function over time).

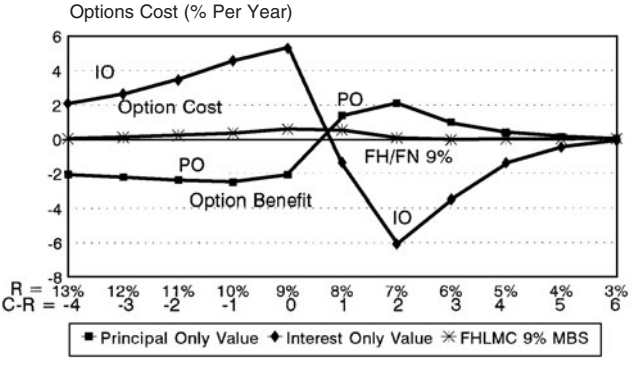


Figure 10.14. FHLMC/FNMA 9% MBS Illustration. Whipsaw Option Costs for Interest Only and Principal Only Strips

Thus, without the benefit of seeing the future, it is entirely reasonable that some researchers projected positive option costs in 1991–1993, while others projected negative option costs for coupons that are slight premiums and near the (changing) cusp of the prepayment function. Note that the shape of the projected option cost function is a bit like a "sine wave" located to cross from positive to negative. (We examine brokers' forecasts and empirical estimates of IOs' option costs in section 4.)

3. EMPIRICAL IO PRICE AND DURATION FUNCTIONS

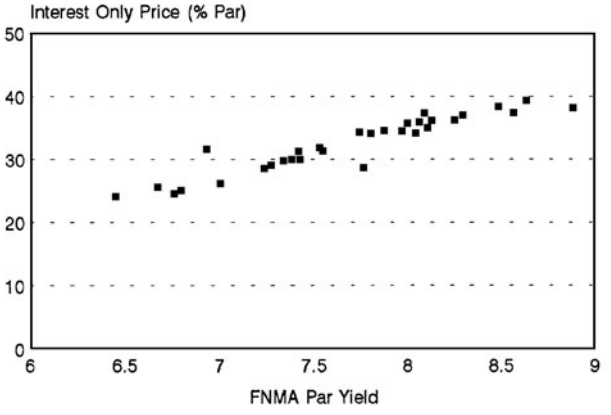
Prices of interest-only strips are expected to form an S-curve, with the steep slope located between par and refinancing rates where the coupon is 200 to 300 basis points over the refi rate, as that's where prepayments are most elastic with respect to interest rates. Figure 10.15 shows the IO strip price curves observed using monthly data for 1987–1996 collected by Smith Breeden Associates from a variety of sources. While the lower coupons trace out only a segment of the price curve, due to their more recent existence, the 9s to 10.5s have data for the entire period, which traces out a greater range in the IO price function. Both the positive and negative convexities anticipated by the theory (Figure 10.12) are demonstrated in these observed price curves for the higher coupons.

Brokers' forecasts of IOs' option-adjusted durations were collected, and table 10.4 presents the medians of brokers' forecasts of those durations quarterly for 1991–1996. Empirical durations for IOs (or "price elasticities") are estimated by regressing 5-day IO returns on 5-day changes in the 10-year Treasury note rate, and they are in table 10.5. Figures 10.16 and 10.17 plot those values sorted by their "relative coupon" (i.e., C – R = Coupon – Refinancing Rate). The V-shaped pattern predicted by the theory is present in the scatter plot.

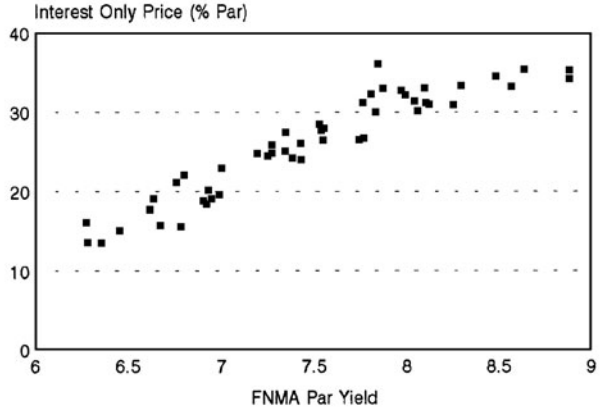
The V-shaped pattern for IO durations is much easier to see in figures 10.18 and 10.19, which graph the mean IO durations for coupons that fall into different "C – R buckets." For example, the 0.5 bucket contains the average duration estimate or empirical duration for coupons with C – R between 0.25 and 0.75. Note that the largest predicted durations are for coupons with C – R = 1–2%, which conforms to the theory's illustration in figure 10.13. Similarly, the empirical durations are also highest for coupons in that cusp range. The major inconsistency is that the brokers' maximum median forecast of duration averages only about –16 years, while the maximum empirical duration averages –28 years.

It is intuitive from the V-pattern of option-adjusted durations in figures 10.18 and 10.19 to see how option costs and benefits are generated by dynamic hedging strategies for IOs. As IOs usually have negative durations, proper hedges will go long bond futures. For IOs on discount and near-par securities, we are traveling along the left side of the pattern, meaning that as rates decrease, C – R increases, and the IO duration increases in absolute value.

(a)



(b)



(c)

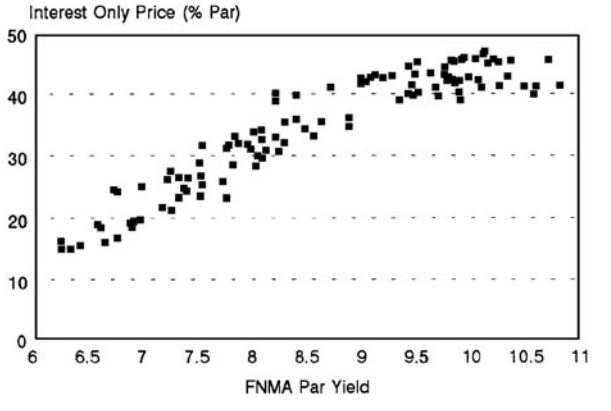
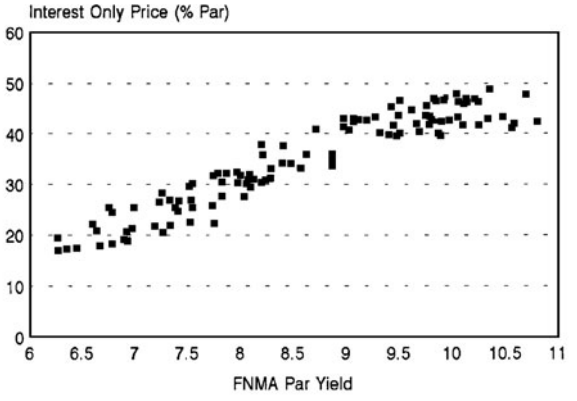
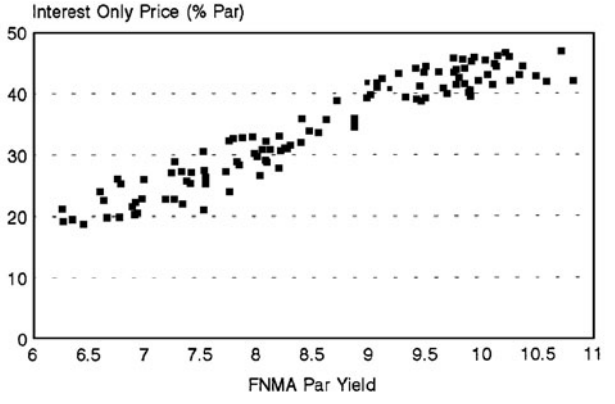


Figure 10.15. (a) FNMA 7.5 Interest Only Prices vs. Rates: Monthly Data January 1994–July 1996. (b) FNMA 8.5 Interest Only Prices vs. Rates: Monthly Data June 1992–July 1996. (c) FNMA 9.0 Interest Only Prices vs. Rates: Monthly Data July 1987–July 1996. (d) FNMA 9.5 Interest Only Prices vs. Rates: Monthly Data May 1987–June 1996. (e) FNMA 10.0 Interest Only Prices vs. Rates: Monthly Data July 1987 1994–July 1996. (f) FNMA 10.5 Interest Only Prices vs. Rates: Monthly Data June 1990–July 1996.

(d)



(e)



(f)

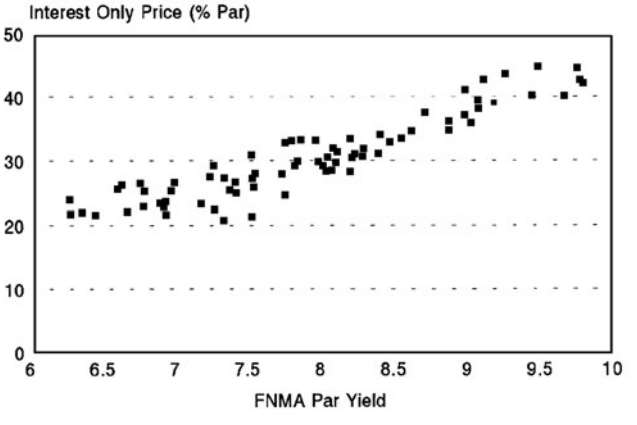


Figure 10.15. ***continued***

Table 10.4. Interest-only strips: median broker forecasts of option-adjusted durations.

End of Quarter	Par Yield FNMA	FN IO 6.5 249	FN IO 7.0 215	FN IO 7.5 218	FN IO 8.0 203/54	FN IO 8.5 7/24	FN IO 9.0 1/6	FN IO 9.5 4	FN IO 10.0 2	FN IO 10.5 50
4Q90	9.27				–1.3	–2.7	–5.7	–8.1	–11.7	–15.9
1Q91	9.08				–1.0	–2.5	–5.7	–8.2	–11.7	–14.1
2Q91	9.19				1.4	0.0	–2.9	–4.9	–7.3	–9.9
3Q91	8.41				–1.2	–3.0	–6.7	–9.6	–11.7	–10.4
4Q91	7.55				–6.1	–8.6	–16.8	–16.9	–12.1	–7.7
1Q92	8.40				–1.8	–3.8	–9.2	–13.3	–16.4	–12.0
2Q92	7.84				–3.3	–6.0	–12.0	–16.9	–16.0	–7.1
3Q92	7.27				–10.4	–16.0	–17.0	–15.3	–15.0	–7.6
4Q92	7.55				–8.1	–9.9	–16.0	–14.7	–11.0	–8.0
1Q93	6.92				–19.4	–18.4	–16.1	–11.4	–10.9	–2.7
2Q93	6.63		–12.2	–20.1	–24.2	–20.5	–15.4	–11.4	–10.4	–4.6
3Q93	6.28		–20.2	–27.5	–29.1	–9.1	–9.3	–5.6	–3.0	–3.6
4Q93	6.67		–16.7	–27.3	–31.8	–14.4	–10.0	–8.3	–5.1	–4.6
1Q94	7.74	–4.3	–6.8	–11.4	–20.2	–22.3	–18.7	–13.6	–11.7	–13.0
2Q94	8.29	–2.1	–2.6	–4.6	–7.6	–9.4	–11.6	–8.9	–11.0	–15.2
3Q94	8.48	2.2	0.2	–2.5	–4.1	–2.8	–5.0	–7.9	–10.5	–12.8
4Q94	8.88	1.6	–0.4	–1.7	–3.1	–2.9	–6.3	–9.9	–11.8	–12.9
1Q95	8.25	1.1	–1.1	–2.8	–6.6	–8.8	–12.1	–14.1	–14.5	–13.5
2Q95	7.42	–1.2	–5.3	–10.6	–17.3	–26.4	–14.8	–12.5	–11.8	–10.2
3Q95	7.34	–0.9	–3.9	–8.6	–17.2	–19.7	–14.8	–12.2	–11.6	–10.7
4Q95	6.79	–6.1	–14.9	–25.0	–37.7	–23.4	–17.9	–14.5	–11.8	–11.2
1Q96	7.53	–1.4	–4.5	–8.0	–12.5	–16.1	–13.3	–12.2	–11.1	–10.3
2Q96	7.80	–1.0	–3.3	–6.7	–11.6	–14.1	–8.0	–6.7	–7.2	–6.5

This means the long hedge will have to get longer by purchasing more bond futures when rates are lower and bond prices higher. Similarly, the hedge will sell futures as rates increase (i.e., when bond prices are low). This dynamic hedging strategy generates "whipsaw" losses due to negative convexity if rates move away from and then back to their starting point. This corresponds to the prediction that discount and current coupon IOs will have positive option costs.

For high-premium IOs (with $C - R > 1.0$), the dynamic hedging strategy is shifting according to the right side of the V in figures 10.18 and 10.19. For these IOs, as rates decrease and $C - R$ increases, the proper hedge becomes smaller. To accomplish this, bond futures are sold when rates decrease and bond prices increase. This dynamic hedging strategy generates

Table 10.5. Interest-only strips: empirical data for option-adjusted durations.

End of Quarter	Par Yield FNMA	FN IO 6.5 249	FN IO 7.0 215	FN IO 7.5 218	FN IO 8.0 203/54	FN IO 8.5 7/24	FN IO 9.0 1/6	FN IO 9.5 4	FN IO 10.0 2	FN IO 10.5	FN IO 11.5 50
1Q92	8.40				–8.6	–14.2	–10.7	–25.3	–18.9		
2Q92	7.84				–39.2	–5.5	–12.8	–17.3	–15.0		
3Q92	7.27				–30.0	–37.4	–48.2	–47.9	–49.1		
4Q92	7.55				–10.9	–24.0	–26.8	–26.3	–18.7	–13.0	
1Q93	6.92				–38.2	–25.0	–23.3	–27.1	–22.8	–29.4	
2Q93	6.63				–41.0	–20.1	–19.9	–11.7	–9.0	–5.0	
3Q93	6.28				–32.6	–17.9	–14.8	–16.0	–9.6	–6.8	
4Q93	6.67		–32.5	–39.7	–34.8	–24.8	–22.0	–15.6	–10.6	–8.4	
1Q94	7.74		–32.7	–44.2	–60.5	–44.6	–37.1	–28.3	–21.5	–13.2	
2Q94	8.29		–8.0	–11.4	–14.6	–21.6	–19.8	–30.2	–26.2	–14.5	
3Q94	8.48	–7.0	–10.1	–10.4	–14.0	–16.3	–15.5	–17.2	–15.2	–15.4	
4Q94	8.88		–12.2	–11.6	–11.3	–13.8	–13.5	–14.9	–15.3	–17.9	–16.8
1Q95	8.25		–5.5	–6.5	–6.1	–17.4	–16.8	–17.1	–16.0	–16.7	–15.8
2Q95	7.42		–8.4	–9.8	–11.7	–22.8	–21.0	–19.4	–19.5	–19.5	–17.7
3Q95	7.34		–15.1	–19.3	–29.6	–31.5	–25.3	–22.5	–21.0	–17.1	–15.5
4Q95	6.79		–21.0	–26.8	–26.0	–29.9	–27.7	–29.2	–25.9	–24.2	–22.2
1Q96	7.53		–19.8	–23.9	–24.0	–30.5	–20.1	–18.1	–17.6	–17.7	–16.3
2Q96	7.80		–21.2	–20.3	–20.2	–21.1	–24.0	–23.0	–22.9	–24.3	–23.1

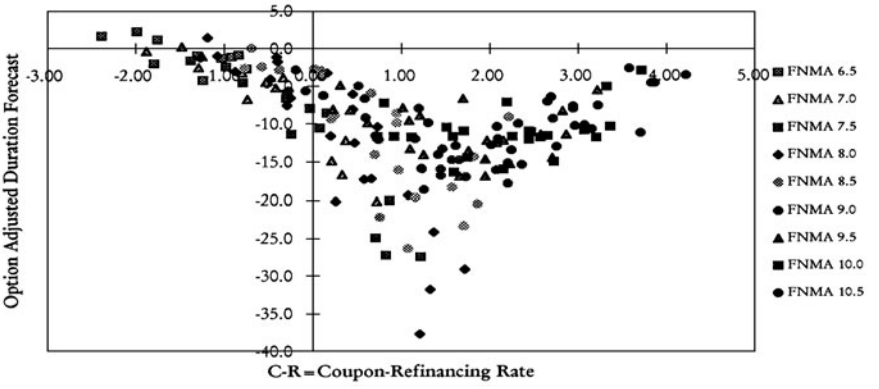


Figure 10.16. FNMA Interest Only Strips: Quarterly by Coupon, 1991–1996. Median Broker Option-Adjusted duration Forecasts vs. C-R

"whipsaw gains" due to the positive convexity of the high-premium IOs. Thus, these IOs have an option benefit, as is discussed further in the next section.

4. BROKER FORECASTS AND EMPIRICAL OPTION COSTS FOR IOS

Medians of brokers' forecasts for the option costs of interest-only strips are in table 10.6, with a time series plot as figure 10.20. A scatter plot of these data sorted by C – R is in figure 10.21, and the means of the C – R buckets are in figure 10.22. (In figures 10.21 and 10.22, 3.00 represents 300 basis points.)

Note that the brokers' forecasts of option costs do have a rational pattern, according to the theory of section 2, in that they have the "sine wave"

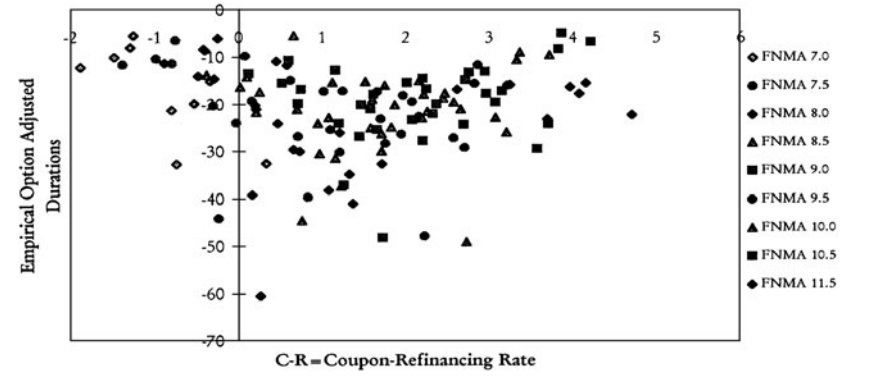


Figure 10.17. FNMA Interest Only Strips: Quarterly by Coupon 1992–1996. Empirical Option Adjusted Durations vs. C-R

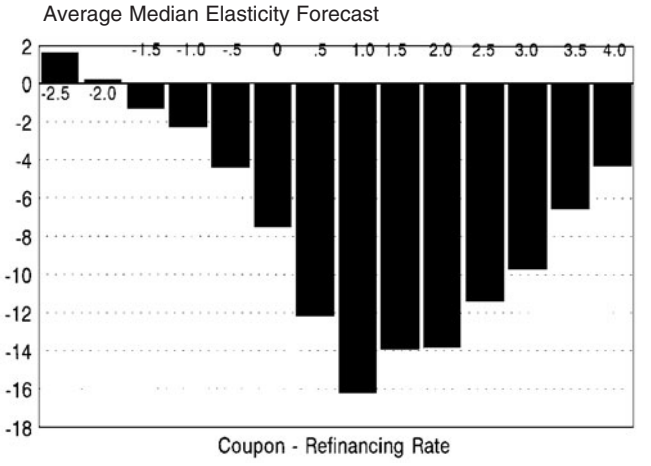


Figure 10.18. Median IO Duration Broker Forecasts vs. Coupon-Refinancing Rate: Composite of FNMA 6.5–10.5 Coupons

shape. Discount and small-premium IOs are predicted to have significant option costs, while higher-premium (C – R > 2) IOs are predicted to have option benefits (positive convexity). This squares with the theory, although the brokers' crossover point appears to be at a slightly higher C – R than the illustration of section 2.

Empirical option creation costs are estimated as in Breeden (1991). For 1991–1996, the median broker forecasts of durations are used quarterly to hedge IO returns with 10-year Treasury note futures, with durations chang-

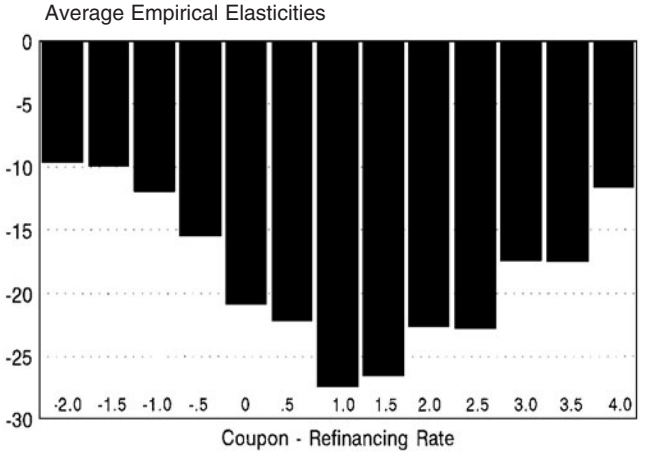


Figure 10.19. Empirical Durations vs. Coupon-Refinancing Rate for IOs: Average of FNMA 6.5–10.5 Coupons, Quarterly Estimates, 1991–1996

Table 10.6. Median broker forecasts: option costs for interest-only strips.

End of Quarter	ParYld FNMA	# of Brokers	6.5	7.0	7.5	8.0	8.5	9.0	9.5	10.0	10.5
1Q91	9.08	1				120	144	203	208	217	208
2Q91	9.19	2				106	154	227	267	285	300
3Q91	8.41	3				171	157	246	170	–8	–120
4Q91	7.55	3				19	–36	–139	–198	–366	–671
1Q92	8.40	3				83	84	114	–13	–100	–706
2Q92	7.84	3				49	–80	–169	–777	–1,086	–919
3Q92	7.27	3				101	132	195	218	–104	–542
4Q92	7.55	3				192	163	116	–70	–76	–375
1Q93	6.92	3				217	276	–4	–33	–31	–296
2Q93	6.63	3				322	62	64	40	–83	
3Q93	6.28	3		349	330	86	–218	–117	–179	–169	
4Q93	6.67	3		529	477	322	405	–93	–105	–38	
1Q94	7.74	3		260	433	389	381	198	97	–72	
2Q94	8.29	2		281	388	552	645	652	489	165	
3Q94	8.48	2		230	291	370	401	436	406	287	
4Q94	8.88	2		202	265	350	365	385	385	314	
1Q95	8.25	2		305	410	535	372	415	294	120	
2Q95	7.42	3		172	276	473	305	174	126	110	
3Q95	7.34	2	8	187	309	456	270	115	71	45	
4Q95	6.79	2	210	414	622	468	234	148	115	102	
1Q96	7.53	2	96	163	257	371	277	195	123	94	
2Q96	7.80	2	89	132	190	260	271	227	127	78	

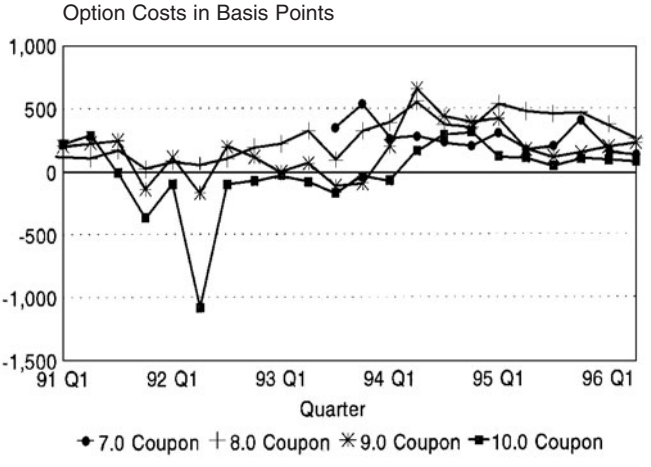


Figure 10.20. Median Broker Forecasts Interest Only Strips: Options Costs. End of Quarter: March 1991–June 1996

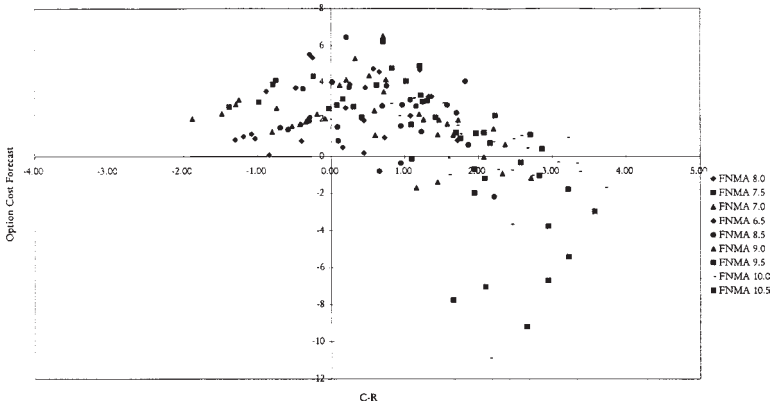


Figure 10.21. Interest Only Strips: Quarterly by Coupon, 1991–1996. Median Broker Option Cost Forecasts vs. C-R

ing quarterly. For 1988–1990, Smith Breeden's IO duration forecasts are used to estimate the IO option cost through the same dynamic hedging strategy. The return from the dynamic hedging strategy is then compared to the return that would have been earned if one had hedged with the ex post average duration forecasted.

Thus, if this were a normal mortgage with negative convexity, the dynamic strategy would do worse due to the "whipsaw" that occurs with buying high and selling low to recreate the prepayment call option in the hedge. Of course, if the underlying instrument had positive convexity, the whipsaw option cost should be negative (i.e., an option benefit).

Table 10.7 shows the numerical empirical option costs for IOs, where 12.93 represents 1,293 basis points. Figure 10.23 shows the scatter plot of these empirical option costs sorted by C – R, and figure 10.24 shows the

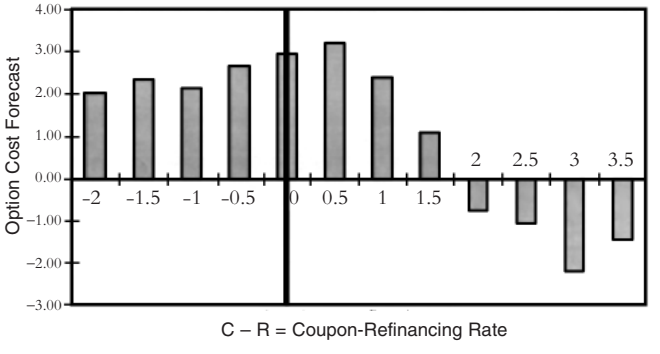


Figure 10.22. Interest Only Strips: Composite of FNMA 6.5–10.5, 1991–1996. Median Broker Option Cost Forecasts vs. C-R

Table 10.7. Interest-only empirical option costs from dynamic hedging; quarterly median broker duration adjustments.

Year End	Par MBS Yield 8.82	TNote Volatility	TNote Rate	FNMA 7.0 T257	FNMA 7.5 T254	FNMA 8.0 T203	FNMA 8.5 T7	FNMA 9.0 T6	FNMA 9.5 T4	FNMA 10.0 T2	FNMA 10.5 T50
1987	10.19	4.2	0.63								
1988	10.56	4.5	–0.02					1.75	5.77	4.40	
1989	9.51	3.8	–1.15					5.10	–1.31	–4.07	
1990	9.33	3.5	0.12					2.64	3.00	3.26	
1991	7.59	2.7	–1.79					1.00	1.06	2.19	4.18
1992	7.59	3.9	–0.63					7.27	1.07	–4.02	–1.49
1993	6.72	2.8	–1.67				–2.69	4.70	–5.68	–5.59	–2.07
1994	8.92	3.5	1.47		12.93	14.00	4.17	0.63	–0.16	–3.14	–5.22
1995	6.82	2.9	–2.67	2.35	4.33	6.09	10.17	2.04	–0.42	–1.16	–1.72
6/95–6/96	7.81	3.2	0.26	8.17	11.93	14.64	0.99	2.08	1.68	0.03	0.40
Average	8.53	3.50	–0.55	5.26	9.73	11.58	3.16	1.98	0.56	–0.90	–0.99

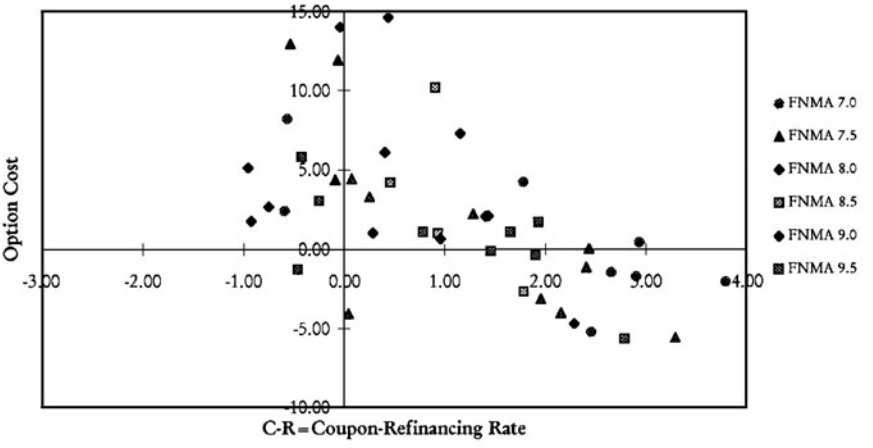


Figure 10.23. FNMA Interest Only Strips: Yearly by Coupon, 1988–1996. Empirical Option Costs vs. C-R

bucketed means. Note that the shape of the relationship is that predicted by the theory, and the crossover point from negative to positive convexity is similar to the brokers' forecasts. Thus, the theory, the brokers' forecasts, and the empirical data all have the same shape for the option costs.

The scale of the option cost is a remaining point of dispute, however. Figures 10.25 and 10.26 show the bucketed means for the option cost from Goldman Sachs and J.P. Morgan. Both have sensible shapes, in accord with the theory and the data. Note that the amplitudes of the option cost fluctuations differ considerably, as Goldman's are in a range of ±200 basis points, while J.P. Morgan's are in a range of ±800 basis points. From figure 10.22,

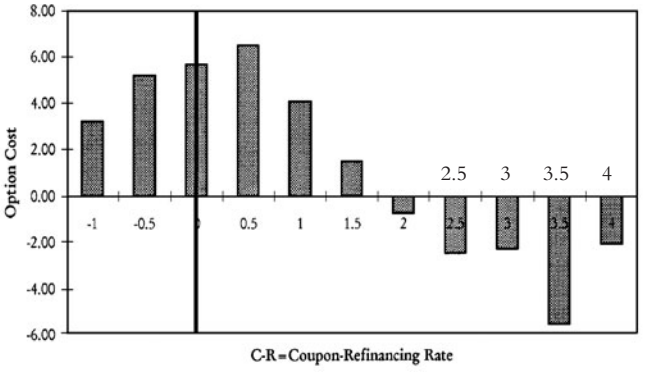


Figure 10.24. FNMA Interest Only Strips: Composite of FNMA 7.0–10.5, 1988–1996. Empirical Option Costs

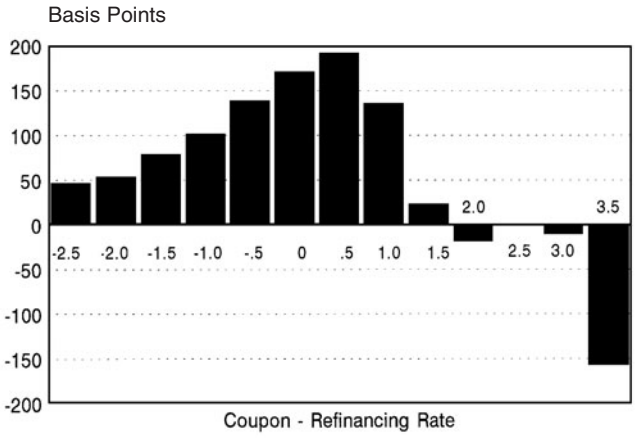


Figure 10.25. Goldman Sachs Option Cost Forecasts vs. C-R for Interest Only Strips: Composite of FNMA 6.5–100 Coupons, 1991–1996

the median broker's forecast is in a range from –200 to +300 basis points. What are the empirical estimates?

From figure 10.24, the empirical estimates of option costs range from approximately +600 basis points to –550 basis points, roughly between the Goldman and the J.P. Morgan estimates, but above the median broker estimates. From the results obtained and shown in table 10.7, forecasts of option costs of 1,000 basis points or more are not out of line in some years for IOs on lower coupons. Option benefits of 500 basis points or more have also occurred. Thus, IOs certainly display nontrivial positive and negative convexities.

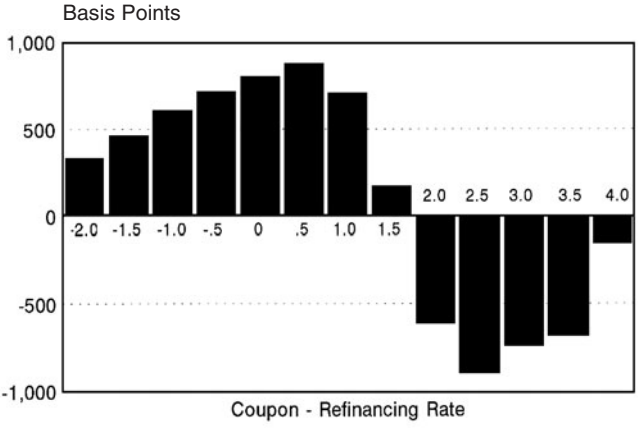


Figure 10.26. J.P. Morgan Option Cost Forecasts vs. C-R for Interest Only Strips: Composite of FNMA 6.5–10.0 Coupons, 1991–1996

Table 10.8. Median broker forecasts: FHLMC/FNMA option-adjusted spreads.

End of Quarter	Par Yld FNMA	# of Brokers	6.0	6.5	7.0	7.5	8.0	8.5	9.0	9.5	10.0	10.5	11.0	11.5	12.0	12.5	13.0
4Q87	10.11	2					112	106	103	102	93	89	85	78	70		
1Q88	9.76	4				99	95	84	75	71	49	36	28	11	32	11	
2Q88	9.83	5				78	74	72	63	54	44	38	32	20	9	9	–54
3Q88	9.91	3				82	86	81	70	60	46	41	29	32	29	68	
4Q88	10.48	4				95	84	88	84	82	74	92	60	79	54	62	
1Q89	10.81	5				97	97	103	102	91	91	64	59	76	71	76	10
2Q89	9.70	5				94	90	93	88	83	76	73	80	62	79	61	
3Q89	9.88	4				104	100	103	94	87	82	83	85	84	50	65	
4Q89	9.50	5				88	95	101	91	82	79	86	94	80	85	91	
1Q90	9.97	5				76	79	87	83	74	70	73	77	85	86	89	
2Q90	9.67	5				74	64	72	72	72	70	72	72	79	81	86	
3Q90	9.78	5				72	66	62	63	62	57	52	71	70	60	97	
4Q90	9.27	6				93	83	88	83	81	83	91	106	105	107	106	
1Q91	9.08	6				97	70	80	70	68	71	72	75	83	89	109	
2Q91	9.19	5				76	67	75	66	66	55	58	75	72	65		
3Q91	8.41	6			74	78	70	68	66	65	61	80	112	121			

4Q91	7.55	8					69	77	75	83	87	88	76	88	105	114	120
1Q92	8.40	7			76	67	64	61	61	69	73	70	46	85	106		
2Q92	7.84	7			79	66	65	61	58	65	72	70	71	104	131		
3Q92	7.27	6			86	85	92	90	82	97	96	96	99	122	179		
4Q92	7.55	7			65	64	68	69	69	70	69	53	60	67			
1Q93	6.92	7			67	77	93	100	80	80	80	80	105	79			
2Q93	6.63	7		58	63	71	85	83	83	85	85	56	77	43			
3Q93	6.28	7	58	72	98	115	126	135	96	102	108	87	121	125			
4Q93	6.67	7	66	68	70	73	68	77	79	71	69	68	106	124			
1Q94	7.74	7	69	69	65	59	57	52	60	59	44	41	52	115			
2Q94	8.29	6	55	54	54	54	56	55	45	42	42	3	23				
3Q94	8.48	6	51	48	52	52	54	52	58	49	42	23	50				
4Q94	8.88	6	48	47	49	47	49	50	45	44	38	45	48	79			
1Q95	8.25	8	50	45	48	45	47	54	55	59	60	45	56				
2Q95	7.42	6	42	44	57	61	47	68	59	41	91	49	36				
3Q95	7.34	7	48	50	57	61	61	57	59	48	75	58					
4Q95	6.79	7	55	59	65	54	50	63	63	99	85	75	175	201			
1Q96	7.53	7	51	49	51	66	65	59	48	12	14	40	123	151			
2Q96	7.80	7	58	63	68	70	75	84	85	64	83	36	34				

Table 10.9. Median broker forecasts: FHLMC/FNMA price elasticities (option-adjusted durations).

End of Quarter	Par Yld FNMA	# of Brokers	Slope	R^2	6.0	6.5	7.0	7.5	8.0	8.5	9.0	9.5	10.0	10.5	11.0	11.5	12.0
1Q87	8.69	2							5.6	5.4	5.2	5.4	3.5	1.8	1.4		
2Q87	9.85	2	0.86	0.86					5.0	5.8	5.4	5.2	4.2	2.9	2.0		
3Q87	10.70	2	0.73	0.83					5.4	5.8	5.6	5.6	5.0	4.4	3.4	2.7	2.0
4Q87	10.11	2	0.58	0.50					6.1	5.8	5.6	5.2	5.0	4.0	2.8	2.2	1.6
1Q88	9.76	3	0.84	0.98				5.2	5.4	5.1	4.9	4.5	3.9	3.4	2.7	1.9	1.9
2Q88	9.83	4	0.87	0.98				5.2	5.5	5.2	5.0	4.6	4.0	3.5	2.8	2.1	1.0
3Q88	9.91	2	0.84	0.98				4.8	5.2	5.0	4.6	4.4	3.8	3.0	2.4	1.0	1.0
4Q88	10.48	4	1.03	0.88				5.8	5.4	5.1	4.7	4.2	4.0	3.2	2.8	2.3	1.8
1Q89	10.81	5	1.15	0.74				5.3	5.4	5.0	4.6	4.3	4.0	3.4	2.9	2.4	2.2
2Q89	9.70	5	0.73	0.97				5.1	5.2	4.6	4.1	3.9	3.5	2.9	2.2	2.0	1.6
3Q89	9.88	4	0.63	0.95				5.0	4.8	4.6	4.3	4.0	3.4	2.9	2.2	1.8	1.7
4Q89	9.50	5	0.63	0.95				5.0	5.1	4.7	4.0	3.7	3.3	2.4	1.8	1.7	1.6
1Q90	9.97	5	0.89	0.94				5.2	5.1	4.8	4.7	4.4	3.8	3.1	2.3	2.0	2.0
2Q90	9.67	4	1.01	0.99				5.2	5.4	5.1	4.8	4.4	3.8	3.2	2.6	2.3	2.4
3Q90	9.78	5	1.12	0.95				5.6	5.5	5.3	5.1	4.8	4.3	3.5	2.7	2.3	2.3
4Q90	9.27	5	1.14	0.98				5.2	5.3	5.2	4.8	4.3	3.6	2.7	2.7	2.8	2.7
1Q91	9.08	6	1.31	0.93				5.3	5.3	5.0	4.8	4.3	3.5	2.6	2.2	2.4	2.2
2Q91	9.19	5	0.96	0.93				5.2	5.2	5.2	5.0	4.4	3.8	3.0	2.4	2.2	2.5

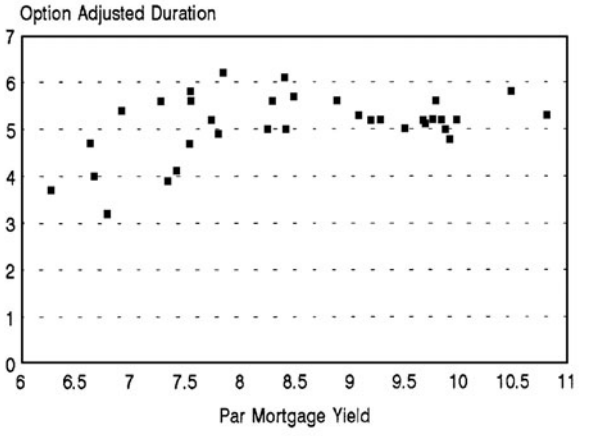
3Q91	8.41	6	0.95	0.96				5.0	5.1	5.0	4.5	3.6	2.6	2.5	2.5	2.6	2.6
4Q91	7.55	8	1.20	0.95				5.6	5.0	4.8	3.9	2.8	2.0	1.8	1.9	1.6	2.4
1Q92	8.40	7	1.10	0.90				6.1	5.5	5.0	4.2	3.5	2.5	2.1	2.1	2.1	2.6
2Q92	7.84	7	1.00	0.93			6.2	6.2	5.7	4.8	3.6	2.7	1.9	1.8	1.4	1.9	2.6
3Q92	7.27	6	0.81	0.97			6.1	5.6	5.0	3.6	2.6	2.2	2.0	1.5	1.6	1.4	
4Q92	7.55	8	1.00	0.93			6.0	5.8	5.0	3.9	3.0	2.2	2.0	1.8	1.5	1.0	
1Q93	6.92	8	0.92	0.99		6.8	6.0	5.4	4.4	2.8	2.0	1.6	1.8	1.6	1.6	1.8	
2Q93	6.63	8	0.67	0.95		6.4	5.8	4.7	3.4	2.4	1.6	1.4	1.5	1.5	1.6	1.8	
3Q93	6.28	7	0.76	0.92	6.8	5.9	5.0	3.7	2.8	2.0	1.9	1.8	1.7	1.3	1.1	1.8	
4Q93	6.67	8	0.92	0.97	6.8	6.2	5.3	4.0	2.8	2.4	2.0	1.7	1.6	1.4	1.4	1.3	
1Q94	7.74	6	0.94	0.98	6.6	6.4	5.9	5.2	4.4	3.0	2.6	2.4	2.3	1.8	1.9	2.1	
2Q94	8.29	7	1.18	0.95	6.5	6.4	5.9	5.6	5.1	4.4	3.6	3.2	3.0	2.2	2.2		
3Q94	8.48	7	1.40	0.97	6.3	6.1	5.9	5.7	5.2	4.6	4.0	3.4	2.8	2.4	2.4		
4Q94	8.88	7			6.3	6.2	5.8	5.6	5.1	4.8	4.1	3.5	3.1	2.1	2.0	3.0	1.9
1Q95	8.25	7			6.3	5.9	5.5	5.0	4.4	3.7	3.7	2.7	2.6	2.3	2.7	2.7	1.8
2Q95	7.42	6			6.0	5.5	4.9	4.1	3.0	1.6	1.9	1.2	2.2	1.8	2.5	2.5	
3Q95	7.34	7			5.8	5.5	4.8	3.9	2.7	2.0	1.1	1.2	2.3	1.8	1.7	2.9	
4Q95	6.79	7			5.6	5.0	3.9	3.2	1.9	1.3	1.2	1.5	2.2	1.8	1.7	2.5	
1Q96	7.53	7			6.2	5.4	5.5	4.7	3.7	2.7	1.9	1.3	2.4	1.7	2.7	3.2	
2Q96	7.80	7			6.4	6.0	5.6	4.9	4.1	3.3	2.5	1.6	2.7	1.7	2.2		

Table 10.10. Median broker forecasts: FHLMC/FNMA option cost (basis points).

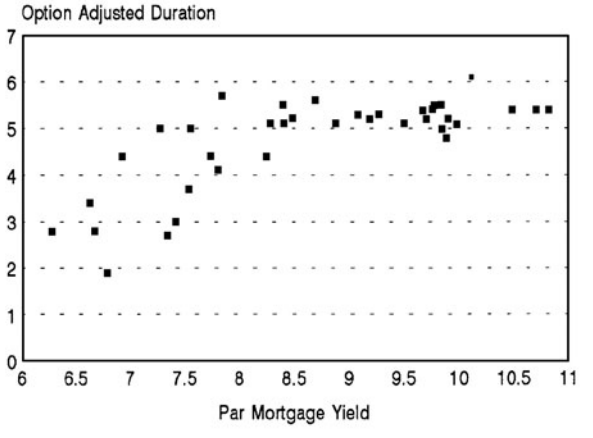
End of Quarter	Par Yld FNMA	# of Brokers	6.0	6.5	7.0	7.5	8.0	8.5	9.0	9.5	10.0	10.5	11.0	11.5	12.0	12.5	13.0
4Q86	8.77																
1Q87	8.69	1															
2Q87	9.85	1				57	62	60	60	65	73	79	85	70	56		
3Q87	10.70	1				62	67	65	65	70	74	75	90	91	92		
4Q87	10.11	1				57	62	61	61	66	72	82	92	88	83		
1Q88	9.76	3				24	30	38	51	58	82	93	86	71	41	38	
2Q88	9.83	3				34	43	53	59	68	78	88	98	93	88	43	
3Q88	9.91	2				33	28	34	50	61	74	85	93	84	75		
4Q88	10.48	3				36	45	50	58	68	82	91	78	56	38	33	
1Q89	10.81	3				32	42	46	57	67	88	91	99	108	116	69	
2Q89	9.70	3				44	48	53	66	79	94	105	112	101	84	70	
3Q89	9.88	2				8	20	23	33	44	56	68	48	45	47	49	
4Q89	9.50	2				19	20	28	34	42	53	62	55	43	39		
1Q90	9.97	2				22	23	25	32	39	49	56	52	40	36		
2Q90	9.67	2				22	23	27	32	40	48	57	50	41	35		
3Q90	9.78	2				22	22	26	30	35	43	52	50	46	40		
4Q90	9.27	3				23	21	27	38	45	50	49	33	31	25	12	
1Q91	9.08	3				16	22	25	36	46	54	54	33	13	24	14	

2Q91	9.19	2				12	16	21	29	39	45	44	32	18	10
3Q91	8.41	3				14	9	21	35	40	38	23	12	10	7
4Q91	7.55	4				4	11	32	37	34	30	28	17	3	1
1Q92	8.40	3				14	20	41	43	42	31	13	4	–15	–24
2Q92	7.84	3				2	42	43	42	26	11	7	3	–37	–49
3Q92	7.27	3			25	4	14	20	36	32	39	31	50	46	
4Q92	7.55	4			9	29	40	54	67	63	59	54	72	75	
1Q93	6.92	4			19	31	44	66	56	58	61	41	49	51	
2Q93	6.63	4		14	20	31	47	50	42	37	34	34			
3Q93	6.28	4	17	20	29	42	50	30	25	15	19	36			
4Q93	6.67	5	10	18	41	50	52	55	49	33	34	50	56		
1Q94	7.74	5	13	24	36	50	64	73	71	56	40	52	52	48	
2Q94	8.29	5	15	22	29	39	53	61	69	63	57	64	58		
3Q94	8.48	5	17	19	23	28	39	50	47	49	41	23	31		
4Q94	8.88	5	10	18	25	35	42	52	63	54	53	52	38	33	
1Q95	8.25	4	18	24	32	43	52	59	72	71	41	33	27	24	
2Q95	7.42	4	37	36	51	68	74	72	77	47	19	14	11	12	
3Q95	7.34	4	22	35	32	69	88	77	77	47	36	14	15	12	
4Q95	6.79	5	45	49	62	80	83	53	15	65	17	39	4		
1Q96	7.53	5	28	30	38	49	59	70	73	78	34	16	11	–16	
2Q96	7.80	5	31	48	53	56	62	75	53	67	53	48	39		

(a)



(b)



(c)

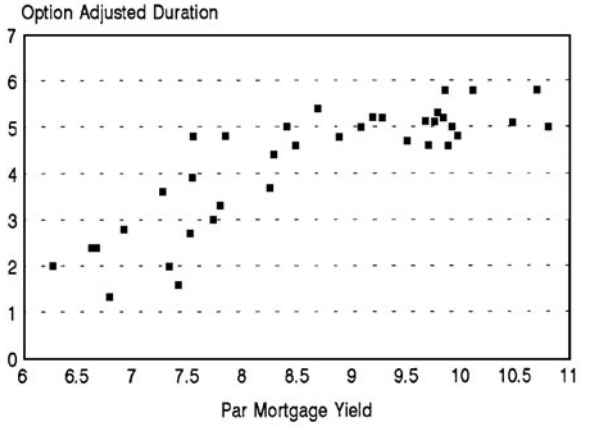
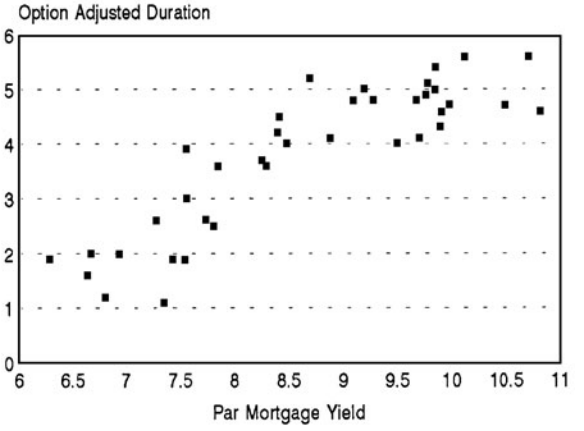
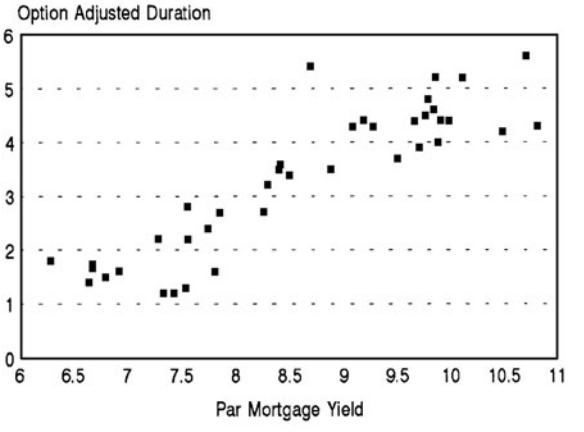


Figure 10.27. (a) Median Broker Forecasts: FNMA 7.5% Option Adjusted Durations.(b) Median Broker Forecasts: FNMA 8.0% Option Adjusted Durations. (c) Median Broker Forecasts: FNMA 8.5% Option Adjusted Durations. (d) Median Broker Forecasts: FNMA 9.0% Option Adjusted Durations. (e) Median Broker Forecasts: FNMA 9.5% Option Adjusted Durations. (f) Median Broker Forecasts: FNMA 10.5% Option Adjusted Durations.

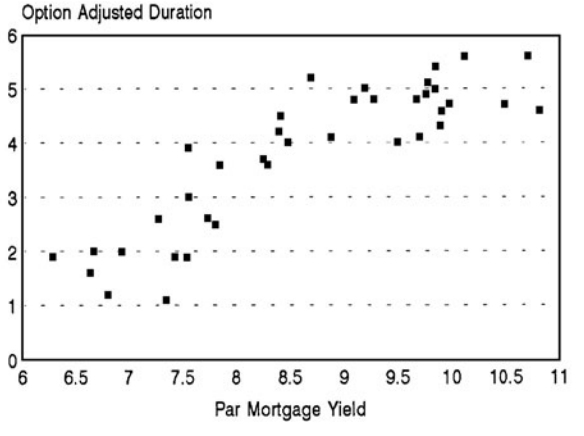
(d)


Option Adjusted Duration
6
5
4
3
2
1
0
6
6.5
7
7.5
8
8.5
9
9.5
10
10.5
11
Par Mortgage Yield

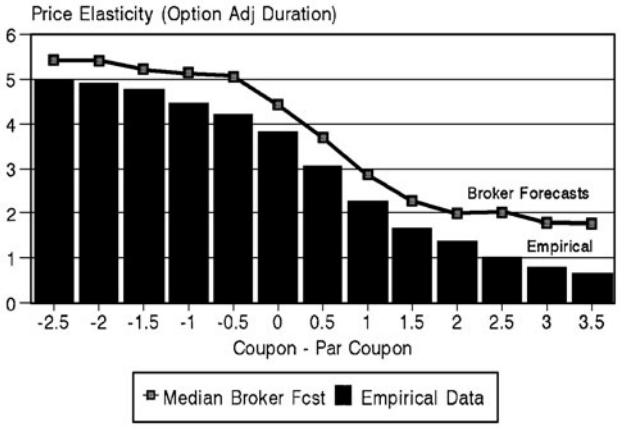
(e)


Option Adjusted Duration
6
5
4
3
2
1
0
6
6.5
7
7.5
8
8.5
9
9.5
10
10.5
11
Par Mortgage Yield

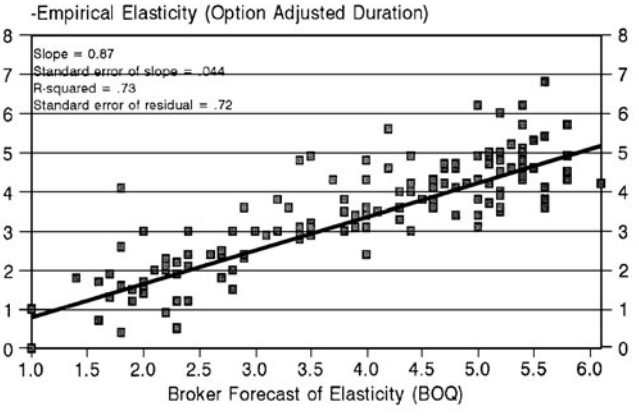
(f)


Option Adjusted Duration
6
5
4
3
2
1
0
6
6.5
7
7.5
8
8.5
9
9.5
10
10.5
11
Par Mortgage Yield

(a)



(b)



(c)

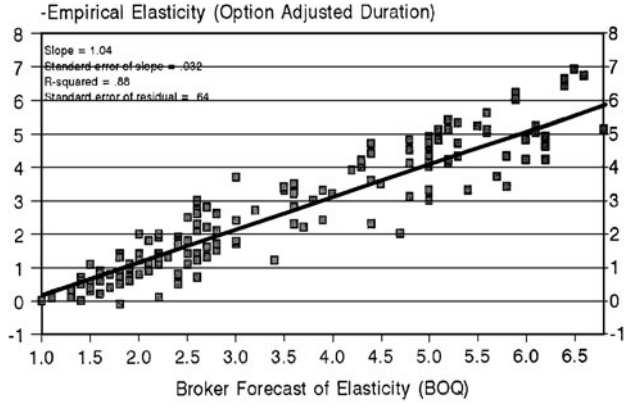
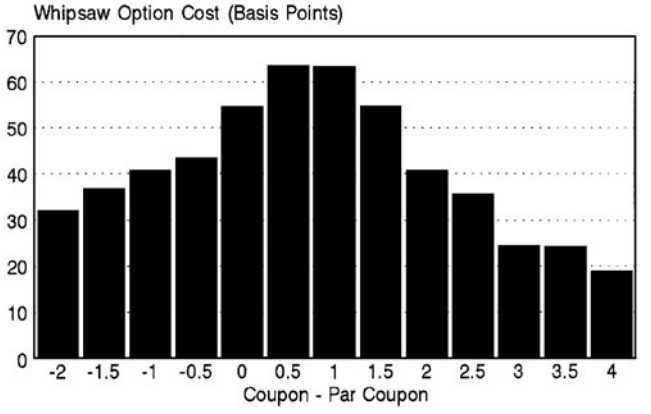
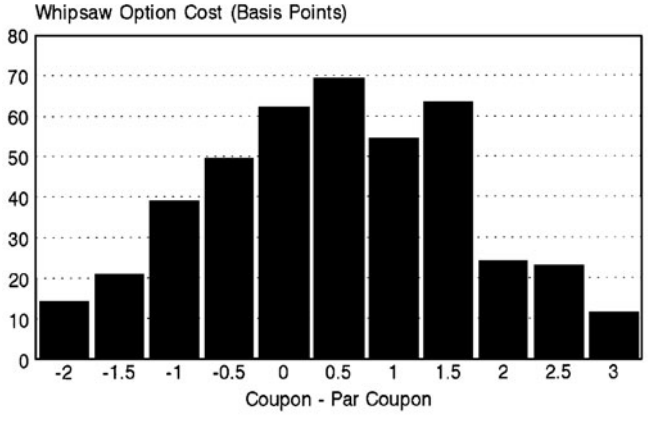


Figure 10.28. (a) **Median Broker Forecasts: Broker vs. Empirical Elasticities by C-R.** (b) **Broker Risk Estimates vs. Empirical Risk: 1987–1990 Quarterly Data, Coupons: 7%–12%.** (c) **Broker Risk Estimates vs. Empirical Risk: 1991–1994 Quarterly, Coupons: 7%–12%.**

(a)



(b)



(c)

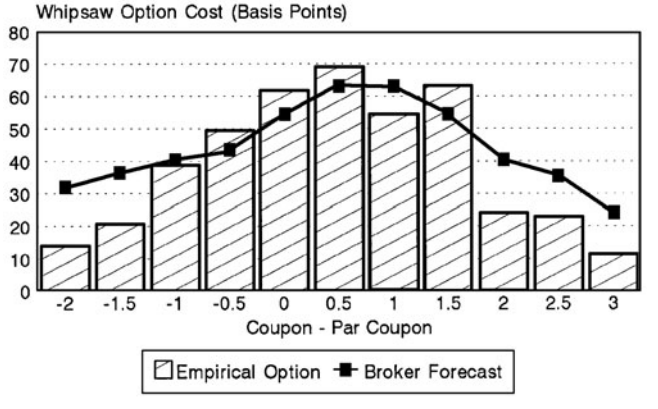


Figure 10.29. (a) Median Broker Forecasts 1987–1996: FHLMC/FNMA Whipsaw Option Costs by C-R. (b) Empirical Option Costs for FNMA 1988–1996: Median Broker Quarterly OA Durations. (c) Empirical Option Cost vs. Broker Forecast: Median Broker Quarterly OA Durations.

Table 10.11. FNMA dynamic option hedging cost ("model whipsaw").

	FNMA 6.5	FNMA 7.0	FNMA 7.5	FNMA 8.0	FNMA 8.5	FNMA 9.0	FNMA 9.5	FNMA 10.0	FNMA 10.5	FNMA 11.0	FNMA 11.5	FNMA 12.0
1987												
1988					0.23	0.29	0.23	0.35		0.11		0.08
1989					0.14	0.20	0.24	0.22		0.37		0.47
1990					0.32	0.72	0.75	0.69		0.51		0.34
1991					0.03	–0.02	–0.22	–0.24		0.04		0.13
1992			0.70	0.78	0.50	0.34	0.44	0.15	0.24	–0.20	0.53	
1993			1.33	1.33	1.06	0.51	0.13	0.14	0.30	0.31		
1994	0.04	0.33	0.90	1.28	1.13	0.96	0.86	0.71	0.48	0.49		
1995	0.24	0.35	0.54	0.81	1.25	0.99	0.88	0.27	0.25			
1996	0.39	0.57	0.42	0.48	0.34	–0.08	–0.24	0.04	0.02			
	0.22	0.42	0.78	0.94	0.56	0.43	0.34	0.26	0.26	0.23	0.53	0.26

5. DURATIONS AND EMPIRICAL OPTION COSTS FOR CONVENTIONAL MBS

This section presents the results of similar analysis for conventional mortgage-backed securities using data on FHLMC and FNMA coupons. Tables 10.8–10.10 present our survey results for the median broker forecasts of option-adjusted spreads, price elasticities (or option-adjusted durations), and option costs. Comparing tables 10.8 and 10.10, we see that the option costs are similar in magnitude to option-adjusted spreads, but estimated option costs are in a much more narrow range for standard MBS than for IOs. Most of the option cost estimates fall between 25 and 75 basis points for conventional MBS.

One of the reasons for this study is the observation from studying portfolio returns in practice that the whipsaw option cost appears often to be greater than this 25–75 basis point range in practice. We'll see what the results are shortly.

If you study table 10.9, you find quite predictably that when rates decline, prepayments surge, and durations (and corresponding dynamic hedge positions) shorten. To make this easier to see, figure 10.27 shows the coupon-by-coupon sensitivities of brokers' duration forecasts to changes in interest rates. As these have a positive slope, short hedges are reduced by buying futures when rates are low and prices high, and hedges are increased by selling futures when rates are high and prices low. This dynamic hedging pattern generates whipsaw costs due to the negative convexities of standard MBS, as expected.

Figure 10.28 shows the relationship of option-adjusted durations (or price elasticities) to relative coupons. The higher coupons (relative to par) have lower elasticities, as is sensible. Figure 10.28 shows that brokers' forecasts of durations are closely and significantly related to empirical durations for MBS, with the fit being tighter in the more recent period (1991–1996) than in earlier years (1987–1990).

Figure 10.29 shows median broker forecasts of option costs for conventional MBS, along with a graph of the empirical estimates of option cost sorted by relative coupon. Note that the empirical whipsaw option costs are quite similar in magnitude to the brokers' forecasts over the entire period from 1988 to 1996. The empirical option costs are slightly higher than the brokers' forecasts on coupons near par, but slightly lower than the brokers' forecasts on both superpremiums and deep discount securities.

Table 10.11 gives the annual data on empirical estimates of the option cost in conventional fixed-rate MBS. From these data, we see that the realized whipsaw cost year by year can be much higher than projected by the brokers, with some years giving whipsaws of 100 to 133 basis points. In other quiet years with little change in rates, whipsaw can be smaller than forecasted, or even slightly negative if model changes occur that generate positive benefits of "model whipsaw." On the whole, the option costs for conventional MBS turn out to be very similar to the median of brokers' forecasts.

It is interesting that when these dynamic hedging costs are estimated using Smith Breeden's empirical duration forecasts, empirical option costs that result are higher than shown here for lower and middle coupons, with a peak of 170 basis points on FNMA 7.5s and 8s in 1993. For GNMAs, the Smith Breeden elasticities yielded whipsaws of as much as 200 basis points in 1994.

Thus, once again we may have evidence of "model whipsaw" that affects option costs in practice. Alternatively, it is possible that taking the medians in the broker survey series tends to lead to underestimates of the whipsaws experienced in practice. This may be plausible, given the lags that may appear in adjustments in our median broker forecasts.

6. CONCLUSION

Theory says that interest-only strips should have large positive and negative option costs and benefits, due to their convexity patterns. While brokers clearly have some difficulty zeroing in on the proper costs of these strips, as well as the empirical durations, their forecasts appear rational in shape across coupons and time. The empirical option costs, however, exhibit somewhat greater amplitudes of fluctuation (larger positive and negative values) than did the median brokers' forecasts.

Given their size, option costs for interest-only strips and other high-risk mortgage derivatives deserve researchers' and traders' attention and continued study. Certainly, it would seem advisable that investors in high-risk derivatives put a band of error around their estimates of option-adjusted durations and whipsaw option costs and benefits. This probably is wise not just for IO strips but also for many mortgage derivatives that have volatile cash flow streams.

The results on option costs for conventional MBS show that the empirical dynamic hedging costs are similar in magnitude to those estimated by the brokers, although there is a great deal of annual fluctuation in the dynamic whipsaw experienced by portfolio managers. These fluctuations in the dynamic whipsaw cost of hedging mortgage prepayments are a return factor of the first order of magnitude for the returns on hedged mortgage portfolios.

ACKNOWLEDGMENTS The author thanks Dan Adler, Jenny Breeden, Andrew Cohen, Kerrie Hillman, and John Huff for research assistance and data for this project, which was presented at the Berkeley Program in Finance Conference in Honor of Fischer Black, Santa Barbara, California, September 29–October 1, 1996.

REFERENCES

Asay, Michael R., and Timothy D. Sears. "Stripped Mortgage-Backed Securities, Part I: Basic Concepts and Pricing Theory; Part II: Trading Strategies and Portfolio Applications." Goldman Sachs & Co., January 1988.

Black, Fischer, Emanuel Derman, and William Toy. "A One-Factor Model of In-

terest Rates and Its Application to Treasury Bond Options." *Financial Analysts Journal,* 46 (1), 1990, pp. 33–39.

Breeden, Douglas T. "Risk, Return, and Hedging of Fixed-Rate Mortgages." *Journal of Fixed Income,* 1 (September 1991), pp. 85–107.

Breeden, Douglas T. "Complexities of Hedging Mortgages." *Journal of Fixed Income,* 4 (December 1994), pp. 6–41.

Diller, Stanley. "Parametric Analysis of Fixed Income Securities: Options, Passthroughs, Convexity, and Asset/Liability Management." Goldman Sachs Financial Strategies Group, June 1984.

Hayre, Lakhbir. "A Simple Statistical Framework for Modeling Burnout and Refinancing Behavior." *Journal of Fixed Income*, 4 (December 1994), pp. 69–74.

Patruno, Gregg. "Mortgage Prepayments: A New Model for a New Era." *Journal of Fixed Income,* 4 (December 1994), pp. 42–56.

Richard, Scott F., and Richard Roll. "Prepayments on Fixed-Rate Mortgage-Backed Securities." *Journal of Portfolio Management,* 15 (Spring 1989), pp. 73–82.

Roll, Richard. "Stripped Mortgage-Backed Securities." Goldman Sachs & Co., Inc., October 1986.

11

The Supply and Demand of Immediacy: Evidence from the NYSE

Roger D. Huang & Hans R. Stoll

1.

Participants in a trading market can be partitioned into suppliers of immediacy and demanders of immediacy, as Demsetz (1968) first noted. Suppliers of immediacy stand ready to trade at the prices they post, and demanders of immediacy place market orders and require immediate execution at posted prices. On the New York Stock Exchange (NYSE), immediacy is supplied by specialists, who post bid and/or ask quotes for their own accounts, by securities firms that place quotes for their own accounts, and by public traders, who place limit orders. The bid–ask spread is often used as a measure of the cost of immediacy to investors, but it is rare that the supplier of immediacy earns the entire spread. A supplier of immediacy earns the difference between the bid (ask) at which he buys (sells) and the subsequent price at which he sells (buys). Because prices tend to move against suppliers of immediacy, revenues on a round-trip trade are less than the spread. Consequently, revenues of suppliers of immediacy are typically less than the spread.

Professional suppliers of immediacy, such as the specialist, must earn enough to cover their operating costs and a normal profit. Public limit order traders may be willing to supply immediacy at a loss so long as limit orders are not more costly than market orders. As markets become automated and public investors have direct access to markets, the role of public limit orders is likely to increase and the role of professional dealers is likely to decrease, a point made by Fischer Black (1971a, 1971b, 1995). Black was intrigued by the interplay of professional dealers, public traders, and automation. He concluded that dealers would not exist in a competitive equilibrium under a technology that gives public investors easy access to exchanges. In Black's view, an exchange is an order-matching system defined by the kinds of orders it accepts. In equilibrium, orders will be designed to offer traders protection against being picked off and to offer informed traders ways to capitalize on their information.

In this study, we measure the average revenues per share earned on the NYSE by suppliers of immediacy as a group by comparing the buying (sell-

ing) price to the average price at a later time, a measure we term the realized spread. Since stock market trading is a zero sum game (over the short run), what suppliers of immediacy earn, demanders of immediacy fail to earn. Consequently, the revenues of immediacy suppliers are the execution costs of investors that place market orders. Execution costs measured by the realized spread are less than execution costs measured in other ways, such as by the quoted spread, by the effective spread, or by the Roll (1984) implied spread. We compare the realized spread with these other measures. The revenues to immediacy suppliers are compensation for services rendered. By contrast, other measures of investor execution costs include a portion that represents losses to informed traders, a phenomenon first discussed by Bagehot (1971) and modeled by Copeland and Galai (1983) and Glosten and Milgrom (1985). When investors differ in their information sets, income is redistributed from those without information to those with information. This redistribution does not reflect the use of real resources (that is, labor and real capital) to provide immediacy.

The existing literature is characterized by different execution cost measures, although neither the appropriateness of the measures nor the relation among them have been addressed. Execution costs viewed from the perspective of institutional investors are analyzed by Beebower, Kamath, and Surz (1985), Berkowitz, Logue, and Noser (1988), Perold (1988), Schwartz and Whitcomb (1988), Arnott and Wagner (1990), Bodurtha and Quinn (1990), Chan and Lakonishok (1993), Perold and Sirri (1993), Wagner and Edwards (1993) and Keim and Madhavan (1996). Execution costs are estimated from observable quotes and transaction prices in a number of studies, including Roll (1984), Stoll (1989), and Hasbrouck (1993). Papers that focus on the differences in execution costs across markets include Blume and Goldstein (1992), McInish and Wood (1992), Christie and Huang (1994), Lee (1993), Petersen and Fialkowski (1994), and Huang and Stoll (1996a,b). Stoll (1979) and Cohen and Conroy (1990) study the impact of regulatory changes on trading costs. Huang and Stoll (1994) show that execution costs computed from transaction data affect the short-run dynamics of stock prices.

The contributions of this study are fourfold. First, we calculate the realized half-spread, a measure of the revenue per share to suppliers of immediacy. While studies of the price impact of block trades by Kraus and Stoll (1972) and Holthausen, Leftwich, and Mayers (1987) have provided similar measures for blocks, this is the first study to provide such measures for all trade sizes. Sofianos (1995) calculates specialist revenues from specialist trading reports, but our estimates encompass all immediacy suppliers. Second, we decompose the average revenue per share into revenues of securities firms and public limit orders. Third, noting that the revenues earned by suppliers of immediacy are the execution costs paid by demanders of immediacy, we provide evidence on the relation between the revenue measure and measures of execution costs such as the quoted spread and the effective spread. We also provide evidence on the Roll implied half-spread and a measure we term the perfect foresight half-spread. Fourth,

our study is distinguished by the large sample size. The database consists 67 million transactions in 343 NYSE stocks in the 5-year period 1987 to 1991.

2. TRANSACTION DATA

The data source is the transactions database for 1987 to 1991 maintained by the Institute for the Study of Security Markets. We restrict our sample to NYSE companies that were continuously in the S&P 500 Index from 1987 to 1991. This results in a sample of 343 securities with more than 67 million transactions.[1] Summary measures of execution costs and other data are computed for each stock in every month except October 1987, producing a sample containing 20,237 stock-months.

All trades that are coded as regular sales and all quotations that are BBO-eligible are included in our data set.[2] We confine the sample to trades with positive prices, positive volumes, and positive bid and ask quotes with nonnegative depths. To further minimize data errors, we apply the following filters:

1. Exclude quotes and prices when their decimal portions are not multiples of 1/16.
2. Exclude bid–ask quotes when the spread is greater than $2.
3. Exclude trade price p_t when $|(p_t - p_{t-1})/p_{t-1}| > .10$.
4. Exclude ask quote a_t when $|(a_t - a_{t-1})/a_{t-1}| > .10$.
5. Exclude bid quote b_t when $|(b_t - b_{t-1})/b_{t-1}| > .10$.
6. Exclude the entire day's data if the first trade price and quotes satisfy $\left|\frac{p_t - \frac{1}{2}(a_t + b_t)}{p_t}\right| > .10$.

Filter 6 is used to eliminate data errors such as a price that is wrongly coded for the entire trading day since filter 3 will not pick it up.

When trade prices are compared to quotes, we use the most recent prior NYSE quote that is time stamped at least 5 seconds earlier than the trade.[3] This 5-second rule is intended to compensate for the speedier reporting of quotes than of trades. If a trade does not have a prior BBO-eligible quote on the same day, it is excluded from the analysis.

2.1 Volume of Trading

Our data set contains a total of 67,153,925 trades. Table 11.1 presents summary statistics on the number of trades and share volumes for our sample. The number of trades in the 343 stocks on all markets, including regional exchanges and NASDAQ, has increased steadily since 1987 and amounted to 16,470,169 trades in 1991. On average, 70% of a stock's trades and 85.8% of its total share volume took place on the NYSE. In the period 1987 to 1991, the NYSE share of small trades declined from 74.2% to 64.4%, and the NYSE share of small trade volume declined from 78% to 68.8%. In the large trade size category, the NYSE share of trades and volume has increased.

Table 11.1. Number of trades and share volume of 343 S&P 500 stocks and percentage of trades and share volume conducted on the NYSE by trade size and year in the period 1987–1991 with October 1987 excluded. A small trade size has 1,000 shares or less, a medium trade size has greater than 1,000 but less than 10,000 shares, and a large trade has 10,000 or more shares.

	Trade Size			
	All	Small	Medium	Large
Number of trades in all markets				
1987	10,692,524	7,975,043	2,374,063	343,418
1988	11,843,178	8,830,727	2,595,486	416,965
1989	13,471,358	10,306,370	2,714,700	450,288
1990	14,676,696	11,346,885	2,876,433	453,378
1991	16,470,169	12,881,716	3,094,049	494,404
Percentage of all trades on NYSE[a]				
1987	0.787	0.742	0.925	0.889
1988	0.774	0.730	0.903	0.893
1989	0.737	0.691	0.885	0.893
1990	0.738	0.692	0.891	0.899
1991	0.700	0.644	*0.885*	0.917
Share volume of trading in all markets (in 100s)				
1987	199,076,076	30,204,014	70,463,104	98,408,702
1988	241,353,320	33,629,868	78,606,819	129,116,374
1989	233,519,518	37,342,106	82,820,458	113,356,537
1990	247,708,082	40,784,732	87,769,408	119,153,574
1991	254,055,237	45,600,250	94,126,963	114,327,538
Percentage of share volume on NYSE[a]				
1987	0.880	0.780	0.931	0.872
1988	0.871	0.757	0.913	0.879
1989	0.858	0.721	0.900	0.877
1990	0.859	0.725	0.905	0.881
1991	0.858	0.688	0.903	0.899

[a]The percentage of NYSE trades or volume is calculated for each stock-month, and the average percentage over all stock-months in the year is reported.

In 1991, small trades of 1,000 shares or less accounted for 78.2% of all trades and 17.9% of share volume. Medium trades of more than 1,000 but less than 10,000 shares accounted for 18.8% of all trades and 37.0% of share volume. Large trades of 10,000 or more shares accounted for 3.0% of all trades and 45.0% of share volume.

2.2 Trade Frequency

Table 11.2 documents the frequency of trades at various price locations: at the bid, at the ask, at the midpoint, between the midpoint and bid–ask quotes, and outside the bid–ask quotes. For example, in 1991, 33.6% of the trades took place inside the quotes—24.2% at the midpoint and 9.4% away from the midpoint but inside the quotes. The percentage of trades at the midpoint has increased over time at the expense of trades inside the quotes but not at the midpoint. This trend is consistent with a decrease in quoted spreads.

Table 11.2. Percentage of trades in 343 S&P 500 stocks by price location and trade size in the years 1987–1991 with October 1987 excluded. A small trade size has 1,000 shares or less, a medium trade size has greater than 1,000 but less than 10,000 shares, and a large trade has 10,000 or more shares. Monthly averages are first calculated across all stocks from the average values for each stock, and the figures below represent means of the monthly averages.

		Trade Size			
		All	Small	Medium	Large
At bid	1987	0.304	0.292	0.336	0.335
	1988	0.333	0.330	0.345	0.346
	1989	0.323	0.325	0.321	0.315
	1990	0.311	0.317	0.299	0.293
	1991	0.309	0.310	0.312	0.312
At ask	1987	0.350	0.334	0.384	0.382
	1988	0.341	0.322	0.386	0.383
	1989	0.343	0.335	0.361	0.365
	1990	0.326	0.326	0.325	0.313
	1991	0.352	0.352	0.357	0.341
At midpoint	1987	0.191	0.208	0.153	0.158
	198	0.208	0.222	0.176	0.181
	1989	0.238	0.242	0.230	0.241
	1990	0.253	0.246	0.268	0.292
	1991	0.242	0.241	0.238	0.260

continued

Table 11.2. continued

		Trade Size			
		All	Small	Medium	Large
Between bid and midpoint	1987	0.068	0.074	0.051	0.049
	1988	0.055	0.060	0.041	0.039
	1989	0.045	0.046	0.041	0.036
	1990	0.052	0.052	0.052	0.047
	1991	0.045	0.045	0.044	0.040
Between midpoint and ask	1987	0.078	0.085	0.057	0.047
	1988	0.059	0.064	0.046	0.037
	1989	0.049	0.051	0.043	0.033
	1990	0.057	0.058	0.053	0.044
	1991	0.049	0.050	0.044	0.036
Less than bid	1987	0.004	0.003	0.009	0.016
	1988	0.001	0.001	0.003	0.007
	1989	0.001	0.000	0.001	0.005
	1990	0.001	0.001	0.001	0.006
	1991	0.001	0.001	0.002	0.005
More than ask	1987	0.005	0.003	0.0 10	0.0 14
	1988	0.002	0.001	0.003	0.006
	1989	0.001	0.001	0.002	0.005
	1990	0.001	0.001	0.002	0.005
	1991	0.001	0.001	0.003	0.005

Surprisingly, trading inside the quotes is just as frequent for large trades as for small trades. We suspect that this reflects the prenegotiation of block trades and the presence of preblock trading to bring market prices in line with block prices.

Finally, table 11.2 discloses a small asymmetry in that the proportion of trades at the bid is less than the proportion of trades at the ask (.309 versus .352 in 1991). This is surprising since sellers are usually thought to be more anxious than buyers, which would imply greater frequency of trades at the bid than at the ask. Since this pattern is present in all trade sizes and since purchases equal sales, the asymmetry implies that trades at the quote midpoint are more likely to be at undisclosed bid prices than at undisclosed ask prices.

3. REVENUES OF IMMEDIACY SUPPLIERS: THE REALIZED HALF-SPREAD

Suppliers of immediacy are passive traders. They place quotes or limit prices and wait for incoming market orders with which to trade. A supplier of immediacy who buys (sells) at the bid (ask) price does so in anticipation of ultimately reversing that position at a higher (lower) price. If trades can be reversed at better prices, suppliers of immediacy earn revenues. If not, they make losses.

The reversal after a trade at the bid or the ask measures the revenue per share realized by a supplier of immediacy in a single trade.[4] We term this measure the realized half-spread. We say "realized" because the measure calculates what suppliers of immediacy actually realize as reflected in subsequent market prices and "half-spread" because we measure revenue on a single trade, whereas the spread is a measure of revenue on a round trip consisting of two trades.

The realized spread measure requires identification of a trade as a purchase or sale. In particular, price reversals cannot be reported for trades at the quote midpoint, where buys and sells cannot be distinguished. We expect that suppliers of immediacy do earn revenues from trades at the midpoint; we just cannot measure them.[5] Our measure of the *realized half-spread*, π, is the average price reversal subsequent to a trade. For trades at the bid, the average reversal is

$$\pi_b = E(\Delta\, p_{t+\tau} \mid b_t) = E(p_{t+\tau} - p_t^b) \tag{1}$$

where p_t is the transaction price at time t, τ is the time until the subsequent trade, b denotes trades at the bid, and $E()$ is the mean operator. For trades at the ask, the average reversal is

$$\pi_a = -E(\Delta\, p_{t+\tau} \mid a_t) = -E(p_{t+\tau} - p_t^a) \tag{2}$$

where a denotes trades at the ask. Note that the subsequent price at $t + \tau$ is a transaction price that may be either at the bid or the ask or inside the quotes. On average, for initial trades at p_t^b, the expected price change is positive. For initial trades at p_t^a, the expected price change is negative. The expected reversal, or realized half-spread, is positive in each case. We assume, in other words, that the suppliers of immediacy are able to liquidate at the average subsequent trade price. Since the trade price will sometimes be at the bid and sometimes at the ask, we expect that suppliers of immediacy are active about half the time and passive about half the time in unwinding their positions.

The reversal compensates the provider of immediacy for costs associated with processing the order and for inventory risk. For professional suppliers of immediacy, order-processing costs include wages, clearing charges, communication costs, and the like. For public limit orders, order-processing costs include the costs of monitoring the order. The inventory cost of providing immediacy arises from the risk of taking on unwanted

inventory or of holding unwanted inventory, which includes the cost of nonexecution.

Two opposing considerations enter in specifying the period τ over which the reversal is measured. On the one hand, τ should be long enough for the price at $t + \tau$ to reflect an independent trade. In particular, we do not look at the immediately following price because trades tend to cluster at the bid or the ask. On the other hand, choosing too large a τ increases the likelihood of confounding the measure with additional information or information-based trades. In our empirical work, τ is 5 minutes or 30 minutes. The choice of 5 minutes is intended to be long enough to overcome clustering yet short enough not to be unduly influenced by random price changes resulting from unexpected news events. Additional estimates based on τ at 30 minutes provide a check on the robustness of our results.

For each trade on the NYSE, we determine the trade price, p_t, and its location with respect to the bid or ask. For the 5-minute horizon, we choose the first trade price on any exchange after 5 minutes but no more than 10 minutes have elapsed, p_{t+5}, and we calculate $p_{t+5} - p_t$. If no subsequent price is available within the specified time span on the same day, no price change is calculated. The 30-minute standard follows the same procedure except that the first price after 30 minutes but no more than 35 minutes have elapsed is used to calculate the price change.[6]

Table 11.3 reports the average 5-minute and 30-minute price changes for transactions taking place at the ask and for transactions taking place at the bid. We present dollar measures of trading costs rather than percentage measures because the cross-sectional variation in stock prices is greater than the cross-sectional variation in execution costs per share. Dividing by price would only increase the difference across stocks in measured costs. The table also presents the results of *t*-tests of the null hypothesis that the trading cost estimate is zero. All values, except those with an asterisk, are significant at the 5% level. The test results using the Fisher sign test are similar and are not reported here.

As expected, the average price change after a transaction at the ask is negative and the average price change after a transaction at the bid is positive. In 1991, for example, the average revenue per share of a trader who bought at the bid and sold 5 minutes later was 3.2 cents. The average revenue of a trader who sold at the ask and covered 5 minutes later was 2.4 cents. The corresponding revenues for the 30-minute horizon were 3.2 and 1.9 cents, respectively.

The average price reversals do not appear to become smaller over the period 1987 to 1991. This implies that any efficiencies in trading and any reductions in spreads have not been reflected in reduced revenues for suppliers of immediacy. It has been suggested by Lee (1993) that competing markets are diverting the easy trades of uninformed investors away from the NYSE. If that were the case, realized spreads should decline over time on the NYSE, and there is no indication of that in our data.

Table 11.3. Price reversal measure of execution costs on the NYSE in 343 S&P 500 stocks by trade size and year in the period 1987–1991 with October 1987 excluded. The figures are in dollars per share. A small trade size has 1,000 shares or less, a medium trade size has greater than 1,000 but less than 10,000 shares, and a large trade has 10,000 or more shares. Monthly averages are first calculated across all stocks from the average values for each stock, and the figures below represent means of the monthly averages. An asterisk represents a *p*-value greater than 5% using the *t*-test.

	Trade Size			
	All	Small	Medium	Large
5-minute price change at bid				
1987	0.032	0.035	0.032	0.039
1988	0.026	0.029	0.020	0.028
1989	0.033	0.037	0.025	0.026
1990	0.032	0.035	0.027	0.029
1991	0.032	0.036	0.026	0.029
5-minute price change at ask				
1987	–0.017	–0.023	–0.009	–0.003*
1988	–0.026	–0.032	–0.016	–0.007
1989	–0.030	–0.037	–0.018	–0.006
1990	–0.027	–0.032	–0.016	–0.011
1991	–0.024	–0.030	–0.013	–0.005
30-minute price change at bid				
1987	0.043	0.043	0.046	0.045
1988	0.024	0.027	0.020	0.021
1989	0.033	0.037	0.025	0.027
1990	0.031	0.034	0.025	0.024
1991	0.032	0.036	0.027	0.028
30-minute price change at ask				
1987	–0.010	–0.017	–0.001*	0.010
1988	–0.024	–0.031	–0.013	0.001*
1989	–0.027	–0.035	–0.015	0.005
1990	–0.027	–0.032	–0.015	–0.005
1991	–0.019	–0.025	–0.006	0.007

Somewhat to our surprise, the per share reversal does not increase with trade size. For trades at the bid, the reversal is about the same for all trade sizes, which means that total revenues increase linearly with the number of shares. Apparently, these revenues, combined with economies of scale in handling a large trade, are enough to offset the higher inventory risk present in a large trade.

Evident in the data is an asymmetry in the price reversals after trades at the ask and at the bid. Formal tests (not reported here) confirm that price reversals at the bid and the ask are reliably different from one another. In the small trade size category, the asymmetry is small. In 1991, for example, the average reversal after trades at the ask was 3 cents and the average reversal after trades at the bid was 3.6 cents. In the large size category, the asymmetry is larger. In 1991, for example, the average reversal after large trades at the ask was 0.5 cents and the average reversal after trades at the bid was 2.9 cents. This difference is too large to be explained by trades inside the quotes. It reflects the well-known asymmetry in the price behavior around block purchases and block sales. Kraus and Stoll (1972) and Holthausen, Leftwich, and Mayers (1987) have found that prices reverse after sales but not after purchases, which implies that sales have a market impact cost but that purchases do not. The data in table 11.3 show this same pattern. Chan and Lakonishok (1993) examined large packages of institutional trades and reached a similar conclusion. They found that prices after the completion of a package of sales tended to reverse by about 0.10%, whereas prices tended to continue after a package of purchases. Assuming an average stock price of $35, a 0.10% reversal is 3.5 cents per share.

4. DECOMPOSITION OF REVENUES: RETURNS TO LIMIT ORDERS AND TO SECURITIES FIRMS

The realized half-spread measures the average revenue to all suppliers of immediacy. On the NYSE, some providers of immediacy are public customers who place limit orders, while others are securities firms that quote bids and offers and that earn a living from supplying immediacy. We include in securities firms not only specialist firms but also other firms that trade for their own accounts. In this section, we make inferences about the revenues of these two categories of immediacy suppliers. Presumably, firms earn larger reversals than public limit orders because they are closer to the market and can more quickly adjust their orders, thereby reducing the chances of being "picked off." If that were not the case, firms could not cover their costs and would exit the market. A decomposition is important when comparing a mixed dealer/auction market like the NYSE with a dealer market like NASDAQ. Even if the two markets exhibit different average reversals, there is still the possibility that dealers in the two markets earn the same return.

We divide the market into active traders, who demand immediacy by placing market orders, and passive traders, who supply immediacy by

placing limit orders. Firms, f, and public investors, p, can in principle be either active or passive. We can represent these possibilities with the following 2×2 trading table in which the cells can represent the proportion of trading, α, or the price reversal, π:

		Active Market orders: Firms, f	Active Market orders: Public, p
Passive Limit orders	Firms, f	α_{ff} π_{ff}	α_{fp} π_{fp}
	Public, p	α_{pf} π_{pf}	α_{pp} π_{pp}

Where α_{ij} is the fraction of shares traded in which i is passive and j is active, where i, j can take on the value f (firm) or p (public), and π_{ij} is the average price reversal after trades in which i is passive and j is active, where i, j can take on the value f (firm) or p (public). A positive reversal indicates earnings by suppliers of immediacy, and a negative reversal indicates losses by suppliers of immediacy.

The trading table can apply to the bid side of the market, where passive buyers and active sellers meet, or to the ask side of the market, where passive sellers and active buyers meet. Our discussion usually refers to the bid side.

To help identify the fractions and the price reversals in the trading table, we tap three different data sources and state three different conditions that must hold on the elements in the trading table. The first condition defines the overall average realized half-spread, $\pi\ldots$, as

$$\pi\ldots = \sum_{ij} \alpha_{ij}\pi_{ij} \tag{3}$$

The first data source, which provides information on $\pi..$, is the aggregate estimate of the average realized half-spread for passive buyers or passive sellers, which can be computed from our data as presented in tables 11.2 and 11.3. These estimates of $\pi..$ are reported as lines 1 and 2 in table 11.4. These estimates account for the frequency of trading at different trade locations and are averaged under the assumption that the reversal for trades at the midpoint is 25% of the reversal for trades at the quote:

$$\pi.. = [w(\Delta P_b) + (.5 - w)0.25(\Delta P_b)]/0.5$$

where ΔP_b is the average price change after a trade at the bid for all trades as reported in table 11.3, w is the fraction of trades at prices below the midquote as reported in table 11.2, $(.5 - w)$ is the fraction of trades at the midquote, and 0.25 is the fraction of the price change after a trade at the bid that is assumed to occur after a trade at the midpoint. The reversal after a trade at the midpoint is nonzero because we assume that immediacy suppliers receive some compensation, even for trades at the midquote. The

Table 11.4. Realized half-spread, securities firms' per share trading revenues, and implied price changes after limit order trades.

	Year				
	1987	1988	1989	1990	1991
1. Avg. ΔP after passive purchase[a] (in cents)	2.6	2.2	2.6	*2.5*	*2.5*
2. Avg. ΔP after passive sale[b] (in cents)	–1.5	–2.2	–2.5	–2.2	–2.0
3. Firms' participation[c]	0.265	0.228	0.237	0.230	0.231
4. Firms' trading gains[d] (in millions of dollars)	1,854	2,255	2,756	830	2, 020
5. Share volume on exchanges[e] (in millions)	64,083	52,666	54,417	53,746	58,296
6. Firms' trading gains per share[f] (in cents)	5.46	9.40	10.70	3.36	7.50
7. Implied ΔP after limit order buy[g] (in cents)	–0.63	–3.84	–4.70	1.77	–1.79
8. Implied ΔP after limit order sell[h] (in cents)	2.79	3.84	4.89	–1.21	2.72

[a]Calculated as $[w(\Delta P_b) + (.5 - w)\ 0.25\ (\Delta P_b)]/0.5$, where ΔP_b is the average price change after a trade at the bid for all trades (table 11.3), w is the fraction of trades at prices below the midquote (table 11.2), $(.5 - w)$ is the fraction of trades at the midquote. and 0.25 is the fraction of the price change after a trade at the bid assumed to occur after a trade at the midpoint.
[b]Calculated in the same way as the average price change after passive purchase described in note a.
[c]Specialists and other member firms' purchases plus sales as a fraction of twice total volume. Taken from the NYSE Fact Book.
[d]Based on FOCUS reports as described in Stoll (1994). The figure represents gains from trading on the NYSE, AMEX, and regional exchanges. Investment gains and underwriting gains of firms are not included.
[e]Share volume of trading in stocks listed on the NYSE, AMEX, and regional exchanges. Taken from SEC Annual Report.
[f]Firms' trading gains divided by twice the firms' volume.
[g]Calculated as described in text so that the average price change is an average of the gains of firms and the gains of limit orders. All firms' volume is assumed to be passive.

proportion of trades at the midpoint is divided between passive purchases and passive sales such that the overall frequencies of passive sales and passive purchases are both 50%. (Implicit in this procedure is the assumption that every trade has an active trader on one side and a passive trader on the other.) We assume that the reversal for trades between the quote and the midpoint is the same as the reversal at the quote reported in table 11.3. In 1991, the average reversal for passive buyers, which we take as the left-hand side of equation (3), is calculated as 2.5 cents.

The second condition defines the participation rate for securities firms—specialists and other NYSE members. The participation rate of firms P_f is defined as firms' purchases plus sales divided by twice the volume. It can be written as

$$P_f = \frac{\alpha_{ff} + \alpha_{fp} + \alpha_{ff} + \alpha_{pf}}{2} = \alpha_{ff} + \frac{\alpha_{fp} + \alpha_{pf}}{2} \tag{4}$$

The second data source is the participation rate as reported by the NYSE. In 1991, the participation rate was $P_f = 0.231$.[7] Corresponding data for other years are in line 3 of table 11.4.

The third condition defines the firms' average trading gain, G_f, as the weighted average of the gains from passive and active trading:

$$G_f = \frac{\alpha_{ff}\pi_{ff} + \alpha_{fp}\pi_{fp} + \alpha_{ff}\left(-\pi_{ff}\right) + \alpha_{pf}\left(-\pi_{pf}\right)}{\alpha_{ff} + \alpha_{fp} + \alpha_{ff} + \alpha_{pf}} = \frac{\alpha_{fp}\pi_{fp} + \alpha_{pf}\left(-\pi_{pf}\right)}{2\alpha_{ff} + \alpha_{fp} + \alpha_{pf}} \tag{5}$$

where the first two terms in the numerator of the expression between the two equal signs represent gains from supplying immediacy to firms and the public and the second two terms represent gains from active trading with firms and the public. Equation (5) says that the firms' total trading gains come from reversals they earn as providers of immediacy and from continuations in price (negative reversals) after active trading.

The third data set, which provides information on G_f, is the FOCUS (Financial and Operational Combined Uniform Single) report. All securities firms report their revenues, expenses, and balance sheet items to the Securities and Exchange Commission on this form. Our interest is in securities firms' trading gains in exchange-listed stocks, something that is reported. Trading gains are earned if price changes are favorable after a trade. As in Stoll (1995), we compute per share trading gains on exchanges from these data and data on securities firms' share volume of trading. Lines 3, 4, and 5 of table 11.4 are used to derive firms' trading gains per share for the years 1987–1991 in line 6. For example, in 1991, securities firms' revenues from trading in exchange-listed stocks were $2020 million (line 4).[8] Dividing by firms' purchases plus sales (the product of twice line 5 and line 3) yields a per share trading gain of 7.5 cents (line 6).[9]

In addition to equations (3), (4), and (5), we have the following add-up constraint on the participation rate of different parties:

$$1 = \sum_{ij} \alpha_{ij}$$

It is easily seen that not all the variables in the 2×2 table can be identified with equations (3) to (5) and the adding-up constraint. Thus, our decomposition of the realized spread between securities firms and public limit orders is one of many possible decompositions. We analyze two informative decompositions.

In the first decomposition, we assume that firms always act as suppliers of immediacy so that $\alpha_{ff} = \alpha_{pf} = 0$. Since π_{ff} and π_{pf} only appear in conjunction with α_{ff} and α_{pf}, respectively, we can ignore the first column of the 2×2 table. Under this scenario, and using the fact that $P_f = 0.231$ in 1991, equation (4) implies that firms' participation as suppliers of immediacy is $\alpha_{fp} = 0.462$, or passive purchases average 0.462 of active sellers. Therefore, using the add-up constraint, immediacy for the remaining fraction of trades, $\alpha_{pp} = 0.538$, is provided by public limit orders. A dealer/auction market like the NYSE is distinguished by the ability of limit orders to provide immediacy to public market orders. By way of contrast, in a dealer market like the NASDAQ Stock Market, $\alpha_{pp} = 0.0$.

Next, using equation (5), with $G_f = 7.5$ in 1991 and $\alpha_{fp} = 0.462$, we calculate that $\pi_{fp} = 7.5$ when firms' trading gains come entirely from providing immediacy to the public. In 1991, the overall reversal was 2.5 cents, and substituting the values of α_{pp}, α_{pp}, and π_{fp} into equation (3) to solve for price reversals after public limit buy orders yields $\pi_{pp} = -1.79$. This figure is reported in line 7 of table 11.4. When firms earn all their trading revenues by supplying immediacy to the public, the overall reversal of 2.5 cents is simply a weighted average of the reversal earned by firms (7.5 cents) and by public limit buy orders (–1.79 cents), where the weights are the fraction of trading by each, 0.462 and 0.538, respectively.

The implied price changes after public limit order purchases (when a positive price change is expected) and sales (when a negative price change is expected), reported for the years 1987–1991 in lines 7 and 8 of table 11.7, indicate that price changes after public limit order trades tend to be adverse. They fall after limit order purchases in four of the five years, and they rise after limit order sales in four of the five years. In other words, these calculations imply that public limit orders face a winner's curse: if they trade, they tend to lose. Securities firms are able to avoid this winner's curse because they are better informed and are closer to the market. Similar conclusions are reached by Simpson (1994) in an analysis of TORQ data. He finds that public limit orders are "picked off" more frequently than specialist limit orders. A second cost facing public limit orders is the cost of nonexecution, which is not reflected in these data. Harris and Hasbrouck (1996) also find that changes in the quote after a trade by a limit order are adverse. For example, after a limit order purchase at the bid when the spread is 1/8, the bid price falls on average 7.3 cents and 9.2 cents depending on order size (their table 7). Their numbers are much larger than ours because they are quotes, not transaction prices. If passive limit order buyers dispose of their position by entering active market orders to sell, the adverse price change would be this large, but this is unlikely. Our estimates are based on transaction prices, and they reflect the fact that, in disposing their positions, suppliers of immediacy may be active or passive.

Despite the winner's curse, and depending on the cost of nonexecution, it may be rational for public investors to place limit orders. We estimate that limit orders to buy lost an average of 1.79 cents per share in 1991, and

that, under the assumptions of our first decomposition, public market orders to buy lost an average of 2.0 cents per share, as shown in line 2 of table 11.4. Not included in the cost of a limit order is the cost of nonexecution. While our results are fairly crude estimates, they are consistent with an equilibrium in which the expected cost to the public investor of trading with a market order equals the expected cost of trading with a limit order.

In the second decomposition, we assume that securities firms trade by market order as well as by placing bids and offers. Since securities firms earn positive revenues, the effect would be to raise the public's cost of trading by market order above the average realized spread and to reduce the cost of trading by limit order. For this scenario, suppose that 5% of the securities firms' active trading is vis-à-vis other firms acting passively and 10% is vis-à-vis public limit orders. In addition, we assume that firms earn the same from passive and active trading with the public (that is, $\pi_{fp} = -\pi_{pf}$) and that firms neither gain nor lose when they trade with each other (that is, $\pi_{ff} = 0.0$).

As in the previous decomposition, we use 1991 values for equations (3) to (5). Substituting $\alpha_{ff} = 0.05$ and $\alpha_{pf} = 0.10$ into equation (4), we can solve for $\alpha_{fp} = 0.262$. Therefore, given the add-up constraint, we can solve for $\alpha_{pp} = 0.588$. Under our assumptions, equation (5) implies that $\pi_{fp} = -\pi_{pf} = 7.5$ when $\alpha_{fp} = 0.262$. Finally, we use equation (3) with the appropriate substitutions to solve for $\pi_{pp} = 2.19$. The trading table can now be completed as follows:

		Active Market orders: Firms, f	Active Market orders: Public, p	Average π_j
Passive Limit orders	Firms, f	$\alpha_{ff} = 0.05$ $\pi_{ff} = 0.0$	$\alpha_{fp} = 0.262$ $\pi_{fp} = 7.5$	6.30
	Public, p	$\alpha_{pf} = 0.10$ $\pi_{pf} = -7.5$	$\alpha_{pp} = 0.588$ $\pi_{pp} = 2.19$	0.78
	Average $\pi_{j.}$	−5.00	3.82	2.5

In this case, public market orders pay an average of 3.82 cents (the price reverses by 3.82 cents), which is larger than under the decomposition that assumes firms act only passively. Firms make an average of 5 cents (the price continues by 5 cents) when trading actively and an average of 6.3 cents (the price reverses by 6.3 cents) when trading passively. Public limit orders earn an average of 0.78 cents. The average price reversal is maintained at 2.5 cents.

In both decompositions, public market orders pay 7.5 cents to firms, while the average cost of immediacy for public market orders is 2.5 cents in the first decomposition and 3.82 cents in the second. These data compare with an average reversal of about 15 cents on the NASDAQ Stock Market in 1991.[10] We assume that the reversal on NASDAQ is earned by dealers because all trades are with dealers. The difference between the

average reversal on the NYSE (2.5 cents) and the average reversal on NASDAQ (15 cents) is striking. The difference between the average reversal earned by firms on the NYSE (7.5 cents) and the average NASDAQ reversal earned by dealers (15 cents) is less striking, yet the difference is still large.[11] One may ask whether the earnings of NYSE firms from providing immediacy to the public are underestimated. First, it is possible that the FOCUS data underestimate firms' trading gains or that other assumptions made in Stoll (1995) lead to an underestimate. Second, the estimate also depends on the degree to which firms trade actively and on the earnings per share from active trading. If firms earn less on active trading than on passive trading, their earnings from providing immediacy to the public must be higher than 7.5 cents.

In summary, we have shown that the realized spread is small yet consistent with data from an entirely different source—FOCUS reports filed by securities firms. The data on securities firms' revenues and trading volume permit inferences as to the revenues of limit orders. Our results imply that limit orders are "picked off," a result consistent with an equilibrium in which public investors choose limit orders and market orders so as to equalize their costs on the margin. Securities firms earn greater revenues per share than the realized half-spread, and public limit orders earn less. Comparisons with the realized spread on the NASDAQ market indicate that suppliers of immediacy on that market earned substantially more per trade than on the NYSE. This is because limit orders were not allowed to provide immediacy on NASDAQ and because NASDAQ dealers earned a larger reversal than NYSE firms.

5. RECONCILING REVENUES OF IMMEDIACY SUPPLIERS AND BID–ASK SPREADS

We have shown that immediacy suppliers earn an average of 2.5 cents per share (or less), a number that seems quite small when compared to half the quoted spread or half the minimum tick size, but we have also shown that the number is consistent with reported trading gains of securities firms and the distribution of volume between securities firms and limit orders. If immediacy suppliers earn 2.5 cents on average, immediacy demanders pay 2.5 cents on average. We now ask whether frequently used measures of market order execution costs—the quoted spread and the effective spread—are consistent with a realized spread of 2.5 cents.

The quoted bid–ask spread is used to measure the trading cost on a round trip of two trades. For a single trade, the execution measure is the *quoted half-spread*:

$$\frac{1}{2}s_t = \frac{1}{2}(a_t - b)_t$$

where s_t is the quoted bid–ask spread, a_t is the quoted ask price for a normal quantity, and b_t is the quoted bid price for a normal quantity.[12] The

half-spread implicitly uses the quote midpoint at the time of a trade as the standard against which trades that occur at the bid or the ask quotes are compared. Given the minimum tick size of 12.5 cents, the minimum quoted half-spread is 6.25 cents, which is more than double the realized half-spread reported in table 11.3. The first panel in table 11.5 shows that the average quoted half-spread on the NYSE for the 343 S&P 500 stocks in our data set amounted to 14.2 cents in 1987 and 11.1 cents in 1991. Since quotes are not available for trades of different sizes, no categorization by trade size is possible.

Is a quoted spread of 11.1 cents consistent with a realized spread of 2.5 cents? The answer is "approximately yes." Quoted spreads overstate the average execution costs to market orders for three reasons. First, some quotes are excessively wide and attract no trades. This can be seen from the second and third panels of table 11.5, which provide data on the average effective half-spread for trades at the bid and at the ask. The *effective half-spread* is defined as

$$z_t = |p_t - q_t|$$

where z_t is the effective half-spread and $q_t = \frac{1}{2}(a_t - b_t)$ is the quote midpoint existing at the time of the trade. Blume and Goldstein (1992), Lee (1993), and Petersen and Fialkowski (1994) base most of their analyses of trading costs on this statistic, a measure termed "liquidity premium" by Lee. The effective half-spread for trades at the bid and the ask is the quoted half-spread weighted by the number of trades taking place at the quotes. If a quote attracts no trades, it receives zero weight. The second and third panels of table 11.5 indicate that the quoted half-spread at which trades took place is about two cents less than the average quoted half-spread. For example, in 1991, the quoted half-spread for trades at the bid and at the ask was 9.1 cents per share as compared with a quoted spread of 11.1 cents.

A second reason that the realized spread is less than the quoted spread is that a substantial fraction of trades take place inside the quotes. Table 11.2 indicates that in 1991 33.6% of the trades took place inside the quotes. McInish and Wood (1992) report that limit orders bettering the displayed quote are frequently not displayed. Consequently, when a market order arrives and trades with the undisplayed limit order, a trade inside the disseminated quotes takes place. Brokers on the floor may also hold orders with undisplayed quotes inside the quoted spread. A second reason for trades inside the quoted spread is that the specialist "stops" stock; that is, he guarantees the posted quote but attempts to better it by negotiating with an incoming order.

The fourth panel of table 11.5 shows that the average effective half-spread was 6.7 cents in 1991, which reflects not only the fact that the quotes at which trades take place are narrower than the set of all quotes but also the fact that a third of the trades take place inside the quotes. Effective spreads increase with trade size, although the increase is not large. For example, in 1991 the effective half-spread was 6.5 cents for small trades,

Table 11.5. Contemporaneous measures of execution costs on the NYSE in 343 S&P 500 stocks by trade size and year in the period 1987–1991 with October 1987 excluded. The figures are in dollars per share. A small trade size has 1,000 shares or less, a medium trade size has greater than 1,000 but less than 10,000 shares, and a large trade has 10,000 or more shares. Monthly averages are first calculated across all stocks from the average values for each stock, and the figures below represent means of the monthly averages. All the estimates are significantly different from zero using the *t*-test.

	Trade Size			
	All	Small	Medium	Large
Quoted half-spread				
1987	0.142			
1988	0.124			
1989	0.113			
1990	0.113			
1991	0.111			
Effective ha1f-spread,[a] trades at the bid				
1987	0.126	0.123	0.131	0.134
1988	0.104	0.102	0.109	0.112
1989	0.092	0.090	0.097	0.101
1990	0.092	0.090	0.096	0.103
1991	0.091	0.089	0.096	0.101
Effective half-spread,[a] trades at the ask				
1987	0.126	0.124	0.129	0.130
1988	0.106	0.104	0.109	0.110
1989	0.093	0.092	0.097	0.098
1990	0.093	0.092	0.096	0.098,
1991	0.091	0.089	0.094	0.098
Effective half-spread,[a] all trades				
1987	0.095	0.090	0.108	0.114
1988	0.079	0.075	0.087	0.092
1989	0.068	0.066	0.073	0.078
1990	0.066	0.065	0.068	0.073
1991	0.067	0.065	0.071	0.076

[a]Equals $|p_t - q_t|$, where p_t is the trade price and q_t is the quote midpoint.

7.1 cents for medium trades, and 7.6 cents for large trades. The middle two panels of table 11.5 show that effective spreads for trades at the quotes increase with trade size, which suggests that quotes are widened in anticipation of large trades.

Over time, effective spreads and quoted spreads have fallen. For example, the overall effective spread declined from 9.5 cents in 1987 to 6.7 cents in 1991, corresponding to a decline in quoted half-spreads from 14.2 cents to 11.1 cents. Corresponding declines in effective spreads are evident in each size category. Possible explanations for the decline are increased competition for order flow (reflected in part in the NYSE's market share decline), new, more efficient trading technology, and the ability to hedge positions in futures and options markets.

A third reason that the realized half-spread is less than the quoted half-spread is that the quoted and effective spreads are wider than out-of-pocket execution costs as protection against informed traders, as described by Bagehot (1971) and modeled by Copeland and Galai (1983) and Glosten and Milgrom (1985). After a sale (purchase) at the bid (ask), quotes adjust downward (upward) to reflect the adverse information conveyed by the trade. Consequently, reversals are attenuated. Suppose the effective half-spread is 6.7 cents and that the expected value of adverse information reflected in a sale is 4 cents.[13] After a trade at the bid, the bid and ask prices would decline by 4 cents to reflect the information contained in the trade. This would cause the average reversal to be 2.7 cents rather than the effective half-spread of 6.7 cents. The adverse information component of the spread is not an execution cost but a redistribution of income from uninformed to informed traders. Active informed traders who place market orders impose losses on suppliers of immediacy, and active uninformed traders pay enough to cover the costs of supplying immediacy plus the losses of immediacy suppliers to informed traders, a point first made by Bagehot (1971).

Different demanders of immediacy incur different execution costs. One expects that securities firms and informed investors who trade actively incur smaller costs than uninformed public investors. Different suppliers of immediacy also earn different revenues. Securities firms must earn enough to cover their costs, and limit orders can be expected to earn less. Trading strategies may be able to reduce trading costs for any particular market participant. An appropriate standard against which execution costs should be measured is the average realized half-spread across all trades (perhaps categorized by trade size), for it reflects the market-wide out-of-pocket costs.

6. RECONCILING REVENUES OF IMMEDIACY SUPPLIERS AND THE ROLL IMPLIED HALF-SPREAD

Another well-known measure of execution costs was proposed by Roll (1984). Roll shows that the presence of a bid–ask bounce induces negative

serial covariance in price changes. He calculates an implied spread from the observed serial correlation of price changes under an assumed price-generating process. The *Roll implied half-spread* is $\sqrt{-\text{cov}}$, where cov is the serial covariance of successive price changes. Using daily data for the NYSE and the AMEX stocks in the period 1963 to 1982, Roll calculates an implied spread of 0.298%, or about 8.9 cents on a $30 stock. He interprets his result as a measure of the effective spread at which transactions actually occur. Roll's measure is particularly useful if data on bid and ask prices, which are required for the other execution cost measures, are not available. Like the reversal measure of trading costs, the implied half-spread is a prospective measure in the sense that it depends on price changes after the trade.

Under Roll's (1984) assumptions that quotes don't change in response to trades, his implied half-spread equals the effective half-spread. But quotes do respond to trades, and this fact causes the Roll implied spread to be greater than the realized spread. Stoll (1989) modifies the Roll (1984) model to allow quotes to adjust in response to trades. Huang and Stoll (1997) provide a general framework that incorporates the Roll (1984) and Stoll (1989) models as special cases.

We can analyze the relation between the effective half-spread, the Roll implied half-spread, and the realized half-spread by using the following model of transaction price changes from Huang and Stoll (1997):

$$\Delta p_t = \frac{S}{2}(Q_t - Q_{t-1}) + \lambda \frac{S}{2} Q_{t-1} + e_t \tag{6}$$

where S is the effective spread, $Q_t = -1$ if the trade is at the bid and 1 if the trade is at the ask, λ is the fraction of the effective half-spread by which quotes adjust in response to a trade, and e_t is the random error reflecting unanticipated news events, etc. Huang and Stoll (1997) show that this model yields the following serial covariance in price changes derived in Stoll (1989):

$$\text{cov}(\Delta p_t, \Delta p_{t-1}) = S^2 \left[\frac{\lambda^2}{4}(1-2\rho) - \rho^2(1-\lambda) \right] \tag{7}$$

where ρ is the probability of a price reversal. Roll assumes $\lambda = 0$, $\rho = 0.5$, which results in cov $= -S^2/4$. Solving for the effective half-spread gives $\frac{1}{2}S = \sqrt{-\text{cov}(\Delta p_t, \Delta p_{t-1})}$, as indicated above.

By taking the conditional expectation of equation (6), the model also yields the following expected price change[14]:

$$E(\Delta p_t | Q_{t-1}) = \left(\frac{\lambda}{2} - \rho \right) S Q_{t-1} \tag{8}$$

Under the Roll assumptions, $\lambda = 0$, $\rho = 0.5$, the expected reversal is $\pi = 0.5S$, the effective half-spread. However, we expect in general that $0 < \lambda < 1$ and $0.5 < \rho < 1$. Under these parameter restrictions, it will be the case that the *effective half-spread* > *Roll implied half-spread* > *realized half-spread*; that is,

$$\frac{S}{2} = \sqrt{-\text{cov}(\Delta p_t, \Delta p_{t-1})} > \pi$$

We examine these implications by estimating the Roll implied half-spread. We calculate a serial covariance for all trades in a given stock in a given month as follows. For each trade, we determine the trade price, p_t, the most recent prior price, p_{t-1}, and the next price, p_{t+1}. The trade at time t with price p_t takes place on the NYSE, and the other trades may take place on any exchange. The three prices provide two price changes, and we use all such pairs of price changes to calculate one serial covariance, cov, for each stock in each month:

$$\text{cov} = E\left[\left(p_{t+1} - p_t - \overline{\Delta p}\right)\left(p_t - p_{t-1} - \overline{\Delta p}\right)\right]$$

where Δp is the mean price change in the stock-month. Thus, there is one covariance estimate for each stock-month. This contrasts with the other execution cost measures, which are averages based on all eligible trades in the stock-month.

The implied half-spread reported in table 11.6 is calculated as the square root of the average covariance over all stock-months in each year. By taking the square root of the average, we avoid the downward bias, due to Jensen's inequality, present in the average of the square root of negative covariances.[15]

In table 11.6, we report the average Roll half-spread by year and for trade size categories. Roll's measure does not provide execution costs at specific price locations. In contrast, the reversal measures are meaningful only when calculated for specific price locations. As predicted, the implied half-spread is less than the quoted or the effective half-spread in table 11.5 and greater than the realized half-spread in table 11.3. For example, in 1991, the overall implied half-spread was 4.2 cents, whereas the effective half-spread was 6.7 cents. The realized half-spread was about 2.5 cents. These differences reflect the fact that prices do not bounce back by the full amount of the effective half-spread. Since trades convey information, prices tend to move in the direction of the trade, attenuating the bid–ask bounce. There is some tendency for Roll's spreads to increase with trade size but, with the exception of 1991, the tendency is not strong. Roll's spreads have also declined over time for small and medium trade size categories.

The differences between the Roll implied half-spread and the realized half-spread are not likely to be due to differences in the estimation procedures or to statistical problems such as those raised by Harris (1990). The covariance underlying the Roll measure is based on successive trade prices, whereas the reversals are based on trade prices that are 5 or 30 minutes apart. Since successive price changes are more likely to be positively correlated, and since positive serial correlation reduces Roll's spread, this is not an explanation for the larger Roll spread. The reversal calculations also omit trades inside the quotes, whereas the covariance

Table 11.6. Roll's half-spread measure of execution costs on the NYSE in 343 S&P 500 stocks by trade size and year in the period 1987–1991 with October 1987 excluded. The figures are in dollars per share. A small trade size has 1,000 shares or less, a medium trade size has greater than 1,000 but less than 10,000 shares, and a large trade has 10,000 or more shares. Monthly averages are first calculated across all stocks from the average values for each stock, and the figures below represent means of the monthly averages. An asterisk represents a *p*-value greater than 5% using the *t*-test.

	Trade Size			
	All	Small	Medium	Large
Roll's half-spread				
1987	0.054	0.053	0.060	0.058
1988	0.049	0.048	0.052	0.049
1989	0.045	0.046	0.048	0.044
1990	0.045	0.045	0.043	0.049
1991	0.042	0.041	0.045	0.131*

is calculated over all trades. Since trades inside the quotes are unlikely to have bigger reversals than trades at the quotes, this is also not an explanation for the larger Roll spread. Finally, as Harris has noted, the Roll measure is likely to be biased downward, yet it still exceeds the reversal measure.

7. MAXIMUM REVENUES OF IMMEDIACY SUPPLIERS

The realized half-spread measures the average price increase after a trade at the bid or the average price decrease after a trade at the ask. Some suppliers of immediacy, such as professional market makers, may be better able to anticipate price changes than others, such as public limit orders, and therefore earn more than the average realized half-spread. Similarly, some demanders of immediacy may pay a higher execution cost than the average. We calculate a perfect foresight realized half-spread to measure the maximum execution costs that an investor might pay. A supplier of immediacy who has perfect foresight would trade so as always to earn the subsequent price change. He would buy before the price goes up and sell before the price goes down. The perfect foresight measure is useful in the sense that it provides an upper bound on estimates of trading costs. It is also useful for assessing the potential costs of trading inside the quotes. Other measures give a zero cost of trading at the midquote, but this need

not be the case if the demander of immediacy trades with someone who can anticipate the subsequent price change.

The *perfect foresight realized half-spread* is defined as the absolute price change after a trade:

$$E|p_{t+\tau} - p_t|$$

where τ is 5 or 30 minutes. This is a measure of the maximum revenues that a market maker could earn and as such serves as an upper bound on the trading costs. Alternatively, this measure can be interpreted as the short-run volatility of prices.[16] Some authors, such as Hasbrouck (1993), have taken unexplained short-run variability of transaction prices as a measure of execution costs and the quality of a market. The perfect foresight price reversal overestimates maximum execution costs because some of the absolute price changes reflect the price volatility induced by new information, and this overestimate will be greater the longer the time period over which it is calculated.

The first two panels of table 11.7 present perfect foresight price reversals for all the trades over 5-minute and 30–minute horizons. In 1987, the average perfect foresight price reversal is 11.5 cents for the 5-minute horizon and 20.2 cents for a 30-minute horizon. The corresponding values for 1991 are 8.1 cents and 14.3 cents. The 5-minute results are less than the quoted half-spread reported in table 11.5. In other words, even a supplier of immediacy with perfect foresight does not earn the quoted half-spread over a 5-minute horizon. As expected, the perfect foresight reversal is substantially larger when calculated over the 30-minute horizon, but the likelihood that a supplier of immediacy would in fact be able to anticipate price changes over this longer horizon is less.

There has been some decline over time in the perfect foresight half-spread, but most of the apparent decline reflects the abnormally large values of the absolute price change in 1987. The perfect foresight reversal is somewhat larger for medium trades than for small trades, but, surprisingly, it tends to be smaller for large trades than for small trades. The process of prenegotiation of large trades and preblock trading seems to attenuate the block's posttrade price volatility.

The last three panels of table 11.7 provide data on the 5-minute perfect foresight spread of trades at the ask, the bid, and the midpoint.[17] A comparison of the three panels indicates that the perfect foresight half-spread is roughly the same for trades at the midquote as for trades at the quotes. For example, in 1991, the average absolute 5-minute price change after a midpoint trade was 7.9 cents, which compares with absolute price changes of 7.8 and 8.1 cents after trades at the ask and the bid, respectively. Measuring execution costs by the effective spread (as in Lee (1993) and Petersen and Fialkowski (1994), for example) assumes that trades at the midquote incur no trading costs, but this need not be the case if market makers can anticipate the near term price change. If market makers can anticipate near-

Table 11.7. Perfect foresight price reversal measure of execution costs on the NYSE in 343 S&P 500 stocks by trade size and year in the period 1987–1991 with October 1987 excluded. The figures are in dollars per share. A small trade size has 1,000 shares or less, a medium trade size has greater than 1,000 but less than 10,000 shares, and a large trade has 10,000 or more shares. Monthly averages are first calculated across all stocks from the average values for each stock, and the figures below represent means of the monthly averages. All the estimates are significantly different from zero using the *t*-test.

	Trade Size			
	All	Small	Medium	Large
5-minute $\lvert\Delta P\rvert$, all trades				
1987	0.115	0.114	0.119	0.106
1988	0.087	0.085	0.092	0.082
1989	0.081	0.079	0.085	0.074
1990	0.084	0.082	0.088	0.078
1991	0.081	0.080	0.085	0.074
30-minute $\lvert\Delta P\rvert$, all trades				
1987	0.202	0.199	0.210	0.195
1988	0.146	0.144	0.152	0.146
1989	0.138	0.135	0.145	0.136
1990	0.146	0.144	0.151	0.143
1991	0.143	0.141	0.150	0.143
5-minute $\lvert\Delta P\rvert$, trades at bid				
1987	0.120	0.119	0.122	0.107
1988	0.087	0.086	0.091	0.080
1989	0.079	0.078	0.083	0.074
1990	0.080	0.079	0.085	0.076
1991	0.081	0.081	0.085	0.075
5-minute $\lvert\Delta P\rvert$, trades at ask				
1987	0.115	0.113	0.117	0.104
1988	0.088	0.087	0.093	0.082
1989	0.081	0.079	0.085	0.073
1990	0.081	0.080	0.086	0.075
1991	0.078	0.077	0.083	0.073
5-minute $\lvert\Delta P\rvert$, trades at midpoint				
1987	0.108	0.107	0.111	0.100
1988	0.082	0.082	0.084	0.075
1989	0.078	0.077	0.080	0.070
1990	0.083	0.083	0.085	0.072
1991	0.079	0.078	0.080	0.072

term price changes, they can earn as much from trades at the midquote as from trades at the bid or ask. We thought that posttrade volatility would be greater for trades at the bid or ask than for trades inside the quotes, but this does not appear to be the case.

8. CORRELATION AMONG EXECUTION COST MEASURES

The average execution cost measures do not tell us how much the various measures covary across stocks. If the measures are highly correlated, they can be assumed to reflect the same phenomena. If not, they can be assumed to reflect different aspects of execution costs. However, caution must be used in interpreting the correlations between pairs of execution cost measures since some costs, such as bid–ask spreads, tend to have low variation by construction. Moreover, the execution costs have different temporal characteristics. Bid–ask spreads and effective half-spreads are contemporaneous measures. Roll's half-spread is measured in transaction time. Realized and perfect foresight price reversals are measured in calendar time.

We examine the correlations of execution cost measures after filtering out the effects of price level and market activity on the measures. Let y_i be the monthly average execution cost measure for stock i in a particular year. We first estimate the following cross-sectional regression for each year:

$$y_i = a_0 + a_1 p_i + a_2 p_i^2 + a_3 V_i + e_i \tag{9}$$

where p_i is the average of monthly stock prices for stock i, and V_i is the natural logarithm of the average of monthly NYSE share volume of stock i. The price-squared variable is used as an explanatory variable to capture nonlinearities in the execution cost and price relation. We then examine the correlations between the residuals from the regressions above for transactions on the NYSE in 1991. These correlations of adjusted execution cost measures avoid correlations that could be attributed to differences in volume and price.

The results reported in table 11.8 show that the contemporaneous measures—the half-spread and the effective half-spread—are positively correlated with each other. The correlation between the adjusted half-spread and the adjusted effective half-spread is 0.593. The high correlation contrasts with those obtained by Petersen and Fialkowski (1994). The perfect foresight spread is also correlated with the effective and quoted spreads because more volatile stocks with larger absolute price changes also tend to have larger spreads. The prospective measures—the implied spread and the realized spread—are also positively correlated with each other. For example, the correlation between the adjusted implied spread and the adjusted realized spreads at the bid and the ask exceed 0.58.

In contrast, the correlation between the contemporaneous and the prospective measures are insignificantly different from zero or are negative. Specifically, negative correlations between the two sets of adjusted

Table 11.8. Correlations between execution costs. The table shows the correlations between various adjusted execution costs for all trade sizes in 1991. The adjusted execution costs are residuals from regressing costs on the natural logarithm of NYSE total share volume, NYSE trade price, and square of NYSE trade price. S denotes the quoted spread, Z denotes half the effective spread, COV denotes the covariance of price changes, Abs5 (Abs30) denotes perfect foresight 5-minute (30-minute) price change, Bid5 (Ask5) denotes the 5-minute price reversals at the bid (ask), and Bid30 (Ask30) denotes 30–minute price reversals at the bid (ask). The sample consists of 343 S&P 500 stocks that were continuously listed on the NYSE between 1987 and 1991 and excludes data from October 1987. An asterisk denotes insignificance at the 5% level.

	S	Z	Abs5	Abs30	Bid5	Ask5	Bid30	Ask30
S								
Z	0.593							
Abs5	0.685	0.567						
Abs30	0.592	0.399	0.905					
Bid5	–0.140	–0.272	–0.279	–0.355				
Ask5	–0.250	–0.356	–0.436	–0.511	0.498			
Bid30	–0.123	–0.207	–0.242	–0.303	0.798	0.437		
Ask30	–0.288	–0.422	–0.417	–0.482	0.448	0.858	0.313	
Cov	–0.219	–0.080*	–0.362	–0.365	0.587	0.590	0.513	0.483

measures are unduly affected by a few outliers, and removing the outliers produces correlations close to zero. The lack of positive correlation between the contemporaneous and the prospective measures indicates that they measure different aspects of execution costs.

9. SUMMARY

This study estimates the average per share revenues of immediacy suppliers for 343 NYSE stocks continuously listed in the period 1987–1991 at about 2 to 3 cents per share. The estimate, which we term the realized half-spread, is simply the average price increase after a trade at the bid, or the negative of the average price decrease after a trade at the ask, estimated over horizons of 5 and 30 minutes. We find that the realized spread, albeit small, is consistent with data from an entirely different source—FOCUS reports filed by securities firms with the Securities and Exchange Commission.

The data on securities firms' revenues and trading volume also make possible inferences as to the revenues of limit orders. Our inferences suggest that limit orders are "picked off," a result consistent with an equilibrium in which public investors choose limit orders and market orders so as to equalize their costs on the margin. Securities firms earn greater revenues per share than the average realized half-spread, and public limit orders earn less than the average. Comparisons with the realized spread on

the NASDAQ market indicate that suppliers of immediacy in that market earned substantially more per trade than on the NYSE. This is because limit orders are not allowed to provide immediacy on NASDAQ and because NASDAQ dealers earned a larger reversal than NYSE firms.

The quoted half-spread [(ask – bid)/2] in 1991 averaged 11.1 cents, and the effective half-spread [absolute value of (trade price – midpoint)] averaged 6.7 cents. Two factors explain the difference between the quoted and effective spreads. First, some quotes attract no trades. Second, many trades take place at the quote midpoint. The realized half-spread is less than the effective half-spread because prices move against suppliers of immediacy (the adverse information effect). We also predict on the basis of Huang and Stoll (1997) that Roll's implied half-spread is less than the effective half-spread but greater than the realized half-spread, and the data support this prediction.

Finally, we calculate the average absolute price change after a trade, a measure we term the perfect foresight half-spread. This measure provides an upper bound on the revenues of immediacy suppliers since only a trader with perfect foresight could consistently predict the direction of the price change. In 1991, the average perfect foresight half-spread, calculated over a 5-minute horizon, was 8.1 cents, less than the quoted half-spread. Surprisingly, the perfect foresight spread is approximately the same for trades at the bid, at the ask, and at the midpoint, which suggests that knowledgeable suppliers of immediacy are able to earn revenues from trades at the midpoint, contrary to the assumption underlying the effective half-spread that such revenues are zero.

ACKNOWLEDGMENTS This study was funded in part by the Financial Markets Research Center at Vanderbilt University with a grant from the NASD and by the Dean's Fund for Faculty Research. A preliminary version of this chapter was presented at a Conference on Global Competition in the Market for Markets in Honor of Kal Cohen at Duke University. We thank Jennifer Conrad and Dennis Sheehan for their comments. Earlier versions of this chapter were titled "Anatomy of Trading Costs: Evidence from the NYSE."

NOTES

1. Hasbrouck, Sofianos, and Sosebee (1993) provide an excellent description of the institutional arrangements that generate the data used in this and other studies based on NYSE transaction data.

2. Quotes that are BBO (best bid or offer)-eligible are used to calculate the inside quote across markets. Certain quotes are not used for this purpose (for example, preopening quote indications or quotes that are not firm). Our ISSM data tapes also exclude autoquotes of the regional exchanges. An autoquote provides no new information since it is simply calculated by adding (subtracting) 1/8 to (from) the NYSE ask (bid) whenever the NYSE quote changes.

3. Lee and Ready (1991) recommend this rule.

4. The price reversal as a measure of market impact was first used by Scholes (1972). See also Kraus and Stoll (1972).

5. Instead, we later report the results of perfect foresight price reversal, defined as the absolute price change, which can be calculated for trades at the midpoint and compared to trades at the quotes.

6. By requiring a subsequent trade to be within a 5-minute interval, we reduce the sample of eligible transactions, especially for low-volume stocks. However, we standardize on the calendar time over which reversals are measured. The smaller sample does not introduce bias as long as the average price change we are unable to observe is not systematically different from the average price change we do observe. We see no reason why this isn't the case. Bias might be introduced if the procedure forces us to eliminate stock-months (our basic unit of observation) from our analysis. However, no stock-month was eliminated by this procedure. We also calculated reversals on the basis of the first trade after 5 minutes or 30 minutes, which produced the same results.

7. From NYSE Fact Book.

8. See Stoll (1994) for details.

9. An alternative approach to the per share trading gains of firms (and limit orders) is to estimate directly the realized half-spread after trades by firms (or limit orders). Unfortunately, this is usually impossible because identification of the parties to a trade is not known. However, limit orders can be identified in the TORQ data, and Simpson (1994) and Harris and Hasbrouck (1996) provide some estimates of price reversals after limit order trades.

10. See Huang and Stoll (1996b), table 3.

11. Demsetz (1995) has pointed out the fact that revenues of specialists could be the same as revenues of NASDAQ dealers, but this is unlikely from our data. The specialist would be required to earn a reversal of more than the quoted half-spread.

12. Perold recommends a measure that is broader than the quoted spread, namely the difference between the trade price and the price of the stock when the decision to trade was reached. In other words, if an investor decides to sell the stock at 9 A.M. when its price is \$30 and actually sells the stock at 11 A.M. at a price of \$29, the cost is \$1. Keim and Madhavan (1996) adopt this approach to measuring execution cost. The Perold approach includes the quoted half-spread as a component, but it also includes any price changes between the decision point and the trade point.

13. See Huang and Stoll (1994, 1997) for the response of quotes to trades and for estimates of the adverse information component (and other components) of the spread.

14. We use the fact that $E(Q_t \mid Q_{t-1}) = (1 - 2\rho)Q_{t-1}$.

15. Harris (1990) describes the Jensen inequality bias and biases arising from small samples, both of which bias downward the estimated serial covariance.

16. One can also think of the perfect foresight price reversal as a measure of price continuity since it gives the average price change over 5 and 30 minutes. The NYSE calculates the proportion of adjacent trades that are within a given range relative to the last price.

17. Only the 5-minute results are presented in table 11.6; the 30-minute results yield qualitatively similar inferences.

REFERENCES

Arnott, Robert D. and Wayne H. Wagner. 1990. "The Measurement and Control of Trading Costs," *Financial Analysts Journal* 46 (November/December), 73–80.

Bagehot, Walter (pseudonym for Jack Treynor). 1971. "The Only Game in Town," *Financial Analysts Journal* 27 (March/April), 31–53.

Beebower, Gilbert, Vasant Kamath, and Ronald Surz. 1985. "Commission and Transaction Costs of Stock Market Trading," working paper, SEI Corporation (July), 28 pages.

Berkowitz, Stephen, Dennis Logue, and Eugene Noser. 1988. "The Total Cost of Transactions on the NYSE," *Journal of Finance* 43 (March), 97–112.

Black, Fischer. 1971a. "Toward a Fully Automated Exchange I," *Financial Analysts Journal* 27 (July/August), 28–35, 44.

Black, Fischer. 1971b. "Toward a Fully Automated Exchange II," *Financial Analysts Journal* 27 (November/December), 24–28, 86–87.

Black, Fischer. 1995. "Equilibrium Exchanges," *Financial Analysts Journal* 51 (May/June), 23–29.

Blume, Marshall and Michael Goldstein. 1992. "Displayed and Effective Spreads by Market," working paper 27–92, Rodney White Center, The Wharton School, University of Pennsylvania (December 23).

Bodurtha, Stephen and Thomas Quinn. 1990. "Does Patient Program Trading Really Pay?" *Financial Analysts Journal* 46 (May/June), 35–42.

Chan, Louis K. C. and Josef Lakonishok. 1993. "Institutional Trades and Intra-day Stock Price Behavior," *Journal of Financial Economics* 33, 173–201.

Christie, William G. and Roger D. Huang. 1994. "Market Structures and Liquidity: A Transactions Data Study of Exchange Listings," *Journal of Financial Intermediation*, 3, 300–326.

Cohen, Kalman and Robert Conroy. 1990. "An Empirical Study of the Effect of Rule 19c-3," *Journal of Law and Economics* 33 (April), 277–305.

Copeland, Tom and Dan Galai. 1983. "Information Effects of the Bid–Ask Spread," *Journal of Finance* 38, 1457–1469.

Demsetz, Harold. 1968. "The Cost of Transacting." *Quarterly Journal of Economics* 82 (February), 33–53.

Demsetz, Harold. 1995. "A Non-collusive Explanation for the Pattern of Spreads on the NASDAQ," Working paper, UCLA Department of Economics (October 19), presented at Shadow SEC Meeting of November 5, 1995.

Glosten, Lawrence R. and Paul Milgrom. 1985. "Bid, Ask and Transaction Prices in a Specialist Market with Heterogeneously Informed Traders," *Journal of Financial Economics* 14 (March), 71–100.

Harris, Lawrence. 1990. "Statistical Properties of the Roll Serial Covariance Bid/Ask Spread Estimator," *Journal of Finance* 45, 579–590.

Harris, Lawrence and Joel Hasbrouck. 1996. "Market vs. Limit Orders: The Super Dot Evidence on Order Submission Strategy," *Journal of Financial and Quantitative Analysis* 31, 213–232.

Hasbrouck, Joel. 1993. "Assessing the Quality of a Security Market: A New Approach to Transaction-Cost Measurement," *The Review of Financial Studies* 6, 192–212.

Hasbrouck, Joel, George Sofianos, and Deborah Sosebee. 1993. "New York Stock Exchange Systems and Trading Procedures," NYSE working paper #93-01 (April 27).

Holthausen, Robert, Richard Leftwich, and David Mayers. 1987. "The Effect of Large Block Transactions on Security Prices: A Cross-sectional Analysis," *Journal of Financial Economics* 19 (December), 237–267.

Huang, Roger D. and Hans R. Stoll, 1994. "Market Microstructure and Stock Return Predictions," *The Review of Financial Studies* 7 (Spring), 179–213.

Huang, Roger D. and Hans R. Stoll. 1996a. "Competitive Trading of NYSE Listed Stocks: Measurement and Interpretation of Trading Costs," *Financial Markets, Institutions & Instruments* 5, 54 pages.

Huang, Roger D. and Hans R. Stoll. 1996b. "Dealer versus Auction Markets: A Paired Comparison of Execution Costs on NASDAQ and the NYSE," *Journal of Financial Economics* 41 (July), 313–357.

Huang, Roger D. and Hans R. Stoll. 1997. "The Components of the Bid–Ask Spread: A General Approach," *The Review of Financial Studies* 10, 995–1034.

Keim, Donald and Ananth Madhavan. 1996. "The Upstairs Market for Large-Block Transactions: Analysis and Measurement of Price Effects," *The Review of Financial Studies* 9 (Winter), 1–36.

Kraus, Alan and Hans Stoll. 1972. "Price Impacts of Block Trading on the New York Stock Exchange," *Journal of Finance* 27 (June), 569–588.
Lee, Charles. 1993. "Market Integration and Price Execution for NYSE Listed Securities," *Journal of Finance* 48 (July), 1009–1038.
Lee, Charles and Mark Ready. 1991. "Inferring Trade Direction from Intraday Data," *Journal of Finance* 46 (June), 733–746.
McInish, T. and R. Wood. 1992. "Hidden Limit Orders on the NYSE," Memphis State University, Department of Finance, working paper (October).
Perold, Andre F. 1988. "The Implementation Shortfall: Paper versus Reality," *Journal of Portfolio Management* 14 (Spring), 4–9.
Perold, Andre and Erik Sirri. 1993. "The Cost of International Equity Trading," Harvard Business School working paper (May).
Petersen, M. and D. Fialkowski. 1994. "Posted versus Effective Spreads: Good Prices or Bad Quotes?" *Journal of Financial Economics* 35 (June), 269–292.
Roll, Richard. 1984. "A Simple Implicit Measure of the Bid–Ask Spread in an Efficient Market," *Journal of Finance* 39 (September), 1127–1139.
Scholes, Myron. 1972. "The Market for Securities: Substitution versus Price Pressure and the Effects of Information on Share Price," *Journal of Business* 45 (April), 179–211.
Schwartz, Robert and David K. Whitcomb. 1988. "Transaction Costs and Institutional Investor Trading Strategies," *Monograph Series in Finance and Economics*, Monograph 1988-2/3, New York University, Leonard Stern School of Business, 72 pages.
Simpson, Rick. 1994. "The Winner's Curse of Supplying Immediacy: The Specialist versus Public Limit Orders," working paper, Vanderbilt University (August).
Sofianos, George. 1995. "Specialist Gross Trading Revenues at the New York Stock Exchange," NYSE working paper #95-01 (March 27).
Stoll, Hans R. 1979. "Regulation of Securities Markets: An Examination of the Effects of Increased Competition," *Monograph Series in Finance and Economics*, Monograph 1979-2, New York University, Graduate School of Business, 82 pages.
Stoll, Hans R. 1989. "Inferring the Components of the Bid–Ask Spread: Theory and Empirical Tests," *The Journal of Finance* 44 (March), 115–134.
Stoll, Hans R. 1995. "The Importance of Equity Trading Costs: Evidence from Securities Firms' Revenues," in Robert Schwartz (ed.), *Global Equity Markets: Technological, Competitive and Regulatory Challenges* (Irwin, Homewood, Ill.).
Wagner, Wayne and Mark Edwards. 1993. "Best Execution," *Financial Analysts Journal*, 49 (January/February), 65–71.

12

Black, Merton, and Scholes—Their Central Contributions to Economics

Darrell Duffie

1. WHY THEY WON THE PRIZE

I will briefly summarize the central contributions to economics of Fischer Black, Robert C. Merton, and Myron S. Scholes.

Of course, the contribution that first comes to mind is the Black–Scholes option-pricing formula, for which Robert Merton and Myron Scholes were awarded the Alfred Nobel Memorial Prize in Economic Sciences in 1997. I have no doubt that, because of his key role in that far-reaching formula, Fischer Black would have shared in that prize but for his untimely death. In this chapter, I will address the contributions of all three of these exceptional economists simultaneously, rather than giving separate treatment to Fischer Black. My goal is to give an objective and concise account of their path-breaking research and what it has offered to the theory and practice of economics.

2. SETTING THE STAGE

Finance is a large, richly interwoven, widely applied, and extremely active area of economics. One of the central issues within finance is the valuation of future cash flows. While there are important alternatives, a current basic paradigm for valuation, in both academia and in practice, is that of competitive market equilibrium: The price that will apply in the market is that price which, taken as given by market participants, equates total demand to total supply.

With 1997's award to Robert Merton and Myron Scholes, three fundamental contributions to the theory of financial valuation that are based on this paradigm of market equilibrium have now been closely linked to Nobel Memorial Prizes. These are:

1. the portion of the Modigliani–Miller (1958) theory that deals with the irrelevance of capital structure or dividend policy for the market value of a corporation;
2. the Capital Asset Pricing Model (CAPM) of William Sharpe[1] (1964);
3. the Black–Scholes (1973) option-pricing theory.

The first and third of these contributions rely on the notion of market equilibrium in only the weakest possible sense, known as "arbitrage reasoning." If, under their respective assumptions, the valuation formulas provided by these theories were not correct, then market participants would have an opportunity to create an "arbitrage"; that is, to trade securities so as to make unbounded profits with no initial investment and no subsequent risk of loss. In particular, if the market price of a financial security were lower than suggested by arbitrage reasoning, arbitrageurs would ask to buy it, and in unbounded quantities. Conversely, if the market price were higher than suggested by theory, arbitrageurs would want to sell, and the more the better. In such situations, markets could not clear, and equilibrium would be impossible. Such "arbitrages" are only prevented, in theory, when the proposed valuation formulas actually apply.

While there are some close precursors in the literature, Modigliani and Miller (1958) essentially established the modern foundation in finance for arbitrage-based valuation reasoning. The Black–Scholes theory provided an extremely powerful extension of arbitrage modeling to dynamic settings.

The assumptions of any model rarely (if ever) apply literally. What might be an arbitrage in theory is sometimes difficult to carry out in practice. For example, arbitrage-based valuation models often rely on the assumptions of perfect information and the absence of transaction costs. No-arbitrage arguments are so compelling, however, that financial economists encounter almost daily reference to the Modigliani–Miller and Black–Scholes theories as central points of departure for model building or reasoned discussion of financial problems.

As shall be discussed below, even though the CAPM does not rely on arbitrage reasoning, it also played a key role in the development of the Black–Scholes formula.

3. ARBITRAGE PRICING OF OPTIONS

Before getting to the focal point of our story, the Black–Scholes formula, it will be useful for readers that are newcomers to finance or unfamiliar with stochastic calculus to see the basic idea of arbitrage-based option pricing in the simplest possible setting. Ironically, this simple introductory model was only developed, by William F. Sharpe, *after* the advent of the Black–Scholes model.

Consider a financial security, say a traded stock, whose price today is 100 and whose price tomorrow will be either 102 or 98. Consider an option that grants its owner the right to purchase the stock tomorrow for 100. If the stock price tomorrow turns out to be 102, the owner of the option will (as we assume rationality and no transaction costs) exercise the right to buy for 100 and thereby benefit from exercising the option to the extent of a cash flow of $102 - 100 = 2$. If, on the other hand, the stock price turns out to be 98 tomorrow, the owner of the option will decline the opportunity to buy at 100, and the option has no cash flow in that event.

Suppose, to keep the numbers simple, that the overnight interest rate is zero. At another interest rate, the following arguments would apply with slightly different numbers.

We claim that, in the absence of arbitrage, the price of the option today is 1. How can one be so precise in the absence of any additional information? Is there no role in this for the risk preferences of market participants or the probabilities that they assign to the event that the stock price goes up? Let's delay an answer to these questions for now.

Before directly addressing the arbitrage valuation, let us first find the number a of shares of stock to buy and the amount b to borrow so that, whether the stock price goes up or down, the net proceeds of the stock portfolio with loan repayment are equal to the cash flow from owning one option. This means that a and b must solve

$$102a - b = 2 \tag{1}$$

$$98a - b = 0 \tag{2}$$

The solution is $a = 0.5$ and $b = 49$. The net initial cost of this option-replicating portfolio is $100a - b = 50 - 49 = 1$. It seems unlikely that brokers would quote a price other than 1 for the option if one can make a "synthetic" version of its cash flows for a cost of 1.

In order to substantiate this claim, a simple proof by contradiction will serve. Suppose that the option were actually trading at a price of $p > 1$. If this were true, an arbitrageur could sell the option for p and replicate, at an initial cost of 1, the option's future cash flows by purchasing 0.5 shares of stock and borrowing 49. The net payoff tomorrow of the replicating portfolio meets any cash flow demanded tomorrow by the purchaser of the option, whether the stock price goes up or down. The arbitrageur has netted an initial gain of $p - 1 > 0$ with no investment and no risk. This, however, is an arbitrage! So, we must have $p \leq 1$. If, however, $p < 1$, then buying the option and selling the option-replicating portfolio (that is, short-selling 0.5 shares of stock and lending 49) constitutes an arbitrage. Thus, $p = 1$ is necessary for the absence of arbitrage. One can easily check that it is sufficient for no arbitrage that $p = 1$.

Those new to this could test their understanding by solving the option valuation problem for a nonzero interest rate. (A simple daily interest rate below –2% or above 2% will not work. Why?)

Before we get to the actual Black–Scholes formula, let us revisit the role in this simple option-pricing example of the risk preferences of investors and the probabilities that they may assign to positive or negative stock returns. A naive objection might be: How could the initial option price be as little as 1, for example, if there is a 99% chance that the stock return will be positive (in which case the option pays 2) and there are investors that are relatively close to indifferent about bearing risk? The natural, and correct, reply is that preferences as well as beliefs about the likelihoods of two states *do* indeed play a role because they determine the initial price of the

stock and the interest rate. If, for example, an investor that is close to risk-neutral believes that it is virtually certain that the stock price will be 102 tomorrow, then the initial price of the stock must be close to 102 today, not 100. As the stock price and interest rate vary with the preferences and beliefs of investors, so will the option price.

4. THE BLACK–SCHOLES FORMULA

Now we are ready to see how the Black–Scholes formula works.

In the Black–Scholes model, the price of a security, say a stock, is assumed to be given at any time $t \geq 0$ by $X_t = x \exp(\alpha t + \sigma B_t)$, where $x > 0$, a and $\sigma > 0$ are constants, and B is a standard Brownian motion.[2] Riskless borrowing and lending is possible at the constant continuously compounding interest rate r. For future reference, x is the initial stock price, σ is referred to as the *volatility* of the stock, and, because $E(X_t) = e^{(\alpha + \sigma^2/2)t}$, we call $\mu \equiv \alpha + \sigma^2/2$ the expected rate of return on the stock.

Consider an option that grants its owner the right, but not the obligation, to buy the stock at a given exercise date T and at a given exercise price K. Trading is permitted at arbitrary frequency, and there are no transaction costs. The information available to investors at any time t is the history of the stock price up to that time. Certain minor technical assumptions apply.

Now, at what price will the option be sold, assuming that there are no arbitrages? For purposes of future reference, let

$$C(x, r, \mu, T, \sigma, K) = e^{-rT}E[\max(X_T - K, 0)] \tag{3}$$

denote the expected discounted payoff of the option for given parameters $(x, r, \mu, T, \sigma, K)$. This is *not*, in general, the price of the option. Sprenkle (1961), in effect, showed that

$$C(x, r, \mu, T, \sigma, K) = xe^{(\mu - r)T}N\big(d(x, \mu, T)\big) - e^{-rT}KN\big(d(x, \mu, T) - \sigma\sqrt{T}\ \big) \tag{4}$$

where $N(\cdot)$ is the cumulative standard normal distribution function and, for any (x, y, T),

$$d(x, y, T) = \frac{\log(x/K) + \big(y + \sigma^2/2\big)T}{\sigma\sqrt{T}}$$

Stochastic calculus[3] can be used to show the following fact: One can invest a total of $C(x, r, r, T, \sigma, K)$ at time 0 and, at each time t between 0 and T, hold $N(d(X_t, r, T - t))$ shares of the stock, always borrowing or lending cash flows as necessary to finance the position between 0 and T, and be left at time T with a position in cash and stock whose market value is exactly $\max(X_T - K, 0)$, the payoff of the option. (This is analogous to the replication strategy shown in section 3.)

From the definition (3) of $C(\cdot)$, this initial cost $C(x, r, r, T, \sigma, K)$ of replicating the option, called the Black–Scholes option-pricing formula, would

be the expected discounted payoff of the option, *if the mean rate of the return of the stock were the riskless rate r*.

Using the same logic as in section 3, assuming that there are no arbitrages, the option must trade in the market for its initial replication cost, $C(x, r, r, T, \sigma, K)$. If the option were selling for some amount p strictly larger than $C(x, r, r, T, \sigma, K)$, then one could sell the option, invest in the replicating strategy, and take away an initial riskless profit of $p - C(x, r, r, T, \sigma, K)$. The net cash flow at expiration is zero since the payoff of the replicating strategy precisely covers the claim against the option. This would be an arbitrage. Conversely, if $p < C(x, r, r, T, \sigma, K)$, the opposite strategy of buying the option and selling the replicating strategy is an arbitrage. Indeed, then, the arbitrage-free price of the option is $C(x, r, r, T, \sigma, K)$, which is familiar from (4) as the Black–Scholes option-pricing formula.

As far as the sufficiency of the Black–Scholes formula for the absence of arbitrage, one must place only some reasonable limits on the class of allowable trading strategies. For example, as shown by Dybvig and Huang (1988), it is enough to insist that an investor should not be given unlimited credit.

5. HISTORY OF THE BLACK–SCHOLES FORMULA

The best two available written sources on the history of the development of the formula are Black (1989) and Bernstein (1992), the latter being based on extensive interviews of those involved. The accounts given in these two sources are consistent with each other, with other published sources including the published form of the original paper by Black and Scholes (1973) presenting the formula, and with what has been told to me anecdotally. The story goes roughly as follows.

Fischer Black, one of many who had looked at this problem,[4] began with the idea of applying the Capital Asset Pricing Model at each instant of time for investments over an infinitesimally small period of time. This allowed him to derive a partial differential equation (PDE) for the option price $c(x, t)$ that would apply at any time $t < T$ and at any stock price x for that time. This now famous PDE is

$$c_t(x,t) + c_x(x,t)rx + \frac{1}{2}c_{xx}(x,t)\sigma^2x^2 - rc(x,t) = 0 \tag{5}$$

where subscripts indicate partial derivatives in the customary way, with the obvious boundary condition

$$c(x, T) = \max(x - K, 0) \tag{6}$$

Black had found this PDE by 1969 or earlier but could not initially solve it. He did note, however, that the solution could not allow any role for the coefficient μ, the expected rate of return on the stock! With this in mind, Black and Scholes teamed up at MIT. They noted that since the PDE did not involve μ, any expected return for the stock would generate the same

option price, including the riskless rate of return r. Then they noted that if the stock could be treated as having a riskless rate of return, then, by applying the CAPM instant-by-instant, so could the option because (under the assumption that the option price is a smooth function of the stock price) changes in the option and stock prices over infinitesimal periods of time are perfectly correlated. Using Sprenkle's calculation (4), this would imply the explicit option valuation $C(x, r, r, T, \sigma, K)$, the discounted expected payoff of the option that would apply if the stock had the riskless expected rate of return r and if the option payoff could be discounted at a risk-free rate. Sure enough, this solution satisfied the PDE (5).

As Black (1989) put it, referring to himself and Scholes, "We had our option formula." He continued, "As we worked on the paper, we had long discussions with Robert Merton, who was also working on option valuation. Merton made a number of suggestions that improved our paper. In particular, he pointed out that if you assume continuous trading in the option or stock, you can maintain a hedged position between them that is literally riskless. In the final version of the paper, we derived the formula that way, because it seemed to be the most general derivation." This generous acknowledgment of Merton's contribution to the more general derivation, indeed the derivation that truly revolutionized modern financial theory, is consistent with the acknowledgment of Merton's contribution given in Black and Scholes (1973) (in their footnote numbered 3). To be precise, there is no need to rely, as Black and Scholes had originally, on market equilibrium under the strong assumptions of the CAPM. Instead, the simple assumption of no arbitrage would suffice. Merton's no-arbitrage argument appears, along with the CAPM-based argument, in the finally[5] published form of Black and Scholes (1973).

In yet another source,[6] Merton's contribution is acknowledged, with Black's statement that "A key part of the option paper that I wrote with Myron Scholes was the arbitrage argument for deriving the formula. Bob gave us that argument. It should probably be called the Black–Merton–Scholes paper."

Merton went on to write his 1973 paper, "Rational Option Pricing," another landmark contribution that elaborated on the Black–Scholes approach to option valuation in many ways. Merton generously attempted to delay the publication of his own paper until the earlier paper of Black and Scholes, after surprising resistance from journal editors, could finally be published (apparently with the help of Merton Miller and Gene Fama) in *The Journal of Political Economy* in 1973. Later, Merton (1977) derived a more theoretically sound and concrete version of his replication argument, based on actual trading strategies rather than the more ephemeral notion of returns over "infinitesimal periods." The argument sketched out in section 2 is essentially that of Merton (1977), and is now the standard derivation.

There is an additional important contribution in Black and Scholes (1973). They observed that because of limited liability, the equity of a corporation may itself be treated as an option on the total asset value of the firm and

thereby priced by the same methodology. This observation is at the core of modern corporate finance and was apparently made independently by Merton (1973a). Because Merton, having overslept, missed a presentation of this idea by Black and Scholes, neither team was aware of the other's progress on this problem.

6. THE SIGNIFICANCE OF BLACK–SCHOLES TODAY

The option-pricing methods of Black, Merton, and Scholes are now being taught to almost every MBA student and to most graduate, and many undergraduate, students in economics. Many investors and major corporations use these methods for planning, purchasing, pricing, or accounting purposes. In addition to valuing straight put and call options, corporations use Black–Scholes modeling to value executive stock compensation plans, real production options, warrants, convertible securities, debt, and so on. (In fact, for many of these applications, the methods are sometimes applied inappropriately.)

Before the advent of Black–Scholes, option markets were sparse and thinly traded. Now they are among the largest and most active security markets. The change is attributed by many to the Black–Scholes model since it provides a benchmark for valuation and (via the arbitrage argument) a method for replicating or hedging options positions. One now can buy options on most of the major exchange-traded commodities, foreign currencies, stock indices, and government bonds. None of these markets existed in any active form before 1973. Over-the-counter options can be obtained from major investment banks on almost any important index, even if there is no commodity or security underlying the index.

The Black–Scholes approach has been extended to a wide variety of instruments with embedded options such as caps, floors, collars, collateralized mortgage obligations, knockout options, swaptions, lookback options, barrier options, compound options, and the list goes on and on. Indeed, there is nothing that restricts the approach to options, as opposed to other contingent claims. For example, the same arguments used in the previous section apply for a contingent claim paying $g(X_T)$ at time T for any function g:$[0, \infty) \to \mathbb{R}$ satisfying technical conditions. One merely substitutes the PDE boundary condition (6) with $c(x, T) = g(x)$. In many cases, of course, the contingent claim's price cannot be computed explicitly, and it is now standard operating procedure to use such numerical techniques as finite-difference solution of the associated PDE or Monte Carlo integration of the associated "risk-neutral" expectation. Almost every major bank and trading firm has a team of specialists that use advanced Black–Scholes methods. All of these methods have their genesis in the work of Black, Merton, and Scholes.

Aside from their use in pricing, the methods developed by Black, Merton, and Scholes are widely applied to financial risk management. The idea that the option can be priced by finding a trading strategy that replicates its

payoff is frequently used to hedge a given security, or even to hedge a given cash flow that is not traded as a security. If one is to receive an untraded option payoff, for example, the risk inherent in that payoff can be eliminated by selling the replicating strategy previously described. This converts the risky cash flow at expiration into an initial cash flow. Indeed, the ability to hedge the value of an option on the entire S&P 500 portfolio in this manner, under the rubric of "portfolio insurance," was accused by some of having contributed significantly to the stock market crash of 1987. Investment banks routinely sell securities with embedded options of essentially any variety requested by their customers and then cover the combined risk associated with their net position by adopting dynamic hedging strategies.

The approach of Black, Merton, and Scholes also allows one to use market option price quotations as a gauge of market volatility. For example, the Black–Scholes formula $C(x, r, r, T, \sigma, K)$ can be inverted to recover the volatility parameter σ implied by the option price. The fact that volatility is not actually constant, but rather varies over time with uncertainty, has instigated a new generation of option-pricing formulas allowing "stochastic volatility" but based on the same Black–Scholes approach.

In addition to these important practical applications (pricing, synthesis of untraded cash flows, hedging, and information discovery), the option-pricing work of Black, Scholes, and Merton has led to important theoretical work on optimal portfolio choice and multiperiod equilibrium in financial markets. The key to this work is the observation that, in a Black–Scholes setting, one can replicate not only the opton payoff but any stream of cash flows that depends on the path taken by the stock price. The required initial investment is the expected discounted cash flow after replacing the expected return of the stock with the riskless rate of return. Harrison and Kreps (1979) later obtained an essentially definitive extension of the Black–Scholes model and the general notion of "risk-neutral valuation," following in part on ideas appearing in Cox and Ross (1976). The Harrison–Kreps generalization of the Black–Scholes modeling approach allowed Cox and Huang (1989) to give important extensions of Merton's (1971) model of optimal portfolio choice in a multiperiod setting.[7] The idea is that the dynamic program that Merton solved can be replaced with a static calculus-of-variations problem. One merely replaces the complicated dynamic budget constraint with a static constraint that the expected discounted consumption payoff of the investor (after replacing the expected rate of return on all securities with the riskless rate) must be equal to the initial wealth of the investor. This approach applies not only in Merton's setting but in significantly more general settings.

The replication arguments used to derive the Black–Scholes formula have also been applied to general equilibrium modeling in multiperiod financial markets. Arrow (1953) showed that security markets are an efficient method for allocating risk because they allow one to replace a complete set of contingent claims markets with a sparser set of financial security markets. The payoff of any contingent claim can be replicated by a portfolio of basis

securities. His one-period model has been extended in a series of papers by various authors to multiperiod settings by using the dynamic replication arguments used to prove the Black–Scholes formula. Literally, an infinite-dimensional space of possible consumption streams that investors might wish to obtain can be synthesized by trading a small number of securities. This allows one to convert a complicated multiperiod general equilibrium problem into a single-period problem.

With some additional concepts, many of these ideas apply even if there are not enough securities to replicate every possible consumption stream, a situation known as "incomplete markets."

7. OTHER MAJOR CONTRIBUTIONS

As indicated in part by the attached list of publications, Black, Merton, and Scholes are responsible for a tremendously large and important body of ideas and papers going well beyond the Black–Scholes formula. I have listed below only those that I think of as extremely important to the development of financial markets or theory. Even without any of these additional contributions, the discovery of the Black–Scholes formula and the method by which it was derived constitute an exceptionally strong justification for the award of the Nobel Memorial Prize in Economic Sciences.

1. Black (1972) developed an extension of the Capital Asset Pricing Model that applies without the existence of a riskless security. The new "zero-beta" model replaces the riskless rate of return in the famous "beta formula" with the expected rate of return on a portfolio uncorrelated with the market portfolio.
2. Black (1976) examined the pricing of commodity contracts and, in particular, extended the Black–Scholes model to the case of options on futures or forwards.
3. Black, Derman, and Toy (1990) developed a model of the valuation of term-structure securities (those whose payoffs depend on the history of the term structure of interest rates) that is now an industry standard. The model is in broad spirit much like the Black–Scholes model, in its binomial form developed by Cox, Ross, and Rubinstein (1979), and has important computational advantages in everyday work on "Wall Street." An important aspect of the model is the fact that it is constructed so that, in principle, its parameters can be computed from the current term structure and from the current prices of options on Treasury bonds, much in the way that the volatility parameter of the Black–Scholes model can be computed from the option price.
4. Merton (1969, 1971) found a path-breaking method of solving the problem of optimal consumption and portfolio choice in a continuous-time setting. His method involved reduction of the problem to a partial differential (Hamilton–Jacobi–Bellman)

equation for the investor's indirect utility function for wealth. This formulation, a breakthrough in its own right, may well have influenced the way that Black and Scholes approached the option-pricing problem, in choosing a continuous setting with the same stock price model assumed by Merton, and reducing the valuation equation to a PDE. To this day, there is a virtual industry of researchers extending Merton's model in many different ways. Merton's model is also widely referred to among specialists working in mathematics as the best and most elegant textbook example of a stochastic control problem.

5. Merton (1973b) offered the first major extension of equilibrium capital asset pricing theory to a multiperiod setting. By taking the approach used in his 1971 paper on optimal investment and consumption behavior, and allowing for a multivariate Markov state process for the market environment, Merton was able to show how the equilibrium expected rate of return of a given security depends not only on the covariance of the return with that of the market portfolio (as in the one-period CAPM) but also on the covariance of the return with changes in the state variables of the economy. This is the essence of the dynamic equilibrium problem: Investors are concerned not only with their wealth in the next period but also with how their opportunities to generate wealth in much later periods will depend on state variables in the next period. Breeden (1979), based in part on work by Rubinstein (1976), was later able to reduce Merton's solution to an elegant formula showing that the multiperiod CAPM is in fact the same as the Sharpe–Lintner CAPM once one substitutes covariance between returns and aggregate consumption for covariance between returns and aggregate wealth (the payoff of the market portfolio). (Of course, aggregate consumption and aggregate wealth are the same in the one-period setting of Sharpe–Lintner.)
6. Among Merton's most important extensions of the Black–Scholes formula are: (i) his work on American options and on options on stocks paying dividends, among many other applications, in Merton (1973a); (ii) his extension to the case of discontinuous stock price processes in Merton (1976), which is important also for showing that the model would not in the future be confined to the setting of Brownian motion; and (iii) the conversion in Merton (1977) of the original Black–Scholes–Merton no-arbitrage pricing argument from one based on instantaneous returns to one based on dynamic replicating strategies.
7. Scholes did important work on dividends and their impact on the valuation of common stock in Black and Scholes (1974), Miller and Scholes (1978), and Miller and Scholes (1982].
8. Scholes and Williams (1977), in a study of how to estimate betas (in the sense of the CAPM) from nonsynchronous data, provided

an important and widely cited (and taught) contribution to empirical methods in finance.

9. Scholes is one of the leading experts on employee stock compensation plans. His textbook *Taxes and Business Strategy* (1992), coauthored with Mark Wolfson, is the first of its kind in a critical and understudied area of finance.

ACKNOWLEDGMENTS My deep admiration of Fischer Black, Bob Merton, and Myron Scholes has been for qualities going well beyond those evident in their exceptional research contributions described in this chapter. I am grateful to Linda Bethel and Aileen Lee for technical assistance.

NOTES

1. The work of John Lintner (1965) is also often cited.

2. Underlying the model is a probability space. On this space, $B_0 = 0$, $Bt - B_s$ is normally distributed with zero expectation and variance $|t - s|$, and the increments of B are independently distributed. Except for technicalities that can be found, for example, in Karatzas and Shreve (1988), these properties define a standard Brownian motion. It is noteworthy that Brownian motion was first given an effective definition by Bachelier (1900) in a study of security price behavior that included an option-pricing formula not entirely unlike that of Black and Scholes.

3. Karatzas and Shreve (1988) offer a good textbook treatment of stochastic calculus.

4. Those who had attacked some version of the problem prior to Black and Scholes included Bachelier (1900), Sprenkle (1961), Samuelson (1965), and Samuelson and Merton (1969), who, in effect, derived the reservation price of an investor with a particular utility function.

5. Because of the difficulty that Black and Scholes had in getting their original paper (1973) published, their second paper on this topic (1972) actually appeared in print before the first.

6. See *MIT Management*, 1988, Fall, p. 28. This quote came to my attention in Bernstein (1992, p. 223).

7. See also Karatzas, Lehoczky, and Shreve (1987).

REFERENCES

Arrow, K.: Le rôle des valeurs boursier pour la repartition la meilleure des risques. *Econométrie Colloques Internationaux du C.N.R.S. 40*, 41–47, Paris, 1952; discussion, 47–48, C.N.R.S., Paris, 1953; English translation: *Review of Economic Studies 31*, 91–96, 1964.

Bachelier, L.: Théorie de la speculation. *Annales Scientifiques de l'Ecole Normale Supérieure*, troisième série *17*, 21–88, 1900; translation in *The Random Character of Stock Market Prices*, Paul Cootner (ed.), MIT Press, Cambridge, Mass., 1967.

Bemstein, P.: *Capital Ideas: The Improbable Origins of Modern Wall Street*. Free Press, New York, 1992.

Black, F.: Capital market equilibrium with restricted borrowing. *Journal of Business 45*, 444–454, 1972.

Black, F.: The pricing of commodity contracts. *Journal of Financial Economics 3*, 167–179, 1976.

Black, F.: How we came up with the option formula. *Journal of Portfolio Management 15* (2), 4–8, 1989.

Black, F., Derman, E., and Toy, W.: A one-factor model of interest rates and its application to treasury bond options. *Financial Analysts Journal 7*, 33–39, 1990.

Black, F., and Scholes M. S.: The valuation of options contracts and a test of market efficiency. *Journal of Finance 27* (*May*), 399–417, 1972.

Black, F., and Scholes M. S.: The pricing of options and corporate liabilities. *Journal of Political Economy 81* (*May/June*), pp. 637–654, 1973.

Black, F., and Scholes M. S.: The effects of dividend yield and dividend policy on common stock prices and returns. *Journal of Financial Economics 1* (*May*), pp. 1–22, 1974.

Breeden, D.: An intertemporal asset pricing model with stochastic consumption and investment opportunities. *Journal of Financial Economics 7*, 265–296, 1979.

Cox, J. and Huang, C.-F.: Optimal consumption and portfolio policies when asset prices follow a diffusion process. *Journal of Economic Theory 49*, 33–83, 1989.

Cox, J. and Ross, S.: The valuation of options for alternative stochastic processes. *Journal of Financial Economics 3*, 145–166, 1976.

Cox, J., Ross, S., and Rubinstein, M.: Option pricing: A simplified approach. *Journal of Financial Economics 7*, 229–263, 1979.

Dybvig, P. and Huang, C.-F.: Non-negative wealth, absence of arbitrage, and feasible consumption plans. *Review of Financial Studies 1*, 377–401, 1988.

Harrison, J. M. and Kreps, D.: Martingales and arbitrage in multiperiod securities markets. *Journal of Economic Theory 20*, 381–408, 1979.

Karatzas, I., Lehoczky, J., and Shreve, S.: Optimal portfolio and consumption decisions for a "small investor" on a finite horizon. *SIAM Journal of Control and Optimization 25*, 1157–1186, 1987.

Karatzas, I. and Shreve, *S.: Brownian Motion and Stochastic Calculus.* Springer-Verlag, New York, 1988.

Lintner, J.: The valuation of risky assets and the selection of risky investment in stock portfolios and capital budgets. *Review of Economics and Statistics 47*, 13–37, 1965.

Merton, R. C. : Lifetime portfolio selection under uncertainty: the continuous-time case. *Review of Economics and Statistics 51* (*August*), pp. 247–257, 1969.

Merton, R. C. : Optimum consumption and portfolio rules in a continuous-time model. *Journal of Economic Theory 3* (*December*) pp. 373–413, 1971; Erratum (1973), pp. 213–214.

Merton, R. C. : Theory of rational option pricing. *Bell Journal of Economics and Management Science 4* (*Spring*), pp. 141–183, 1973a.

Merton, R. C. : An intertemporal capital asset pricing model. *Econometrica 41* (*September*), pp. 867–888, 1973b.

Merton, R. C.: Option pricing when underlying stock returns are discontinuous. *Journal of Financial Economics 3* (*January-February*), pp. 125–144, 1976.

Merton, R. C.: On the pricing of contingent claims and the Modigliani-Miller theorem. *Journal of Financial Economics 5* (*November*), pp. 241–250, 1977.

Miller, M. and Scholes, M.: Dividends and taxes. *Journal of Financial Economics 6*, pp. 333–364, 1978.

Miller, M. and Scholes, M.: Dividends and taxes: some empirical evidence. *Journal of Political Economy 90*, pp. 1118–1141, 1982.

Modigliani, F. and Miller, M.: The cost of capital, corporation finance, and the theory of investment. *American Economic Review 48*, 261–297, 1958.

Rubinstein, M.: The valuation of uncertain income streams and the pricing of options. *Bell Journal of Economics 7*, 407–425, 1976.
Samuelson, P.: Rational theory of warrant pricing. *Industrial Management Review 6*, 13–31, 1965.
Samuelson, P. A., and Merton, R. C.: A complete model of warrant pricing that maximizes utility. *Industrial Management Review 10* (*Winter*), pp. 17–46, 1969.
Scholes, M. S., and Williams, J.: Estimating betas for non-synchronous data. *Journal of Financial Economics 5*, pp. 309–327, 1977.
Scholes, M. S. and Wolfson, M. A: *Taxes and Business Strategy: A Planning Approach*. Englewood Cliffs, NJ: Prentice-Hall, 1992.
Sharpe, W.: Capital asset prices: A theory of market equilibrium under conditions of risk. *Journal of Finance 19*, 425–442, 1964.
Sprenkle, C.: Warrant prices as indications of expectations. *Yale Economics* Essays *1*, 139–232, 1961; reprinted in *The Random Character of Stock Market Prices*, Paul Cootner (ed.), Cambridge, MA: MIT Press, 1967.

Index

abandonment put, 41, 47n.21
absolute smile model, 142–44, 147–49
accounting
  Black on, 33, 43–44
  definition of earnings, 43
  and portfolio risk, 55, 60–61
active traders, 265–66
agency business, 53
agency theories of capital structure, 42
agents, markets for, 96–117
aggregation, 93n.6, 93n.8
AMEX, 167, 267
APT. *See* Arbitrage Pricing Theory (APT) models
arbitrage
  pricing of options, 287–89
  riskless opportunities, 125
  and risk-neutral probability distribution, 132
  tax in pension funding, 42–43, 47n.26
Arbitrage Pricing Theory (APT) models, 9, 21, 161, 162, 170, 188
arbitrage reasoning, 287
Arrow-Debreu model, 22–23, 28
Asay, Michael, 227
asset class benchmark, 60, 93n.6
asset pricing, 21–30
  linear, 188–90
  models, 39, 163, 295
  and volatility, 141, 143, 144–45
  See also Capital Asset Pricing Model
at-the-money implied volatility, 150
average trading gain, 268

balance of payments, 14
balance sheet, 34, 36, 40, 50
Bank for International Settlements (BIS), 51, 53
banks, 49
Bayesian updating, 29, 115
bearish view, 88, 89
best hedge position, 75, 83–84, 88, 89
Best Hedges report, 56, 82–83, 92
Best Replicating Portfolio, 56, 84, 92
beta, 46n.10
  in cross-sectional regressions, 186n.12
  estimation of, 36, 295
  relationship with return, 35–36
  time-varying or uncertain, 38
  zero assets, 19, 20, 21
  zero CAPM, 7–11, 12
  zero rate, 13
Bibliography of Finance, A (Brealey and Edwards), 3
bid-ask bounce, 274
bid-ask spread, 164, 166, 168, 256, 271–74, 280
binomial tree, 159
  extension to include mean reversion, 197, 199, 201
  in forecasts of risk-neutral probabilities, 141–44, 148, 149
  implied, 128–29, 136–38, 141–44, 148, 149, 160n.4
  recombining, 136, 199
  in risk-neutral stochastic processes, 136–38
BIS. *See* Bank for International Settlements
Black, Fischer, 187, 189, 194, 195–96n.1
  on accounting, 33, 43–44
  on applied theory and empirical work, 26
  on consumption smoothing, 29
  contributions to corporate finance, 33–44, 294

Black, Fischer (*continued*)
contributions to financial economics, 3–4, 96–97, 286, 294
contributions to international asset pricing, 11–19
on corporate investment decisions, 36–39
discounting rule, 37, 38–39
on dividends, 42
on hedging, 18–19
implied volatility, 198, 208–9
theory of business cycles, 22
on valuation, 3–31
See also Black, Derman, Toy model; Black-Scholes model
Black, Derman, Toy model, 197, 199, 201, 227, 294
Black-Scholes model, 40, 289–90
and Capital Asset Pricing Model, 19–21
contemporary significance of, 292–94
economic causes of departure from, 140
as extension of arbitrage modeling, 287
history of, 290–92
of option pricing, 33, 129–32, 142–50, 160n.2
and risk-neutral probabilities, 127, 132, 134, 150
and volatility, 131, 138–39, 143, 289
bonds, 17, 19, 26, 49–50, 90, 198, 201–2
book-to-market ratio, 164, 166, 186n.5
borrowing, 10
Brealey, Richard, 3
Brownian motion, 16, 295, 298n.2
bullish view, 88, 90
business cycles, 21–30

call options, 58, 127
Canabarro model, 210, 210n.1
capital, 22, 46n.2
Capital Asset Pricing Model (CAPM), 33, 156, 164, 189, 194, 195
anomalies in, 10
application to option valuation, 19
Black's application of, 290, 291, 294
in general equilibrium, 3–31
global, 13, 18
international, 15, 18
linkage to Nobel Prize, 286
mean-variance analysis as basis for, 56
multiperiod, 295
pricing, 19–21, 23, 26, 161
progression to ICAPM, 30
Sharpe-Lintner, 4–10, 12, 14, 18, 21, 28, 295
and standard framework for corporate finance, 34–36
zero beta, 7–11, 12, 294
See also ICAPM
capitalism, 24
capital market theory, 6
CAPM. *See* Capital Asset Pricing Model
cap prices, 197, 198, 208
cash flow, 36–39, 46n.10, 46n.12, 286, 287, 293
CEV model, 129, 144–49
Chalke, Inc., 197
Chan model, 197, 210, 210n.1
commodity contracts, 294
common stock, 24, 295
competitive market equilibrium, 286
Connor-Korajczyk factors, 162, 163, 170–74, 177–79, 181–82, 186n.10, 188, 189, 191
consumption, 5, 8, 12, 21–22, 161, 294
consumption smoothing, 26–27, 29
continuous-time setting, 294
corporate finance, 33–44
corporate investment decisions, 36–39
correlation effect, 140
correlation matrix, 198, 199
counterintuitive views, 88–89
covariance matrix, 66, 94n.11, 94n.13, 162, 185n.1, 198, 203, 205–8, 210
credit spreads, 49, 50
crises. *See* financial crises
cross-sectional determinants, of expected returns, 161–96
data, 166–70
empirical results, 170–81
Fama-MacBeth portfolio regressions, 171–73
generalized least squares regressions, 173–75
hypotheses, 165–66

individual security Fama-MacBeth regressions, 175–81
linear asset-pricing relations, 188–90
measurement of risk premiums, 190–94
currency, 16, 62, 93n.5, 93n.7

data-snooping bias, 163
DCF. *See* discounted cash flow
debt, 41–42, 47n.24
debt-to-equity ratio, 140
decomposition of revenues, 265–71
default, 49
demand. *See* supply and demand
deposit insurance, 42
derivatives, 19–21, 26, 55, 65
differential taxation hypothesis, 165
diffusion-based option models, 144–45, 148
diffusion process, 16, 19, 20, 21, 23, 26
Dimson procedure, 186n.18
disaggregation, 93n.6
discounted cash flow (DCF), 34, 35, 36–41, 46n.12, 47n.21
discounting rule, 37, 38–39
discount rates, 37, 38, 46n.12
displaced diffusion model, 144, 148
distribution, 128, 145, 153
of gains and losses, 57, 68
historical frequency, 129
lognormal, 132, 134, 136, 145, 160nn.2–3, 160n.5
of outcomes, 65, 68
recovered, 132
risk-neutral, 132–36, 150–51
diversification, 68
dividends, 10–11, 33, 42, 165, 166, 295
Dothan model, 197
duration risk adjustment, 212

economy, 129
Edwards, Helen, 3
effective half-spread, 272–75, 280, 281
efficient market, 33
employee stock compensation plans, 296
equilibrium
analysis, 104–6
arbitrage reasoning, 287
Capital Asset Pricing Model, 3–31, 295
competitive market, 286
fee structure, 107–9
in financial markets, 96
institutional instability of, 100–101
market for fund management, 101–4
market risk premium, 106–7
models of financial, 125–26, 151–53
noisy rational expectations model, 97, 98–116
pooling, 114–16, 117
separating, 112–14, 115
and viable agents, 104–6
"Equilibrium in the Creation of Investment Goods under Uncertainty" (Black), 4–5
equity premium puzzle, 26
errors-in-variables problems, 163, 173
evidence, 31
exchange rate, 14, 15–17, 18, 19, 93n.5
exchanges, 256
execution costs, 257, 264, 273, 277–82, 285n.12
expectations, 27, 30
expected returns. *See* cross-sectional determinants, of expected returns
expiration date, 128, 129, 141
explicit numerical technique, 199
exposure-accounting system, 50

factor analysis, 210n.4
factor loadings, 176
factor-pricing models, 161–96
Fama-French factors, 161, 162, 164, 165, 167, 170, 178, 180, 181, 185n.3, 186n.12
Fama-MacBeth procedures, 163, 171–73, 175–83, 190, 192–94, 196n.6
Federal Reserve Bank, 48
fee structure, 107–9
finance
closed and open economy, 14
corporate, 33–44
data mining in, 10
international, 11–19
valuation of cash flow within, 286
financial asset flows, 14
financial crises, 48–54

financial economics, 3–4, 25
financial institutions, 96–117
financial modeling, 48
financing, 41–42
fixed income, 93n.8
  accounting approach, 60–61
  Best Hedges report, 82–83
  Best Replicating Portfolio, 84, 86
  implied views for, 90, 91
  market exposure summary, 85
  Markovian, lognormal models for, 197–209
  portfolio hot spots, 80
  stress tests for Eurodollar futures and options, 62
flexibility, 40
FOCUS (Financial and Operational Combined Uniform Single) report, 267, 268, 271, 281
four-standard-deviation rule, 58, 66, 93n.5
fund management, 96–117
fuzziness, 41

gains and losses, 57
general equilibrium, 3–31
generalized least squares regressions, 173–75, 196n.6
geometric random walk, 37, 46n.12
gestation lags, 22
global asset allocation, 97
global equity portfolio, 59, 60, 66, 75, 82
global volatility, 137–38
Goldman Sachs, 18, 55, 56, 58, 60, 61, 66, 76, 77, 83, 87, 224–26, 240–41
Gorgias, 31
"Greek letter" exposures, 61, 63
Green's functions, 201

Harrison-Kreps generalization, 293
hedging
  best hedge position, 75, 83–84, 88, 89
  Black's work on, 18–19
  of currency exposures, 93n.7
  exchange rate, 17, 19
  Markovian, lognormal model, 197–209
  of mortgage securities, 212–54
  on options, 293
  portfolio, 56, 75, 83–84, 88, 89, 92
  of risks, 53, 68, 71–72, 75
historical frequency distributions, 129
historical simulations, 66, 67
Hot Spots report, 77–80, 86–87, 92
human capital, 22, 23, 24

ICAPM, 23, 24–25, 30
IMF. *See* International Monetary Fund
immediacy, 195–96n.1, 256–82
implied binomial tree, 128–29, 136–38, 141–44, 148, 149, 159, 160n.4
implied half-spread, 257, 274–77, 280, 281
implied views, 56, 87–90, 92
implied volatility smile model, 129–32, 143
independent state variables, 198, 202
individual security Fama-MacBeth regressions, 175–81
information flows, 8, 20, 28
institutional investors, 166, 257
interest-only strips (IOs), 212–14, 216–30
  broker forecasts and empirical option costs, 235–41, 243
  empirical price and duration functions, 230–35
  option-adjusted spreads, 216–17, 222–23
  predictions of risks, 227–30
interest rates
  and arbitrage pricing of options, 288
  and prepayment of mortgage securities, 212–54
  riskless, 5, 7
  term structure of, 197–209, 211nn.6–7, 294
international asset pricing, 11–19
International Monetary Fund (IMF), 52
IOs. *See* interest-only strips
irrational pricing, 36

Jensen's inequality, 17, 276, 285n.15
J.P. Morgan (co.), 224, 225, 240–41
jump-diffusion model, 143–45, 147, 148
jumps, 140

left skewness, 128, 139, 145, 150, 159
leptokurtosis, 128, 145, 159
leverage, 46n.9, 52, 140
LIBOR. *See* London Interbank Offered Rate
limited-information model, 166
limit orders, 196n.1, 256, 265–71, 285n.9
linear approximation, 56, 57
linear asset pricing, 188–90
linear payoff, 58, 59
Lintner, John, 46n.4, 297n.1
liquidity, 48–54, 164, 165–66, 195–96n.1
liquidity premium, 272
lognormal distribution, 132, 134, 136, 160nn.2–3, 160n.5
lognormal model, 197–209
London Interbank Offered Rate (LIBOR), 49–50, 198, 203, 215
Long-Term Capital Management (LTCM), 48–49
losses. *See* gains and losses; profits and losses; whipsaw losses
LTCM. *See* Long-Term Capital Management

macroeconomic factors, 38, 47n.15, 161, 162, 185n.1
marginal analyses, 90
market exposure, 83–84, 92
Market Exposure report, 56, 84, 85
market portfolio, 127
market risk premium, 106–7
market-value balance sheet, 34
market-value maximization, 34
Markovian, lognormal model, 197–209
Markowitz, Harry, 56
matched book business, 53
maximum entropy criterion, 135
maximum smoothness criterion, 128, 135
MBS. *See* mortgage-backed securities
mean reversion, 37, 197, 198, 199, 201
mean-variance analysis, 56
mean-variance-efficient portfolio, 6, 8–9, 18, 19, 186n.21
Merton, Robert C., 286, 291, 292, 293, 294–95
Merton's ICAPM, 23, 295
Merton's option valuation model, 21, 291
minimum-variance portfolios, 13
model whipsaw, 217, 243, 254
Modern Portfolio Theory, 56, 97
Modigliani-Miller theory, 286, 287
Monte Carlo models, 66, 227, 292
mortgage-backed securities (MBS), 213, 220–23, 226–27, 229, 242–43
mortgages
  broker forecasts and empirical option costs for interest-only strips, 235–41
  convexity and empirical option costs, 212–54
  empirical interest-only price and duration functions, 230–35
  environment and brokers' forecasts, 213–26
  predictions of risk in interest-only strips, 227–30
  prepayment, 212–17, 224, 227–30, 242
  spreads, 49
mutual funds, 97

naïve trader models, 129, 142, 143, 146, 149, 159
NASDAQ Stock Market, 167, 265, 269, 270–71, 281, 285n.11
negative skewness, 140, 212
"neglected firm" effect, 186n.9
net present value (NPV), 40
New York Stock Exchange. *See* NYSE
noise, 9, 29, 42
noisy rational expectations model, 97, 98–116
nonlinear payoff, 58, 59, 61
NPV. *See* net present value
NYSE (New York Stock Exchange), 166, 168, 176, 181, 256–82

October, 1987, stock market crash, 27–28, 93n.4, 127, 140, 293
one-dimensional, recombining binomial tree, 199
one-factor, Markovian, lognormal model, 197, 199–202
option-adjusted durations, 216, 218–20, 230–35, 246–47, 250–52

option-adjusted spreads (OAS), 215–17, 222–24, 244–45
option costs (whipsaw costs), 217, 221–26, 229, 235–43, 248–49, 253–54
option risk adjustment, 212
options
  arbitrage pricing of, 287–89
  call, 58, 127
  hedging on, 293
  Merton's work on, 21, 291, 295
  pricing models, 20, 39, 40–41, 48, 125–59
  put, 52, 127, 129, 156, 158
  real, 39–41
  valuation, 19, 20
  See also Black-Scholes model
out-of-the-money puts, 129, 156, 158

partial differential equation (PDE), 290–91, 292, 294–95
participation rate, 268
passive traders, 265–66
PDE. *See* partial differential equation
pensions, 42–43, 47nn.26–27
perfect foresight price reversal, 278–79, 285n.5, 285.16
perfect foresight realized half-spread, 257, 278–82
Perold approach, 285n.12
physical capital, 22, 24
portfolio
  analysis tools, 56, 77–92
  cross-sectional determinants of expected returns, 161–96
  effects, 35
  fixed-income, 197
  formation, 163, 181, 186n.15
  global equity, 59, 60, 66, 75, 82
  hedges, 56, 75, 83–84, 88, 89, 92
  "hot spots," 77–81, 84, 86–87, 92
  implied views of, 56, 87–90, 91, 92
  market, 127
  mean-variance-efficient, 6, 8–9, 18, 19, 186n.21
  measures of risk, 57–59
  minimum-variance, 13
  optimal, 90, 92
  proprietary, 61, 63, 64, 67, 68
  quantitative management, 18
  replicating, 84, 86–87
  risk analysis, 56–57
  risk management, 55–95
  selection, 46n.4
  theory, 53, 57, 96
  understanding risk, 59–68
  volatility, 55, 56, 66, 68–77, 94nn.14–15
  world market, 13
  zero beta, 19
POs. *See* principal-only strips
positive skewness, 212
present value of growth opportunities, 40
price-earnings ratios, 43
price elasticities. *See* option-adjusted durations
pricing factors, 161–96
principal-only strips (POs), 212–14, 227–29
probabilities, 125–59, 201
probability of shortfall, 94n.18
production, 5, 22
profits and losses, 51, 63, 64, 65, 68
proprietary portfolio, 61, 63, 64, 67, 68
Prudential (co.), 224–25
put-call parity, 41
put options, 52, 127, 129, 156, 158
PVGO (present value of growth opportunities), 40

quadratic optimization, 128
quoted half-spread, 271–74, 281, 285n.12

R&D. *See* research and development
"Rational Option Pricing" (Merton), 291
realized half-spread, 257, 262–67, 274–82, 285n.9
realized spread, 257, 262, 272, 280–81
real options, 39–41
relative price, 41
relative smile model, 142–44, 147–49
research and development, 40, 41
reverse repurchase agreement (reverse REPO), 49, 50
risk
  analysis, 56–57
  asset, 47n.24, 56
  aversion, 26–27, 29, 98, 102, 109–10, 125–59

cross-sectional determinants of expected returns, 161–96
decomposition, 56, 68, 72–79, 94n.17, 95n.21
definition, 57
effect, 140
hedging, 53, 68, 71–72, 75
incremental impact on, 94n.17
in interest-only strips, 227–30
market price, 27
measures, 57–59
monitoring, 56, 68–77
premiums, 18, 25–26, 97, 106–7, 190–94, 196n.9
reduction potential, 72
understanding, 59–68
volatility as measure of, 56, 58, 65, 74, 78, 79
See also risk management
risk-adjusted returns, 164
risk-free rate, 13, 26
riskless arbitrage opportunities, 125
riskless rate of interest, 5, 7
riskless return, 125–26
risk management, 50–54
accounting approach, 60–61
best hedge position, 75, 83–84, 88, 89
Black, Merton, Scholes methods of, 292–93
"hot spots," 77–81, 84, 86–87, 92
portfolio, 55–95
Trade Risk Profile, 56, 70, 71, 75, 76, 79–82, 90, 92
versus risk monitoring, 56, 68–77
risk-neutral probabilities, 125–59
distributions, 132–36
empirical tests of alternative forecasts, 141–50
risk-neutral valuation, 20, 293
risky assets, 4, 5, 8, 17, 19, 28, 98–100, 102, 109
Roll, Richard, 227
Roll implied half-spread, 257, 274–77, 280, 281
roundabout production, 22

Salomon Brothers, 197, 225
Scholes, Myron S., 286, 294, 295–96
See also Black-Scholes model
Sears, Timothy, 227
Security Market Line, 8, 13, 14, 19, 36, 38, 46n.9
Sharpe, William, 56, 286, 287
Sharpe-Lintner CAPM, 4–10, 12, 14, 18, 21, 28, 295
Siegel's paradox, 18
size factor, 164, 165
small-firm effect, 10, 165
smiles and smile models, 129–32, 135, 141–44, 147–49
smooth information flows, 24
spread levels, 49, 50
Standard and Poor's 500, 49, 127, 130–32, 134–39, 146–47, 150, 153–56, 164, 166, 169, 182, 258–60, 273, 277, 279
state-contingent prices, 125–26
state dependence, 8, 21–30
statistical measures, 55, 65, 66, 93–94n.10
stochastic growth theory, 26
stochastic information flows, 8, 20
stochastic processes, 128, 136–41, 145
stochastic volatility, 144, 148, 149, 293
stock market
cross-sectional determinants of expected returns, 161–96
October, 1987, crash, 27–28, 93n.4, 127, 140, 293
share prices, 140, 165
supply and demand of immediacy, 256–82
See also Standard and Poor's 500
stock volatility, 289
stop-loss policy, 52–53
strategic investments, 40, 41
stress-loss cushions, 52, 53
stress-loss limits, 51, 52
stress tests, 55, 57, 61, 63–65
strike price, 128, 132, 137, 141, 208–9
stripped mortgage-backed securities, 227
subjective probability, 125–27, 130
supply and demand, 5, 22, 256–82
swap spreads, 49, 50, 51
swaptions, 197, 198, 209

tail exposures, 51
tangency portfolio. *See* mean-variance-efficient portfolio

taxation
  arbitrage, 42–43, 47n.26
  and dividends, 42, 165, 166
  and international asset pricing, 12–13, 14
  and pensions, 42–43, 47n.26
Taxes and Business Strategy (Scholes and Wolfson), 296
Tenney model, 197, 210, 210n.1
term structure, 197–209, 211nn.6–7, 294
tracking error, 66, 69–70, 94n.12
Trade Risk Profile, 56, 70, 71, 75, 76, 79–82, 90, 92
trader's option, 117
trading volume, 164, 165
transaction costs, 52, 79
transition probabilities, 201
t-statistic, 164, 172, 174, 175, 177–79, 181–83, 185n.4
turnover. *See* trading volume
two-factor, Markovian, lognormal model, 197–99, 202–9

upside potential, 58

valuation
  arbitrage-based reasoning, 287
  Black on, 3–31
  within finance, 296
  Markovian, lognormal model, 197–209
  prepayment of mortgage securities, 212–54
  risk-neutral, 20, 293
  of term-structure securities, 197–209, 211nn.6–7, 294
value additivity, 34
value-at-risk (VAR), 50–53, 55, 57–59, 65, 66, 68, 74, 79, 93nn.1–2
value-relevant states, 24
VAR. *See* value-at-risk
viable agents, 104–6, 110
volatility
  and asset price, 141, 143, 144–45
  in Black-Scholes formula, 131, 138–39, 289
  definition of, 57
  expectations regarding, 27
  and general equilibrium model, 9
  global, 137–38
  implied, 140, 143, 147, 148, 150
  and implied binomial tree, 128
  implied Black, 198, 208–9
  implied smile model, 129–32, 143
  at Long-Term Capital Management, 48
  as measure of risk, 56, 58, 65, 74, 78, 79
  of portfolio, 55, 56, 66, 68–77, 94nn.14–15
  of profits and losses, 51
  on Standard and Poor's index, 49
  stochastic, 144, 148, 149, 293
  of stock, 289
  and stop-loss technology, 53
  term structure of, 198
  wealth, 26

wealth, 26, 27, 140
weighted average cost of capital, 46n.3
whipsaw costs. *See* option costs
whipsaw losses, 233
Wiener process, 200
winner's curse, 269
Wolfson, Mark, 296

yield curve shift, 198, 210n.3
yield curve twist, 198

zero beta assets, 19, 20, 21
zero beta CAPM, 7–11, 12, 294
zero beta rate, 13
zero-coupon bonds, 198, 201–2
zero net supply, 13, 19, 24, 29, 30