Francis X. Diebold

Elements of FORECASTING

Fourth Edition

Elements of FORECASTING

Fourth Edition

Francis X. Diebold
University of Pennsylvania

THOMSON
SOUTH-WESTERN

Australia · Brazil · Canada · Mexico · Singapore · Spain · United Kingdom · United States

THOMSON
SOUTH-WESTERN

Elements of Forecasting, Fourth Edition
Francis X. Diebold

VP/Editorial Director:
Jack W. Calhoun

VP/Editor-in-Chief:
Alex von Rosenberg

Sr. Acquisitions Editor:
Mike Worls

Developmental Editor:
Leslie Kauffman, LEAP

Marketing Manager:
Jenny Garamy

Editorial Assistant:
Jessica Hartman

Content Project Manager:
Elycia Arendt

Manager of Technology, Editorial:
John Barans

Technology Project Manager:
Dana Cowden

Sr. Manufacturing Coordinator:
Sandee Milewski

Production House:
Interactive Composition Corporation

Art Director:
Michelle Kunkler

Cover Designer:
Craig Ramsdell, Ramsdell Design

Cover Image:
© Getty Images, Inc.

Printer:
Thomson/West

COPYRIGHT © 2007, 2004
Thomson South-Western, a part of The Thomson Corporation. Thomson, the Star logo, and South-Western are trademarks used herein under license.

Printed in the United States of America
3 4 5 09 08 07

ISBN-13: 978-0-324-32359-7 (package)
ISBN-10: 0-324-32359-X (package)
ISBN-13: 978-0-324-35904-6 (book only)
ISBN-10: 0-324-35904-7 (book only)

ALL RIGHTS RESERVED.
No part of this work covered by the copyright hereon may be reproduced or used in any form or by any means—graphic, electronic, or mechanical, including photocopying, recording, taping, Web distribution or information storage and retrieval systems, or in any other manner—without the written permission of the publisher.

For permission to use material from this text or product, submit a request online at http://www.thomsonrights.com.

Library of Congress Control Number:
2006909336

For more information about our products, contact us at:

Thomson Learning Academic
Resource Center
1-800-423-0563

Thomson Higher Education
5191 Natorp Boulevard
Mason, OH 45040
USA

THE LIBRARY
UNIVERSITY OF GUELPH

To Lawrence Klein, Marc Nerlove, and Peter Pauly,
who taught me forecasting.

Most good texts arise from the desire to leave one's stamp on a discipline by training future generations of students, coupled with the recognition that existing texts are inadequate in various respects. My motivation is no different.

There is a real need for a concise and modern introductory forecasting text. A number of features distinguish this book. First, although it uses only elementary mathematics, it conveys a strong feel for the important advances made since the work of Box and Jenkins more than 30 years ago. In addition to standard models of trend, seasonality, and cycles, it touches—sometimes extensively—upon topics such as

- data mining and in-sample overfitting
- statistical graphics and exploratory data analysis
- model selection criteria
- recursive techniques for diagnosing structural change
- nonlinear models, including neural networks
- regime-switching models
- unit roots and stochastic trends
- smoothing techniques in their relation to stochastic-trend unobserved-components models
- vector autoregressions
- cointegration and error correction
- predictive causality
- forecast evaluation and combination
- simulation and simulation-based methods
- volatility measurement, modeling, and forecasting

Much of that material appears in the "Exercises, Problems, and Complements" following each chapter, which form an integral part of the book. The Exercises, Problems, and Complements are organized so that instructors and students can pick and choose according to their backgrounds and interests.

Second, the book does not attempt to be exhaustive in coverage. In fact, the coverage is intentionally selective, focusing on the core techniques with the widest applicability. The book is designed so that it can be covered realistically in a one-semester course. Core material appears in the main text, and additional material that expands on the depth and breadth of coverage is provided in the Exercises, Problems, and Complements, as well as the Bibliographical and Computational Notes, at the end of each chapter.

Third, the book is applications-oriented. It illustrates all methods with detailed real-world applications designed to mimic typical forecasting situations. In many chapters, the application is the centerpiece of the presentation. In various places, the book uses applications not simply to illustrate the methods but also to drive home an important lesson, the limitations of forecasting, by presenting truly realistic examples in which not everything works perfectly!

Fourth, the book is in touch with modern modeling and forecasting software. It uses Eviews, which is a good modern computing environment for forecasting, throughout. At the same time, I am not a software salesman, so the discussion is not wed to any particular software. Students and instructors can use whatever computing environment they like best.

The book has found wide use among students in a variety of fields, including business, finance, economics, public policy, statistics, and even engineering. *The book is directly accessible at the undergraduate and master's levels;* the only prerequisite is an introductory statistics course that includes linear regression. To help refresh students' memories, Chapter 2 reviews linear regression from a forecasting perspective. The book is also of interest to those with more advanced preparation, because of the hard-to-find direct focus on forecasting (as opposed, for example, to general statistics, econometrics, or time series analysis). I have used it successfully for many years as the primary text in my undergraduate forecasting course, as a background text for various other undergraduate and graduate courses, and as the primary text for master's-level Executive Education courses given to professionals in business, finance, economics, and government.

SUPPLEMENTS

Data Sets and Eviews Programs

Selected data and Eviews programs, as used both in the text chapters and in the Exercises, Problems, and Complements at the end of each chapter, are available on the text Web site at www.thomsonedu.com/economics/diebold.

Text Web Site

The text Web site at www.thomsonedu.com/economics diebold provides teaching resources, including the solutions manual for instructors; learning resources, including data sets and Eviews programs; and many more features.

Solutions Manual

Prepared by Francis Diebold, University of Pennsylvania, the solutions manual contains remarks, suggestions, hints, and solutions for many of the end-of-chapter exercises, problems, and complements. It is available on the text Web site and may be downloaded for use by adopting instructors.

Eviews Software

Upon the instructor's request, Eviews Student Version can be bundled with the text. With Eviews, students can do homework anywhere they have access to a PC. For more information on this special Eviews offer, contact your Thomson South-Western representative or call the Academic Resource Center at 1-800-423-0563.

Economic Applications

Economic Applications includes South-Western's dynamic Web features: EconNews, EconDebate, and EconData Online. Organized by pertinent economic topics and searchable by topic or feature, these features are easy to integrate into the classroom. EconNews, EconDebate, and EconData all deepen students' understanding of theoretical concepts through hands-on exploration and analysis for the latest economic news stories, policy debates, and data. These features are updated on a regular basis. For more information, visit www.thomsonedu.com.

InfoTrac

With InfoTrac College Edition, students can receive anytime, anywhere online access to a database of full-text articles from thousands of popular and scholarly periodicals, such as *Newsweek*, *Fortune*, and *Nation's Business*, among others. InfoTrac is a great way to expose students to online research techniques, with the security that the content is academically based and reliable. For more information, visit www.thomsonedu.com.

You can start using many of these resources right away by following the directions on the access card that came with the purchase of a new book. Get started today at www.thomsonedu.com!

ACKNOWLEDGEMENTS

Many people have contributed to the development of this book—some explicitly, some without knowing it. One way or another, all of the following deserve thanks:

Joan B. Anderon
University of San Diego

Scott Armstrong
University of Pennsylvania

Alan Auerbach
University of California, Berkeley

David Bivin
Indiana University–Purdue University at Indianapolis

Gregory A. Charles
Oregon Health & Science University

Chris Chatfield
University of Bath

Jen-Chi Cheng
Wichita State University

Sidhartha Chib
Washington University in St. Louis

Peter Christoffersen
McGill University

Jörg Clostermann
University of Applied Sciences, Fachhochschule Ingolstadt

Dean Croushore
Federal Reserve Bank of Philadelphia

Robert A. Dickler
IMADEC University

Tom Doan
Estima

Michael Donihue
Colby College

Jean-Marie DuFour
University of Montreal

Jeffrey Edwards
Texas Tech University

Robert F. Engle
University of California, San Diego

Farzad Farsio
Montana State University, Billings

Robert Fildes
University of Lancaster

Jessica Gamburg
Heitman

Antonio Garcia-Ferrer
Universidad Autonoma de Madrid

Patrick A. Gaughan
Farleigh-Dickenson University

Clive Granger
University of California, San Diego

Craig Hakkio
Federal Reserve Bank of Kansas City

Eric Hillebrand
Louisiana State University

Eric C. Howe
University of Saskatchewan

Der An Hsu
University of Wisconsin, Milwaukee

Lawrence R. Klein
University of Pennsylvania

James Kozik
SPSS, Inc.

Junsoo Lee
University of Alabama

Tae-Hwy Lee
University of California, Riverside

David Lilien
University of California, Irvine

Jose Lopez
Federal Reserve Bank of New York

Ron Michener
University of Virginia

Ray Nelson
Brigham Young University

Caitlin O'Neil
Goldman, Sachs & Co.

Llad Phillips
University of California, Santa Barbara

W. Robert Reed
University of Oklahoma

Russell Robins
Tulane University

Philip Rothman
East Carolina University

Glenn D. Rudebusch
Federal Reserve Bank of San Francisco

Robert Rycroft
Mary Washington College

Richard Saba
Auburn University

John H. Shannon
Royal Melbourne Institute of Technology

Steven Shwiff
Texas A&M University–Commerce

Gokce Soydemir
University of Texas, PanAmerican

Robert Stine
University of Pennsylvania

James H. Stock
Harvard University

Mark Strazicich
University of Central Florida

Norman Swanson
Texas A&M University

Hirokuni Tamura
University of Washington

George Tavlas
Bank of Greece

Hiroki Tsurumi
Rutgers University

William Veloce
Brock University

Mark W. Watson
Princeton University

Barry Weller
Penn State University, Erie

Kenneth D. West
University of Wisconsin

Koichi Yoshimine
University of British Columbia

Toshiyuki Yuasa
University of Houston

Tao Zha
Federal Reserve Bank of Atlanta

I am especially grateful to all members of the Thomson South-Western team, past and present, including Elycia Arendt, Jennifer Baker, Jack Calhoun, Jennifer Garamy, Dennis Hanseman, Leslie Kauffman, and Michael Worls, without whose encouragement and guidance this book would not have been written. I am similarly grateful to the many energetic undergraduate and graduate student assistants that I have had over the years, who read and improved much of the manuscript, including Boragan Aruoba, Adam Buresh, Morris Davis, Atsushi Inoue, John Schindler, Eric Schwartz, Chiara Scotti, Georg Strasser, Anthony Tay, Karen Toll, and Ginger Wu.

Finally, I apologize and accept full responsibility for the many errors and shortcomings that undoubtedly remain—minor and major—despite ongoing efforts to eliminate them.

The fourth edition maintains the emphasis of earlier editions on providing an intuitive building-block approach to the development of modern and practical methods for producing, evaluating, and combining forecasts. Within that framework, several improvements have been implemented, including

1. Enhanced and extended discussion of the elements of probability and statistics of maximal relevance to forecasting, now included as a separate Chapter 2,
2. Many new exercises, problems, and complements, which emphasize practical implementation of the methods developed in the text, including simple drills to check understanding,
3. Selectively reworked and/or rearranged material, to maximize clarity and pedagogical effectiveness.

Throughout, my intent has been to insert and delete where needed, sparingly, avoiding the temptation to fix parts "that ain't broke." Hopefully I have moved forward.

F.X.D
August 2006

FRANCIS X. DIEBOLD is W.P. Carey Professor of Economics, and Professor of Finance and Statistics, at the University of Pennsylvania and its Wharton School, and a Research Associate at the National Bureau of Economic Research in Cambridge, Massachusetts. A leader in forecasting and modeling in business, economics, and finance, Diebold has published widely and served on numerous editorial boards, including *Econometrica* and *Review of Economics and Statistics*. He is an elected Fellow of the Econometric Society and the American Statistical Association, and the recipient of Sloan, Guggenheim, and Humboldt awards. A prize-winning teacher and popular lecturer, he lectures worldwide and has held visiting appointments in finance and economics at Princeton University, the University of Chicago, Cambridge University, Johns Hopkins University, and New York University. Diebold also has extensive experience in corporate and policy environments; he is consulted regularly by leading financial firms, central banks, and policy organizations, and he has served on a variety of advisory and corporate boards. From 1986 to 1989 he served as an economist at the Federal Reserve Board in Washington D.C., working first with Paul Volcker and then with Alan Greenspan. You can find him on the web at www.ssc.upenn.edu/~fdiebold.

PART I: GETTING STARTED

Chapter 1: Introduction to Forecasting: Applications, Methods, Books, Journals, and Software 1

1. Forecasting in Action 1
2. Forecasting Methods: An Overview of the Book 3
3. Useful Books, Journals, Software, and Online Information 6
4. Looking Ahead 9
   - Exercises, Problems, and Complements 9
     - Forecasting in daily life: We are all forecasting, all the time 9
     - Forecasting in business, finance, economics, and government 9
     - The basic forecasting framework 10
     - Degrees of forecastability 10
     - Data on the web 10
     - Univariate and multivariate forecasting models 10
   - Concepts for Review 11
   - References and Additional Readings 11

Chapter 2: A Brief Review of Probability, Statistics, and Regression for Forecasting 13

1. Why This Chapter? 13
2. Random Variables, Distributions, and Moments 14
3. Multivariate Random Variables 15
4. Statistics 16
5. Regression Analysis 18

Exercises, Problems, and Complements 30
  Interpreting distributions and densities 30
  Covariance and correlation 30
  Conditional expectations versus linear projections 30
  Conditional mean and variance 30
  Scatterplots and regression lines 30
  Desired values of regression diagnostic statistics 31
  Mechanics of fitting a linear regression 31
  Regression with and without a constant term 31
  Interpreting coefficients and variables 31
  Nonlinear least squares 31
  Regression semantics 32
Bibliographical and Computational Notes 32
Concepts for Review 32
References and Additional Readings 33

Chapter 3. Six Considerations Basic to Successful Forecasting 34

1. The Decision Environment and Loss Function 35
2. The Forecast Object 39
3. The Forecast Statement 40
4. The Forecast Horizon 43
5. The Information Set 45
6. Methods and Complexity, the Parsimony Principle, and the Shrinkage Principle 46
7. Concluding Remarks 47

Exercises, Problems, and Complements 47
  Data and forecast timing conventions 47
  Properties of loss functions 47
  Relationships among point, interval, and density forecasts 47
  Forecasting at short through long horizons 47
  Forecasting as an ongoing process in organizations 48
  Assessing forecasting situations 48
Bibliographical and Computational Notes 49
Concepts for Review 49
References and Additional Readings 50

PART II: BUILDING USING, AND EVALUATING FORECASTING MODELS

Chapter 4. Statistical Graphics for Forecasting 51

1. The Power of Statistical Graphics 51
2. Simple Graphical Techniques 55
3. Elements of Graphical Style 59

4. Application: Graphing Four Components of Real GDP 63
5. Concluding Remarks 66

Exercises, Problems, and Complements 67
  Outliers 67
  Simple versus partial correlation 67
  Graphical regression diagnostic 1: time series plot of y_t, $\hat{y}_t$, and e_t 67
  Graphical regression diagnostic 2: time series plot of e_t^2 or $|e_t|$ 68
  Graphical regression diagnostic 3: scatterplot of e_t versus x_t 68
  Graphical analysis of foreign exchange rate data 68
  Common scales 69
  Graphing real GDP, continued from Section 4 69
  Color 69
  Regression, regression diagnostics, and regression graphics in action 69

Bibliographical and Computational Notes 70
Concepts for Review 71
References and Additional Readings 71

Chapter 5: Modeling and Forecasting Trend 72

1. Modeling Trend 72
2. Estimating Trend Models 80
3. Forecasting Trend 81
4. Selecting Forecasting Models Using the Akaike and Schwarz Criteria 82
5. Application: Forecasting Retail Sales 87

Exercises, Problems, and Complements 94
  Calculating forecasts from trend models 94
  Identifying and testing trend models 94
  Understanding model selection criteria 94
  Mechanics of trend estimation and forecasting 95
  Properties of polynomial trends 95
  Specialized nonlinear trends 95
  Moving average smoothing for trend estimation 95
  Bias corrections when forecasting from logarithmic models 96
  Model selection for long-horizon forecasting 97
  The variety of "information criteria" reported across software packages 97

Bibliographical and Computational Notes 97
Concepts for Review 98
References and Additional Readings 98

Chapter 6: Modeling and Forecasting Seasonality 99

1. The Nature and Sources of Seasonality 99
2. Modeling Seasonality 101
3. Forecasting Seasonal Series 103
4. Application: Forecasting Housing Starts 104

Exercises, Problems, and Complements 108
  Log transformations in seasonal models 108
  Seasonal adjustment 108
  Selecting forecasting models involving calendar effects 108
  Testing for seasonality 109
  Seasonal regressions with an intercept and $s-1$ seasonal dummies 109
  Applied trend and seasonal modeling 109
  Periodic models 109
  Interpreting dummy variables 110
  Constructing seasonal models 110
  Calendar effects 110
Bibliographical and Computational Notes 111
Concepts for Review 111
References and Additional Readings 111

Chapter 7 Characterizing Cycles 112

1. Covariance Stationary Time Series 113
2. White Noise 117
3. The Lag Operator 123
4. Wold's Theorem, the General Linear Process, and Rational Distributed Lags 124
5. Estimation and Inference for the Mean, Autocorrelation, and Partial Autocorrelation Functions 127
6. Application: Characterizing Canadian Employment Dynamics 130

Exercises, Problems, and Complements 132
  Lag operator expressions 1 132
  Lag operator expressions 2 133
  Autocorrelation functions of covariance stationary series 133
  Autocorrelation vs. partial autocorrelation 133
  Conditional and unconditional means 133
  White noise residuals 133
  Selecting an employment forecasting model with the AIC and SIC 134
  Simulation of a time series process 134
  Sample autocorrelation functions for trending series 134
  Sample autocorrelation functions for seasonal series 134
  Volatility dynamics: correlograms of squares 135
Bibliographical and Computational Notes 135
Concepts for Review 135
References and Additional Readings 136

Chapter 8 Modeling Cycles: MA, AR, and ARMA Models 137

1. Moving Average (MA) Models 138
2. Autoregressive (AR) Models 145
3. Autoregressive Moving Average (ARMA) Models 152

4. Application: Specifying and Estimating Models for Employment Forecasting 154

Exercises, Problems, and Complements 163
ARMA lag inclusion 163
Shapes of correlograms 163
The autocovariance function of the MA(1) process, revisited 163
ARMA algebra 163
Diagnostic checking of model residuals 163
Mechanics of fitting ARMA models 165
Modeling cyclical dynamics 165
Aggregation and disaggregation: top-down forcasting model vs. bottom-up forecasting model 165
Nonlinear forecasting models: regime switching 165
Difficulties with nonlinear optimization 166

Bibliographical and Computational Notes 167
Concepts for Review 168
References and Additional Readings 169

Chapter 9: Forecasting Cycles 171

1. Optimal Forecasts 171
2. Forecasting Moving Average Processes 172
3. Making the Forecasts Operational 176
4. The Chain Rule of Forecasting 177
5. Application: Forecasting Employment 180

Exercises, Problems, and Complements 184
Forecast accuracy across horizons 184
Mechanics of forecasting with ARMA models: BankWire continued 184
Forecasting an AR(1) process with known and unknown parameters 185
Forecasting an ARMA(2, 2) process 185
Optimal forecasting under asymmetric loss 186
Truncation of infinite distributed lags, state space representations, and the Kalman filter 187
Point and interval forecasts allowing for serial correlation—Nile.com continued 187
Bootstrapping simulation to acknowledge innovation distribution uncertainty and parameter estimation uncertainty 188

Bibliographical and Computational Notes 189
Concepts for Review 190
References and Additional Readings 190

Chapter 10: Putting It All Together: A Forecasting Model with Trend, Seasonal, and Cyclical Components 191

1. Assembling What We've Learned 191
2. Application: Forecasting Liquor Sales 193

3. Recursive Estimation Procedures for Diagnosing and Selecting Forecasting Models 207
4. Liquor Sales, Continued 212
Exercises, Problems, and Complements 214
Serially correlated disturbances vs. lagged dependent variables 214
Assessing the adequacy of the liquor sales forecasting model trend specification 214
Improving nontrend aspects of the liquor sales forecasting model 214
CUSUM analysis of the housing starts model 215
Model selection based on simulated forecasting performance 215
Seasonal models with time-varying parameters: forecasting AirSpeed passenger-miles 215
Formal models of unobserved components 216
The restrictions associated with unobserved-components structures 216
Additive unobserved-components decomposition and multiplicative unobserved-components decomposition 217
Signal, noise, and overfitting 217
Bibliographical and Computational Notes 217
Concepts for Review 218
References and Additional Readings 218

Chapter 11: Forecasting with Regression Models 219

1. Conditional Forecasting Models and Scenario Analysis 220
2. Accounting for Parameter Uncertainty in Confidence Intervals for Conditional Forecasts 220
3. Unconditional Forecasting Models 223
4. Distributed Lags, Polynomial Distributed Lags, and Rational Distributed Lags 224
5. Regressions with Lagged Dependent Variables, Regressions with ARMA Disturbances, and Transfer Function Models 225
6. Vector Autoregressions 228
7. Predictive Causality 230
8. Impulse-Response Functions and Variance Decompositions 231
9. Application: Housing Starts and Completions 235
Exercises, Problems, and Complements 249
Econometrics, time series analysis, and forecasting 249
Forecasting crop yields 249
Regression forecasting models with expectations, or anticipatory, data 249
Business cycle analysis and forecasting: expansions, contractions, turning points, and leading indicators 250
Subjective information, Bayesian VARs, and the Minnesota prior 251
Housing starts and completions, continued 251
Nonlinear regression models 1: functional form and Ramsey's test 251
Nonlinear regression models 2: logarithmic regression models 252
Nonlinear regression models 3: neural networks 252
Spurious regression 253
Comparative forecasting performance of VAR and univariate models 254

Bibliographical and Computational Notes 254
Concepts for Review 255
References and Additional Readings 255

Chapter 12 Evaluating and Combining Forecasts 257

1. Evaluating a Single Forecast 257
2. Evaluating Two or More Forecasts: Comparing Forecast Accuracy 260
3. Forecast Encompassing and Forecast Combination 263
4. Application: OverSea Shipping Volume on the Atlantic East Trade Lane 268
Exercises, Problems, and Complements 280
  Forecast evaluation in action 280
  Forecast error analysis 280
  Combining forecasts 280
  Quantitative forecasting, judgmental forecasting, forecast combination, and shrinkage 281
  The algebra of forecast combination 281
  The mechanics of practical forecast evaluation and combination 282
  What are we forecasting? Preliminary series, revised series, and the limits to forecast accuracy 282
  Ex post versus real-time forecast evaluation 283
  What do we know about the accuracy of macroeconomic forecasts? 283
  Forecast evaluation when realizations are unobserved 283
  Forecast error variances in models with estimated parameters 283
  The empirical success of forecast combination 284
  Forecast combination and the Box-Jenkins paradigm 284
  Consensus forecasts 285
  The Delphi method for combining experts' forecasts 285
Bibliographical and Computational Notes 285
Concepts for Review 286
References and Additional Readings 286

PART III MORE ADVANCED TOPICS

Chapter 13 Unit Roots, Stochastic Trends, ARIMA Forecasting Models, and Smoothing 288

1. Stochastic Trends and Forecasting 288
2. Unit Roots: Estimation and Testing 295
3. Application: Modeling and Forecasting the Yen/Dollar Exchange Rate 302
4. Smoothing 312
5. Exchange Rates, Continued 318
Exercises, Problems, and Complements 320
  Modeling and forecasting the deutschemark/dollar (DEM/USD) exchange rate 320

Housing starts and completions, continued 320
ARIMA models, smoothers, and shrinkage 320
Using stochastic trend unobserved-components models to implement smoothing techniques in a probabilistic framework 320
Automatic ARIMA modeling 321
The multiplicative seasonal ARIMA(p, d, q) × (P, D, Q) model 321
The Dickey-Fuller regression in the AR(2) case 321
Holt-Winters smoothing with multiplicative seasonality 322
Cointegration 323
Error correction 323
Forecast encompassing tests for $I(1)$ series 324
Evaluating forecasts of integrated series 324
Theil's U-statistic 324

Bibliographical and Computational Notes 325
Concepts for Review 326
References and Additional Readings 326

Chapter 14: Volatility Measurement, Modeling, and Forecasting 329

1. The Basic ARCH Process 330
2. The GARCH Process 333
3. Extensions of ARCH and GARCH Models 337
4. Estimating, Forecasting, and Diagnosing GARCH Models 340
5. Application: Stock Market Volatility 341

Exercises, Problems, and Complements 349
Removing conditional mean dynamics before modeling volatility dynamics 349
Variations on the basic ARCH and GARCH models 349
Empirical performance of pure ARCH models as approximations to volatility dynamics 349
Direct modeling of volatility proxies 350
GARCH volatility forecasting 350
Assessing volatility dynamics in observed returns and in standardized returns 350
Allowing for leptokurtic conditional densities 351
Optimal prediction under asymmetric loss 351
Multivariate GARCH models 351

Bibliographical and Computational Notes 352
Concepts for Review 352
References and Additional Readings 353

Bibliography 355

Name Index 361

Subject Index 363

Introduction to Forecasting: Applications, Methods, Books, Journals, and Software

Forecasting is important. Forecasts are constantly made in business, finance, economics, government, and many other fields, and much depends on them. As with anything else, there are good and bad ways to forecast. This book is about the good ways—modern, quantitative, statistical/econometric methods of producing and evaluating forecasts.

1. Forecasting in Action

Forecasts are made to guide decisions in a variety of fields. To develop a feel for the tremendous diversity of forecasting applications, let's sketch some of the areas where forecasts are used and the corresponding diversity of decisions aided by forecasts.

a. *Operations planning and control.* Firms routinely forecast sales to help guide decisions in inventory management, sales force management, and production planning, as well as strategic planning regarding product lines, new market entry, and so on. Firms use forecasts to decide what to produce (What product or mix of products should be produced?), when to produce (Should we build up inventories now in anticipation of high future demand? How many shifts should be run?), how much to produce and how much capacity to build (What are the trends in market size and market share? Are there cyclical or seasonal effects? How quickly and with what pattern will a newly built plant or a newly installed technology depreciate?), and where to produce (Should we have one plant or many? If many, where should we locate them?). Firms also use forecasts of future prices and availability of inputs to guide production decisions.

b. *Marketing.* Forecasting plays a key role in many marketing decisions. Pricing decisions, distribution path decisions, and advertising expenditure decisions all rely heavily on forecasts of responses of sales to different marketing schemes.

c. *Economics.* Governments, policy organizations, and private forecasting firms around the world routinely forecast the major economic variables, such as gross domestic product (GDP), unemployment, consumption, investment, the price level, and interest rates. Governments use such forecasts to guide monetary and fiscal policy, and private firms use them for strategic planning, because economy-wide economic fluctuations typically have industry-level and firm-level effects. In addition to forecasting "standard" variables such as GDP, economists sometimes make more exotic forecasts, such as the stage of the business cycle that we'll be in 6 months from now (expansion or contraction), the state of future stock market activity (bull or bear), or the state of future foreign exchange market activity (appreciation or depreciation). Again, such forecasts are of obvious use to both governments and firms—if they're accurate!

d. *Financial asset management.* Portfolio managers have an interest in forecasting asset returns (stock returns, interest rates, exchange rates, and commodity prices), and such forecasts are made routinely. There is endless debate about the success of forecasts of asset returns. On the one hand, asset returns should be very hard to forecast; if they were easy to forecast, you could make a fortune easily, and any such "get rich quick" opportunities would already have been exploited. On the other hand, those who exploited them along the way may well have gotten rich! Thus, we expect that simple, widely available methods for forecasting should have little success in financial markets, but there may well be profits to be made from using new and sophisticated techniques to uncover and exploit previously unnoticed patterns in financial data (at least for a short time, until other market participants catch on or your own trading moves the market).

e. *Financial risk management.* The forecasting of asset return volatility is related to the forecasting of asset returns. In the last 10 years, practical methods for volatility forecasting have been developed and widely applied. Volatility forecasts are crucial for evaluating and insuring risks associated with asset portfolios. Volatility forecasts are also crucial for firms and investors who need to price assets such as options and other derivatives.

f. *Business and government budgeting.* Businesses and governments of all sorts must constantly plan and justify their expenditures. A major component of the budgeting process is the revenue forecast. Large parts of firms' revenues typically come from sales, and large parts of governments' revenues typically come from tax receipts, both of which exhibit cyclical and long-term variation.

g. *Demography.* Demographers routinely forecast the populations of countries and regions all over the world, often in disaggregated form, such as by age, sex, and race. Population forecasts are crucial for planning government expenditure on health care, infrastructure, social insurance, antipoverty programs, and so forth. Many private sector decisions, such as strategic product line decisions by businesses, are guided by demographic forecasts of particular targeted population subgroups. Population in turn depends on births, deaths, immigration, and emigration, which are also forecasted routinely.

h. *Crisis management.* A variety of events corresponding to crises of various sorts are frequently forecast. Such forecasts are routinely issued as probabilities. For example, in both consumer and commercial lending, banks generate default probability forecasts and refuse loans if the probability is deemed too high. Similarly, international investors of various sorts are concerned with probabilities of default, currency devaluations, military coups, and so forth, and use forecasts of such events to inform their portfolio allocation decisions.

The variety of forecasting tasks that we've just sketched was selected to help you begin to get a feel for the depth and breadth of the field. Surely you can think of many more situations in which forecasts are made and used to guide decisions.

With so many different forecasting applications, you might think that a huge variety of forecasting techniques exists and that you'll have to master all of them. Fortunately, that's not the case. Instead, a relatively small number of tools form the common core of almost all forecasting methods. Needless to say, the details differ if one is forecasting Intel's stock price one day and the population of Scotland the next, but the principles underlying the forecasts are identical. Thus, we'll focus on the underlying core principles that drive all applications.

2. Forecasting Methods: An Overview of the Book

To give you a broad overview of the forecasting landscape, let's sketch what's to follow in the chapters ahead. If some of the terms and concepts seem unfamiliar, rest assured that we'll be studying them in depth in later chapters.

Forecasting is inextricably linked to the building of **statistical models.** Before we can forecast a variable of interest, we must build a model for it and estimate the model's parameters using observed historical data. Typically, the estimated model summarizes dynamic patterns in the data; that is, the estimated model provides a statistical characterization of the links between the present and the past. More formally, an estimated **forecasting model** provides a characterization of what we expect in the *present*, conditional on the *past*, from which we infer what to expect in the future, conditional on the present and past. Quite simply, we use the estimated forecasting model to extrapolate the observed historical data.

In this book, we focus on core modeling and forecasting methods that are very widely applicable; variations on them can be applied in almost any forecasting situation. The book is divided into two parts. The first provides background and introduces various fundamental issues relevant to any forecasting exercise. The second treats the construction, use, and evaluation of modern forecasting models. We give special attention to basic methods of forecasting trend, seasonality, and cycles, in both univariate and multivariate contexts.[1] We also discuss special topics in forecasting with regression models, as well as forecast evaluation and combination. Along the way, we introduce a number of modern developments, sometimes in the text and sometimes in the Exercises, Problems, and Complements that follow each chapter. These include model selection criteria, recursive estimation and analysis, ARMA and ARIMA models, unit roots and cointegration, volatility models, simulation, vector autoregressions, and nonlinear forecasting models. Every chapter contains a detailed application; examples include forecasting retail sales, housing starts, employment, liquor sales, exchange rates, and shipping volume.

In this chapter, we provide a broad overview of the forecasting landscape. In Chapter 2 we review probability, statistics, and regression from a forecasting perspective. In Chapter 3, we highlight six considerations relevant to all forecasting tasks: the decision-making environment, the nature of the object to be forecast, the way the forecast will be stated, the forecast horizon, the information on which the forecast will be based, and the choice of forecasting method.

In Chapter 4, we introduce certain aspects of statistical graphics relevant for forecasting. Graphing data is a useful first step in any forecasting project, as it can often reveal features of the data relevant for modeling and forecasting. We discuss a variety of graphical techniques of use in modeling and forecasting, and we conclude with a discussion of the elements of graphical style—what makes good graphics good and bad graphics bad.

After Chapter 4, the chapters proceed differently—each treats a specific set of tools applicable in a specific and important forecasting situation. We exploit the fact that a useful approach to forecasting consists of separately modeling the unobserved components underlying an observed **time series**—trend components, seasonal components, and cyclical components.[2] Trend is that part of a series' movement that corresponds to long-term, slow evolution.

---

[1] See the Exercises, Problems, and Complements at the end of this chapter for a discussion of the meanings of *univariate* and *multivariate*.

[2] We'll define the idea of a time series more precisely in subsequent chapters, but for now just think of a *time series* as a variable of interest that has been recorded over time. For example, the annual rainfall in Brazil from 1950 to 2006, a string of 57 numbers, is a time series. On the basis of that historical data, one might want to forecast Brazilian rainfall for the years 2007–2010.

Seasonality is that part of a series' movement that repeats each year. Cycle is a catchall term for various forms of dynamic behavior that link the present to the past and hence the future to the present.

In Chapter 5, we discuss trend—what it is, where it comes from, why it's important, how to model it, and how to forecast it. We do the same for seasonality in Chapter 6. Next we provide an extensive discussion of cycles; indeed, cycles are so important that we split the discussion into three parts. In Chapter 7, we introduce the idea of a cycle in the context of analysis of covariance stationary time series, and we discuss methods for the quantitative characterization of cyclical dynamics. In Chapter 8, we describe explicit models for cyclical series, focusing on autoregressive (AR), moving average (MA), and mixed (ARMA) processes. Relying heavily on the foundation built in Chapters 7 and 8, we explicitly treat the model-based forecasting of cyclical series in Chapter 9. Finally, in Chapter 10, we assemble what we learned in earlier chapters, modeling and forecasting series with trend, seasonality, and cycles simultaneously present.

In Chapter 11, we consider multiple regression models in greater detail, focusing on nuances of particular relevance for forecasting. In particular, we make the distinction between "conditional" forecasting models, useful for answering "what if" questions (e.g., What will happen to my sales if I lower my price by 10%?) but not directly useful for forecasting, and "unconditional" forecasting models, which are directly useful for forecasting. We also treat issues concerning the proper dynamic specification of such models, including distributed lags, lagged dependent variables, and serially correlated errors, and we study and apply vector autoregressive models in detail.

In Chapter 12, in contrast to our earlier development of methods for constructing and using various forecasting models, we consider the *evaluation* of forecasting performance once a track record of forecasts and realizations has been established. That is, we show how to assess the accuracy of forecasts and how to determine whether a forecast can be improved. We also show how to combine a set of forecasts to produce a potentially superior composite forecast.

Chapters 1–12 form a coherent whole, and some courses may end with Chapter 12, depending on time constraints and course emphasis. For those so inclined to proceed to more advanced material, we include two such chapters.

First, in Chapter 13, we introduce the idea of **stochastic** trend, meaning that the trend can be affected by random disturbances.[3] We show how to forecast in models with stochastic trends and highlight the differences between forecasts from stochastic trend and **deterministic** trend models. Finally, we discuss "smoothing" methods for producing forecasts, which turn out to be optimal for forecasting series with certain types of stochastic trend.

Second, in Chapter 14, we introduce models of time-varying volatility, which have found wide application, especially in financial asset management and risk management. We focus on the so-called ARCH family of volatility models, including several important variations and extensions.

[3] The word *stochastic* simply means "involving randomness." A process is called *deterministic* if it is not stochastic.

3. Useful Books, Journals, Software, and Online Information

As you begin your study of forecasting, it's important that you begin to develop an awareness of a variety of useful and well-known forecasting textbooks, professional forecasting journals where original forecasting research is published, and forecasting software.

BOOKS

A number of good books exist that complement this one; some are broader, some are more advanced, and some are more specialized. Here we'll discuss a few that are more broad or more advanced, in order to give you a feel for the relevant literature. More specialized books will be discussed in subsequent chapters when appropriate.

Wonnacott and Wonnacott (1990) is a well-written and popular statistics book, which you may wish to consult to refresh your memory on statistical distributions, estimation, and hypothesis testing. It also contains a thorough and very accessible discussion of linear regression, which we use extensively throughout this book.[4] Another good source is Anderson, Sweeney, and Williams (2006).

Pindyck and Rubinfeld (1997) is a well-written general statistics and econometrics text, and you'll find it a very useful refresher for basic statistical topics, as well as a good introduction to more advanced **econometric models.** Similarly useful books include Maddala (2001) and Kennedy (1998).

As a student of forecasting, you'll want to familiarize yourself with the broader time series analysis literature.[5] Chatfield (1996) is a good introductory book, which you'll find useful as a background reference. More advanced books, which you may want to consult later, include Granger and Newbold (1986) and Harvey (1993). Granger and Newbold, in particular, is packed with fine insights and explicitly oriented toward those areas of time series analysis that are relevant for forecasting. Hamilton (1994) is a more advanced book suitable for Ph.D.-level study.

---

[4] You'll also want to explore Chapter 2, which provides a concise review of the regression model as relevant for forecasting.

[5] Most forecasting methods are concerned with forecasting time series. The modeling and forecasting of time series are so important that an entire field called **time series analysis** has arisen. Although the origins of the field go back hundreds of years, major advances have occurred in the last 50 years. Time series analysis is intimately related to forecasting, because quantitative time series forecasting techniques require that quantitative time series models first be fit to the series of interest. Thus, forecasting requires knowledge of time series modeling techniques. A substantial portion of this book is therefore devoted to time series modeling.

A number of specialized books are also of interest. Makridakis and Wheelwright (1997) and Bails and Peppers (1997) display good business sense, with interesting discussions, for example, of the different forecasting needs of the subunits of a typical business firm, and of communicating forecasts to higher management. Taylor (1996) provides a nice introduction to modeling and forecasting techniques of particular relevance in finance.

Finally, Makridakis and Wheelwright (1987), Armstrong (2001), Clements and Hendry (2002), and Elliott, Granger, and Timmermann (2005) are informative and well-written collections of articles by experts in various subfields of forecasting, dealing with both forecasting applications and methods. They provide a nice complement to this book, with detailed descriptions of forecasting in action in various business, economic, financial, and governmental settings.

JOURNALS

A number of journals cater to the forecasting community. The leading academic forecasting journals, which contain a mixture of newly proposed methods, evaluation of existing methods, practical applications, and book and software reviews, are *Journal of Forecasting* and *International Journal of Forecasting*. In addition, *Journal of Business Forecasting* is a good source for case studies of forecasting in various corporate and government environments.

Although a number of journals are devoted to forecasting, its interdisciplinary nature results in a rather ironic outcome: A substantial fraction of the best forecasting research is published not in the forecasting journals but rather in the broader applied econometrics and statistics journals, such as *Journal of Business and Economic Statistics*, *Review of Economics and Statistics*, and *Journal of Applied Econometrics*, among many others. Several recent journal symposia have focused on forecasting—see, for example, Diebold and Watson (1996); Diebold and West (1998); Diebold, Stock, and West (1999); Diebold and West (2001); and Diebold, Engle, Favero, Gallo, and Schorfheide (2005).

SOFTWARE

Just as some journals specialize exclusively in forecasting, so too do some software packages. But just as important forecasting articles appear regularly in journals much broader than the specialized forecasting journals, so, too, are forecasting tools scattered throughout econometric/statistical software packages with capabilities much broader than forecasting alone.[6]

One of the best such packages is Eviews, a modern Windows environment with extensive time series, modeling, and forecasting capabilities.[7] Eviews can implement almost all of the methods described in this book (and many more). Most of the examples in this book are done in Eviews, which reflects a balance of generality and specialization that makes it ideal for the sorts of tasks that will concern us.[8] If you feel more comfortable with another package, however, that's fine—none of our discussion is wed to Eviews in any way, and most of our techniques can be implemented in a variety of packages, including Minitab, SAS, and many others.[9]

---

[6] Rycroft (1993) provides a thorough comparison of several forecasting software environments.

[7] The Eviews web page is at www.eviews.com.

If you go on to more advanced modeling and forecasting, you'll probably want to have available an open-ended high-level computing environment in which you can quickly program, evaluate, and apply new tools and techniques. Matlab is one very good such environment.[10] Matlab is particularly well suited for time series modeling and forecasting.[11]

Although most forecasting is done in time series environments, some is done in "cross sections," which refers to examination of a population at one point in time. Stata is an outstanding package for cross-section modeling, with strengths in areas such as qualitative response modeling, Poisson regression, quantile regression, and survival analysis.[12]

Before proceeding, and at the risk of belaboring the obvious, it is important to note that no software is perfect. In fact, all software is highly imperfect! The results obtained when modeling or forecasting in different software environments may differ—sometimes a little and sometimes a lot—for a variety of reasons. The details of implementation may differ across packages, for example, and small differences in details can sometimes produce large differences in results. Hence, it is important that you understand *precisely* what your software is doing (insofar as possible, as some software documentation is more complete than others). And, of course, quite apart from correctly implemented differences in details, always remember that deficient implementations occur: There is no such thing as bug-free software.

ONLINE INFORMATION

A variety of information of interest to forecasters is available on the web. The best way to learn about what's out there in cyberspace is to spend a few hours searching the web for whatever interests you. However, any list of good web sites for forecasters is likely to be outdated shortly after its compilation. Hence, we mention just one, which is regularly updated and tremendously authoritative: Resources for Economists, at www.rfe.org. It contains hundreds of links to data sources, journals, professional organizations, and so on. Frankly, the Resources for Economists page is all you need to start on your way.

---

[8] A number of other good software packages are reviewed by Kim and Trivedi (1995).

[9] S+ also deserves mention as a fine computing environment with special strengths in graphical data analysis and modern statistical methods. See Hallman (1993) for a review.

[10] Matlab maintains a web page that contains material on product availability, user-written add-ons, and more, at www.mathworks.com.

[11] Rust (1993) provides a comparative review of Matlab and one of its competitors, Gauss.

[12] For a review of Stata, see Ferrall (1994). The Stata web page is at www.stata.com. The page has product information, user-supplied routines, course information, and so forth, as well as links to other statistical software products, many of which are useful for forecasting.

4. Looking Ahead

A forecast is little more than a guess about the future. Because forecasts guide decisions, good forecasts help to produce good decisions. In the remainder of this book, we'll motivate, describe, and compare modern forecasting methods. You'll learn how to build and evaluate forecasts and forecasting models, and you'll be able to use them to improve your decisions. Enjoy!

Exercises, Problems, and Complements

1. (Forecasting in daily life: We are all forecasting, all the time)
   a. Sketch in detail three forecasts that you make routinely, and probably informally, in your daily life. What makes you believe that the forecast object is predictable? What factors might introduce error into your forecasts?
   b. What decisions are aided by your three forecasts? How might the degree of predictability of the forecast object affect your decisions?
   c. How might you measure the "goodness" of your three forecasts?
   d. For each of your forecasts, what is the value to you of a "good" as opposed to a "bad" forecast?

2. (Forecasting in business, finance, economics, and government) What sorts of forecasts would be useful in the following decision-making situations? Why? What sorts of data might you need to produce such forecasts?
   a. Shop-All-The-Time Network (SATTN) needs to schedule operators to receive incoming calls. The volume of calls varies depending on the time of day, the quality of the TV advertisement, and the price of the good being sold. SATTN must schedule staff to minimize the loss of sales (too few operators leads to long hold times, and people hang up if put on hold) while also considering the loss associated with hiring excess employees.
   b. You're a U.S. investor holding a portfolio of Japanese, British, French, and German stocks and government bonds. You're considering broadening your portfolio to include corporate stocks of Tambia, a developing economy with a risky emerging stock market. You're only willing to do so if the Tambian stocks produce higher portfolio returns sufficient to compensate you for the higher risk. There are rumors of an impending military coup, in which case your Tambian stocks would likely become worthless. There is also a chance of a major Tambian currency depreciation, in which case the dollar value of your Tambian stock returns would be greatly reduced.
   c. You are an executive with Grainworld, a huge corporate farming conglomerate with grain sales both domestically and abroad. You have no control over the price of your grain, which is determined in the competitive market, but you must decide what to plant and how much, over the next 2 years. You are paid in foreign currency for all grain sold abroad, which you subsequently convert to dollars. Until now, the government has bought all unsold grain to keep the price you receive stable, but the agricultural lobby is weakening, and you are concerned that the government subsidy may be

reduced or eliminated in the next decade. Meanwhile, the price of fertilizer has risen because the government has restricted production of ammonium nitrate, a key ingredient in both fertilizer and terrorist bombs.

   d. You run BUCO, a British utility supplying electricity to the London metropolitan area. You need to decide how much capacity to have on line, and two conflicting goals must be resolved in order to make an appropriate decision. On the one hand, you obviously want to have enough capacity to meet average demand, but that's not enough, because demand is uneven throughout the year. In particular, demand skyrockets during summer heat waves—which occur randomly—as more and more people run their air conditioners constantly. If you don't have sufficient capacity to meet peak demand, you get bad press. On the other hand, if you have a large amount of excess capacity over most of the year, you also get bad press.

3. (The basic forecasting framework) True or false (explain your answers):
   a. The underlying principles of time series forecasting differ radically depending on the time series being forecast.
   b. Ongoing improvements in forecasting methods will eventually enable perfect prediction.
   c. There is no way to learn from a forecast's historical performance whether and how it could be improved.

4. (Degrees of forecastability) Which of the following can be forecast perfectly? Which cannot be forecast at all? Which are somewhere in between? Explain your answers, and be careful!
   a. The direction of change tomorrow in a country's stock market
   b. The eventual lifetime sales of a newly introduced automobile model
   c. The outcome of a coin flip
   d. The date of the next full moon
   e. The outcome of a (fair) lottery

5. (Data on the web) A huge amount of data of all sorts is available on the web. Frumkin (2004) and Baumohl (2005) provide useful and concise introductions to the construction, accuracy, and interpretation of a variety of economic and financial indicators, many of which are available on the web. Search the web for information on U.S. retail sales, U.K. stock prices, German GDP, and Japanese federal government expenditures. (The Resources for Economists page is a fine place to start: www.rfe.org.) Using graphical methods, compare and contrast the movements of each series and speculate about the relationships that may be present.

6. (Univariate and multivariate forecasting models) In this book, we consider both univariate and multivariate forecasting models. In a **univariate model,** a single variable is modeled and forecast solely on the basis of its own past. Univariate approaches to forecasting may seem simplistic, and in some situations they are, but they are tremendously important and worth studying for at least two reasons. First, although they are simple, they are not necessarily simplistic, and a large amount of accumulated experience suggests that they often perform admirably. Second, it's necessary to understand univariate forecasting models before tackling more complicated multivariate models.

   In a **multivariate model,** a variable (or each member of a set of variables) is modeled on the basis of its own past, as well as the past of other variables, thereby accounting for and exploiting cross-variable interactions. Multivariate models have the *potential* to produce forecast improvements relative to univariate models, because they exploit more information to produce forecasts.

a. Determine which of the following are examples of univariate or multivariate forecasting:
   - Using a stock's price history to forecast its price over the next week
   - Using a stock's price history and volatility history to forecast its price over the next week
   - Using a stock's price history and volatility history to forecast its price and volatility over the next week
b. Keeping in mind the distinction between univariate and multivariate models, consider a wine merchant seeking to forecast the price per case at which 1990 Chateau Latour, one of the greatest Bordeaux wines ever produced, will sell in 2015, at which time it will be fully mature.
   - What sorts of univariate forecasting approaches can you imagine that might be relevant?
   - What sorts of multivariate forecasting approaches can you imagine that might be relevant? What other variables might be used to predict the Latour price?
   - What are the comparative costs and benefits of the univariate and multivariate approaches to forecasting the Latour price?
   - Would you adopt a univariate or multivariate approach to forecasting the Latour price? Why?

Concepts for Review

Forecasting
Statistical model
Forecasting model
Time series
Stochastic
Deterministic
Econometric model
Time series analysis
Univariate model
Multivariate model

References and Additional Readings

Anderson, D. R., Sweeney, D. J., and Williams, T. A. (2006). *Statistics for Business and Economics.* 4th ed. Cincinnati: South-Western.

Armstrong, J. S., ed. (2001). *The Principles of Forecasting.* Norwell, Mass.: Kluwer Academic Forecasting.

Bails, D. G., and Peppers, L. C. (1997). *Business Fluctuations.* 2nd ed. Englewood Cliffs, N.J.: Prentice Hall.

Baumohl, B. (2005). *Secrets of Economic Indicators: The Hidden Clues to Future Economic Trends and Investment Opportunities.* Philadelphia: Wharton School Publishing.

Chatfield, C. (1996). *The Analysis of Time Series: An Introduction,* 5th ed. London: Chapman and Hall.

Clements, M. P., and Hendry, D. F., eds. (2002). *A Companion to Economic Forecasting.* Oxford: Blackwell.

Diebold, F. X., Engle, R. F., Favero, C., Gallo, G., and Schorfheide, F. (2005). *The Econometrics of Macroeconomics, Finance, and the Interface.* Special issue of *Journal of Econometrics.*

Diebold, F. X., and Watson, M. W., eds. (1996). *New Developments in Economic Forecasting.* Special issue of *Journal of Applied Econometrics,* 11, 453–594.

Diebold, F. X., Stock, J. H., and West, K. D., eds. (1999). *Forecasting and Empirical Methods in Macroeconomics and Finance, II.* Special issue of *Review of Economics and Statistics,* 81, 553–673.

Diebold, F. X., and West, K. D., eds. (1998). *Forecasting and Empirical Methods in Macroeconomics and Finance.* Special issue of *International Economic Review,* 39, 811–1144.

Diebold, F. X., and West, K. D., eds. (2001). *Forecasting and Empirical Methods in Macroeconomics and Finance III.* Special issue of *Journal of Econometrics,* 105, 1–308.

Elliott, G., Granger, C. W. J., and Timmermann, A., eds. (2005). *Handbook of Economic Forecasting.* Amsterdam: North-Holland.

Ferrall, C. (1994). "A Review of Stata 3.1." *Journal of Applied Econometrics,* 9, 469–478.

Frumkin, N. (2004). *Tracking America's Economy.* 4th ed. Armonk, N.Y.: Sharpe.

Granger, C. W. J., and Newbold, P. (1986). *Forecasting Economic Time Series.* 2nd ed. Orlando, Fl.: Academic Press.

Hallman, J. (1993). "Review of S+." *Journal of Applied Econometrics,* 8, 213–220.

Hamilton, J. D. (1994). *Time Series Analysis.* Princeton, N.J.: Princeton University Press.

Harvey, A. C. (1993). *Time Series Models.* 2nd ed. Cambridge, Mass.: MIT Press.

Kennedy, P. (1998). *A Guide to Econometrics.* 4th ed. Cambridge, Mass.: MIT Press.

Kim, J., and Trivedi, P. (1995). "Econometric Time Series Analysis Software: A Review." *American Statistician,* 48, 336–346.

Maddala, G. S. (2001). *Introduction to Econometrics.* 3rd ed. New York: Macmillan.

Makridakis, S., and Wheelwright S. (1987). *The Handbook of Forecasting: A Manager's Guide.* 2nd ed. New York: Wiley.

Makridakis, S., and Wheelwright S. C. (1997). *Forecasting: Methods and Applications.* 3rd ed. New York: Wiley.

Pindyck, R. S., and Rubinfeld, D. L. (1997). *Econometric Models and Economic Forecasts.* 4th ed. New York: McGraw-Hill.

Rust, J. (1993). "Gauss and Matlab: A Comparison." *Journal of Applied Econometrics,* 8, 307–324.

Rycroft, R. S. (1993). "Microcomputer Software of Interest to Forecasters in Comparative Review: An Update." *International Journal of Forecasting,* 9, 531–575.

Taylor, S. (1996). *Modeling Financial Time Series.* 2nd ed. New York: Wiley.

Wonnacott, T. H., and Wonnacott, R. J. (1990). *Introductory Statistics.* 5th ed. New York: Wiley.

A Brief Review of Probability, Statistics, and Regression for Forecasting

1. Why This Chapter?

The role of this chapter is threefold. First, it reviews some familiar material. You've already studied some probability and statistics, but chances are that you could use a bit of review, so this chapter supplies it.[1]

Second, although this chapter largely reviews familiar material, it does so from a new perspective. That is, it begins developing the material from the explicit perspective of forecasting, which involves special considerations and nuances. For example, we motivate the regression model as a model of a conditional expectation, which turns out to be an intuitive and appealing forecast.

Third, the chapter foreshadows *new* material subsequently developed in greater detail. It begins to introduce tools that are new but that are related to things you learned earlier and very important for building forecasting models, such as information criteria for model selection. Hence, you should not worry if some of the material looks unfamiliar!

---

[1] Be warned, however: This chapter is no substitute for a full-course introduction to probability and statistics. If the bulk of it looks unfamiliar to you, you're in trouble and should speak with your instructor immediately.

2. Random Variables, Distributions, and Moments

Consider an experiment with a set O of possible outcomes. A random variable Y is simply a mapping from O to the real numbers. For example, the experiment might be flipping a coin twice, in which case $O = \{$(Heads, Heads), (Tails, Tails), (Heads, Tails), (Tails, Heads)$\}$. We might define a random variable Y to be the number of heads observed in the two flips, in which case Y could assume three values, $y = 0$, $y = 1$, and $y = 2$.[2]

Discrete random variables—that is, random variables with **discrete probability distributions**—can assume only a countable number of values y_i, $i = 1, 2, \ldots$, each with positive probability p_i such that $\sum_i p_i = 1$. The probability distribution $f(y)$ assigns a probability p_i to each such value y_i. In the example at hand, Y is a discrete random variable, and $f(y) = 0.25$ for $y = 0$, $f(y) = 0.50$ for $y = 1$, $f(y) = 0.25$ for $y = 2$, and $f(y) = 0$ otherwise.

In contrast, **continuous random variables** can assume a continuum of values, and the **probability density function** $f(y)$ is a nonnegative continuous function such that the area under $f(y)$ between any points a and b is the probability that Y assumes a value between a and b.[3]

In what follows we will simply speak of a "distribution" $f(y)$. It will be clear from context whether we are in fact speaking of a discrete random variable with probability distribution $f(y)$ or a continuous random variable with probability density $f(y)$.

Moments provide important summaries of various aspects of distributions. Roughly speaking, moments are simply expectations of powers of random variables, and expectations of different powers convey different sorts of information. You are already familiar with two crucially important moments, the mean and variance. In what follows we shall consider the first four moments: mean, variance, skewness, and kurtosis.[4]

The **mean,** or **expected value,** of a discrete random variable is a probability-weighted average of the values it can assume,[5]

$$E(y) = \sum_i p_i y_i .$$

Often we use the Greek letter μ to denote the mean. The mean measures the **location,** or **central tendency,** of y.

The **variance** of y is its expected squared deviation from its mean,

$$\sigma^2 = \text{var}(y) = E(y - \mu)^2 .$$

It measures the **dispersion,** or **scale,** of y around its mean.

---

[2] Note that we use capitals for random variables (Y) and lowercase letters for their realizations (y). We will often neglect this formalism, however, as the meaning will be clear from context.

[3] In addition, the total area under $f(y)$ must be 1.

[4] In principle, we could of course consider moments beyond the fourth, but in practice, only the first four are typically examined.

[5] A similar formula holds in the continuous case.

Often we assess dispersion using the square root of the variance, which is called the **standard deviation,**

$$\sigma = \mathrm{std}(y) = \sqrt{E(y-\mu)^2}\,.$$

The standard deviation is more easily interpreted than the variance, because it has the same units of measurement as y. That is, if y is measured in dollars (say), then var(y) is in dollars squared, but std(y) is again in dollars.

The **skewness** of y is its expected cubed deviation from its mean (scaled by σ^3 for technical reasons),

$$S = \frac{E(y-\mu)^3}{\sigma^3}\,.$$

Skewness measures the amount of **asymmetry** in a distribution. The larger the absolute size of the skewness, the more asymmetrical is the distribution. A large positive value indicates a long right tail, and a large negative value indicates a long left tail. A zero value indicates symmetry around the mean.

The **kurtosis** of y is the expected fourth power of the deviation of y from its mean (scaled by σ^4),

$$K = \frac{E(y-\mu)^4}{\sigma^4}\,.$$

Kurtosis measures the thickness of the tails of a distribution. A kurtosis above 3 indicates "fat tails," or **leptokurtosis,** relative to the **normal,** or **Gaussian, distribution** that you studied in earlier course work. Hence, a kurtosis above 3 indicates that extreme events are more likely to occur than would be the case under normality.

3. Multivariate Random Variables

Suppose now that instead of a single random variable Y, we have two random variables Y and X.[6] We can examine the distributions of Y or X in isolation, which are called **marginal distributions.** This is effectively what we've already studied. But now there's more: Y and X may be related and therefore move together in various ways, characterization of which requires a **joint distribution.** In the discrete case, the joint distribution $f(y, x)$ gives the probability associated with each possible pair of y and x values; in the continuous case, the joint density $f(y, x)$ is such that the area under it in any region is the probability of a (y, x) realization in that region.

We can examine the moments of y or x in isolation, such as mean, variance, skewness, and kurtosis. But, again, there's more: To help assess the dependence between y and x, we often examine a key moment of relevance in multivariate environments, the **covariance.** The covariance between y and x is simply the expected product of the deviations of y and x from their respective means,

$$\mathrm{cov}(y, x) = E((y_t - \mu_y)(x_t - \mu_x))\,,$$

---

[6] We could of course consider more than two variables, but for pedagogical reasons, we presently limit ourselves to two.

A positive covariance means that y and x are positively related. That is, when y is above its mean, x tends to be above its mean; and when y is below its mean, x tends to be below its mean. Conversely, a negative covariance means that y and x are inversely related: When y is below its mean x tends to be above its mean, and vice versa. The covariance can take any value in the real numbers.

Frequently, we convert the covariance to a **correlation** by standardizing by the product of σ_y and σ_x,

$$\operatorname{corr}(y, x) = \frac{\operatorname{cov}(y, x)}{\sigma_y \sigma_x}.$$

The correlation takes values in $[-1, 1]$. Note that covariance depends on units of measurement (such as dollars, cents, and billions of dollars), but correlation does not. Hence, correlation is more immediately interpretable, which is the reason for its popularity.

Note also that covariance and correlation measure only *linear* dependence; in particular, a zero covariance or correlation between y and x does not necessarily imply that y and x are independent. That is, they may be *non*linearly related. If, however, two random variables are jointly *normally* distributed with zero covariance, then they are independent.

Our multivariate discussion has focused on the joint distribution $f(y, x)$. In later chapters we will also make heavy use of the **conditional distribution** $f(y \mid x)$—that is, the distribution of the random variable Y *conditional* on $X = x$. Conditional distributions are tremendously important for forecasting, in which a central concern is the distribution of future values of a series conditional on past values. **Conditional moments** are similarly important. In particular, the **conditional mean** and **conditional variance** play key roles in forecasting, in which attention often centers on the mean or variance of a series conditional on its past values.

4. Statistics

Thus far, we've reviewed aspects of known **population distributions** of random variables. Often, however, we have a sample of data drawn from an unknown population distribution f,

$$\{y_t\}_{t=1}^{T} \sim f(y),$$

and we want to learn from the **sample** about various aspects of f, such as its moments. To do so, we use various estimators.[7] We can obtain estimators by replacing population expectations with sample averages, because the arithmetic average is the sample analog of the population expectation. Such "analog estimators" turn out to have good properties quite generally.

---

[7] An **estimator** is an example of a **statistic**, or **sample statistic**, which is simply a function of the sample observations.

The **sample mean** is simply the arithmetic average,

$$\bar{y} = \frac{1}{T}\sum_{t=1}^{T} y_t .$$

It provides an empirical measure of the location of y.

The **sample variance** is the average squared deviation from the sample mean,

$$\hat{\sigma}^2 = \frac{\sum_{t=1}^{T}(y_t - \bar{y})^2}{T} .$$

It provides an empirical measure of the dispersion of y around its mean.

We commonly use a slightly different version of $\hat{\sigma}^2$, which corrects for the 1 degree of freedom used in the estimation of $\bar{y}$, thereby producing an unbiased estimator of σ^2,

$$s^2 = \frac{\sum_{t=1}^{T}(y_t - \bar{y})^2}{T-1} .$$

Similarly, the **sample standard deviation** is defined as either

$$\hat{\sigma} = \sqrt{\hat{\sigma}^2} = \sqrt{\frac{\sum_{t=1}^{T}(y_t - \bar{y})^2}{T}}$$

or

$$s = \sqrt{s^2} = \sqrt{\frac{\sum_{t=1}^{T}(y_t - \bar{y})^2}{T-1}} .$$

It provides an empirical measure of dispersion in the same units as y.

The **sample skewness** is

$$\hat{S} = \frac{\frac{1}{T}\sum_{t=1}^{T}(y_t - \bar{y})^3}{\hat{\sigma}^3} .$$

It provides an empirical measure of the amount of asymmetry in the distribution of y.

The **sample kurtosis** is

$$\hat{K} = \frac{\frac{1}{T}\sum_{t=1}^{T}(y_t - \bar{y})^4}{\hat{\sigma}^4} .$$

It provides an empirical measure of the fatness of the tails of the distribution of y relative to a normal distribution.

Many of the most famous and important statistical sampling distributions arise in the context of sample moments, and the normal distribution is the father of them all. In particular, the celebrated central limit theorem establishes that under quite general conditions, the sample mean $\bar{y}$ will have a normal distribution as the sample size gets large. The **χ^2 distribution** arises from squared normal random variables, the **t distribution** arises from ratios of normal and χ^2 variables, and the **F distribution** arises from ratios of χ^2 variables.

Because of the fundamental nature of the normal distribution as established by the central limit theorem, it has been studied intensively, a great deal is known about it, and a variety of powerful tools have been developed for use in conjunction with it. Hence, it is often of interest to assess whether the normal distribution governs a given sample of data. A simple strategy is to check various implications of normality, such as $S = 0$ and $K = 3$, via informal examination of $\hat{S}$ and $\hat{K}$. Alternatively and more formally, the **Jarque-Bera test** (JB) effectively aggregates the information in the data about both skewness and kurtosis to produce an overall test of the hypothesis that $S = 0$ and $K = 3$, based on $\hat{S}$ and $\hat{K}$.[8] The test statistic is

$$\mathrm{JB} = \frac{T}{6}\left(\hat{S}^2 + \frac{1}{4}(\hat{K} - 3)^2\right),$$

where T is the number of observations.[9] Under the null hypothesis of independent normally distributed observations, the Jarque-Bera statistic is distributed in large samples as a χ^2 random variable with 2 degrees of freedom. We will use the Jarque-Bera test in various places throughout this book.

5. Regression Analysis

Ideas that fall under the general heading of "**regression analysis**" are crucial for building forecasting models, using them to produce forecasts, and evaluating those forecasts. Here we provide a brief review of linear regression to refresh your memory and provide motivation from a forecasting perspective.

Suppose that we have data on two variables, y and x, as in Figure 2.1, and suppose that we want to find the linear function of x that gives the best forecast of y, where "best forecast" means that the sum of squared forecast errors, for the sample of data at hand, is as small as possible. This amounts to finding the line that best fits the data points, in the sense that the sum of squared vertical distances of the data points from the fitted line is minimized. When we "run a regression" or "fit a regression line," that's what we do. The estimation

---

[8] Other tests of conformity to the normal distribution exist and may of course be used, such as the Kolmogorov-Smirnov test. We use the Jarque-Bera test in this book because of its simplicity and because of its convenient and intuitive decomposition into skewness and leptokurtosis components.

[9] The formula given is for an observed time series. If the series being tested for normality is the residual from a model, then T should be replaced with $T - k$, where k is the number of parameters estimated.

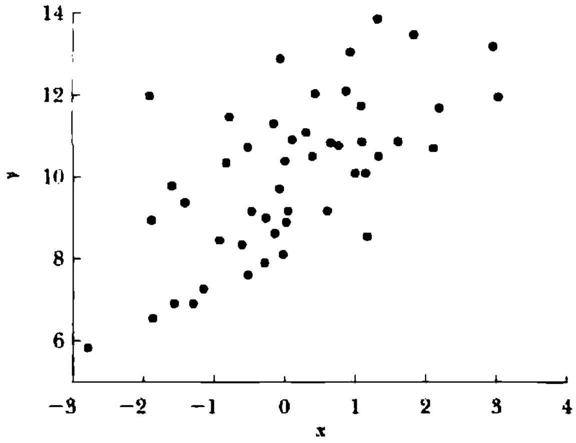


FIGURE 2.1
Scatterplot of y versus x

strategy is called **least squares.** The least squares estimator has a well-known mathematical formula. We won't reproduce it here; suffice it to say that we simply use the computer to evaluate the formula.

In Figure 2.2, we illustrate graphically the results of regressing y on x. The best-fitting line slopes upward, reflecting the positive correlation between y and x. Note that the data points don't satisfy the fitted linear relationship exactly; rather, they satisfy it on average. To forecast y for any given value of x, we use the fitted line to find the value of y that corresponds to the given value of x.

Thus far, we haven't postulated a probabilistic model that relates y and x; instead, we simply ran a mechanical regression of y on x to find the best forecast of y formed as a linear function of x. It's easy, however, to construct a probabilistic framework that lets us make statistical assessments about the

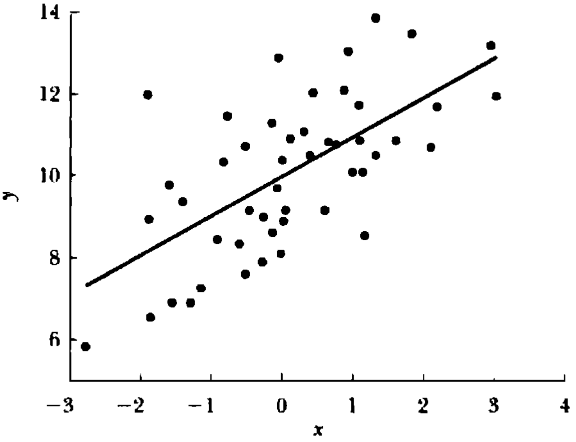


FIGURE 2.2
Scatterplot of y versus x, Regression Line Superimposed

properties of the fitted line and the corresponding forecasts. We assume that y is linearly related to an exogenously determined variable x, and we add an independent and identically distributed (iid) **disturbance** with zero mean and constant variance:

$$y_t = \beta_0 + \beta_1 x_t + \varepsilon_t$$

$$\varepsilon_t \overset{\text{iid}}{\sim} (0, \sigma^2)\,,$$

$t = 1, \ldots, T$. The intercept of the line is β_0, the slope is β_1, and the variance of the disturbance is σ^2.[10] Collectively, β_0, β_1, and σ^2 are called the model's **parameters.** The index t keeps track of time; the data sample begins at some time we've called "1" and ends at some time we've called "T."

If the regression model postulated here holds true, then the expected value of y conditional on x taking a particular value, say x^*, is

$$E(y \mid x^*) = \beta_0 + \beta_1 x^*\,.$$

That is, the **regression function** is the **conditional expectation** of y. As we'll see in detail later in the book, the expectation of future y conditional on available information is a particularly good forecast. In fact, under fairly general conditions, it is the best possible forecast. The intimate connection between regression and optimal forecasts makes regression an important tool for forecasting.

We assume that the model sketched here is true in population. If we knew β_0 and β_1, we could make a forecast of y for any given value of x_t^*, and the variance of the corresponding forecast error would be σ^2. The problem, of course, is that we don't *know* the values of the model's parameters. When we run the regression, or "estimate the regression model," we use a computer to *estimate* the unknown parameters by solving the problem $\min\limits_{\beta} \sum\limits_{t=1}^{T} (y_t - \beta_0 - \beta_1 x_t)^2$ (or, equivalently, $\min\limits_{\beta} \sum\limits_{t=1}^{T} \varepsilon_t^2$, because $y_t - \beta_0 - \beta_1 x_t = \varepsilon_t$), where β is shorthand notation for the set of two parameters, β_0 and β_1.[11] We denote the set of estimated parameters by $\hat{\beta}$ and its elements by $\hat{\beta}_0$ and $\hat{\beta}_1$. Each estimated coefficient gives the weight put on the corresponding variable in forming the best linear forecast of y. We can think of β_0 as the coefficient on a "constant" variable that's always equal to 1. The estimated coefficient on the constant variable is the best forecast in the event that x is 0. In that sense, it's a baseline forecast. We use the set of estimated parameters, $\hat{\beta}_0$ and $\hat{\beta}_1$, to make forecasts that improve on the baseline. The **fitted values,** or **in-sample forecasts,** are

$$\hat{y}_t = \hat{\beta}_0 + \hat{\beta}_1 x_t\,,$$

$t = 1, \ldots, T$.

Forecasts are rarely perfect; instead, we make errors. The **residuals,** or **in-sample forecast errors,** are

$$e_t = y_t - \hat{y}_t\,,$$

---

[10] We speak of the **regression intercept** and the **regression slope.**

[11] Shortly we'll show how to estimate σ^2 as well.

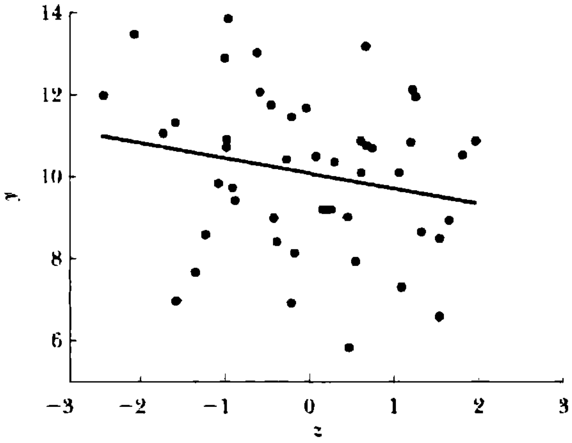


FIGURE 2.3 *Scatterplot of y versus z, Regression Line Superimposed*

$t = 1, \ldots, T$. Forecasters are keenly interested in studying the properties of their forecast errors. Systematic patterns in forecast errors indicate that the forecasting model is inadequate; forecast errors from a good forecasting model must be unforecastable!

Now suppose we have a second exogenous variable, z, which we could also use to forecast y. In Figure 2.3, we show a scatterplot of y against z, with the regression line superimposed. This time the slope of the fitted line is negative. The regressions of y on x and y on z are called **simple linear regressions;** they are potentially useful, but ultimately we'd like to regress y on *both* x and z. Fortunately, the idea of linear regression readily generalizes to accommodate more than one right-hand-side variable. We write

$$y_t = \beta_0 + \beta_1 x_t + \beta_2 z_t + \varepsilon_t \,,$$

$t = 1, \ldots, T$. This is called a **multiple linear regression model.** Again, we use the computer to find

$$\varepsilon_t \overset{\text{iid}}{\sim} (0, \sigma^2) \,,$$

the values of β_0, β_1, and β_2 that produce the best forecast of y; that is, we find the β values that solve the problem $\min_{\beta} \sum_{t=1}^{T} (y_t - \beta_0 - \beta_1 x_t - \beta_2 z_t)^2$, where β denotes the set of three model parameters. We denote the set of estimated parameters by $\hat{\beta}$, with elements $\hat{\beta}_0$, $\hat{\beta}_1$, and $\hat{\beta}_2$. The fitted values are

$$\hat{y}_t = \hat{\beta}_0 + \hat{\beta}_1 x_t + \hat{\beta}_2 z_t \,,$$

and the residuals are

$$e_t = y_t - \hat{y}_t \,,$$

$t = 1, \ldots, T$. Extension to the general multiple linear regression model, with an arbitrary number of right-hand-side variables (k, including the constant), is immediate.

TABLE 2.1
Regression of y on x and z

```
LS // Dependent variable is y.

Sample: 1960 2007
Included observations: 48

Variable      Coefficient     Std. Error    t-Statistic      Prob.
-------------------------------------------------------------------------------
C               9.884732      0.190297      51.94359         .0000
X               1.073140      0.150341      7.138031         .0000
Z              -0.638011      0.172499     -3.698642         .0006

R-squared                      0.552928     Mean dependent var.       10.08241
Adjusted R-squared             0.533059     SD dependent var.         1.908842
SE of regression               1.304371     Akaike info criterion     3.429780
Sum squared resid.             76.56223     Schwarz criterion         3.546730
Log likelihood                -79.31472     F-statistic               27.82752
Durbin-Watson stat.            1.506278     Prob(F-statistic)          .000000
```

This time, let's do more than a simple graphical analysis of the regression fit. Instead, let's look in detail at the computer output, which we show in Table 2.1. We do so dozens of times in this book, and the output format and interpretation are always the same, so it's important to get comfortable with it quickly. The output is in Eviews format. Other software will produce more or less the same information, which is fundamental and standard.

The printout begins by reminding us that we're running a least-squares (LS) regression and that the left-hand-side variable (the "dependent variable"—see the Exercises, Problems, and Complements at the end of this chapter) is y. It then shows us the sample range of the historical data, which happens to be 1960 to 2007, for a total of 48 observations.

Next comes a table listing each right-hand-side variable together with four statistics. The right-hand-side variables x and z need no explanation, but the variable C does. C is notation for the earlier-mentioned constant variable. The C variable always equals 1, so the estimated coefficient on C is the estimated intercept of the regression line.[12]

The four statistics associated with each right-hand-side variable are the estimated coefficient ("Coefficient"), its **standard error** ("Std. Error"), a ***t*-statistic**, and a corresponding **probability value** ("Prob."). The standard errors of the estimated coefficients indicate their likely sampling variability and hence their reliability. The estimated coefficient plus or minus 1 standard error is approximately a 68% confidence interval for the true but unknown population parameter, and the estimated coefficient plus or minus 2 standard errors is approximately a 95% confidence interval, assuming that the estimated coefficient is approximately normally distributed.[13] Thus, large, coefficient standard errors translate into wide confidence intervals.

---

[12] Sometimes the population coefficient on C is called the **constant term**, and the regression estimate, the estimated constant term.

Each *t*-statistic provides a test of the hypothesis of variable irrelevance: that the true but unknown population parameter is 0, so that the corresponding variable contributes nothing to the forecasting regression and can therefore be dropped. One way to test variable irrelevance, with, say, a 5% probability of incorrect rejection, is to check whether 0 is outside the 95% confidence interval for the parameter. If so, we reject irrelevance. The *t*-statistic is just the ratio of the estimated coefficient to its standard error, so if 0 is outside the 95% confidence interval, then the *t*-statistic must be bigger than 2 in absolute value. Thus, we can quickly test irrelevance at the 5% level by checking whether the *t*-statistic is greater than 2 in absolute value.[14]

Finally, associated with each *t*-statistic is a *probability value*, which is the probability of getting a value of the *t*-statistic at least as large in absolute value as the one actually obtained, assuming that the irrelevance hypothesis is true. Hence, if a *t*-statistic were 2, the corresponding probability value would be approximately .05. The smaller the probability value, the stronger the evidence against irrelevance. There's no magic cutoff, but typically probability values less than .1 are viewed as strong evidence against irrelevance, and probability values below .05 are viewed as very strong evidence against irrelevance. Probability values are useful because they eliminate the need for consulting tables of the *t* distribution. Effectively the computer does it for us and tells us the significance level at which the irrelevance hypothesis is just rejected.

Now let's interpret the actual estimated coefficients, standard errors, *t*-statistics, and probability values. The estimated intercept is approximately 10, so that conditional on x and z both being 0, our best forecast of y would be 10. Moreover, the intercept is very precisely estimated, as evidenced by the small standard error of 0.19 relative to the estimated coefficient. An approximate 95% confidence interval for the true but unknown population intercept is $10 \pm 2(0.19)$, or $[9.62, 10.38]$. Zero is far outside that interval, so the corresponding *t*-statistic is huge, with a probability value that's zero to four decimal places.

The estimated coefficient on x is 1.07, and the standard error is again small in relation to the size of the estimated coefficient, so the *t*-statistic is large, and its probability value small. The coefficient is positive, so that y tends to rise when x rises. In fact, the interpretation of the estimated coefficient of 1.07 is that, holding everything else constant, we forecast that a one-unit increase in x will produce a 1.07-unit increase in y.

The estimated coefficient on z is -0.64. Its standard error is larger relative to the estimated parameter, and its *t*-statistic smaller, than those of the other coefficients. The standard error is nevertheless small, and the absolute value of the *t*-statistic is still well above 2, with a small probability value of .06%. Hence, at conventional levels we reject the hypothesis that z contributes nothing to the forecasting regression. The estimated coefficient is negative, so y tends to fall when z rises. We forecast that a one-unit increase in z will produce a 0.64-unit *decrease* in y.

---

[13] The coefficient will be normally distributed if the regression disturbance is normally distributed, or if the sample size is large.

[14] If the sample size is small, or if we want a significance level other than 5%, we must refer to a table of critical values of the *t* distribution. It should also be pointed out that use of the *t* distribution in small samples also requires an assumption of normally distributed disturbances.

A variety of diagnostic statistics follow: they help us to evaluate the adequacy of the regression. We provide detailed discussions of many of them elsewhere. Here we introduce them very briefly.

MEAN DEPENDENT VAR. 10.08

The **sample mean of the dependent variable** is

$$\bar{y} = \frac{1}{T}\sum_{t=1}^{T} y_t \,.$$

It measures the central tendency, or location, of y.

SD DEPENDENT VAR. 1.91

The **sample standard deviation of the dependent variable** is

$$\text{SD} = \sqrt{\frac{\sum_{t=1}^{T}(y_t - \bar{y})^2}{T-1}} \,.$$

It measures the dispersion, or scale, of y.

SUM SQUARED RESID. 76.56

Minimizing the **sum of squared residuals** is the objective of least-squares estimation. It's natural, then, to record the minimized value of the sum of squared residuals. In isolation it's not of much value, but it serves as an input to other diagnostics that we'll discuss shortly. Moreover, it's useful for comparing models and testing hypotheses. The formula is

$$\text{SSR} = \sum_{t=1}^{T} e_t^2 \,.$$

LOG LIKELIHOOD −79.31

The **likelihood function** is the joint density function of the data, viewed as a function of the model parameters. Hence, a natural estimation strategy, called **maximum likelihood** estimation, is to find (and use as estimates) the parameter values that maximize the likelihood function. After all, by construction, those parameter values maximize the likelihood of obtaining the data that were actually obtained. In the leading case of normally distributed regression disturbances, maximizing the likelihood function turns out to be equivalent to minimizing the sum of squared residuals, hence the maximum likelihood parameter estimates are identical to the least-squares parameter estimates. The number reported is the maximized value of the log of the likelihood function.[15] Like the sum of squared residuals, it's not of direct use, but it's useful for comparing models and testing hypotheses. We will rarely use the likelihood function directly; instead, we'll focus for the most part on the sum of squared residuals.

F-STATISTIC 27.83

We use the ***F*-statistic** to test the hypothesis that the coefficients of all variables in the regression except the intercept are jointly 0.[16] That is, we test whether, taken jointly as a set, the variables included in the forecasting model have any predictive value. This contrasts with the *t*-statistics, which we use to examine the predictive worth of the variables one at a time.[17] If no variable has predictive value, the *F*-statistic follows an *F* distribution with $k - 1$ and $T - k$ degrees of freedom. The formula is

$$F = \frac{(\mathrm{SSR}_{\mathrm{res}} - \mathrm{SSR})/(k-1)}{\mathrm{SSR}/(T-k)},$$

where $\mathrm{SSR}_{\mathrm{res}}$ is the sum of squared residuals from a *restricted* regression that contains only an intercept. Thus, the test proceeds by examining how much the SSR increases when all the variables except the constant are dropped. If it increases by a great deal, there's evidence that at least one of the variables has predictive content.

PROB(*F*-STATISTIC) .000000

The probability value for the *F*-statistic, or **Prob(*F*-statistic)**, gives the significance level at which we can just reject the hypothesis that the set of right-hand-side variables has no predictive value. Here, the value is indistinguishable from 0, so we reject the hypothesis overwhelmingly.

---

[15] Throughout this book, *log* refers to a natural (base e) logarithm.

[16] We don't want to restrict the intercept to be 0, because under the hypothesis that all the other coefficients are 0, the intercept would equal the mean of *y*, which in general is not 0.

[17] In the degenerate case of only one right-hand-side variable, the *t*- and *F*-statistics contain exactly the same information, and $F = t^2$. When there are two or more right-hand-side variables, however, the hypotheses tested differ, and $F \neq t^2$.

SE OF REGRESSION 1.30

If we knew the elements of β, then our forecast errors would be the ε_t's, with variance σ^2. We'd like an estimate of σ^2, because it tells us whether our forecast errors are likely to be large or small. The observed residuals, the e_t's, are effectively estimates of the unobserved population disturbances, the ε_t's. Thus the sample variance of the e's, which we denote s^2 (read "s-squared"), is a natural estimator of σ^2:

$$s^2 = \frac{\sum_{t=1}^{T} e_t^2}{T-k}.$$

s^2 is an estimate of the dispersion of the regression disturbance and hence is used to assess goodness of fit of the model, as well as the magnitude of forecast errors that we're likely to make. The larger s^2 is, the worse the model's fit, and the larger the forecast errors we're likely to make. s^2 involves a degrees-of-freedom correction (division by $T-k$ rather than by T or $T-1$), which is an attempt to get a good estimate of the out-of-sample forecast error variance on the basis of the in-sample residuals.

The standard error of the regression (SER) conveys the same information; it's an estimator of σ rather than σ^2, so we simply use s rather than s^2. The formula is

$$\text{SER} = \sqrt{s^2} = \sqrt{\frac{\sum_{t=1}^{T} e_t^2}{T-k}}.$$

The standard error of the regression is easier to interpret than s^2, because its units are the same as those of the e's, whereas the units of s^2 are not. If the e's are in dollars, then the squared e's are in dollars squared, so s^2 is in dollars squared. By taking the square root at the end of it all, SER converts the units back to dollars.

It's often informative to compare the standard error of the regression with the mean of the dependent variable. As a rough rule of thumb, the SER of a good forecasting model shouldn't be more than 10% or 15% of the mean of the dependent variable. For the present model, the SER is about 13% of the mean of the dependent variable, so it just squeaks by.

Sometimes it's informative to compare the standard error of the regression (or a close relative) with the standard deviation of the dependent variable (or a close relative). The *standard error of the regression* is an estimate of the standard deviation of forecast errors from the regression model, and the standard deviation of the dependent variable is an estimate of the standard deviation of the forecast errors from a simpler forecasting model, in which the forecast of each period is simply $\bar{y}$. If the ratio is small, the variables in the model appear very helpful in forecasting y. ***R*-squared** measures, to which we now turn, are based on precisely that idea.

R-SQUARED 0.55

If an intercept is included in the regression, as is almost always the case, *R*-squared must be between 0 and 1. In that case, *R*-squared, usually written R^2, is the percent of the variance of y explained by the variables included in the regression. R^2 measures the in-sample success of the regression equation in forecasting y; hence, it is widely used as a quick check of **goodness of fit**, or forecastability of y based on the variables included in the regression. Here the R^2 is about 55%—good but not great. The formula is

$$R^2 = 1 - \frac{\sum_{t=1}^{T} e_t^2}{\sum_{t=1}^{T} (y_t - \bar{y})^2}.$$

We can write R^2 in a more roundabout way as

$$R^2 = 1 - \frac{\frac{1}{T}\sum_{t=1}^{T} e_t^2}{\frac{1}{T}\sum_{t=1}^{T} (y_t - \bar{y})^2},$$

which makes clear that the numerator in the large fraction is very close to s^2, and the denominator is very close to the sample variance of y.

ADJUSTED *R*-SQUARED 0.53

The interpretation is the same as that of R^2, but the formula is a bit different. **Adjusted R^2** incorporates adjustments for degrees of freedom used in fitting the model, in an attempt to offset the inflated appearance of good fit, or high forecastability of y, if a variety of right-hand-side variables are tried and the "best model" selected. Hence, adjusted R^2 is a more trustworthy goodness-of-fit measure than R^2. As long as there is more than one right-hand-side variable in the model fitted, adjusted R^2 is smaller than R^2; here, however, the two are quite close (53% versus 55%). Adjusted R^2 is often denoted $\bar{R}^2$; the formula is

$$\bar{R}^2 = 1 - \frac{\frac{1}{T-k}\sum_{t=1}^{T} e_t^2}{\frac{1}{T-1}\sum_{t=1}^{T} (y_t - \bar{y})^2},$$

where k is the number of right-hand-side variables, including the constant term. Here the numerator in the large fraction is precisely s^2, and the denominator is precisely the sample variance of y.

AKAIKE INFO CRITERION 3.43

The **Akaike information criterion,** or AIC, is effectively an estimate of the out-of-sample forecast error variance, as is s^2, but it penalizes degrees of freedom more harshly. It is used to select among competing forecasting models. The formula is

$$\mathrm{AIC} = e^{\left(\frac{2k}{T}\right)} \frac{\sum_{t=1}^{T} e_t^2}{T} .$$

SCHWARZ CRITERION 3.55

The **Schwarz information criterion,** or SIC, is an alternative to the AIC with the same interpretation but a still harsher degrees-of-freedom penalty. The formula is

$$\mathrm{SIC} = T^{\left(\frac{k}{T}\right)} \frac{\sum_{t=1}^{T} e_t^2}{T} .$$

As they arise in the course of our discussion, we will discuss in detail the sum of squared residuals, the standard error of the regression, R^2, adjusted R^2, the AIC, and the SIC, the relationships among them, and their role in selecting forecasting models. Thus, we'll say no more here. It is worth noting, however, that other formulas, slightly different from the ones given here, are sometimes used for AIC and SIC, as discussed in greater detail in Chapter 5.

DURBIN-WATSON STAT. 1.51

We mentioned earlier that we're interested in examining whether there are patterns in our forecast errors, because errors from a good forecasting model should be unforecastable. The **Durbin-Watson statistic** tests for correlation over time, called **serial correlation,** in regression disturbances. If the errors made by a forecasting model are serially correlated, then they are forecastable, and we could improve the forecasts by forecasting the forecast errors. The Durbin-Watson test works within the context of the model

$$\begin{aligned} y_t &= \beta_0 + \beta_1 x_t + \beta_2 z_t + \varepsilon_t \\ \varepsilon_t &= \varphi \varepsilon_{t-1} + v_t, \\ v_t &\overset{\text{iid}}{\sim} N(0, \sigma^2). \end{aligned}$$

The regression disturbance is serially correlated when $\varphi \neq 0$. The hypothesis of interest is that $\varphi = 0$. When $\varphi = 0$, the ideal conditions hold, but when $\varphi \neq 0$, the disturbance is serially correlated. More specifically, when $\varphi \neq 0$, we say that ε_t follows an autoregressive process of order 1, or AR(1) for short.[18] If

---

[18] The Durbin-Watson test is designed to be very good at detecting serial correlation of the AR(1) type. Many other types of serial correlation are possible: we'll discuss them extensively in Chapter 8.

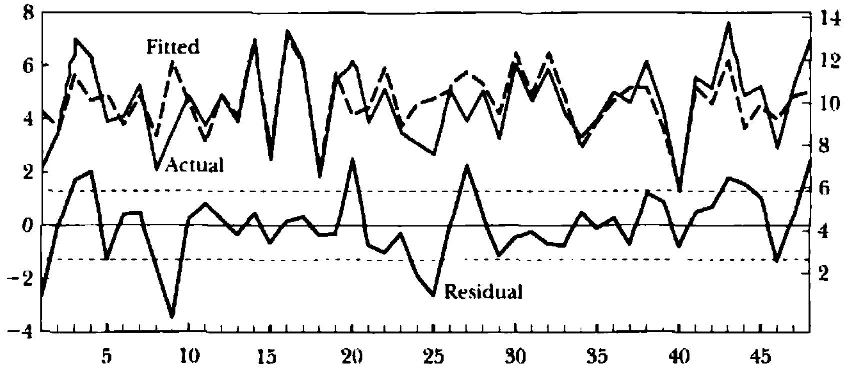


FIGURE 2.4 *Residual Plot, Regression of y on x and z*

$\varphi > 0$, the disturbance is positively serially correlated, and if $\varphi < 0$, the disturbance is negatively serially correlated. **Positive serial correlation** is typically the relevant alternative in the applications that will concern us. The formula for the Durbin-Watson (DW) statistic is

$$DW = \frac{\sum_{t=2}^{T} (e_t - e_{t-1})^2}{\sum_{t=1}^{T} e_t^2}.$$

DW takes values in the interval [0, 4], and if all is well, DW should be around 2. If DW is substantially less than 2, there is evidence of positive serial correlation. As a rough rule of thumb, if DW is less than 1.5, there may be cause for alarm, and we should consult the tables of the DW statistic, available in many statistics and econometrics texts. Here the DW statistic is very close to 1.5. A look at the tables of the DW statistic reveals, however, that we would not reject the null hypothesis at the 5% level.

After running a regression, it's usually a good idea to assess the adequacy of the model by plotting and examining the actual data (y_t's), the fitted values ($\hat{y}_t$'s), and the residuals (e_t's). Often we'll refer to such plots, shown together in a single graph, as a **residual plot.**[19] In Figure 2.4, we show the residual plot for the regression of y on x and z. The actual (short-dashed line) and fitted (long-dashed line) values appear at the top of the graph; their scale is on the right. The fitted values track the actual values fairly well. The residuals appear at the bottom of the graph (solid line); their scale is on the left. It's important to note that the scales differ; the e_t's are in fact substantially smaller and less variable than either the y_t's or the $\hat{y}_t$'s. We draw the zero line through the residuals for visual comparison. There are no obvious patterns in the residuals.

---

[19] Sometimes, however, we'll use *residual plot* to refer to a plot of the residuals alone. The intended meaning will be clear from context.

Exercises, Problems, and Complements

1. (Interpreting distributions and densities) The Sharpe Pencil Company has a strict quality control monitoring program. As part of that program, it has determined that the distribution of the amount of graphite in each batch of 100 pencil leads produced is continuous and uniform between 1 and 2 grams. That is, $f(y) = 1$ for y in $[1, 2]$, and 0 otherwise, where y is the graphite content per batch of 100 leads.
   a. Is y a discrete or continuous random variable?
   b. Is $f(y)$ a probability distribution or a density?
   c. What is the probability that y is between 1 and 2? Between 1 and 1.3? Exactly equal to 1.67?
   d. For high-quality pencils, the desired graphite content per batch is 1.8 grams, with low variation across batches. With that in mind, discuss the nature of the density $f(y)$.

2. (Covariance and correlation) Suppose that the annual revenues of the world's two top oil producers have a covariance of 1,735,492.
   a. Based on the covariance, the claim is made that the revenues are "very strongly positively related." Evaluate the claim.
   b. Suppose instead that, again based on the covariance, the claim is made that the revenues are "positively related." Evaluate the claim.
   c. Suppose you learn that the revenues have a *correlation* of 0.93. In light of that new information, reevaluate the claims in parts a and b.

3. (Conditional expectations versus linear projections) It is important to note the distinction between a conditional mean and a **linear projection.**
   a. The conditional mean is not necessarily a linear function of the conditioning variable(s). In the Gaussian case, the conditional mean is a linear function of the conditioning variables, so it coincides with the linear projection. In non-Gaussian cases, however, linear projections are best viewed as approximations to generally nonlinear conditional mean functions.
   b. The U.S. Congressional Budget Office (CBO) is helping the president to set tax policy. In particular, the president has asked for advice on where to set the average tax rate to maximize the tax revenue collected per taxpayer. For each of 23 countries, the CBO has obtained data on the tax revenue collected per taxpayer and the average tax rate. Is tax revenue likely related to the tax rate? Is the relationship likely linear? (*Hint:* How much revenue would be collected at tax rates of 0% or 100%?) If not, is a linear regression nevertheless likely to produce a good approximation to the true relationship?

4. (Conditional mean and variance) Given the regression model,

$$y_t = \beta_0 + \beta_1 x_t + \beta_2 x_t^2 + \beta_3 z_t + \varepsilon_t$$

$$\varepsilon_t \overset{\text{iid}}{\sim} (0, \sigma^2),$$

   find the mean and variance of y_t conditional on $x_t = x_t^*$ and $z_t = z_t^*$. Does the conditional mean adapt to the conditioning information? Does the conditional variance adapt to the conditioning information?

5. (Scatterplots and regression lines) Draw qualitative scatterplots and regression lines for each of the following two-variable datasets, and state the R^2 in each case:

a. Dataset 1: y and x have correlation 1.
b. Dataset 2: y and x have correlation -1.
c. Dataset 3: y and x have correlation 0.

6. (Desired values of regression diagnostic statistics) For each of the diagnostic statistics listed here, indicate whether, other things the same, "bigger is better," "smaller is better," or neither. Explain your reasoning. (*Hint:* Be careful, think before you answer, and be sure to qualify your answers as appropriate.)
   a. Coefficient
   b. Standard error
   c. t-statistic
   d. Probability value of the t-statistic
   e. R^2
   f. Adjusted R^2
   g. Standard error of the regression
   h. Sum of squared residuals
   i. Log likelihood
   j. Durbin-Watson statistic
   k. Mean of the dependent variable
   l. Standard deviation of the dependent variable
   m. Akaike information criterion
   n. Schwarz information criterion
   o. F-statistic
   p. Probability value of the F-statistic

7. (Mechanics of fitting a linear regression) On the book's web page, you will find a second set of data on y, x, and z similar to, but different from, the data that underlie the analysis performed in this chapter. Using the new data, repeat the analysis and discuss your results.

8. (Regression with and without a constant term) Consider Figure 2.2, in which we showed a scatterplot of y versus x with a fitted regression line superimposed.
   a. In fitting that regression line, we included a constant term. How can you tell?
   b. Suppose that we had not included a constant term. How would the figure look?
   c. We almost always include a constant term when estimating regressions. Why?
   d. When, if ever, might you explicitly want to exclude the constant term?

9. (Interpreting coefficients and variables) Let $y_t = \beta_0 + \beta_1 x_t + \beta_2 z_t + \varepsilon_t$, where y_t is the number of hot dogs sold at an amusement park on a given day, x_t is the number of admission tickets sold that day, z_t is the daily maximum temperature, and ε_t is a random error.
   a. State whether each of y_t, x_t, z_t, β_0, β_1, and β_2 is a coefficient or a variable.
   b. Determine the units of β_0, β_1, and β_2, and describe the physical meaning of each.
   c. What does the sign of a coefficient tell you about how its corresponding variable affects the number of hot dogs sold? What are your expectations for the signs of the various coefficients—negative, zero, positive, or unsure?
   d. Is it sensible to entertain the possibility of a nonzero intercept—that is, $\beta_0 \neq 0$? $\beta_0 > 0$? $\beta_0 < 0$?

10. (Nonlinear least squares) The least-squares estimator discussed in this chapter is often called "ordinary" least squares. The adjective *ordinary* distinguishes the ordinary least-squares estimator from fancier estimators, such as the nonlinear

least-squares estimator. When we estimate by **nonlinear least squares,** we use a computer to find the minimum of the sum of squared residual function directly, using numerical methods. For the simple regression model discussed in this chapter, ordinary and nonlinear least squares produce the same result, and ordinary least squares is simpler to implement, so we prefer ordinary least squares. As we will see, however, some intrinsically nonlinear forecasting models can't be estimated using ordinary least squares but can be estimated using nonlinear least squares. We use nonlinear least squares in such cases.

For each of the following models, determine whether ordinary least squares may be used for estimation (perhaps after transforming the data).

a. $y_t = \beta_0 + \beta_1 x_t + \varepsilon_t$
b. $y_t = \beta_0 e^{\beta_1 x_t} \varepsilon_t$
c. $y_t = \beta_0 + e^{\beta_1 x_t} + \varepsilon_t$

11. (Regression semantics) Regression analysis is so important, and used so often by so many people, that a variety of associated terms have evolved over the years, all of which are the same for our purposes. You may encounter them in your reading, so it's important to be aware of them. Some examples:
    a. Ordinary least squares, least squares, OLS, LS
    b. *y*, left-hand-side variable, regressand, dependent variable, endogenous variable
    c. *x*'s, right-hand-side variables, regressors, independent variables, exogenous variables, predictors
    d. Probability value, prob-value, *p*-value, marginal significance level
    e. Schwarz criterion, Schwarz information criterion, SIC, Bayes information criterion, BIC

Bibliographical and Computational Notes

See any good introductory statistics or econometrics book for much more thorough discussions of probability, statistics, and regression and for tables of significance points of the normal, *t*, *F*, and Durbin-Watson distributions. Possibilities include Anderson, Sweeney, and Williams (2006), Maddala (2001), Pindyck and Rubinfeld (1997), and Wonnacott and Wonnacott (1990).

The Jarque-Bera test is developed in Jarque and Bera (1987).

Dozens of software packages—including spreadsheets—implement various statistical and linear regression analyses. Most automatically include an intercept in linear regressions unless explicitly instructed otherwise. That is, they automatically create and include a *C* variable.

Concepts for Review

Discrete random variable
Discrete probability distribution
Continuous random variable
Probability density function
Moment
Mean
Expected value
Location
Central tendency
Variance
Dispersion
Scale

Standard deviation
Skewness
Asymmetry
Kurtosis
Leptokurtosis
Normal distribution
Gaussian distribution
Marginal distribution
Joint distribution
Covariance
Correlation
Conditional distribution
Conditional moment
Conditional mean
Conditional variance
Population distribution
Sample
Estimator
Statistic
Sample statistic
Sample mean
Sample variance
Sample standard deviation
Sample skewness
Sample kurtosis
χ^2 distribution
t distribution
F distribuition
Jarque-Bera test
Regression analysis
Least squares
Disturbance
Parameter
Regression function
Conditional expectation
Fitted value
In-sample forecast
Residual
In-sample forecast error
Regression intercept
Regression slope
Simple linear regression
Multiple linear regression model
Standard error
t-statistic
Probability value
Constant term
Sample mean of the dependent variable
Sample standard deviation of the dependent variable
Sum of squared residuals
Likelihood function
Maximum likelihood estimation
F-statistic
Prob(F-statistic)
Standard error of the regression
R^2
Goodness of fit
Adjusted R^2
Akaike information criterion
Schwarz information criterion
Durbin-Watson statistic
Serial correlation
Positive serial correlation
Residual plot
Linear projection
Nonlinear least squares

References and Additional Readings

Anderson, D. R., Sweeney, D. J., and Williams, T. A. (2006). *Statistics for Business and Economics.* 4th ed. Cincinnati: South-Western.

Jarque, C. M., and Bera, A. K. (1987). "A Test for Normality of Observations and Regression Residuals." *International Statistical Review,* 55, 163–172.

Maddala, G. S. (2001). *Introduction to Econometrics.* 3rd ed. New York: Macmillan.

Pindyck, R. S., and Rubinfeld, D. L. (1997). *Econometric Models and Economic Forecasts,* 4th ed. New York: McGraw-Hill.

Wonnacott, T. H., and Wonnacott, R. J. (1990). *Introductory Statistics.* 5th ed. New York: Wiley.

Six Considerations Basic to Successful Forecasting

In Chapter 1, we sketched a variety of areas where forecasts are used routinely, and we took a brief tour of the basic forecasting tools that you'll master as you progress through this book. Now let's back up and consider six types of questions that are relevant for *any* forecasting task.[1]

- **Decision environment** and **loss function.** What decision will the forecast guide, and what are the implications for the design, use, and evaluation of the forecasting model? Related, how do we quantify what we mean by a "good" forecast and, in particular, the cost or loss associated with forecast errors of various signs and sizes? How should we define optimality of a forecast in a particular situation? How do we compute optimal forecasts?
- **Forecast object.** What is the object that we need to forecast? Is it a time series, such as sales of a firm recorded over time, or an event, such as devaluation of a currency? And what is the quantity and quality of the data? How long is the sample of available data? Are we forecasting one object or many (such as sales of each of 350 products)? Are there missing observations? Unusual observations?
- **Forecast statement.** How do we wish to state our forecasts? If, for example, the object to be forecast is a time series, are we interested in a single "best

---

[1] There are of course many possible variations, combinations, and extensions of the questions; you should try to think of some as you read through them.

guess" forecast, a "reasonable range" of possible future values that reflects the underlying uncertainty associated with the forecasting problem, or a probability distribution of possible future values? What are the associated costs and benefits?

- **Forecast horizon.** What is the forecast horizon of interest, and what determines it? Are we interested, for example, in forecasting 1 month ahead, 1 year ahead, or 10 years ahead? The best modeling and forecasting strategy will likely vary with the horizon.
- **Information set.** On what information will the forecast be based? Are the available data simply the past history of the series to be forecast, or are other series available that may be related to the series of interest?
- **Methods and complexity,** the **parsimony principle,** and the **shrinkage principle.** What forecasting method is best suited to the needs of a particular forecasting problem? How complex should the forecasting model be? More generally, what sorts of models, in terms of complexity, tend to do best for forecasting in business, finance, economics, and government? The phenomena that we model and forecast are often tremendously complex, but does it necessarily follow that our forecasting models should be complex?

1. The Decision Environment and Loss Function

Forecasts are not made in a vacuum. The key to generating good and useful forecasts, which we will stress now and throughout, is recognizing that forecasts are made to guide decisions. The link between forecasts and decisions sounds obvious—and it is—but it's worth thinking about in some depth. Forecasts are made in a wide variety of situations, but in every case forecasts are of value because they aid in decision making. Quite simply, good forecasts help to produce good decisions. Recognition and awareness of the decision-making environment is the key to effective design, use, and evaluation of forecasting models.

Consider the following stylized problem: You have started a firm and must decide how much inventory to hold going into the next sales period. If you knew that demand would be high next period, then you'd like to have a lot of inventory on hand. If you knew that demand would be slack, then you would like to deplete your inventories because it costs money to store unnecessary inventories. Of course, the problem is that you don't know next period's demand, and you've got to make your inventory stocking decision *now*!

There are four possible combinations of inventory decisions and demand outcomes: in two we make the correct decision, and in two we make the incorrect decision. We show the four possible outcomes in Table 3.1. Each entry of the table contains a "cost" or "loss" to you corresponding to the associated decision/outcome pair. The good pairs on the diagonal have zero loss—you did the right thing, building inventory when demand turned out to be high or

TABLE 3.1
Decision Making with Symmetric Loss

	Demand High	Demand Low
Build Inventory	0	$10,000
Reduce Inventory	$10,000	0

contracting inventory when demand turned out to be low. The bad pairs off the diagonal have positive loss—you did the wrong thing, building inventory when demand turned out to be low or contracting inventory when demand turned out to be high.

In Table 3.1, the loss associated with each incorrect decision is $10,000. We call such a loss structure *symmetric*, because the loss is the same for both of the bad outcomes. In many important decision environments, a **symmetric loss** structure closely approximates the true losses of the forecaster. In other decision environments, however, symmetric loss may *not* be realistic; in general, there's no reason for loss to be symmetric.

In Table 3.2, we summarize a decision environment with an **asymmetric loss** structure. As before, each entry of the table contains a loss corresponding to the associated decision/outcome pair. The good pairs on the diagonal have zero loss for the same reason as before—when you do the right thing, you incur no loss. The bad pairs off the diagonal again have positive loss—when you do the wrong thing, you suffer—but now the amount of the loss differs depending on what sort of mistake you make. If you reduce inventories and demand turns out to be high, then you have insufficient inventories to meet demand, and you miss out on a lot of business, which is very costly ($20,000). If you build inventories and demand turns out to be low, then you must carry unneeded inventories, which is not as costly ($10,000).

To recap: For every decision-making problem, there is an associated *loss structure;* for each decision/outcome pair, there is an associated loss. We can think of zero loss as associated with the correct decision and positive loss as associated with the incorrect decision.

Recall that *forecasts* are made to help guide *decisions*. Thus, the loss structure associated with a particular decision induces a similar loss structure for forecasts used to inform that decision. Continuing with our example, we might forecast sales to help us decide whether to build or reduce inventory, and the loss we incur depends on the divergence between actual and predicted sales. To keep things simple, imagine that sales forecasts and sales realizations are either "high" or "low." Table 3.3 illustrates a symmetric forecasting loss structure, and

TABLE 3.2
Decision Making with Asymmetric Loss

	Demand High	Demand Low
Build Inventory	0	$10,000
Reduce Inventory	$20,000	0

TABLE 3.3
Forecasting with Symmetric Loss

	High Actual Sales	Low Actual Sales
High Forecasted Sales	0	$10,000
Low Forecasted Sales	$10,000	0

Table 3.4 illustrates an asymmetric forecasting loss structure. Note that a forecast of high sales implies the decision "build inventory" (likewise for low sales and "reduce inventory"); thus, we derive the loss structure associated with a forecast from the loss structure of decisions based on the forecasts.

This example is highly simplified: Forecasts are either "up" or "down," and realizations are similarly "up" or "down." In the important case of time series forecasting, both the forecast and the realization can typically assume a continuous range of values, so a more general notion of loss function is needed.

Let y denote a series and $\hat{y}$ its forecast. The corresponding **forecast error,** e, is the difference between the realization and the previously made forecast:

$$e = y - \hat{y}\,.$$

We consider loss functions of the form $L(e)$. This means that the loss associated with a forecast depends only on the size of the forecast error. We require the loss function $L(e)$ to satisfy three conditions:

- $L(0) = 0$. That is, no loss is incurred when the forecast error is 0. (A 0 forecast error, after all, corresponds to a perfect forecast!)
- $L(e)$ is continuous. That is, nearly identical forecast errors should produce nearly identical losses.
- $L(e)$ is increasing on each side of the origin. That is, the bigger the absolute value of the error, the bigger the loss.

Apart from these three requirements, we impose no restrictions on the form of the loss function.

The **quadratic loss** function is tremendously important in practice, both because it is often an adequate approximation to realistic loss structures and because it is mathematically convenient. Quadratic loss is given by

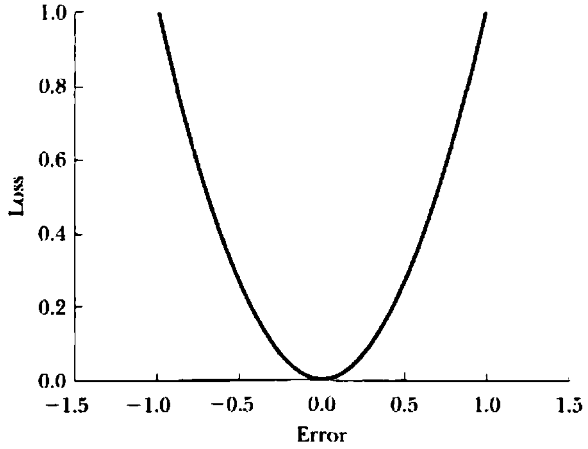
$$L(e) = e^2,$$

and we graph it as a function of the forecast error in Figure 3.1. Because of the squaring associated with the quadratic loss function, it is symmetric around the origin; in addition, it increases at an increasing rate on each side of the origin, so that large errors are penalized *much* more severely than small ones.

TABLE 3.4
Forecasting with Asymmetric Loss

	High Actual Sales	Low Actual Sales
High Forecasted Sales	0	$10,000
Low Forecasted Sales	$20,000	0

FIGURE 3.1
Quadratic Loss



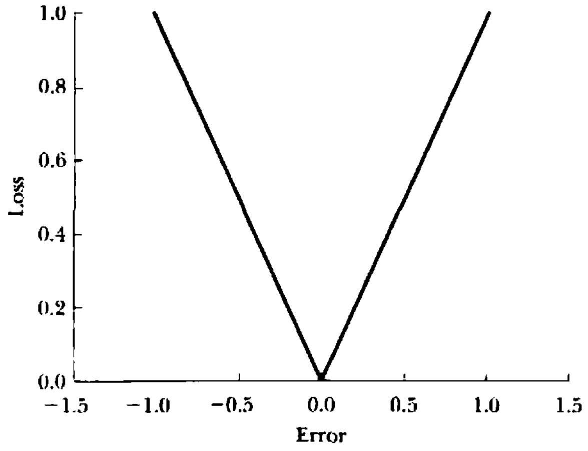
Another important symmetric loss function is **absolute loss,** or **absolute error loss,** given by

$$L(e) = |e| \, .$$

Like quadratic loss, absolute loss is increasing on each side of the origin, but loss increases at a constant (linear) rate with the size of the error. We illustrate absolute loss in Figure 3.2.

In certain contexts, symmetric loss functions may not be an adequate distillation of the forecast/decision environment. In Figure 3.3, for example, we show a particular asymmetric loss function for which negative forecast errors are less costly than positive errors.

FIGURE 3.2
Absolute Loss



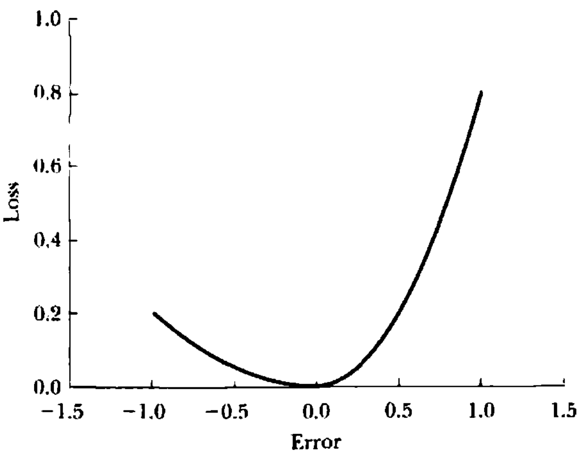


FIGURE 3.3
Asymmetric Loss

In some situations, even the $L(e)$ form of the loss function is too restrictive. Although loss will always be of the form $L(y, \hat{y})$, there's no reason that y and $\hat{y}$ should necessarily enter as $y - \hat{y}$. In predictions of financial asset returns, for example, interest sometimes focuses on direction of change. A **direction-of-change forecast** takes one of two values—up or down. The loss function associated with a direction of change forecast might be[2]

$$L(y, \hat{y}) = \begin{cases} 0, & \text{if} \quad \text{sign}(\Delta y) = \text{sign}(\Delta \hat{y}) \\ 1, & \text{if} \quad \text{sign}(\Delta y) \neq \text{sign}(\Delta \hat{y}) \,. \end{cases}$$

With this loss function, if you predict the direction of change correctly, you incur no loss; but if your prediction is wrong, you're penalized.

Much of this book is about how to produce **optimal forecasts.** What precisely do we mean by an optimal forecast? That's where the loss function comes in. We'll work with a wide class of symmetric loss functions, and we'll learn how to produce forecasts that are optimal in the sense that they minimize expected loss for any such loss function.[3]

2. The Forecast Object

There are many objects that we might want to forecast. In business and economics, the forecast object is typically one of three types: event outcome, event timing, or time series.

---

[2] The operator Δ means "change." Thus, Δy_t is the change in y from period $t-1$ to period t, or $y_t - y_{t-1}$.

[3] As noted earlier, not all relevant loss functions need be symmetric. Symmetric loss, however, is usually a reasonable approximation, and symmetric loss is used routinely for practical forecasting.

Event outcome forecasts are relevant to situations in which an event is certain to take place at a given time but the outcome is uncertain. For example, many people are interested in whether the current chairman of the Board of Governors of the U.S. Federal Reserve System will eventually be reappointed. The "event" is the reappointment decision; the decision will occur at the end of the term. The outcome of this decision is confirmation or denial of the reappointment.

Event timing forecasts are relevant when an event is certain to take place and the outcome is known, but the timing is uncertain. A classic example of an event timing forecast concerns business cycle turning points. There are two types of turning points: peaks and troughs. A peak occurs when the economy moves from expansion into recession, and a trough occurs when the economy moves from recession into expansion. If, for example, the economy is currently in an expansion, then there is no doubt that the next turning point will be a peak, but there is substantial uncertainty as to its *timing*. Will the peak occur this quarter, this year, or 10 years from now?

Time series forecasts are relevant when the future value of a time series is of interest and must be projected. As we'll see, there are many ways to make such forecasts, but the basic forecasting setup doesn't change much. Based on the history of the time series (and possibly a variety of other types of information as well, such as the histories of related time series or subjective considerations), we want to project future values of the series. For example, we may have data on the number of Apple computers sold in Germany in each of the last 60 months, and we may want to use that data to forecast the number of Apple computers to be sold in Germany in each month of the next year.

Time series forecasts are by far the most frequently encountered in practice for at least two reasons. First, most business, economic, and financial data are time series; thus, the general scenario of projecting the future of a series for which we have historical data arises constantly. Second, the technology for making and evaluating time series forecasts is well developed, and the typical time series forecasting scenario is precise, so time series forecasts can be made and evaluated routinely. In contrast, the situations associated with event outcome and event timing forecasts arise less frequently and are often less amenable to quantitative treatment.

3. The Forecast Statement

When we make a forecast, we must decide whether the forecast will be (1) a single number (a "best guess"), (2) a range of numbers, into which the future value can be expected to fall a certain percentage of the time, or (3) an entire probability distribution for the future value. In short, we need to decide on the forecast type.

More precisely, we must decide whether the forecast will be (1) a point forecast, (2) an interval forecast, or (3) a density forecast. A **point forecast** is a single number. For example, one possible point forecast of the growth rate of the total number of web pages over the next year might be +23.3%; likewise, a point forecast of the growth rate of U.S. real GDP over the next year might be +1.3%. Point forecasts are made routinely in numerous applications, and the methods used to construct them vary in difficulty from simple to sophisticated. The defining characteristic of a point forecast is simply that it is a single number.

A good point forecast provides a simple and easily digested guide to the future of a time series. However, random and unpredictable "shocks" affect all of the series that we forecast. As a result of such shocks, we expect nonzero forecast errors, even from very good forecasts. Thus, we may want to know the degree of confidence we have in a particular point forecast. Stated differently, we may want to know how much uncertainty is associated with a particular point forecast. The uncertainty surrounding point forecasts suggests the usefulness of an interval forecast.

An **interval forecast** is not a single number; rather, it is a range of values in which we expect the realized value of the series to fall with some (prespecified) probability.[4] Continuing with our examples, a 90% interval forecast for the growth rate of web pages might be the interval [11.3%, 35.3%] (23.3% ± 12%). That is, the forecast states that with probability 90%, the future growth rate of web pages will be in the interval [11.3%, 35.3%]. Similarly, a 90% interval forecast for the growth rate of U.S. real GDP might be [−2.3%, 4.3%] (1.3% ± 3%); that is, the forecast states that with probability 90% the future growth rate of U.S. real GDP will be in the interval [−2.3%, 4.3%].

A number of remarks are in order regarding interval forecasts. First, the length (size) of the intervals conveys information regarding forecast uncertainty. The GDP growth rate interval is much shorter then the web page growth rate interval; this reflects the fact that there is less uncertainty associated with the real GDP growth rate forecast than the web page growth rate forecast. Second, interval forecasts convey more information than point forecasts. Given an interval forecast, you can construct a point forecast by using the midpoint of the interval.[5] Conversely, given only a point forecast, there is no way to infer an interval forecast.

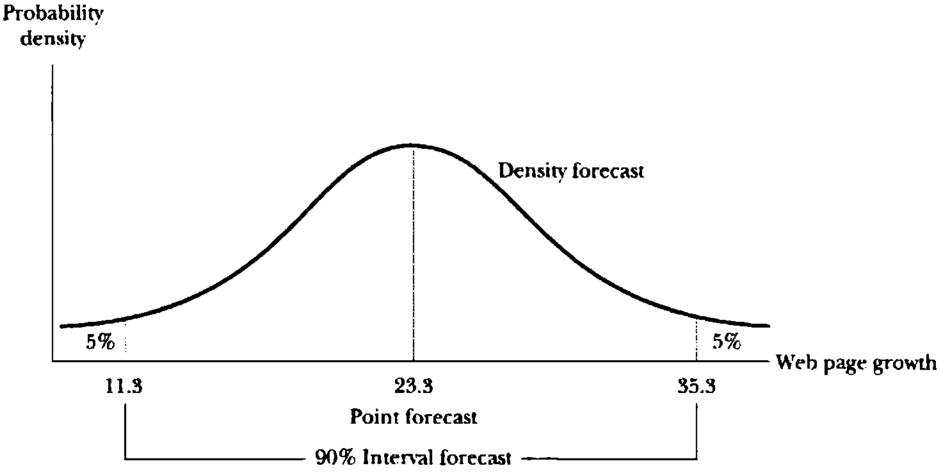
Finally, we consider density forecasts. A **density forecast** gives the entire density (or probability distribution) of the future value of the series of interest. For example, the density forecast of future web page growth might be normally distributed with a mean of 23.3% and a standard deviation of 7.32%.

---

[4] An interval forecast is very similar to the more general idea of a *confidence interval* that you studied in statistics. An interval forecast is simply a confidence interval for the true (but unknown) future value of a series, computed using a sample of historical data. We'll say that [a, b] is a $100(1-\alpha)\%$ interval forecast if the probability of the future value being less than a is $\alpha/2$, and the probability of the future value being greater than b is also $\alpha/2$.

[5] An interval forecast doesn't *have* to be symmetric around the point forecast, so that we wouldn't *necessarily* infer a point forecast as the midpoint of the interval forecast, but in many cases such a procedure is appropriate.

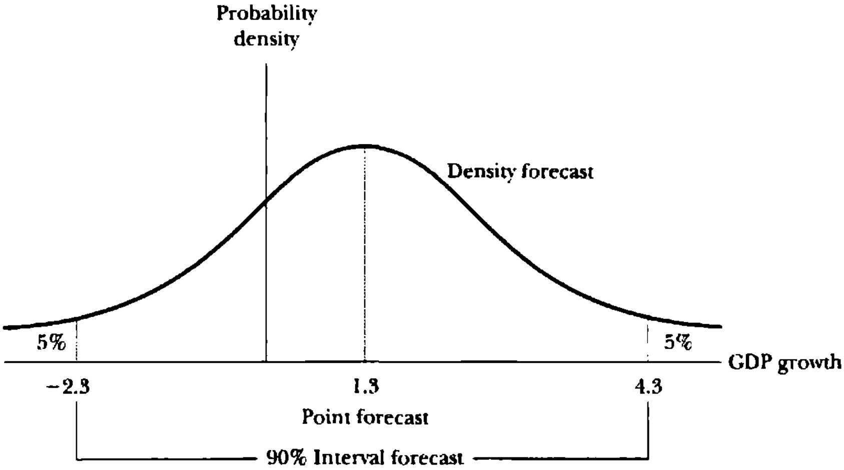
FIGURE 3.4
Web Page Growth: Point, Interval, and Density Forecasts


Probability density
Density forecast
5%
5%
Web page growth
11.3
23.3
35.3
Point forecast
90% Interval forecast

Likewise, the density forecast of future real GDP growth might be normally distributed with a mean of 1.3% and a standard deviation of 1.83%.

As with interval forecasts, density forecasts convey more information than point forecasts. Density forecasts also convey more information than interval forecasts, because given a density, interval forecasts at any desired confidence level are readily constructed. For example, if the future value of some series x is distributed as $N(\mu, \sigma^2)$ then a 95% interval forecast of x is $\mu \pm 1.96\sigma$, a 90% interval forecast of x is $\mu \pm 1.64\sigma$, and so forth. Continuing with our example, the relationships among density, interval, and point forecasts are made clear in Figure 3.4 (web page growth) and Figure 3.5 (U.S. real GDP growth).

FIGURE 3.5
U.S. Real GDP Growth: Point, Interval, and Density Forecasts


Probability density
Density forecast
5%
5%
GDP growth
−2.3
1.3
4.3
Point forecast
90% Interval forecast

To recap, there are three time series forecast types: point, interval, and density. Density forecasts convey more information than interval forecasts, which in turn convey more information than point forecasts. This may seem to suggest that density forecasts are always the preferred forecast, that density forecasts are the most commonly used forecasts in practice, and that we should focus most of our attention in this book on density forecasts.

In fact, the opposite is true. Point forecasts are the most commonly used forecasts in practice, interval forecasts are a rather distant second, and density forecasts are rarely made—for at least two reasons. First, the construction of interval and density forecasts requires either (1) additional and possibly incorrect assumptions relative to those required for construction of point forecasts or (2) advanced and computer-intensive methods involving—for example, extensive simulation. Second, point forecasts are often easier to understand and act on than interval or density forecasts. That is, the extra information provided by interval and density forecasts is not necessarily an advantage when information processing is costly.

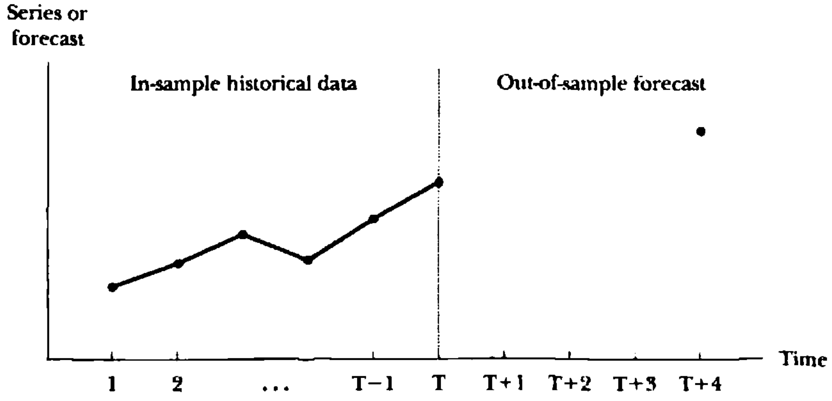
Thus far, we have focused exclusively on types of time series forecasts, because time series are so prevalent and important in numerous fields. It is worth mentioning another forecast type of particular relevance to event outcome and event timing forecasting, the **probability forecast.** To understand the idea of a probability forecast, consider forecasting which of two politicians, Mr. Liar or Ms. Cheat, will win an election. (This is an event outcome forecasting situation.) If our calculations tell us that the odds favor Mr. Liar, we might issue the forecast simply as "Mr. Liar will win." This is roughly analogous to the time series point forecasts discussed earlier, in the sense that we're not reporting any measure of the *uncertainty* associated with our forecast. Alternatively, we could report the probabilities associated with each of the possible outcomes; for example, "Mr. Liar will win with probability .6, and Ms. Cheat will win with probability .4." This is roughly analogous to the time series interval or density forecasts discussed earlier, in the sense that it explicitly quantifies the uncertainty associated with the future event with a probability distribution.

Event outcome and timing forecasts, although not as common as time series forecasts, do nevertheless arise in certain important situations and are often stated as probabilities. For example, when a bank assesses the probability of default on a new loan or a macroeconomist assesses the probability that a business cycle turning point will occur in the next 6 months, the banker or macroeconomist will often use a probability forecast.

4. The Forecast Horizon

The forecast horizon is defined as the number of periods between today and the date of the forecast we make. For example, if we have annual data, and it's now year T, then a forecast of GDP for year $T + 2$ has a forecast horizon of 2 steps. The meaning of a step depends on the frequency of observation of the

FIGURE 3.6
4-Step-Ahead Point Forecast



data. For monthly data, a step is one month; for quarterly data, a step is one quarter (3 months); and so forth. In general, we speak of an **h-step-ahead forecast,** where the horizon h is at the discretion of the user.[6]

The horizon is important for at least two reasons. First, of course, the forecast changes with the forecast horizon. Second, the best forecasting model will often change with the forecasting horizon as well. All of our forecasting models are approximations to the underlying dynamic patterns in the series we forecast; there's no reason why the best approximation for one purpose (such as short-term forecasting) should be the same as the best approximation for another purpose (such as long-term forecasting).

In closing this section, let's distinguish between what we've called h-step-ahead forecasts and what we'll call **h-step-ahead extrapolation forecasts.** In h-step-ahead forecasts, the horizon is always fixed at the same value, h. For example, every month we might make a 4-month-ahead forecast. Alternatively, in extrapolation forecasts, the horizon includes all steps from 1-step-ahead to h-steps-ahead. There's nothing particularly deep or difficult about the distinction, but it's useful to make it, and we'll use it subsequently.

Suppose, for example, that you observe a series from some initial time 1 to some final time T, and you plan to forecast the series.[7] We illustrate the difference between h-step-ahead and h-step-ahead extrapolation forecasts in Figures 3.6 and 3.7. In Figure 3.6, we show a 4-step-ahead point forecast; in Figure 3.7, we show a 4-step-ahead extrapolation point forecast. The

---

[6] The choice of h depends on the decision that the forecast will guide. The nature of the decision environment typically dictates whether short-term, medium-term, or long-term forecasts are needed.

[7] For a sample of data on a series y, we'll typically write $\{y_t\}_{t=1}^{T}$. This notation means, "We observe the series y from some beginning time $t = 1$ to some ending time $t = T$."

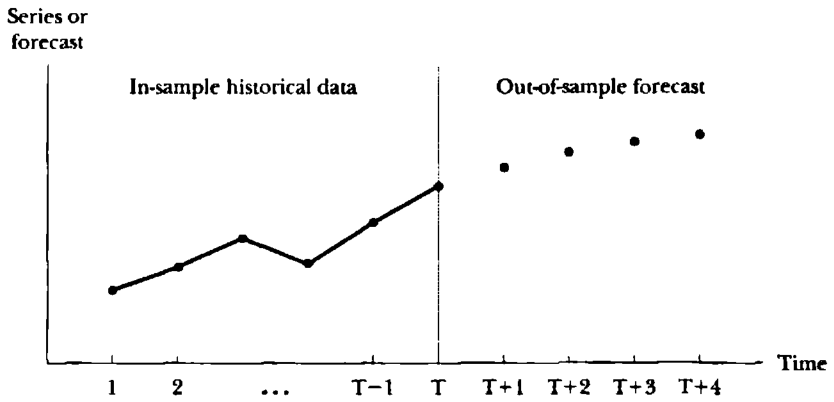


FIGURE 3.7
4-Step-Ahead Extrapolation Point Forecast

extrapolation forecast is nothing more than a set consisting of 1-, 2-, 3-, and 4-step-ahead forecasts.

5. The Information Set

The quality of our forecasts is limited by the quality and quantity of information available when forecasts are made. Any forecast we produce is conditional on the information used to produce it, whether explicitly or implicitly.

The idea of an information set is fundamental to constructing good forecasts. In forecasting a series, y, using historical data from time 1 to time T, sometimes we use the univariate information set, which is the set of historical values of y up to and including the present,

$$\Omega_T^{\text{univariate}} = \{y_T, y_{T-1}, \ldots, y_1\}.$$

Alternatively, sometimes we use the multivariate information set

$$\Omega_T^{\text{multivariate}} = \{y_T, x_T, y_{T-1}, x_{T-1}, \ldots, y_1, x_1\},$$

where the x's are a set of additional variables potentially related to y. Regardless, it's always important to think hard about what information is available, what additional information could be collected or made available, the form of the information (for example, quantitative or qualitative), and so on.

The idea of an information set is also fundamental for evaluating forecasts. When evaluating a forecast, we're sometimes interested in whether the forecast could be improved by using a given set of information more efficiently, and we're sometimes interested in whether the forecast could be improved by using more information. Either way, the ideas of information and information sets play crucial roles in forecasting.

6. Methods and Complexity, the Parsimony Principle, and the Shrinkage Principle

It's crucial to tailor forecasting tools to forecasting tasks, and doing so is partly a matter of judgment. Typically the specifics of the situation (such as decision environment, forecast object, forecast statement, forecast horizon, information set, and so forth) will indicate the desirability of a specific method or modeling strategy. Moreover, as we'll see, formal statistical criteria exist to guide model selection within certain classes of models.

We've stressed that a variety of forecasting applications use a small set of common tools and models. You might guess that those models are tremendously complex, because of the obvious complexity of the real-world phenomena that we seek to forecast. Fortunately, such is not the case. In fact, decades of professional experience suggest just the opposite: Simple, parsimonious models tend to be best for out-of-sample forecasting in business, finance, and economics. Hence, the *parsimony principle:* Other things being the same, simple models are usually preferable to complex models.

A number of reasons explain why smaller, simpler models are often more attractive than larger, more complicated ones. First, by virtue of their parsimony, we can estimate the parameters of simpler models more precisely. Second, because simpler models are more easily interpreted, understood, and scrutinized, anomalous behavior is more easily spotted. Third, it's easier to communicate an intuitive feel for the behavior of simple models, which makes them more useful in the decision-making process. Finally, enforcing simplicity lessens the scope for "data mining"—tailoring a model to maximize its fit to historical data. Data mining often results in models that fit historical data beautifully (by construction) but perform miserably in out-of-sample forecasting, because it tailors models in part to the *idiosyncracies* of historical data, which have no relationship to unrealized future data.

The parsimony principle is related to, but distinct from, the *shrinkage principle,* which codifies the idea that imposing restrictions on forecasting models often improves forecast performance. The name *shrinkage* comes from the notion of coaxing, or "shrinking," forecasts in certain directions by imposing restrictions of various sorts on the models used to produce the forecasts.[8] The reasoning behind the shrinkage principle is subtle, but it permeates forecasting. By the time you've completed this book, you'll have a firm grasp of it.

Finally, note that simple models should not be confused with naive models. All of this is well formalized in the **KISS principle** (appropriately modified for forecasting): "Keep It Sophisticatedly Simple." We'll attempt to do so throughout.

---

[8] One such possible restriction is that, loosely speaking, forecasting models be simple, hence the link to the parsimony principle.

7. Concluding Remarks

This chapter, like Chapter 1, deals with broad issues of general relevance. For the most part, it avoids detailed discussion of specific modeling or forecasting techniques. In the next chapter, we begin to change the mix toward specific tools with specific applications. In the broad-brush tradition of Chapters 1, 2, and 3, we focus on principles of statistical graphics, which are relevant in any forecasting situation, but we also introduce a variety of specific graphical techniques, which are useful in a variety of situations.

Exercises, Problems, and Complements

1. (Data and forecast timing conventions) Suppose that, in a particular monthly dataset, time $t = 10$ corresponds to September 1960.
   a. Name the month and year of each of the following times: $t + 5$, $t + 10$, $t + 12$, $t + 60$.
   b. Suppose that a series of interest follows the simple process $y_t = y_{t-1} + 1$, for $t = 1, 2, 3, \ldots$, meaning that each successive month's value is one higher than the previous month's. Suppose that $y_0 = 0$, and suppose that at present $t = 10$. Calculate the forecasts $y_{t+5,t}$, $y_{t+10,t}$, $y_{t+12,t}$, $y_{t+60,t}$ where, for example, $y_{t+5,t}$ denotes a forecast made at time t for future time $t + 5$, assuming that $t = 10$ at present.

2. (Properties of loss functions) State whether the following potential loss functions meet the criteria introduced in the text and, if so, whether they are symmetric or asymmetric:
   a. $L(e) = e^2 + e$
   b. $L(e) = e^4 + 2e^2$
   c. $L(e) = 3e^2 + 1$
   d. $L(e) = \begin{cases} \sqrt{e} & \text{if} \quad e > 0 \\ |e| & \text{if} \quad e \leq 0 \end{cases}$.

3. (Relationships among point, interval, and density forecasts) For each of the following density forecasts, how might you infer "good" point and 90% interval forecasts? Conversely, if you started with your point and interval forecasts, could you infer "good" density forecasts? Be sure to defend your definition of "good."
   a. Future y is distributed as $N(10, 2)$.
   b. $P(y) = \begin{cases} \dfrac{y-5}{25} & \text{if} \quad 5 < y < 10 \\ -\dfrac{y-15}{25} & \text{if} \quad 10 < y < 15 \\ 0 & \text{otherwise} \end{cases}$.

4. (Forecasting at short through long horizons) Consider the claim "The distant future is harder to forecast than the near future." Is it sometimes true? Usually true? Always true? Why or why not? Discuss in detail. Be sure to define "harder."

5. (Forecasting as an ongoing process in organizations) We could add another very important item to this chapter's list of considerations basic to successful forecasting: Forecasting in organizations is an ongoing process of building, using, evaluating, and improving forecasting models. Provide a concrete example of a forecasting model used in business, finance, economics, or government, and discuss ways in which each of the following questions might be resolved prior to, during, or after its construction.
   a. Are the data "dirty"? For example, are there "**ragged edges**"? That is, do the starting and ending dates of relevant series differ? Are there **missing observations?** Are there aberrant observations, called **outliers,** perhaps due to **measurement error?** Are the data stored in a format that inhibits computerized analysis?
   b. Has software been written for importing the data in an ongoing forecasting operation?
   c. Who will build and maintain the model?
   d. Are sufficient resources available (time, money, staff) to facilitate model building, use, evaluation, and improvement on a routine and ongoing basis?
   e. How much time remains before the first forecast must be produced?
   f. How many series must be forecast, and how often must ongoing forecasts be produced?
   g. What level of data **aggregation** or **disaggregation** is desirable?
   h. To whom does the forecaster or forecasting group report, and how will the forecasts be communicated?
   i. How might you conduct a "forecasting audit"?

6. (Assessing forecasting situations) For each of the following scenarios, discuss the decision environment, the nature of the object to be forecast, the forecast type, the forecast horizon, the loss function, the information set, and what sorts of simple or complex forecasting approaches you might entertain.
   a. You work for Airborne Analytics, a highly specialized mutual fund investing exclusively in airline stocks. The stocks held by the fund are chosen based on your recommendations. You learn that a newly rich oil-producing country has requested bids on a huge contract to deliver 30 state-of-the-art fighter planes, but that only two companies submitted bids. The stock of the successful bidder is likely to rise.
   b. You work for the Office of Management and Budget in Washington, D.C., and must forecast tax revenues for the upcoming fiscal year. You work for a president who wants to maintain funding for his pilot social programs, and high revenue forecasts ensure that the programs keep their funding. However, if the forecast is too high, and the president runs a large deficit at the end of the year, he will be seen as fiscally irresponsible, which will lessen his probability of reelection. Furthermore, your forecast will be scrutinized by the more conservative members of Congress; if they find fault with your procedures, they might have fiscal grounds to undermine the president's planned budget.
   c. You work for D&D, a major Los Angeles advertising firm, and you must create an ad for a client's product. The ad must be targeted toward teenagers, because they constitute the primary market for the product. You must (somehow) find out what kids currently think is cool, incorporate that information into your ad, and make your client's product attractive to the new

generation. If your hunch is right, your firm basks in glory, and you can expect multiple future clients from this one advertisement. If you miss, however, and the kids don't respond to the ad, then your client's sales fall, and the client may reduce or even close its account with you.

Bibliographical and Computational Notes

Klein (1971) and Granger and Newbold (1986) contain a wealth of insightful (but more advanced) discussion of many of the topics discussed in this chapter. The links between forecasts and decisions are clearly displayed in many of the chapters of Makridakis and Wheelwright (1987). Armstrong (1978) provides entertaining and insightful discussion of many of the specialized issues and techniques relevant in long-horizon forecasting. Several of the essays in Diebold and Watson (1996) concern the use of loss functions tailored to the decision-making situation of interest, both with respect to the forecast horizon and with respect to the shape of the loss function, as does Christoffersen and Diebold (1997). Zellner (1992) provides an insightful statement of the KISS principle, which is very much related to the parsimony principle of Box and Jenkins (see Box, Jenkins, and Reinsel, 1994). Levenbach and Cleary (1984) contains useful discussion of forecasting as an ongoing process.

Concepts for Review

Decision environment
Loss function
Forecast object
Forecast statement
Forecast horizon
Information set
Methods and complexity
Parsimony principle
Shrinkage principle
Symmetric loss
Asymmetric loss
Forecast error
Quadratic loss
Absolute loss
Absolute error loss
Direction-of-change forecast
Optimal forecast
Event outcome forecast
Event timing forecast
Time series forecast
Point forecast
Interval forecast
Density forecast
Probability forecast
h-step-ahead forecast
h-step-ahead extrapolation forecast
KISS principle
Ragged edges
Missing observations
Outlier
Measurement error
Aggregation
Disaggregation

References and Additional Readings

Armstrong, J. S. (1978). *Long Run Forecasting: From Crystal Ball to Computer.* New York: Wiley.

Box, G. E. P., Jenkins, G. W., and Reinsel, G. (1994). *Time Series Analysis, Forecasting and Control.* 3rd ed. Englewood Cliffs, N. J.: Prentice Hall.

Christoffersen, P. F., and Diebold, F. X. (1997). "Optimal Prediction under Asymmetric Loss." *Econometric Theory,* 13, 808–817.

Diebold, F. X., and Watson, M. W., eds. (1996). *New Developments in Economic Forecasting.* Special issue of *Journal of Applied Econometrics,* 11, 453–594.

Granger, C. W. J., and Newbold, P. (1986). *Forecasting Economic Time Series.* 2nd ed. Orlando, Fl.: Academic Press.

Klein, L. R. (1971). *An Essay on the Theory of Economic Prediction.* Chicago: Markham.

Levenbach, H., and Cleary, J. P. (1984). *The Modern Forecaster.* Belmont, Calif.: Lifetime Learning Publications.

Makridakis, S., and Wheelwright, S. (1987). *The Handbook of Forecasting: A Manager's Guide.* 2nd ed. New York: Wiley.

Zellner, A. (1992). "Statistics, Science and Public Policy." *Journal of the American Statistical Association,* 87, 1–6.

Statistical Graphics for Forecasting

It's almost always a good idea to begin forecasting projects with graphical data analysis. When compared with the modern array of statistical modeling methods, graphical analysis might seem trivially simple, perhaps even so simple as to be incapable of delivering serious insights into the series to be forecast. Such is not the case: In many respects the human eye is a far more sophisticated tool for data analysis and modeling than even the most sophisticated modern modeling techniques. That's certainly not to say that graphical analysis alone will get the job done—certainly, graphical analysis has its limitations—but it's usually the best place to *start.* With that in mind, we introduce in this chapter some simple graphical techniques, and we consider some basic elements of graphical style.

1. The Power of Statistical Graphics

The four datasets shown in Table 4.1, known as **Anscombe's quartet,** provide stark illustration of the power of statistical graphics. Each dataset consists of 11 observations on two variables. Simply glancing at the data—or even studying it with some care—yields little insight. Of course, you say, but that's why we have powerful modern statistical techniques, such as the linear regression model.

TABLE 4.1
Anscombe's Quartet

(1)		(2)		(3)		(4)	
x1	**y1**	**x2**	**y2**	**x3**	**y3**	**x4**	**y4**
10.0	8.04	10.0	9.14	10.0	7.46	8.0	6.58
8.0	6.95	8.0	8.14	8.0	6.77	8.0	5.76
13.0	7.58	13.0	8.74	13.0	12.74	8.0	7.71
9.0	8.81	9.0	8.77	9.0	7.11	8.0	8.84
11.0	8.33	11.0	9.26	11.0	7.81	8.0	8.47
14.0	9.96	14.0	8.10	14.0	8.84	8.0	7.04
6.0	7.24	6.0	6.13	6.0	6.08	8.0	5.25
4.0	4.26	4.0	3.10	4.0	5.39	19.0	12.50
12.0	10.84	12.0	9.13	12.0	8.15	8.0	5.56
7.0	4.82	7.0	7.26	7.0	6.42	8.0	7.91
5.0	5.68	5.0	4.74	5.0	5.73	8.0	6.89

So let's regress y on x for each of the four datasets. The results appear in Table 4.2. Interestingly enough, although the four datasets certainly contain different numerical data values, the standard linear regression output is identical in each case. First, the fitted regression line is the same in each case, $y = 3 + \frac{1}{2}x$. Second, the uncertainty associated with the estimated parameters, as summarized by standard errors, is also the same in each dataset. Hence, the t-statistics, which are simply ratios of estimated coefficients to their standard errors, are also identical across datasets. Third, R^2, which is the percentage of variation in y explained by variation in x, is identical across datasets. Fourth, the sum of squared residuals, and hence the standard error of the regression (the estimated standard deviation of the stochastic disturbance to the linear regression relationship), is the same in each dataset.

That's all fine, too, you say—the relationship between y and x is simply the same in each dataset, even though the specific data differ due to random influences. The assertion that the relationship between y and x is the same in each dataset *could* be correct, but graphical examination of the data reveals immediately that it's *not* correct. In Figure 4.1, we show graphs of y versus x (called **pairwise scatterplots** or **bivariate scatterplots**) for each of the four datasets, with fitted regression lines superimposed. Although the fitted regression line *is* the same in each case, the reasons differ greatly, and it's clear that for most of the datasets the linear regression model is not appropriate.

In dataset 1, all looks well. We see that $y1$ and $x1$ are clearly positively correlated, and they appear to conform rather well to a linear relationship, although the relationship is certainly not perfect. In short, all the conditions of the classical linear regression model appear satisfied in dataset 1.

In dataset 2, the situation is very different. The graph reveals that there's certainly a relationship between $y2$ and $x2$—perhaps even a deterministic

TABLE 4.2 *Anscombe's Quartet: Regressions of y_i on x_i, $i = 1, \ldots, 4$*

LS // Dependent variable is y1.

Variable	**Coefficient**	**Std. Error**	***t*-Statistic**	
C	3.00	1.12	2.67	
x1	0.50	0.12	4.24	
R^2	0.67		S.E. of regression	1.24

LS // Dependent variable is y2.

Variable	**Coefficient**	**Std. Error**	***t*-Statistic**	
C	3.00	1.12	2.67	
x2	0.50	0.12	4.24	
R^2	0.67		S.E. of regression	1.24

LS // Dependent Variable is y3.

Variable	**Coefficient**	**Std. Error**	***t*-Statistic**	
C	3.00	1.12	2.67	
x3	0.50	0.12	4.24	
R^2	0.67		S.E. of regression	1.24

LS // Dependent Variable is y4.

Variable	**Coefficient**	**Std. Error**	***t*-Statistic**	
C	3.00	1.12	2.67	
x4	0.50	0.12	4.24	
R^2	0.67		S.E. of regression	1.24

relationship—but it also makes clear that the relationship is not at all linear. Thus, the use of the linear regression model is not desirable in dataset 2.

In dataset 3, the graphics indicate that although y and x do seem to conform to a linear relationship, there is one key (y3, x3) pair that doesn't conform well to the linear relationship. Most likely you never noticed that data point when you simply examined the raw data in tabular form, in spite of the fact that it's visually obvious when we make use of graphics.

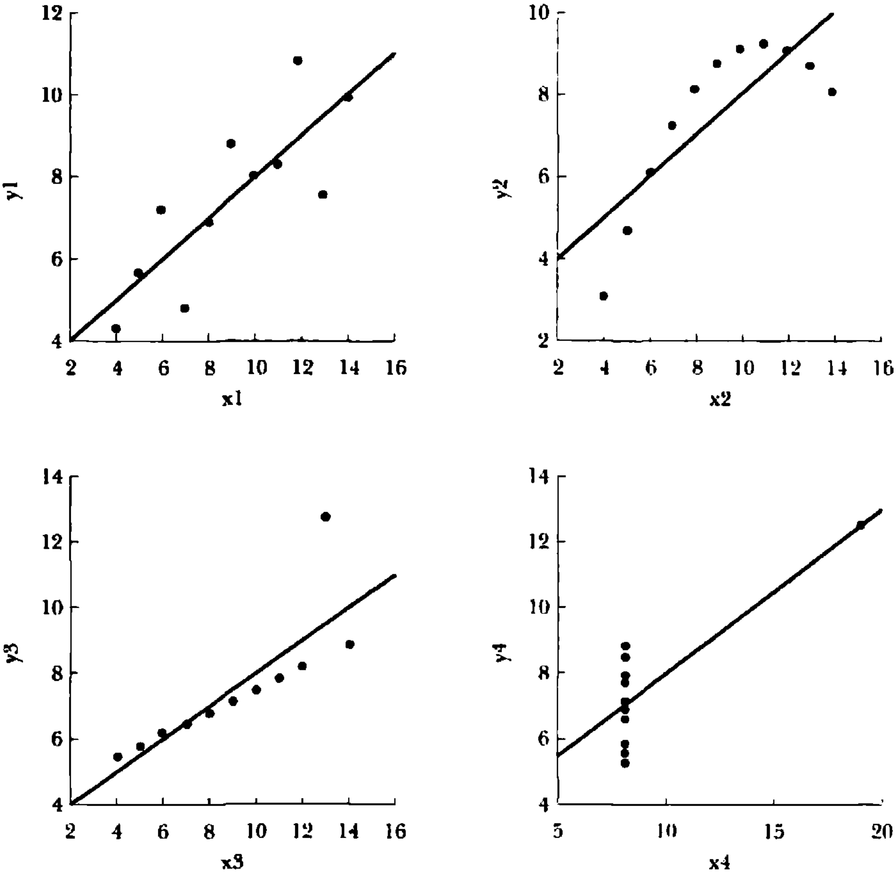
Dataset 4 is rather odd—the (y4, x4) pairs are all stacked vertically, with the exception of one point, which exerts a huge influence on the fitted regression line. At any rate, the graphics once again make the anomalous nature of this situation immediately apparent.

Let's summarize what we've learned about the power of graphics:

a. *Graphics helps us to summarize and reveal patterns in data,* as, for example, with linear versus nonlinear functional form in the first and second Anscombe datasets. That's key in any forecasting project.

b. *Graphics helps us identify anomalies in data,* as in the third Anscombe dataset. That's also key in forecasting, because we'll produce our forecasts from models fit to the historical data, and the dictum "garbage in, garbage out" most definitely applies.

FIGURE 4.1
Anscombe's Quartet: Bivariate Scatterplots



c. Less obvious, but most definitely relevant, is the fact that *graphics facilitates and encourages comparison of different pieces of data.* That's why, for example, we graphed all four datasets in one big figure. By doing so, we facilitate effortless and instantaneous cross-dataset comparison of statistical relationships. This technique is called **multiple comparisons.**

d. There's one more aspect of the power of statistical graphics. It comes into play in the analysis of large datasets, so it wasn't revealed in the analysis of the Anscombe datasets, which are not large, but it's nevertheless tremendously important. *Graphics enables us to present a huge amount of data in a small space, and it enables us to make huge datasets coherent.* We might, for example, have supermarket scanner data, recorded in 5-minute intervals for a year, on the quantities of goods sold in each of four food categories: dairy, meat, grains, and vegetables. Tabular or similar analysis of such data is simply out of the question, but graphics is still straightforward and can reveal important patterns.

2. Simple Graphical Techniques

As we discussed in Chapter 3, time series are by far the most common objects for which forecasts are made. Thus, we will focus primarily on graphics useful for modeling and forecasting time series. The dimensionality of the data—the number of time series we wish to examine—plays a key role. Because graphical analysis "lets the data speak for themselves," it is most useful when the dimensionality of the data is low. We will segment our discussion into two parts: univariate and multivariate.

UNIVARIATE GRAPHICS

First and foremost, graphics is used to reveal the patterns in time series data. *We use graphical analysis to get a preliminary and informal idea of the nature* of trend, seasonality, and cycles, as well as the nature and location of any unusual or aberrant observations, structural breaks, and so forth. The great workhorse of univariate time series graphics is the simple time series plot, in which the series of interest is graphed against time.

In Figure 4.2, for example, we present a **time series plot** of the 1-year U.S. Treasury bond rate, 1960.01–2005.03.[1] A number of important features of the series are apparent. Among other things, its movements appear sluggish and persistent, it appears to trend gently upward until about 1980, and it appears to trend gently downward thereafter.

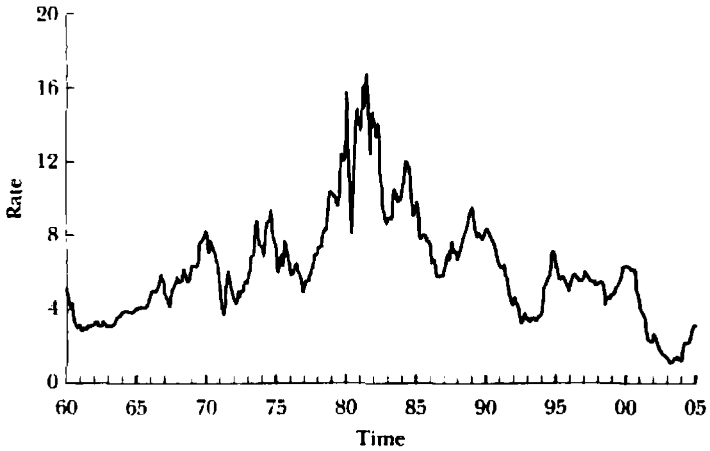


FIGURE 4.2
1-Year Treasury Bond Rate

---

[1] The notation "1960.01–2005.03" means the first month of 1960 through the third month of 2005.

FIGURE 4.3
Change in 1-Year Treasury Bond Rate

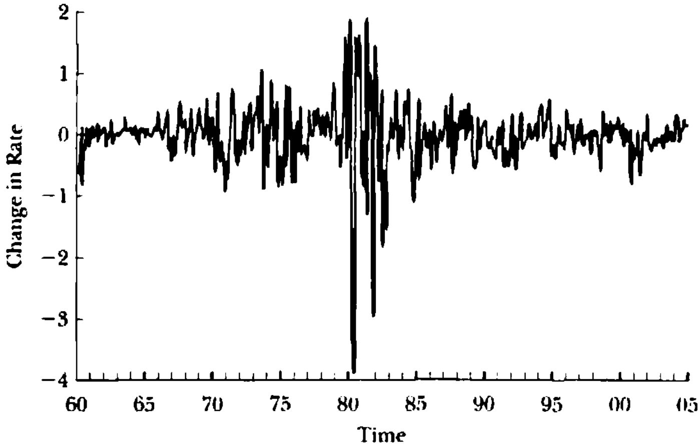


Figure 4.3 provides a different perspective; we plot the *change* in the 1-year T-bond rate, which highlights volatility fluctuations. Interest rate volatility appears low in the 1960s, a bit higher in the 1970s, and very high from late 1979 through late 1982 (the period during which the Federal Reserve targeted a monetary aggregate, which had the side effect of increasing interest rate volatility), after which volatility gradually declines.

Time series plots are helpful for learning about other features of time series as well. In Figure 4.4, for example, we show a time series plot of U.S. liquor sales, 1960.01–2001.03. Clearly they're trending upward, but the plot indicates that a break in the trend may occur sometime during the 1980s. In addition,

FIGURE 4.4
Liquor Sales

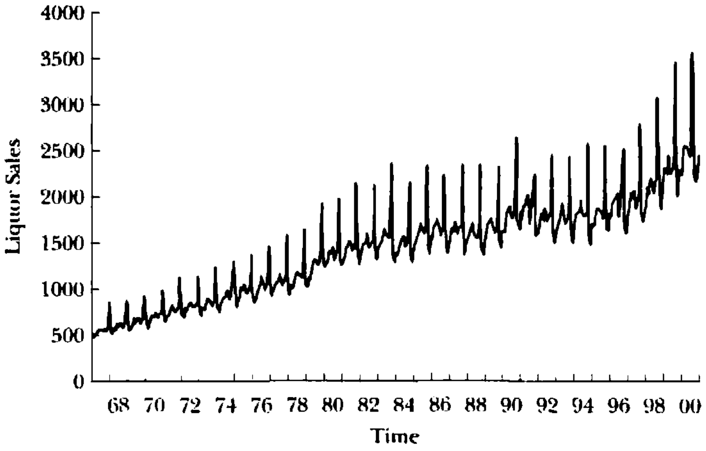
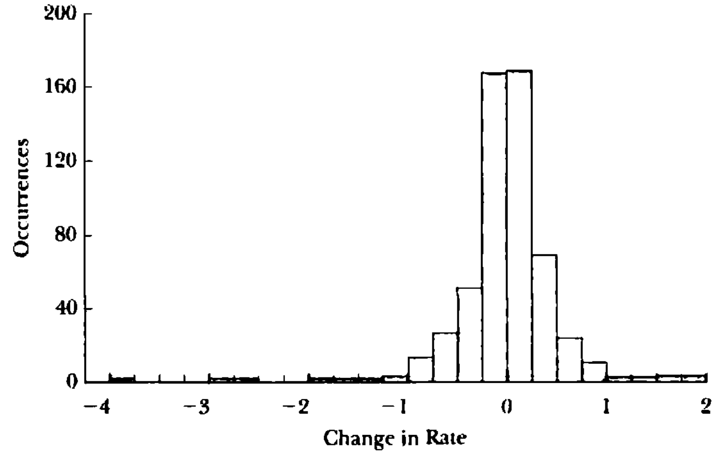


FIGURE 4.5
Histogram and Descriptive Statistics: Change in 1-Year Treasury Bond Rate



```
Sample 1960M01 2005M03
Observations 543

Mean          -0.003389
Median         0.000000
Maximum        1.900000
Minimum       -3.910000
Std. dev.      0.470731
Skewness      -1.380580
Kurtosis      17.44767

Jarque-Bera   4895.124
Probability   0.000000
```

the plot makes clear the pronounced seasonality in the series—liquor sales skyrocket every December—and moreover that the volatility of the seasonal fluctuations grows over time as the level of the series increases.

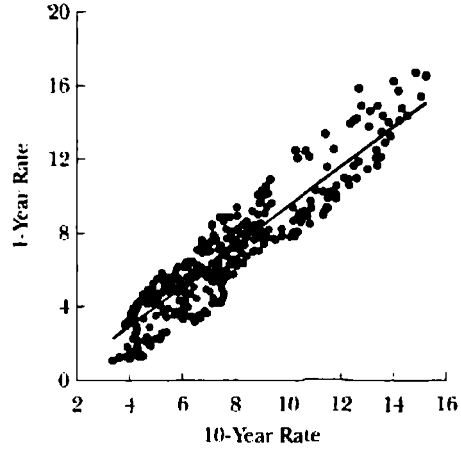
Univariate graphical techniques are also routinely used to assess distributional shape. A **histogram,** for example, provides a simple estimate of the probability density of a random variable. The observed range of variation of the series is split into a number of segments of equal length, and the height of the bar placed at a segment is the percentage of observations falling in that segment.[2] Figure 4.5 shows a histogram for the change in the 1-year T-bond rate with related diagnostic information. The histogram indicates that the series is roughly symmetrically distributed, and the additional statistics such as the sample mean, median, maximum, minimum, and standard deviation convey important additional information about the distribution.

For example, a key feature of the distribution of T-bond rate changes, which may not have been immediately apparent from the histogram, is that it has fatter tails than would be the case under normality. This is at once apparent from the kurtosis statistic, which would be approximately 3 if the data were normally distributed. Instead, it's about 10, indicating much fatter tails than the normal, which is very common in high-frequency financial data. The skewness statistic is modestly negative, indicating a rather long left tail. The Jarque-Bera normality test rejects the hypothesis of independent normally distributed

---

[2] In some software packages (for example, Eviews), the height of the bar placed at a segment is simply the number, not the percentage, of observations falling in that segment. Strictly speaking, such histograms are not density estimators, because the "area under the curve" doesn't add to 1, but they are equally useful for summarizing the shape of the density.

FIGURE 4.6
Scatterplot: 1-Year versus 10-Year Treasury Bond Rate



observations. The rejection occurs because the interest rate changes are not independent, not normally distributed, or both. It's likely both, and the deviation from normality is due more to leptokurtosis than to asymmetry.[3]

MULTIVARIATE GRAPHICS

When two or more variables are available, the possibility of relations between the variables becomes important, and we use graphics to uncover the existence and nature of such relationships. We use **relational graphics** to display relationships and flag anomalous observations. You already understand the idea of a bivariate scatterplot—we used it extensively to uncover relationships and anomalies in the Anscombe data.[4] In Figure 4.6, for example, we show a bivariate scatterplot of the 1-year U.S. Treasury bond rate versus the 10-year U.S. Treasury bond rate, 1960.01–2005.03. The scatterplot indicates that the two move closely together. Although each of the rates is individually highly persistent, the deviations from the superimposed regression line appear transient. You can think of the line as perhaps representing long-run equilibrium relationships, to which the variables tend to cling.

The regression line that we superimpose on a scatterplot of y versus x is an attempt to summarize how the conditional mean of y (given x) varies with x.

---

[3] The rejection could also occur because the sample size is too small to invoke the large-sample theory on which the Jarque-Bera test is based, but that's not likely in the present application, for which we have quite a large sample of data.

[4] Just as in our analysis of the Anscombe data, we often make bivariate scatterplots with fitted regression lines superimposed, to help us to visually assess the adequacy of a linear model. Note that although superimposing a regression line is helpful in bivariate scatterplots, "connecting the dots" is not. This contrasts to time series plots, for which connecting the dots is fine and is typically done.

Under certain conditions that we'll discuss in later chapters, this conditional mean is the best point forecast of y. Thus, you can think of the regression line as summarizing how our best point forecast of y varies with x. The linear regression model involves a lot of structure (it assumes that $E(y|x)$ is a linear function of x), but less structured approaches exist and are often used to provide potentially nonlinear estimates of conditional mean functions for superimposition on scatterplots.

Thus far, all our discussion of multivariate graphics has been bivariate. That's because graphical techniques are best suited to low-dimensional data. Much recent research has been devoted to graphical techniques for high-dimensional data, but all such high-dimensional graphical analysis is subject to certain inherent limitations. Here we'll discuss just one simple and popular scatterplot technique for high-dimensional data—and one that's been around for a long time—the **scatterplot matrix, or multiway scatterplot.** The scatterplot matrix is just the set of all possible bivariate scatterplots, arranged in the upper-right or lower-left part of a matrix to facilitate multiple comparisons. If we have data on N variables, there are $\frac{N^2-N}{2}$ such pairwise scatterplots. In Figure 4.7, for example, we show a scatterplot matrix for the 1-year, 10-year, 20-year, and 30-year U.S. Treasury bond rates, 1960.01–2005.03. There are a total of six pairwise scatterplots, and the multiple comparison makes clear that although the interest rates are closely related in each case, with a regression slope of approximately 1, the relationship is more precise in some cases (such as 20- and 30-year rates) than in others (such as 1- and 30-year rates).

3. Elements of Graphical Style

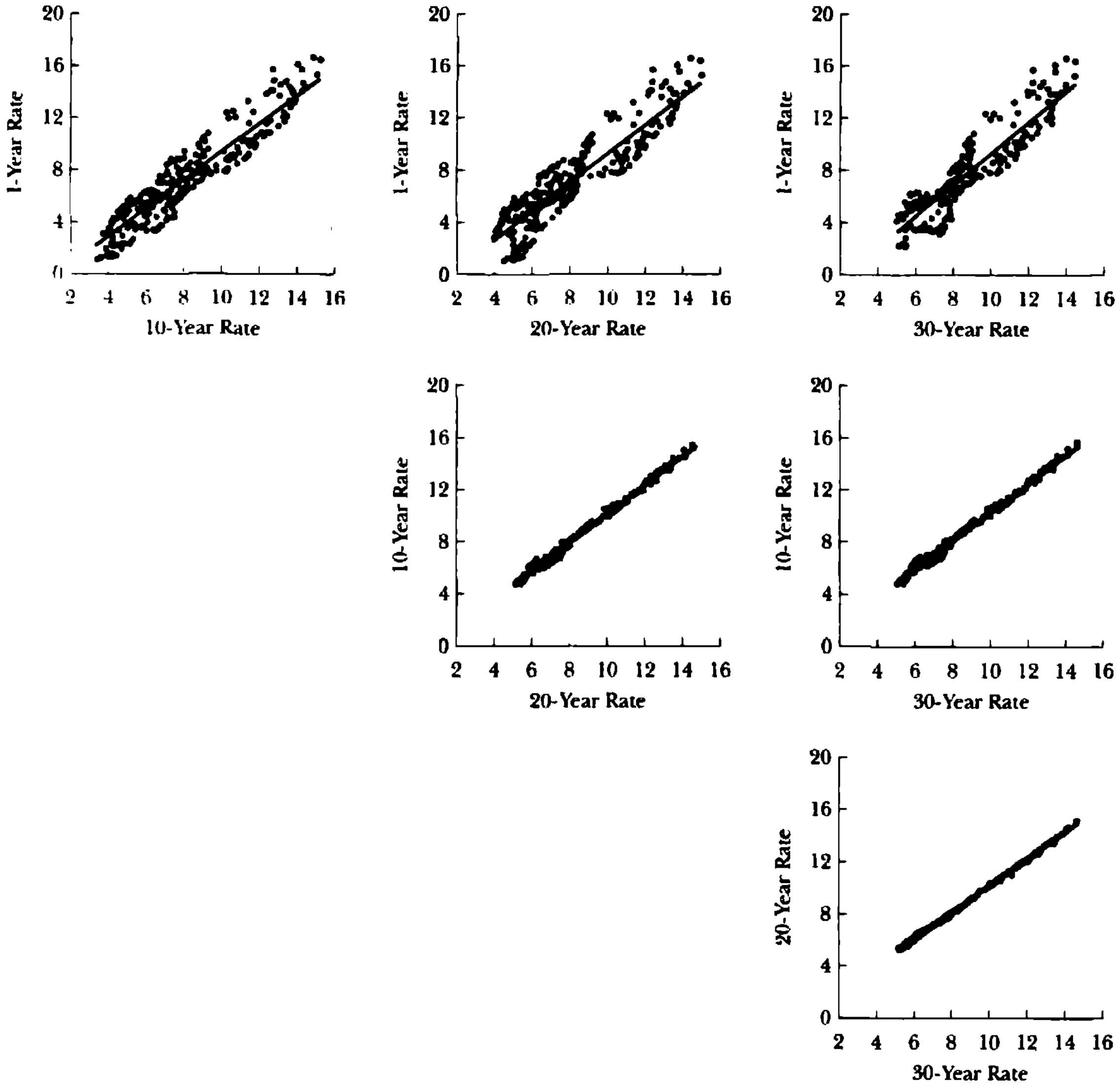
In the preceding section, we discussed various graphical tools. As with all tools, however, graphical tools can be used effectively or ineffectively. In this section, you'll learn what makes good graphics good and bad graphics bad. In doing so, you'll learn to use graphical tools effectively.

Bad graphics is like obscenity: It's hard to define, but you know it when you see it. Conversely, producing good graphics is like good writing: It's an iterative, trial-and-error procedure and very much an art rather than a science. But that's not to say that anything goes; as with good writing, good graphics requires discipline. There are at least three keys to good graphics:

a. Know your audience, and know your goals.
b. Understand and follow two fundamental principles: show the data, and appeal to the viewer.
c. Revise and edit, again and again.

We can use a number of devices to *show the data.* First, avoid distorting the data or misleading the viewer. Thus, for example, avoid changing scales in midstream, use common scales when performing multiple comparisons,

FIGURE 4.7 *Scatterplot Matrix: 1-, 10-, 20-, and 30-Year Treasury Bond Rates*



and so on. Second, minimize, within reason, **nondata ink.**[5] Avoid **chartjunk** (elaborate shadings and grids, decoration, and related nonsense), erase unnecessary axes, refrain from use of artificial three-dimensional perspective, and so forth.

---

[5] *Nondata ink* is ink used to depict anything other than data points.

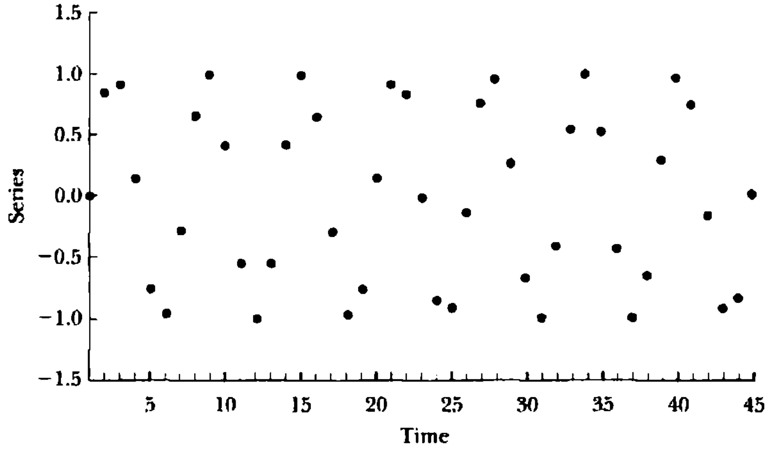


FIGURE 4.8
Time Series Plot, Aspect Ratio 1:1.6

Other guidelines help us *appeal to the viewer.* First, use clear and modest type, avoid mnemonics and abbreviations, and use labels rather then legends when possible. Second, make graphics self-contained; a knowledgeable reader should be able to understand your graphics without reading pages of accompanying text. Third, as with our prescriptions for showing the data, avoid chartjunk.

An additional aspect of creating graphics that show the data and appeal to the viewer is selection of a graph's aspect ratio. The **aspect ratio** is the ratio of the graph's height, h, to its width, w, and it should be selected such that the graph reveals patterns in the data and is visually appealing. One time-honored approach geared toward visual appeal is to use an aspect ratio such that height is to width as width is to the sum of height and width. Algebraically,

$$\frac{h}{w} = a = \frac{w}{h+w} .$$

Dividing numerator and denominator of the right side by w yields

$$a = \frac{1}{a+1} ,$$

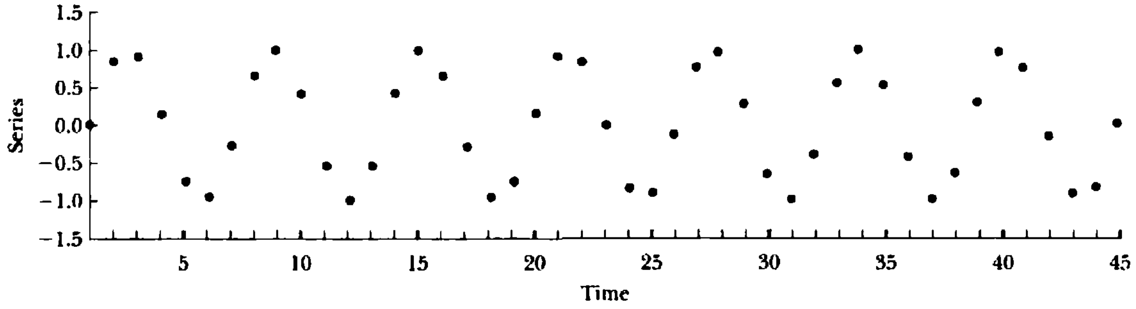
or

$$a^2 + a - 1 = 0 .$$

The positive root of this quadratic polynomial is $a = 0.618$, the so-called **golden ratio.** Graphics that conform to the golden ratio, with height a bit less than two-thirds of width, are visually appealing. In Figure 4.8, for example, we show a plot whose dimensions roughly correspond to the golden aspect ratio.

Other things the same, it's a good idea to keep the golden ratio in mind when producing graphics. Other things are not always the same, however. In particular, the golden aspect ratio may not be the one that maximizes pattern revelation. Consider Figure 4.9, for example, in which we plot exactly the

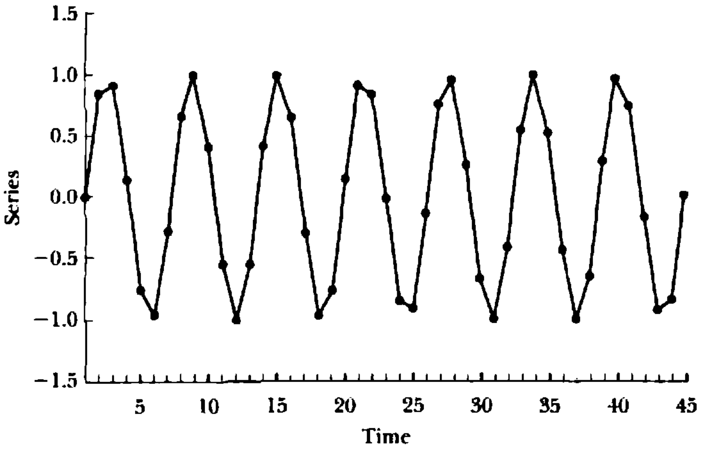
FIGURE 4.9 *Time Series Plot, Banked to 45 Degrees*



same data as in Figure 4.8, but with a smaller aspect ratio. The new plot reveals an obvious pattern in the data, which you probably didn't notice before, and is therefore a superior graphic.

The improved aspect ratio of Figure 4.9 was selected to make the average absolute slope of the line segments connecting the data points approximately equal to 45 degrees. This procedure, **banking to 45 degrees,** is useful for selecting a revealing aspect ratio. As in Figure 4.9, the most revealing aspect ratio for time series—especially long time series—is often less than the golden ratio. Sometimes, however, various devices can be used to maintain the golden aspect ratio while nevertheless clearly revealing patterns in the data. In Figure 4.10, for example, we use the golden aspect ratio but connect the data points, which makes the pattern clear.

FIGURE 4.10
Time Series Plot, Aspect Ratio 1:1.6



4. Application: Graphing Four Components of Real GDP

As with writing, the best way to learn graphics is to do it, so let's proceed immediately with an application that illustrates various points of graphical style. We'll examine four key components of U.S. real GDP: manufacturing, retail, services, and agriculture, recorded annually from 1960 to 2001 in millions of current dollars.

We begin in Figure 4.11 with a set of bar graphs. The value of each series in each year is represented by the height of a vertical bar, with different bar shadings for the different series. It's repugnant and unreadable, with no title, no axis numbering or labels, bad mnemonics, and so on. The good news is there's plenty of room for improvement.

We continue in Figure 4.12 with a set of stacked bar graphs, which are a bit easier to read because there's only one bar at each time point rather than four, but otherwise they suffer from all the defects of the bar graphs in Figure 4.11. Typically, bar graphs are simply not good graphical tools for time series. We therefore switch in Figure 4.13 to a time series plot with different types of lines and symbols for each series, which is a big improvement, but there's still room for additional improvement.

In Figure 4.14, we drop the symbols and add axis numbering. This figure is a major improvement, but the plot is still poor. In particular, it still has bad mnemonics, no title, and no axis labels. Moreover, it's not clear that dropping

FIGURE 4.11

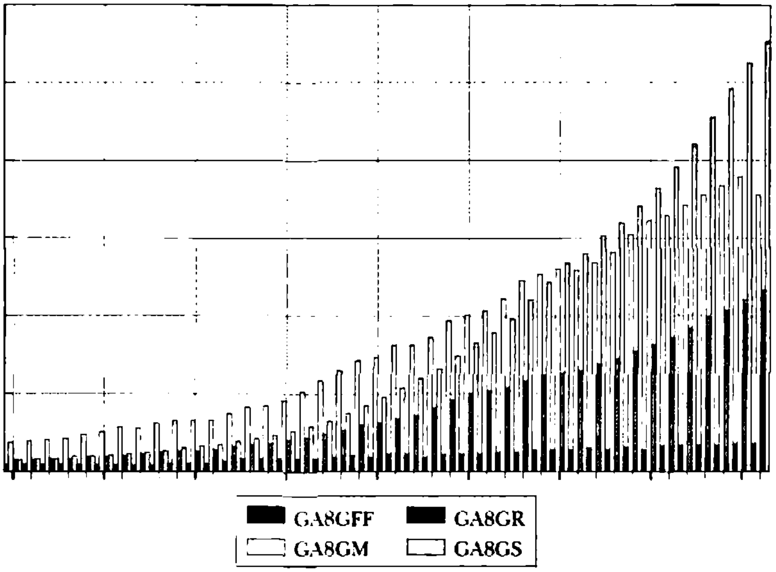


FIGURE 4.12

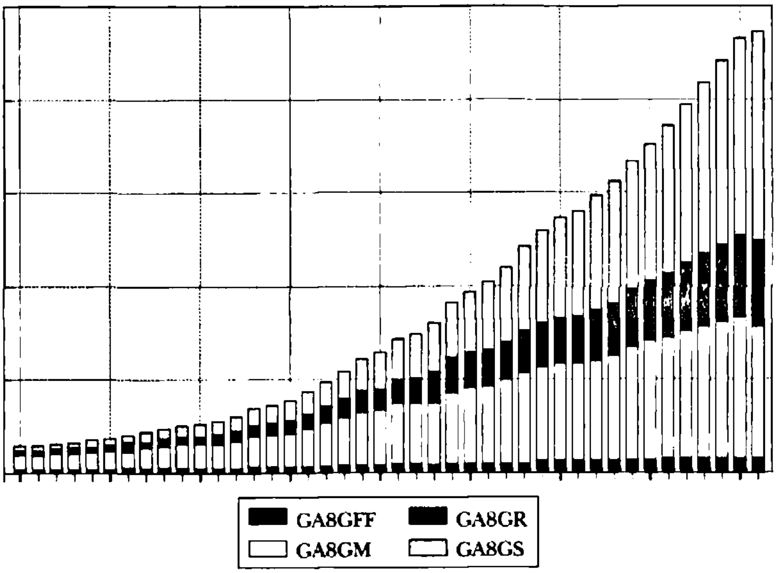
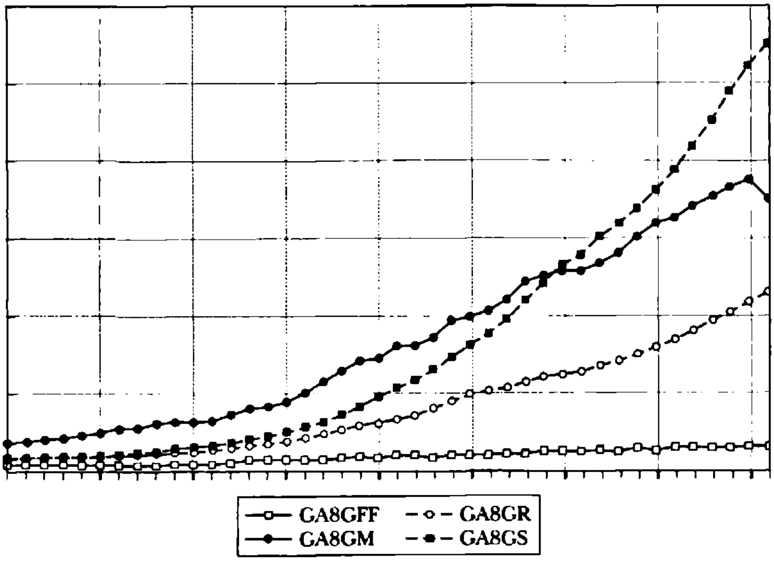


FIGURE 4.13



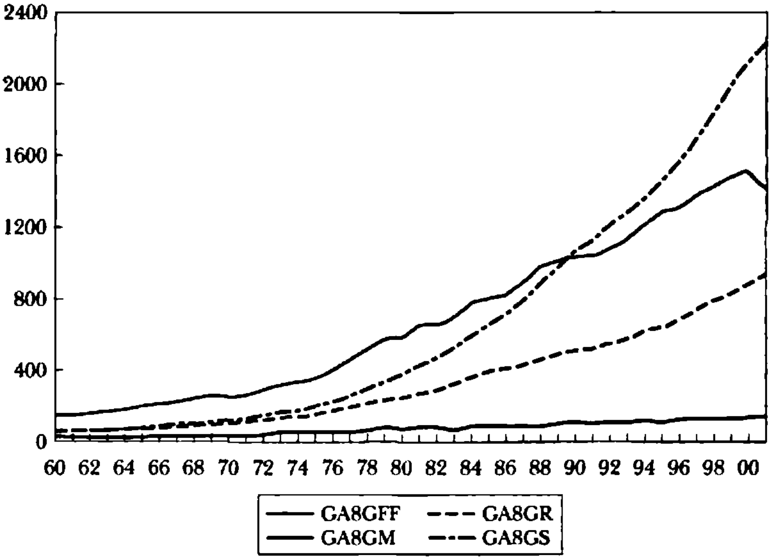


FIGURE 4.14

the plotting symbols produced an improvement, even though they are nondata ink. (Why?)

In Figure 4.15, we drop the different plotting lines and symbols altogether. Instead, we simply plot all the series with solid lines and label them directly. This approach produces a much more informative and appealing plot, in large part because there's no longer a need for the hideous legend and associated

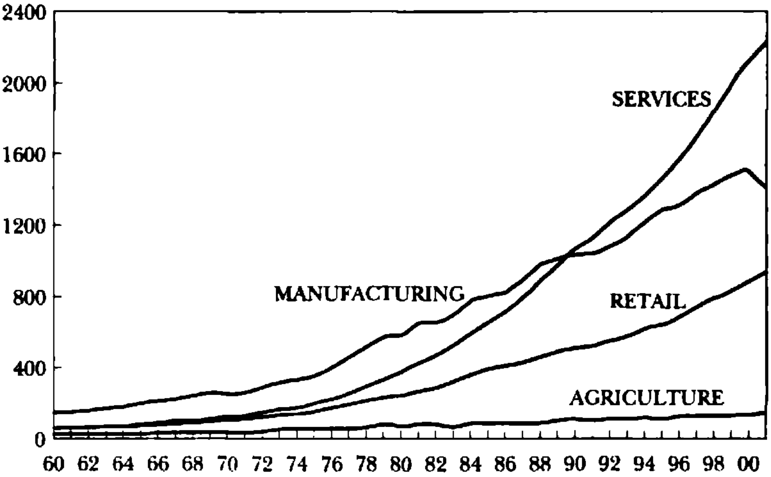
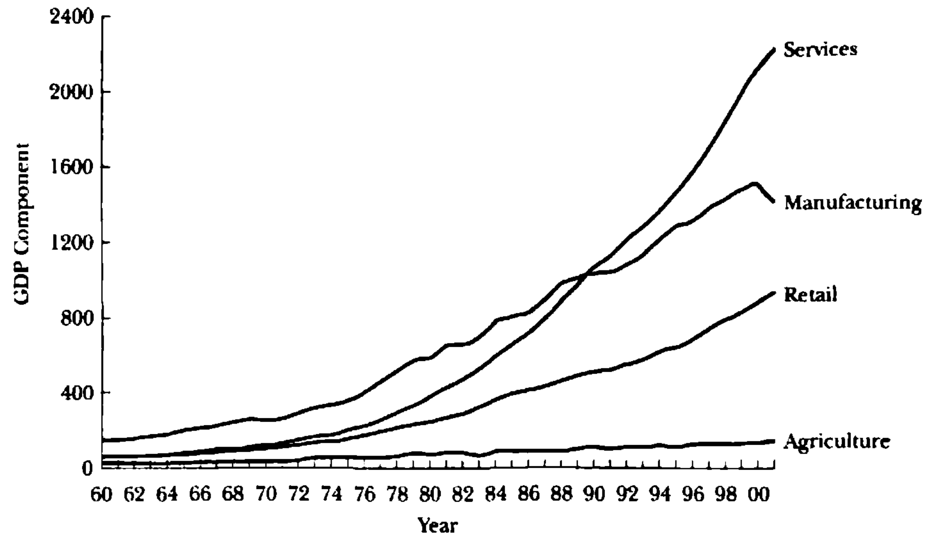


FIGURE 4.15

FIGURE 4.16
Components of Real GDP (Millions of Current Dollars, Annual)



mnemonics. However, a new annoyance has been introduced; the CAPITAL series labeling repels the viewer.

In Figure 4.16, we attempt to remedy the remaining defects of the plot. Both the horizontal and vertical axes are labeled, all labeling makes use of both capital and lowercase type as appropriate, the northern and eastern box lines have been eliminated (they're nondata ink and serve no useful purpose), the plot has a descriptive title, and, for visual reference, we have added shading indicating recessions.

5. Concluding Remarks

We've emphasized in this chapter that graphics is a powerful tool with a variety of uses in the construction and evaluation of forecasts and forecasting models. We hasten to add, however, that graphics has its limitations. In particular, graphics loses a lot of its power as the dimension of the data grows. If we have data in 10 dimensions, and we try to squash it into 2 or 3 dimensions to make a graph, we're bound to lose some information. That's also true of the models we fit; a linear regression model with 10 right-hand-side variables, for example, assumes that the data tend to lie in a small subset of 10-dimensional space.

Thus, in contrast to the analysis of data in two or three dimensions, in which case learning about data by fitting models involves a loss of information whereas graphical analysis does not, graphical methods lose their comparative advantage in higher dimensions. In higher dimensions, *both* graphics and models lose information, and graphical analysis can become comparatively laborious and less insightful. The conclusion, however, is straightforward: Graphical analysis and model fitting are complements, not substitutes, and when used together they can make valuable contributions to forecasting.

Exercises, Problems, and Complements

1. (Outliers) Recall the lower-left panel of the multiple comparison plot of the Anscombe data (Figure 4.1), which made clear that dataset 3 contained a severely anomalous observation. We call such data points **outliers.**
    a. Outliers require special attention because they can have substantial influence on the fitted regression line. Regression parameter estimates obtained by least squares are particularly susceptible to such distortions. Why?
    b. Outliers can arise for a number of reasons. Perhaps the outlier is simply a mistake because of a clerical recording error, in which case you'd want to replace the incorrect data with the correct data. We'll call such outliers **measurement outliers,** because they simply reflect measurement errors. If a particular value of a recorded series is plagued by a measurement outlier, there's no reason that observations at other times should necessarily be affected. But they *might* be affected. Why?
    c. Alternatively, outliers in time series may be associated with large unanticipated shocks, the effects of which may linger. If, for instance, an adverse shock hits the U.S. economy this quarter (for example, the price of oil on the world market triples) and the U.S. plunges into a severe depression, then it's likely that the depression will persist for some time. Such outliers are called **innovation outliers,** because they're driven by shocks, or "innovations," whose effects naturally last more than one period because of the dynamics operative in business, economic, and financial series.
    d. How to identify and treat outliers is a time-honored problem in data analysis, and there's no easy answer. What factors would you, as a forecaster, examine when deciding what to do with an outlier?

2. (Simple versus partial correlation) The set of pairwise scatterplots that comprises a multiway scatterplot provides useful information about the joint distribution of the N variables, but it's incomplete information and should be interpreted with care. A pairwise scatterplot summarizes information regarding the simple correlation between, say, x and y. But x and y may appear highly related in a pairwise scatterplot even if they are in fact unrelated, if each depends on a third variable—say, z. The crux of the problem is that there's no way in a pairwise scatterplot to examine the correlation between x and y *controlling* for z, which we call **partial correlation.** When interpreting a scatterplot matrix, keep in mind that the pairwise scatterplots provide information only on **simple correlation.**

3. (Graphical regression diagnostic 1: time series plot of y_t, $\hat{y}_t$, and e_t) After estimating a forecasting model, we often make use of graphical techniques to provide important diagnostic information regarding the adequacy of the model.

Often the graphical techniques involve the residuals from the model. Throughout, let the regression model be

$$y_t = \sum_{i=1}^{k} \beta_i x_{it} + \varepsilon_t \ ,$$

and let the fitted values be

$$\hat{y}_t = \sum_{i=1}^{k} \hat{\beta}_i x_{it} \ .$$

The difference between the actual and fitted values is the residual,

$$e_t = y_t - \hat{y}_t \ .$$

a. Superimposed time series plots of y_t and $\hat{y}_t$ help us to assess the overall fit of a forecasting model and to assess variations in its performance at different times (for example, performance in tracking peaks versus troughs in the business cycle).
b. A time series plot of e_t (a so-called residual plot) helps to reveal patterns in the residuals. Most important, it helps us to assess whether the residuals are correlated over time—that is, whether the residuals are serially correlated, as well as whether there are any anomalous residuals. Note that even though there might be many right-hand-side variables in this regression model, the actual values of y, the fitted values of y, and the residuals are simple univariate series that can be plotted easily. We'll make use of such plots throughout this book.

4. (Graphical regression diagnostic 2: time series plot of e_t^2 or $|e_t|$) Plots of e_t^2 or $|e_t|$ reveal patterns (most notably serial correlation) in the *squared* or *absolute* residuals, which correspond to nonconstant volatility, or heteroskedasticity, in the levels of the residuals. As with the standard residual plot, the squared or absolute residual plot is always a simple univariate plot, even when there are many right-hand-side variables. Such plots feature prominently, for example, in tracking and forecasting time-varying volatility.

5. (Graphical regression diagnostic 3: scatterplot of e_t versus x_t) This plot helps us to assess whether the relationship between y and the set of x's is truly linear, as assumed in linear regression analysis. If not, the linear regression residuals will depend on x. In the case where there is only one right-hand-side variable, as earlier, we can simply make a scatterplot of e_t versus x_t. When there is more than one right-hand-side variable, we can make separate plots for each, although the procedure loses some of its simplicity and transparency.

6. (Graphical analysis of foreign exchange rate data) Magyar Select, a marketing firm representing a group of Hungarian wineries, is considering entering into a contract to sell 8,000 cases of premium Hungarian dessert wine to AMI Imports, a worldwide distributor based in New York and London. The contract must be signed now, but payment and delivery is 90 days hence. Payment is to be in U.S. dollars; Magyar is therefore concerned about U.S. dollar/Hungarian forint (USD/HUF) exchange rate volatility over the next 90 days. Magyar has hired you to analyze and forecast the exchange rate, on which it has collected data for the

last 620 days. Naturally, you suggest that Magyar begin with a graphical examination of the data. (The USD/HUF exchange rate data is on the book's web page.)

a. Why might we be interested in examining data on the log rather than the level of the USD/HUF exchange rate?
b. Take logs and produce a time series plot of the log of the USD/HUF exchange rate. Discuss.
c. Produce a scatterplot of the log of the USD/HUF exchange rate against the lagged log of the USD/HUF exchange rate. Discuss.
d. Produce a time series plot of the change in the log USD/HUF exchange rate, and also produce a histogram, normality test, and other descriptive statistics. Discuss. (For small changes, the change in the logarithm is approximately equal to the percent change, expressed as a decimal.) Do the log exchange rate changes appear normally distributed? If not, what is the nature of the deviation from normality? Why do you think we computed the histogram and so forth for the differenced log data, rather than for the original series?
e. Produce a time series plot of the *square* of the change in the log USD/HUF exchange rate. Discuss and compare with the earlier series of log changes. What do you conclude about the volatility of the exchange rate, as proxied by the squared log changes?

7. (Common scales) Redo the multiple comparison of the Anscombe data in Figure 4.1 using **common scales.** Do you prefer the original or your newly created graphic? Why or why not?

8. (Graphing real GDP, continued from Section 4)
   a. Consider the final plot at which we arrived when graphing four components of U.S. real GDP. What do you like about the plot? What do you dislike about the plot? How could you make it still better? Do it!
   b. To help sharpen your eye (or so I claim), some of the graphics in this book fail to adhere strictly to the elements of graphical style that we emphasized. Pick and critique three graphs from anywhere in the book (apart from this chapter), and produce improved versions.

9. (Color)
   a. Color can aid graphics both in showing the data and in appealing to the viewer. How?
   b. Color can also confuse. How?
   c. Keeping in mind the principles of graphical style, formulate as many guidelines for color graphics as you can.

10. (Regression, regression diagnostics, and regression graphics in action) You're a new financial analyst at a major investment house, tracking and forecasting earnings of the health care industry. At the end of each quarter, you forecast industry earnings for the next quarter. Experience has revealed that your clients care about your forecast accuracy—that is, they want small errors—but that they are not particularly concerned with the sign of your error. (Your clients use your forecast to help to allocate their portfolios, and if your forecast is way off, they lose money, regardless of whether you're too optimistic or too pessimistic.) Your immediate predecessor has bequeathed to you a forecasting model in which

current earnings (y_t) are explained by one variable lagged by one quarter (x_{t-1}). (Both are on the book's web page.)

a. Suggest and defend some candidate x variables. Why might lagged x, rather than current x, be included in the model?
b. Graph y_t versus x_{t-1} and discuss.
c. Regress y_t on x_{t-1} and discuss (including related regression diagnostics that you deem relevant).
d. Assess the entire situation in light of the "six considerations basic to successful forecasting" emphasized in Chapter 3: the decision environment and loss function, the forecast object, the forecast statement, the forecast horizon, the information set, and the parsimony principle.
e. Consider as many variations as you deem relevant on the general theme. At a minimum, you will want to consider the following:
    - Does it appear necessary to include an intercept in the regression?
    - Does the functional form appear adequate? Might the relationship be nonlinear?
    - Do the regression residuals seem random, and, in particular, do they appear serially correlated or heteroskedastic?
    - Are there any outliers? If so, does the estimated model appear robust to their inclusion/exclusion?
    - Do the regression disturbances appear normally distributed?
    - How might you assess whether the estimated model is structurally stable?

Bibliographical and Computational Notes

A subfield of statistics called **exploratory data analysis** (EDA) focuses on learning about patterns in data without pretending to have too much a priori theory. As you would guess, EDA makes heavy use of graphical and related techniques. For an introduction, see Tukey (1977), a well-known book by a pioneer in the area.

This chapter has been heavily influenced by Tufte (1983), as are all modern discussions of statistical graphics. Tufte's book is an insightful and entertaining masterpiece on graphical style that I recommend enthusiastically. Our discussion of Anscombe's quartet follows Tufte's; the original paper is Anscombe (1973).

Cleveland (1993, 1994) and Cook and Weisberg (1994) are fine examples of modern graphical techniques. Cleveland (1993) stresses tools for revealing information in high-dimensional data, as well as techniques that aid in showing the data and appealing to the viewer in standard low-dimensional situations. It also contains extensive discussion of banking to 45 degrees. Cook and Weisberg (1994) develop powerful graphical tools useful in the specification and evaluation of regression models.

Details of the Jarque-Bera test may be found in Jarque and Bera (1987).

All graphics in this chapter were done using Eviews. S+ implements a variety of more sophisticated graphical techniques and in many respects represents the cutting edge of statistical graphics software.

Concepts for Review

Anscombe's quartet
Pairwise scatterplot
Bivariate scatterplot
Multiple comparison
Time series plot
Histogram
Relational graphics
Scatterplot matrix
Multiway scatterplot
Nondata ink
Chartjunk
Aspect ratio
Golden ratio
Banking to 45 degrees
Outlier
Measurement outlier
Innovation outlier
Simple correlation
Partial correlation
Common scales
Exploratory data analysis

References and Additional Readings

Anscombe, F. J. (1973). "Graphs in Statistical Analysis." *American Statistician*, 27, 17–21.

Cleveland, W. S. (1993). *Visualizing Data*. Summit, N.J.: Hobart Press.

Cleveland, W. S. (1994). *The Elements of Graphing Data*. 2nd ed. Monterey Park, Calif.: Wadsworth.

Cook, R. D., and Weisberg, S. (1994). *An Introduction to Regression Graphics*. New York: Wiley.

Jarque, C. M., and Bera, A. K. (1987). "A Test for Normality of Observations and Regression Residuals." *International Statistical Review*, 55, 163–172.

Tufte, E. R. (1983). *The Visual Display of Quantitative Information*. Cheshire, Conn.: Graphics Press.

Tukey, J. W. (1977). *Exploratory Data Analysis*. Reading, Mass.: Addison-Wesley.

Modeling and Forecasting Trend

1. Modeling Trend

The series that we want to forecast vary over time, and we often mentally attribute that variation to unobserved underlying components, such as trends, seasonals, and cycles. In this chapter, we focus on trend.[1] **Trend** is slow, long-run evolution in the variables that we want to model and forecast. In business, finance, and economics, for example, trend is produced by slowly evolving preferences, technologies, institutions, and demographics. We'll focus here on models of **deterministic trend,** in which the trend evolves in a perfectly predictable way. Deterministic trend models are tremendously useful in practice.[2]

Existence of trend is empirically obvious. Numerous series in diverse fields display trends. In Figure 5.1, we show the U.S. labor force participation rate for females aged 16 and over, the trend in which appears roughly *linear,* meaning that it increases or decreases like a straight line. That is, a simple linear function of time,

$$T_t = \beta_0 + \beta_1 \text{TIME}_t \, ,$$

---

[1] Later we'll define and study seasonals and cycles. Not all components need be present in all observed series.

[2] Later we'll broaden our discussion to allow for **stochastic trend.**

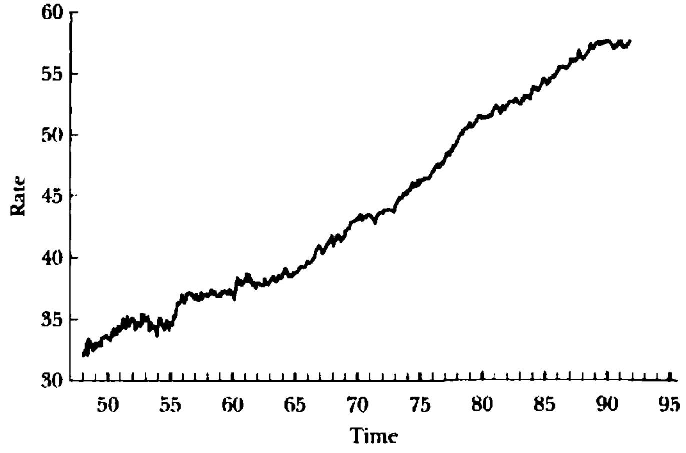


FIGURE 5.1
Labor Force Participation Rate, Females

provides a good description of the trend. The variable TIME is constructed artificially and is called a *time trend* or **time dummy**. Time equals 1 in the first period of the sample, 2 in the second period, and so on. Thus, for a sample of size T, TIME $= (1, 2, 3, \ldots, T-1, T)$; put differently, $\text{TIME}_t = t$. β_0 is the **regression intercept**; it's the value of the trend at time $t = 0$. β_1 is the **regression slope**; it's positive if the trend is increasing and negative if the trend is decreasing. The larger the absolute value of β_1, the steeper the trend's slope. In Figure 5.2, for example, we show two linear trends, one increasing and one decreasing. The increasing trend has an intercept of $\beta_0 = -50$ and a slope of $\beta_1 = 0.8$, whereas the decreasing trend has an intercept of $\beta_0 = 10$ and a gentler absolute slope of $\beta_1 = -0.25$.

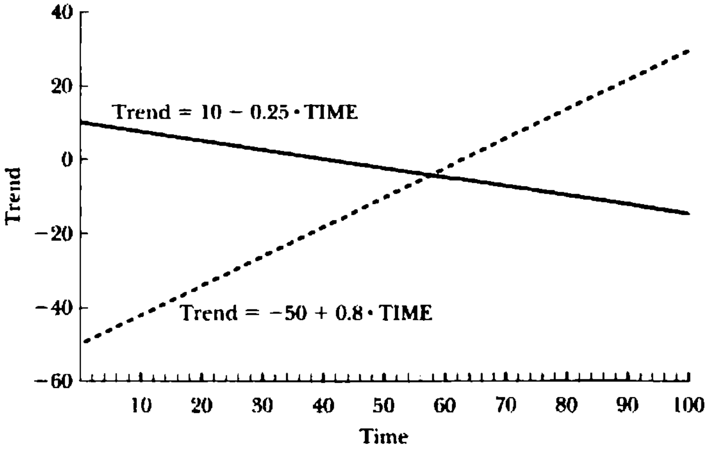
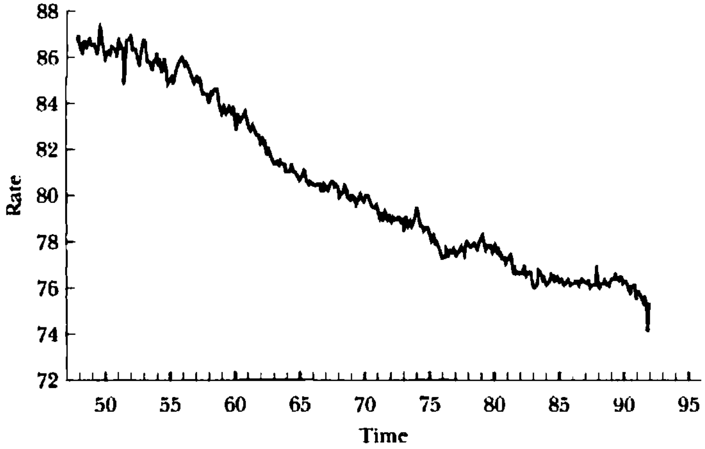


FIGURE 5.2
Increasing and Decreasing Linear Trends

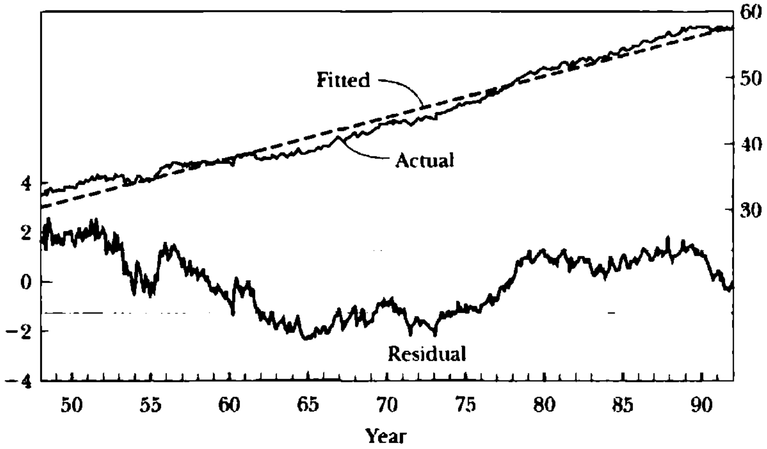
FIGURE 5.3
Labor Force Participation Rate, Males



In business, finance, and economics, linear trends are typically increasing, corresponding to growth, but such need not be the case. In Figure 5.3, for example, we show the U.S. labor force participation rate for *males* aged 16 and over, which displays linearly *decreasing* trend.

To provide a visual check of the adequacy of linear trends for the labor force participation rates, we show them with linear trends superimposed in Figures 5.4 and 5.5.[3] In each case, we show the actual participation rate series

FIGURE 5.4
Linear Trend, Female Labor Force Participation Rate



---

[3] Shortly we'll discuss how we estimated the trends. For now, just take them as given.

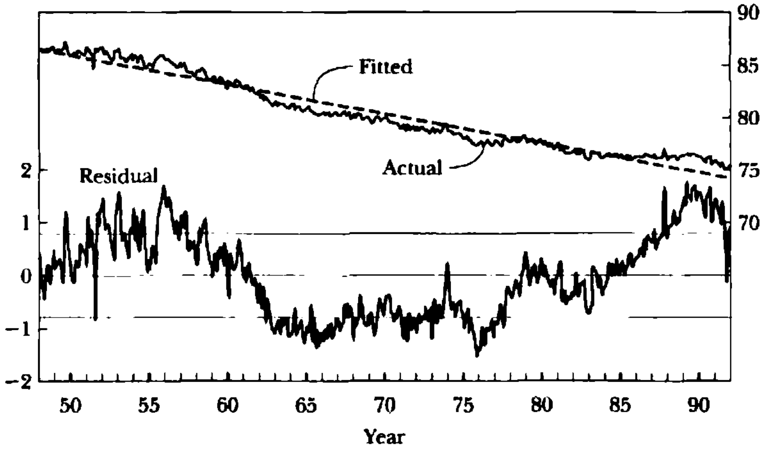


FIGURE 5.5
Linear Trend, Male Labor Force Participation Rate

together with the fitted trend, and we also show the residual—the deviation of the actual participation rate from the trend. The linear trends seem adequate. There are still obvious dynamic patterns in the residuals, but that's to be expected—persistent dynamic patterns are typically observed in the deviations of variables from trend.

Sometimes trend appears *nonlinear*, or curved, as, for example, when a variable increases at an increasing or decreasing rate. Ultimately, we don't require that trends be linear, only that they be *smooth*. Figure 5.6 shows the monthly volume of shares traded on the New York Stock Exchange (NYSE). Volume increases at an increasing rate; the trend is therefore nonlinear.

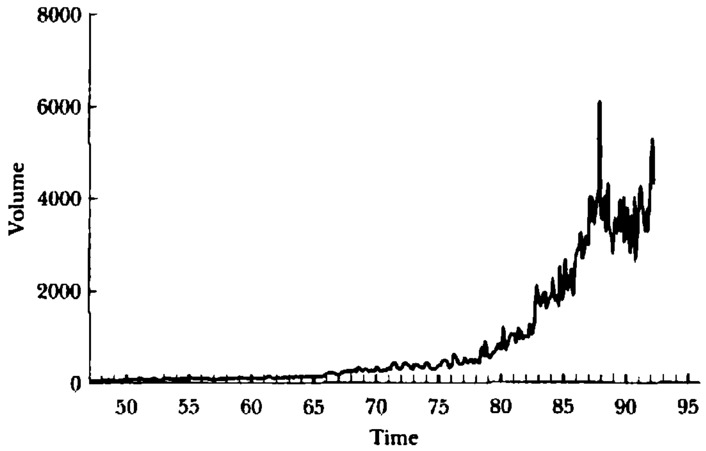
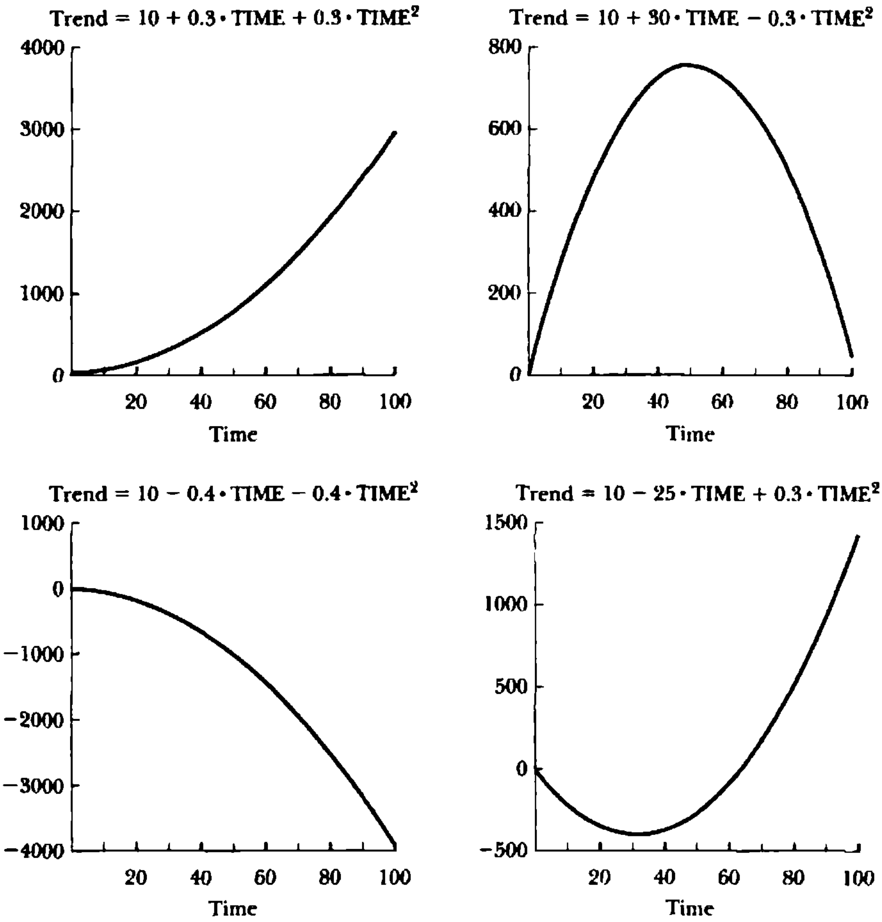


FIGURE 5.6
Volume on the New York Stock Exchange

FIGURE 5.7
Various Shapes of Quadratic Trends



Quadratic trend models can potentially capture nonlinearities such as those observed in the volume series. Such trends are quadratic, as opposed to linear, functions of time,

$$T_t = \beta_0 + \beta_1 \mathrm{TIME}_t + \beta_2 \mathrm{TIME}_t^2 .$$

Linear trend emerges as a special (and potentially restrictive) case when $\beta_2 = 0$. Higher-order polynomial trends are sometimes entertained, but it's important to use low-order polynomials to maintain smoothness.

A variety of different nonlinear quadratic trend shapes are possible, depending on the signs and sizes of the coefficients; we show several in Figure 5.7. In particular, if $\beta_1 > 0$ and $\beta_2 > 0$ as in the upper-left panel, the trend is monotonically, but nonlinearly, increasing, Conversely, if $\beta_1 < 0$ and $\beta_2 < 0$, the trend is monotonically decreasing. If $\beta_1 < 0$ and $\beta_2 > 0$, the trend has a U shape; and if $\beta_1 > 0$ and $\beta_2 < 0$, the trend has an inverted U shape. Keep in mind that quadratic trends are used to provide local approximations; one

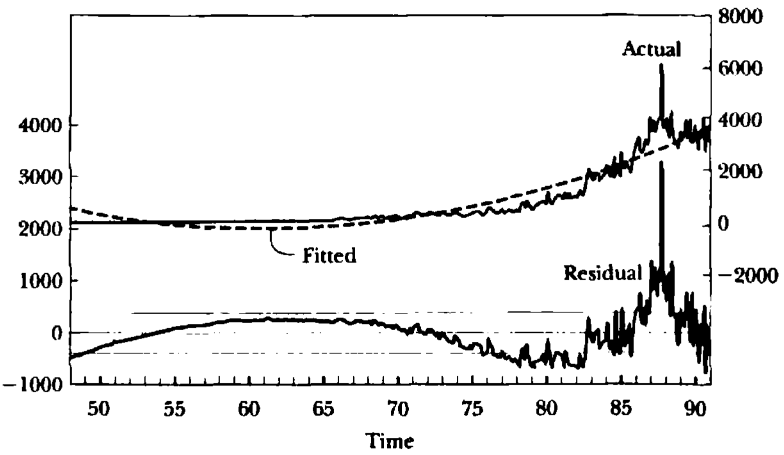


FIGURE 5.8
Quadratic Trend, Volume on the New York Stock Exchange

rarely has a "U-shaped" trend, for example. Instead, all of the data may lie on one or the other side of the U.

Figure 5.8 presents the stock market volume data with a superimposed quadratic trend. The quadratic trend fits better than the linear trend, but it still has some awkward features. The best-fitting quadratic trend is still a little more U-shaped than the volume data, resulting in an odd pattern of deviations from trend, as reflected in the residual series.

Other types of nonlinear trend are sometimes appropriate. Consider the NYSE volume series once again. In Figure 5.9, we show the *logarithm* of volume, the trend of which appears approximately *linear.*[4] This situation, in which a trend appears nonlinear in levels but linear in logarithms, is called

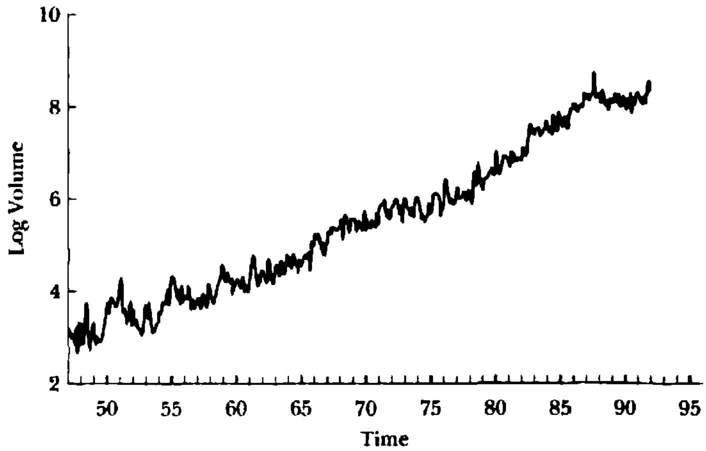
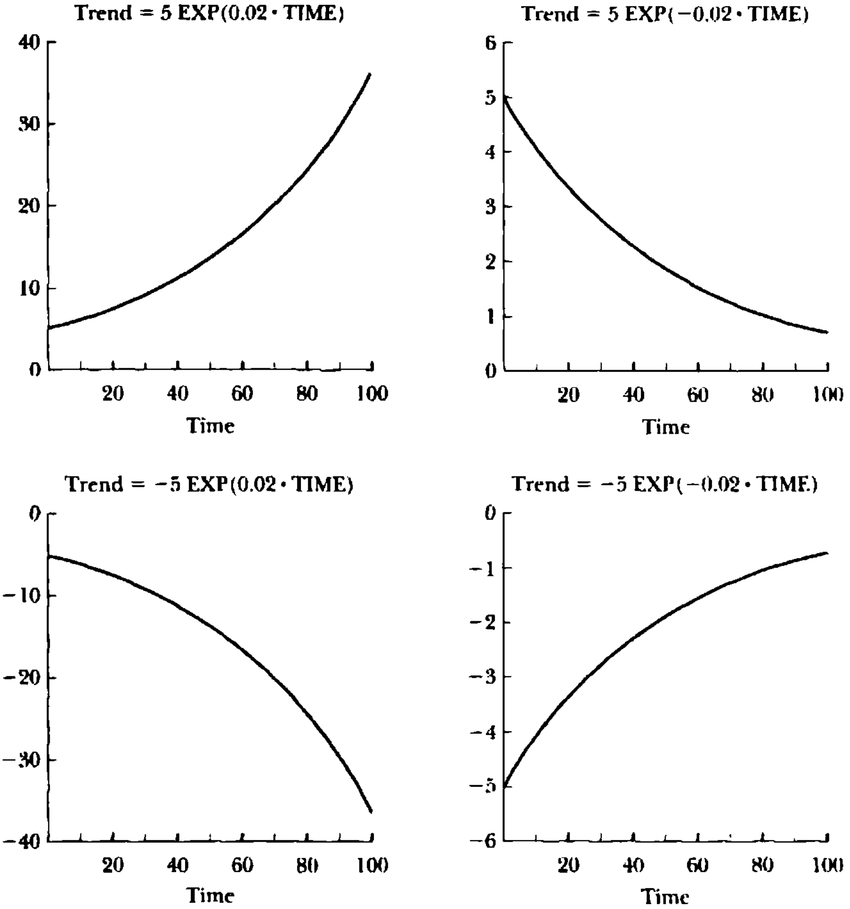


FIGURE 5.9
Log Volume on the New York Stock Exchange

---

[4] Throughout this book, logarithms are *natural* (base e) logarithms.

FIGURE 5.10
Various Shapes of Exponential Trends



exponential trend, or **log-linear trend,** and is very common in business, finance, and economics. That's because economic variables often display roughly constant growth rates (for example, 3% per year). If trend is characterized by constant growth at rate β_1, then we can write

$$T_t = \beta_0 e^{\beta_1 \text{TIME}_t} .$$

The trend is a nonlinear (exponential) function of time in levels, but in logarithms we have

$$\ln(T_t) = \ln(\beta_0) + \beta_1 \text{TIME}_t .$$

Thus, $\ln(T_t)$ is a linear function of time.

Figure 5.10 shows the variety of exponential trend shapes that can be obtained depending on the parameters. As with quadratic trend, depending on the signs and sizes of the parameter values, exponential trend can achieve a variety of patterns, increasing or decreasing at an increasing or decreasing rate.

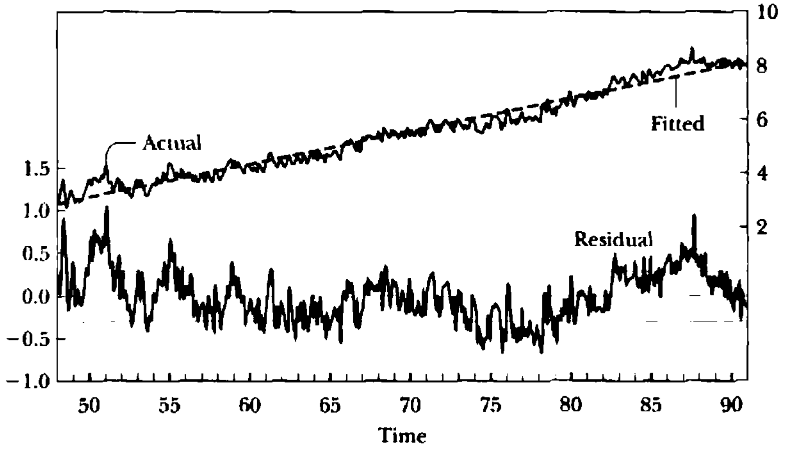


FIGURE 5.11 *Linear Trend, Log Volume on the New York Stock Exchange*

It's important to note that, although the same sorts of qualitative trend shapes can be achieved with quadratic and exponential trends, there are subtle differences between them. The nonlinear trends in some series are well approximated by quadratic trend, whereas the trends in other series are better approximated by exponential trend. We have already seen, for example, that although quadratic trend looked better than linear trend for the NYSE volume data, the quadratic fit still had some undesirable features. Let's see how an exponential trend compares. In Figure 5.11, we plot the *log* volume data with linear trend superimposed; the log-linear trend looks quite good. Equivalently, Figure 5.12 shows the volume data in *levels* with exponential trend superimposed; the exponential trend looks much better than did the quadratic.

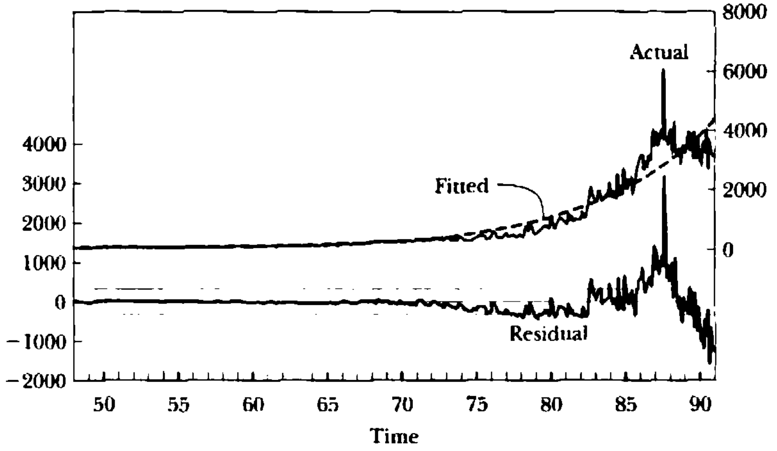


FIGURE 5.12 *Exponential Trend, Volume on the New York Stock Exchange*

2. Estimating Trend Models

Before we can estimate trend models, we need to create and store on the computer variables such as TIME and its square. Fortunately, we don't have to type the trend values (1, 2, 3, 4, . . .) in by hand; in most good software packages, a command exists to create the trend automatically, after which we can immediately compute derived variables such as the square of TIME, or TIME^2. Because, for example, $\text{TIME} = (1, 2, \ldots, T)$, $\text{TIME}^2 = (1, 4, \ldots, T^2)$; that is, $\text{TIME}_t^2 = t^2$.

We fit our various trend models to data on a time series y using **least-squares regression.** That is, we use a computer to find[5]

$$\hat{\theta} = \underset{\theta}{\operatorname{argmin}} \sum_{t=1}^{T} (y_t - T_t(\theta))^2 ,$$

where θ denotes the set of parameters to be estimated. A linear trend, for example, has $T_t(\theta) = \beta_0 + \beta_1 \text{TIME}_t$ and $\theta = (\beta_0, \beta_1)$, in which case the computer finds

$$(\hat{\beta}_0, \hat{\beta}_1) = \underset{\beta_0, \beta_1}{\operatorname{argmin}} \sum_{t=1}^{T} (y_t - \beta_0 - \beta_1 \text{TIME}_t)^2 .$$

Similarly, in the quadratic trend case, the computer finds

$$(\hat{\beta}_0, \hat{\beta}_1, \hat{\beta}_2) = \underset{\beta_0, \beta_1, \beta_2}{\operatorname{argmin}} \sum_{t=1}^{T} (y_t - \beta_0 - \beta_1 \text{TIME}_t - \beta_2 \text{TIME}_t^2)^2 .$$

We can estimate an exponential trend in two ways. First, we can proceed directly from the exponential representation and let the computer find

$$(\hat{\beta}_0, \hat{\beta}_1) = \underset{\beta_0, \beta_1}{\operatorname{argmin}} \sum_{t=1}^{T} (y_t - \beta_0 e^{\beta_1 \text{TIME}_t})^2 .$$

Alternatively, because the nonlinear exponential trend is nevertheless linear in logs, we can estimate it by regressing log y on an intercept and TIME. Thus, we let the computer find

$$(\hat{\beta}_0, \hat{\beta}_1) = \underset{\beta_0, \beta_1}{\operatorname{argmin}} \sum_{t=1}^{T} (\ln y_t - \ln \beta_0 - \beta_1 \text{TIME}_t)^2 .$$

Note that the fitted values from this regression are the fitted values of log y, so they must be exponentiated to get the fitted values of y.

---

[5] **Argmin** just means "the argument that minimizes." Least squares proceeds by finding the argument (in this case, the value of θ) that minimizes the sum of squared residuals; thus, the least squares estimator is the "argmin" of the sum of squared residuals function.

3. Forecasting Trend

Consider first the construction of point forecasts. Suppose we're presently at time T, and we want to use a trend model to forecast the h-step-ahead value of a series y. For illustrative purposes, we'll work with a linear trend, but the procedures are identical with more complicated trends. The linear trend model, which holds for any time t, is

$$y_t = \beta_0 + \beta_1 \mathrm{TIME}_t + \varepsilon_t \ .$$

In particular, at time $T + h$, the future time of interest,

$$y_{T+h} = \beta_0 + \beta_1 \mathrm{TIME}_{T+h} + \varepsilon_{T+h} \ .$$

Two future values of series appear on the right side of the equation, TIME_{T+h} and ε_{T+h}. If TIME_{T+h} and ε_{T+h} were known at time T, we could immediately crank out the forecast. In fact, TIME_{T+h} *is* known at time T, because the artificially constructed time variable is perfectly predictable; specifically, $\mathrm{TIME}_{T+h} = T + h$. Unfortunately ε_{T+h} is not known at time T, so we replace it with an optimal forecast of ε_{T+h} constructed using information only up to time T.[6] Under the assumption that ε is simply independent zero-mean random noise, the optimal forecast of ε_{T+h} for any future period is 0, yielding the point forecast[7]

$$y_{T+h,T} = \beta_0 + \beta_1 \mathrm{TIME}_{T+h} \ .$$

The subscript "$T + h, T$" on the forecast reminds us that the forecast is for time $T + h$ and is made at time T.

The point forecast formula just given is not of practical use, because it assumes known values of the trend parameters β_0 and β_1. But it's a simple matter to make it operational—we just replace unknown parameters with their least squares estimates, yielding

$$\hat{y}_{T+h,T} = \hat{\beta}_0 + \hat{\beta}_1 \mathrm{TIME}_{T+h} \ .$$

To form an interval forecast, we assume further that the trend regression disturbance is normally distributed, in which case a 95% interval forecast ignoring parameter estimation uncertainty is $y_{T+h,T} \pm 1.96\sigma$, where σ is the standard deviation of the disturbance in the trend regression.[8] To make this operational, we use $\hat{y}_{T+h,T} \pm 1.96\hat{\sigma}$, where $\hat{\sigma}$ is the standard error of the trend regression, an estimate of σ.

---

[6] More formally, we say that we're "projecting ε_{T+h} on the time-T information set," which we'll discuss in detail in Chapter 9.

[7] "Independent zero-mean random noise" is just a fancy way of saying that the regression disturbances satisfy the usual assumptions—they are identically and independently distributed.

[8] When we say that we ignore parameter estimation uncertainty, we mean that we use the estimated parameters as if they were the true values, ignoring the fact that they are only estimates and subject to sampling variability. Later we'll see how to account for parameter estimation uncertainty by using simulation techniques.

To form a density forecast, we again assume that the trend regression disturbance is normally distributed. Then, ignoring parameter estimation uncertainty, we have the density forecast $N(y_{T+h,T}, \sigma^2)$, where σ is the standard deviation of the disturbance in the trend regression. To make this operational, we use the density forecast $N(\hat{y}_{T+h,T}, \hat{\sigma}^2)$.

4. Selecting Forecasting Models Using the Akaike and Schwarz Criteria

We've introduced a number of trend models, but how do we select among them when fitting a trend to a specific series? What are the consequences, for example, of fitting a number of trend models and selecting the model with highest R^2? Is there a better way? This issue of **model selection** is of tremendous importance in all of forecasting, so we introduce it now.

It turns out that model selection strategies such as selecting the model with highest R^2 do *not* produce good out-of-sample forecasting models. Fortunately, however, a number of powerful modern tools exist to assist with model selection. Here we digress to discuss some of the available methods, which will be immediately useful in selecting among alternative trend models, as well as many other situations.

Most model selection criteria attempt to find the model with the smallest out-of-sample 1-step-ahead mean squared prediction error. The criteria we examine fit this general approach; the differences among criteria amount to different penalties for the number of degrees of freedom used in estimating the model (that is, the number of parameters estimated). Because all of the criteria are effectively estimates of out-of-sample mean square prediction error, they have a negative orientation—the smaller, the better.

First, consider the **mean squared error** (MSE),

$$\text{MSE} = \frac{\sum_{t=1}^{T} e_t^2}{T},$$

where T is the sample size and

$$e_t = y_t - \hat{y}_t,$$

where

$$\hat{y}_t = \hat{\beta}_0 + \hat{\beta}_1 \text{TIME}_t .$$

MSE is intimately related to two other diagnostic statistics routinely computed by regression software, the sum of squared residuals and R^2. Looking at the MSE formula reveals that the model with the smallest MSE is also the model with smallest sum of squared residuals, because scaling the sum of squared residuals by $1/T$ doesn't change the ranking. So selecting the model with the smallest MSE is equivalent to selecting the model with the smallest sum of squared residuals. Similarly, recall the formula for R^2,

$$R^2 = 1 - \frac{\sum_{t=1}^{T} e_t^2}{\sum_{t=1}^{T} (y_t - \bar{y})^2} .$$

The denominator of the ratio that appears in the formula is just the sum of squared deviations of y from its sample mean (the so-called total sum of squares), which depends only on the data, not on the particular model fit. Thus, selecting the model that minimizes the sum of squared residuals—which, as we saw, is equivalent to selecting the model that minimizes MSE—is also equivalent to selecting the model that maximizes R^2.

Selecting forecasting models on the basis of MSE or any of the equivalent forms discussed—that is, using in-sample MSE to estimate the out-of-sample 1-step-ahead MSE—turns out to be a bad idea. In-sample MSE *can't* rise when more variables are added to a model, and typically it will fall continuously as more variables are added. To see why, consider the fitting of polynomial trend models. In that context, the number of variables in the model is linked to the degree of the polynomial (call it p):

$$T_t = \beta_0 + \beta_1 \text{TIME}_t + \beta_2 \text{TIME}_t^2 + \cdots + \beta_p \text{TIME}_t^p .$$

We've already considered the cases of $p = 1$ (linear trend) and $p = 2$ (quadratic trend), but there's nothing to stop us from fitting models with higher powers of time included. As we include higher powers of time, the sum of squared residuals *can't* rise, because the estimated parameters are explicitly chosen to *minimize* the sum of squared residuals. The last-included power of time could always wind up with an estimated coefficient of 0; to the extent that the estimate is anything else, the sum of squared residuals must have fallen. Thus, the more variables we include in a forecasting model, the lower the sum of squared residuals will be, and therefore the lower MSE will be, and the higher R^2 will be. The reduction in MSE as higher powers of time are included in the model occurs even if they are in fact of no use in forecasting the variable of interest. Again, the sum of squared residuals can't rise, and because of sampling error, it's very unlikely that we'd get a coefficient of exactly 0 on a newly included variable even if the coefficient is 0 in population.

The effects described here go under various names, including **in-sample overfitting** and **data mining,** reflecting the idea that including more variables in a forecasting model won't necessarily improve its out-of-sample forecasting performance, although it will improve the model's "fit" on historical data. The upshot is that MSE is a biased estimator of **out-of-sample 1-step-ahead prediction error variance,** and the size of the bias increases with the number of variables included in the model. The direction of the bias is downward—in-sample MSE provides an overly optimistic (that is, too small) assessment of out-of-sample prediction error variance.

To reduce the bias associated with MSE and its relatives, we need to penalize for degrees of freedom used. Thus, let's consider the mean squared error corrected for degrees of freedom,

$$s^2 = \frac{\sum_{t=1}^{T} e_t^2}{T-k},$$

where k is the number of degrees of freedom used in model fitting,[9] and s^2 is just the usual unbiased estimate of the regression disturbance variance. That is, it is the square of the usual standard error of the regression. So selecting the model that minimizes s^2 is also equivalent to selecting the model that minimizes the standard error of the regression. Also, s^2 is intimately connected to the R^2 adjusted for degrees of freedom (the adjusted R^2, or $\bar{R}^2$). Recall that

$$\bar{R}^2 = 1 - \frac{\sum_{t=1}^{T} e_t^2/(T-k)}{\sum_{t=1}^{T}(y_t - \bar{y})^2/(T-1)} = 1 - \frac{s^2}{\sum_{t=1}^{T}(y_t - \bar{y})^2/(T-1)}.$$

The denominator of the $\bar{R}^2$ expression depends only on the data, not the particular model fit, so the model that minimizes s^2 is also the model that maximizes $\bar{R}^2$. In short, the strategies of selecting the model that minimizes s^2, or the model that minimizes the standard error of the regression, or the model that maximizes $\bar{R}^2$, are equivalent, and they *do* penalize for degrees of freedom used.

To highlight the degree-of-freedom penalty, let's rewrite s^2 as a penalty factor times the MSE,

$$s^2 = \left(\frac{T}{T-k}\right)\frac{\sum_{t=1}^{T} e_t^2}{T}.$$

Note in particular that including more variables in a regression will not necessarily lower s^2 or raise $\bar{R}^2$—the MSE will fall, but the degrees-of-freedom penalty will rise, so the product could go either way.

As with s^2, many of the most important forecast model selection criteria are of the form "penalty factor times MSE." The idea is simply that if we want to get an accurate estimate of the 1-step-ahead out-of-sample prediction error variance, we need to penalize the in-sample residual variance (the MSE) to reflect the degrees of freedom used. Two very important such criteria are the

---

[9] The degrees of freedom used in model fitting is simply the number of parameters estimated.

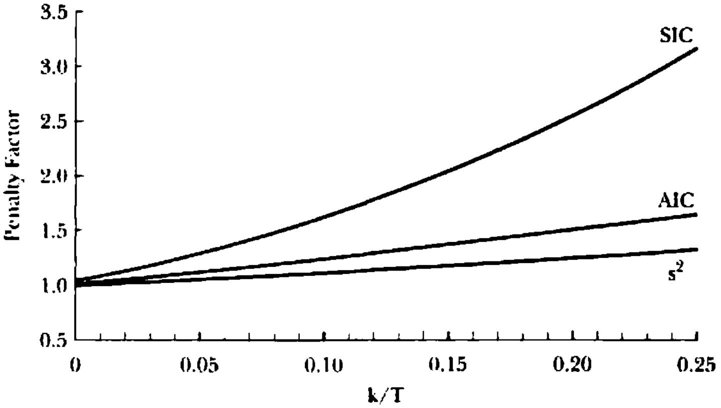


FIGURE 5.13 *Degrees-of-Freedom Penalties, Various Model Selection Criteria*

Akaike information criterion (AIC) and the Schwarz information criterion (SIC). Their formulas are

$$\mathrm{AIC} = e^{\left(\frac{2k}{T}\right)} \frac{\sum_{t=1}^{T} e_t^2}{T}$$

and

$$\mathrm{SIC} = T^{\left(\frac{k}{T}\right)} \frac{\sum_{t=1}^{T} e_t^2}{T}.$$

How do the penalty factors associated with MSE, s^2, AIC, and SIC compare in terms of severity? All of the penalty factors are functions of k/T, the number of parameters estimated per sample observation, and we can compare the penalty factors graphically as k/T varies. In Figure 5.13, we show the penalties as k/T moves from 0 to 0.25, for a sample size of $T = 100$. The s^2 penalty is small and rises slowly with k/T; the AIC penalty is a bit larger and still rises only slowly with k/T. The SIC penalty, on the other hand, is substantially larger and rises at a slightly increasing rate with k/T.

It's clear that the different criteria penalize degrees of freedom differently. In addition, we could propose many other criteria by altering the penalty. How, then, do we select among the criteria? More generally, what properties might we expect a "good" model selection criterion to have? Are s^2, AIC, and SIC "good" model selection criteria?

We evaluate model selection criteria in terms of a key property called **consistency**. A model selection criterion is consistent if the following conditions are met:

a. when the true model—that is, the **data-generating process (DGP)**—is among the models considered, the probability of selecting the true DGP approaches 1 as the sample size gets large; and

b. when the true model is *not* among those considered, so that it's impossible to select the true DGP, the probability of selecting the best *approximation* to the true DGP approaches 1 as the sample size gets large.[10]

Consistency is, of course, desirable. If the DGP is among those considered, then we'd hope that as the sample size gets large, we'd eventually select it. Of course, all of our models are false—they're intentional simplifications of a much more complex reality. Thus, the second notion of consistency is the more compelling.

MSE is inconsistent, because it doesn't penalize for degrees of freedom; that's why it's unattractive. s^2 does penalize for degrees of freedom but, as it turns out, not enough to render it a consistent model selection procedure. The AIC penalizes degrees of freedom more heavily than s^2, but it, too, remains inconsistent; even as the sample size gets large, the AIC selects models that are too large ("overparameterized"). The SIC, which penalizes degrees of freedom most heavily, *is* consistent.

The discussion thus far conveys the impression that SIC is unambiguously superior to AIC for selecting forecasting models, but such is not the case. Until now, we've implicitly assumed that either the true DGP or the best approximation to the true DGP is in the fixed set of models considered. In that case, SIC *is* a superior model selection criterion. However, a potentially more compelling view for forecasters is that both the true DGP and the best approximation to it are much more complicated than any model we fit, in which case we may want to expand the set of models we entertain as the sample size grows. We're then led to a different optimality property, called **asymptotic efficiency.** An asymptotically efficient model selection criterion chooses a sequence of models, as the sample size get large, whose 1-step-ahead forecast error variances approach the one that would be obtained using the true model with known parameters at a rate at least as fast as that of any other model selection criterion. The AIC, although inconsistent, *is* asymptotically efficient, whereas the SIC is not.

In practical forecasting, we usually report and examine both AIC and SIC. Most often they select the same model. When they don't, and in spite of the theoretical asymptotic efficiency property of AIC, I recommend use of the more parsimonious model selected by the SIC, other things being equal. This approach is in accord with the KISS principle of Chapter 3 and with the results of studies comparing out-of-sample forecasting performance of models selected by various criteria.

The AIC and SIC have enjoyed widespread popularity, but they are not universally applicable, and we're still learning about their performance in specific situations. However, the general principle that we need to correct somehow for degrees of freedom when estimating out-of-sample MSE on the basis of in-sample MSE *is* universally applicable. Judicious use of criteria like the AIC and SIC, in conjunction with knowledge about the nature of the system being forecast, is helpful in a variety of forecasting situations.

[10] Most model selection criteria—including all of those discussed here—assess goodness of approximation in terms of 1-step-ahead mean squared forecast error.

5. Application: Forecasting Retail Sales

We'll illustrate trend modeling with an application to forecasting U.S. current-dollar retail sales. The data are monthly from 1955.01 through 1994.12 and have been seasonally adjusted.[11] We'll use the period 1955.01–1993.12 to estimate our forecasting models, and we'll use the "holdout sample" 1994.01–1994.12 to examine their out-of-sample forecasting performance.

In Figure 5.14, we provide a time series plot of the retail sales data, which display a clear nonlinear trend and not much else. Cycles are probably present but are not easily visible, because they account for a comparatively minor share of the series' variation.

In Table 5.1, we show the results of fitting a linear trend model by regressing retail sales on a constant and a linear time trend. The trend appears highly significant as judged by the *p*-value of the *t*-statistic on the time trend, and the regression's R^2 is high. Moreover, the Durbin-Watson statistic indicates that the disturbances are positively serially correlated, so that the disturbance at any time *t* is positively correlated with the disturbance at time $t - 1$. In later chapters, we'll show how to model such **residual serial correlation** and exploit it for forecasting purposes, but for now we'll ignore it and focus only on the trend.[12]

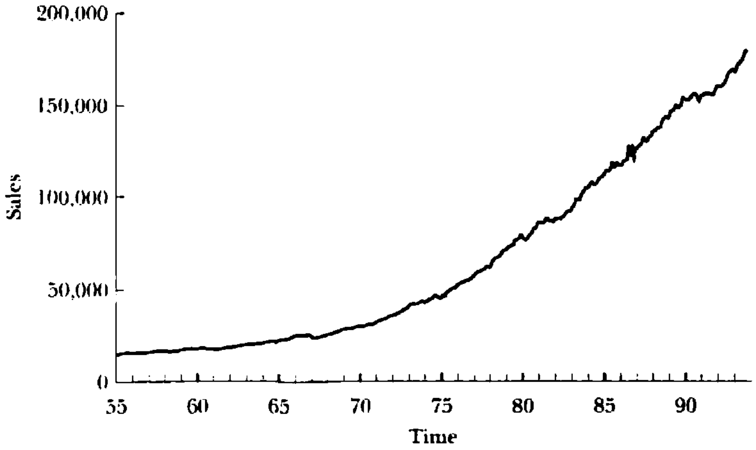


FIGURE 5.14
Retail Sales

---

[11] When we say that the data have been "seasonally adjusted," we simply mean that they have been smoothed in a way that eliminates seasonal variation. We'll discuss seasonality in detail in Chapter 6.

[12] Such residual serial correlation may, however, render the standard errors of estimated coefficients (and the associated *t*-statistics) untrustworthy. Here that's not a big problem, because it's visually obvious that trend is important in retail sales, but in other situations it may well be. Typically, when constructing forecasting models, we're concerned more with point estimation than with inference.

TABLE 5.1
Retail Sales, Linear Trend Regression

Dependent variable is RTRR.
Sample: 1955:01 1993:12
Included observations: 448

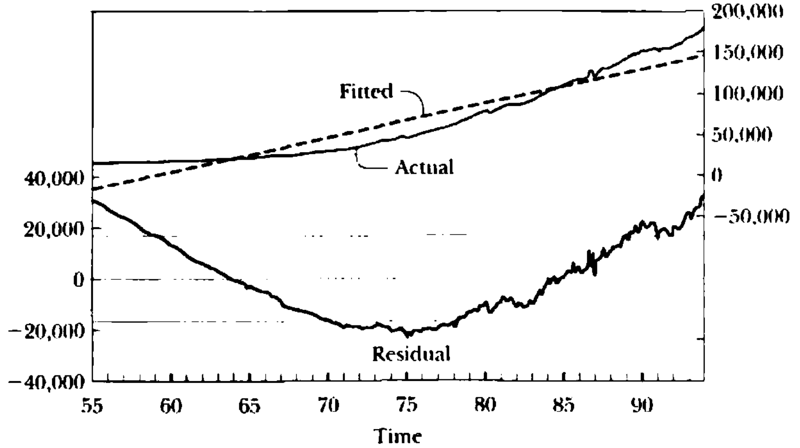
Variable	Coefficient	Std. Error	*t*-Statistic	Prob.
C	−16391.25	1469.177	−11.15676	0.0000
TIME	349.7731	5.428670	64.43073	0.0000

R^2	0.899076	Mean dependent var.	65630.56
Adjusted R^2	0.898859	SD dependent var.	49889.26
SE of regression	15866.12	Akaike info criterion	19.34815
Sum squared resid.	1.17E+11	Schwarz criterion	19.36587
Log likelihood	−5189.529	*F*-statistic	4151.319
Durbin-Watson stat.	0.004682	Prob(*F*-statistic)	0.000000

The residual plot in Figure 5.15 makes clear what's happening. The linear trend is simply inadequate, because the actual trend is nonlinear. That's one key reason why the residuals are so highly serially correlated—first the data are all above the linear trend, then below, and then above. Along with the residuals, we plot ±1 standard error of the regression, for visual reference.

Table 5.2 presents the results of fitting a quadratic trend model. Both the linear and quadratic terms appear highly significant.[13] R^2 is now almost 1.

FIGURE 5.15
Retail Sales, Linear Trend Residual Plot



---

[13] The earlier caveat regarding the effects of serial correlation on inference applies, however.

TABLE 5.2
Retail Sales, Quadratic Trend Regression

Dependent variable is RTRR.
Sample: 1955:01 1993:12
Included observations: 468

Variable	**Coefficient**	**Std. Error**	***t*-Statistic**	**Prob.**
C	18708.70	379.9566	49.23905	0.0000
TIME	−98.31130	3.741388	−26.27669	0.0000
TIME2	0.955404	0.007725	123.6754	0.0000

R^2	0.997022	Mean dependent var.	65630.56
Adjusted R^2	0.997010	SD dependent var.	49889.26
SE of regression	2728.205	Akaike info criterion	15.82919
Sum squared resid.	3.46E+09	Schwarz criterion	15.85578
Log likelihood	−4365.093	*F*-statistic	77848.80
Durbin-Watson stat.	0.151089	Prob(*F*-statistic)	0.000000

Figure 5.16 shows the residual plot, which now looks very nice, as the fitted nonlinear trend tracks the evolution of retail sales well. The residuals still display persistent dynamics (indicated as well by the still-low Durbin-Watson statistic), but there's little scope for explaining such dynamics with trend, because they're related to the business cycle, not the growth trend.

Now let's estimate a different type of nonlinear trend model, the exponential trend. First, we'll do it by OLS regression of the log of retail sales on a constant and linear time trend variable. We show the estimation results and

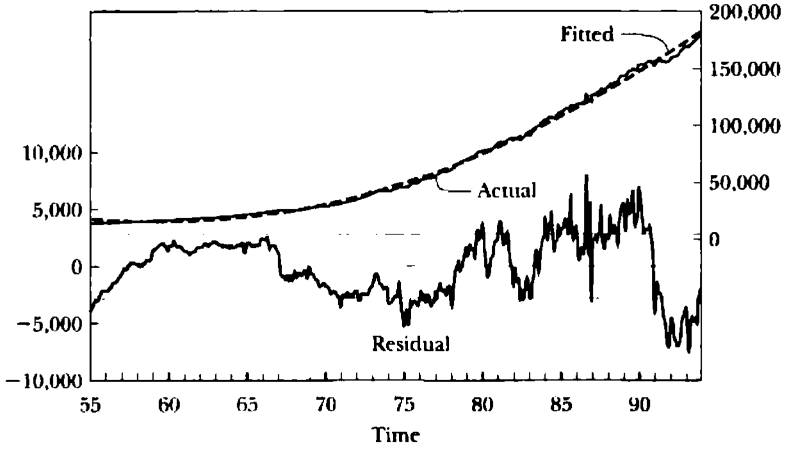


FIGURE 5.16
Retail Sales, Quadratic Trend Residual Plot

TABLE 5.3
Retail Sales, Log-Linear Trend Regression

Dependent variable is LRTRR.
Sample: 1955:01 1993:12
Included observations: 468

Variable	Coefficient	Std. Error	t-Statistic	Prob.
C	9.389975	0.008508	1103.684	0.0000
TIME	0.005931	3.14E-05	188.6541	0.0000

R^2	0.987076	Mean dependent var.	10.78072
Adjusted R^2	0.987048	SD dependent var.	0.807325
SE of regression	0.091879	Akaike info criterion	−4.770302
Sum squared resid.	3.933853	Schwarz criterion	−4.752573
Log likelihood	454.1874	F-statistic	35590.36
Durbin-Watson stat.	0.019949	Prob(F-statistic)	0.000000

residual plot in Table 5.3 and Figure 5.17. As with the quadratic nonlinear trend, the exponential nonlinear trend model seems to fit well, apart from the low Durbin-Watson statistic.

In sharp contrast to the results of fitting a linear trend to retail sales, which were poor, the results of fitting a linear trend to the *log* of retail sales seem much improved. But it's hard to compare the log-linear trend model with the linear and quadratic models because they're in levels, not logs, which renders diagnostic statistics like R^2 and the standard error of the regression incomparable. One way around this problem is to estimate the exponential trend

FIGURE 5.17
Retail Sales, Log-Linear Trend Residual Plot

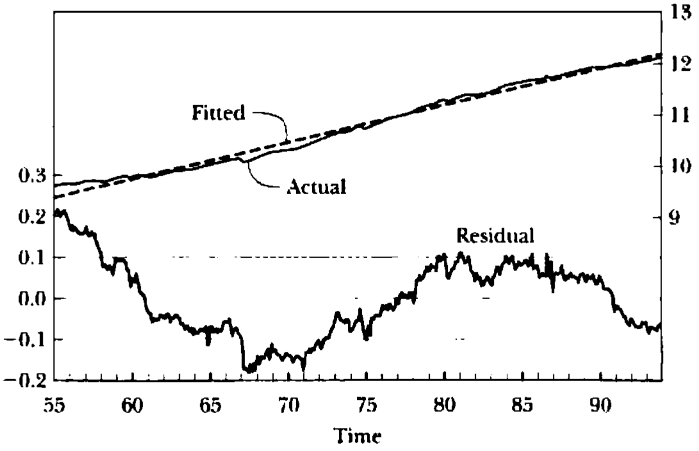


TABLE 5.4
Retail Sales, Exponential Trend Regression

Dependent variable is RTRR.
Sample: 1955:01 1993:12
Included observations: 468
Convergence achieved after 1 iteration
RTRR=C(1)*EXP(C(2)*TIME)

	Coefficient	**Std. Error**	***t*-Statistic**	**Prob.**
C(1)	11967.80	177.9598	67.25003	0.0000
C(2)	0.005944	3.77E-05	157.7469	0.0000

R^2	0.988796	Mean dependent var.	65630.56
Adjusted R^2	0.988772	SD dependent var.	49889.26
SE of regression	5286.406	Akaike info criterion	17.15005
Sum squared resid.	1.30E+10	Schwarz criterion	17.16778
Log likelihood	−4675.175	*F*-statistic	41126.02
Durbin-Watson stat.	0.040527	Prob(*F*-statistic)	0.000000

model directly in levels, using nonlinear least squares. In Table 5.4 and Figure 5.18, we show the nonlinear least-squares estimation results and residual plot for the exponential trend model. The diagnostic statistics and residual plot indicate that the exponential trend fits better than the linear but worse than the quadratic.

Thus far, we've been informal in our comparison of the linear, quadratic, and exponential trend models for retail sales. We've noticed, for example, that the quadratic trend seems to fit the best. The quadratic trend model, however,

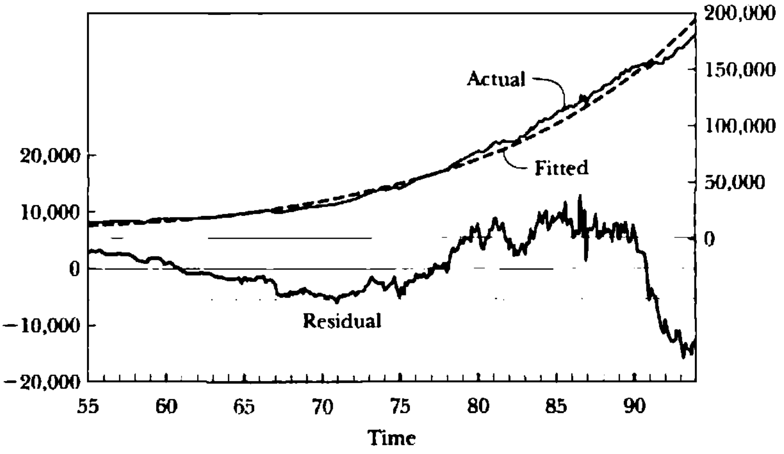


FIGURE 5.18
Retail Sales, Exponential Trend Residual Plot

TABLE 5.5 *Model Selection Criteria, Linear, Quadratic, and Exponential Trend Models*

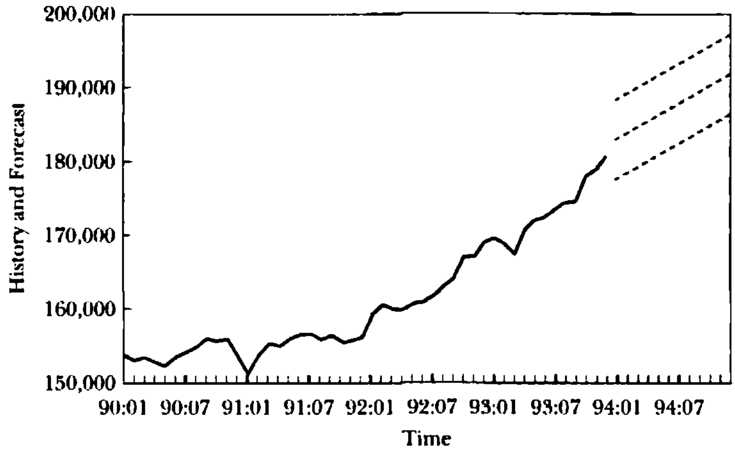
	Linear Trend	Quadratic Trend	Exponential Trend
AIC	19.35	15.83	17.15
SIC	19.37	15.86	17.17

contains one more parameter than the other two, so it's not surprising that it fits a little better, and there's no guarantee that its better fit on historical data will translate into better out-of-sample forecasting performance. (Recall the KISS principle.) To settle on a final model, we examine the AIC or SIC, which are summarized in Table 5.5 for the three trend models.[14] Both the AIC and SIC indicate that nonlinearity is important in the trend, as both rank the linear trend last. Both, moreover, favor the quadratic trend model. So let's use the quadratic trend model.

Figure 5.19 shows the history of retail sales, 1990.01–1993.12, together with out-of-sample point and 95% interval extrapolation forecasts, 1994.01–1994.12. The point forecasts look reasonable. The interval forecasts are computed under the (incorrect) assumption that the deviation of retail sales from trend is random noise, which is why they're of equal width throughout. Nevertheless, they look reasonable.

In Figure 5.20, we show the history of retail sales through 1993, the quadratic trend forecast for 1994, *and* the realization for 1994. The forecast is quite good, as the realization hugs the forecasted trend line quite closely. All of the realizations, moreover, fall inside the 95% forecast interval.

FIGURE 5.19 *Retail Sales: History, 1990.01–1993.12; and Quadratic Trend Forecast, 1994.01–1994.12*



---

[14] It's important that the exponential trend model be estimated in levels, in order to maintain comparability of the exponential trend model AIC and SIC with those of the other trend models.

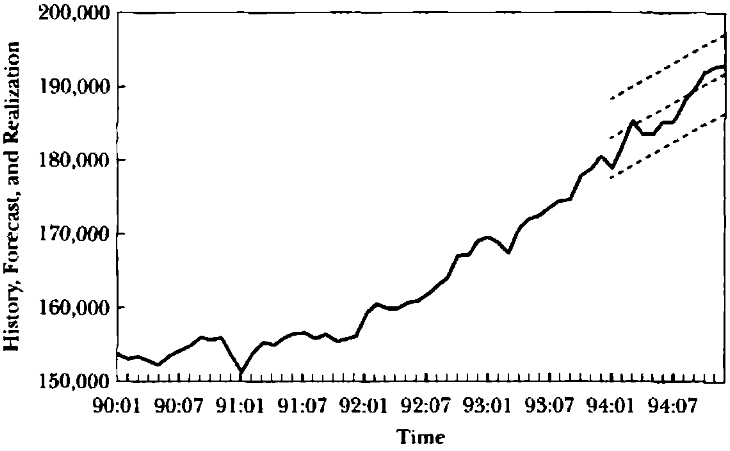


FIGURE 5.20
Retail Sales: History, 1990.01–1993.12; and Quadratic Trend Forecast and Realization, 1994.01–1994.12

For comparison, we examine the forecasting performance of a simple linear trend model. Figure 5.21 presents the history of retail sales and the out-of-sample point and 95% interval extrapolation forecasts for 1994. The point forecasts look very strange. The huge drop forecasted relative to the historical sample path occurs because the linear trend is far below the sample path by the end of the sample. The confidence intervals are very wide, reflecting the large standard error of the linear trend regression relative to the quadratic trend regression.

Finally, Figure 5.22 shows the history, the linear trend forecast for 1994, and the realization. The forecast is terrible—far below the realization. Even the

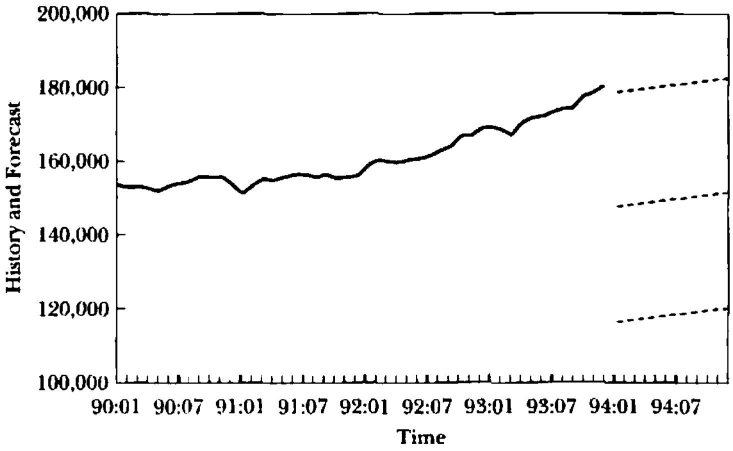
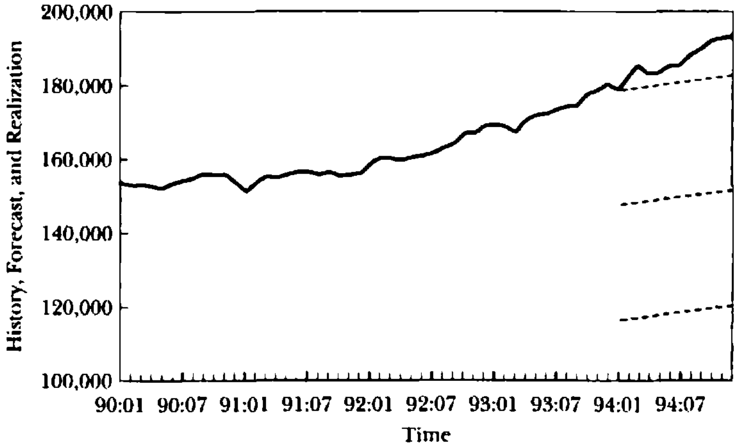


FIGURE 5.21
Retail Sales: History, 1990.01–1993.12; and Linear Trend Forecast, 1994.01–1994.12

FIGURE 5.22 *Retail Sales: History, 1990.01–1993.12; and Linear Trend Forecast and Realization, 1994.01–1994.12*



very wide interval forecasts fail to contain the realizations. The reason for the failure of the linear trend forecast is that the forecasts (point and interval) are computed under the assumption that the linear trend model is actually the true DGP, whereas in fact the linear trend model is a very poor approximation to the trend in retail sales.

Exercises, Problems, and Complements

1. (Calculating forecasts from trend models) You work for the International Monetary Fund in Washington, D.C., monitoring Singapore's real consumption expenditures. Using a sample of real consumption data (measured in billions of 2005 Singapore dollars), y_t, $t = 1990{:}Q1, \ldots, 2006{:}Q4$, you estimate the linear consumption trend model, $y_t = \beta_0 + \beta_1 \mathrm{TIME}_t + \varepsilon_t$, where $\varepsilon_t \sim N(0, \sigma^2)$, obtaining the estimates $\hat{\beta}_0 = 0.51$, $\hat{\beta}_1 = 2.30$, and $\sigma^2 = 16$. Based on your estimated trend model, construct feasible point, interval, and density forecasts for $2010{:}Q1$.

2. (Identifying and testing trend models) In 1965, Intel cofounder Gordon Moore predicted that the number of transistors that one could place on a square inch integrated circuit would double every 12 months.
   a. What sort of trend is this?
   b. Given a monthly series containing the number of transistors per square inch for the latest integrated circuit, how would you test Moore's prediction? How would you test the currently accepted form of Moore's Law—namely, that the number of transistors actually doubles every 18 months?

3. (Understanding model selection criteria) You are tracking and forecasting the earnings of a new company developing and applying proprietary nanotechnology. The earnings are trending upward. You fit linear, quadratic, and exponential

trend models, yielding sums of squared residuals of 4352, 2791, and 2749, respectively. Which trend model would you select, and why?

4. (Mechanics of trend estimation and forecasting) Obtain from the web an upward-trending monthly series that interests you. Choose your series such that it spans at least 10 years and ends at the end of a year (that is, in December).
   a. What is the series, and why does it interest you? Produce a time series plot of it. Discuss.
   b. Fit linear, quadratic, and exponential trend models to your series. Discuss the associated diagnostic statistics and residual plots.
   c. Select a trend model using the AIC and one using the SIC. Do the selected models agree? If not, which do you prefer?
   d. Use your preferred model to forecast each of the 12 months of the next year. Discuss.
   e. The *residuals* from your fitted model are effectively a *detrended* version of your original series. Why? Plot them and discuss.

5. (Properties of polynomial trends) Consider a sixth-order deterministic **polynomial trend:**

$$T_t = \beta_0 + \beta_1 \mathrm{TIME}_t + \beta_2 \mathrm{TIME}_t^2 + \cdots + \beta_6 \mathrm{TIME}_t^6 .$$

   a. How many local maxima or minima may such a trend display?
   b. Plot the trend for various values of the parameters to reveal some of the different possible trend shapes.
   c. Is this an attractive trend model in general? Why or why not?
   d. Fit the sixth-order polynomial trend model to the NYSE volume series. How does it perform in that particular case?

6. (Specialized nonlinear trends) The **logistic trend** is

$$T_t = \frac{1}{a + br^t} ,$$

with $0 < r < 1$.
   a. Display the trend shape for various a and b values. When might such a trend shape be useful?
   b. Can you think of other specialized situations in which other specialized trend shapes might be useful? Produce mathematical formulas for the additional specialized trend shapes you suggest.

7. (Moving average smoothing for trend estimation) The trend regression technique is one way to estimate and forecast trend. Another way to estimate trend is by **smoothing** techniques, which we briefly introduce here. We'll focus on three: **two-sided moving averages, one-sided moving averages,** and **one-sided weighted moving averages.** Here we present them as ways to estimate and examine the trend in a time series; later we'll see how they can actually be used to *forecast* time series.

   Denote the original data by $\{y_t\}_{t=1}^{T}$ and the smoothed data by $\{\bar{y}_t\}$. Then the two-sided moving average is $\bar{y}_t = (2m+1)^{-1}\sum_{i=-m}^{m} y_{t-i}$, the one-sided moving average is $\bar{y}_t = (m+1)^{-1}\sum_{i=0}^{m} y_{t-i}$, and the one-sided weighted moving average is $\bar{y}_t = \sum_{i=0}^{m} w_i y_{t-i}$, where the w_i are weights and m is an integer chosen by the user.

The "standard" one-sided moving average corresponds to a one-sided weighted moving average with all weights equal to $(m + 1)^{-1}$.

a. For each of the smoothing techniques, discuss the role played by m. What happens as m gets very large? Very small? In what sense does m play a role similar to p, the order of a polynomial trend?
b. If the original data runs from time 1 to time T, over what range can smoothed values be produced using each of the three smoothing methods? What are the implications for **real-time (on-line) smoothing** versus **ex post (off-line) smoothing?**
c. You've been hired as a consultant by ICSB, a major international bank, to advise its management on trends in North American and European stock markets and to help them to allocate their capital. You have extracted from your database the recent history of EUROStar, an index of 11 major European stock markets. Smooth the EUROStar data using equally weighted one-sided and two-sided moving averages, for a variety of m values, until you have found values of m that work well. What do we mean by "work well"? Must the chosen value of m be the same for the one- and two-sided smoothers? For your chosen m values, plot the two-sided smoothed series against the actual, and plot the one-sided smoothed series against the actual. Do you notice any systematic difference in the relationship of the smoothed to the actual series depending on whether you do a two-sided or one-sided smooth? Explain.
d. Moving average procedures can also be used to **detrend** a series—we simply subtract the estimated trend from the series. Sometimes, but not usually, it's appropriate and desirable to detrend a series before modeling and forecasting it. Why might it sometimes be appropriate? Why is it not usually appropriate?

8. (Bias corrections when forecasting from logarithmic models)
   a. In Chapter 3, we introduced squared error loss, $L(e) = e^2$. A popular measure of forecast accuracy is out-of-sample mean squared error, $\text{MSE} = E(e^2)$.[15] The more accurate the forecast, the smaller is MSE. Show that MSE is equal to the sum of the variance of the error and the square of the mean error.
   b. A forecast is *unbiased* if the mean forecast error is 0. Why might unbiased forecasts be desirable? Are they *necessarily* desirable?
   c. Suppose that $(\log y)_{t+h,t}$ is an unbiased forecast of $(\log y)_{t+h}$. Then $\exp((\log y)_{t+h,t})$ is a *biased* forecast of y_{t+h}. More generally, if $(f(y))_{t+h,t}$ is an unbiased forecast of $(f(y))_{t+h}$, then $f^{-1}((f(y))_{t+h,t})$ is a biased forecast of y_{t+h}, for the arbitrary nonlinear function f. Why? (*Hint:* Is the expected value of a nonlinear function of the random variable the same as the nonlinear function of the expected value?)
   d. Various "corrections" for the bias $\exp((\log y)_{t+h,t})$ have been proposed. In practice, however, **bias corrections** may increase the variance of the forecast error even if they succeed in reducing bias. Why? (*Hint:* In practice, the corrections involve estimated parameters.)
   e. In practice, will bias corrections necessarily reduce the forecast MSE? Why or why not?

---

[15] The MSE introduced earlier in the context of model selection is the mean of the *in-sample* residuals, as opposed to out-of-sample prediction errors. The distinction is crucial.

9. (Model selection for long-horizon forecasting) Suppose that you want to forecast monthly inventory of Lamborgini autos at an exclusive Manhattan dealership.
   a. Using the true data-generating process is best for forecasting at any horizon. Unfortunately, we never know the true data-generating process! All our models are approximations to the true but unknown data-generating process, in which case the best forecasting model may change with the horizon. Why?
   b. At what horizon are the forecasts generated by models selected by the AIC and SIC likely to be most accurate? Why?
   c. How might you proceed to select a 1-month-ahead forecasting model? A 2-month-ahead model? A 3-month-ahead model? A 4-month-ahead model?
   d. What are the implications of your answer for construction of an extrapolation forecast, at horizons 1-month-ahead through 4-months-ahead?
   e. In constructing our extrapolation forecasts for retail sales, we used the AIC and SIC to select one model, which we then used to forecast all horizons. Why do you think we didn't adopt a more sophisticated strategy?

10. (The variety of "information criteria" reported across software packages) Some authors, and software packages, examine and report the logarithms of the AIC and SIC as

$$\ln(\mathrm{AIC}) = \ln\left(\frac{\sum_{t=1}^{T} e_t^2}{T}\right) + \left(\frac{2k}{T}\right)$$

and

$$\ln(\mathrm{SIC}) = \ln\left(\frac{\sum_{t=1}^{T} e_t^2}{T}\right) + \frac{k\ln(T)}{T}.$$

The practice is so common that log(AIC) and log(SIC) are often simply called the "AIC" and "SIC." AIC and SIC must be greater than 0, so log(AIC) and log(SIC) are always well defined and can take on any real value. Other authors and packages use other variants, based, for example, on the value of the maximized likelihood or log likelihood function. Some software packages have even changed definitions of AIC and SIC across releases! The important insight, however, is that although these variations will of course change the numerical values of AIC and SIC produced by your computer, they will not change the *rankings* of models under the various criteria. Consider, for example, selecting among three models. If $\mathrm{AIC}_1 < \mathrm{AIC}_2 < \mathrm{AIC}_3$, then it must be true as well that $\ln(\mathrm{AIC}_1) < \ln(\mathrm{AIC}_2) < \ln(\mathrm{AIC}_3)$, so we would select model 1 regardless of the "definition" of the information criterion used.

Bibliographical and Computational Notes

The AIC and SIC trace at least to Akaike (1974) and Schwarz (1978). Granger, King, and White (1995) provide insightful discussion of consistency of model selection criteria, and the key (and difficult) reference on efficiency is Shibata (1980). Engle and Brown (1986) find that criteria with comparatively harsh degrees-of-freedom penalties (for example, the SIC) select the best forecasting models.

Kennedy (1992) reviews a number of corrections for the bias in $\exp((\log y)_{t+h,t})$.

A number of authors have investigated the use of multiple models for multiple horizons, including Findley (1983) and Tiao and Tsay (1994). Findley (1985) develops criteria for selection of multi-step-ahead forecasting models.

Concepts for Review

- Trend
- Deterministic trend
- Stochastic trend
- Time dummy
- Regression intercept
- Regression slope
- Quadratic trend
- Exponential trend
- Log-linear trend
- Least-squares regression
- Argmin
- Model selection
- Mean squared error
- In-sample overfitting
- Data mining
- Out-of-sample 1-step-ahead prediction error variance
- Consistency
- Data-generating process (DGP)
- Asymptotic efficiency
- Residual serial correlation
- Polynomial trend
- Logistic trend
- Smoothing
- Two-sided moving average
- One-sided moving average
- One-sided weighted moving average
- Real-time (on-line) smoothing
- Ex post (off-line) smoothing
- Detrending
- Bias correction

References and Additional Readings

Akaike, H. (1974). "A New Look at the Statistical Model Identification." *IEEE Transactions on Automatic Control*, AC-19, 716–723.

Engle, R. F., and Brown, S. J. (1986). "Model Selection for Forecasting." *Applied Mathematics and Computation*, 20, 313–327.

Findley, D. F. (1983). "On the Use of Multiple Models for Multi-Period Forecasting." *Proceedings of the American Statistical Association, Business and Economic Statistics Section*, 528–531.

Findley, D. F. (1985). "Model Selection for Multi-Step-Ahead Forecasting." In *Identification and System Parameter Estimation*, 7th IFAC/FORS Symposium, 1039–1044.

Granger, C. W. J., King, M. L., and White, H. (1995). "Comments on the Testing of Economic Theories and the Use of Model Selection Criteria." *Journal of Econometrics*, 67, 173–187.

Schwarz, G. (1978). "Estimating the Dimension of a Model." *Annals of Statistics*, 6, 461–464.

Shibata, R. (1980). "Asymptotically Efficient Selection of the Order of the Model for Estimating the Parameters of a Linear Process." *Annals of Statistics*, 8, 147–164.

Tiao, G. C., and Tsay, R. S. (1994). "Some Advances in Non-Linear and Adaptive Modeling in Time Series." *Journal of Forecasting*, 13, 109–131.

Modeling and Forecasting Seasonality

1. The Nature and Sources of Seasonality

In the last chapter, we focused on the trends; now we'll focus on **seasonality.** A *seasonal* pattern is one that repeats itself every year.[1] The annual repetition can be *exact,* in which case we speak of **deterministic seasonality,** or approximate, in which case we speak of **stochastic seasonality.** Just as we focused exclusively on deterministic trend in Chapter 5, reserving stochastic trend for subsequent treatment, so shall we focus exclusively on deterministic seasonality here.

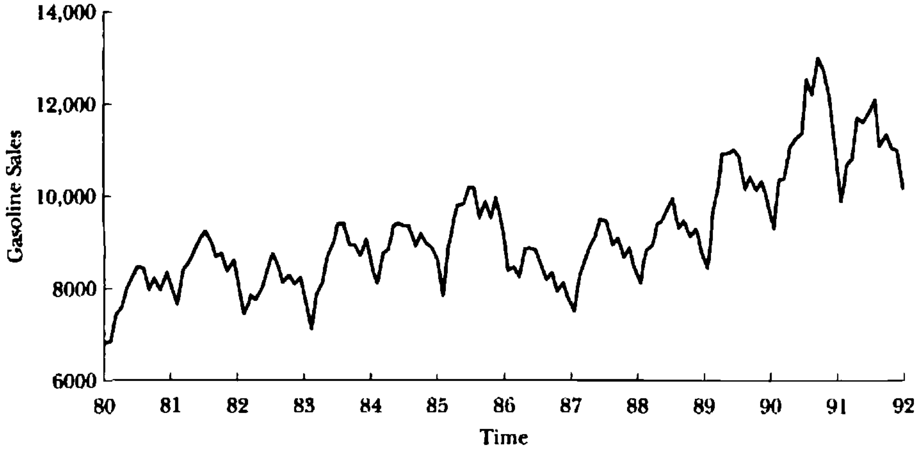
Seasonality arises from links of technologies, preferences, and institutions to the calendar. The weather (for example, daily high temperature in Tokyo) is a trivial but very important seasonal series, as it's always hotter in the summer than in the winter. Any technology that involves the weather, such as production of agricultural commodities, is likely to be seasonal as well.

Preferences may also be linked to the calendar. Consider, for example, gasoline sales. In Figure 6.1, we show monthly U.S. current-dollar gasoline sales, 1980.01–1992.01. People want to do more vacation travel in the summer, which tends to increase both the price and quantity of summertime gasoline sales, both of which feed into higher current-dollar sales.

---

[1] Note, therefore, that seasonality is impossible, and thus not an issue, in data recorded once per year or less often than once per year.

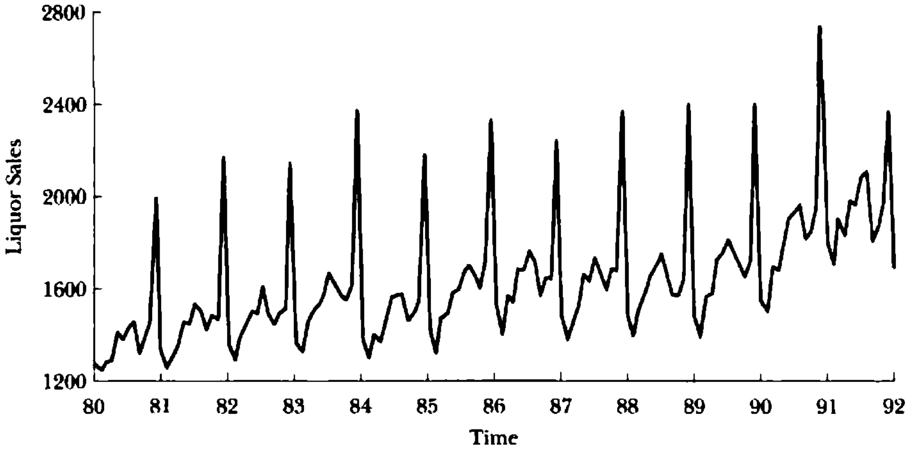
FIGURE 6.1
Gasoline Sales



Finally, social institutions that are linked to the calendar, such as holidays, are responsible for seasonal variation in a variety of series. Purchases of retail goods skyrocket, for example, every Christmas season. In Figure 6.2, we plot monthly U.S. current-dollar liquor sales, 1980.01–1992.01, which are very high in November and December. In contrast, sales of durable goods fall in December, as holiday purchases tend to be nondurables. This emerges clearly in Figure 6.3, in which we show monthly U.S. current-dollar durable goods sales, 1980.01–1992.01.

You might imagine that, although certain series are seasonal for obvious reasons, seasonality is nevertheless uncommon. On the contrary, and perhaps

FIGURE 6.2
Liquor Sales



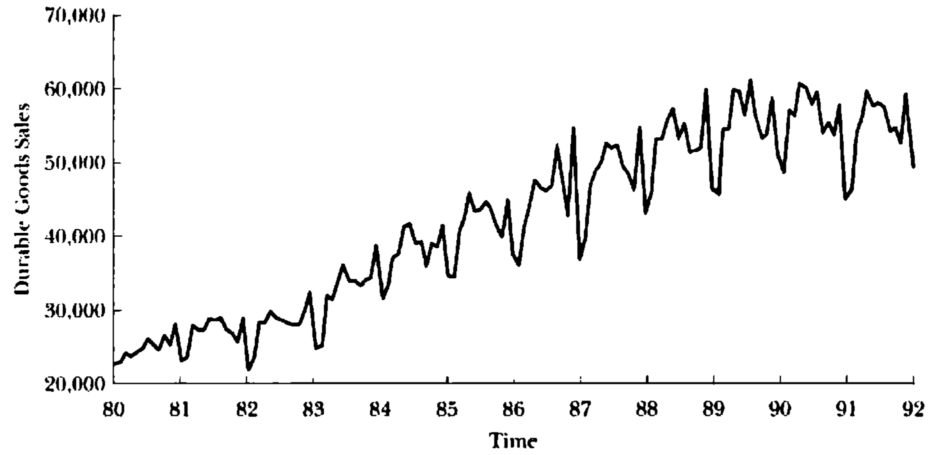


FIGURE 6.3
Durable Goods Sales

surprisingly, seasonality is pervasive in business and economics. Many industrialized economies, for example, expand briskly every fourth quarter and contract every first quarter.

One way to deal with seasonality in a series is simply to remove it and then to model and forecast the **seasonally adjusted time series.**[2] This strategy is perhaps appropriate in certain situations, such as when interest centers explicitly on forecasting **nonseasonal fluctuations,** as is often the case in macroeconomics. Seasonal adjustment is often inappropriate in business forecasting situations, however, precisely because interest typically centers on forecasting *all* the variation in a series, not just the nonseasonal part. If seasonality is responsible for a large part of the variation in a series of interest, the last thing a forecaster wants to do is discard it and pretend it isn't there.

2. Modeling Seasonality

A key technique for modeling seasonality is **regression on seasonal dummies.** Let s be the number of seasons in a year. Normally we'd think of four seasons in a year, but that notion is too restrictive for our purposes. Instead, think of s as the number of observations on a series in each year. Thus, $s = 4$ if we have quarterly data, $s = 12$ if we have monthly data, $s = 52$ if we have weekly data, and so forth.

---

[2] Removal of seasonality is called **seasonal adjustment.**

Now let's construct **seasonal dummy variables,** which indicate which season we're in. If, for example, there are four seasons, we create

$$
\begin{aligned}
D_1 &= (1, 0, 0, 0, 1, 0, 0, 0, 1, 0, 0, 0, \ldots)\ ; \\
D_2 &= (0, 1, 0, 0, 0, 1, 0, 0, 0, 1, 0, 0, \ldots)\ ; \\
D_3 &= (0, 0, 1, 0, 0, 0, 1, 0, 0, 0, 1, 0, \ldots)\ ; \\
D_4 &= (0, 0, 0, 1, 0, 0, 0, 1, 0, 0, 0, 1, \ldots)\ .
\end{aligned}
$$

D_1 indicates whether we're in the first quarter (it's 1 in the first quarter and 0 otherwise), D_2 indicates whether we're in the second quarter (it's 1 in the second quarter and 0 otherwise), and so on. At any given time, we can be in only one of the four quarters, so one seasonal dummy is 1, and all others are 0.

The pure seasonal dummy model is

$$y_t = \sum_{i=1}^{s} \gamma_i D_{it} + \varepsilon_t\ .$$

Effectively, we're just regressing on an intercept, but we allow for a different intercept in each season. Those different intercepts, the γ_i's, are called the *seasonal factors;* they summarize the seasonal pattern over the year. In the absence of seasonality, the γ_i's are all the same, so we can drop all the seasonal dummies and instead simply include an intercept in the usual way.

Instead of including a full set of s seasonal dummies, we can include any $s - 1$ seasonal dummies and an intercept. Then the constant term is the intercept for the omitted season, and the coefficients on the seasonal dummies give the seasonal increase or decrease relative to the omitted season. In no case, however, should we include s seasonal dummies *and* an intercept. Including an intercept is equivalent to including a variable in the regression whose value is always 1, but note that the full set of s seasonal dummies sums to a variable whose value is always 1. Thus, inclusion of an intercept and a full set of seasonal dummies produces perfect multicollinearity, and your computer will scream at you if you run such a regression. (Try it!)

Trend may be included as well, in which case the model is[3]

$$y_t = \beta_1 \mathrm{TIME}_t + \sum_{i=1}^{s} \gamma_i D_{it} + \varepsilon_t\ .$$

In fact, you can think of what we're doing in this chapter as a generalization of what we did in the last, in which we focused exclusively on trend. We *still* want to account for trend, if it's present, but we want to expand the model so that we can account for seasonality as well.

The idea of seasonality may be extended to allow for more general **calendar effects.** "Standard" seasonality is just one type of calendar effect. Two additional important calendar effects are holiday variation and trading-day variation.

Holiday variation refers to the fact that some holidays' dates change over time. That is, although they arrive at approximately the same time each year, the exact dates differ. Easter is a common example. Because the behavior of many series, such as sales, shipments, inventories, hours worked, and so on, depends in part on the timing of such holidays, we may want to keep track of them in our forecasting models. As with seasonality, holiday effects may be handled with dummy variables. In a monthly model, for example, in addition to a full set of seasonal dummies, we might include an "Easter dummy," which is 1 if the month contains Easter and 0 otherwise.

---

[3] For simplicity, we have included only a linear trend, but more complicated models of trend, such as quadratic, exponential, or logistic, could of course be used.

Trading-day variation refers to the fact that different months contain different numbers of trading days or business days, which is an important consideration when modeling and forecasting certain series. For example, in a monthly forecasting model of volume traded on the London Stock Exchange, in addition to a full set of seasonal dummies, we might include a trading-day variable, whose value each month is the number of trading days that month.

Allowing for the possibility of holiday or trading-day variation gives the complete model

$$y_t = \beta_1 \mathrm{TIME}_t + \sum_{i=1}^{s} \gamma_i D_{it} + \sum_{i=1}^{v_1} \delta_i^{\mathrm{HD}} \mathrm{HDV}_{it} + \sum_{i=1}^{v_2} \delta_i^{\mathrm{TD}} \mathrm{TDV}_{it} + \varepsilon_t ,$$

where the HDVs are the relevant holiday variables (there are v_1 of them), and the TDVs are the relevant trading-day variables (here we've allowed for v_2 of them, but in most applications, $v_2 = 1$ will be adequate). This is just a standard regression equation and can be estimated by ordinary least squares.

3. Forecasting Seasonal Series

Now consider constructing an h-step-ahead point forecast, $y_{T+h,T}$, at time T. As with the pure trend models discussed in the previous chapter, there's no problem of forecasting the right-hand-side variables, because of the special (perfectly predictable) nature of trend and seasonal variables, so point forecasts are easy to generate.

The full model is

$$y_t = \beta_1 \mathrm{TIME}_t + \sum_{i=1}^{s} \gamma_i D_{it} + \sum_{i=1}^{v_1} \delta_i^{\mathrm{HD}} \mathrm{HDV}_{it} + \sum_{i=1}^{v_2} \delta_i^{\mathrm{TD}} \mathrm{TDV}_{it} + \varepsilon_t ,$$

so that at time $T + h$,

$$y_{T+h} = \beta_1 \mathrm{TIME}_{T+h} + \sum_{i=1}^{s} \gamma_i D_{i,T+h} + \sum_{i=1}^{v_1} \delta_i^{\mathrm{HD}} \mathrm{HDV}_{i,T+h} + \sum_{i=1}^{v_2} \delta_i^{\mathrm{TD}} \mathrm{TDV}_{i,T+h} + \varepsilon_{T+h} .$$

As with the pure trend model of Chapter 5, we project the right side of the equation on what's known at time T (that is, the time-T information set, Ω_T) to obtain the forecast

$$y_{T+h,T} = \beta_1 \mathrm{TIME}_{T+h} + \sum_{i=1}^{s} \gamma_i D_{i,T+h} + \sum_{i=1}^{v_1} \delta_i^{\mathrm{HD}} \mathrm{HDV}_{i,T+h} + \sum_{i=1}^{v_2} \delta_i^{\mathrm{TD}} \mathrm{TDV}_{i,T+h} .$$

As always, we make this point forecast operational by replacing unknown parameters with estimates,

$$\hat{y}_{T+h,T} = \hat{\beta}_1 \mathrm{TIME}_{T+h} + \sum_{i=1}^{s} \hat{\gamma}_i D_{i,T+h} + \sum_{i=1}^{v_1} \hat{\delta}_i^{\mathrm{HD}} \mathrm{HDV}_{i,T+h} + \sum_{i=1}^{v_2} \hat{\delta}_i^{\mathrm{TD}} \mathrm{TDV}_{i,T+h} .$$

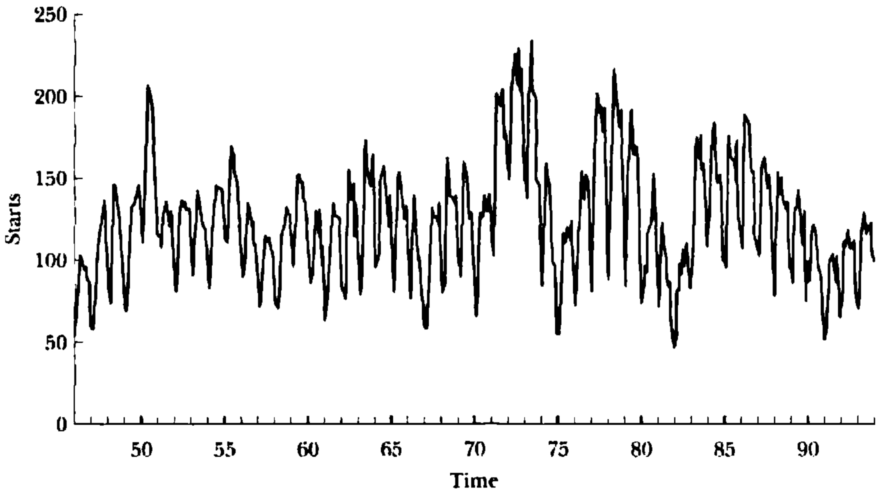
To form an interval forecast, we proceed precisely as in pure trend models we studied earlier. We assume that the regression disturbance is normally distributed, in which case a 95% interval forecast ignoring parameter estimation uncertainty is $y_{T+h,T} \pm 1.96\sigma$, where σ is the standard deviation of the regression disturbance. To make the interval forecast operational, we use $\hat{y}_{T+h,T} \pm 1.96\hat{\sigma}$, where $\hat{\sigma}$ is the standard error of the regression.

To form a density forecast, we again assume that the trend regression disturbance is normally distributed. Then, ignoring parameter estimation uncertainty, the density forecast is $N(y_{T+h,T}, \sigma^2)$, where σ is the standard deviation of the disturbance in the trend regression. The operational density forecast is then $N(\hat{y}_{T+h,T}, \hat{\sigma}^2)$.

4. Application: Forecasting Housing Starts

We'll use the seasonal modeling techniques that we've developed in this chapter to build a forecasting model for housing starts. Housing starts are seasonal because it's usually preferable to start houses in the spring, so that they're completed before winter arrives. We have monthly data on U.S. housing starts; we'll use the 1946.01–1993.12 period for estimation and the 1994.01–1994.11 period for out-of-sample forecasting. We show the entire series in Figure 6.4,

FIGURE 6.4
Housing Starts, 1946.01–1994.11



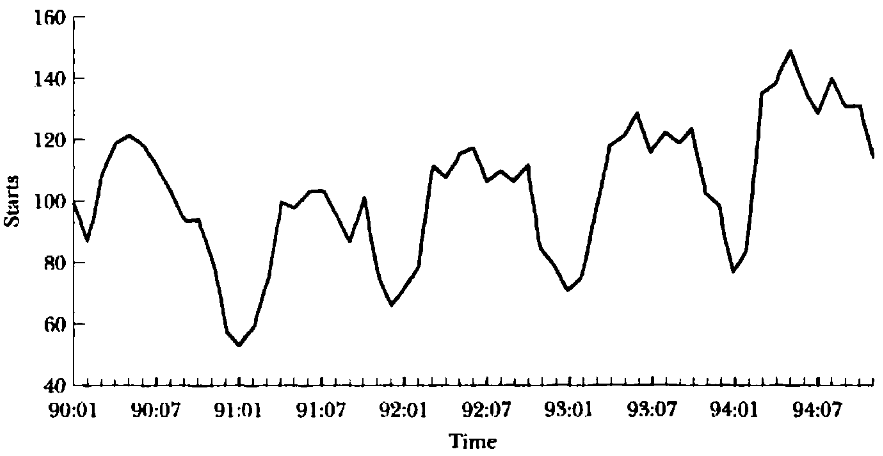


FIGURE 6.5
Housing Starts, 1990.01–1994.11

and we zoom in on the 1990.01–1994.11 period in Figure 6.5 in order to reveal the seasonal pattern in better detail.

The figures reveal that there is no trend, so we'll work with the pure seasonal model,

$$y_t = \sum_{i=1}^{s} \gamma_i D_{it} + \varepsilon_t .$$

Table 6.1 shows the estimation results. The 12 seasonal dummies account for more than a third of the variation in housing starts, as $R^2 = 0.38$. At least some of the remaining variation is cyclical, which the model is not designed to capture. (Note the very low Durbin-Watson statistic.)

The residual plot in Figure 6.6 makes clear the strengths and limitations of the model. First compare the actual and fitted values. The fitted values go through the same seasonal pattern every year—there's nothing in the model other than deterministic seasonal dummies—but that rigid seasonal pattern picks up a lot of the variation in housing starts. It doesn't pick up *all* of the variation, however, as evidenced by the serial correlation that's apparent in the residuals. Note the dips in the residuals, for example, in recessions (for example 1990, 1982, 1980, and 1975), and the peaks in booms.

The estimated seasonal factors are just the 12 estimated coefficients on the seasonal dummies; we graph them in Figure 6.7. The seasonal effects are very low in January and February and then rise quickly and peak in May, after which they decline, at first slowly and then abruptly in November and December.

In Figure 6.8, we see the history of housing starts through 1993, together with the out-of-sample point and 95% interval extrapolation forecasts for the first 11 months of 1994. The forecasts look reasonable, as the model has

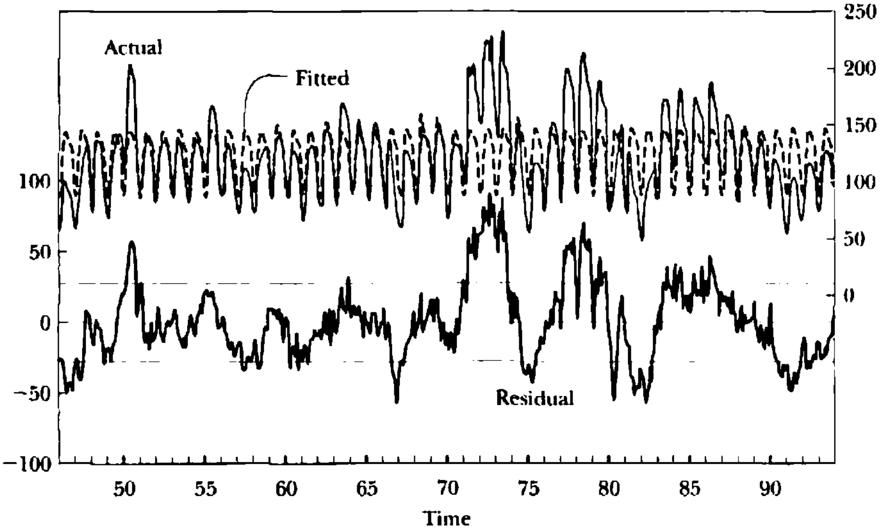
TABLE 6.1
Regression Results: Seasonal Dummy Variable Model, Housing Starts

LS//Dependent variable is STARTS.
Sample: 1946:01 1993:12
Included observations: 576

Variable	Coefficient	Std. Error	*t*-Statistic	Prob.
*D*1	86.50417	4.029055	21.47009	0.0000
*D*2	89.50417	4.029055	22.21468	0.0000
*D*3	122.8833	4.029055	30.49929	0.0000
*D*4	142.1687	4.029055	35.28588	0.0000
*D*5	147.5000	4.029055	36.60908	0.0000
*D*6	145.9979	4.029055	36.23627	0.0000
*D*7	139.1125	4.029055	34.52733	0.0000
*D*8	138.4167	4.029055	34.35462	0.0000
*D*9	130.5625	4.029055	32.40524	0.0000
*D*10	134.0917	4.029055	33.28117	0.0000
*D*11	111.8333	4.029055	27.75671	0.0000
*D*12	92.15833	4.029055	22.87344	0.0000

R^2	0.383780	Mean dependent var.	123.3944
Adjusted R^2	0.371762	SD dependent var.	35.21775
SE of regression	27.91411	Akaike info criterion	6.678878
Sum squared resid.	439467.5	Schwarz criterion	6.769630
Log likelihood	−2728.825	*F*-statistic	31.93250
Durbin-Watson stat.	0.154140	Prob(*F*-statistic)	0.000000

FIGURE 6.6
Residual Plot



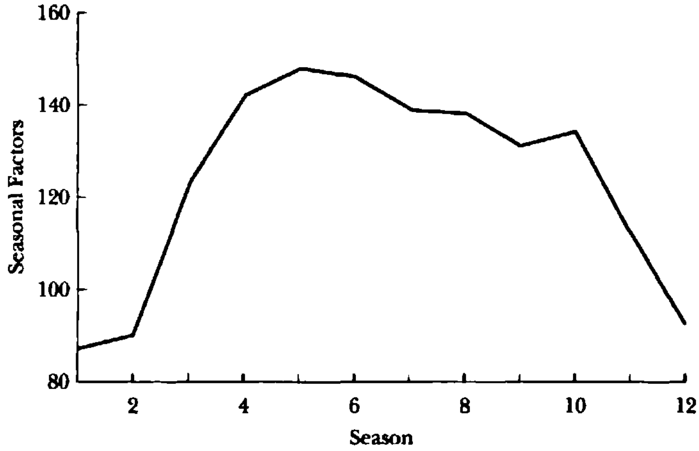


FIGURE 6.7
Estimated Seasonal Factors, Housing Starts

evidently done a good job of capturing the seasonal pattern. The forecast intervals are quite wide, however, reflecting the fact that the seasonal effects captured by the forecasting model are responsible for only about a third of the variation in the variable being forecast.

In Figure 6.9, we include the 1994 realization. The forecast appears highly accurate, as the realization and forecast are quite close throughout. Moreover, the realization is everywhere well within the 95% interval.

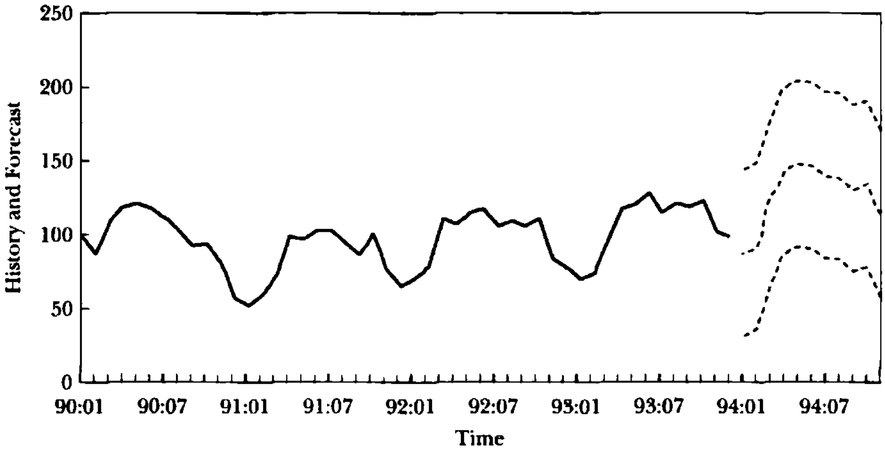


FIGURE 6.8
Housing Starts: History, 1990.01–1993.12; and Forecast, 1994.01–1994.11

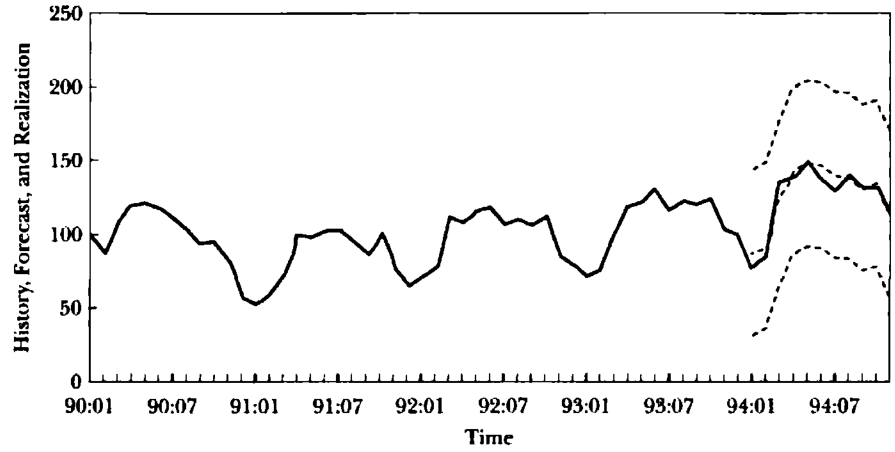


FIGURE 6.9 *Housing Starts: History, 1990.01–1993.12; and Forecast and Realization, 1994.01–1994.11*

Exercises, Problems, and Complements

1. (Log transformations in seasonal models) Just as log transformations were useful in trend models to allow for nonlinearity, so too are they useful in seasonal models, although for a somewhat different purpose: **stabilization of variance.** Often log transformations stabilize seasonal patterns whose variance is growing over time. Explain and illustrate.

2. (Seasonal adjustment) Just as we sometimes want to remove the trend from a series, sometimes we want to seasonally adjust a series before modeling and forecasting it. Seasonal adjustment may be done with moving-average methods analogous to those used for detrending in Chapter 5, or with the dummy variable methods discussed in this chapter, or with sophisticated hybrid methods like the *X*-11 procedure developed at the U.S. Census Bureau.
   a. Discuss in detail how you'd use dummy variable regression methods to seasonally adjust a series. (*Hint:* The seasonally adjusted series is closely related to the residual from the seasonal dummy variable regression.)
   b. Seasonally adjust the housing starts series using dummy variable regression. Discuss the patterns present and absent from the seasonally adjusted series.
   c. Search the web (or the library) for information on the latest U.S. Census Bureau seasonal adjustment procedure, and report what you learned.

3. (Selecting forecasting models involving calendar effects) You're sure that a series you want to forecast is trending and that a linear trend is adequate, but you're not sure whether seasonality is important. To be safe, you fit a forecasting model with both trend and seasonal dummies,

$$y_t = \beta_1 \mathrm{TIME}_t + \sum_{i=1}^{s} \gamma_i D_{it} + \varepsilon_t .$$

a. The hypothesis of no seasonality, in which case you could drop the seasonal dummies, corresponds to equal seasonal coefficients across seasons, which is a set of $s - 1$ linear restrictions:

$$\gamma_1 = \gamma_2, \gamma_2 = \gamma_3, \ldots, \gamma_{s-1} = \gamma_s .$$

How would you perform an F-test of the hypothesis? What assumptions are you implicitly making about the regression's disturbance term?

b. Alternatively, how would you use forecast model selection criteria to decide whether to include the seasonal dummies?

c. What would you do in the event that the results of the "hypothesis testing" and "model selection" approaches disagree?

d. How, if at all, would your answers change if instead of considering whether to include seasonal dummies you were considering whether to include holiday dummies? Trading-day dummies?

4. (Testing for seasonality) Using the housing starts data:
   a. As in the chapter, construct and estimate a model with a full set of seasonal dummies.
   b. Test the hypothesis of no seasonal variation. Discuss your results.
   c. Test for the equality of the coefficients on March and November and the coefficients on all the months in between, and construct a model that uses three dummy variables, one for December, January, and February, one for March and November, and one for the remaining months.

5. (Seasonal regressions with an intercept and $s - 1$ seasonal dummies) Reestimate the housing starts model using an intercept and 11 seasonal dummies, rather than the full set of seasonal dummies as in the text. Compare and contrast your results with those reported in the text. What is the interpretation of the intercept? What are the interpretations of the coefficients on the 11 included seasonal dummies? Does it matter which month's dummy you drop?

6. (Applied trend and seasonal modeling) Nile.com, a successful online bookseller, monitors and forecasts the number of hits per day to its web page. You have daily hits data for January 1, 1998, through September 28, 1998.
   a. Fit and assess the standard linear, quadratic, and log linear trend models.
   b. For a few contiguous days roughly in late April and early May, hits were much higher than usual during a big sale. Do you find evidence of a corresponding group of outliers in the residuals from your trend models? Do they influence your trend estimates much? How should you treat them?
   c. Model and assess the significance of day-of-week effects in Nile.com web page hits.
   d. Select a final model, consisting only of trend and seasonal components, to use for forecasting.
   e. Use your model to forecast Nile.com hits through the end of 1998.

7. (Periodic models) We introduced the seasonal dummy model as a natural and simple method for generalizing a simple "mean plus noise" model,

$$y_t = \mu + \varepsilon_t ,$$

to allow the mean to vary with the seasons,

$$y_t = \sum_{i=1}^{s} \gamma_i D_{it} + \varepsilon_t .$$

More generally, we can also allow the coefficients of richer models to vary with the seasons, as, for example, when we move from the fixed-coefficient regression model,

$$y_t = \beta_0 + \beta_1 x_t + \varepsilon_t ,$$

to the model with **time-varying parameters**

$$y_t = \left(\sum_{i=1}^{s} \gamma_i^{0} D_{it}\right) + \left(\sum_{i=1}^{s} \gamma_i^{1} D_{it}\right) x_t + \varepsilon_t .$$

This model, which permits not only a seasonally varying intercept but also a seasonally varying slope, is an example of a **periodic regression model.** The word *periodic* refers to the coefficients, which vary regularly with a fixed seasonal periodicity.

8. (Interpreting dummy variables) You fit a purely seasonal model with a full set of standard monthly dummy variables to a monthly series of employee hours worked. Discuss how the estimated dummy variable coefficients $\hat{\gamma}_1, \hat{\gamma}_2, \ldots$ would change if you changed the first dummy variable $D_1 = (1, 0, 0, \ldots)$ (with all the other dummy variables remaining the same) to
   a. $D_1 = (2, 0, 0, \ldots)$;
   b. $D_1 = (-10, 0, 0, \ldots)$;
   c. $D_1 = (1, 1, 0, \ldots)$.

9. (Constructing seasonal models) Describe how you would construct a purely seasonal model for the following monthly series. In particular, what dummy variable(s) would you use to capture the relevant effects?
   a. A sporting goods store finds that detrended monthly sales are roughly the same for each month in a given 3-month season. For example, sales are similar in the winter months of January, February, and March; in the spring months of April, May, and June; and so on.
   b. A campus bookstore finds that detrended sales are roughly the same for all first, all second, all third, and all fourth months of each trimester. For example, sales are similar in January, May, and September, the first months of the first, second, and third trimesters, respectively.
   c. A Christmas ornament store is only open in November and December, so sales are zero in all other months.

10. (Calendar effects) You run a large catering firm, specializing in Sunday brunches and weddings. You model the firm's monthly income as $y_t = \beta_0 + \delta_s S_t + \delta_w W_t + \varepsilon_t$, where y is monthly income, and S and W are calendar effect variables indicating the number of Sundays and weddings in a month.
    a. What are the units of β_0, δ_s, and δ_w?
    b. How could you estimate the average income the firm receives per wedding?
    c. Over the past 30 years, you have regularly increased your prices to keep pace with inflation. How would you modify the model to account for the effects of such increases?

Bibliographical and Computational Notes

Nerlove, Grether, and Carvalho (1979), Hylleberg (1986), and Ghysels and Osborne (2001) discuss seasonality as relevant for forecasting (and much else). Franses and Paap (2004) provide a detailed overview of periodic time series models.

Concepts for Review

Seasonality
Deterministic seasonality
Stochastic seasonality
Seasonally adjusted time series
Nonseasonal fluctuations
Regression on seasonal dummies
Seasonal adjustment
Seasonal dummy variables
Calendar effects
Holiday variation
Trading-day variation
Stabilization of variance
Time-varying parameters
Periodic regression model

References and Additional Readings

Franses, P. H., and Paap, R. (2004). *Periodic Time Series Models.* Oxford: Oxford University Press.

Ghysels, E., and Osborne, D. R. (2001). *The Econometric Analysis of Seasonal Time Series.* Cambridge: Cambridge University Press.

Hylleberg, S. (1986). *Seasonality in Regression.* Orlando, Fl.: Academic Press.

Nerlove, M., Grether, D. M., and Carvalho, J. L. (1979). *Analysis of Economic Time Series: A Synthesis.* (2nd ed., 1996). New York: Academic Press.

Characterizing Cycles

We've already built forecasting models with trend and seasonal components. In this chapter, as well as the next two, we consider a crucial third component, **cycles.** When you think of a "cycle," you probably think of the sort of rigid up-and-down pattern depicted in Figure 7.1. Such cycles can sometimes arise, but cyclical fluctuations in business, finance, economics, and government are typically much less rigid. In fact, when we speak of cycles, we have in mind a much more general, all-encompassing notion of cyclicality: any sort of dynamics not captured by trends or seasonals.

Cycles, according to our broad interpretation, may display the sort of back-and-forth movement characterized in Figure 7.1, but they don't have to. All we require is that there be some dynamics, some persistence, some way in which the present is linked to the past and the future to the present. Cycles are present in most of the series that concern us, and it's crucial that we know how to model and forecast them, because their history conveys information regarding their future.

Trend and seasonal dynamics are simple, so we can capture them with simple models. Cyclical dynamics, however, are more complicated. Because of the wide variety of cyclical patterns, the sorts of models we need are substantially more involved. Thus, we split the discussion into three parts. Here in Chapter 7 we develop methods for *characterizing* cycles, in Chapter 8 we discuss *models* of cycles, and following that, in Chapter 9, we show how to use those models

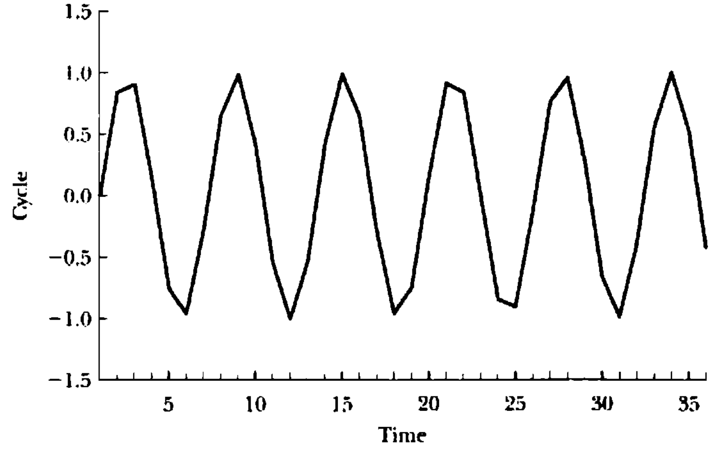


FIGURE 7.1
A Rigid Cyclical Pattern

to *forecast* cycles. All of the material is crucial to a real understanding of forecasting and forecasting models, and it's also a bit difficult the first time around because it's unavoidably rather mathematical, so careful, systematic study is required. The payoff will be large when we arrive at Chapter 10, in which we assemble and apply extensively the ideas for modeling and forecasting trends, seasonals, and cycles developed in Chapters 5–9.

1. Covariance Stationary Time Series

A **realization** of a time series is an ordered set, $\{\ldots, y_{-2}, y_{-1}, y_0, y_1, y_2, \ldots\}$. Typically the observations are ordered in time—hence the name **time series**—but they don't have to be. We could, for example, examine a spatial series, such as office space rental rates as we move along a line from a point in Midtown Manhattan to a point in the New York suburbs 30 miles away. But the most important case for forecasting, by far, involves observations ordered in time, so that's what we'll stress.

In theory, a time series realization begins in the infinite past and continues into the infinite future. This perspective may seem abstract and of limited practical applicability, but it will be useful in deriving certain very important properties of the forecasting models we'll be using soon. In practice, of course, the data we observe are just a finite subset of a realization, $\{y_1, \ldots, y_T\}$, called a **sample path.**

Shortly we'll be building forecasting models for cyclical time series. If the underlying probabilistic structure of the series were changing over time, we'd be doomed—there would be no way to predict the future accurately on the basis of the past, because the laws governing the future would differ from those governing the past. If we want to forecast a series, at a minimum we'd like its mean and its covariance structure (i.e., the covariances between current and past values) to be stable over time, in which case we say that the series is **covariance stationary.**

Let's discuss covariance stationarity in greater depth. The first requirement for a series to be covariance stationary is that the mean of the series be stable over time. The mean of the series at time t is

$$E(y_t) = \mu_t .$$

If the mean is stable over time, as required by covariance stationarity, then we can write

$$E(y_t) = \mu ,$$

for all t. Because the mean is constant over time, there's no need to put a time subscript on it.

The second requirement for a series to be covariance stationary is that its covariance structure be stable over time. Quantifying stability of the covariance structure is a bit tricky, but tremendously important, and we do it using the **autocovariance function.** The autocovariance at displacement τ is just the covariance between y_t and $y_{t-\tau}$. It will of course depend on τ, and it may also depend on t, so in general we write

$$\gamma(t, \tau) = \text{cov}(y_t, y_{t-\tau}) = E(y_t - \mu)(y_{t-\tau} - \mu) .$$

If the covariance structure is stable over time, as required by covariance stationarity, then the autocovariances depend only on displacement, τ, not on time, t, and we write

$$\gamma(t, \tau) = \gamma(\tau) ,$$

for all t.

The autocovariance function is important because it provides a basic summary of cyclical dynamics in a covariance stationary series. By examining the autocovariance structure of a series, we learn about its dynamic behavior. We graph and examine the autocovariances as a function of τ. Note that the autocovariance function is symmetric; that is,

$$\gamma(\tau) = \gamma(-\tau) ,$$

for all τ. Typically, we'll consider only nonnegative values of τ. Symmetry reflects the fact that the autocovariance of a covariance stationary series depends only on displacement; it doesn't matter whether we go forward or backward. Note also that

$$\gamma(0) = \text{cov}(y_t, y_t) = \text{var}(y_t) .$$

There is one more technical requirement of covariance stationarity: We require that the variance of the series—the autocovariance at displacement 0, $\gamma(0)$—be finite. It can be shown that no autocovariance can be larger in absolute value than $\gamma(0)$, so if $\gamma(0) < \infty$, then so, too, are all the other autocovariances.

It may seem that the requirements for covariance stationarity are quite stringent, which would bode poorly for our forecasting models, almost all of which invoke covariance stationarity in one way or another. It is certainly true that many economic, business, financial, and government series are not covariance stationary. An upward trend, for example, corresponds to a steadily increasing mean, and seasonality corresponds to means that vary with the season, both of which are violations of covariance stationarity.

But appearances can be deceptive. Although many series are not covariance stationary, it is frequently possible to work with models that give special treatment to nonstationary components such as trend and seasonality, so that the cyclical component that's left over is likely to be covariance stationary. We'll often adopt that strategy. Alternatively, simple transformations often appear to transform nonstationary series to covariance stationarity. For example, many series that are clearly nonstationary in levels appear covariance stationary in growth rates.

In addition, note that although covariance stationarity requires means and covariances to be stable and finite, it places no restrictions on other aspects of the distribution of the series, such as skewness and kurtosis.[1] The upshot is simple: Whether we work directly in levels and include special components for the nonstationary elements of our models, or we work on transformed data such as growth rates, the covariance stationarity assumption is not as unrealistic as it may seem.

Recall that the correlation between two random variables x and y is defined by

$$\mathrm{corr}(x, y) = \frac{\mathrm{cov}(x, y)}{\sigma_x \sigma_y} .$$

That is, the correlation is simply the covariance, "normalized" or "standardized," by the product of the standard deviations of x and y. Both the correlation and the covariance are measures of linear association between two random variables. The correlation is often more informative and easily interpreted, however, because the construction of the correlation coefficient guarantees that $\mathrm{corr}(x, y) \in [-1, 1]$, whereas the covariance between the same two random variables may take any value. The correlation, moreover, does not depend on the units in which x and y are measured, whereas the covariance does. Thus, for example, if x and y have a covariance of 10 million, they're not necessarily very strongly associated, whereas if they have a correlation of .95, it is unambiguously clear that they are very strongly associated.

In light of the superior interpretability of correlations as compared with covariances, we often work with the correlation, rather than the covariance, between y_t and $y_{t-\tau}$. That is, we work with the **autocorrelation function, $\rho(\tau)$**, rather than the autocovariance function, $\gamma(\tau)$. The autocorrelation function is obtained by dividing the autocovariance function by the variance,

$$\rho(\tau) = \frac{\gamma(\tau)}{\gamma(0)}, \quad \tau = 0, 1, 2, \ldots .$$

---

[1] For that reason, covariance stationarity is sometimes called **second-order stationarity** or **weak stationarity.**

The formula for the autocorrelation is just the usual correlation formula, specialized to the correlation between y_t and $y_{t-\tau}$. To see why, note that the variance of y_t is $\gamma(0)$, and by covariance stationarity, the variance of y at any other time $y_{t-\tau}$ is also $\gamma(0)$. Thus,

$$\rho(\tau) = \frac{\operatorname{cov}(y_t, y_{t-\tau})}{\sqrt{\operatorname{var}(y_t)}\sqrt{\operatorname{var}(y_{t-\tau})}} = \frac{\gamma(\tau)}{\sqrt{\gamma(0)}\sqrt{\gamma(0)}} = \frac{\gamma(\tau)}{\gamma(0)},$$

as claimed. Note that we always have $\rho(0) = \frac{\gamma(0)}{\gamma(0)} = 1$, because any series is perfectly correlated with itself. Thus, the autocorrelation at displacement 0 isn't of interest; rather, only the autocorrelations *beyond* displacement 0 inform us about a series' dynamic structure.

Finally, the **partial autocorrelation function,** $p(\tau)$, is sometimes useful. $p(\tau)$ is just the coefficient of $y_{t-\tau}$ in a population linear regression of y_t on $y_{t-1}, \ldots, y_{t-\tau}$.[2] We call such a regression an **autoregression,** because the variable is regressed on lagged values of itself. It's easy to see that the autocorrelations and partial autocorrelations, although related, differ in an important way. The autocorrelations are just the "simple" or "regular" correlations between y_t and $y_{t-\tau}$. The partial autocorrelations, on the other hand, measure the association between y_t and $y_{t-\tau}$ after *controlling* for the effects of $y_{t-1}, \ldots, y_{t-\tau+1}$; that is, they measure the partial correlation between y_t and $y_{t-\tau}$.

As with the autocorrelations, we often graph the partial autocorrelations as a function of τ and examine their qualitative shape, which we'll do soon. Like the autocorrelation function, the partial autocorrelation function provides a summary of a series' dynamics, but as we'll see, it does so in a different way.[3]

All of the covariance stationary processes that we will study subsequently have autocorrelation and partial autocorrelation functions that approach 0, one way or another, as the displacement gets large. In Figure 7.2 we show an autocorrelation function that displays gradual one-sided damping, and in Figure 7.3 we show a constant autocorrelation function; the latter could not be the autocorrelation function of a stationary process, whose autocorrelation function must eventually decay. The precise decay patterns of autocorrelations and partial autocorrelations of a covariance stationary series, however, depend on the specifics of the series, as we'll see in detail in the next chapter. In Figure 7.4, for example, we show an autocorrelation function that displays damped oscillation—the autocorrelations are positive at first, then become negative for a while, then positive again, and so on, while continuously getting

---

[2] To get a feel for what we mean by "population regression," imagine that we have an infinite sample of data at our disposal, so that the parameter estimates in the regression are not contaminated by sampling variation; that is, they're the true population values. The thought experiment just described is a **population regression.**

[3] Also in parallel to the autocorrelation function, the partial autocorrelation at displacement 0 is always 1 and is therefore uninformative and uninteresting. Thus, when we graph the autocorrelation and partial autocorrelation functions, we'll begin at displacement 1 rather than displacement 0.

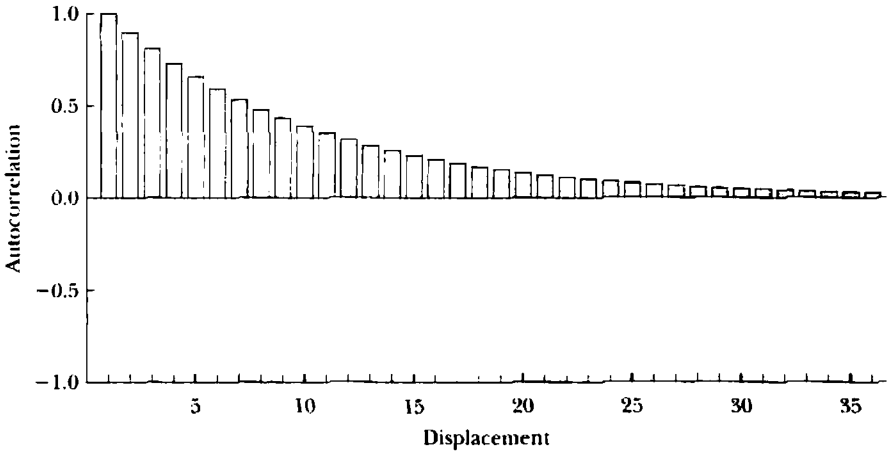


FIGURE 7.2
Autocorrelation Function, One-Sided Gradual Damping

smaller in absolute value. Finally, in Figure 7.5 we show an autocorrelation function that differs in the way it approaches 0—the autocorrelations drop abruptly to 0 beyond a certain displacement.

2. White Noise

In this section and throughout the next chapter, we'll study the population properties of certain time series models, or **time series processes,** which are very important for forecasting. Before we estimate time series forecasting models, we need to understand their population properties, assuming that the

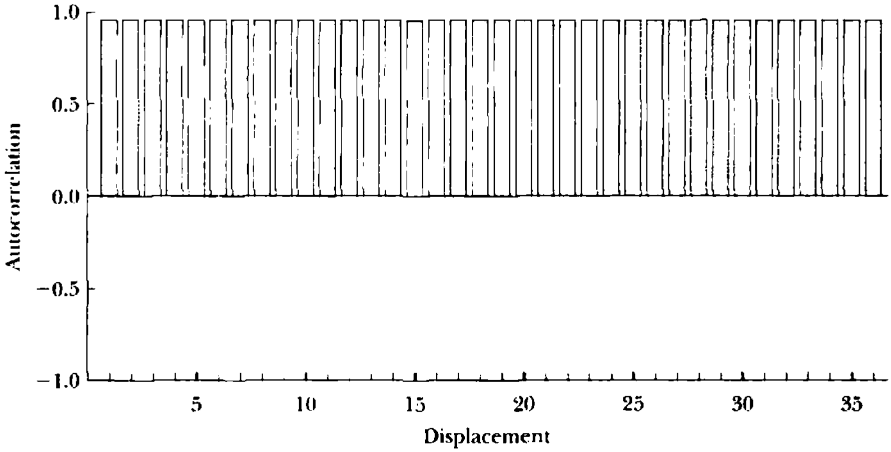
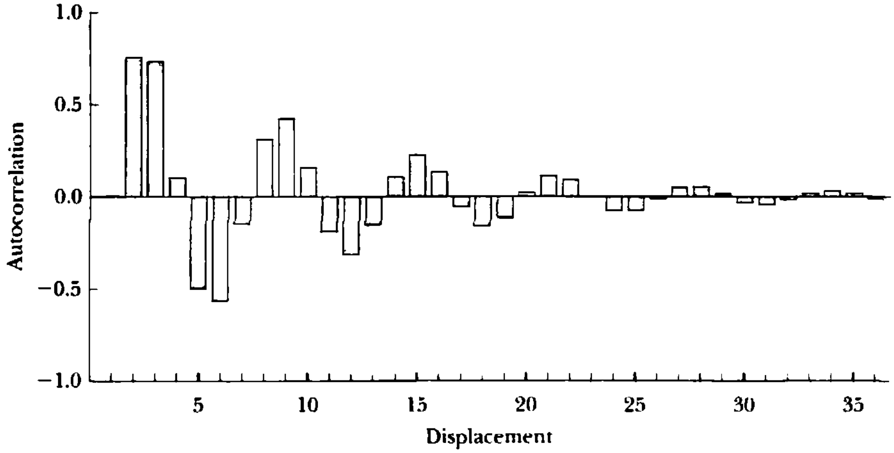


FIGURE 7.3
Autocorrelation Function, Nondamping

FIGURE 7.4
Autocorrelation Function, Gradual Damped Oscillation



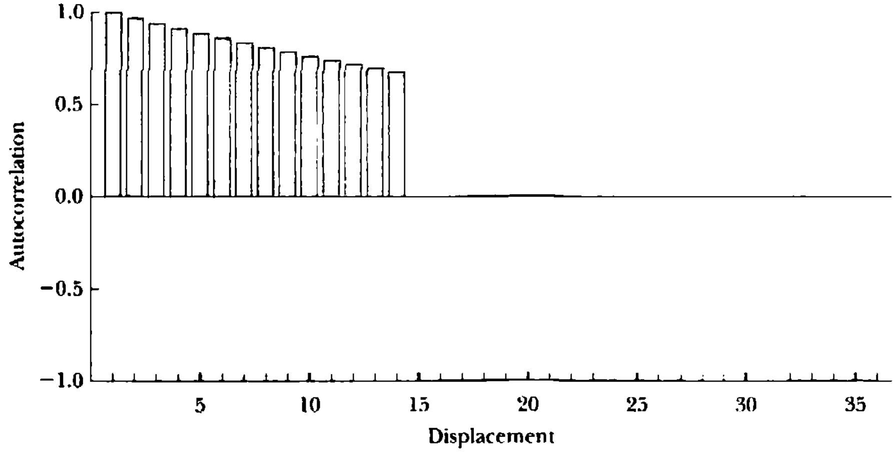
postulated model is true. The simplest of all such time series processes is the fundamental building block from which all others are constructed. In fact, it's so important that we introduce it now. We use y to denote the observed series of interest. Suppose that

$$y_t = \varepsilon_t$$

$$\varepsilon_t \sim (0, \sigma^2),$$

where the "shock," ε_t, is uncorrelated over time. We say that ε_t, and hence y_t, is **serially uncorrelated.** Throughout, unless explicitly stated otherwise, we assume that $\sigma^2 < \infty$. Such a process, with zero mean, constant variance, and

FIGURE 7.5
Autocorrelation Function, Sharp Cutoff



no serial correlation. is called **zero-mean white noise,** or simply **white noise.**[4] Sometimes for short we write

$$\varepsilon_t \sim WN(0, \sigma^2)$$

and hence

$$y_t \sim WN(0, \sigma^2) \, .$$

Note that, although ε_t and hence y_t are serially uncorrelated, they are not necessarily serially independent, because they are not necessarily normally distributed.[5] If in addition to being serially uncorrelated, y is serially independent, then we say that y is **independent white noise.**[6] We write

$$y_t \overset{\text{iid}}{\sim} (0, \sigma^2) \, ,$$

and we say that "y is independently and identically distributed with zero mean and constant variance." If y is serially uncorrelated and normally distributed, then it follows that y is also serially independent, and we say that y is **normal white noise** or **Gaussian white noise.**[7] We write

$$y_t \overset{\text{iid}}{\sim} N(0, \sigma^2) \, .$$

We read "y is independently and identically distributed as normal, with zero mean and constant variance" or simply "y is Gaussian white noise." In Figure 7.6 we show a sample path of Gaussian white noise, of length $T = 150$, simulated on a computer. There are no patterns of any kind in the series due to the independence over time.

You're already familiar with white noise, although you may not realize it. Recall that the disturbance in a regression model is typically assumed to be white noise of one sort or another. There's a subtle difference here, however. Regression disturbances are not observable, whereas we're working with an observed series. Later, however, we'll see how all of our models for observed series can be used to model unobserved variables such as regression disturbances.

Let's characterize the dynamic stochastic structure of white noise, $y_t \sim WN(0, \sigma^2)$. By construction the unconditional mean of y is

$$E(y_t) = 0 \, ,$$

and the unconditional variance of y is

$$\text{var}(y_t) = \sigma^2 \, .$$

---

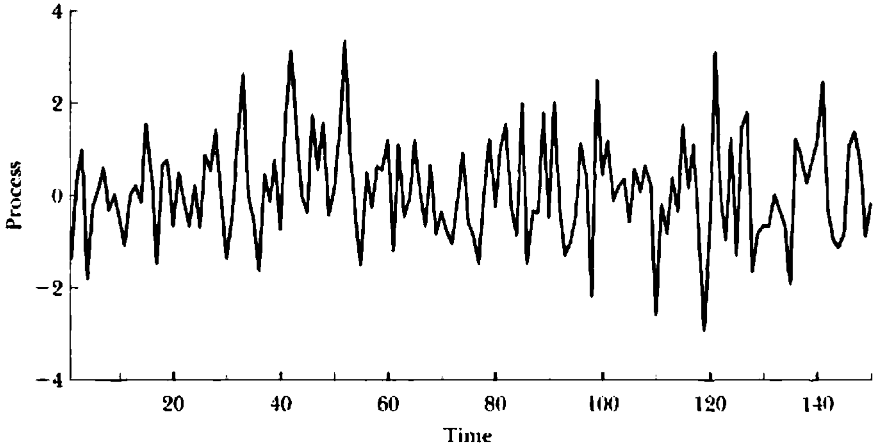
[4] It's called white noise by analogy with white light, which is composed of all colors of the spectrum, in equal amounts. We can think of white noise as being composed of a wide variety of cycles of differing periodicities, in equal amounts.

[5] Recall that zero correlation implies independence only in the normal case.

[6] Another name for independent white noise is **strong white noise,** in contrast to standard serially uncorrelated **weak white noise.**

[7] Karl Friedrich Gauss, one of the greatest mathematicians of all time, discovered the normal distribution some 200 years ago—hence the adjective *Gaussian.*

FIGURE 7.6
Realization of White Noise Process



Note that the unconditional mean and variance are constant. In fact, the unconditional mean and variance must be constant for any covariance stationary process. The reason is that constancy of the unconditional mean was our first explicit requirement of covariance stationarity and that constancy of the unconditional variance follows implicitly from the second requirement of covariance stationarity—that the autocovariances depend only on displacement, not on time.[8]

To understand fully the linear dynamic structure of a covariance stationary time series process, we need to compute and examine its mean and its autocovariance function. For white noise, we've already computed the mean and the variance, which is the autocovariance at displacement 0. We have yet to compute the rest of the autocovariance function; fortunately, however, it's very simple. Because white noise is, by definition, uncorrelated over time, all the autocovariances, and hence all the autocorrelations, are 0 beyond displacement 0.[9] Formally, then, the autocovariance function for a white noise process is

$$\gamma(\tau) = \begin{cases} \sigma^2, & \tau = 0 \\ 0, & \tau \geq 1 \end{cases},$$

and the autocorrelation function for a white noise process is

$$\rho(\tau) = \begin{cases} 1, & \tau = 0 \\ 0, & \tau \geq 1 \end{cases}.$$

In Figure 7.7 we plot the white noise autocorrelation function.

Finally, consider the partial autocorrelation function for a white noise series. For the same reason that the autocorrelation at displacement 0 is always 1, so, too, is the partial autocorrelation at displacement 0. For a white noise

---

[8] Recall that $\sigma^2 = \gamma(0)$.

[9] If the autocovariances are all 0, so are the autocorrelations, because the autocorrelations are proportional to the autocovariances.

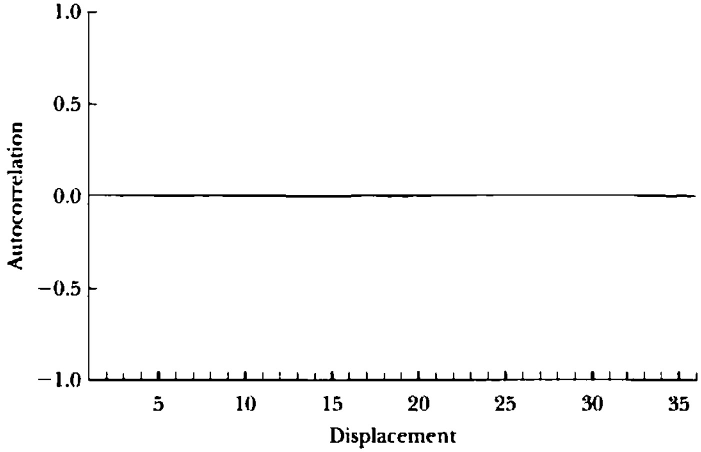


FIGURE 7.7 *Population Autocorrelation Function, White Noise Process*

process, all partial autocorrelations beyond displacement 0 are 0, which again follows from the fact that white noise, by construction, is serially uncorrelated. Population regressions of y_t on y_{t-1}, or on y_{t-1} and y_{t-2} or on any other lags, produce nothing but 0 coefficients, because the process is serially uncorrelated. Formally, the partial autocorrelation function of a white noise process is

$$p(\tau) = \begin{cases} 1, & \tau = 0 \\ 0, & \tau \geq 1 \end{cases}.$$

We show the partial autocorrelation function of a white noise process in Figure 7.8. Again, it's degenerate and exactly the same as the autocorrelation function!

By now you've surely noticed that if you were assigned the task of forecasting independent white noise, you'd likely be doomed to failure. What happens

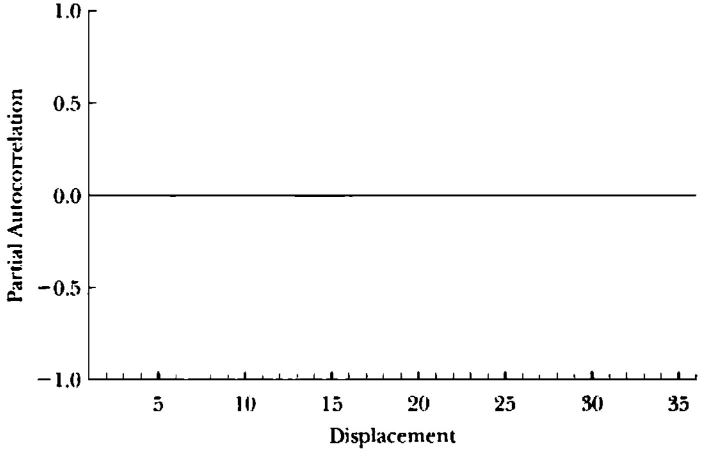


FIGURE 7.8 *Population Partial Autocorrelation Function, White Noise Process*

to a white noise series at any time is uncorrelated with anything in the past; similarly, what happens in the future is uncorrelated with anything in the present or past. But understanding white noise is tremendously important for at least two reasons. First, as already mentioned, processes with much richer dynamics are built up by taking simple transformations of white noise. Second, 1-step-ahead forecast errors from good models should be white noise. After all, if such forecast errors aren't white noise, then they're serially correlated, which means that they're forecastable; and if forecast errors are forecastable, then the forecast can't be very good. Thus, it's important that we understand and be able to recognize white noise.

Thus far we've characterized white noise in terms of its mean, variance, autocorrelation function, and partial autocorrelation function. Another characterization of dynamics, with important implications for forecasting, involves the mean and variance of a process, *conditional* on its past. In particular, we often gain insight into the dynamics in a process by examining its conditional mean, which is a key object for forecasting.[10] In fact, throughout our study of time series, we'll be interested in computing and contrasting the **unconditional mean and variance** and the **conditional mean and variance** of various processes of interest. Means and variances, which convey information about location and scale of random variables, are examples of what statisticians call **moments.** For the most part, our comparisons of the conditional and unconditional moment structure of time series processes will focus on means and variances (they're the most important moments), but sometimes we'll be interested in higher-order moments, which are related to properties such as skewness and kurtosis.

For comparing conditional and unconditional means and variances, it will simplify our story to consider independent white noise, $y_t \overset{iid}{\sim} (0, \sigma^2)$. By the same arguments as before, the unconditional mean of y is 0, and the unconditional variance is σ^2. Now consider the conditional mean and variance, where the information set Ω_{t-1} on which we condition contains either the past history of the observed series, $\Omega_{t-1} = \{y_{t-1}, y_{t-2}, \ldots\}$, or the past history of the shocks, $\Omega_{t-1} = \{\varepsilon_{t-1}, \varepsilon_{t-2}, \ldots\}$. (They're the same in the white noise case.) In contrast to the unconditional mean and variance, which must be constant by covariance stationarity, the conditional mean and variance need not be constant, and in general we'd expect them *not* to be constant. The unconditionally expected growth of laptop computer sales next quarter may be 10%, but expected sales growth may be much higher, *conditional* on knowledge that sales grew this quarter by 20%. For the independent white noise process, the conditional mean is

$$E(y_t \mid \Omega_{t-1}) = 0 ,$$

and the conditional variance is

$$\mathrm{var}(y_t \mid \Omega_{t-1}) = E((y_t - E(y_t \mid \Omega_{t-1}))^2 \mid \Omega_{t-1}) = \sigma^2 .$$

---

[10] If you need to refresh your memory on conditional means, consult any good introductory statistics book, such as Wonnacott and Wonnacott (1990).

Conditional and unconditional means and variances are identical for an independent white noise series; there are no dynamics in the process and hence no dynamics in the conditional moments to exploit for forecasting.

3. The Lag Operator

The **lag operator** and related constructs are the natural language in which forecasting models are expressed. If you want to understand and manipulate forecasting models—indeed, even if you simply want to be able to read the software manuals—you have to be comfortable with the lag operator. The lag operator, L, is very simple: It "operates" on a series by lagging it. Hence,

$$Ly_t = y_{t-1} .$$

Similarly,

$$L^2 y_t = L(L(y_t)) = L(y_{t-1}) = y_{t-2} ,$$

and so on. Typically we'll operate on a series not with the lag operator but with a **polynomial in the lag operator.** A lag operator polynomial of degree m is just a linear function of powers of L, up through the mth power,

$$B(L) = b_0 + b_1 L + b_2 L^2 + \cdots b_m L^m .$$

To take a very simple example of a lag operator polynomial operating on a series, consider the mth-order lag operator polynomial L^m, for which

$$L^m y_t = y_{t-m} .$$

A well-known operator, the first-difference operator Δ, is actually a first-order polynomial in the lag operator; you can readily verify that

$$\Delta y_t = (1 - L)y_t = y_t - y_{t-1} .$$

As a final example, consider the second-order lag operator polynomial $(1 + 0.9L + 0.6L^2)$ operating on y_t. We have

$$(1 + 0.9L + 0.6L^2)y_t = y_t + 0.9y_{t-1} + 0.6y_{t-2} ,$$

which is a weighted sum, or **distributed lag,** of current and past values. All forecasting models, one way or another, must contain such distributed lags, because they've got to quantify how the past evolves into the present and future; hence, lag operator notation is a useful shorthand for stating and manipulating forecasting models.

Thus far, we've considered only finite-order polynomials in the lag operator; it turns out that infinite-order polynomials are also of great interest. We write the infinite-order lag operator polynomial as

$$B(L) = b_0 + b_1 L + b_2 L^2 + \cdots = \sum_{i=0}^{\infty} b_i L^i .$$

Thus, for example, to denote an infinite distributed lag of current and past shocks, we might write

$$B(L)\varepsilon_t = b_0\varepsilon_t + b_1\varepsilon_{t-1} + b_2\varepsilon_{t-2} + \cdots = \sum_{i=0}^{\infty} b_i\varepsilon_{t-i} .$$

At first sight, infinite distributed lags may seem esoteric and of limited practical interest, because models with infinite distributed lags have infinitely many parameters ($b_0, b_1, b_2, \ldots$) and therefore can't be estimated with a finite sample of data. On the contrary, and surprisingly, it turns out that models involving infinite distributed lags are central to time series modeling and forecasting. Wold's theorem, to which we now turn, establishes that centrality.

4. Wold's Theorem, the General Linear Process, and Rational Distributed Lags[11]

WOLD'S THEOREM

Many different dynamic patterns are consistent with covariance stationarity. Thus, if we know only that a series is covariance stationary, it's not at all clear what sort of model we might fit to describe its evolution. The trend and seasonal models that we've studied aren't of use; they're models of specific nonstationary components. Effectively, what we need now is an appropriate model for what's left after fitting the trend and seasonal components—a model for a covariance stationary residual. **Wold's representation theorem** points to the appropriate model.

THEOREM

Let $\{y_t\}$ be any zero-mean covariance-stationary process.[12] Then we can write it as

$$y_t = B(L)\varepsilon_t = \sum_{i=0}^{\infty} b_i\varepsilon_{t-i}$$

$$\varepsilon_t \sim WN(0, \sigma^2) ,$$

where $b_0 = 1$ and $\sum_{i=0}^{\infty} b_i^2 < \infty$. In short, the correct "model" for any covariance stationary series is some infinite distributed lag of white noise, called the **Wold representation.** The ε_t's are often called **innovations,** because (as we'll see in Chapter 9) they correspond to the 1-step-ahead forecast errors that we'd make if we were to use a particularly good forecast. That is, the ε_t's represent that part of the evolution of y that's linearly unpredictable on the basis of the past of y. Note also that the ε_t's, although uncorrelated, are not necessarily independent. Again, it's only for Gaussian random variables that lack of correlation implies independence, and the innovations are not necessarily Gaussian.

---

[11] This section is a bit more abstract than others, but don't be put off. On the contrary, you may want to read it several times. The material in it is crucially important for time series modeling and forecasting and is therefore central to our concerns.

[12] Moreover, we require that the covariance stationary processes don't contain any deterministic components.

In our statement of Wold's theorem we assumed a zero mean. That may seem restrictive, but it's not. Rather, whenever you see y_t, just read $y_t - \mu$, so that the process is expressed in deviations from its mean. The deviation from the mean has a zero mean, by construction. Working with zero-mean processes therefore involves no loss of generality while facilitating notational economy. We'll use this device frequently.

THE GENERAL LINEAR PROCESS

Wold's theorem tells us that when formulating forecasting models for covariance stationary time series, we need only consider models of the form

$$y_t = B(L)\varepsilon_t = \sum_{i=0}^{\infty} b_i\varepsilon_{t-i}$$

$$\varepsilon_t \sim WN(0, \sigma^2),$$

where the b_i are coefficients with $b_0 = 1$ and $\sum_{i=0}^{\infty} b_i^2 < \infty$. We call this the **general linear process,** "general" because any covariance stationary series can be written that way, and "linear" because the Wold representation expresses the series as a linear function of its innovations.

The general linear process is so important that it's worth examining its unconditional and conditional moment structure in some detail. Taking means and variances, we obtain the unconditional moments

$$E(y_t) = E\left(\sum_{i=0}^{\infty} b_i\varepsilon_{t-i}\right) = \sum_{i=0}^{\infty} b_i E(\varepsilon_{t-i}) = \sum_{i=0}^{\infty} b_i \cdot 0 = 0$$

and

$$\operatorname{var}(y_t) = \operatorname{var}\left(\sum_{i=0}^{\infty} b_i\varepsilon_{t-i}\right) = \sum_{i=0}^{\infty} b_i^2 \operatorname{var}(\varepsilon_{t-i}) = \sum_{i=0}^{\infty} b_i^2\sigma^2 = \sigma^2\sum_{i=0}^{\infty} b_i^2 .$$

At this point, in parallel to our discussion of white noise, we could compute and examine the autocovariance and autocorrelation functions of the general linear process. Those calculations, however, are rather involved, and not particularly revealing, so we'll proceed instead to examine the conditional mean and variance, where the information set Ω_{t-1} on which we condition contains past innovations; that is, $\Omega_{t-1} = \{\varepsilon_{t-1}, \varepsilon_{t-2}, \ldots\}$. In this manner, we can see how dynamics are modeled via conditional moments.[13] The conditional mean is

---

[13] Although Wold's theorem guarantees only serially uncorrelated white noise innovations, we shall sometimes make a stronger assumption of independent white noise innovations to focus the discussion. We do so, for example, in the following characterization of the conditional moment structure of the general linear process.

$$\begin{aligned} E(y_t \mid \Omega_{t-1}) &= E(\varepsilon_t \mid \Omega_{t-1}) + b_1 E(\varepsilon_{t-1} \mid \Omega_{t-1}) + b_2 E(\varepsilon_{t-2} \mid \Omega_{t-1}) + \cdots \\ &= 0 + b_1\varepsilon_{t-1} + b_2\varepsilon_{t-2} + \cdots = \sum_{i=1}^{\infty} b_i \varepsilon_{t-i} \,, \end{aligned}$$

and the conditional variance is

$$\operatorname{var}(y_t \mid \Omega_{t-1}) = E((y_t - E(y_t \mid \Omega_{t-1}))^2 \mid \Omega_{t-1}) = E(\varepsilon_t^2 \mid \Omega_{t-1}) = E(\varepsilon_t^2) = \sigma^2 .$$

The key insight is that the conditional mean *moves* over time in response to the evolving information set. The model captures the dynamics of the process, and the evolving conditional mean is one crucial way of summarizing them. An important goal of time series modeling, especially for forecasters, is capturing such conditional mean dynamics—the unconditional mean is constant (a requirement of stationarity), but the conditional mean varies in response to the evolving information set.[14]

RATIONAL DISTRIBUTED LAGS

As we've seen, the Wold representation points to the crucial importance of models with infinite distributed lags. Infinite distributed lag models, in turn, are stated in terms of infinite polynomials in the lag operator, which are therefore very important as well. Infinite distributed lag models are not of immediate practical use, however, because they contain infinitely many parameters, which certainly inhibits practical application! Fortunately, infinite polynomials in the lag operator needn't contain infinitely many free parameters. The infinite polynomial $B(L)$ may, for example, be a ratio of finite-order (and perhaps very low-order) polynomials. Such polynomials are called **rational polynomials,** and distributed lags constructed from them are called **rational distributed lags.**

Suppose, for example, that

$$B(L) = \frac{\Theta(L)}{\Phi(L)} \,,$$

where the numerator polynomial is of degree q,

$$\Theta(L) = \sum_{i=0}^{q} \theta_i L^i \,,$$

and the denominator polynomial is of degree p,

$$\Phi(L) = \sum_{i=0}^{p} \varphi_i L^i .$$

There are *not* infinitely many free parameters in the $B(L)$ polynomial; instead, there are only $p + q$ parameters (the θ's and the φ's). If p and q are small—say,

---

[14] Note, however, an embarrassing asymmetry: the conditional variance, like the unconditional variance, is a fixed constant. However, models that allow the conditional variance to change with the information set have been developed recently, as discussed in detail in Chapter 14.

0, 1, or 2—then what seems like a hopeless task—estimation of $B(L)$—may actually be easy.

More realistically, suppose that $B(L)$ is not exactly rational but is approximately rational,

$$B(L) \approx \frac{\Theta(L)}{\Phi(L)} .$$

Then we can find an **approximation of the Wold representation** using a rational distributed lag. Rational distributed lags produce models of cycles that economize on parameters (they're **parsimonious**), while nevertheless providing accurate approximations to the Wold representation. The popular ARMA and ARIMA forecasting models, which we'll study shortly, are simply rational approximations to the Wold representation.

5. Estimation and Inference for the Mean, Autocorrelation, and Partial Autocorrelation Functions

Now suppose we have a sample of data on a time series, and we don't know the true model that generated the data, or the mean, autocorrelation function, or partial autocorrelation function associated with that true model. Instead, we want to use the data to *estimate* the mean, autocorrelation function, and partial autocorrelation function, which we might then use to help us learn about the underlying dynamics and to decide on a suitable model or set of models to fit to the data.

SAMPLE MEAN

The mean of a covariance stationary series is $\mu = Ey_t$. A fundamental principle of estimation, called the **analog principle,** suggests that we develop estimators by replacing expectations with sample averages. Thus, our estimator for the population mean, given a sample of size T, is the **sample mean,**

$$\bar{y} = \frac{1}{T}\sum_{t=1}^{T} y_t .$$

Typically we're not directly interested in the estimate of the mean, but it's needed for estimation of the autocorrelation function.

SAMPLE AUTOCORRELATIONS

The autocorrelation at displacement τ for the covariance stationary series y is

$$\rho(\tau) = \frac{E((y_t - \mu)(y_{t-\tau} - \mu))}{E((y_t - \mu)^2)} .$$

Application of the analog principle yields a natural estimator,

$$\hat{\rho}(\tau) = \frac{\frac{1}{T}\sum_{t=\tau+1}^{T}((y_t - \bar{y})(y_{t-\tau} - \bar{y}))}{\frac{1}{T}\sum_{t=1}^{T}(y_t - \bar{y})^2} = \frac{\sum_{t=\tau+1}^{T}((y_t - \bar{y})(y_{t-\tau} - \bar{y}))}{\sum_{t=1}^{T}(y_t - \bar{y})^2} .$$

This estimator, viewed as a function of τ, is called the **sample autocorrelation function** or correlogram. Note that some of the summations begin at $t = \tau + 1$, not at $t = 1$; this is necessary because of the appearance of $y_{t-\tau}$ in the sum. Note that we divide those same sums by T, even though only $(T - \tau)$ terms appear in the sum. When T is large relative to τ (which is the relevant case), division by T or by $T - \tau$ will yield approximately the same result, so it won't make much difference for practical purposes; moreover, there are good mathematical reasons for preferring division by T.[15]

It's often of interest to assess whether a series is reasonably approximated as white noise, which is to say whether all its autocorrelations are 0 in population. A key result, which we simply assert, is that if a series is white noise, then the distribution of the sample autocorrelations in large samples is

$$\hat{\rho}(\tau) \sim N\left(0, \frac{1}{T}\right) .$$

Note how simple the result is. The sample autocorrelations of a white noise series are approximately normally distributed, and the normal is always a convenient distribution to work with. Their mean is 0, which is to say the sample autocorrelations are unbiased estimators of the true autocorrelations, which are in fact 0. Finally, the variance of the sample autocorrelations is approximately $1/T$ (equivalently, the standard deviation is $1/\sqrt{T}$), which is easy to construct and remember. Under normality, taking plus or minus two standard errors yields an approximate 95% confidence interval. Thus, if the series is white noise, then approximately 95% of the sample autocorrelations should fall in the interval $\pm\frac{2}{\sqrt{T}}$. In practice, when we plot the sample autocorrelations for a sample of data, we typically include the "two-standard-error bands," which are useful for making informal graphical assessments of whether and how the series deviates from white noise.

The two-standard-error bands, although very useful, only provide 95% bounds for the sample autocorrelations taken one at a time. Ultimately, we're often interested in whether a series is white noise—that is, whether *all* its autocorrelations are *jointly* 0. A simple extension lets us test that hypothesis. Rewrite the expression

$$\hat{\rho}(\tau) \sim N\left(0, \frac{1}{T}\right)$$

as

$$\sqrt{T}\hat{\rho}(\tau) \sim N(0, 1) .$$

---

[15] For additional discussion, consult any of the more advanced time series texts mentioned in Chapter 1.

Squaring both sides yields[16]

$$T\hat{\rho}^2(\tau) \sim \chi_1^2 .$$

It can be shown that, in addition to being approximately normally distributed, the sample autocorrelations at various displacements are approximately independent of one another. Recalling that the sum of independent χ^2 variables is also χ^2 with degrees of freedom equal to the sum of the degrees of freedom of the variables summed, we have shown that the **Box-Pierce Q-statistic,**

$$Q_{BP} = T\sum_{\tau=1}^{m} \hat{\rho}^2(\tau) ,$$

is approximately distributed as a χ_m^2 random variable under the null hypothesis that y is white noise.[17] A slight modification of this, designed to follow more closely the χ^2 distribution in small samples, is

$$Q_{LB} = T(T+2)\sum_{\tau=1}^{m} \left(\frac{1}{T-\tau}\right) \hat{\rho}^2(\tau) .$$

Under the null hypothesis that y is white noise, Q_{LB} is approximately distributed as a χ_m^2 random variable. Note that the **Ljung-Box Q-statistic** is the same as the Box-Pierce Q-statistic, except that the sum of squared autocorrelations is replaced by a weighted sum of squared autocorrelations, where the weights are $(T+2)/(T-\tau)$. For moderate and large T, the weights are approximately 1, so that the Ljung-Box statistic differs little from the Box-Pierce statistic.

Selection of m is done to balance competing criteria. On the one hand, we don't want m too small, because, after all, we're trying to do a joint test on a large part of the autocorrelation function. On the other hand, as m grows relative to T, the quality of the distributional approximations we've invoked deteriorates. In practice, focusing on m in the neighborhood of $\sqrt{T}$ is often reasonable.

SAMPLE PARTIAL AUTOCORRELATIONS

Recall that the partial autocorrelations are obtained from population linear regressions, which correspond to a thought experiment involving linear regression using an infinite sample of data. The sample partial autocorrelations correspond to the same thought experiment, except that the linear regression is now done on the (feasible) sample of size T. If the fitted regression is

$$\hat{y}_t = \hat{c} + \hat{\beta}_1 y_{t-1} + \cdots + \hat{\beta}_\tau y_{t-\tau} ,$$

then the **sample partial autocorrelation** at displacement τ is

$$\hat{p}(\tau) \equiv \hat{\beta}_\tau .$$

---

[16] Recall that the square of a standard normal random variable is a χ^2 random variable with 1 degree of freedom. We square the sample autocorrelations $\hat{\rho}(\tau)$ so that positive and negative values don't cancel when we sum across various values of τ, as we will soon do.

[17] m is a maximum displacement selected by the user. Shortly we'll discuss how to choose it.

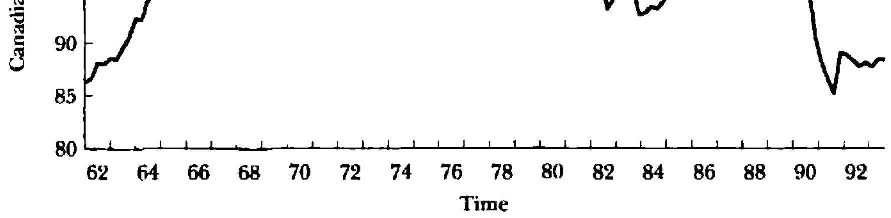
Distributional results identical to those we discussed for the sample autocorrelations hold as well for the sample *partial* autocorrelations. That is, if the series is white noise, approximately 95% of the sample partial autocorrelations should fall in the interval $\pm \frac{2}{\sqrt{T}}$. As with the sample autocorrelations, we typically plot the sample partial autocorrelations along with their two-standard-error bands.

6. Application: Characterizing Canadian Employment Dynamics

To illustrate the ideas we've introduced, we examine a quarterly, seasonally adjusted index of Canadian employment, 1962.1–1993.4, which we plot in Figure 7.9. The series displays no trend, and of course it displays no seasonality because it's seasonally adjusted. It does, however, appear highly serially correlated. It evolves in a slow, persistent fashion—high in business cycle booms and low in recessions.

To get a feel for the dynamics operating in the employment series, we perform a correlogram analysis.[18] The results appear in Table 7.1. Consider first the Q-statistic.[19] We compute the Q-statistic and its p-value under the null hypothesis of white noise for values of m (the number of terms in the sum that

FIGURE 7.9
Canadian Employment Index



---

[18] A **correlogram analysis** simply means examination of the sample autocorrelation and partial autocorrelation functions (with two-standard-error bands), together with related diagnostics, such as Q-statistics.

[19] We show the Ljung-Box version of the Q-statistic.

TABLE 7.1
Canadian Employment Index, Correlogram

Sample: 1962:1 1993:4
Included observations: 128

	Acorr.	P. Acorr.	Std. Error	Ljung-Box	*p*-value
1	0.949	0.949	.088	118.07	0.000
2	0.877	−0.244	.088	219.66	0.000
3	0.795	−0.101	.088	303.72	0.000
4	0.707	−0.070	.088	370.82	0.000
5	0.617	−0.063	.088	422.27	0.000
6	0.526	−0.048	.088	460.00	0.000
7	0.438	−0.033	.088	486.32	0.000
8	0.351	−0.049	.088	503.41	0.000
9	0.258	−0.149	.088	512.70	0.000
10	0.163	−0.070	.088	516.43	0.000
11	0.073	−0.011	.088	517.20	0.000
12	−0.005	0.016	.088	517.21	0.000

underlies the Q-statistic) ranging from 1 through 12. The *p*-value is consistently 0 to four decimal places, so the null hypothesis of white noise is decisively rejected.

Now we examine the sample autocorrelations and partial autocorrelations. The sample autocorrelations are very large relative to their standard errors and display slow one-sided decay.[20] The sample partial autocorrelations, in contrast, are large relative to their standard errors at first (particularly for the one-quarter displacement) but are statistically negligible beyond displacement 2.[21] In Figure 7.10 we plot the sample autocorrelations and partial autocorrelations along with their two-standard-error bands.

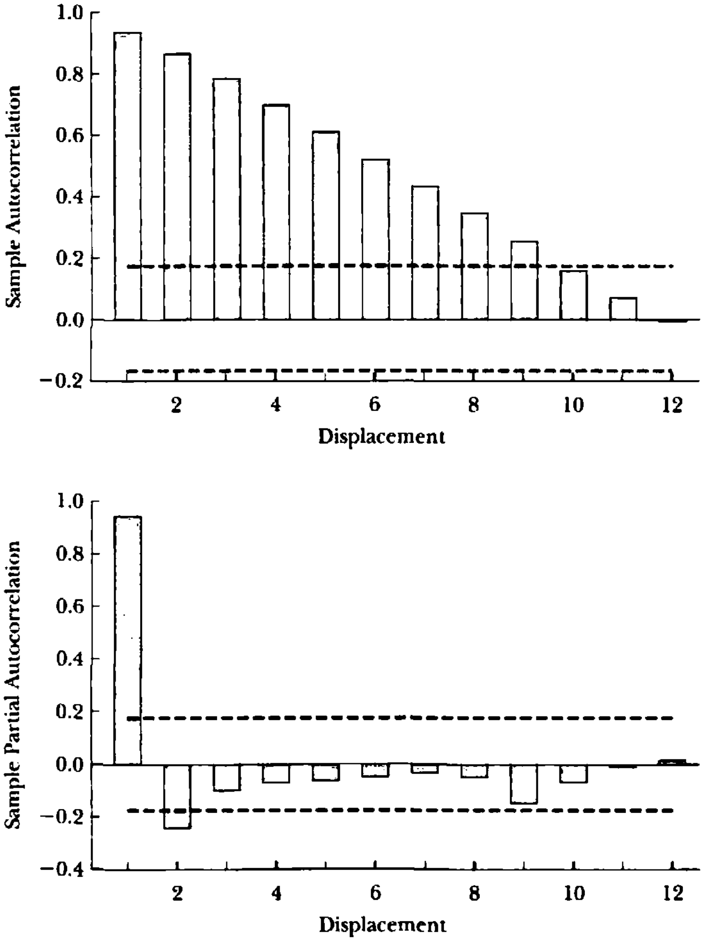
It's clear that employment has a strong cyclical component; all diagnostics reject the white noise hypothesis immediately. Moreover, the sample autocorrelation and partial autocorrelation functions have particular shapes—the autocorrelation function displays slow one-sided damping, while the partial autocorrelation function cuts off at displacement 2. You might guess that such patterns, which summarize the dynamics in the series, might be useful for suggesting candidate forecasting models. Such is indeed the case, as we'll see in the next chapter.

---

[20] We don't show the sample autocorrelation or partial autocorrelation at displacement 0, because as we mentioned earlier, they equal 1.0, by construction, and therefore convey no useful information. We'll adopt this convention throughout.

[21] Note that the sample autocorrelation and partial autocorrelation are identical at displacement 1. That's because at displacement 1, there are no earlier lags to control for when computing the sample partial autocorrelation, so it equals the sample autocorrelation. At higher displacements, of course, the two diverge.

FIGURE 7.10 *Canadian Employment Index: Sample Autocorrelation and Partial Autocorrelation Functions, with Plus or Minus Two-Standard-Error Bands*



Exercises, Problems, and Complements

1. (Lag operator expressions 1) Rewrite the following expressions without using the lag operator.
   a. $(L^{\tau})y_t = \varepsilon_t$
   b. $y_t = \left(\dfrac{2 + 5L + 0.8L^2}{L - 0.6L^3}\right)\varepsilon_t$
   c. $y_t = 2\left(1 + \dfrac{L^3}{L}\right)\varepsilon_t.$

2. (Lag operator expressions 2) Rewrite the following expressions in lag operator form.
   a. $y_t + y_{t-1} + \cdots + y_{t-N} = \alpha + \varepsilon_t + \varepsilon_{t-1} + \cdots + \varepsilon_{t-N}$, where α is a constant
   b. $y_t = \varepsilon_{t-2} + \varepsilon_{t-1} + \varepsilon_t$.

3. (Autocorrelation functions of covariance stationary series) While interviewing at a top investment bank, your interviewer is impressed by the fact that you have taken a course on time series forecasting. She decides to test your knowledge of the autocovariance structure of covariance stationary series and lists four autocovariance functions:
   a. $\gamma(t, \tau) = \alpha$,
   b. $\gamma(t, \tau) = e^{-\alpha\tau}$,
   c. $\gamma(t, \tau) = \alpha\tau$, and
   d. $\gamma(t, \tau) = \dfrac{\alpha}{\tau}$,

   where α is a positive constant. Which autocovariance function(s) are consistent with covariance stationarity, and which are not? Why?

4. (Autocorrelation vs. partial autocorrelation) Describe the difference between autocorrelations and partial autocorrelations. How can autocorrelations at certain displacements be positive while the partial autocorrelations at those same displacements are negative?

5. (Conditional and unconditional means) As head of sales of the leading technology and innovation magazine publisher TECCIT, your bonus is dependent on the firm's revenue. Revenue changes from season to season, as subscriptions and advertising deals are entered or renewed. From your experience in the publishing business, you know that the revenue in a season is a function of the number of magazines sold in the previous season and can be described as $y_t = 1000 + 0.9x_{t-1} + \varepsilon_t$, with uncorrelated residuals $\varepsilon_t \sim N(0, 1000)$, where y is revenue and x is the number of magazines sold.
   a. What is the expected revenue for next season conditional on total sales of 6340 this season?
   b. What is unconditionally expected revenue if unconditionally expected sales are 8500?
   c. A rival publisher offers you a contract identical to your current contract (same base pay and bonus). Based on a confidential interview, you know that the same revenue model with identical coefficients is appropriate for your rival. The rival has sold an average of 9000 magazines in previous seasons but only 5650 this season. Will you accept the offer? Why or why not?

6. (White noise residuals) You work for a top five consulting firm and are in the middle of a 1-week vacation, when one of the directors calls you and urges you immediately to join a turnaround project at Stardust Cinemas. You are briefed that despite its bad financial condition, the recently fired CEO had planned to increase Stardust's market share by renovating every theater to include a bar, an arcade, and a restaurant. Your task on the team is to assess whether this renovation should be scrapped or included in a future value creation project. To do so, you spend a long night fitting a trend + seasonal model to a sample of $T = 100$ observations of Stardust's recent box office income data. You find that the residuals (e) from your model approximately follow $e_t = 0.5e_{t-1} + v_t$, where $v_t \overset{iid}{\sim} N(0, 1)$. At 4 A.M. you send your results to your project manager.

a. The next morning you receive an e-mail from your project manager. He thinks that your residuals do not look like white noise. Why? Why care?

b. Assuming that the residuals do indeed follow $e_t = 0.5e_{t-1} + v_t$, what is their autocorrelation function? Discuss.

c. What type of model might be useful for describing the historical path of box office income and its likely future path in the absence of renovations? How would you use it to assess the efficacy of the renovation project, if implemented?

7. (Selecting an employment forecasting model with the AIC and SIC) Use the AIC and SIC to assess the necessity and desirability of including trend and seasonal components in a forecasting model for Canadian employment.

   a. Display the AIC and SIC for a variety of specifications of trend and seasonality. Which would you select using the AIC? SIC? Do the AIC and SIC select the same model? If not, which do you prefer?

   b. Discuss the estimation results and residual plot from your preferred model, and perform a correlogram analysis of the residuals. Discuss, in particular, the patterns of the sample autocorrelations and partial autocorrelations, and their statistical significance.

   c. How, if at all, are your results different from those reported in the text? Are the differences important? Why or why not?

8. (**Simulation of a time series process**) Many cutting-edge estimation and forecasting techniques involve simulation. Moreover, simulation is often a good way to get a feel for a model and its behavior. White noise can be simulated on a computer using **random number generators,** which are available in most statistics, econometrics, and forecasting packages.

   a. Simulate a Gaussian white noise realization of length 200. Call the white noise ε_t. Compute the correlogram. Discuss.

   b. Form the distributed lag $y_t = \varepsilon_t + 0.9\varepsilon_{t-1}$, $t = 2, 3, \ldots, 200$. Compute the sample autocorrelations and partial autocorrelations. Discuss.

   c. Let $y_1 = 1$ and $y_t = 0.9y_{t-1} + \varepsilon_t$, $t = 2, 3, \ldots, 200$. Compute the sample autocorrelations and partial autocorrelations. Discuss.

9. (Sample autocorrelation functions for trending series) A telltale sign of the slowly evolving nonstationarity associated with trend is a sample autocorrelation function that damps extremely slowly.

   a. Find three trending series, compute their sample autocorrelation functions, and report your results. Discuss.

   b. Fit appropriate trend models, obtain the model residuals, compute their sample autocorrelation functions, and report your results. Discuss.

10. (Sample autocorrelation functions for seasonal series) A telltale sign of seasonality is a sample autocorrelation function with sharp peaks at the seasonal displacements (4, 8, 12, etc., for quarterly data; 12, 24, 36, etc., for monthly data; etc.).

    a. Find a series with both trend and seasonal variation. Compute its sample autocorrelation function. Discuss.

    b. Detrend the series. Discuss.

    c. Compute the sample autocorrelation function of the detrended series. Discuss.

    d. Seasonally adjust the detrended series. Discuss.

    e. Compute the sample autocorrelation function of the detrended, seasonally-adjusted series. Discuss.

11. (Volatility dynamics: correlograms of squares) In Chapter 4's Exercises, Problems, and Complements, we suggested that a time series plot of a squared residual, e_t^2, can reveal serial correlation in squared residuals, which corresponds to nonconstant volatility, or heteroskedasticity, in the levels of the residuals. Financial asset returns often display little systematic variation, so instead of examining residuals from a model of returns, we often examine returns directly. In what follows, we will continue to use the notation e_t, but you should interpret e_t as an observed asset return.
    a. Find a high-frequency (e.g., daily) financial asset return series, e_t, plot it, and discuss your results.
    b. Perform a correlogram analysis of e_t, and discuss your results.
    c. Plot e_t^2, and discuss your results.
    d. In addition to plotting e_t^2, examining the correlogram of e_t^2 often proves informative for assessing volatility persistence. Why might that be so? Perform a correlogram analysis of e_t^2, and discuss your results.

Bibliographical and Computational Notes

Wold's theorem was originally proved in a 1938 monograph, later revised as Wold (1954). Rational distributed lags have long been used in engineering, and their use in econometric modeling dates at least to Jorgenson (1966).

Bartlett (1946) derived the standard errors of the sample autocorrelations and partial autocorrelations of white noise. In fact, the plus-or-minus two-standard-error bands are often called the "**Bartlett bands.**"

The two variants of the Q-statistic that we introduced were developed in the 1970s by Box and Pierce (1970) and by Ljung and Box (1978). Some packages compute both variants, and some compute only one (typically Ljung-Box, because it's designed to be more accurate in small samples). In practice, the Box-Pierce and Ljung-Box statistics usually lead to the same conclusions.

For concise and insightful discussion of random number generation, as well as a variety of numerical and computational techniques, see Press *et al.* (1992).

Concepts for Review

Cycle
Realization
Sample path
Covariance stationarity
Autocovariance function
Second-order stationarity
Weak stationarity
Autocorrelation function
Partial autocorrelation function
Population regression
Autoregression
Time series process
Serially uncorrelated
Zero-mean white noise
White noise
Strong white noise
Weak white noise
Independent white noise
Normal white noise
Gaussian white noise
Unconditional mean and variance
Conditional mean and variance

Moments
Lag operator
Polynomial in the lag operator
Distributed lag
Wold's representation theorem
Wold representation
Innovation
General linear process
Rational polynomial
Rational distributed lag
Approximation of the Wold representation
Parsimonious
Analog principle
Sample mean
Sample autocorrelation function
Box-Pierce Q-statistic
Ljung-Box Q-statistic
Sample partial autocorrelation
Correlogram analysis
Simulation of a time series process
Random number generator
Bartlett bands

References and Additional Readings

Bartlett, M. (1946). "On the Theoretical Specification of Sampling Properties of Autocorrelated Time Series." *Journal of the Royal Statistical Society* B, 8, 27–41.

Box, G. E. P., and Pierce, D. A. (1970). "Distribution of Residual Autocorrelations in ARIMA Time-Series Models." *Journal of the American Statistical Association*, 65, 1509–1526.

Jorgenson, D. (1966). "Rational Distributed Lag Functions." *Econometrica*, 34, 135–149.

Ljung, G. M., and G. E. P. Box (1978). "On a Measure of Lack of Fit in Time-Series Models." *Biometrika*, 65, 297–303.

Press, W. H., et al. (1992). *Numerical Recipes: The Art of Scientific Computing.* Cambridge: Cambridge University Press.

Wold, H. O. (1954). *A Study in the Analysis of Stationary Time Series.* 2nd ed. Uppsala, Sweden: Almquist and Wicksell.

Wonnacott, T. H., and Wonnacott, R. J. (1990). *Introductory Statistics.* 5th ed. New York: Wiley.

Modeling Cycles: MA, AR, and ARMA Models

When building forecasting models, we don't want to pretend that the model we fit is true. Instead, we want to be aware that we're *approximating* a more complex reality. That's the modern view, and it has important implications for forecasting. In particular, we've seen that the key to successful time series modeling and forecasting is parsimonious, yet accurate, approximation of the Wold representation. In this chapter, we consider three approximations: **moving average (MA) models, autoregressive (AR) models,** and **autoregressive moving average (ARMA) models.** The three models vary in their specifics and have different strengths in capturing different sorts of autocorrelation behavior.

We begin by characterizing the autocorrelation functions and related quantities associated with each model, under the assumption that the model is "true." We do this separately for MA, AR, and ARMA models.[1] These characterizations have nothing to do with data or estimation, but they're crucial for developing a basic understanding of the properties of the models, which is necessary to perform intelligent modeling and forecasting. They enable us to make statements such as "If the data were really generated by an autoregressive process, then we'd expect its autocorrelation function to have property x." Armed with that knowledge, we use the *sample* autocorrelations and partial autocorrelations, in conjunction with the AIC and the SIC, to suggest candidate forecasting models, which we then estimate.

---

[1] Sometimes, especially when characterizing population properties under the assumption that the models are correct, we refer to them as *processes*, which is short for **stochastic processes**—hence the terms *moving average process, autoregressive process,* and *ARMA process.*

1. Moving Average (MA) Models

The finite-order moving average process is a natural and obvious approximation to the Wold representation, which is an infinite-order moving average process. Finite-order moving average processes also have direct motivation. The fact that all variation in time series, one way or another, is driven by shocks of various sorts suggests the possibility of modeling time series directly as distributed lags of current and past shocks—that is, as moving average processes.[2]

THE MA(1) PROCESS

The first-order moving average process, or **MA(1) process,** is

$$y_t = \varepsilon_t + \theta\varepsilon_{t-1} = (1 + \theta L)\varepsilon_t$$

$$\varepsilon_t \sim WN(0, \sigma^2).$$

The defining characteristic of the MA process in general, and the MA(1) in particular, is that the current value of the observed series is expressed as a function of current and lagged unobservable shocks. Think of it as a regression model with nothing but current and lagged disturbances on the right-hand side.

To help develop a feel for the behavior of the MA(1) process, we show two simulated realizations of length 150 in Figure 8.1. The processes are

$$y_t = \varepsilon_t + 0.4\varepsilon_{t-1}$$

and

$$y_t = \varepsilon_t + 0.95\varepsilon_{t-1},$$

where in each case $\varepsilon_t \overset{iid}{\sim} N(0, 1)$. To construct the realizations, we used the same series of underlying white noise shocks; the only difference in the realizations comes from the different coefficients. Past shocks feed *positively* into the current value of the series, with a small weight of $\theta = 0.4$ in one case and a large weight of $\theta = 0.95$ in the other. You might think that $\theta = 0.95$ would induce much more persistence than $\theta = 0.4$, but it doesn't. The structure of the

---

[2] Economic equilibria, for example, may be disturbed by shocks that take some time to be fully assimilated.

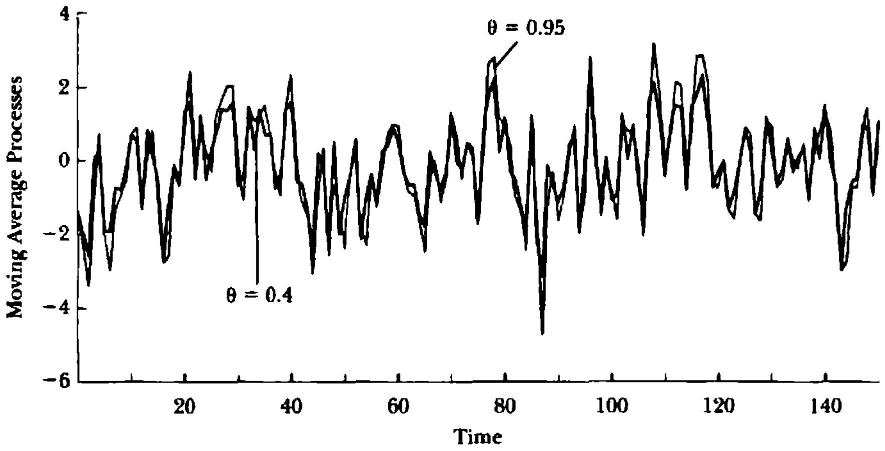


FIGURE 8.1
Realizations of Two MA(1) Processes

MA(1) process, in which only the first lag of the shock appears on the right, forces it to have a very short memory, and hence weak dynamics, regardless of the parameter value.

The unconditional mean and variance are

$$E(y_t) = E(\varepsilon_t) + \theta E(\varepsilon_{t-1}) = 0$$

and

$$\text{var}(y_t) = \text{var}(\varepsilon_t) + \theta^2 \text{var}(\varepsilon_{t-1}) = \sigma^2 + \theta^2\sigma^2 = \sigma^2(1+\theta^2).$$

Note that for a fixed value of σ, as θ increases in absolute value, so, too, does the unconditional variance. That's why the MA(1) process with parameter $\theta = 0.95$ varies a bit more than the process with a parameter of $\theta = 0.4$.

The conditional mean and variance of an MA(1), where the conditioning information set is $\Omega_{t-1} = \{\varepsilon_{t-1}, \varepsilon_{t-2}, \ldots\}$, are

$$E(y_t \mid \Omega_{t-1}) = E(\varepsilon_t + \theta\varepsilon_{t-1} \mid \Omega_{t-1}) = E(\varepsilon_t \mid \Omega_{t-1}) + \theta E(\varepsilon_{t-1} \mid \Omega_{t-1}) = \theta\varepsilon_{t-1}$$

and

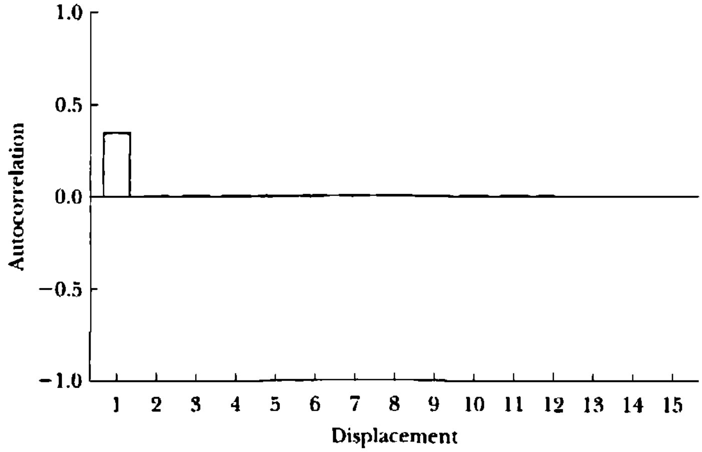
$$\text{var}(y_t \mid \Omega_{t-1}) = E((y_t - E(y_t \mid \Omega_{t-1}))^2 \mid \Omega_{t-1}) = E(\varepsilon_t^2 \mid \Omega_{t-1}) = E(\varepsilon_t^2) = \sigma^2.$$

The conditional mean explicitly adapts to the information set, in contrast to the unconditional mean, which is constant. Note, however, that only the first lag of the shock enters the conditional mean—more distant shocks have no effect on the current conditional expectation. This is indicative of the one-period memory of MA(1) processes, which we'll now characterize in terms of the autocorrelation function.

To compute the autocorrelation function for the MA(1) process, we must first compute the autocovariance function. We have

$$\gamma(\tau) = E(y_t y_{t-\tau}) = E((\varepsilon_t + \theta\varepsilon_{t-1})(\varepsilon_{t-\tau} + \theta\varepsilon_{t-\tau-1})) = \begin{cases} \theta\sigma^2, & \tau = 1 \\ 0, & \text{otherwise}. \end{cases}$$

FIGURE 8.2
Population Autocorrelation Function, MA(1) Process, $\theta = 0.4$

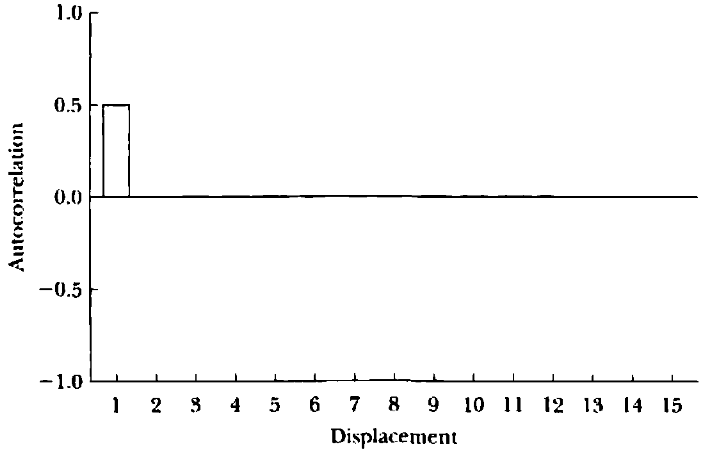


(The proof is left as a problem.) The autocorrelation function is just the autocovariance function scaled by the variance,

$$\rho(\tau) = \frac{\gamma(\tau)}{\gamma(0)} = \begin{cases} \frac{\theta}{1+\theta^2}, & \tau = 1 \\ 0, & \text{otherwise} . \end{cases}$$

The key feature here is the sharp **cutoff in the autocorrelation function.** All autocorrelations are 0 beyond displacement 1, the order of the MA process. In Figures 8.2 and 8.3, we show the autocorrelation functions for our two MA(1) processes with parameters $\theta = 0.4$ and $\theta = 0.95$. At displacement 1, the process with parameter $\theta = 0.4$ has a smaller autocorrelation (0.34)

FIGURE 8.3
Population Autocorrelation Function, MA(1) Process, $\theta = 0.95$



than the process with parameter $\theta = 0.95$ (0.50), but both drop to 0 beyond displacement 1.

Note that the requirements of covariance stationarity (constant unconditional mean, constant and finite unconditional variance, autocorrelation dependent only on displacement) are met for any MA(1) process, *regardless* of the values of its parameters. If, moreover, $|\theta| < 1$, then we say that the MA(1) process is **invertible.** In that case, we can "invert" the MA(1) process and express the current value of the series not in terms of a current shock and a lagged shock but rather in terms of a current shock *and lagged values of the series.* That's called an **autoregressive representation.** An autoregressive representation has a current shock and lagged observable values of the series on the right, whereas a moving average representation has a current shock and lagged unobservable shocks on the right.

Let's compute the autoregressive representation. The process is

$$y_t = \varepsilon_t + \theta\varepsilon_{t-1}$$

$$\varepsilon_t \sim WN(0, \sigma^2).$$

Thus, we can solve for the innovation as

$$\varepsilon_t = y_t - \theta\varepsilon_{t-1}.$$

Lagging by successively more periods gives expressions for the innovations at various dates,

$$\varepsilon_{t-1} = y_{t-1} - \theta\varepsilon_{t-2}$$

$$\varepsilon_{t-2} = y_{t-2} - \theta\varepsilon_{t-3}$$

$$\varepsilon_{t-3} = y_{t-3} - \theta\varepsilon_{t-4},$$

and so forth. Making use of these expressions for lagged innovations, we can substitute backward in the MA(1) process, yielding

$$y_t = \varepsilon_t + \theta y_{t-1} - \theta^2 y_{t-2} + \theta^3 y_{t-3} - \cdots.$$

In lag operator notation, we write the infinite autoregressive representation as

$$\frac{1}{1+\theta L} y_t = \varepsilon_t.$$

Note that the back substitution used to obtain the autoregressive representation only makes sense, and in fact a convergent autoregressive representation only exists, if $|\theta| < 1$, because in the back substitution we raise θ to progressively higher powers.

We can restate the invertibility condition in another way: The inverse of the root of the moving average lag operator polynomial $(1 + \theta L)$ must be less than 1 in absolute value. Recall that a polynomial of degree m has m roots. Thus, the MA(1) lag operator polynomial has one root, which is the solution to

$$1 + \theta L = 0.$$

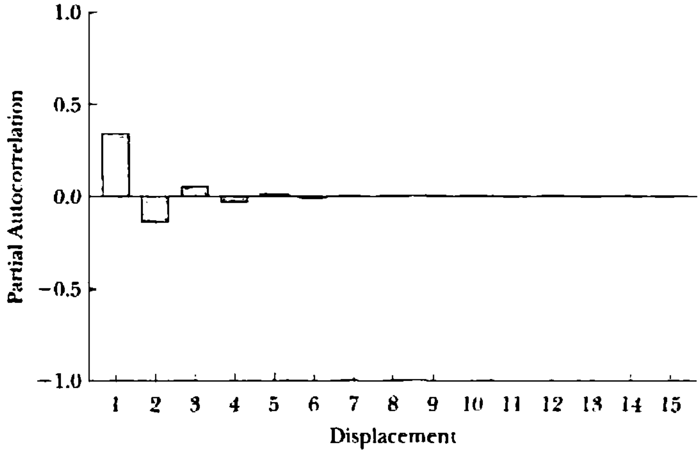
The root is $L = -1/\theta$, so its inverse will be less than 1 in absolute value if $|\theta| < 1$, and the two invertibility conditions are equivalent. The "inverse root" way of stating invertibility conditions seems tedious, but it turns out to be of greater applicability than the $|\theta| < 1$ condition, as we'll see shortly.

Autoregressive representations are appealing to forecasters, because one way or another, if a model is to be used for real-world forecasting, it must link the present observables to the past history of observables, so that we can extrapolate to form a forecast of future observables based on present and past observables. Superficially, moving average models don't seem to meet that requirement, because the current value of a series is expressed in terms of current and lagged unobservable shocks, not observable variables. But under the invertibility conditions that we've described, moving average processes have equivalent autoregressive representations. Thus, although we want autoregressive representations for forecasting, we don't have to start with an autoregressive model. However, we typically restrict ourselves to invertible processes, because for forecasting purposes we want to be able to express current observables as functions of past observables.

Finally, let's consider the partial autocorrelation function for the MA(1) process. From the infinite autoregressive representation of the MA(1) process, we see that the partial autocorrelation function will decay gradually to 0. As we discussed in Chapter 7, the partial autocorrelations are just the coefficients on the last included lag in a sequence of progressively higher-order autoregressive approximations. If $\theta > 0$, then the pattern of decay will be one of damped oscillation; otherwise, the decay will be one-sided.

In Figures 8.4 and 8.5 we show the partial autocorrelation functions for our example MA(1) processes. For each process, $|\theta| < 1$, so that an autoregressive representation exists, and $\theta > 0$, so that the coefficients in the autoregressive

FIGURE 8.4
Population Partial Autocorrelation Function, MA(1) Process, $\theta = 0.4$



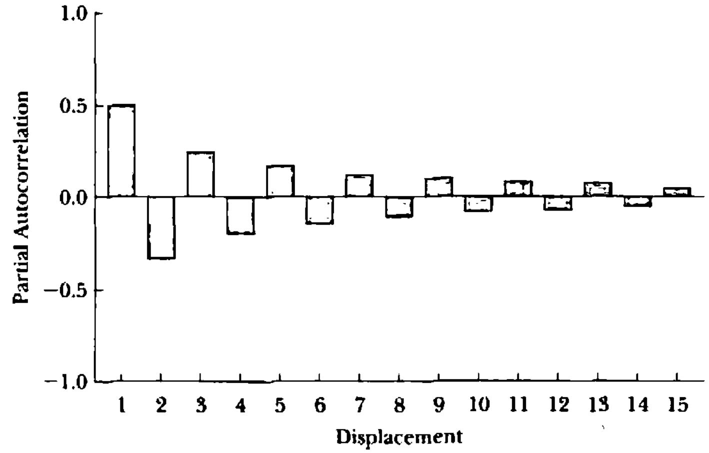


FIGURE 8.5 *Population Partial Autocorrelation Function, MA(1) Process, $\theta = 0.95$*

representations alternate in sign. Specifically, we showed the general autoregressive representation to be

$$y_t = \varepsilon_t + \theta y_{t-1} - \theta^2 y_{t-2} + \theta^3 y_{t-3} - \cdots,$$

so the autoregressive representation for the process with $\theta = 0.4$ is

$$y_t = \varepsilon_t + 0.4 y_{t-1} - 0.4^2 y_{t-2} + \cdots = \varepsilon_t + 0.4 y_{t-1} - 0.16 y_{t-2} + \cdots,$$

and the autoregressive representation for the process with $\theta = 0.95$ is

$$y_t = \varepsilon_t + 0.95 y_{t-1} - 0.95^2 y_{t-2} + \cdots = \varepsilon_t + 0.95 y_{t-1} - 0.9025 y_{t-2} + \cdots.$$

The partial autocorrelations display a similar damped oscillation.[3] The decay, however, is slower for the $\theta = 0.95$ case.

THE MA(q) PROCESS

Now consider the general finite-order moving average process of order q, or MA(q) for short,

$$y_t = \varepsilon_t + \theta_1 \varepsilon_{t-1} + \cdots + \theta_q \varepsilon_{t-q} = \Theta(L)\varepsilon_t$$
$$\varepsilon_t \sim WN(0, \sigma^2),$$

where

$$\Theta(L) = 1 + \theta_1 L + \cdots + \theta_q L^q$$

is a qth-order lag operator polynomial. The **MA(q) process** is a natural generalization of the MA(1). By allowing for more lags of the shock on the right side of the equation, the MA(q) process can capture richer dynamic patterns, which we can potentially exploit for improved forecasting. The MA(1) process is of course a special case of the MA(q), corresponding to $q = 1$.

---

[3] Note, however, that the partial autocorrelations are *not* the successive coefficients in the infinite autoregressive representation. Rather, they are the coefficients on the last included lag in sequence of progressively longer autoregressions. The two are related but distinct.

The properties of the MA(q) processes parallel those of the MA(1) process in all respects, so in what follows we'll refrain from grinding through the mathematical derivations. Instead, we'll focus on the key features of practical importance. Just as the MA(1) process was covariance stationary for any value of its parameters, so, too, is the finite-order MA(q) process. As with the MA(1) process, the MA(q) process is *invertible* only if a root condition is satisfied. The MA(q) lag operator polynomial has q roots; when $q > 1$, the possibility of **complex roots** arises. The **condition for invertibility of the MA(q) process** is that the inverses of all of the roots must be inside the unit circle, in which case we have the convergent autoregressive representation,

$$\frac{1}{\Theta(L)} y_t = \varepsilon_t .$$

The conditional mean of the MA(q) process evolves with the information set, in contrast to the unconditional moments, which are fixed. In contrast to the MA(1) case, in which the conditional mean depends on only the first lag of the innovation, in the MA(q) case the conditional mean depends on q lags of the innovation. Thus, the MA(q) process has the potential for longer memory.

The potentially longer memory of the MA(q) process emerges clearly in its autocorrelation function. In the MA(1) case, all autocorrelations beyond displacement 1 are 0; in the MA(q) case, all autocorrelations beyond displacement q are 0. This autocorrelation cutoff is a distinctive property of moving average processes. The partial autocorrelation function of the MA(q) process, in contrast, decays gradually, in accord with the infinite autoregressive representation, in either an oscillating or a one-sided fashion, depending on the parameters of the process.

In closing this section, let's step back for a moment and consider in greater detail the precise way in which finite-order moving average processes approximate the Wold representation. The Wold representation is

$$y_t = B(L)\varepsilon_t ,$$

where $B(L)$ is of infinite order. The MA(1), in contrast, is simply a first-order moving average, in which a series is expressed as a one-period moving average of current and past innovations. Thus, when we fit an MA(1) model, we're using the first-order polynomial $1 + \theta L$ to approximate the infinite-order polynomial $B(L)$. Note that $1 + \theta L$ is a rational polynomial with numerator polynomial of degree 1 and degenerate denominator polynomial (degree 0).

MA(q) processes have the potential to deliver better approximations to the Wold representation, at the cost of more parameters to be estimated. The Wold representation involves an infinite moving average; the MA(q) process approximates the infinite moving average with a *finite-order* moving average,

$$y_t = \Theta(L)\varepsilon_t ,$$

whereas the MA(1) process approximates the infinite moving average with only a *first-order* moving average, which can sometimes be very restrictive.

2. Autoregressive (AR) Models

The autoregressive process is also a natural approximation to the Wold representation. We've seen, in fact, that under certain conditions a moving average process has an autoregressive representation, so an autoregressive process is in a sense the same as a moving average process. Like the moving average process, the autoregressive process has direct motivation; it's simply a *stochastic difference equation*, a simple mathematical model in which the current value of a series is linearly related to its past values, plus an additive stochastic shock. Stochastic difference equations are a natural vehicle for discrete-time stochastic dynamic modeling.

THE AR(1) PROCESS

The first-order autoregressive process, AR(1) for short, is

$$y_t = \varphi y_{t-1} + \varepsilon_t$$

$$\varepsilon_t \sim WN(0, \sigma^2) .$$

In lag operator form, we write

$$(1 - \varphi L)y_t = \varepsilon_t .$$

In Figure 8.6 we show simulated realizations of length 150 of two AR(1) processes; the first is

$$y_t = 0.4y_{t-1} + \varepsilon_t ,$$

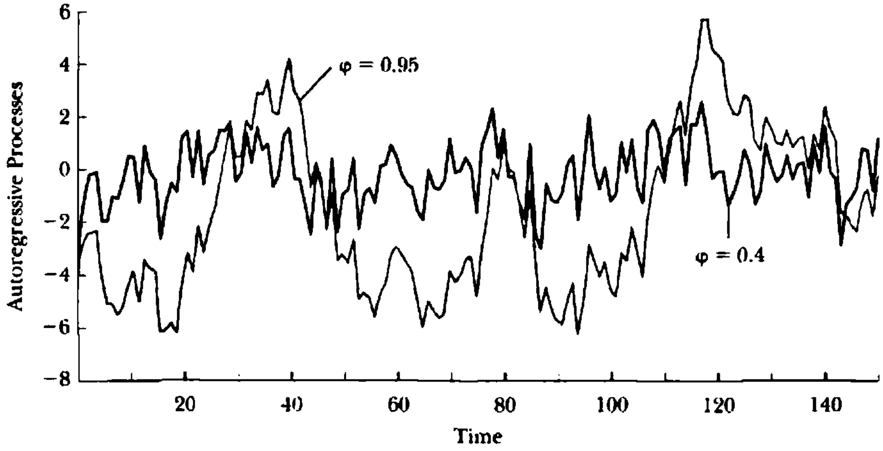


FIGURE 8.6
Realizations of Two AR(1) Processes

and the second is

$$y_t = 0.95y_{t-1} + \varepsilon_t,$$

where in each case $\varepsilon_t \overset{\text{iid}}{\sim} N(0, 1)$, and the same innovation sequence underlies each realization. The fluctuations in the AR(1) with parameter $\varphi = 0.95$ appear much more persistent that those of the AR(1) with parameter $\varphi = 0.4$. This contrasts sharply with the MA(1) process, which has a very short memory regardless of parameter value. Thus, the AR(1) model is capable of capturing much more persistent dynamics than is the MA(1).

Recall that a finite-order moving average process is always covariance stationary but that certain conditions must be satisfied for invertibility, in which case an autoregressive representation exists. For autoregressive processes, the situation is precisely the reverse. Autoregressive processes are always invertible—in fact, invertibility isn't even an issue, as finite-order autoregressive processes *already are* in autoregressive form—but certain conditions must be satisfied for an autoregressive process to be covariance stationary.

If we begin with the AR(1) process,

$$y_t = \varphi y_{t-1} + \varepsilon_t,$$

and substitute backward for lagged y's on the right side, we obtain

$$y_t = \varepsilon_t + \varphi\varepsilon_{t-1} + \varphi^2\varepsilon_{t-2} + \cdots .$$

In lag operator form, we write

$$y_t = \frac{1}{1-\varphi L}\varepsilon_t .$$

This moving average representation for y is convergent if and only if $|\varphi| < 1$; thus, $|\varphi| < 1$ is the condition for covariance stationarity in the AR(1) case. Equivalently, the condition for covariance stationarity is that the inverse of the root of the autoregressive lag operator polynomial be less than 1 in absolute value.

From the moving average representation of the covariance stationary AR(1) process, we can compute the unconditional mean and variance,

$$\begin{aligned} E(y_t) &= E(\varepsilon_t + \varphi\varepsilon_{t-1} + \varphi^2\varepsilon_{t-2} + \cdots) \\ &= E(\varepsilon_t) + \varphi E(\varepsilon_{t-1}) + \varphi^2 E(\varepsilon_{t-2}) + \cdots \\ &= 0 \end{aligned}$$

and

$$\begin{aligned} \operatorname{var}(y_t) &= \operatorname{var}(\varepsilon_t + \varphi\varepsilon_{t-1} + \varphi^2\varepsilon_{t-2} + \cdots) \\ &= \sigma^2 + \varphi^2\sigma^2 + \varphi^4\sigma^2 + \cdots \\ &= \sigma^2 \sum_{i=0}^{\infty} \varphi^{2i} \\ &= \frac{\sigma^2}{1-\varphi^2} . \end{aligned}$$

The conditional moments, in contrast, are

$$\begin{aligned}E(y_t \mid y_{t-1}) &= E(\varphi y_{t-1} + \varepsilon_t \mid y_{t-1})\\ &= \varphi E(y_{t-1} \mid y_{t-1}) + E(\varepsilon_t \mid y_{t-1})\\ &= \varphi y_{t-1} + 0\\ &= \varphi y_{t-1}\end{aligned}$$

and

$$\begin{aligned}\operatorname{var}(y_t \mid y_{t-1}) &= \operatorname{var}(\varphi y_{t-1} + \varepsilon_t \mid y_{t-1})\\ &= \varphi^2 \operatorname{var}(y_{t-1} \mid y_{t-1}) + \operatorname{var}(\varepsilon_t \mid y_{t-1})\\ &= 0 + \sigma^2\\ &= \sigma^2 .\end{aligned}$$

Note in particular the simple way in which the conditional mean adapts to the changing information set as the process evolves.

To find the autocovariances, we proceed as follows. The process is

$$y_t = \varphi y_{t-1} + \varepsilon_t ,$$

so that, multiplying both sides of the equation by $y_{t-\tau}$ we obtain

$$y_t y_{t-\tau} = \varphi y_{t-1} y_{t-\tau} + \varepsilon_t y_{t-\tau} .$$

For $\tau \geq 1$, taking expectations of both sides gives

$$\gamma(\tau) = \varphi\gamma(\tau - 1) .$$

This is called the **Yule-Walker equation.** It is a recursive equation; that is, given $\gamma(\tau)$, for any τ, the Yule-Walker equation immediately tells us how to get $\gamma(\tau + 1)$. If we knew $\gamma(0)$ to start things off (an "initial condition"), we could use the Yule-Walker equation to determine the entire autocovariance sequence. And we *do* know $\gamma(0)$; it's just the variance of the process, which we already showed to be $\gamma(0) = \frac{\sigma^2}{1-\varphi^2}$. Thus, we have

$$\begin{aligned}\gamma(0) &= \frac{\sigma^2}{1-\varphi^2}\\ \gamma(1) &= \varphi\frac{\sigma^2}{1-\varphi^2}\\ \gamma(2) &= \varphi^2\frac{\sigma^2}{1-\varphi^2} ,\end{aligned}$$

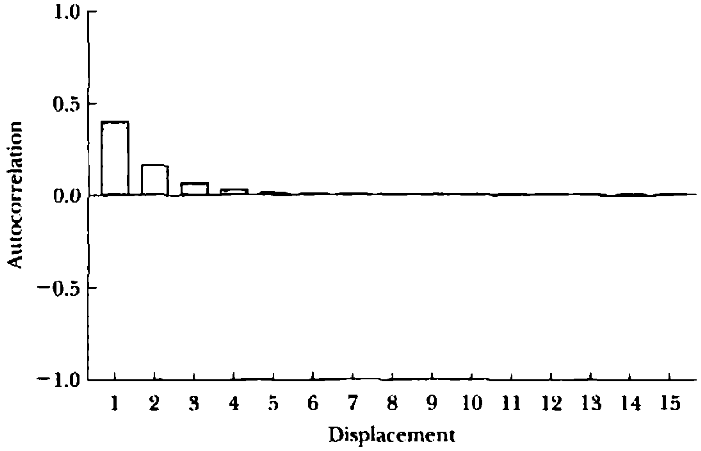
and so on. In general, then,

$$\gamma(\tau) = \varphi^\tau \frac{\sigma^2}{1-\varphi^2}, \quad \tau = 0, 1, 2, \ldots .$$

Dividing through by $\gamma(0)$ gives the autocorrelations,

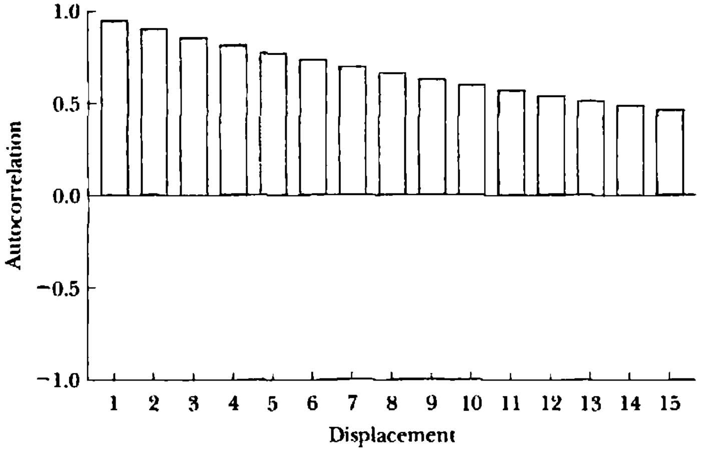
$$\rho(\tau) = \varphi^\tau, \quad \tau = 0, 1, 2, \ldots .$$

FIGURE 8.7
Population Autocorrelation Function, AR(1) Process, $\varphi = 0.4$



Note the gradual autocorrelation decay, which is typical of autoregressive processes. The autocorrelations approach 0, but only in the limit as the displacement approaches infinity. In particular, they don't cut off to 0, as is the case for moving average processes. If φ is positive, the autocorrelation decay is one-sided. If φ is negative, the decay involves back-and-forth oscillations. The relevant case in business and economics is $\varphi > 0$, but either way, the autocorrelations damp gradually, not abruptly. In Figures 8.7 and 8.8, we show the autocorrelation functions for AR(1) processes with parameters $\varphi = 0.4$ and $\varphi = 0.95$. The persistence is much stronger when $\varphi = 0.95$, in contrast to the MA(1) case, in which the persistence was weak regardless of the parameter.

FIGURE 8.8
Population Autocorrelation Function, AR(1) Process, $\varphi = 0.95$



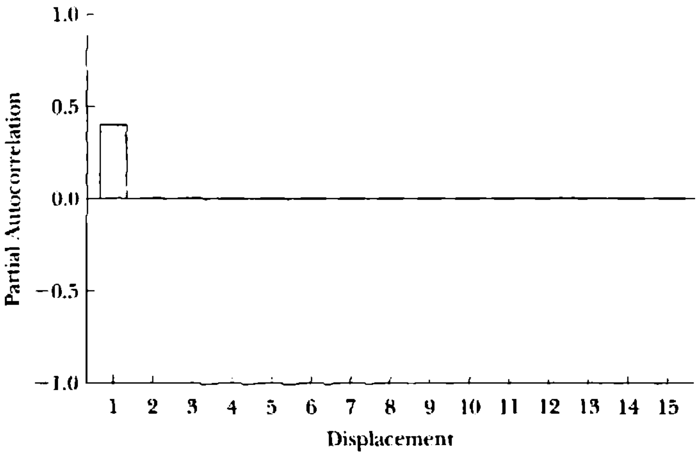


FIGURE 8.9
Population Partial Autocorrelation Function, AR(1) Process, $\varphi = 0.4$

Finally, the partial autocorrelation function for the AR(1) process cuts off abruptly; specifically,

$$p(\tau) = \begin{cases} \varphi, & \tau = 1 \\ 0, & \tau > 1 \end{cases}$$

It's easy to see why. The partial autocorrelations are just the last coefficients in a sequence of successively longer population autoregressions. If the true process is in fact an AR(1), the first partial autocorrelation is just the autoregressive coefficient, and coefficients on all longer lags are 0.

In Figures 8.9 and 8.10 we show the partial autocorrelation functions for our two AR(1) processes. At displacement 1, the partial autocorrelations are simply the parameters of the process (0.4 and 0.95, respectively), and at longer displacements, the partial autocorrelations are 0.

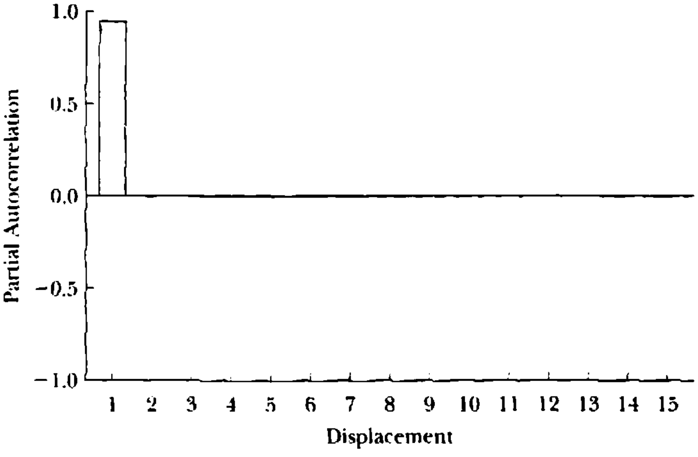


FIGURE 8.10
Population Partial Autocorrelation Function, AR(1) Process, $\varphi = 0.95$

THE AR(p) PROCESS

The general pth order autoregressive process, or AR(p) for short, is

$$y_t = \varphi_1 y_{t-1} + \varphi_2 y_{t-2} + \cdots + \varphi_p y_{t-p} + \varepsilon_t$$
$$\varepsilon_t \sim WN(0, \sigma^2)\,.$$

In lag operator form, we write

$$\Phi(L)y_t = (1 - \varphi_1 L - \varphi_2 L^2 - \cdots - \varphi_p L^p)y_t = \varepsilon_t\,.$$

As with our discussion of the MA(q) process, in our discussion of the **AR(p) process,** we dispense here with mathematical derivations and instead rely on parallels with the AR(1) case to establish intuition for its key properties.

An AR(p) process is covariance stationary if and only if the inverses of all roots of the autoregressive lag operator polynomial $\Phi(L)$ are inside the unit circle.[4] In the covariance stationary case, we can write the process in the convergent infinite moving average form

$$y_t = \frac{1}{\Phi(L)}\varepsilon_t\,.$$

The autocorrelation function for the general AR(p) process, as with that of the AR(1) process, decays gradually with displacement. Finally, the AR(p) partial autocorrelation function has a sharp cutoff at displacement p, for the same reason that the AR(1) partial autocorrelation function has a sharp cutoff at displacement 1.

Let's discuss the AR(p) autocorrelation function in a bit greater depth. The key insight is that, in spite of the fact that its qualitative behavior (gradual damping) matches that of the AR(1) autocorrelation function, it can nevertheless display a richer variety of patterns, depending on the order and parameters of the process. It can, for example, have damped monotonic decay, as in the AR(1) case with a positive coefficient, but it can also have damped oscillation in ways that AR(1) can't have. In the AR(1) case, the only possible oscillation occurs when the coefficient is negative, in which case the autocorrelations switch signs at each successively longer displacement. In higher-order autoregressive models, however, the autocorrelations can oscillate with much richer patterns reminiscent of cycles in the more traditional sense. This occurs when some roots of the autoregressive lag operator polynomial are complex.[5]

Consider, for example, the AR(2) process,

$$y_t = 1.5y_{t-1} - 0.9y_{t-2} + \varepsilon_t\,.$$

The corresponding lag operator polynomial is $1 - 1.5L + 0.9L^2$, with two complex conjugate roots, $0.83 \pm 0.65i$. The inverse roots are $0.75 \pm 0.58i$, both of

---

[4] A necessary **condition for covariance stationarity,** which is often useful as a quick check, is $\sum_{i=1}^{p} \varphi_i < 1$. If the condition is satisfied, the process may or may not be stationary; but if the condition is violated, the process can't be stationary.

[5] Note that complex roots can't occur in the AR(1) case.

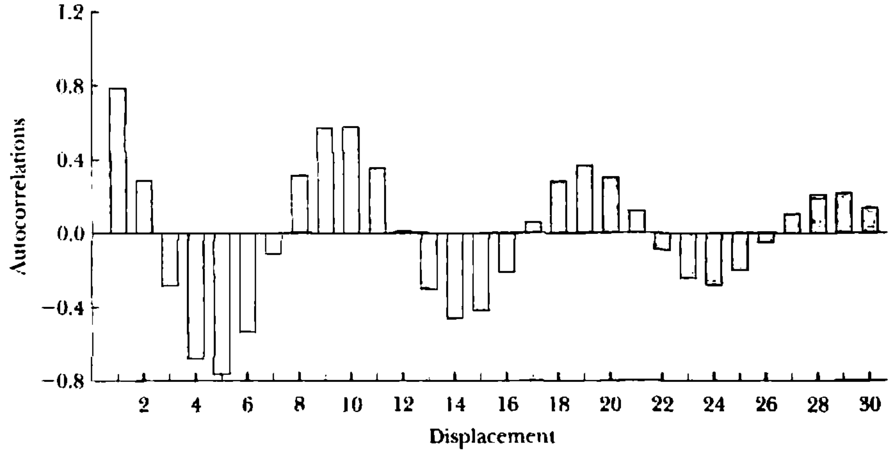


FIGURE 8.11 *Population Autocorrelation Function, AR(2) Process with Complex Roots*

which are close to, but inside, the unit circle; thus, the process is covariance stationary. It can be shown that the autocorrelation function for an AR(2) process is

$$\rho(0) = 1$$

$$\rho(1) = \frac{\varphi_1}{1 - \varphi_2}$$

$$\rho(\tau) = \varphi_1 \rho(\tau - 1) + \varphi_2 \rho(\tau - 2), \quad \tau = 2, 3, \ldots .$$

Using this formula, we can evaluate the autocorrelation function for the process at hand; we plot it in Figure 8.11. Because the roots are complex, the autocorrelation function oscillates, and because the roots are close to the unit circle, the oscillation damps slowly.

Finally, let's step back once again to consider in greater detail the precise way that finite-order autoregressive processes approximate the Wold representation. As always, the Wold representation is

$$y_t = B(L)\varepsilon_t ,$$

where $B(L)$ is of infinite order. The AR(1), as compared to the MA(1), is simply a different approximation to the Wold representation. The moving average representation associated with the AR(1) process is

$$y_t = \frac{1}{1 - \varphi L}\varepsilon_t .$$

Thus, when we fit an AR(1) model, we're using $\frac{1}{1-\varphi L}$, a rational polynomial with degenerate numerator polynomial (degree 0) and denominator polynomial of degree 1, to approximate $B(L)$. The moving average representation associated with the AR(1) process is of infinite order, as is the Wold representation, but it does not have infinitely many free coefficients. In fact, only one parameter, φ, underlies it.

The AR(p) is an obvious generalization of the AR(1) strategy for approximating the Wold representation. The moving average representation associated with the AR(p) process is

$$y_t = \frac{1}{\Phi(L)} \varepsilon_t .$$

When we fit an AR(p) model to approximate the Wold representation we're still using a rational polynomial with degenerate numerator polynomial (degree 0), but the denominator polynomial is of higher degree.

3. Autoregressive Moving Average (ARMA) Models

Autoregressive and moving average models are often combined in attempts to obtain better and more parsimonious approximations to the Wold representation, yielding the autoregressive moving average process, **ARMA(p, q) process** for short. As with moving average and autoregressive processes, ARMA processes also have direct motivation.[6] First, if the random shock that drives an autoregressive process is itself a moving average process, then it can be shown that we obtain an ARMA process. Second, ARMA processes can arise from aggregation. For example, sums of AR processes, or sums of AR and MA processes, can be shown to be ARMA processes. Finally, AR processes observed subject to measurement error also turn out to be ARMA processes.

The simplest ARMA process that's not a pure autoregression or pure moving average is the ARMA(1, 1), given by

$$y_t = \varphi y_{t-1} + \varepsilon_t + \theta \varepsilon_{t-1}$$

$$\varepsilon_t \sim WN(0, \sigma^2) ,$$

or, in lag operator form,

$$(1 - \varphi L) y_t = (1 + \theta L) \varepsilon_t ,$$

where $|\theta| < 1$ is required for stationarity and $|\theta| < 1$ is required for invertibility.[7] If the covariance stationarity condition is satisfied, then we have the moving average representation

$$y_t = \frac{(1 + \theta L)}{(1 - \varphi L)} \varepsilon_t ,$$

which is an infinite distributed lag of current and past innovations. Similarly, if the invertibility condition is satisfied, then we have the infinite autoregressive representation,

$$\frac{(1 - \varphi L)}{(1 + \theta L)} y_t = \varepsilon_t .$$

---

[6] For more extensive discussion, see Granger and Newbold (1986).

[7] Both stationarity and invertibility need to be checked in the ARMA case, because both autoregressive and moving average components are present.

The ARMA(p, q) process is a natural generalization of the ARMA(1, 1) that allows for multiple moving average and autoregressive lags. We write

$$y_t = \varphi_1 y_{t-1} + \cdots + \varphi_p y_{t-p} + \varepsilon_t + \theta_1 \varepsilon_{t-1} + \cdots + \theta_q \varepsilon_{t-q}$$
$$\varepsilon_t \sim WN(0, \sigma^2),$$

or

$$\Phi(L) y_t = \Theta(L) \varepsilon_t,$$

where

$$\Phi(L) = 1 - \varphi_1 L - \varphi_2 L^2 - \cdots - \varphi_p L^p$$

and

$$\Theta(L) = 1 + \theta_1 L + \theta_2 L^2 + \cdots + \theta_q L^q .$$

If the inverses of all roots of $\Phi(L)$ are inside the unit circle, then the process is covariance stationary and has convergent infinite moving average representation

$$y_t = \frac{\Theta(L)}{\Phi(L)} \varepsilon_t .$$

If the inverses of all roots of $\Theta(L)$ are inside the unit circle, then the process is invertible and has convergent infinite autoregressive representation

$$\frac{\Phi(L)}{\Theta(L)} y_t = \varepsilon_t .$$

As with autoregressions and moving averages, ARMA processes have a fixed unconditional mean but a time-varying conditional mean. In contrast to pure moving average or pure autoregressive processes, however, neither the autocorrelation nor partial autocorrelation functions of ARMA processes cut off at any particular displacement. Instead, each damps gradually, with the precise pattern depending on the process.

ARMA models approximate the Wold representation by a ratio of two finite-order lag operator polynomials, neither of which is degenerate. Thus, ARMA models use ratios of full-fledged polynomials in the lag operator to approximate the Wold representation,

$$y_t = \frac{\Theta(L)}{\Phi(L)} \varepsilon_t .$$

ARMA models, by allowing for both moving average and autoregressive components, often provide accurate approximations to the Wold representation that nevertheless have just a few parameters. That is, ARMA models are often both highly accurate and highly parsimonious. In a particular situation, for example, it might take an AR(5) to get the same approximation accuracy as could be obtained with an ARMA(2, 1), but the AR(5) has five parameters to be estimated, whereas the ARMA(2, 1) has only three.

4. Application: Specifying and Estimating Models for Employment Forecasting

In Chapter 7, we examined the correlogram for the Canadian employment series, and we saw that the sample autocorrelations damp slowly and the sample partial autocorrelations cut off, just the opposite of what's expected for a moving average. Thus, the correlogram indicates that a finite-order moving average process would not provide a good approximation to employment dynamics. Nevertheless, nothing stops us from fitting moving average models, so let's fit them and use the AIC and the SIC to guide model selection.

Moving average models are nonlinear in the parameters; thus, estimation proceeds by nonlinear least squares (numerical minimization). The idea is the same as when we encountered nonlinear least squares in our study of nonlinear trends—pick the parameters to minimize the sum of squared residuals—but finding an expression for the residual is a little bit trickier. To understand why moving average models are nonlinear in the parameters, and to get a feel for how they're estimated, consider an invertible MA(1) model, with a nonzero mean explicitly included for added realism,

$$y_t = \mu + \varepsilon_t + \theta\varepsilon_{t-1} .$$

Substitute backward m times to obtain the autoregressive approximation

$$y_t \approx \frac{\mu}{1+\theta} + \theta y_{t-1} - \theta^2 y_{t-2} + \cdots + (-1)^{m+1}\theta^m y_{t-m} + \varepsilon_t .$$

Thus, an invertible moving average can be approximated as a finite-order autoregression. The larger is m, the better the approximation. This lets us (approximately) express the residual in terms of observed data, after which we can use a computer to solve for the parameters that minimize the sum of squared residuals,

$$\hat{\mu}, \hat{\theta} = \underset{\mu,\theta}{\operatorname{argmin}} \sum_{t=1}^{T} \left(y_t - \left(\frac{\mu}{1+\theta} + \theta y_{t-1} - \theta^2 y_{t-2} + \cdots + (-1)^{m+1}\theta^m y_{t-m} \right) \right)^2$$

$$\hat{\sigma}^2 = \frac{1}{T} \sum_{t=1}^{T} \left(y_t - \left(\frac{\hat{\mu}}{1+\hat{\theta}} + \hat{\theta} y_{t-1} - \hat{\theta}^2 y_{t-2} + \cdots + (-1)^{m+1}\hat{\theta}^m y_{t-m} \right) \right)^2 .$$

The parameter estimates must be found using numerical optimization methods, because the parameters of the autoregressive approximation are restricted. The coefficient of the second lag of y is the square of the coefficient on the first lag of y, and so on. The parameter restrictions must be imposed in estimation, which is why we can't simply run an ordinary least-squares regression of y on lags of itself.

The next step would be to estimate MA(q) models, $q = 1, 2, 3, 4$. Both the AIC and the SIC suggest that the MA(4) is best. To save space, we report only

TABLE 8.1
Employment MA(4) Model

LS // Dependent variable is CANEMP.
Sample: 1962:1 1993:4
Included observations: 128
Convergence achieved after 49 iterations

Variable	**Coefficient**	**Std. Error**	***t*-Statistic**	**Prob.**
C	100.5438	0.843322	119.2234	0.0000
MA(1)	1.587641	0.063908	24.84246	0.0000
MA(2)	0.994369	0.089995	11.04917	0.0000
MA(3)	−0.020305	0.046550	−0.436189	0.6635
MA(4)	−0.298387	0.020489	−14.56311	0.0000

R^2	0.849951	Mean dependent var.	101.0176
Adjusted R^2	0.845071	SD dependent var.	7.499163
SE of regression	2.951747	Akaike info criterion	2.203073
Sum squared resid.	1071.676	Schwarz criterion	2.314481
Log likelihood	−317.6208	*F*-statistic	174.1826
Durbin-Watson stat.	1.246600	Prob(*F*-statistic)	0.000000

Inverted MA roots	.41	−.56 + .72i	−.56 − .72i	−.87

the results of MA(4) estimation in Table 8.1. The results of the MA(4) estimation, although better than lower-order MAs, are nevertheless poor. The R^2 of 0.84 is rather low, for example, and the Durbin-Watson statistic indicates that the MA(4) model fails to account for all the serial correlation in employment. The residual plot, which we show in Figure 8.12, clearly indicates a neglected cycle, an impression confirmed by the residual correlogram (Table 8.2 and Figure 8.13).

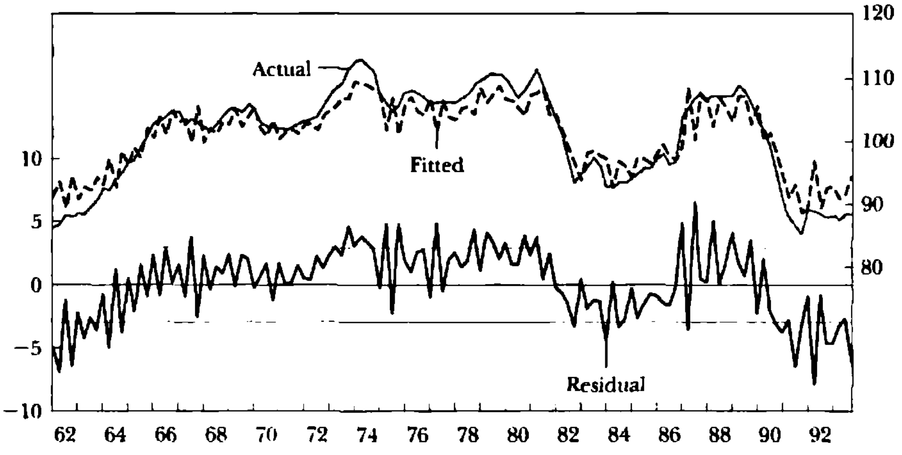


FIGURE 8.12
Employment MA(4) Model, Residual Plot

TABLE 8.2
Employment MA(4) Model, Residual Correlogram

Sample: 1962:1 1993:4
Included observations: 128
Q-statistic probabilities adjusted for 4 ARMA term(s)

	Acorr.	P. Acorr.	Std. Error	Ljung-Box	*p*-value
1	0.345	0.345	.088	15.614	
2	0.660	0.614	.088	73.089	
3	0.534	0.426	.088	111.01	
4	0.427	−0.042	.088	135.49	
5	0.347	−0.398	.088	151.79	0.000
6	0.484	0.145	.088	183.70	0.000
7	0.121	−0.118	.088	185.71	0.000
8	0.348	−0.048	.088	202.46	0.000
9	0.148	−0.019	.088	205.50	0.000
10	0.102	−0.066	.088	206.96	0.000
11	0.081	−0.098	.088	207.89	0.000
12	0.029	−0.113	.088	208.01	0.000

If we insist on using a moving average model, we'd want to explore orders greater than 4, but all the results thus far indicate that moving average processes don't provide good approximations to employment dynamics. Thus, let's consider alternative approximations, such as autoregressions. Autoregressions can be conveniently estimated by ordinary least-squares regression. Consider, for example, the AR(1) model,

$$(y_t - \mu) = \varphi(y_{t-1} - \mu) + \varepsilon_t$$

$$\varepsilon_t \sim WN(0, \sigma^2).$$

We can write it as

$$y_t = c + \varphi y_{t-1} + \varepsilon_t,$$

where $c = \mu(1 - \varphi)$. The least-squares estimators are

$$\hat{c}, \hat{\varphi} = \underset{c,\varphi}{\operatorname{argmin}} \sum_{t=1}^{T} (y_t - c - \varphi y_{t-1})^2$$

$$\hat{\sigma}^2 = \frac{1}{T} \sum_{t=1}^{T} (y_t - \hat{c} - \hat{\varphi} y_{t-1})^2.$$

The implied estimate of μ is $\hat{\mu} = \hat{c}/(1 - \hat{\varphi})$. Unlike the moving average case, for which the sum-of-squares function is nonlinear in the parameters,

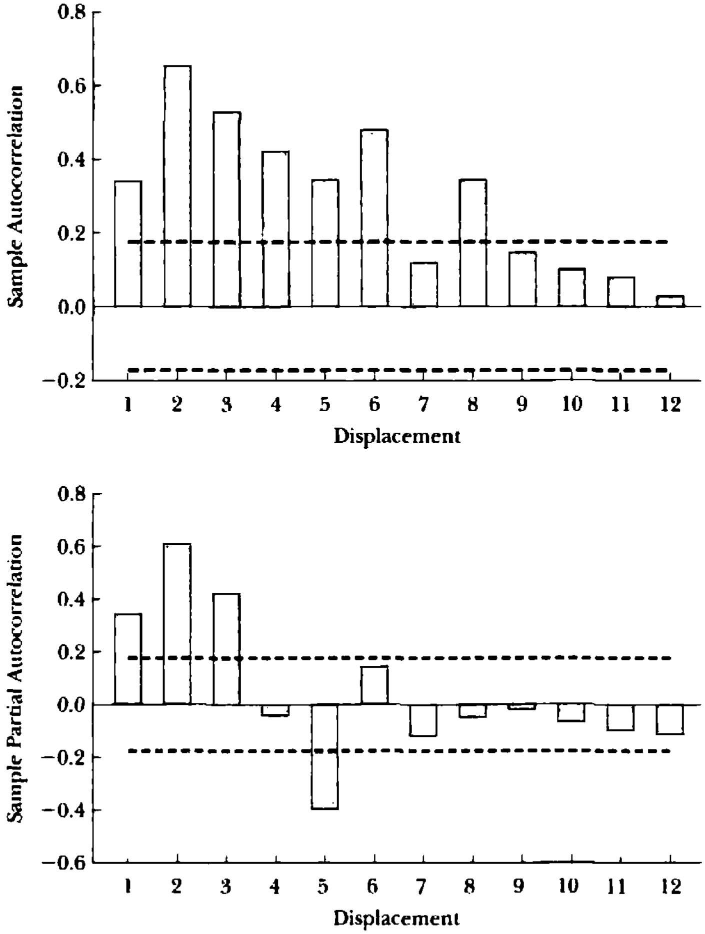


FIGURE 8.13 *Employment MA(4) Model: Residual Sample Autocorrelation and Partial Autocorrelation Functions, with Plus or Minus Two-Standard-Error Bands*

requiring the use of numerical minimization methods, the sum of squares function for autoregressive processes is linear in the parameters, so that estimation is particularly stable and easy. In the AR(1) case, we simply run an ordinary least-squares regression of y on one lag of y; in the AR(p) case, we regress y on p lags of y.

We estimate AR(p) models, $p = 1, 2, 3, 4$. Both the AIC and the SIC suggest that the AR(2) is best. To save space, we report only the results of AR(2) estimation in Table 8.3. The estimation results look good, and the residuals (Figure 8.14) look like white noise. The residual correlogram (Table 8.4 and Figure 8.15) supports that conclusion.

Finally, we consider ARMA(p, q) approximations to the Wold representation. ARMA models are estimated in a fashion similar to moving average

TABLE 8.3
Employment AR(2) Model

LS // Dependent variable is CANEMP.
Sample: 1962:1 1993:4
Included observations: 128
Convergence achieved after 3 iterations

Variable	**Coefficient**	**Std. Error**	***t*-Statistic**	**Prob.**
C	101.2413	3.399620	29.78017	0.0000
AR(1)	1.438810	0.078487	18.33188	0.0000
AR(2)	−0.476451	0.077902	−6.116042	0.0000

R^2	0.963372	Mean dependent var.	101.0176
Adjusted R^2	0.962786	SD dependent var.	7.499163
SE of regression	1.446663	Akaike info criterion	0.761677
Sum squared resid.	261.6041	Schwarz criterion	0.828522
Log likelihood	−227.3715	*F*-statistic	1643.837
Durbin-Watson stat.	2.067024	Prob(*F*-statistic)	0.000000
Inverted AR roots	.92	.52	

models; they have autoregressive approximations with nonlinear restrictions on the parameters, which we impose when doing a numerical sum of squares minimization. We examine all ARMA(p, q) models with p and q less than or equal to 4; the SIC and AIC values appear in Tables 8.5 and 8.6. The SIC selects the AR(2) (an ARMA(2, 0)), which we've already discussed. The AIC, which penalizes degrees of freedom less harshly, selects an ARMA(3, 1) model.

FIGURE 8.14
Employment AR(2) Model, Residual Plot

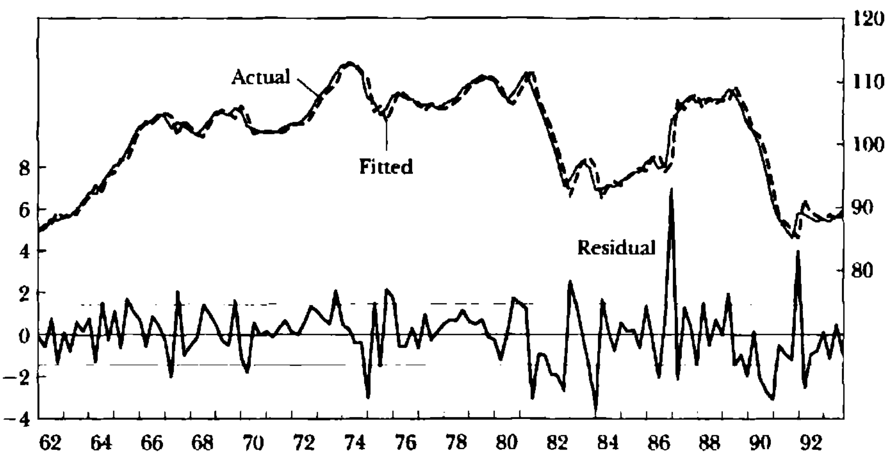


TABLE 8.4
Employment AR(2) Model, Residual Correlogram

Sample: 1962:1 1993:4
Included observations: 128
Q-statistic probabilities adjusted for 2 ARMA term(s)

	Acorr.	P. Acorr.	Std. Error	Ljung-Box	p-value
1	−0.035	−0.035	.088	0.1606	
2	0.044	0.042	.088	0.4115	
3	0.011	0.014	.088	0.4291	0.512
4	0.051	0.050	.088	0.7786	0.678
5	0.002	0.004	.088	0.7790	0.854
6	0.019	0.015	.088	0.8272	0.935
7	−0.024	−0.024	.088	0.9036	0.970
8	0.078	0.072	.088	1.7382	0.942
9	0.080	0.087	.088	2.6236	0.918
10	0.050	0.050	.088	2.9727	0.936
11	−0.023	−0.027	.088	3.0504	0.962
12	−0.129	−0.148	.088	5.4385	0.860

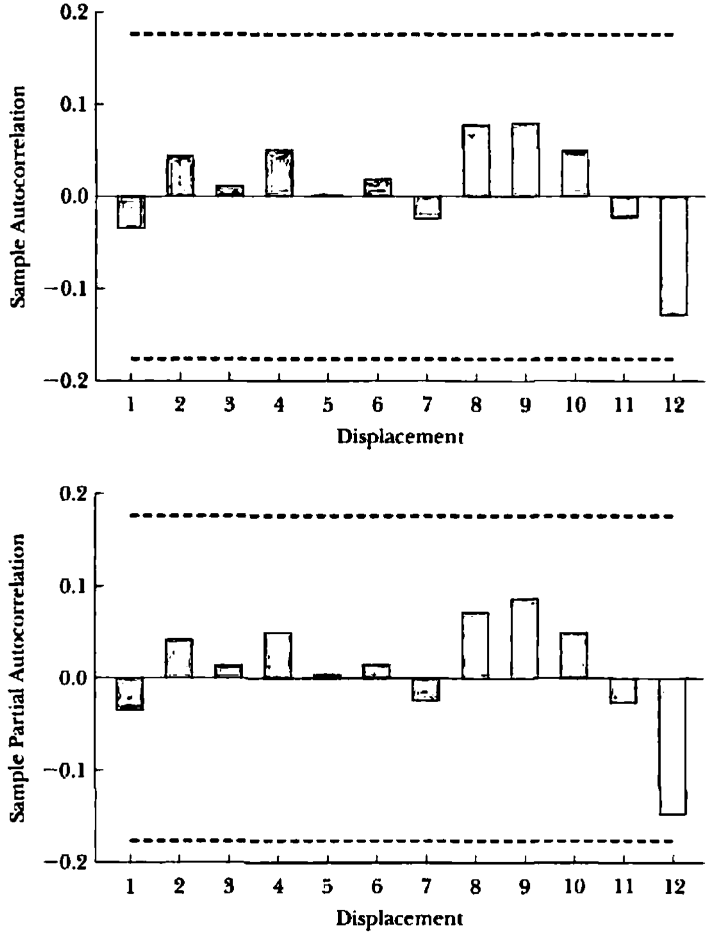
TABLE 8.5
Employment AIC Values, Various ARMA Models

		MA Order				
		0	1	2	3	4
	0		2.86	2.32	2.47	2.20
	1	1.01	0.83	0.79	0.80	0.81
AR Order	2	0.762	0.77	0.78	0.80	0.80
	3	0.77	0.761	0.77	0.78	0.79
	4	0.79	0.79	0.77	0.79	0.80

TABLE 8.6
Employment SIC Values, Various ARMA Models

		MA Order				
		0	1	2	3	4
	0		2.91	2.38	2.56	2.31
	1	1.05	0.90	0.88	0.91	0.94
AR Order	2	0.83	0.86	0.89	0.92	0.96
	3	0.86	0.87	0.90	0.94	0.96
	4	0.90	0.92	0.93	0.97	1.00

FIGURE 8.15 *Employment AR(2) Model: Residual Sample Autocorrelation and Partial Autocorrelation Functions, with Plus or Minus Two-Standard-Error Bands*



The ARMA(3, 1) model looks good: the estimation results appear in Table 8.7, the residual plot in Figure 8.16, and the residual correlogram in Table 8.8 and Figure 8.17.

Although the ARMA(3, 1) looks good, apart from its lower AIC, it looks no better than the AR(2), which basically seemed perfect. In fact, there are at least three reasons to prefer the AR(2). First, for the reasons discussed in Chapter 5, when the AIC and the SIC disagree, we recommend using the more parsimonious model selected by the SIC. Second, if we consider a model selection strategy involving examination of not just the AIC and SIC but also autocorrelations and partial autocorrelations, which we advocate, we're led to the AR(2). Finally, and importantly, the impression that the ARMA(3, 1) provides a richer approximation to employment dynamics is likely spurious in this case. The

TABLE 8.7
Employment ARMA(3, 1) Model

LS // Dependent variable is CANEMP.
Sample: 1962:1 1993:4
Included observations: 128
Convergence achieved after 17 iterations

Variable	Coefficient	Std. Error	*t*-Statistic	Prob.
C	101.1378	3.538602	28.58130	0.0000
AR(1)	0.500493	0.087503	5.719732	0.0000
AR(2)	0.872194	0.067096	12.99917	0.0000
AR(3)	−0.443355	0.080970	−5.475560	0.0000
MA(1)	0.970952	0.035015	27.72924	0.0000

R^2	0.964535	Mean dependent var.	101.0176
Adjusted R^2	0.963381	SD dependent var.	7.499163
SE of regression	1.435043	Akaike info criterion	0.760668
Sum squared resid.	253.2997	Schwarz criterion	0.872076
Log likelihood	−225.3069	*F*-statistic	836.2912
Durbin-Watson stat.	2.057302	Prob(*F*-statistic)	0.000000
Inverted AR roots	.93	.51	−.94
Inverted MA roots	−.97		

ARMA(3, 1) has a inverse autoregressive root of −0.94 and an inverse moving average root of −0.97. Those roots are of course just *estimates*, subject to sampling uncertainty, and are likely to be statistically indistinguishable from one another, in which case we can *cancel* them, which brings us down to an ARMA(2, 0), or AR(2), model with roots virtually indistinguishable from those

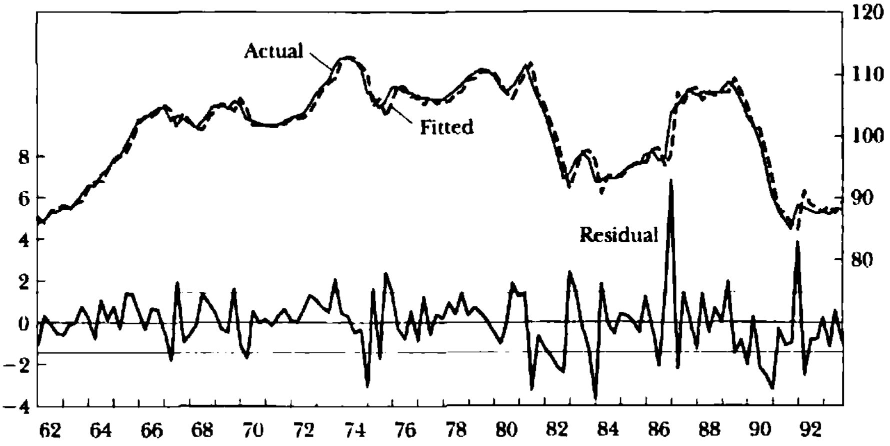


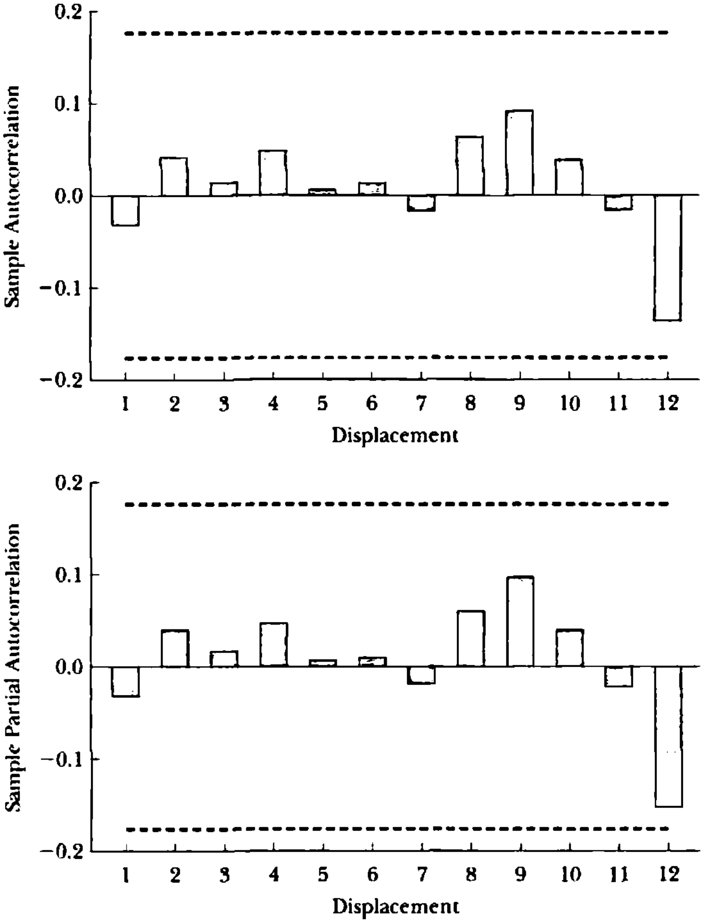
FIGURE 8.16
Employment ARMA(3, 1) Model, Residual Plot

TABLE 8.8
Employment ARMA(3, 1) Model, Residual Correlogram

Sample: 1962:1 1993:4
Included observations: 128
Q-statistic probabilities adjusted for four ARMA term(s)

	Acorr.	P. Acorr.	Std. Error	Ljung-Box	*p*-value
1	−0.032	−0.032	.09	0.1376	
2	0.041	0.040	.09	0.3643	
3	0.014	0.017	.09	0.3904	
4	0.048	0.047	.09	0.6970	
5	0.006	0.007	.09	0.7013	0.402
6	0.013	0.009	.09	0.7246	0.696
7	−0.017	−0.019	.09	0.7650	0.858
8	0.064	0.060	.09	1.3384	0.855
9	0.092	0.097	.09	2.5182	0.774
10	0.039	0.040	.09	2.7276	0.842
11	−0.016	−0.022	.09	2.7659	0.906
12	−0.137	−0.153	.09	5.4415	0.710

FIGURE 8.17
Employment ARMA(3, 1) Model: Residual Sample Autocorrelation and Partial Autocorrelation Functions, with Plus or Minus Two-Standard-Error Bands



of our earlier-estimated AR(2) process! We refer to this situation as one of **common factors** in an ARMA model. Be on the lookout for such situations, which arise frequently and can lead to substantial model simplification.

Thus, we arrive at an AR(2) model for employment. In the next chapter, we'll learn how to use it to produce point and interval forecasts.

Exercises, Problems, and Complements

1. (ARMA lag inclusion) Review Table 8.1. Why is the MA(3) term included even though the *p*-value indicates that it is not significant? What would be the costs and benefits of dropping the insignificant MA(3) term?

2. (Shapes of correlograms) Given the following ARMA processes, sketch the expected forms of the autocorrelation and partial autocorrelation functions. (*Hint:* Examine the roots of the various autoregressive and moving average lag operator polynomials.)
   a. $y_t = \left(\frac{1}{1 - 1.05L - 0.09L^2}\right)\varepsilon_t$
   b. $y_t = (1 - 0.4L)\varepsilon_t$
   c. $y_t = \left(\frac{1}{1 - 0.7L}\right)\varepsilon_t.$

3. (The autocovariance function of the MA(1) process, revisited) In the text, we wrote

$$\gamma(\tau) = E(y_t\, y_{t-\tau}) = E((\varepsilon_t + \theta\varepsilon_{t-1})(\varepsilon_{t-\tau} + \theta\varepsilon_{t-\tau-1})) = \begin{cases} \theta\sigma^2, & \tau = 1 \\ 0, & \text{otherwise}. \end{cases}$$

   Fill in the missing steps by evaluating explicitly the expectation $E((\varepsilon_t + \theta\varepsilon_{t-1})(\varepsilon_{t-\tau} + \theta\varepsilon_{t-\tau-1}))$.

4. (ARMA algebra) Derive expressions for the autocovariance function, autocorrelation function, conditional mean, unconditional mean, conditional variance, and unconditional variance of the following processes:
   a. $y_t = \mu + \varepsilon_t + \theta_1\varepsilon_{t-1} + \theta_2\varepsilon_{t-2}$
   b. $y_t = \varphi y_{t-1} + \varepsilon_t + \theta\varepsilon_{t-1}$

5. (Diagnostic checking of model residuals) If a forecasting model has extracted all the systematic information from the data, then what's left—the residual—should be white noise. More precisely, the true innovations are white noise, and if a model is a good approximation to the Wold representation, then its 1-step-ahead forecast errors should be approximately white noise. The model residuals are the in-sample analog of out-of-sample 1-step-ahead forecast errors—hence the usefulness of various tests of the hypothesis that residuals are white noise.

   The Durbin-Watson test is the most popular. Recall the Durbin-Watson test statistic, discussed in Chapter 2,

$$\mathrm{DW} = \frac{\sum_{t=2}^{T}(e_t - e_{t-1})^2}{\sum_{t=1}^{T} e_t^2} .$$

Note that

$$\sum_{t=2}^{T}(e_t - e_{t-1})^2 \approx 2\sum_{t=2}^{T} e_t^2 - 2\sum_{t=2}^{T} e_t e_{t-1} .$$

Thus,

$$\mathrm{DW} \approx 2(1 - \hat{\rho}(1)) ,$$

so that the Durbin-Watson test is effectively based only on the first sample autocorrelation and really only tests whether the first autocorrelation is 0. We say therefore that the Durbin-Watson is a test for **first-order serial correlation.** In addition, the Durbin-Watson test is not valid in the presence of lagged dependent variables.[8] On both counts, we'd like a more general and flexible framework for diagnosing serial correlation. The residual correlogram, comprised of the residual sample autocorrelations, the sample partial autocorrelations, and the associated Q-statistics, delivers the goods.

a. When we discussed the correlogram in the text, we focused on the case of an observed time series, in which case we showed that the Q-statistics are distributed as χ^2_m. Now, however, we want to assess whether unobserved model disturbances are white noise. To do so, we use the model residuals, which are estimates of the unobserved disturbances. Because we fit a model to get the residuals, we need to account for the degrees of freedom used. The upshot is that the distribution of the Q-statistics under the white noise hypothesis is better approximated by a χ^2_{m-k} random variable, where k is the number of parameters estimated. That's why, for example, we don't report (and in fact the software doesn't compute) the p-values for the Q-statistics associated with the residual correlogram of our employment forecasting model until $m > k$.

b. **Durbin's *h*-test** is an alternative to the Durbin-Watson test. As with the Durbin-Watson test, it's designed to detect first-order serial correlation, but it's valid in the presence of lagged dependent variables. Do some background reading on Durbin's *h*-test, and report what you learned.

c. The **Breusch-Godfrey test** is another alternative to the Durbin-Watson test. It's designed to detect *p*th-order serial correlation, where p is selected by the user, and is also valid in the presence of lagged dependent variables. Do some background reading on the Breusch-Godfrey procedure, and report what you learned.

d. Which do you think is likely to be most useful to you in assessing the properties of residuals from forecasting models: the residual correlogram, Durbin's *h*-test, or the Breusch-Godfrey test? Why?

---

[8] Following standard, if not strictly appropriate, practice, in this book we often report and examine the Durbin-Watson statistic even when lagged dependent variables are included. We always supplement the Durbin-Watson statistic, however, with other diagnostics such as the residual correlogram, which remain valid in the presence of lagged dependent variables and which almost always produce the same inference as the Durbin-Watson statistic.

6. (Mechanics of fitting ARMA models) The book's web page presents data for daily transfers over BankWire, a financial wire transfer system in a country responsible for much of the world's finance, over a recent span of 200 business days.
   a. Is trend or seasonality operative? Defend your answer.
   b. Using the methods developed in Chapters 7 and 8, find a parsimonious ARMA(p, q) model that fits well, and defend its adequacy.

7. (Modeling cyclical dynamics) As a research analyst at the U.S. Department of Energy, you have been asked to model nonseasonally adjusted U.S. imports of crude oil.
   a. Find a suitable time series on the web.
   b. Create a model that captures the trend in the series.
   c. Adding to the model from part *b*, create a model with trend and a full set of seasonal dummy variables.
   d. Observe the residuals of the model from part *b* and their correlogram. Is there evidence of neglected dynamics? If so, what to do?

8. (**Aggregation** and **disaggregation: top-down forecasting model** vs. **bottom-up forecasting model**) Related to the issue of methods and complexity discussed in Chapter 3 is the question of aggregation. Often we want to forecast an aggregate, such as total sales of a manufacturing firm, but we can take either an aggregated or disaggregated approach.

   Suppose, for example, that total sales is composed of sales of three products. The aggregated, or top-down or macro, approach is simply to model and forecast total sales. The disaggregated, or bottom-up or micro, approach is to model and forecast separately the sales of the individual products and then to add them together.

   Perhaps surprisingly, it's impossible to know in advance whether the aggregated or disaggregated approach is better. It all depends on the specifics of the situation; the only way to tell is to try both approaches and compare the forecasting results.

   However, in real-world situations characterized by likely model misspecification and parameter estimation uncertainty, there are reasons to suspect that the aggregated approach may be preferable. First, standard (e.g., linear) models fit to aggregated series may be less prone to specification error, because aggregation can produce approximately linear relationships even when the underlying disaggregated relationships are not linear. Second, if the disaggregated series depends in part on a common factor (e.g., general business conditions), then it will emerge more clearly in the aggregate data. Finally, modeling and forecasting of one aggregated series, as opposed to many disaggregated series, rely on far fewer parameter estimates.

   Of course, if our interest centers on the disaggregated components, then we have no choice but to take a disaggregated approach.

   It is possible that an aggregate forecast may be useful in forecasting disaggregated series. Why? (*Hint:* See Fildes and Stekler, 2000.)

9. (Nonlinear forecasting models: regime switching) In this chapter, we've studied dynamic **linear models**, which are tremendously important in practice. They're called *linear* because y_t is a simple linear function of past y's or past ε's. In some forecasting situations, however, good statistical characterization of dynamics may require some notion of regime switching, as between "good" and "bad" states, which is a type of **nonlinear model**.

Models incorporating **regime switching** have a long tradition in business cycle analysis, in which expansion is the good state, and contraction (recession) is the bad state. This idea is also manifest in the great interest in the popular press; for example, in identifying and forecasting turning points in economic activity. It is only within a regime-switching framework that the concept of a turning point has intrinsic meaning; turning points are naturally and immediately defined as the times separating expansions and contractions.

Threshold models are squarely in line with the regime-switching tradition. The following threshold model, for example, has three regimes, two thresholds, and a *d*-period delay regulating the switches:

$$y_t = \begin{cases} c^{(u)} + \varphi^{(u)} y_{t-1} + \varepsilon_t^{(u)}, & \theta^{(u)} < y_{t-d} \\ c^{(m)} + \varphi^{(m)} y_{t-1} + \varepsilon_t^{(m)}, & \theta^{(l)} < y_{t-d} < \theta^{(u)} \\ c^{(l)} + \varphi^{(l)} y_{t-1} + \varepsilon_t^{(l)}, & \theta^{(l)} > y_{t-d} . \end{cases}$$

The superscripts indicate "upper," "middle," and "lower" regimes, and the regime operative at any time t depends on the observable past history of y—in particular, on the value of y_{t-d}.

Although observable threshold models are of interest, models with *latent* (or unobservable) states as opposed to observed states may be more appropriate in many business, economic, and financial contexts. In such a setup, time series dynamics are governed by a finite-dimensional parameter vector that switches (potentially each period) depending on which of two unobservable states is realized, with state transitions governed by a first-order Markov process (meaning that the state at any time t depends only on the state at time $t-1$, not at time $t-2$, $t-3$, etc.).

To make matters concrete, let's take a simple example. Let $\{s_t\}_{t=1}^{T}$ be the (latent) sample path of a two-state first-order autoregressive process, taking just the two values 0 or 1, with the transition probability matrix given by

$$M = \begin{pmatrix} p_{00} & 1 - p_{00} \\ 1 - p_{11} & p_{11} \end{pmatrix} .$$

The ijth element of M gives the probability of moving from state i (at time $t-1$) to state j (at time t). Note that there are only two free parameters, the staying probabilities, p_{00} and p_{11}. Let $\{y_t\}_{t=1}^{T}$ be the sample path of an observed time series that depends on $\{s_t\}_{t=1}^{T}$ such that the density of y_t conditional on s_t is

$$f(y_t \mid s_t; \theta) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left(\frac{-(y_t - \mu_{s_t})^2}{2\sigma^2} \right) .$$

Thus, y_t is Gaussian white noise with a potentially switching mean. The two means around which y_t moves are of particular interest and may, for example, correspond to episodes of differing growth rates ("booms" and "recessions," "bull" and "bear" markets, etc.).

10. (Difficulties with nonlinear optimization) Nonlinear optimization is a tricky business, fraught with problems. Some eye-opening reading includes Newbold, Agiakloglou, and Miller (1994) and McCullough and Vinod (1999).

Some problems are generic. It's relatively easy to find a local optimum, for example, but much harder to be confident that the local optimum is global. Simple checks such as trying a variety of startup values and checking the optimum to which convergence occurs are used routinely, but the problem nevertheless remains. Other problems may be software specific. For example, some software may use highly accurate analytic derivatives, whereas other software uses approximate numerical derivatives. Even the same software package may change algorithms or details of implementation across versions, leading to different results. Software for ARMA model estimation is unavoidably exposed to all such problems, because estimation of any model involving MA terms requires numerical optimization of a likelihood or sum-of-squares function.

Bibliographical and Computational Notes

Characterization of time series by means of autoregressive, moving average, or ARMA models was suggested, more or less simultaneously, by the Russian statistician and economist E. Slutsky and the British statistician G. U. Yule. Slutsky (1927) remains a classic. The Slutsky-Yule framework was modernized, extended, and made part of an innovative and operational modeling and forecasting paradigm in a more recent classic, a 1970 book by Box and Jenkins, the latest edition of which is Box, Jenkins, and Reinsel (1994). In fact, ARMA and related models are often called **Box-Jenkins models.**

Granger and Newbold (1986) contains more detailed discussion of a number of topics that arose in this chapter, including the idea of moving average processes as describing economic equilibrium disturbed by transient shocks, the Yule-Walker equation, and the insight that aggregation and measurement error lead naturally to ARMA processes.

The sample autocorrelations and partial autocorrelations, together with related diagnostics, provide graphical aids to model selection that complement the Akaike and Schwarz information criteria introduced earlier. Not long ago, the sample autocorrelation and partial autocorrelation functions were often used *alone* to guide forecast model selection, a tricky business that was more art than science. Use of the Akaike and Schwarz criteria results in more systematic and replicable model selection, but the sample autocorrelation and partial autocorrelation functions nevertheless remain important as basic graphical summaries of dynamics in time series data. The two approaches are complements, not substitutes.

Our discussion of estimation was a bit fragmented; we discussed estimation of moving average and ARMA models using nonlinear least squares, whereas we discussed estimation of autoregressive models using ordinary least squares. A more unified approach proceeds by writing each model as a regression on an intercept, with a serially correlated disturbance. Thus, the moving average model is

$$
\begin{aligned}
y_t &= \mu + \varepsilon_t \\
\varepsilon_t &= \Theta(L)v_t \\
v_t &\sim WN(0, \sigma^2)\ ,
\end{aligned}
$$

the autoregressive model is

$$\begin{aligned} y_t &= \mu + \varepsilon_t \\ \Phi(L)\varepsilon_t &= v_t \\ v_t &\sim WN(0, \sigma^2), \end{aligned}$$

and the ARMA model is

$$\begin{aligned} y_t &= \mu + \varepsilon_t \\ \Phi(L)\varepsilon_t &= \Theta(L)v_t \\ v_t &\sim WN(0, \sigma^2). \end{aligned}$$

We can estimate each model in identical fashion using nonlinear least squares. Eviews and other forecasting packages proceed in precisely that way.[9]

This framework—regression on a constant with serially correlated disturbances—has a number of attractive features. First, the mean of the process is the regression constant term.[10] Second, it leads us naturally toward regression on more than just a constant, as other right-hand-side variables can be added as desired. Finally, it exploits the fact that because autoregressive and moving average models are special cases of the ARMA model, their estimation is also a special case of estimation of the ARMA model.

Our description of estimating ARMA models—compute the autoregressive representation, truncate it, and estimate the resulting approximate model by nonlinear least squares—is conceptually correct but intentionally simplified. The actual estimation methods implemented in modern software are more sophisticated, and the precise implementations vary across software packages. Beneath it all, however, all estimation methods are closely related to our discussion, whether implicitly or explicitly. You should consult your software manual for details. (Hopefully they're provided!)

Pesaran, Pierse, and Kumar (1989) and Granger (1990) study the question of top-down versus bottom-up forecasting. For a comparative analysis in the context of forecasting Euro-area macroeconomic activity, see Stock and Watson (2003).

Our discussion of regime-switching models draws heavily on Diebold and Rudebusch (1996). Tong (1983) is a key reference on observable-state threshold models, as is Hamilton (1989) for latent-state threshold models. There are a number of extensions of those basic regime-switching models of potential interest for forecasters, such as allowing for smooth as opposed to abrupt transitions in threshold models with observed states (Granger and Teräsvirta, 1993) and allowing for time-varying transition probabilities in threshold models with latent states (Diebold, Lee, and Weinbach, 1994).

Concepts for Review

Moving Average (MA) model
Autoregressive (AR) model
Autoregressive moving average (ARMA) model
Stochastic process
MA(1) process
Cutoff in the autocorrelation function
Invertibility

---

[9] That's why, for example, information on the number of iterations required for convergence is presented even for estimation of the autoregressive model.

[10] Hence the notation "μ" for the intercept.

Autoregressive representation
MA(q) process
Complex roots
Condition for invertibility of the MA(q)
Yule-Walker equation
AR(p) process
Condition for covariance stationarity
ARMA(p, q) process
Common factors
First-order serial correlation
Durbin's h-test
Breusch-Godfrey test
Aggregation
Disaggregation
Top-down forecasting model
Bottom-up forecasting model
Linear model
Nonlinear model
Regime switching
Threshold model
Box-Jenkins model

References and Additional Readings

Bollerslev, T. (1986). "Generalized Autoregressive Conditional Heteroskedasticity." *Journal of Econometrics*, 31, 307–327.

Bollerslev, T., Chou, R. Y., and Kroner, K. F. (1992). "ARCH Modeling in Finance: A Selective Review of the Theory and Empirical Evidence." *Journal of Econometrics*, 52, 5–59.

Box, G. E. P., Jenkins, G. W., and Reinsel, G. (1994). *Time Series Analysis, Forecasting and Control.* 3rd ed. Englewood Cliffs, N.J.: Prentice Hall.

Burns, A. F., and Mitchell, W. C. (1946). *Measuring Business Cycles.* New York: National Bureau of Economic Research.

Diebold, F. X., Lee, J.-H., and Weinbach, G. (1994). "Regime Switching with Time-Varying Transition Probabilities." In C. Hargreaves (ed.), *Nonstationary Time Series Analysis and Cointegration.* Oxford: Oxford University Press, 283–302. Reprinted in Diebold and Rudebusch (1999).

Diebold, F. X., and Lopez, J. (1995). "Modeling Volatility Dynamics." In Kevin Hoover (ed.), *Macroeconometrics: Developments, Tensions and Prospects.* Boston: Kluwer Academic Press, 427–472.

Diebold, F. X., and Rudebusch, G. D. (1996). "Measuring Business Cycles: A Modern Perspective." *Review of Economics and Statistics*, 78, 67–77. Reprinted in Diebold and Rudebusch (1999).

Diebold, F. X., and Rudebusch, G. D. (1999). *Business Cycles: Durations, Dynamics, and Forecasting.* Princeton, N.J.: Princeton University Press.

Engle, R. F. (1982). "Autoregressive Conditional Heteroskedasticity with Estimates of the Variance of U.K. Inflation." *Econometrica*, 50, 987–1008.

Fildes, R., and Steckler, H. (2000). "The State of Macroeconomic Forecasting." Manuscript.

Granger, C. W. J. (1990). "Aggregation of Time Series Variables: A Survey." In T. Barker and M. H. Pesaran (eds.), *Disaggregation in Econometric Modelling.* London: Routledge.

Granger, C. W. J., and Newbold, P. (1986). *Forecasting Economic Time Series.* 2nd ed. Orlando, Fla.: Academic Press.

Granger, C. W. J., and Teräsvirta, Y. (1993). *Modelling Nonlinear Economic Relationships.* Oxford: Oxford University Press.

Hamilton, J. D. (1989). "A New Approach to the Economic Analysis of Nonstationary Time Series and the Business Cycle." *Econometrica*, 57, 357–384.

McCullough, B. D., and Vinod, H. D. (1999). "The Numerical Reliability of Econometric Software." *Journal of Economic Literature*, 37, 633–665.

Newbold, P., Agiakloglou, C., and Miller, J. P. (1994). "Adventures with ARIMA Software." *International Journal of Forecasting*, 10, 573–581.

Pesaran, M. H., Pierse, R. G., and Kumar, M. S. (1989). "Econometric Analysis of Aggregation in the Context of Linear Prediction Models." *Econometrica*, 57, 861–888.

Slutsky, E. (1927). "The Summation of Random Causes as the Source of Cyclic Processes." *Econometrica*, 5, 105–146.

Stock, J. H., and Watson, M. W. (2003). "Macroeconomic Forecasting in the Euro Area: Country-Specific versus Area-Wide Information." *European Economic Review*, 47, 1–18.

Taylor, S. (1996). *Modeling Financial Time Series*. 2nd ed. New York: Wiley.

Tong, H. (1990). *Non-linear Time Series*. Oxford: Clarendon Press.

Forecasting Cycles

1. Optimal Forecasts

By now you've gotten comfortable with the idea of an **information set.** Here we'll use that idea extensively. We denote the time T information set by Ω_T. At first pass it seems most natural to think of the information set as containing the available past history of the series,

$$\Omega_T = \{y_T, y_{T-1}, y_{T-2}, \ldots\},$$

where, for theoretical purposes, we imagine history as having begun in the infinite past.

So long as y is covariance stationary, however, we can just as easily express the information available at time T in terms of current and past shocks,

$$\Omega_T = \{\varepsilon_T, \varepsilon_{T-1}, \varepsilon_{T-2}, \ldots\}.$$

Suppose, for example, that the process to be forecast is a covariance stationary AR(1),

$$y_t = \varphi y_{t-1} + \varepsilon_t .$$

Then immediately,

$$\begin{aligned} \varepsilon_T &= y_T - \varphi y_{T-1} \\ \varepsilon_{T-1} &= y_{T-1} - \varphi y_{T-2} \\ \varepsilon_{T-2} &= y_{T-2} - \varphi y_{T-3}, \end{aligned}$$

and so on. In other words, we can figure out the current and lagged ε's from the current and lagged y's. More generally, for any covariance stationary and invertible series, we can infer the history of ε from the history of y and the history of y from the history of ε.

Assembling the discussion thus far, we can view the time-T information set as containing the current and past values of y and ε,

$$\Omega_T = \{y_T, y_{T-1}, y_{T-2}, \ldots, \varepsilon_T, \varepsilon_{T-1}, \varepsilon_{T-2}, \ldots\} .$$

Based on that information set, we want to find the **optimal forecast** of y at some future time $T + h$. The optimal forecast is the one with the smallest loss on average—that is, the forecast that minimizes **expected loss.** It turns out that under reasonably weak conditions, the optimal forecast is the **conditional mean,** $E(y_{T+h} \mid \Omega_T)$, the expected value of the future value of the series being forecast, conditional on available information.

In general, the conditional mean need not be a linear function of the elements of the information set. Because linear functions are particularly tractable, we prefer to work with **linear forecasts**—forecasts that are linear in the elements of the information set—by finding the best linear approximation to the conditional mean, called the **linear projection,** denoted $P(y_{T+h} \mid \Omega_T)$. This explains the common term "**linear least-squares forecast.**" The linear projection is often very useful and accurate, because the conditional mean is often close to linear. In fact, in the Gaussian case, the conditional expectation is exactly linear, so that $E(y_{T+h} \mid \Omega_T) = P(y_{T+h} \mid \Omega_T)$.

2. Forecasting Moving Average Processes

OPTIMAL POINT FORECASTS FOR FINITE-ORDER MOVING AVERAGES

Our forecasting method is always the same: We write out the process for the future time period of interest, $T + h$, and project it on what's known at time T, when the forecast is made. This process is best learned by example. Consider an MA(2) process,

$$y_t = \varepsilon_t + \theta_1 \varepsilon_{t-1} + \theta_2 \varepsilon_{t-2}$$
$$\varepsilon_t \sim WN(0, \sigma^2) .$$

Suppose we're standing at time T, and we want to forecast y_{T+1}. First we write out the process for $T + 1$,

$$y_{T+1} = \varepsilon_{T+1} + \theta_1 \varepsilon_T + \theta_2 \varepsilon_{T-1} .$$

Then we project on the time-T information set, which simply means that all future innovations are replaced by 0. Thus,

$$y_{T+1,T} = P(y_{T+1} \mid \Omega_T) = \theta_1 \varepsilon_T + \theta_2 \varepsilon_{T-1} .$$

To forecast two steps ahead, we note that

$$y_{T+2} = \varepsilon_{T+2} + \theta_1 \varepsilon_{T+1} + \theta_2 \varepsilon_T ,$$

and we project on the time-T information set to get

$$y_{T+2,T} = \theta_2 \varepsilon_T .$$

Continuing in this fashion, we see that

$$y_{T+h,T} = 0 ,$$

for all $h > 2$.

Now let's compute the corresponding **forecast errors.**[1] We have

$$\begin{aligned} e_{T+1,T} &= \varepsilon_{T+1} && \text{(white noise)} \\ e_{T+2,T} &= \varepsilon_{T+2} + \theta_1 \varepsilon_{T+1} && \text{(MA(1))} \\ &\cdots \\ e_{T+h,T} &= \varepsilon_{T+h} + \theta_1 \varepsilon_{T+h-1} + \theta_2 \varepsilon_{T+h-2} && \text{(MA(2))} , \end{aligned}$$

for all $h > 2$.

Finally, the forecast error variances are

$$\begin{aligned} \sigma_1^2 &= \sigma^2 \\ \sigma_2^2 &= \sigma^2(1 + \theta_1^2) \\ &\cdots \\ \sigma_h^2 &= \sigma^2(1 + \theta_1^2 + \theta_2^2) , \end{aligned}$$

for all $h > 2$. Moreover, the forecast error variance for $h > 2$ is just the unconditional variance of y_t.

Now consider the general MA(q) case. The model is

$$y_t = \varepsilon_t + \theta_1 \varepsilon_{t-1} + \cdots \theta_q \varepsilon_{t-q} .$$

First, consider the forecasts. If $h \le q$, then the forecast has the form

$$y_{T+h,T} = 0 \ + \ \text{“adjustment,”}$$

whereas if $h > q$, then the forecast is

$$y_{T+h,T} = 0 .$$

Thus, an MA(q) process is not forecastable (apart from the unconditional mean) more than q steps ahead. All the dynamics in the MA(q) process, which we exploit for forecasting, "wash out" by the time we get to horizon q, which reflects the autocorrelation structure of the MA(q) process. (Recall that, as we showed in Chapter 8, it cuts off at displacement q.)

Second, consider the corresponding forecast errors. They are

$$e_{T+h,T} = \text{MA}(h-1)$$

---

[1] Recall that the forecast error is simply the difference between the actual and forecasted values. That is, $e_{T+h,T} = y_{T+h} - y_{T+h,T}$.

for $h \leq q$ and

$$e_{T+h,T} = \mathrm{MA}(q)$$

for $h > q$. The h-step-ahead forecast error for $h > q$ is just the process itself, minus its mean.

Finally, consider the **forecast error variances.** For $h \leq q$,

$$\sigma_h^2 \leq \mathrm{var}(y_t)\,;$$

whereas for $h > q$,

$$\sigma_h^2 = \mathrm{var}(y_t)\,.$$

In summary, we've thus far studied the MA(2) and then the general MA(q) processes, computing the optimal h-step-ahead forecast, the corresponding forecast error, and the forecast error variance. As we'll now see, the emerging patterns that we cataloged turn out to be quite general.

OPTIMAL POINT FORECASTS FOR INFINITE-ORDER MOVING AVERAGES

By now you're getting the hang of it, so let's consider the general case of an infinite-order MA process. The infinite-order moving average process may seem like a theoretical curiosity, but precisely the opposite is true. Any covariance stationary process can be written as a (potentially infinite-order) moving average process, and moving average processes are easy to understand and manipulate, because they are written in terms of white noise shocks, which have very simple statistical properties. Thus, if you take the time to understand the mechanics of constructing optimal forecasts for infinite moving average processes, you'll understand everything, and you'll have some powerful technical tools and intuition at your command.

Recall from Chapter 7 that the general linear process is

$$y_t = \sum_{i=0}^{\infty} b_i \varepsilon_{t-i}\,,$$

where $\varepsilon_t \sim WN(0, \sigma^2)$, $b_0 = 1$, and $\sigma^2 \sum_{i=0}^{\infty} b_i^2 < \infty$. We proceed in the usual way. We first write out the process at the future time of interest:

$$y_{T+h} = \varepsilon_{T+h} + b_1\varepsilon_{T+h-1} + \cdots + b_h\varepsilon_T + b_{h+1}\varepsilon_{T-1} + \cdots\,.$$

Then we project y_{T+h} on the time-T information set. The projection yields 0s for all of the future ε's (because they are white noise and hence unforecastable), leaving

$$y_{T+h,T} = b_h\varepsilon_T + b_{h+1}\varepsilon_{T-1} + \cdots\,.$$

It follows that the *h*-step-ahead forecast error is serially correlated; it follows an MA($h-1$) process,

$$e_{T+h,T} = (y_{T+h} - y_{T+h,T}) = \sum_{i=0}^{h-1} b_i \varepsilon_{T+h-i} \, ,$$

with mean 0 and variance

$$\sigma_h^2 = \sigma^2 \sum_{i=0}^{h-1} b_i^2 \, .$$

A number of remarks are in order concerning the optimal forecasts of the general linear process and the corresponding forecast errors and forecast error variances. First, the 1-step-ahead forecast error is simply ε_{T+1}. ε_{T+1} is that part of y_{T+1} that can't be linearly forecast on the basis of Ω_t (which, again, is why it is called the innovation). Second, although it might at first seem strange that an *optimal* forecast error would be serially correlated, as is the case when $h > 1$, nothing is awry. The serial correlation can't be used to improve forecasting performance, because the autocorrelations of the MA($h-1$) process cut off just before the beginning of the time-T information set $\{\varepsilon_T, \varepsilon_{T-1}, \ldots\}$. This is a general and tremendously important property of the errors associated with optimal forecasts: *Errors from optimal forecasts can't be forecast using information available when the forecast was made.* If you can forecast the forecast error, then you can improve the forecast, which means that it couldn't have been optimal. Finally, note that as h approaches infinity $y_{T+h,T}$ approaches 0, the unconditional mean of the process, and σ_h^2 approaches $\sigma^2 \sum_{i=0}^{\infty} b_i^2$, the unconditional variance of the process, which reflects the fact that as h approaches infinity, the conditioning information on which the forecast is based becomes progressively less useful. In other words, the distant future is harder to forecast than the near future!

INTERVAL AND DENSITY FORECASTS

Now we construct interval and density forecasts. Regardless of whether the moving average is finite or infinite, we proceed in the same way, as follows. The definition of the *h*-step-ahead forecast error is

$$e_{T+h,T} = y_{T+h} - y_{T+h,T} \, .$$

Equivalently, the *h*-step-ahead realized value, y_{T+h}, equals the forecast plus the error,

$$y_{T+h} = y_{T+h,T} + e_{T+h,T} \, .$$

If the innovations are normally distributed, then the future value of the series of interest is also normally distributed, conditional on the information set available at the time the forecast was made, and so we have the 95% h-step-ahead interval forecast $y_{T+h,T} \pm 1.96\sigma_h$.[2] In similar fashion, we construct the h-step-ahead density forecast as $N(y_{T+h,T}, \sigma_h^2)$. The mean of the conditional distribution of y_{T+h} is $y_{T+h,T}$, which of course must be the case because we constructed the point forecast as the conditional mean, and the variance of the conditional distribution is σ_h^2, the variance of the forecast error.

As an example of interval and density forecasting, consider again the MA(2) process,

$$y_t = \varepsilon_t + \theta_1\varepsilon_{t-1} + \theta_2\varepsilon_{t-2}$$

$$\varepsilon_t \sim WN(0, \sigma^2)\,.$$

Assuming normality, the 1-step-ahead 95% interval forecast is $y_{T+1,T} = (\theta_1\varepsilon_T + \theta_2\varepsilon_{T-1}) \pm 1.96\sigma$, and the 1-step-ahead density forecast is $N(\theta_1\varepsilon_T + \theta_2\varepsilon_{T-1}, \sigma^2)$.

3. Making the Forecasts Operational

So far we've assumed that the parameters of the process being forecast are known. In practice, of course, they must be estimated. To make our forecasting procedures operational, we simply replace the unknown parameters in our formulas with estimates and the unobservable innovations with residuals.

Consider, for example, the MA(2) process,

$$y_t = \varepsilon_t + \theta_1\varepsilon_{t-1} + \theta_2\varepsilon_{t-2}\,.$$

As you can readily verify using the methods we've introduced, the 2-step-ahead optimal forecast, assuming known parameters, is

$$y_{T+2,T} = \theta_2\varepsilon_T\,,$$

with corresponding forecast error

$$e_{T+2,T} = \varepsilon_{T+2} + \theta_1\varepsilon_{T+1}$$

and forecast error variance

$$\sigma_2^2 = \sigma^2(1 + \theta_1^2)\,.$$

---

[2] Confidence intervals at any other desired confidence level may be constructed in similar fashion, by using a different critical point of the standard normal distribution. A 90% interval forecast, for example, is $y_{T+h,T} \pm 1.64\sigma_h$. In general, for a Gaussian process, a $(1-\alpha)100\%$ confidence interval is $y_{T+h,T} \pm z_{\alpha/2}\sigma_h$, where $z_{\alpha/2}$ is that point on the $N(0, 1)$ distribution such that prob$(z > z_{\alpha/2}) = \alpha/2$.

To make the forecast operational, we replace unknown parameters with estimates and the time-T innovation with the time-T residual, yielding

$$\hat{y}_{T+2,T} = \hat{\theta}_2 \hat{\varepsilon}_T$$

and forecast error variance

$$\hat{\sigma}_2^2 = \hat{\sigma}^2\left(1 + \hat{\theta}_1^2\right).$$

Then, if desired, we can construct operational 2-step-ahead interval and density forecasts, as $\hat{y}_{T+2,T} \pm z_{\alpha/2}\hat{\sigma}_2$ and $N(\hat{y}_{T+2,T}, \hat{\sigma}_2^2)$.

The strategy of taking a forecast formula derived under the assumption of known parameters, and replacing unknown parameters with estimates, is a natural way to operationalize the construction of point forecasts. However, using the same strategy to produce operational interval or density forecasts involves a subtlety that merits additional discussion. The forecast error variance estimate so obtained can be interpreted as one that ignores parameter estimation uncertainty, as follows. Recall once again that the actual future value of the series is

$$y_{T+2} = \varepsilon_{T+2} + \theta_1 \varepsilon_{T+1} + \theta_2 \varepsilon_T$$

and that the operational forecast is

$$\hat{y}_{T+2,T} = \hat{\theta}_2 \varepsilon_T \ .$$

Thus, the exact forecast error is

$$\hat{e}_{T+2,T} = y_{T+2} - \hat{y}_{T+2,T} = \varepsilon_{T+2} + \theta_1 \varepsilon_{T+1} + (\theta_2 - \hat{\theta}_2)\varepsilon_T\,,$$

the variance of which is very difficult to evaluate. So we make a convenient approximation: We ignore parameter estimation uncertainty by assuming that estimated parameters equal true parameters. We therefore set $(\theta_2 - \hat{\theta}_2)$ to 0, which yields

$$\hat{e}_{T+2,T} = \varepsilon_{T+2} + \theta_1 \varepsilon_{T+1}\,,$$

with variance

$$\sigma_2^2 = \sigma^2\left(1 + \theta_1^2\right),$$

which we make operational as

$$\hat{\sigma}_2^2 = \hat{\sigma}^2\left(1 + \hat{\theta}_1^2\right).$$

4. The Chain Rule of Forecasting

POINT FORECASTS OF AUTOREGRESSIVE PROCESSES

Because any covariance stationary AR(p) process can be written as an infinite moving average, there's no need for specialized forecasting techniques for autoregressions. Instead, we can simply transform the autoregression into a moving average and then use the techniques we developed for forecasting moving averages. It turns out, however, that a very simple recursive method for computing the optimal forecast is available in the autoregressive case.

The recursive method, called the **chain rule of forecasting,** is best learned by example. Consider the AR(1) process,

$$y_t = \varphi y_{t-1} + \varepsilon_t$$
$$\varepsilon_t \sim WN(0, \sigma^2) .$$

First we construct the optimal 1-step-ahead forecast, and then we construct the optimal 2-step-ahead forecast, which depends on the optimal 1-step-ahead forecast, which we've already constructed. Then we construct the optimal 3-step-ahead forecast, which depends on the already-computed 2-step-ahead forecast, which we've already constructed, and so on.

To construct the 1-step-ahead forecast, we write out the process for time $T + 1$,

$$y_{T+1} = \varphi y_T + \varepsilon_{T+1} .$$

Then, projecting the right-hand side on the time-T information set, we obtain

$$y_{T+1,T} = \varphi y_T .$$

Now let's construct the 2-step-ahead forecast. Write out the process for time $T + 2$,

$$y_{T+2} = \varphi y_{T-1} + \varepsilon_{T+2} .$$

Then project directly on the time-T information set to get

$$y_{T+2,T} = \varphi y_{T+1,T} .$$

Note that the future innovation is replaced by 0, as always, and that we have directly replaced the time $T + 1$ value of y with its earlier-constructed optimal forecast. Now let's construct the 3-step-ahead forecast. Write out the process for time $T + 3$,

$$y_{T+3} = \varphi y_{T+2} + \varepsilon_{T+3} .$$

Then project directly on the time-T information set,

$$y_{T+3,T} = \varphi y_{T+2,T} .$$

The required 2-step-ahead forecast was already constructed.

Continuing in this way, we can recursively build up forecasts for any and all future periods—hence the name (chain rule of forecasting). Note that for the AR(1) process, only the most recent value of y is needed to construct optimal forecasts, for any horizon, and for the general AR(p) process, only the p most recent values of y are needed.

POINT FORECASTS OF ARMA PROCESSES

Now we consider forecasting covariance-stationary ARMA processes. Just as with autoregressive processes, we could always convert an ARMA process to an infinite moving average and then use our earlier-developed methods for forecasting moving averages. But also as with autoregressive processes, a simpler method is available for forecasting ARMA processes directly, by *combining* our earlier approaches to moving average and autoregressive forecasting.

As always, we write out the ARMA(p, q) process for the future period of interest,

$$y_{T+h} = \varphi_1 y_{T+h-1} + \cdots + \varphi_p y_{T+h-p} + \varepsilon_{T+h} + \theta_1 \varepsilon_{T+h-1} + \cdots + \theta_q \varepsilon_{T+h-q} .$$

On the right-hand side we have various future values of y and ε, and perhaps also past values, depending on the forecast horizon. We replace everything on the right-hand side with its projection on the time-T information set. That is, we replace all future values of y with optimal forecasts (built up recursively using the chain rule) and all future values of ε with optimal forecasts (0), yielding

$$y_{T+h,T} = \varphi_1 y_{T+h-1,T} + \cdots + \varphi_p y_{T+h-p,T} + \varepsilon_{T+h,T} + \theta_1 \varepsilon_{T+h-1,T} + \cdots + \theta_q \varepsilon_{T+h-q,T} .$$

When evaluating this formula, note that the optimal time-T "forecast" of any value of y or ε dated time T or earlier is just y or ε itself.

As an example, consider forecasting the ARMA (1, 1) process,

$$y_t = \varphi y_{t-1} + \varepsilon_t + \theta \varepsilon_{t-1}$$
$$\varepsilon_t \sim WN(0, \sigma^2) .$$

Let's find $y_{T+1,T}$. The process at time $T+1$ is

$$y_{T+1} = \varphi y_T + \varepsilon_{T+1} + \theta \varepsilon_T .$$

Projecting the right-hand side on Ω_T yields

$$y_{T+1,T} = \varphi y_T + \theta \varepsilon_T .$$

Now let's find $y_{T+2,T}$. The process at time $T+2$ is

$$y_{T+2} = \varphi y_{T+1} + \varepsilon_{T+2} + \theta \varepsilon_{T+1} .$$

Projecting the right-hand side on Ω_T yields

$$y_{T+2,T} = \varphi y_{T+1,T} .$$

Substituting our earlier-computed 1-step-ahead forecast yields

$$\begin{aligned} y_{T+2,T} &= \varphi(\varphi y_T + \theta \varepsilon_T) \\ &= \varphi^2 y_T + \varphi\theta \varepsilon_T . \end{aligned}$$

Continuing, it is clear that

$$y_{T+h,T} = \varphi y_{T+h-1,T} ,$$

for all $h > 1$.

INTERVAL AND DENSITY FORECASTS

The chain rule, whether applied to pure autoregressive models or to ARMA models, is a device for simplifying the computation of *point* forecasts. Interval and density forecasts require the *h*-step-ahead forecast error variance, which we get from the moving average representation, as discussed earlier. It is

$$\sigma_h^2 = \sigma^2 \sum_{i=0}^{h-1} b_i^2 ,$$

which we operationalize as

$$\hat{\sigma}_h^2 = \hat{\sigma}^2 \sum_{i=0}^{h-1} \hat{b}_i^2 .$$

Note that we don't actually estimate the moving average representation; rather, we solve backward for as many *b*'s as we need, *in terms of the original model parameters*, which we then replace with estimates.

Let's illustrate by constructing a 2-step-ahead 95% interval forecast for the ARMA(1, 1) process. We already constructed the 2-step-ahead point forecast, $y_{T+2,T}$; we need only compute the 2-step-ahead forecast error variance. The process is

$$y_t = \varphi y_{t-1} + \varepsilon_t + \theta \varepsilon_{t-1} .$$

Substitute backward for y_{t-1} to get

$$\begin{aligned} y_t &= \varphi(\varphi y_{t-2} + \varepsilon_{t-1} + \theta\varepsilon_{t-2}) + \varepsilon_t + \theta\varepsilon_{t-1} \\ &= \varepsilon_t + (\varphi + \theta)\varepsilon_{t-1} + \cdots . \end{aligned}$$

We need not substitute back any farther, because the 2-step-ahead forecast error variance is $\sigma_2^2 = \sigma^2(1 + b_1^2)$, where b_1 is the coefficient on ε_{t-1} in the moving average representation of the ARMA(1, 1) process, which we just calculated to be $(\varphi + \theta)$. Thus, the 2-step-ahead interval forecast is $y_{T+2,T} \pm 1.96\sigma_2$, or $(\varphi^2 y_T + \varphi\theta\varepsilon_T) \pm 1.96\sigma\sqrt{1 + (\varphi + \theta)^2}$. We make this operational as $(\hat{\varphi}^2 y_T + \hat{\varphi}\hat{\theta}\varepsilon_T) \pm 1.96\hat{\sigma}\sqrt{1 + (\hat{\varphi} + \hat{\theta})^2}$.

5. Application: Forecasting Employment

Now we put our forecasting technology to work to produce point and interval forecasts for Canadian employment. Recall that the best moving average model was an MA(4), while the best autoregressive model, as well as the best ARMA model and the best model overall, was an AR(2).

First, consider forecasting with the MA(4) model. In Figure 9.1, we show the employment history together with operational 4-quarter-ahead point and interval extrapolation forecasts. The 4-quarter-ahead extrapolation forecast reverts very quickly to the mean of the employment index. In 1993.4, the last quarter of historical data, employment is well below its mean, but the forecast calls for a quick rise. The forecasted quick rise seems unnatural because employment dynamics are historically very persistent. If employment is well below its mean in 1993.4, we'd expect it to stay well below its mean for some time.

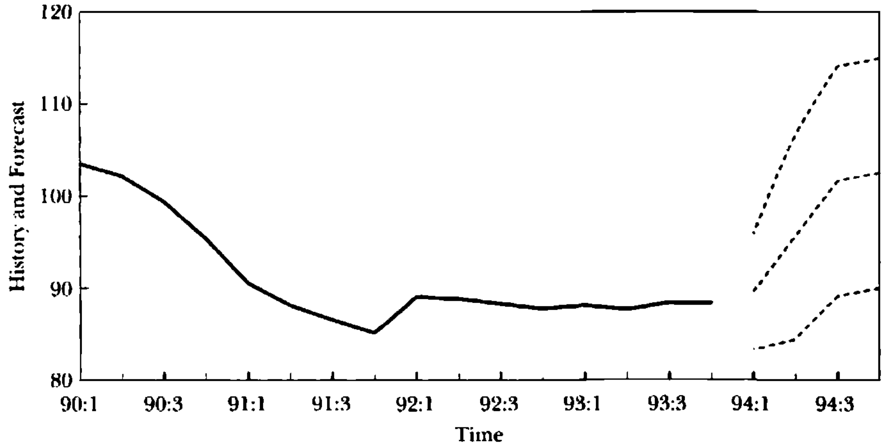


FIGURE 9.1 *Employment History and Forecast MA(4) Model*

The MA(4) model is unable to capture such persistence. The quick reversion of the MA(4) forecast to the mean is a manifestation of the short memory of moving average processes. Recall, in particular, that an MA(4) process has a four-period memory—all autocorrelations are 0 beyond displacement 4. Thus, all forecasts more than four steps ahead are simply equal to the unconditional mean (100.2), and all 95% interval forecasts more than four steps ahead are plus or minus 1.96 unconditional standard deviations. All of this is made clear in Figure 9.2, in which we show the employment history together with 12-step-ahead point and interval extrapolation forecasts.

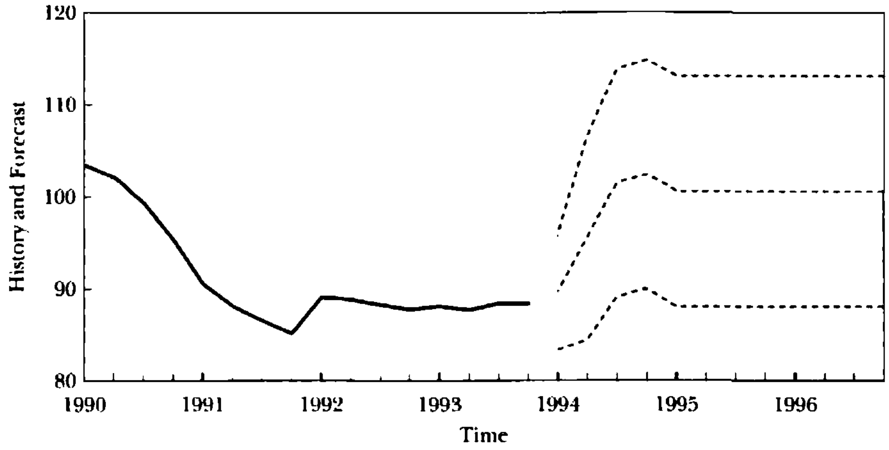
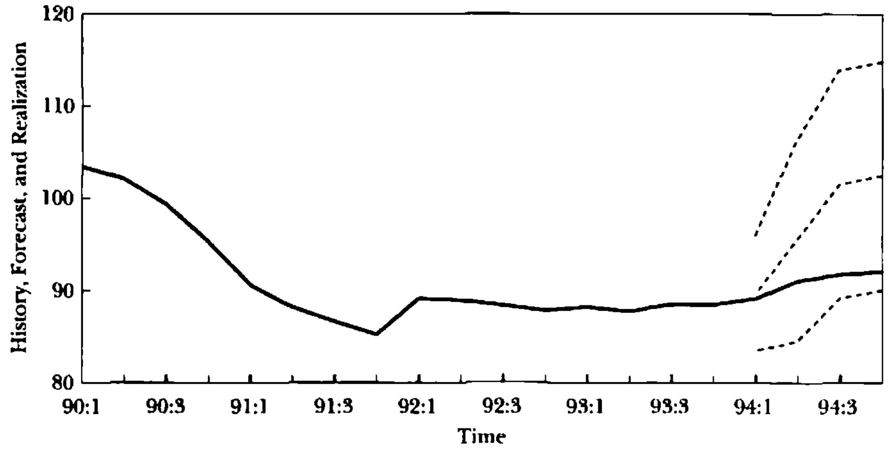


FIGURE 9.2 *Employment History and Long Horizon Forecast MA(4) Model*

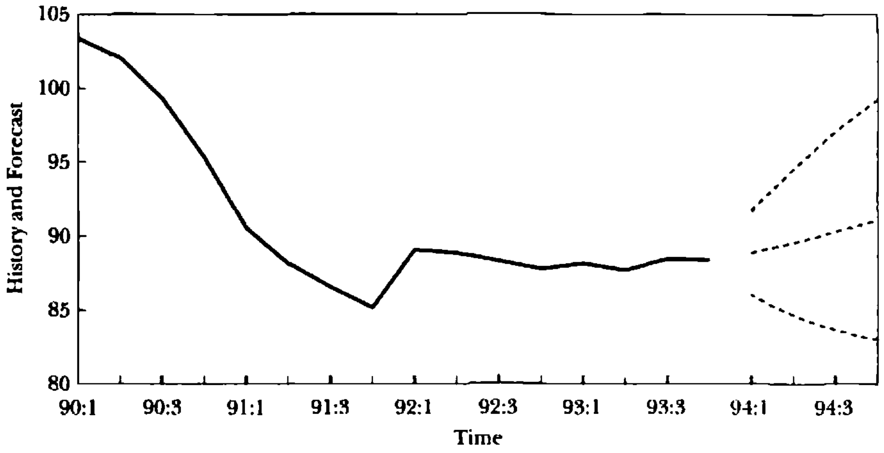
FIGURE 9.3 *Employment History, Forecast, and Realization MA(4) Model*



In Figure 9.3, we show the 4-quarter-ahead forecast and realization. Our suspicions are confirmed. The actual employment series stays well below its mean over the forecast period, whereas the forecast rises quickly back to the mean. The mean squared forecast error is a large 55.9.

Now consider forecasting with the AR(2) model. In Figure 9.4, we show the 4-quarter-ahead extrapolation forecast, which reverts to the unconditional mean much less quickly, as seems natural given the high persistence of employment. The 4-quarter-ahead point forecast, in fact, is still well below the mean. Similarly, the 95% error bands grow gradually and haven't approached their long-horizon values by four quarters out.

FIGURE 9.4 *Employment History and Forecast AR(2) Model*



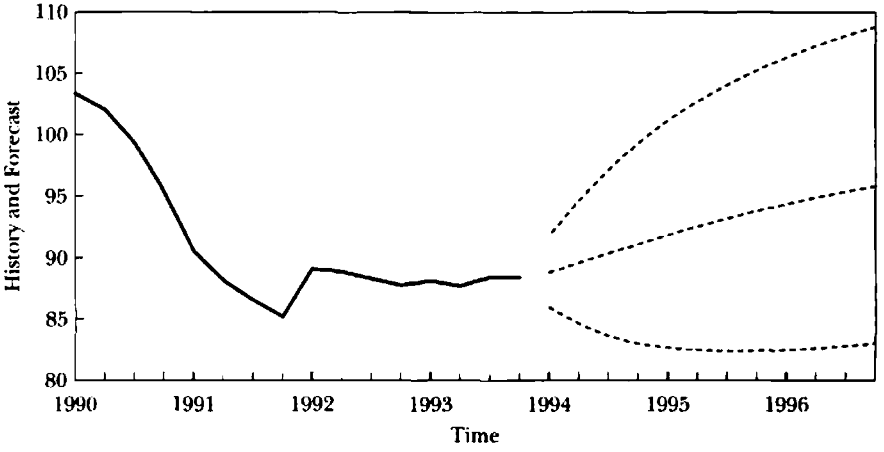


FIGURE 9.5
Employment History and Long-Horizon Forecast AR(2) Model

Figures 9.5 and 9.6 make clear the very different nature of the autoregressive forecasts. Figure 9.5 presents the 12-step-ahead extrapolation forecast, and Figure 9.6 presents a much longer-horizon extrapolation forecast. Eventually the unconditional mean *is* approached, and eventually the error bands *do* go flat, but only for very long-horizon forecasts, due to the high persistence in employment, which the AR(2) model captures.

Figure 9.7 shows the employment history, 4-quarter-ahead AR(2) extrapolation forecast, and the realization. The AR(2) forecast appears quite accurate; the mean squared forecast error is 1.3, drastically smaller than that of the MA(4) forecast.

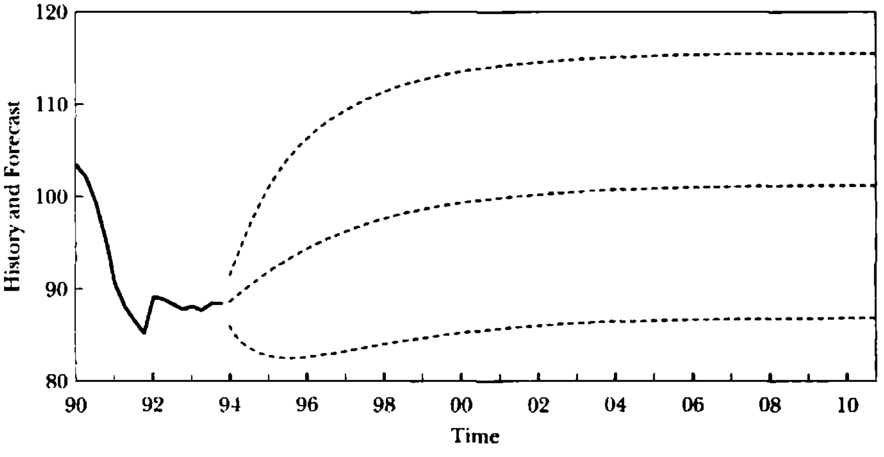
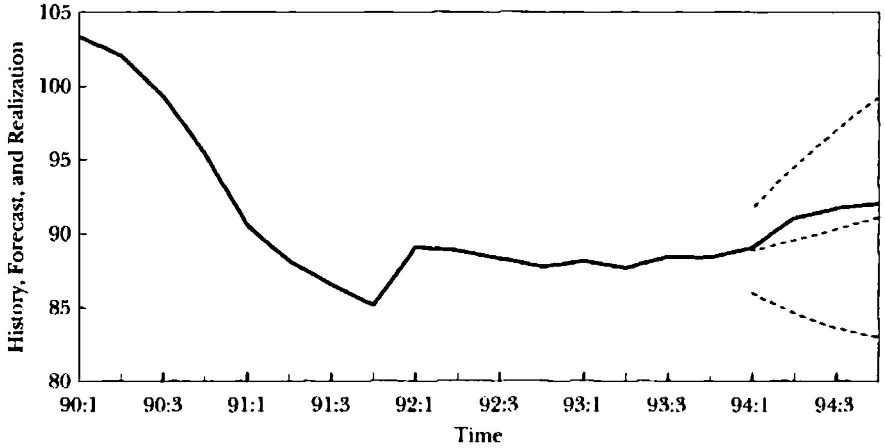


FIGURE 9.6
Employment History and Very Long-Horizon Forecast AR(2) Model

FIGURE 9.7
Employment History, Forecast, and Realization AR(2) Model



Exercises, Problems, and Complements

1. (Forecast accuracy across horizons) You are a consultant to MedTrax, a large pharmaceutical company, which released a new ulcer drug 3 months ago and is concerned about recovering research and development costs. Accordingly, MedTrax has approached you for drug sales projections at 1- through 12-month-ahead horizons, which it will use to guide potential sales force realignments. In briefing you, MedTrax indicated that it expects your long-horizon forecasts (e.g., 12-month-ahead) to be just as accurate as your short-horizon forecasts (e.g., 1-month-ahead). Explain to MedTrax why that is not likely to be the case, even if you do the best forecasting job possible.

2. (Mechanics of forecasting with ARMA models: BankWire continued) On the book's web page, you will find data for daily transfers over BankWire, a wire transfer system in a country responsible for much of the world's finance, over a recent span of 200 business days.
   a. In Chapter 8's Exercises, Problems, and Complements, you were asked to find a parsimonious ARMA(p, q) model that fits the transfer data well and to defend its adequacy. Repeat the exercise, this time using only the first 175 days for model selection and fitting. Is it necessarily the case that the selected ARMA model will remain the same as when all 200 days are used? Does yours?
   b. Use your estimated model to produce point and interval forecasts for days 176 through 200. Plot them and discuss the forecast pattern.
   c. Compare your forecasts with the actual realizations. Do the forecasts perform well? Why or why not?
   d. Discuss precisely how your software constructs point and interval forecasts. It should certainly match our discussion in spirit, but it may differ in some of the details. Are you uncomfortable with any of the assumptions made? How, if at all, could the forecasts be improved?

3. (Forecasting an AR(1) process with known and unknown parameters) Use the chain rule to forecast the AR(1) process,

$$y_t = \varphi y_{t-1} + \varepsilon_t \,.$$

For now, assume that all parameters are known.

a. Show that the optimal forecasts are

$$y_{T+1,T} = \varphi y_T$$

$$y_{T+2,T} = \varphi^2 y_T$$

$$\dots$$

$$y_{T+h,T} = \varphi^h y_T \,.$$

b. Show that the corresponding forecast errors are

$$e_{T+1,T} = (y_{T+1} - y_{T+1,T}) = \varepsilon_{T+1}$$

$$e_{T+2,T} = (y_{T+2} - y_{T+2,T}) = \varphi\, \varepsilon_{T+1} + \varepsilon_{T+2}$$

$$\dots$$

$$e_{T+h,T} = (y_{T+h} - y_{T+h,T}) = \varepsilon_{T+h} + \varphi \varepsilon_{T+h-1} + \cdots + \varphi^{h-1}\varepsilon_{T+1} \,.$$

c. Show that the forecast error variances are

$$\sigma_1^2 = \sigma^2$$

$$\sigma_2^2 = \sigma^2(1 + \varphi^2)$$

$$\dots$$

$$\sigma_h^2 = \sigma^2 \sum_{i=0}^{h-1} \varphi^{2i} \,.$$

d. Show that the limiting forecast error variance is

$$\lim_{h\to\infty} \sigma_h^2 = \frac{\sigma^2}{1-\varphi^2} \,,$$

the unconditional variance of the AR(1) process.

Now assume that the parameters are unknown and so must be estimated.

e. Make your expressions for both the forecasts and the forecast error variances operational, by inserting least-squares estimates where unknown parameters appear, and use them to produce an operational point forecast and an operational 90% interval forecast for $y_{T+2,T}$.

4. (Forecasting an ARMA(2, 2) process) Consider the ARMA(2, 2) process:

$$y_t = \varphi_1 y_{t-1} + \varphi_2 y_{t-2} + \varepsilon_t + \theta_1 \varepsilon_{t-1} + \theta_2 \varepsilon_{t-2} \,.$$

a. Verify that the optimal 1-step-ahead forecast made at time T is

$$y_{T+1,T} = \varphi_1 y_T + \varphi_2 y_{T-1} + \theta_1 \varepsilon_T + \theta_2 \varepsilon_{T-1} \,.$$

b. Verify that the optimal 2-step-ahead forecast made at time T is

$$y_{T+2,T} = \varphi_1 y_{T+1,T} + \varphi_2 y_T + \theta_2 \varepsilon_T \,,$$

and express it purely in terms of elements of the time-T information set.

c. Verify that the optimal 3-step-ahead forecast made at time T is

$$y_{T+3,T} = \varphi_1 y_{T+2,T} + \varphi_2 y_{T+1,T}\,,$$

and express it purely in terms of elements of the time-T information set.

d. Show that for any forecast horizon h greater than or equal to 3,

$$y_{T+h,T} = \varphi_1 y_{T+h-1,T} + \varphi_2 y_{T+h-2,T}\,.$$

5. (Optimal forecasting under asymmetric loss) One of the conditions required for optimality of the conditional mean forecast is **symmetric loss.** We make that assumption for a number of reasons. First, the conditional mean is usually easy to compute. In contrast, optimal forecasting under **asymmetric loss** is rather involved, and the tools for doing so are still under development. (See, e.g., Christoffersen and Diebold, 1997.) Second, and more importantly, symmetric loss often provides a good approximation to the loss structure relevant in a particular decision environment.

Symmetric loss is not *always* appropriate, however. Here we discuss some aspects of forecasting under asymmetric loss. Under asymmetric loss, optimal forecasts are biased, whereas the conditional mean forecast is unbiased.[3] Bias is optimal under asymmetric loss because we can gain on average by pushing the forecasts in the direction such that we make relatively few errors of the more costly sign.

Many asymmetric loss functions are possible. A few, however, have proved particularly useful, because of their flexibility and tractability. One is the **linex loss function,**

$$L(e) = b(\exp(ae) - ae - 1), \quad a \neq 0,\, b > 0\,.$$

It's called *linex* because when $a > 0$, loss is approximately linear to the left of the origin and approximately exponential to the right, and conversely when $a < 0$. Another is the **linlin loss function,** given by

$$L(e) = \begin{cases} a|e|, & \text{if } e > 0 \\ b|e|, & \text{if } e \leq 0 \end{cases}.$$

Its name comes from the linearity on each side of the origin.

a. Discuss three practical forecasting situations in which the loss function might be asymmetric. Give detailed reasons for the asymmetry, and discuss how you might produce and evaluate forecasts.

b Explore and graph the linex and linlin loss functions for various values of a and b. Discuss the roles played by a and b in each loss function. In particular, which parameter or combination of parameters governs the degree of asymmetry? What happens to the linex loss function as a gets smaller? What happens to the linlin loss function as a/b approaches 1?

---

[3] A forecast is unbiased if its error has zero mean. The error from the conditional mean forecast has zero mean, by construction.

6. (Truncation of infinite distributed lags, state space representations, and the Kalman filter) This complement concerns practical implementation of formulas that involve innovations (ε's). Earlier we noted that as long as a process is invertible, we can express the ε's in terms of the y's. If the process involves a moving average component, however, the ε's will depend on the infinite past history of the y's, so we need to truncate to make it operational. Suppose, for example, that we're forecasting the MA(1) process,

$$y_t = \varepsilon_t + \theta\varepsilon_{t-1} .$$

The operational 1-step-ahead forecast is

$$y_{t+1,T} = \hat{\theta}\hat{\varepsilon}_T .$$

But what, precisely, do we insert for the residual, $\hat{\varepsilon}_T$? Back substitution yields the autoregressive representation,

$$\varepsilon_t = y_t + \theta y_{t-1} - \theta^2 y_{t-2} + \cdots .$$

Thus,

$$\varepsilon_T = y_T + \theta y_{T-1} - \theta^2 y_{T-2} + \cdots ,$$

which we are forced to truncate at time $T = 1$, when the data begin. This yields the approximation

$$\varepsilon_T \approx y_T + \theta y_{T-1} - \theta^2 y_{T-2} + \cdots + \theta^T y_1 .$$

Unless the sample size is very small, or θ is very close to 1, the approximation will be very accurate, because θ is less than 1 in absolute value (by invertibility), and we're raising it to higher and higher powers. Finally, we make the expression operational by replacing the unknown moving average parameter with an estimate, yielding

$$\hat{\varepsilon}_T \approx y_T + \hat{\theta} y_{T-1} - \hat{\theta}^2 y_{T-2} + \cdots + \hat{\theta}^T y_1 .$$

In the engineering literature of the 1960s and then in the statistics and econometrics literatures of the 1970s, important tools called *state-space representations* and the Kalman filter were developed. Those tools provide a convenient and powerful framework for estimating a wide variety of forecasting models and constructing optimal forecasts, and they enable us to tailor the forecasts precisely to the sample of data at hand, so that no truncation is necessary.

7. (Point and interval forecasts allowing for serial correlation—Nile.com continued) The book's website has data for the Internet retailer Nile.com, giving the number of hits at its website each day from January 1, 1998, through September 28, 1998. Your marketing firm, CyberMedia, which specializes in developing quick, intensive marketing strategies based on short-term projections, is hired to develop a forecasting model for hits at the Nile.com website.
   a. In Chapter 6, Problem 6, you estimated a trend + seasonal model for Nile.com hits, ignoring the possible presence of cyclical dynamics. Now generalize your earlier model to allow for cyclical dynamics, if present, via AR(p) disturbances. Write the full specification of your model in general notation (e.g., with p left unspecified).

b. Estimate three versions of your full model, corresponding to $p = 0, 1, 2, 3$, while leaving the original trend and seasonal specifications intact, and select the one that optimizes SIC.

c. Using the model selected in part *b*, write theoretical expressions for the 1- and 2-day-ahead point forecasts and 95% interval forecasts, using estimated parameters.

d. Calculate those point and interval forecasts for Nile.com for 9/29 and 9/30.

8. (**Bootstrapping** simulation to acknowledge **innovation distribution uncertainty** and **parameter estimation uncertainty**) A variety of simulation-based methods fall under the general heading of "bootstrap." Their common element, and the reason for the name *bootstrap*, is that they build up an approximation to an object of interest directly from the data. Hence, they "pull themselves up by their own bootstraps." For example, the object of interest might be the distribution of a random disturbance, which has implications for interval and density forecasts, and about which we might sometimes feel uncomfortable making a possibly erroneous assumption such as normality.

a. The density and interval forecasts that we've discussed rely crucially on normality. In many situations, normality is a perfectly reasonable and useful assumption; after all, that's why we call it the "normal" distribution. Sometimes, however, such as when we forecast high-frequency financial asset returns, normality may be unrealistic. Using bootstrap methods, we can relax the normality assumption. Suppose, for example, that we want a 1-step-ahead interval forecast for an AR(1) process. We know that the future observation of interest is

$$y_{T+1} = \varphi y_T + \varepsilon_{T+1} .$$

We know y_T, and we can estimate φ and then proceed as if φ were known, using the operational point forecast, $\hat{y}_{T+1,T} = \hat{\varphi} y_T$. If we want an operational interval forecast, however, we've thus far relied on a normality assumption, in which case we use $\hat{y}_{T+1,T} \pm z_{\alpha/2}\hat{\sigma}$. To relax the normality assumption, we can proceed as follows. Imagine that we could sample from the distribution of ε_{T+1}—whatever that distribution might be. Take R draws, $\{\varepsilon_{T+1}^{(i)}\}_{i=1}^{R}$, where R is a large number, such as 10,000. For each such draw, construct the corresponding forecast of y_{T+1} as

$$\hat{y}_{T+1,T}^{(i)} = \hat{\varphi} y_T + \varepsilon_{T+1}^{(i)} .$$

Then form a histogram of the $\hat{y}_{T+1,T}^{(i)}$ values, which is the density forecast. And given the density forecast, we can of course construct interval forecasts at any desired level. If, for example, we want a 90% interval, we can sort the $\hat{y}_{T+1,T}^{(i)}$ values from smallest to largest, find the 5th percentile (call it a) and the 95th percentile (call it b), and use the 90% interval forecast $[a, b]$.

b. The only missing link in this strategy is how to sample from the distribution of ε_{T+1}. It turns out that it's easy to do—we simply assign probability $1/T$ to each of the observed residuals (which are estimates of the unobserved ε's) and draw from them R times with replacement. Describe how you might do so.

c. Note that the interval and density forecasts we've constructed thus far—even the one earlier based on bootstrap techniques—make no attempt to account for parameter estimation uncertainty. Intuitively, we would expect confidence intervals obtained by ignoring parameter estimation uncertainty to be more narrow than they would be if parameter uncertainty were accounted for, thereby producing an artificial appearance of precision. In spite of this defect, parameter uncertainty is usually ignored in practice, for a number of reasons. The uncertainty associated with estimated parameters vanishes as the sample size grows; in fact, it vanishes quickly. Furthermore, the fraction of forecast error attributable to the difference between estimated and true parameters is likely to be small compared with the fraction of forecast error coming from other sources, such as using a model that does a poor job of approximating the dynamics of the variable being forecast.

d. Quite apart from the reasons already given for ignoring parameter estimation uncertainty, the biggest reason is probably that, until very recently, mathematical and computational difficulties made attempts to account for parameter uncertainty infeasible in many situations of practical interest. Modern computing speed, however, lets us use the bootstrap to approximate the effects of parameter estimation uncertainty. To continue with the AR(1) example, suppose that we know that the disturbances are Gaussian but that we want to attempt to account for the effects of parameter estimation uncertainty when we produce our 1-step-ahead density forecast. How could we use the bootstrap to do so?

e. The "real sample" of data ends with observation y_T, and the optimal point forecast depends only on y_T. It would therefore seem desirable that all of your *R* "bootstrap samples" of data also end with y_T. Do you agree? How might you enforce that property while still respecting the AR(1) dynamics? (This is tricky.)

f. Can you think of a way to assemble the results thus far to produce a density forecast that acknowledges both innovation distribution uncertainty and parameter estimation uncertainty? (This is very challenging.)

Bibliographical and Computational Notes

The methods discussed in this chapter were developed by Wiener, Kolmogorov, and Wold more than 50 years ago, and they underlie all modern forecasting software. It's important to understand them so that you're the master of your software, not the opposite.

For a proof of our assertion of optimality of the conditional mean forecast, as well as a precise statement of the conditions under which the result holds, see any good advanced text, such as Hamilton (1994).

Linex loss was introduced by Varian (1974) in the context of real estate assessment and was further studied by Zellner (1986). Harvey (1993) gives a lucid exposition of state-space representations and the Kalman filter. Efron and Tibshirani (1993) offer a good introduction to the bootstrap and its many uses. Stine (1987) and Breidt, Davis, and Dunsmuir (1995) show how to use the bootstrap to produce interval and density forecasts under weak assumptions. Chatfield (1993, 1995) argues that the fraction of forecast error attributable to the difference between estimated and true parameters is likely much smaller than the fraction of forecast error coming from other sources, such as model misspecification. Clements and Hendry (1994, 1998) provide insightful discussion of a variety of advanced topics in applied forecasting.

Concepts for Review

Expected loss
Linear forecast
Linear least-squares forecast
Forecast error variance
Chain rule of forecasting
Linex loss function
Linlin loss function
Bootstrapping
Innovation distribution uncertainty
Parameter estimation uncertainty

References and Additional Readings

Breidt, F. J., Davis, R. A., and Dunsmuir, W. T. M. (1995). "Improved Bootstrap Prediction Intervals for Autoregressions." *Journal of Time Series Analysis,* 16, 177–200.

Chatfield, C. (1993). "Calculating Interval Forecasts (with Discussion)." *Journal of Business and Economic Statistics,* 11, 121–144.

Chatfield, C. (1995). "Model Uncertainty, Data Mining and Statistical Inference (with Discussion)." *Journal of the Royal Statistical Society* A, 158, 419–466.

Christoffersen, P. F., and Diebold, F. X. (1997). "Optimal Prediction Under Asymmetric Loss." *Econometric Theory,* 13, 808–817.

Clements, M. P., and Hendry, D. F. (1994). "Towards a Theory of Economic Forecasting." In C. P. Hargreaves (ed.), *Nonstationary Times Series Analysis and Cointegration.* Oxford: Oxford University Press.

Clements, M. P., and Hendry, D. F. (1998). *Forecasting Economic Time Series* (The Marshall Lectures in Economic Forecasting). Cambridge: Cambridge University Press.

Efron, B., and Tibshirani, R. J. (1993). *An Introduction to the Bootstrap.* London: Chapman and Hall.

Hamilton, J. D. (1994). *Time Series Analysis.* Princeton, N. J.: Princeton University Press.

Harvey, A. C. (1993). *Time Series Models.* 2nd ed. Cambridge, Mass.: MIT Press.

Stine, R. A. (1987). "Estimating Properties of Autoregressive Forecasts." *Journal of the American Statistical Association,* 82, 1072–1078.

Varian, H. (1974). "A Bayesian Approach to Real Estate Assessment." In S. E. Feinberg and A. Zellner (eds.), *Studies in Bayesian Econometrics and Statistics in Honor of L. J. Savage.* Amsterdam: North-Holland.

Wonnacott, T. H., and Wonnacott, R. J. (1990). *Introductory Statistics.* 5th ed. New York: Wiley.

Zellner, A. (1986). "Bayesian Estimation and Prediction Using Asymmetric Loss Functions." *Journal of the American Statistical Association,* 81, 446–451.

Putting It All Together: A Forecasting Model with Trend, Seasonal, and Cyclical Components

1. Assembling What We've Learned

Thus far, we've focused on modeling trend, seasonals, and cycles one at a time. In Chapter 5, we introduced models and forecasts of trend. We forecasted retail sales, and we used a model that included only trend. The data were seasonally adjusted, so it wasn't necessary to model seasonality, and, although cycles were likely present, we simply ignored them. In Chapter 6, we introduced models and forecasts of seasonality. We forecasted housing starts, and we used a model that included only seasonal dummies. We didn't need a trend, and again we simply ignored cycles. In Chapters 7–9, we introduced models and forecasts of cycles. We forecasted employment, and we used autoregressive, moving average, and ARMA models. We didn't need trends or seasonals, because employment had no trend and had been seasonally adjusted.

In many forecasting situations, however, more than one component is needed to capture the dynamics in a series to be forecast; frequently they're *all* needed. Here we assemble our tools for forecasting trends, seasonals, and cycles; we use regression on a trend and seasonal dummies, and we capture cyclical dynamics by allowing for ARMA effects in the regression disturbances. The full model is

$$y_t = T_t(\theta) + \sum_{i=1}^{s} \gamma_i D_{it} + \sum_{i=1}^{v_1} \delta_i^{\mathrm{HD}} \mathrm{HDV}_{it} + \sum_{i=1}^{v_2} \delta_i^{\mathrm{TD}} \mathrm{TDV}_{it} + \varepsilon_t$$

$$\Phi(L)\varepsilon_t = \Theta(L)v_t$$

$$\Phi(L) = 1 - \varphi_1 L - \cdots - \varphi_p L^p$$

$$\Theta(L) = 1 + \theta_1 L + \cdots + \theta_q L^q$$

$$v_t \sim WN(0, \sigma^2)\ .$$

$T_t(\theta)$ is a trend, with underlying parameters θ. For example, linear trend has $\theta = \beta_1$ and

$$T_t(\theta) = \beta_1 \mathrm{TIME}_t\ ,$$

and quadratic trend has $\theta = (\beta_1, \beta_2)$ and

$$T_t(\theta) = \beta_1 \mathrm{TIME}_t + \beta_2 \mathrm{TIME}_t^2\ .$$

In addition to the trend, we include seasonal dummies, holiday dummies, and trading-day dummies.[1] The disturbances follow an ARMA(p, q) process, of which pure autoregressions and pure moving averages are special cases. In any particular application, of course, various trend effects, seasonal and other calendar effects, and ARMA cyclical effects may not be needed and so could be dropped.[2] Finally, v_t is the underlying innovation that drives everything.

Now consider constructing an h-step-ahead point forecast at time T, $y_{T+h,T}$. At time $T + h$,

$$y_{T+h} = T_{T+h}(\theta) + \sum_{i=1}^{s} \gamma_i D_{i,T+h} + \sum_{i=1}^{v_1} \delta_i^{\mathrm{HD}} \mathrm{HDV}_{i,T+h} + \sum_{i=1}^{v_2} \delta_i^{\mathrm{TD}} \mathrm{TDV}_{i,T+h} + \varepsilon_{T+h}\ .$$

Projecting the right-hand-side variables on what's known at time T (i.e., the time-T information set, Ω_T), yields the point forecast

$$y_{T+h,T} = T_{T+h}(\theta) + \sum_{i=1}^{s} \gamma_i D_{i,T+h} + \sum_{i=1}^{v_1} \delta_i^{\mathrm{HD}} \mathrm{HDV}_{i,T+h} + \sum_{i=1}^{v_2} \delta_i^{\mathrm{TD}} \mathrm{TDV}_{i,T+h} + \varepsilon_{T+h,T}\ .$$

As with the pure trend and seasonal models discussed earlier, the trend and seasonal variables on the right-hand side are perfectly predictable. The only twist concerns the cyclical behavior that may be lurking in the disturbance term, future values of which don't necessarily project to 0, because the disturbance is not necessarily white noise. Instead, we construct $\varepsilon_{T+h,T}$ using the methods we developed for forecasting cycles.

---

[1] Note that, because we include a full set of seasonal dummies, the trend does not contain an intercept, and we don't include an intercept in the regression.

[2] If the seasonal dummies were dropped, then we'd include an intercept in the regression.

As always, we make the point forecast operational by replacing unknown parameters with estimates, yielding

$$\hat{y}_{T+h,T} = T_{T+h}(\hat{\theta}) + \sum_{i=1}^{s} \hat{\gamma}_i D_{i,T+h} + \sum_{i=1}^{v_1} \hat{\delta}_i^{\mathrm{HD}} \mathrm{HDV}_{i,T+h} + \sum_{i=1}^{v_2} \hat{\delta}_i^{\mathrm{TD}} \mathrm{TDV}_{i,T+h} + \hat{\varepsilon}_{T+h,T}.$$

To construct $\hat{\varepsilon}_{T+h,T}$, in addition to replacing the parameters in the formula for $\varepsilon_{T+h,T}$ with estimates, we replace the unobservable disturbances, the ε_t's, with the observable residuals, the e_t's.

We use our earlier-developed operational expressions for cycle forecast error variances to produce an h-step-ahead interval forecast; it's simply $\hat{y}_{T+h,T} \pm z_{\alpha/2}\hat{\sigma}_h$, where $\hat{\sigma}_h^2$ is the operational estimate of the variance of the error in forecasting ε_{T+h}, and $z_{\alpha/2}$ is the appropriate critical point of the $N(0, 1)$ density. For example, a 95% interval forecast is $\hat{y}_{T+h,T} \pm 1.96\hat{\sigma}_h$. Finally, the complete h-step-ahead density forecast is $N(\hat{y}_{T+h,T}, \hat{\sigma}_h^2)$.

Once again, we don't actually have to *do* any of the computations just discussed; rather, the computer does them all for us. So let's get on with an application, now that we know what we're doing.

2. Application: Forecasting Liquor Sales

We'll forecast monthly U.S. liquor sales. We graphed a short span of the series in Chapter 6 and noted its pronounced seasonality—sales skyrocket during the Christmas season. In Figure 10.1, we show a longer history of liquor sales, 1968.01–1993.12. In Figure 10.2, we show log liquor sales; we take logs to stabilize the variance, which grows over time.[3] The variance of

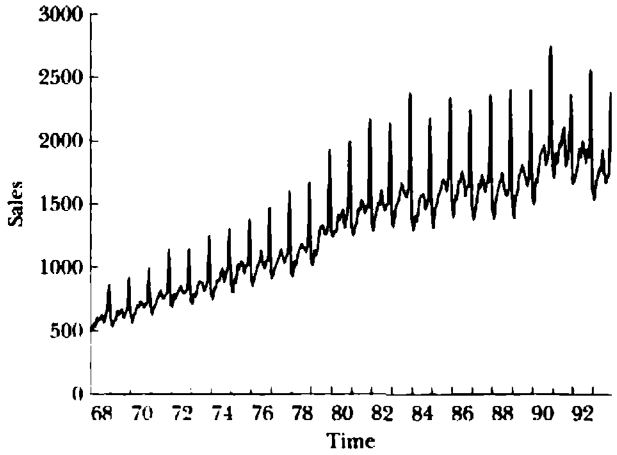
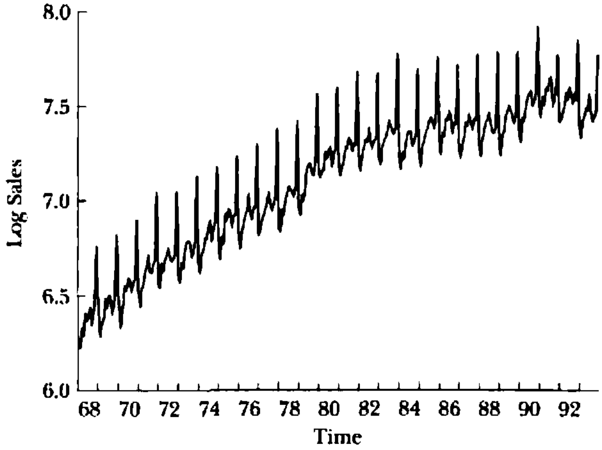


FIGURE 10.1
Liquor Sales, 1968.01–1993.12

---

[3] The nature of the logarithmic transformation is such that it "compresses" an increasing variance. Make a graph of $\log(x)$ as a function of x, and you'll see why.

FIGURE 10.2
Log Liquor Sales, 1968.01–1993.12



log liquor sales is more stable, and it's the series for which we'll build forecasting models.[4]

Liquor sales dynamics also feature prominent trend and cyclical effects. Liquor sales trend upward, and the trend appears nonlinear in spite of the fact that we're working in logs. To handle the nonlinear trend, we adopt a quadratic trend model (in logs). The estimation results are in Table 10.1. The residual plot (Figure 10.3) shows that the fitted trend increases at a decreasing rate; both the linear and quadratic terms are highly significant. The adjusted R^2 is 89%, reflecting the fact that trend is responsible for a large part of the variation in liquor sales. The standard error of the regression is 0.125; it's an estimate of the standard deviation of the error we'd expect to make in forecasting liquor sales if we accounted for trend but ignored seasonality and serial correlation. The Durbin-Watson statistic provides no evidence against the hypothesis that the regression disturbance is white noise.

The residual plot, however, shows obvious residual seasonality. The Durbin-Watson statistic missed it, evidently because it's not designed to have power against seasonal dynamics.[5] The residual plot also suggests that there may be a cycle in the residual, although it's hard to tell (hard for the Durbin-Watson statistic as well), because the pervasive seasonality swamps the picture and makes it hard to infer much of anything.

The residual correlogram (Table 10.2) and its graph (Figure 10.4) confirm the importance of the neglected seasonality. The residual sample autocorrelation function has large spikes, far exceeding the Bartlett bands, at

---

[4] From this point onward, for brevity we'll simply refer to "liquor sales," but remember that we've taken logs.

[5] Recall that the Durbin-Watson test is designed to detect simple AR(1) dynamics. It also has the ability to detect other sorts of dynamics, but evidently not those relevant to the present application, which are very different from a simple AR(1).

TABLE 10.1
Log Liquor Sales Quadratic Trend Regression

LS // Dependent variable is LSALES.
Sample: 1968:01 1993:12
Included observations: 312

Variable	Coefficient	Std. Error	*t*-Statistic	Prob.
C	6.237356	0.024496	254.6267	0.0000
TIME	0.007690	0.000336	22.91552	0.0000
TIME2	−1.14E-05	9.74E-07	−11.72695	0.0000

R^2	0.892394	Mean dependent var.	7.112383
Adjusted R^2	0.891698	SD dependent var.	0.379308
SE of regression	0.124828	Akaike info criterion	−4.152073
Sum squared resid.	4.814823	Schwarz criterion	−4.116083
Log likelihood	208.0146	*F*-statistic	1281.296
Durbin-Watson stat.	1.752858	Prob(*F*-statistic)	0.000000

the seasonal displacements 12, 24, and 36. It indicates some cyclical dynamics as well; apart from the seasonal spikes, the residual sample autocorrelation and partial autocorrelation functions oscillate, and the Ljung-Box statistic rejects the white noise null hypothesis even at very small, nonseasonal, displacements.

In Table 10.3 we show the results of regression on quadratic trend and a full set of seasonal dummies. The quadratic trend remains highly significant. The adjusted R^2 rises to 99%, and the standard error of the regression falls to 0.046, which is an estimate of the standard deviation of the forecast error we expect to make if we account for trend and seasonality but ignore serial correlation. The Durbin-Watson statistic, however, has greater ability to detect serial

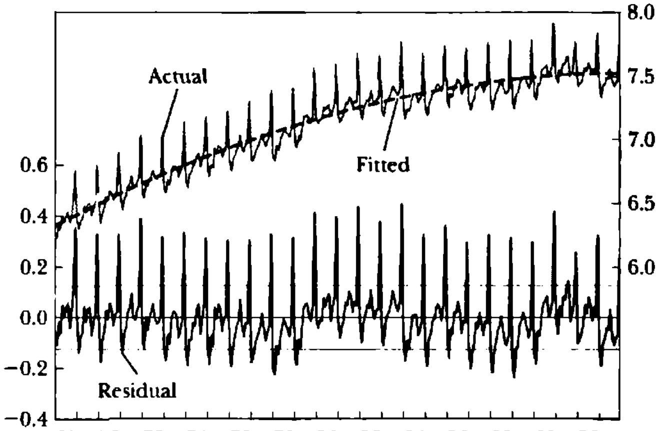


FIGURE 10.3
Log Liquor Sales Quadratic Trend Regression, Residual Plot

TABLE 10.2
Log Liquor Sales Quadratic Trend Regression, Residual Correlogram

	Acorr.	P. Acorr.	Std. Error	Ljung-Box	*p*-value
1	0.117	0.117	.056	4.3158	0.038
2	−0.149	−0.165	.056	11.365	0.003
3	−0.106	−0.069	.056	14.943	0.002
4	−0.014	−0.017	.056	15.007	0.005
5	0.142	0.125	.056	21.449	0.001
6	0.041	−0.004	.056	21.979	0.001
7	0.134	0.175	.056	27.708	0.000
8	−0.029	−0.046	.056	27.975	0.000
9	−0.136	−0.080	.056	33.944	0.000
10	−0.205	−0.206	.056	47.611	0.000
11	0.056	0.080	.056	48.632	0.000
12	0.888	0.879	.056	306.26	0.000
13	0.055	−0.507	.056	307.25	0.000
14	−0.187	−0.159	.056	318.79	0.000
15	−0.159	−0.144	.056	327.17	0.000
16	−0.059	−0.002	.056	328.32	0.000
17	0.091	−0.118	.056	331.05	0.000
18	−0.010	−0.055	.056	331.08	0.000
19	0.086	−0.032	.056	333.57	0.000
20	−0.066	0.028	.056	335.03	0.000
21	−0.170	0.044	.056	344.71	0.000
22	−0.231	0.180	.056	362.74	0.000
23	0.028	0.016	.056	363.00	0.000
24	0.811	−0.014	.056	586.50	0.000
25	0.013	−0.128	.056	586.56	0.000
26	−0.221	−0.136	.056	603.26	0.000
27	−0.196	−0.017	.056	616.51	0.000
28	−0.092	−0.079	.056	619.42	0.000
29	0.045	−0.094	.056	620.13	0.000
30	−0.043	0.045	.056	620.77	0.000
31	0.057	0.041	.056	621.89	0.000
32	−0.095	−0.002	.056	625.07	0.000
33	−0.195	0.026	.056	638.38	0.000
34	−0.240	0.088	.056	658.74	0.000
35	0.006	−0.089	.056	658.75	0.000
36	0.765	0.076	.056	866.34	0.000

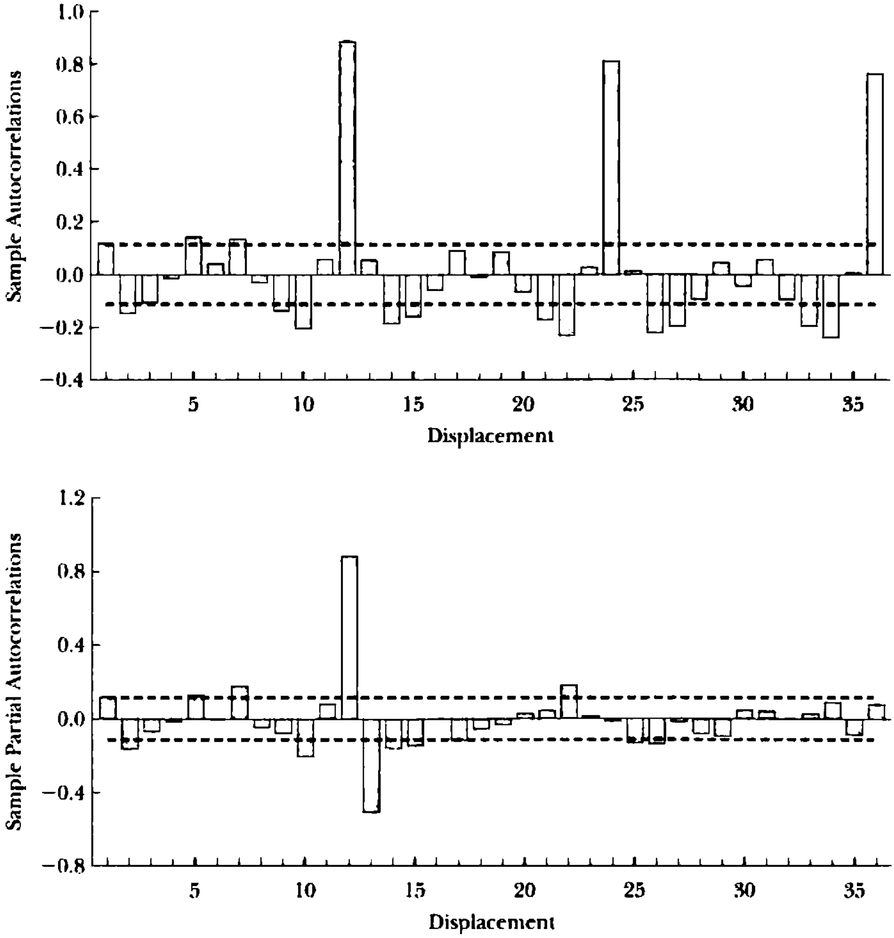


FIGURE 10.4 *Log Liquor Sales Quadratic Trend Regression, Residual Sample Autocorrelation and Partial Autocorrelation Functions*

correlation now that the residual seasonality has been accounted for, and it sounds a loud alarm.

The residual plot of Figure 10.5 shows no seasonality, as that's now picked up by the model, but it confirms the Durbin-Watson's warning of serial correlation. The residuals are highly persistent and hence predictable. We show the residual correlogram in tabular and graphical form in Table 10.4 and Figure 10.6. The residual sample autocorrelations oscillate and decay slowly, and they exceed the Bartlett standard errors throughout. The Ljung-Box test strongly rejects the white noise null at all displacements. Finally, the residual sample partial autocorrelations cut off at displacement 3. All of this suggests that an AR(3) would provide a good approximation to the disturbance's Wold representation.

TABLE 10.3
Log Liquor Sales Quadratic Trend Regression with Seasonal Dummies

LS // Dependent variable is LSALES.
Sample: 1968:01 1993:12
Included observations: 312

Variable	Coefficient	Std. Error	*t*-Statistic	Prob.
TIME	0.007656	0.000123	62.35882	0.0000
TIME2	−1.14E-05	3.56E-07	−32.06823	0.0000
*D*1	6.147456	0.012340	498.1699	0.0000
*D*2	6.088653	0.012353	492.8890	0.0000
*D*3	6.174127	0.012366	499.3008	0.0000
*D*4	6.175220	0.012378	498.8970	0.0000
*D*5	6.246086	0.012390	504.1398	0.0000
*D*6	6.250387	0.012401	504.0194	0.0000
*D*7	6.295979	0.012412	507.2402	0.0000
*D*8	6.268043	0.012423	504.5509	0.0000
*D*9	6.203832	0.012433	498.9630	0.0000
*D*10	6.229197	0.012444	500.5968	0.0000
*D*11	6.259770	0.012453	502.6602	0.0000
*D*12	6.580068	0.012463	527.9819	0.0000

R^2	0.986111	Mean dependent var.	7.112383
Adjusted R^2	0.985505	SD dependent var.	0.379308
SE of regression	0.045666	Akaike info criterion	−6.128963
Sum squared resid.	0.621448	Schwarz criterion	−5.961008
Log likelihood	527.4094	*F*-statistic	1627.567
Durbin-Watson stat.	0.586187	Prob(*F*-statistic)	0.000000

FIGURE 10.5
Log Liquor Sales Quadratic Trend Regression with Seasonal Dummies, Residual Plot

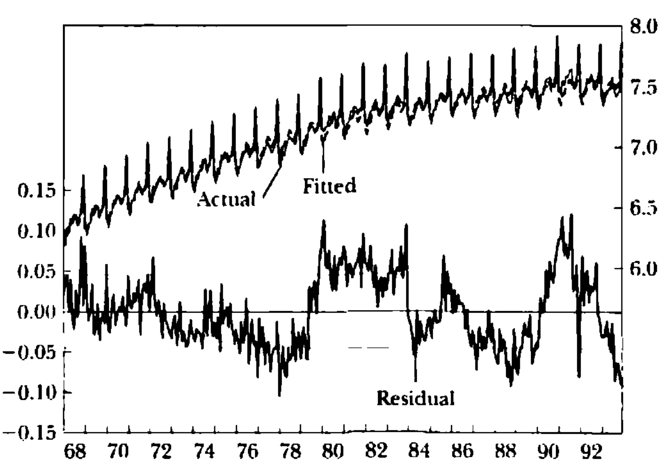
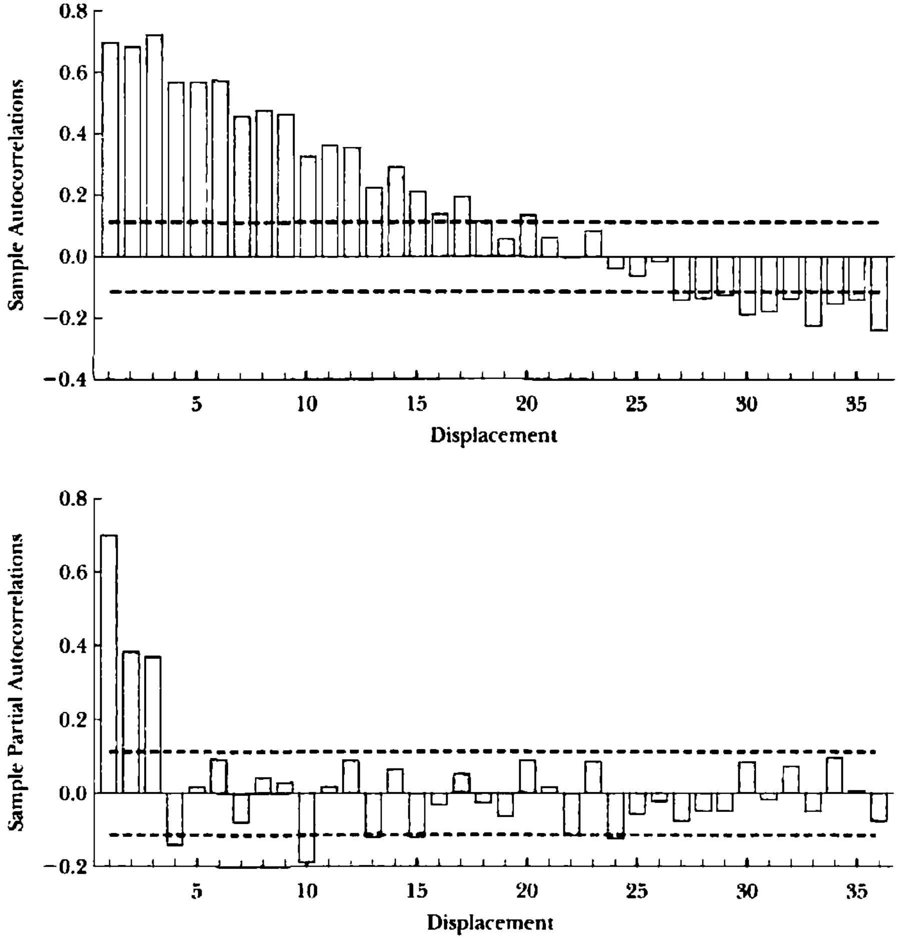


TABLE 10.4
Log Liquor Sales Quadratic Trend Regression with Seasonal Dummies, Residual Correlogram

	Acorr.	P. Acorr.	Std. Error	Ljung-Box	*p*-value
1	0.700	0.700	.056	154.34	0.000
2	0.686	0.383	.056	302.86	0.000
3	0.725	0.369	.056	469.36	0.000
4	0.569	−0.141	.056	572.36	0.000
5	0.569	0.017	.056	675.58	0.000
6	0.577	0.093	.056	782.19	0.000
7	0.460	−0.078	.056	850.06	0.000
8	0.480	0.043	.056	924.38	0.000
9	0.466	0.030	.056	994.46	0.000
10	0.327	−0.188	.056	1029.1	0.000
11	0.364	0.019	.056	1072.1	0.000
12	0.355	0.089	.056	1113.3	0.000
13	0.225	−0.119	.056	1129.9	0.000
14	0.291	0.065	.056	1157.8	0.000
15	0.211	−0.119	.056	1172.4	0.000
16	0.138	−0.031	.056	1178.7	0.000
17	0.195	0.053	.056	1191.4	0.000
18	0.114	−0.027	.056	1195.7	0.000
19	0.055	−0.063	.056	1196.7	0.000
20	0.134	0.089	.056	1202.7	0.000
21	0.062	0.018	.056	1204.0	0.000
22	−0.006	−0.115	.056	1204.0	0.000
23	0.084	0.086	.056	1206.4	0.000
24	−0.039	−0.124	.056	1206.9	0.000
25	−0.063	−0.055	.056	1208.3	0.000
26	−0.016	−0.022	.056	1208.4	0.000
27	−0.143	−0.075	.056	1215.4	0.000
28	−0.135	−0.047	.056	1221.7	0.000
29	−0.124	−0.048	.056	1227.0	0.000
30	−0.189	0.086	.056	1239.5	0.000
31	−0.178	−0.017	.056	1250.5	0.000
32	−0.139	0.073	.056	1257.3	0.000
33	−0.226	−0.049	.056	1275.2	0.000
34	−0.155	0.097	.056	1283.7	0.000
35	−0.142	0.008	.056	1290.8	0.000
36	−0.242	−0.074	.056	1311.6	0.000

FIGURE 10.6
Log Liquor Sales Quadratic Trend Regression with Seasonal Dummies, Residual Sample Autocorrelation and Partial Autocorrelation Functions



In Table 10.5, then, we report the results of estimating a liquor sales model with quadratic trend, seasonal dummies, and AR(3) disturbances. The R^2 is now 100%, and the Durbin-Watson is fine. One inverse root of the AR(3) disturbance process is estimated to be real and close to the unit circle (0.95), and the other two inverse roots are a complex conjugate pair farther from the unit circle. The standard error of this regression is an estimate of the standard deviation of the forecast error we'd expect to make after modeling the residual serial correlation, as we've now done; that is, it's an estimate of the standard deviation of v.[6] It's a very small 0.027, roughly half that obtained when we ignored serial correlation.

---

[6] Recall that v is the innovation that drives the ARMA process for the regression disturbance, ε.

TABLE 10.5
Log Liquor Sales Quadratic Trend Regression with Seasonal Dummies and AR(3) Disturbances

LS // Dependent variable is LSALES.
Sample: 1968:01 1993:12
Included observations: 312
Convergence achieved after 4 iterations

Variable	Coefficient	Std. Error	*t*-Statistic	Prob.
TIME	0.008606	0.000981	8.768212	0.0000
TIME2	−1.41E-05	2.53E-06	−5.556103	0.0000
*D*1	6.073054	0.083922	72.36584	0.0000
*D*2	6.013822	0.083942	71.64254	0.0000
*D*3	6.099208	0.083947	72.65524	0.0000
*D*4	6.101522	0.083934	72.69393	0.0000
*D*5	6.172528	0.083946	73.52962	0.0000
*D*6	6.177129	0.083947	73.58364	0.0000
*D*7	6.223323	0.083939	74.14071	0.0000
*D*8	6.195681	0.083943	73.80857	0.0000
*D*9	6.131818	0.083940	73.04993	0.0000
*D*10	6.157592	0.083934	73.36197	0.0000
*D*11	6.188480	0.083932	73.73176	0.0000
*D*12	6.509106	0.083928	77.55624	0.0000
AR(1)	0.268805	0.052909	5.080488	0.0000
AR(2)	0.239688	0.053697	4.463723	0.0000
AR(3)	0.395880	0.053109	7.454150	0.0000

R^2	0.995069	Mean dependent var.	7.112383
Adjusted R^2	0.994802	SD dependent var.	0.379308
SE of regression	0.027347	Akaike info criterion	−7.145319
Sum squared resid.	0.220625	Schwarz criterion	−6.941373
Log likelihood	688.9610	*F*-statistic	3720.875
Durbin-Watson stat.	1.886119	Prob(*F*-statistic)	0.000000
Inverted AR roots	.95	−.34 + .55i	−.34 − .55i

We show the residual plot in Figure 10.7 and the residual correlogram in Table 10.6 and Figure 10.8. The residual plot reveals no patterns; instead, the residuals look like white noise, as they should. The residual sample autocorrelations and partial autocorrelations display no patterns and are mostly inside the Bartlett bands. The Ljung-Box statistics also look good for small and moderate displacements, although their *p*-values decrease for longer displacements.

All things considered, the quadratic trend, seasonal dummy, AR(3) specification seems tentatively adequate. We also perform a number of additional checks. In Figure 10.9, we show a histogram and normality test applied to the residuals. The histogram looks symmetric, as confirmed by the skewness

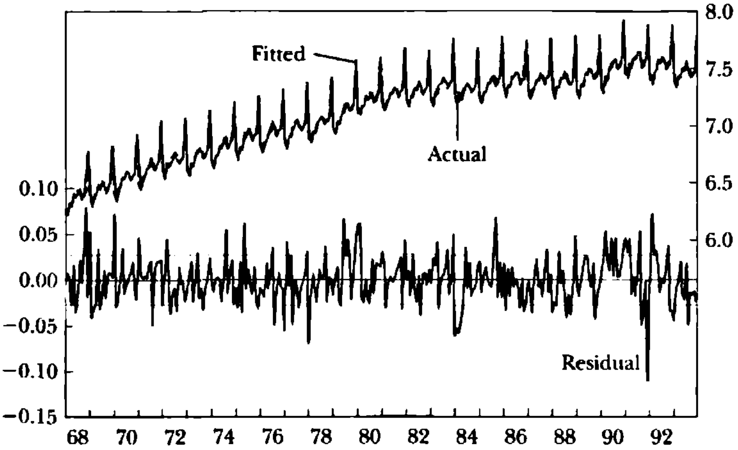


FIGURE 10.7
Log Liquor Sales Quadratic Trend Regression with Seasonal Dummies and AR(3) Disturbances, Residual Plot

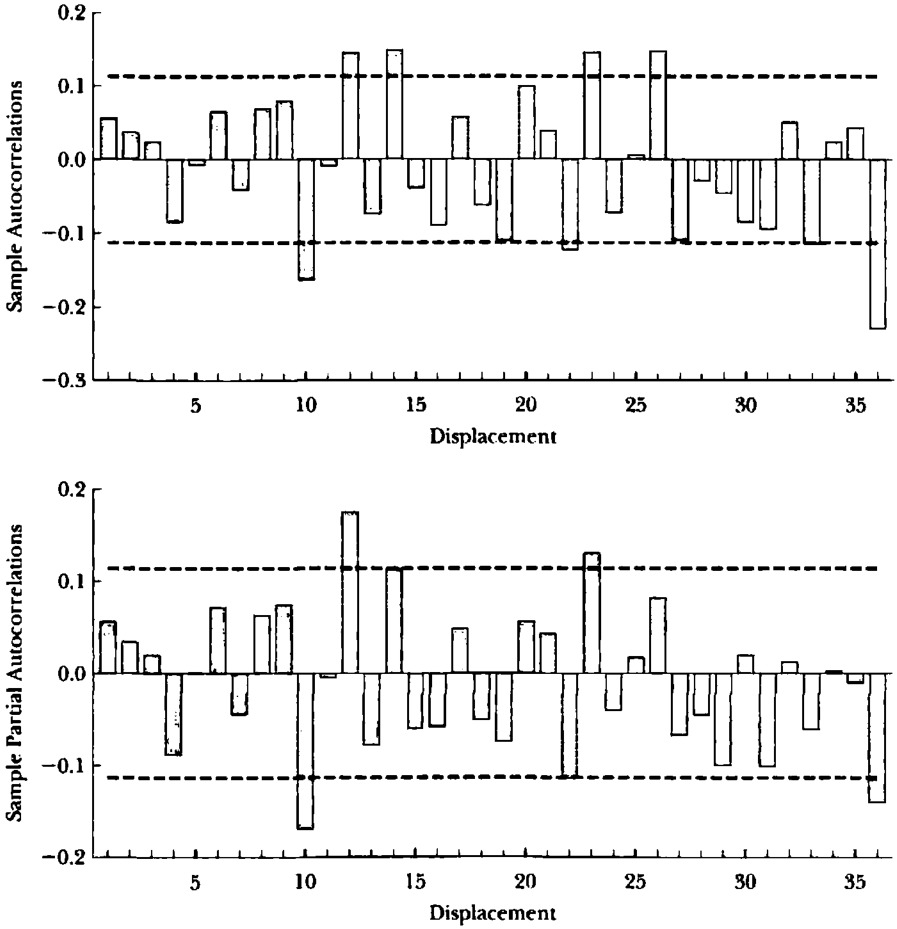
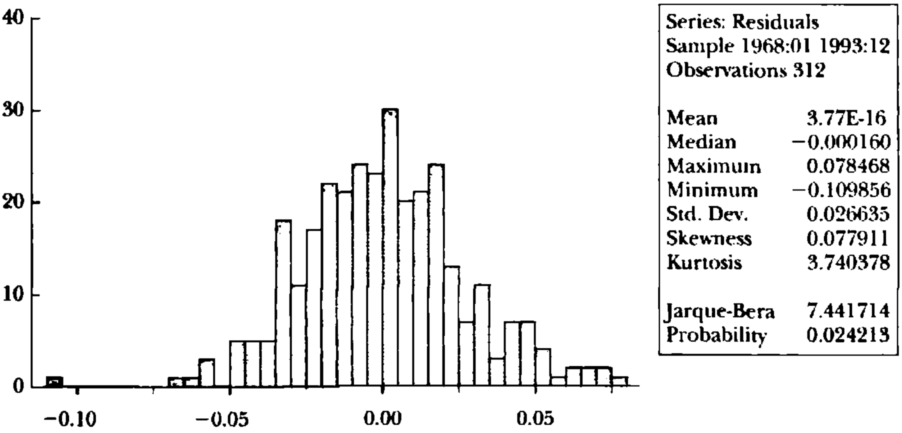


FIGURE 10.8
Log Liquor Sales Quadratic Trend Regression with Seasonal Dummies and AR(3) Disturbances, Residual Sample Autocorrelation and Partial Autocorrelation Functions

TABLE 10.8
Log Liquor Sales Quadratic Trend Regression with Seasonal Dummies and AR(3) Disturbances, Residual Correlogram

	Acorr.	P. Acorr.	Std. Error	Ljung-Box	*p*-value
1	0.056	0.056	.056	0.9779	0.323
2	0.037	0.034	.056	1.4194	0.492
3	0.024	0.020	.056	1.6032	0.659
4	−0.084	−0.088	.056	3.8256	0.430
5	−0.007	0.001	.056	3.8415	0.572
6	0.065	0.072	.056	5.1985	0.519
7	−0.041	−0.044	.056	5.7288	0.572
8	0.069	0.063	.056	7.2828	0.506
9	0.080	0.074	.056	9.3527	0.405
10	−0.163	−0.169	.056	18.019	0.055
11	−0.009	−0.005	.056	18.045	0.081
12	0.145	0.175	.056	24.938	0.015
13	−0.074	−0.078	.056	26.750	0.013
14	0.149	0.113	.056	34.034	0.002
15	−0.039	−0.060	.056	34.532	0.003
16	−0.089	−0.058	.056	37.126	0.002
17	0.058	0.048	.056	38.262	0.002
18	−0.062	−0.050	.056	39.556	0.002
19	−0.110	−0.074	.056	43.604	0.001
20	0.100	0.056	.056	46.935	0.001
21	0.039	0.042	.056	47.440	0.001
22	−0.122	−0.114	.056	52.501	0.000
23	0.146	0.130	.056	59.729	0.000
24	−0.072	−0.040	.056	61.487	0.000
25	0.006	0.017	.056	61.500	0.000
26	0.148	0.082	.056	69.024	0.000
27	−0.109	−0.067	.056	73.145	0.000
28	−0.029	−0.045	.056	73.436	0.000
29	−0.046	−0.100	.056	74.153	0.000
30	−0.084	0.020	.056	76.620	0.000
31	−0.095	−0.101	.056	79.793	0.000
32	0.051	0.012	.056	80.710	0.000
33	−0.114	−0.061	.056	85.266	0.000
34	0.024	0.002	.056	85.468	0.000
35	0.043	−0.010	.056	86.116	0.000
36	−0.229	−0.140	.056	104.75	0.000

FIGURE 10.9 *Log Liquor Sales Quadratic Trend Regression with Seasonal Dummies and AR(3) Disturbances, Residual Histogram and Normality Test*



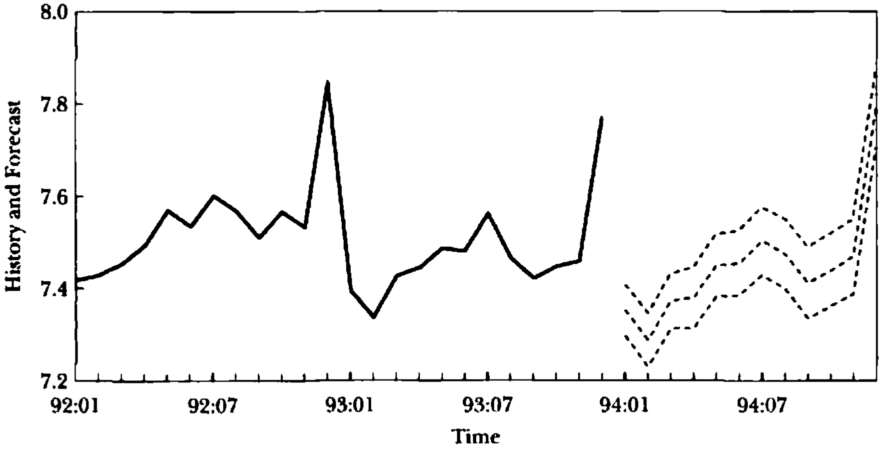
Series: Residuals
Sample 1968:01 1993:12
Observations 312

Mean	3.77E-16
Median	−0.000160
Maximum	0.078468
Minimum	−0.109856
Std. Dev.	0.026635
Skewness	0.077911
Kurtosis	3.740378
Jarque-Bera	7.441714
Probability	0.024213

near 0. The residual kurtosis is a bit higher than 3 and causes the Jarque-Bera test to reject the normality hypothesis with a *p*-value of .02, but the residuals nevertheless appear to be fairly well approximated by a normal distribution, even if they may have slightly fatter tails.

Now we use the estimated model to produce forecasts. In Figure 10.10 we show the history of liquor sales and a 12-month-ahead extrapolation forecast for 1994.[7] To aid visual interpretation, we show only 2 years of history. The

FIGURE 10.10 *Log Liquor Sales History and 12-Month-Ahead Forecast*



---

[7] We show the point forecast together with 95% intervals.

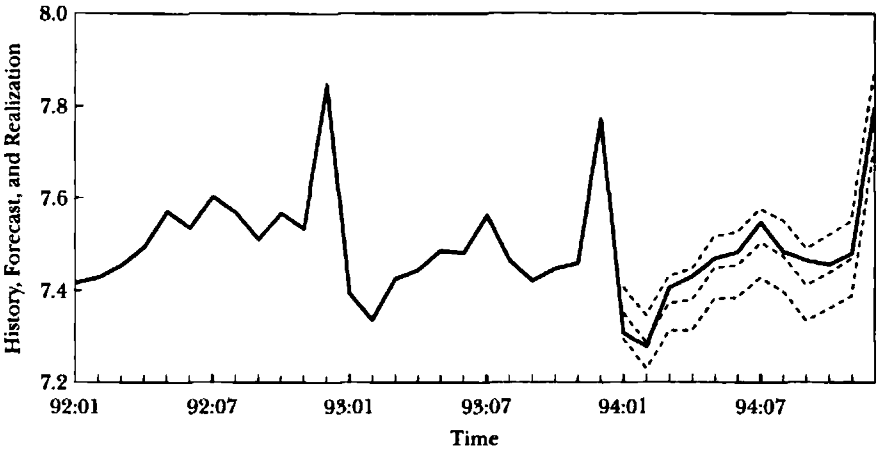


FIGURE 10.11 *Log Liquor Sales History, 12-Month-Ahead Forecast, and Realization*

forecast looks reasonable. It's visually apparent that the model has done a good job of picking up the seasonal pattern, which dominates the local behavior of the series. In Figure 10.11, we show the history, the forecast, and the 1994 realization. The forecast was very good!

Figure 10.12 shows 4 years of history together with a 60-month-ahead (5-year) extrapolation forecast, to provide a better feel for the dynamics in the forecast. The figure also makes clear the trend forecast is slightly *downward.* To

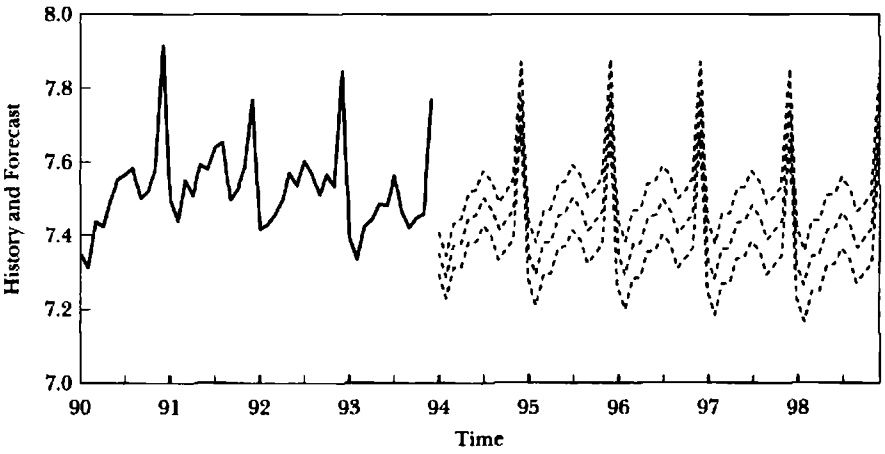
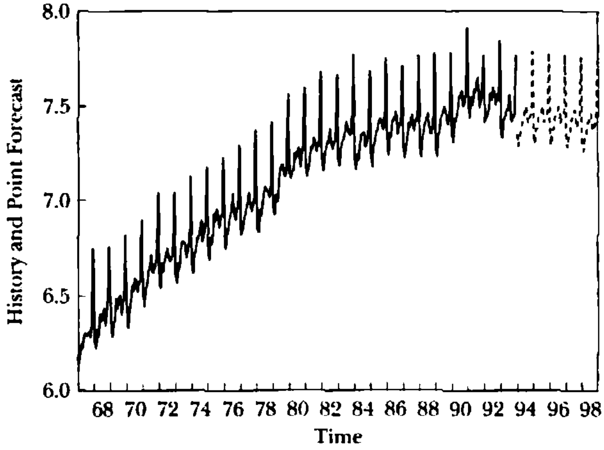


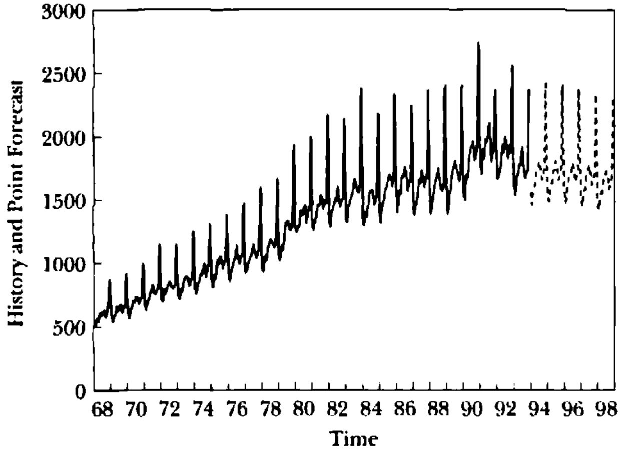
FIGURE 10.12 *Log Liquor Sales History and 60-Month-Ahead Forecast*

FIGURE 10.13
Log Liquor Sales Long History and 60-Month-Ahead Forecast



put the long-horizon forecast in historical context, we show in Figure 10.13 the 60-month-ahead forecast together with the complete history. Finally, in Figure 10.14, we show the history and point forecast of the level of liquor sales (as opposed to log liquor sales), which we obtain by exponentiating the forecast of log liquor sales.[8]

FIGURE 10.14
Liquor Sales Long History and 60-Month-Ahead Forecast



---

[8] Recall that exponentiating "undoes" a natural logarithm.

3. Recursive Estimation Procedures for Diagnosing and Selecting Forecasting Models

Recursive estimation means beginning with a small sample of data, estimating a model, adding an observation and reestimating the model, and continuing in that fashion until the sample is exhausted.[9] Recursive estimation and related techniques are useful in a variety of situations of importance in forecasting, including stability assessment and model selection. On both counts, it's natural to introduce them now.

ASSESSING THE STABILITY OF FORECASTING MODELS: RECURSIVE PARAMETER ESTIMATION AND RECURSIVE RESIDUALS

Business and economic relationships often vary over time; sometimes parameters evolve slowly, and sometimes they break sharply. If a forecasting model displays such instability, it's not likely to produce good forecasts, so it's important that we have tools that help us to diagnose the instability. Recursive estimation procedures allow us to assess and track time-varying parameters and are therefore useful in the construction and evaluation of a variety of forecasting models.

First we introduce the idea of recursive parameter estimation. We work with the standard linear regression model.

$$y_t = \sum_{i=1}^{k} \beta_i x_{i,t} + \varepsilon_t$$

$$\varepsilon_t \overset{\text{iid}}{\sim} N(0, \sigma^2),$$

$t = 1, \ldots, T$, and we estimate it using least squares. Instead of immediately using all the data to estimate the model, however, we begin with a small subset. If the model contains k parameters, begin with the first k observations and estimate the model. Then we estimate it using the first $k + 1$ observations and so on, until the sample is exhausted. At the end we have a set of recursive parameter estimates $\hat{\beta}_{i,t}$, for $t = k, \ldots, T$ and $i = 1, \ldots, k$. It often pays to compute and examine recursive estimates, because they convey important information about parameter stability—they show how the estimated parameters move as more and more observations are accumulated. It's often informative to plot the recursive estimates, to help answer the obvious questions of interest: Do the coefficient estimates stabilize as the sample size grows? Or do they wander around, or drift in a particular direction, or break sharply at one or more points?

---

[9] Strictly speaking, *sequential* might be a more descriptive adjective than *recursive*. *Recursive updating* refers to the fact that an estimate based on $t + 1$ observations can sometimes be computed simply by appropriately combining the old estimate based on t observations with the new observation. (This is possible, for example, with linear least-squares regression.) Recursive updating achieves a drastic reduction in computational requirements relative to complete reestimation of the model each time the sample is updated, which we might call "brute force updating." For our purposes, it's inconsequential whether we do recursive updating or brute force updating (and the speed of modern computers often makes brute force attractive); we use *recursive estimation* as a blanket term for any sequential estimation procedure, whether the computations are done by recursive or brute force techniques.

Now let's introduce the **recursive residuals.** At each t, $t = k, \ldots, T-1$, we can compute a 1-step-ahead forecast, $\hat{y}_{t+1,t} = \sum_{i=1}^{k} \hat{\beta}_{i,t} x_{i,t+1}$. The corresponding forecast errors, or recursive residuals, are $\hat{e}_{t+1,t} = y_{t+1} - \hat{y}_{t+1,t}$. The variance of these 1-step-ahead forecast errors changes as the sample size grows, because under the maintained assumptions the model parameters are estimated more precisely as the sample size grows. Specifically,

$$\hat{e}_{t+1,t} \sim N(0, \sigma^2 r_t) ,$$

where $r_t > 1$ for all t and r_t is a somewhat complicated function of the data.[10]

As with recursive parameter estimates, recursive residuals can reveal **parameter instability** in forecasting models. Often we'll examine a plot of the recursive residuals and estimated two standard-error bands $(\pm 2\hat{\sigma}\sqrt{r_t})$.[11] This has an immediate forecasting interpretation and is sometimes called a sequence of 1-step forecast tests—we make recursive 1-step-ahead 95% interval forecasts and then check where the subsequent realizations fall. If many of them fall outside the intervals, one or more parameters may be unstable, and the locations of the violations of the interval forecasts give some indication as to the nature of the instability.

Sometimes it's helpful to consider the **standardized recursive residuals,**

$$w_{t+1,t} \equiv \frac{\hat{e}_{t+1,t}}{\sigma\sqrt{r_t}} ,$$

$t = k, \ldots, T-1$. Under the maintained assumptions,

$$w_{t+1,t} \overset{\text{iid}}{\sim} N(0, 1) .$$

If any of the maintained model assumptions are violated, the standardized recursive residuals will fail to be iid normal, so we can learn about various model inadequacies by examining them. The cumulative sum ("**CUSUM**") of the standardized recursive residuals is particularly useful in assessing parameter stability. Because $w_{t+1,t} \overset{\text{iid}}{\sim} N(0, 1)$, it follows that

---

[10] Derivation of a formula for r_t is beyond the scope of this book. Ordinarily we'd ignore the inflation of $\text{var}(\hat{e}_{t+1,t})$ due to parameter estimation, which vanishes with sample size so that $r_t \to 1$, and simply use the large-sample approximation $\hat{e}_{t+1,t} \sim N(0, \sigma^2)$. Presently, however, we're estimating the regression recursively, so the initial regressions will always be performed on very small samples, thereby rendering large-sample approximations unpalatable.

[11] $\hat{\sigma}$ is just the usual standard error of the regression, estimated from the full sample of data.

$$\text{CUSUM}_t \equiv \sum_{\tau=k}^{t} w_{\tau+1,\tau}, \quad t = k, \ldots, T-1$$

is just a sum of iid $N(0, 1)$ random variables.[12] Probability bounds for the CUSUM have been tabulated, and we often examine time series plots of the CUSUM and its 95% probability bounds, which grow linearly and are centered at 0.[13] If the CUSUM violates the bounds at any point, there is evidence of parameter instability. Such an analysis is called a *CUSUM analysis.*

As an illustration of the use of recursive techniques for detecting structural change, we consider in Figures 10.15 and 10.16 two stylized data-generating processes (bivariate regression models, satisfying the classical assumptions apart from the possibility of a time-varying parameter). The first has a constant parameter, and the second has a sharply breaking parameter. For each we show a scatterplot of y versus x, recursive parameter estimates, recursive residuals, and a **CUSUM plot.**

We show the constant parameter model in Figure 10.15. As expected, the scatterplot shows no evidence of instability, the recursive parameter estimate stabilizes quickly, its variance decreases quickly, the recursive residuals look like zero-mean random noise, and the CUSUM plot shows no evidence of instability.

We show the breaking parameter model in Figure 10.16; the results are different yet again. The true relationship between y and x is one of proportionality, with the constant of proportionality jumping in midsample. The jump is clearly evident in the scatterplot, in the recursive residuals, and in the recursive parameter estimate. The CUSUM remains near 0 until midsample, at which time it shoots through the 95% probability limit.

Model Selection Based on Simulated Forecasting Performance

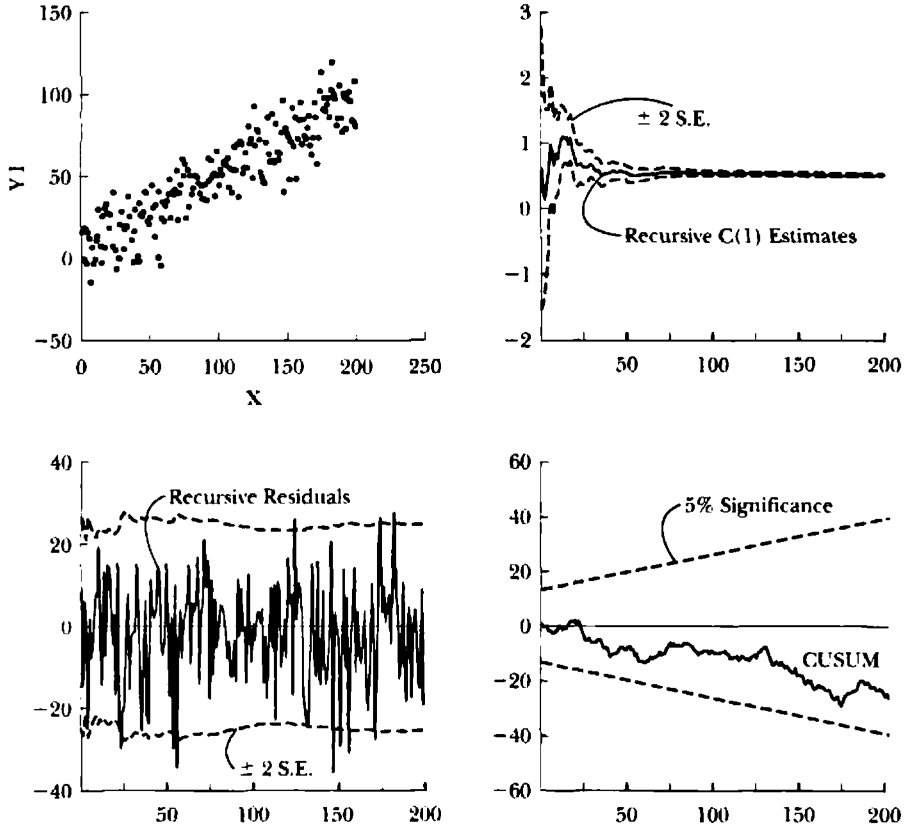
All the forecast model selection strategies that we've studied amount to strategies for finding the model that's most likely to perform well in terms of out-of-sample 1-step-ahead mean-squared forecast error. In every case, we effectively estimate *out-of-sample* 1-step-ahead mean-squared forecast error by adjusting the *in-sample* mean-squared error with a degrees-of-freedom penalty. The important insight is that we estimate out-of-sample forecast accuracy using

---

[12] Sums of zero-mean iid random variables are very important. In fact, they're so important that they have their own name, **random walks.** We'll study them in detail in Chapter 13.

[13] To make the standardized recursive residuals, and hence the CUSUM statistic, operational, we replace σ with $\hat{\sigma}$.

FIGURE 10.15
Recursive Analysis Constant Parameter Model



in-sample residuals. Recursive estimation suggests a different approach, which is also more direct and flexible. Recursive estimation lets us estimate out-of-sample forecast accuracy directly, using out-of-sample forecast errors.

We first introduce a procedure called **cross validation,** in reference to the fact that the predictive ability of the model is evaluated on observations different from those on which the model is estimated, thereby incorporating an automatic degrees-of-freedom penalty. It's actually not based on recursive estimation, because we don't let the estimation sample expand. Instead, we obtain the various estimation samples by sequentially deleting observations. As we'll see, however, it provides a natural introduction to a closely related recursive model selection procedure that we'll introduce subsequently, which we call *recursive cross validation.*

Cross validation proceeds as follows. Consider selecting among J forecasting models. Start with model 1, estimate it using all data observations except the first, use it to forecast the first observation, and compute the associated

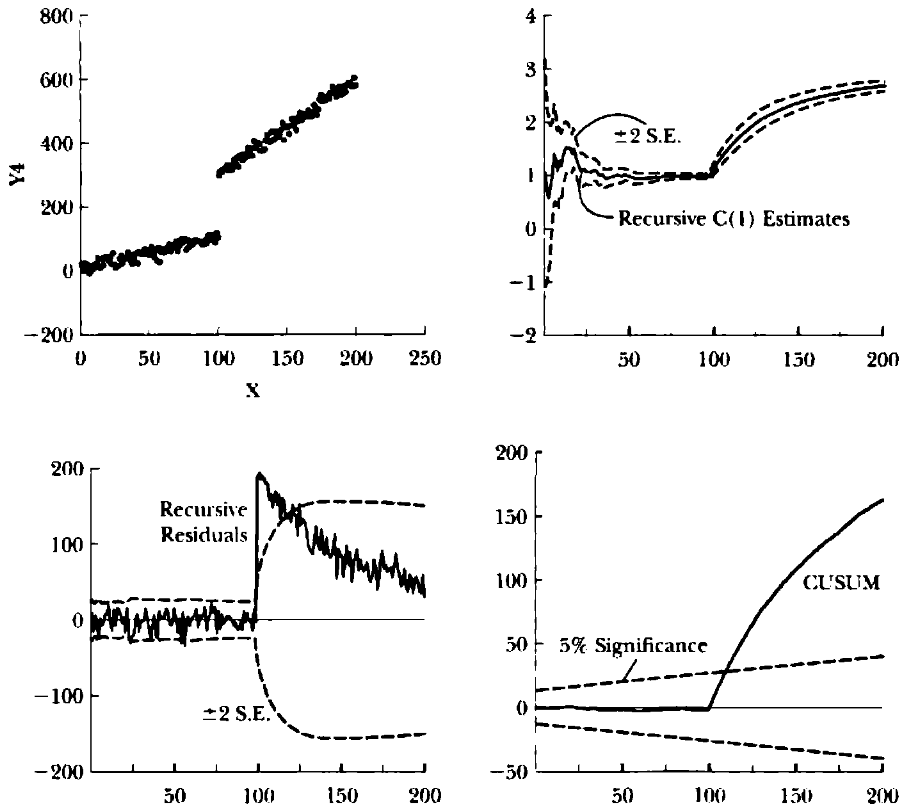


FIGURE 10.16
Recursive Analysis Breaking Parameter Model

squared forecast error. Then estimate it using all observations except the second, use it to forecast the second observation, and compute the associated squared error. Keep doing this—estimating the model with one observation deleted and then using the estimated model to forecast the deleted observation—until each observation has been sequentially deleted, and average the squared errors in predicting each of the T sequentially deleted observations. Repeat the procedure for the other models, $j = 2, \ldots, J$, and select the model with the smallest average squared forecast error.

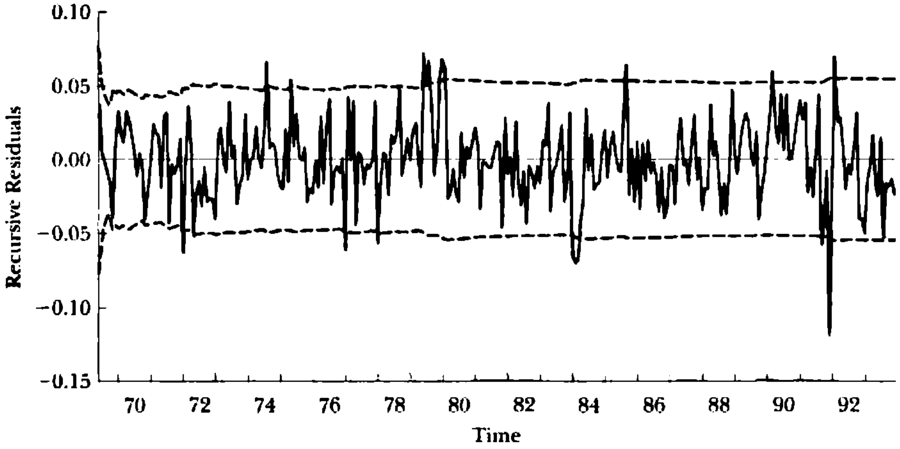
As we've described it here, cross validation is mainly of use in cross section, as opposed to time series, forecasting environments, because the "leave one out" estimations required for cross validation only make sense in the absence of dynamics. That is, it's only in the absence of dynamics that we can simply pluck out an observation, discard it, and proceed to estimate the model with the remaining observations without further adjustment. It's easy to extend the basic idea of cross validation to the time series case, however, which leads to the idea of recursive cross validation.

Recursive cross validation proceeds as follows. Let the initial estimation sample run from $t = 1, \ldots, T^*$, and let the "holdout sample" used for comparing predictive performance run from $t = T^* + 1, \ldots, T$. For each model, proceed as follows. Estimate the model using observations $t = 1, \ldots, T^*$. Then use it to forecast observation $T^* + 1$, and compute the associated squared error. Next, update the sample by one observation (observation $T^* + 1$), estimate the model using the updated sample $t = 1, \ldots, T^* + 1$, forecast observation $T^* + 2$, and compute the associated squared error. Continue this recursive reestimation and forecasting until the sample is exhausted, and then average the squared errors in predicting observations $T^* + 1$ through T. Select the model with the smallest average squared forecast error.

4. Liquor Sales, Continued

In Figures 10.17–10.19, we show the results of a recursive analysis. Figure 10.17 presents the recursive residuals and their two-standard-error bands under the joint null hypothesis of correct specification and parameter constancy. The recursive residuals rarely violate the 95% bands. Figure 10.18 shows the recursive parameter estimates together with recursively computed standard errors. The top row shows the 2 trend parameters, the next three rows show the 12 seasonal dummy parameters, and the last row shows the 3 autoregressive parameters. All parameter estimates seem to stabilize as the sample size grows. Finally, Figure 10.19 is a CUSUM chart, which reveals no evidence against the hypothesis of correct specification and structural stability; the CUSUM never even approaches the 5% significance boundary.

FIGURE 10.17
Log Liquor Sales Quadratic Trend Regression with Seasonal Dummies and AR(3) Disturbances, Recursive Residuals and Two-Standard-Error Bands



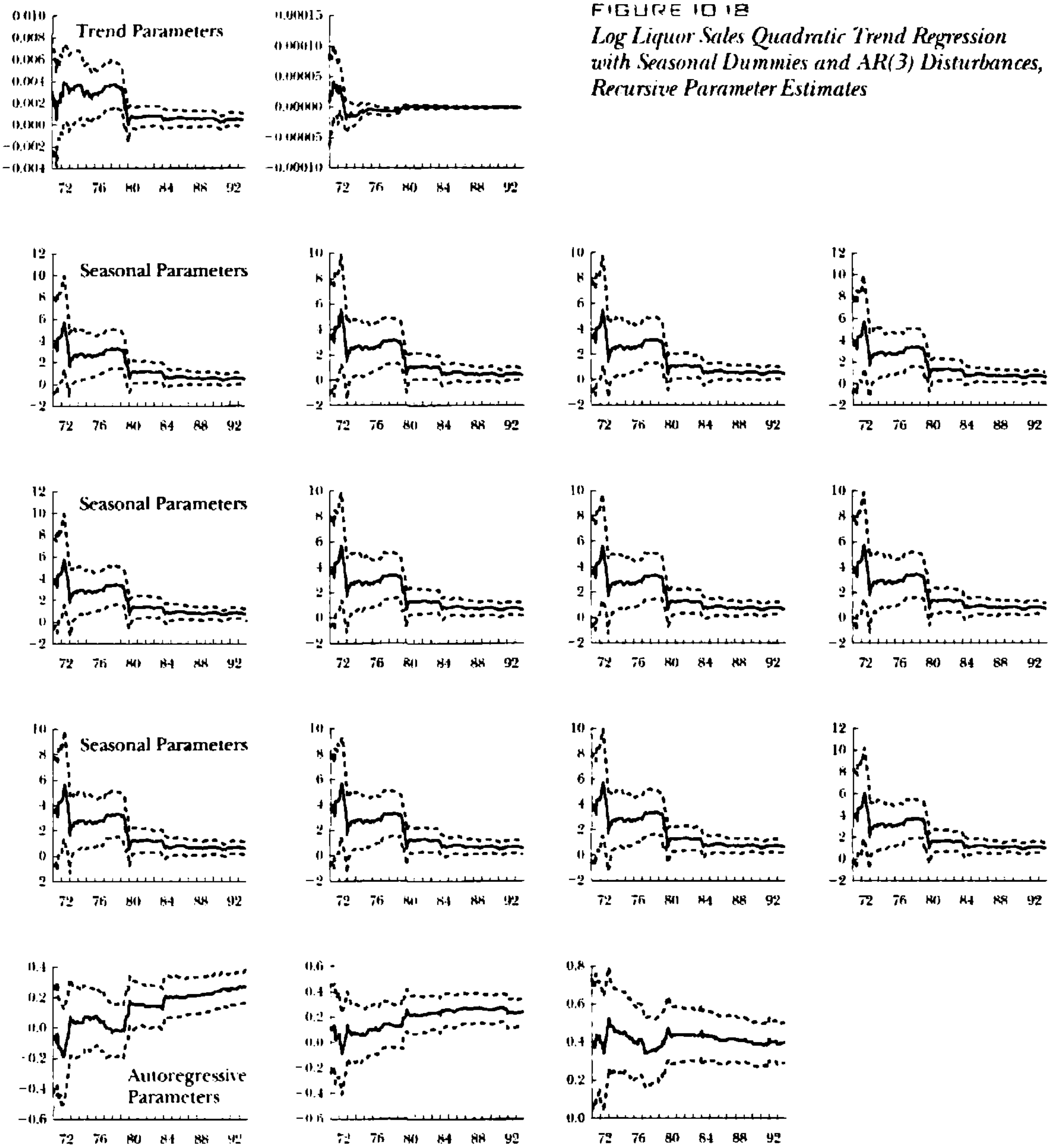


FIGURE 10.18

Log Liquor Sales Quadratic Trend Regression with Seasonal Dummies and AR(3) Disturbances, Recursive Parameter Estimates

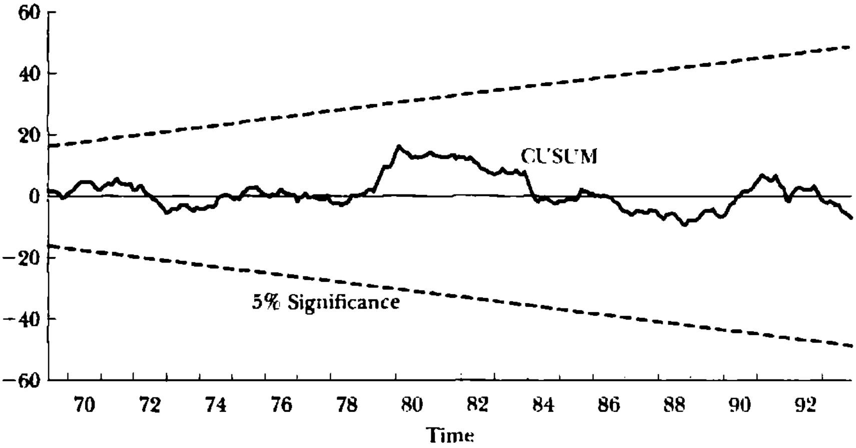


FIGURE 10.19
Log Liquor Sales Quadratic Trend Regression with Seasonal Dummies and AR(3) Disturbances, CUSUM Analysis

Exercises, Problems, and Complements

1. (Serially correlated disturbances vs. lagged dependent variables) Estimate the quadratic trend model for log liquor sales with seasonal dummies and three lags of the dependent variable included directly. Discuss your results and compare them with those we obtained when we instead allowed for AR(3) disturbances in the regression.

2. (Assessing the adequacy of the liquor sales forecasting model trend specification) Critique the liquor sales forecasting model that we adopted (log liquor sales with quadratic trend, seasonal dummies, and AR(3) disturbances).
   a. If the trend is not a good approximation to the actual trend in the series, would it greatly affect short-run forecasts? Long-run forecasts?
   b. Fit and assess the adequacy of a model with loglinear trend.
   c. How might you fit and assess the adequacy of a *broken* linear trend? How might you decide on the location of the break point?

3. (Improving nontrend aspects of the liquor sales forecasting model)[14]
   a. Recall our earlier argument from Chapter 8 that best practice requires using a χ^2_{m-k} distribution rather than a χ^2_m distribution to assess the significance of Q-statistics for model residuals, where m is the number of autocorrelations included in the Box-Pierce statistic and k is the number of parameters estimated. In several places in this chapter, we failed to heed this advice when evaluating the liquor sales model. If we were instead to compare the residual Q-statistic *p*-values with a χ^2_{m-k} distribution, how, if at all, would our assessment of the model's adequacy change?

---

[14] I thank Ron Michener, University of Virginia, for suggesting parts *a* and *c*.

   b. Return to the log-quadratic trend model with seasonal dummies, allow for ARMA(p, q) disturbances, and do a systematic selection of p and q using the AIC and SIC. Do AIC and SIC select the same model? If not, which do you prefer? If your preferred forecasting model differs from the AR(3) that we used, replicate the analysis in the text using your preferred model, and discuss your results.
   c. Discuss and evaluate another possible model improvement: inclusion of an additional dummy variable indicating the number of Fridays and/or Saturdays in the month. Does this model have lower AIC or SIC than the final model used in the text? Do you prefer it to the one in the text? Why or why not?

4. (CUSUM analysis of the housing starts model) Consider the housing starts forecasting model that we built in Chapter 6.
   a. Perform a CUSUM analysis of a housing starts forecasting model that does not account for cycles. (Recall that our model in Chapter 6 did not account for cycles). Discuss your results.
   b. Specify and estimate a model that *does* account for cycles.
   c. Do a CUSUM analysis of the model that accounts for cycles. Discuss your results and compare them with those of part *a*.

5. (Model selection based on simulated forecasting performance)
   a. Return to the retail sales data of Chapter 5, and use recursive cross validation to select between the linear trend forecasting model and the quadratic trend forecasting model. Which do you select? How does it compare with the model selected by the AIC and SIC?
   b. How did you decide on a value of T^* when performing the recursive cross validation on the retail sales data? What are the relevant considerations?
   c. One virtue of recursive cross validation procedures is their flexibility. Suppose that your loss function is not 1-step-ahead mean squared error; instead, suppose it's an asymmetric function of the 1-step-ahead error. How would you modify the recursive cross validation procedure to enforce the asymmetric loss function? How would you proceed if the loss function were 4-step-ahead squared error? How would you proceed if the loss function were an average of 1-step-ahead through 4-step-ahead squared error?

6. (Seasonal models with **time-varying parameters:** forecasting AirSpeed passenger-miles) You work for a hot new startup airline, AirSpeed, modeling and forecasting the miles per person ("passenger-miles") traveled on their flights through the four quarters of the year. During the past 15 years for which you have data, it's well known in the industry that trend passenger-miles have been flat (i.e., there is no trend); similarly, there have been no cyclical effects. Industry experts, however, believe that there are strong seasonal effects, which you think might be very important for modeling and forecasting passenger-miles.
   a. Why might airline passenger-miles be seasonal?
   b. Fit a quarterly seasonal model to the AirSpeed data, and assess the importance of seasonal effects. Do the t- and F-tests indicate that seasonality is important? Do the Akaike and Schwarz criteria indicate that seasonality is important? What is the estimated seasonal pattern?
   c. Use recursive procedures to assess whether the seasonal coefficients are evolving over time. Discuss your results.

d. If the seasonal coefficients are evolving over time, how might you model that evolution and thereby improve your forecasting model? (*Hint:* Allow for trends in the seasonal coefficients themselves.)

e. Compare 4-quarter-ahead extrapolation forecasts from your models with and without evolving seasonality.

7. (**Formal models of unobserved components**) We've used the idea of unobserved components as informal motivation for our models of trends, seasonals, and cycles. Although we will not do so, it's possible to work with formal unobserved components models, such as

$$y_t = T_t + S_t + C_t + I_t ,$$

where T is the trend component; S is the seasonal component; C is the cyclical component; and I is the remainder, or "irregular," component, which is white noise. Typically we'd assume that each component is uncorrelated with all other components at all leads and lags. Typical models for the various components include the following:

Trend

$$T_t = \beta_0 + \beta_1 \text{TIME}_t \qquad \text{(deterministic)}$$

$$T_t = \beta_1 + T_{t-1} + \varepsilon_{1t} \qquad \text{(stochastic)}$$

Seasonal

$$S_t = \sum_{i=1}^{s} \gamma_i D_{it} \qquad \text{(deterministic)}$$

$$S_t = \frac{1}{1 - \gamma L^s} \varepsilon_{2t} \qquad \text{(stochastic)}$$

Cycle

$$C_t = \frac{1}{(1 - \alpha_1 L)} \varepsilon_{3t} \qquad \text{(AR(1))}$$

$$C_t = \frac{1 + \beta_1 L + \beta_2 L^2}{(1 - \alpha_1 L)(1 - \alpha_2 L)} \varepsilon_{3t} \qquad \text{(ARMA(2, 2))}$$

Irregular

$$I_t = \varepsilon_{4t} .$$

8. (The restrictions associated with unobserved-components structures) The restrictions associated with formal unobserved-components models are surely false, in the sense that real-world dynamics are not likely to be decomposable in such a sharp and tidy way. Rather, the decomposition is effectively an accounting framework that we use simply because it's helpful to do so. Trend, seasonal, and cyclical variation are so different—and so important in business, economic, and financial series—that it's often helpful to model them separately to help ensure that we model each adequately. A consensus has not yet emerged as to whether it's more effective to exploit the unobserved components perspective for intuitive motivation, as we do throughout this book, or to enforce formal unobserved components decompositions in hopes of benefitting from considerations related to the shrinkage principle.

9. **(Additive unobserved-components decomposition** and **multiplicative unobserved-components decomposition)** We introduced the formal unobserved components decomposition,

$$y_t = T_t + S_t + C_t + I_t ,$$

where T is the trend component, S is the seasonal component, C is the cyclical component, and I is the remainder, or "irregular," component. Alternatively, we could have introduced a *multiplicative* decomposition,

$$y_t = T_t \, S_t \, C_t \, I_t .$$

   a. Begin with the multiplicative decomposition and take logs. How does your result relate to our original additive decomposition?
   b. Does the exponential (loglinear) trend fit more naturally in the additive or multiplicative decomposition framework? Why?

10. **(Signal, noise, and overfitting)** Using our unobserved-components perspective, we've discussed trends, seasonals, cycles, and noise. We've modeled and forecasted each, with the exception of noise. Clearly we *can't* model or forecast the noise; by construction, it's unforecastable. Instead, the noise is what *remains* after accounting for the other components. We call the other components *signals,* and the signals are buried in noise. Good models fit signals, not noise. Data-mining expeditions, in contrast, lead to models that often fit very well over the historical sample but that fail miserably for out-of-sample forecasting. That's because such data mining effectively tailors the model to fit the idiosyncracies of the in-sample noise, which improves the in-sample fit but is of no help in out-of-sample forecasting.
   a. Choose your favorite trending (but not seasonal) series, and select a sample path of length 100. Graph it.
   b. Regress the first 20 observations on a fifth-order polynomial time trend, and allow for five autoregressive lags as well. Graph the actual and fitted values from the regression. Discuss.
   c. Use your estimated model to produce an 80-step-ahead extrapolation forecast. Graphically compare your forecast with the actual realization. Discuss.

Bibliographical and Computational Notes

Nerlove, Grether, and Carvalho (1979) discuss unobserved components models and their relationship to ARMA models. They also provide an insightful history of the use of unobserved components decompositions for data description and forecasting.

Harvey (1990) derives and presents the formula for r_t, the key element of the variance of the recursive residual. We suggested using the standard error of the regression to estimate σ, the standard deviation of the nonrecursive regression disturbance, as suggested in the original work by Brown, Durbin, and Evans (1975). Since then, a number of authors have used an alternative estimator of σ based on the recursive residuals, which may lead to CUSUM tests with better small-sample power. For a discussion in the context of the dynamic models useful for forecasting, see Krämer, Ploberger, and Alt (1988).

Efron and Tibshirani (1993) give an insightful discussion of forecasting model selection criteria as estimates of out-of-sample MSE, and the natural attractiveness in that regard of numerical methods such as cross validation and its relatives.

Recursive cross validation is often called *predictive stochastic complexity;* the basic theory was developed by Rissanen (1989). Kuan and Liu (1995) make good use of recursive cross validation to select models for forecasting exchange rates, and they provide additional references to the literature on the subject.

Recursive estimation and related techniques are implemented in a number of modern software packages.

Concepts for Review

Recursive estimation
Recursive residuals
Parameter instability
Standardized recursive residuals
CUSUM
CUSUM plot
Random walk
Cross validation
Recursive cross validation
Time-varying parameters
Formal model of unobserved components
Additive unobserved-components decomposition
Multiplicative unobserved-components decomposition
Signal, noise, and overfitting

References and Additional Readings

Brown, R. L., Durbin, J., and Evans, J. M. (1975). "Techniques for Testing the Constance of Regression Relationships over Time." *Journal of the Royal Statistical Society* B, 37, 149–163.

Efron, B., and Tibshirani, R. J. (1993). *An Introduction to the Bootstrap*. New York: Chapman and Hall.

Harvey, A. C. (1990). *The Econometric Analysis of Time Series*. 2nd ed. Cambridge, Mass.: MIT Press.

Krämer, W., Ploberger, W., and Alt, R. (1988). "Testing for Structural Change in Dynamic Models." *Econometrica*, 56, 1355–1369.

Kuan, C. M., and Liu, Y. (1995). "Forecasting Exchange Rates Using Feedforward and Recurrent Neural Networks." *Journal of Applied Econometrics*, 10, 347–364.

Nerlove, M., Grether, D. M., and Carvalho, J. L. (1979). *Analysis of Economic Time Series: A Synthesis* (2nd ed., 1996). New York: Academic Press.

Rissanen, J. (1989). *Stochastic Complexity in Statistical Inquiry*. Singapore: World Science.

Forecasting with Regression Models

The regression model is an explicitly multivariate model, in which variables are explained and forecast on the basis of their own history *and* the histories of other, related variables. Exploiting such cross-variable linkages may lead to good and intuitive forecasting models and to better forecasts than those obtained from univariate models.

Regression models are often called *causal* or *explanatory models.* For example, in the linear regression model,

$$y_t = \beta_0 + \beta_1 x_t + \varepsilon_t$$
$$\varepsilon_t \sim WN(0, \sigma^2)\,,$$

the presumption is that x helps determine, or cause, y, not the other way around. For this reason, the left-hand-side variable is sometimes called the *endogenous* variable, and the right-hand-side variables are called *exogenous* or *explanatory* variables.

But ultimately regression models, like all statistical models, are models of correlation, not causation. Except in special cases, all variables are endogenous, and it's best to admit as much from the outset. Toward the end of this chapter we'll explicitly do so; we'll work with systems of regression equations called *vector autoregressions* (VARs). For now, however, we'll work with the standard single-equation linear regression model, a great workhorse of forecasting, which we can interpret as one equation of a larger system.

1. Conditional Forecasting Models and Scenario Analysis

A **conditional forecasting model** is one that can be used to produce forecasts for a variable of interest, *conditional* on assumptions about other variables. With the regression model, for example, we can forecast y conditional on an assumed future value of x.[1] This sort of conditional forecasting is often called **scenario analysis** or **contingency analysis,** because a conditional forecasting model helps us answer the "what if" questions that often arise. If we condition on the assumption, for example, that the h-step-ahead value of x is x^*_{T+h}, then our h-step-ahead conditional forecast for y is

$$y_{T+h,T} \mid x^*_{T+h} = \beta_0 + \beta_1 x^*_{T+h} \,.$$

Assuming normality, we use the conditional density forecast $N(y_{T+h,T} \mid x^*_{T+h}, \sigma^2)$, and conditional interval forecasts follow immediately from the conditional density forecast. As always, we make the procedure operational by replacing unknown parameters with estimates.

2. Accounting for Parameter Uncertainty in Confidence Intervals for Conditional Forecasts

Forecasts are of course subject to error, and scenario forecasts are no exception. There are at least three sources of such error. One important source of forecast error is **specification uncertainty.** All our models are intentional simplifications, which hopefully capture the salient properties of the data for forecasting purposes. By using modern tools such as information criteria, residual correlograms, and so on, in conjunction with intuition and theory, we attempt to minimize specification uncertainty.

A second source of forecast error is **innovation uncertainty,** which reflects the fact that future innovations are not known when the forecast is made. This is the source of forecast error that we've explicitly acknowledged in our computations of interval and density forecasts. We've seen, for example, that the cumulative effect of innovation uncertainty tends to grow with the forecast horizon, resulting in interval and density forecasts that widen with the horizon.

A third source of forecast error is **parameter uncertainty.** The coefficients that we use to produce forecasts are of course just *estimates,* and the estimates are subject to sampling variability. Specification and innovation uncertainty are likely more important than parameter uncertainty (which vanishes as the sample size grows); in addition, the effect of parameter uncertainty on forecast uncertainty is difficult to quantify in many situations. For both these reasons, parameter uncertainty is often ignored, as we have done thus far.

---

[1] To enhance pedagogical clarity, we work throughout this chapter with regression models containing only one right-hand-side variable. Extensions to models with more than one right-hand-side variable are straightforward.

When using a conditional forecasting model, however, simple calculations allow us to quantify both innovation and parameter uncertainty. Consider, for example, the very simple case in which x has a zero mean and

$$y_t = \beta x_t + \varepsilon_t \, .$$

Suppose we want to predict y_{T+h} at $x_{T+h} = x^*_{T+h}$. If $x_{T+h} = x^*_{T+h}$, then

$$y_{T+h} = \beta x^*_{T+h} + \varepsilon_{T+h} \, .$$

Thus,

$$\hat{y}_{T+h,T} \,|\, x^*_{T+h} = \hat{\beta} x^*_{T+h} \, ,$$

with corresponding error

$$\hat{e}_{T+h,T} = y_{T+h} - \hat{y}_{T+h,T} \,|\, x^*_{T+h} = (\beta - \hat{\beta}) x^*_{T+h} + \varepsilon_{T+h} \, .$$

Therefore,

$$\mathrm{var}(\hat{e}_{T+h,T}) = x^{*2}_{T+h} \, \mathrm{var}(\hat{\beta}) + \sigma^2 \, .$$

We won't do so here, but it can be shown that[2]

$$\mathrm{var}(\hat{\beta}) = \frac{\sigma^2}{\sum_{t=1}^{T} x_t^2} \, .$$

Thus, we arrive at the final formula,

$$\mathrm{var}(\hat{e}_{T+h,T}) = \frac{\sigma^2}{\sum_{t=1}^{T} x_t^2} x^{*2}_{T+h} + \sigma^2 \, .$$

In this expression, the first term accounts for parameter uncertainty, while the second accounts for the usual innovation uncertainty. Taken together, the results suggest an operational density forecast that accounts for parameter uncertainty,

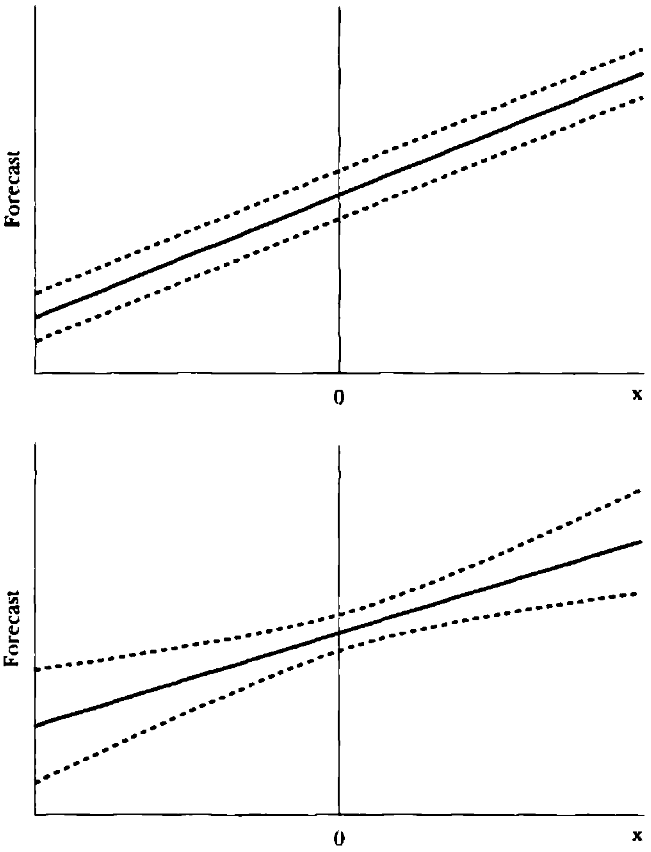
$$N\left(\hat{\beta} x^*_{T+h}, \frac{\hat{\sigma}^2}{\sum_{t=1}^{T} x_t^2} x^{*2}_{T+h} + \hat{\sigma}^2 \right) ,$$

from which interval forecasts may be constructed as well.

---

[2] See any of the elementary statistics or econometrics texts cited in Chapter 1.

FIGURE 11.1
Point and Interval Forecasts



Note: Top panel interval forecasts *don't* acknowledge parameter uncertainty; bottom panel interval forecasts *do* acknowledge parameter uncertainty. To produce the figure, we set $\hat{\beta} = 0$, $\sigma^2 = 1$, and $\sum x_t^2 = 50$.

Note that when parameter uncertainty exists, the closer x^*_{T+h} is to its mean (0), the smaller is the prediction-error variance. The idea can be shown to carry over to more complicated situations when y and x don't necessarily have zero means and to models with more than one regressor: The closer x is to its mean, the tighter is the prediction interval. We illustrate the situation in Figure 11.1; the top panel shows constant intervals ($\pm 1.96\sigma$) that fail to account for parameter uncertainty, and the bottom panel shows the intervals of varying width that account for parameter uncertainty. Finally, note that as the sample size gets large, $\sum_{t=1}^{T} x_t^2$ gets large as well, so the adjustment for parameter uncertainty vanishes, and the formula collapses to our old one.

The discussion of this section depends on the future value of x being known with certainty, which is acceptable in the case of conditional forecasts, in which case we're simply conditioning on an assumption about future x.[3] If we don't want to condition on an assumption about future x, or if we're using certain more complicated models (e.g., those with dynamics), the formula does not apply. We now turn to such situations and models.

3. Unconditional Forecasting Models

Notwithstanding the usefulness of scenario analyses, often we don't want to make forecasts of y conditional on assumptions about x; rather, we just want the best possible forecast of y—an **unconditional forecast.** To get an unconditional forecast from a regression model, we often encounter the **forecasting the right-hand-side variables problem.** That is, to get an optimal unconditional point forecast for y, we can't insert an arbitrary value for future x; rather, we need to insert the optimal point forecast, $x_{T+h,T}$, which yields the unconditional forecast

$$y_{T+h,T} = \beta_0 + \beta_1 x_{T+h,T} .$$

Of course, we usually don't have such a forecast for x, and the model at hand doesn't help us. (It's a model for y—we don't have a model for x.)

One thing we might do is fit a univariate model to x (e.g., an autoregressive model), forecast x (i.e., form $\hat{x}_{T+h,T}$), and then use that forecast of x to forecast y. But just as easily—in fact, preferably—we can estimate all the parameters simultaneously by regressing y on $x_{t-h}, x_{t-h-1}, \ldots$. If we want to forecast only one step ahead, we could use the model

$$y_t = \beta_0 + \delta x_{t-1} + \varepsilon_t .$$

The right-hand-side variable is lagged by one period, so the model is immediately useful for 1-step-ahead unconditional forecasting. More lags of x can of course be included; the key is that all variables on the right are lagged by at least one period. Forecasting more than one step ahead, however, again leads to the problem of forecasting the right-hand-side variables—if we want to forecast h steps ahead, all variables on the right must be lagged by at least h periods.

In a few important special cases, the problem of forecasting the right-hand-side variables doesn't arise, because the regressors are perfectly deterministic, so we know exactly what they'll be at any future time. The trend and seasonal models discussed in Chapters 5 and 6 are leading examples. Such cases are atypical, however.

---

[3] The discussion also applies to forecasting in cross-sectional environments, in which forecasts are almost always conditional. Suppose, for example, that we estimate a regression model relating expenditure on restaurant meals to income, using cross-section data on 1000 households for 1997. Then, if we get 1997 income data for an additional set of people, we can use it to forecast their restaurant expenditures.

4. Distributed Lags, Polynomial Distributed Lags, and Rational Distributed Lags

An unconditional forecasting model like

$$y_t = \beta_0 + \delta x_{t-1} + \varepsilon_t$$

can be immediately generalized to the **distributed lag model,**

$$y_t = \beta_0 + \sum_{i=1}^{N_x} \delta_i x_{t-i} + \varepsilon_t \, .$$

We say that y depends on a distributed lag of past x's. The coefficients on the lagged x's are called *lag weights,* and their pattern is called the *lag distribution.*

One way to estimate a distributed lag model is simply to include all N_x lags of x in the regression, which can be estimated by least squares in the usual way. In many situations, however, N_x might be quite a large number, in which case we'd have to use many degrees of freedom to estimate the model, violating the parsimony principle. Often we can recover many of those degrees of freedom without seriously worsening the model's fit by constraining the lag weights to lie on a low-order polynomial. Such **polynomial distributed lags** promote smoothness in the lag distribution and may lead to sophisticatedly simple models with improved forecasting performance.

Polynomial distributed lag models are estimated by minimizing the sum of squared residuals in the usual way, subject to the constraint that the lag weights follow a low-order polynomial whose degree must be specified. Suppose, for example, that we constrain the lag weights to follow a second-degree polynomial. Then we find the parameter estimates by solving the problem

$$\min_{\beta_0, \delta_i} \sum_{t=N_x+1}^{T} \left(y_t - \beta_0 - \sum_{i=1}^{N_x} \delta_i x_{t-i} \right)^2 ,$$

subject to

$$\delta_i = P(i) = a + bi + ci^2, \quad i = 1, \ldots, N_x \, .$$

This converts the estimation problem from one of estimating $1 + N_x$ parameters, $\beta_0, \delta_1, \ldots, \delta_{N_x}$, to one of estimating four parameters, β_0, a, b, and c. Sometimes additional constraints are imposed on the shape of the polynomial, such as

$$P(N_x) = 0 \, ,$$

which enforces the idea that the dynamics have been exhausted by lag N_x.

Polynomial distributed lags produce aesthetically appealing, but basically ad hoc, lag distributions. After all, why should the lag weights necessarily follow a low-order polynomial? An alternative and often preferable approach makes use of the **rational distributed lags** that we introduced in Chapter 7 in the context of univariate ARMA modeling. Rational distributed lags promote parsimony, and hence smoothness in the lag distribution, but they do so in a way that's potentially much less restrictive than requiring the lag weights to follow a low-order polynomial. We might, for example, use a model like

$$y_t = \frac{A(L)}{B(L)} x_t + \varepsilon_t ,$$

where $A(L)$ and $B(L)$ are low-order polynomials in the lag operator. Equivalently, we can write

$$B(L) y_t = A(L) x_t + B(L) \varepsilon_t ,$$

which emphasizes that the rational distributed lag of x actually brings both lags of x and lags of y into the model. One way or another, it's crucial to allow for lags of y, and we now study such models in greater depth.

5. Regressions with Lagged Dependent Variables, Regressions with ARMA Disturbances, and Transfer Function Models

There's something missing in distributed lag models of the form

$$y_t = \beta_0 + \sum_{i=1}^{N_x} \delta_i x_{t-i} + \varepsilon_t .$$

A multivariate model (in this case, a regression model) should relate the current value y to its own past and to the past of x. But as the model is presently written, we've left out the past of y! Even in distributed lag models, we always want to allow for the presence of the usual univariate dynamics. Put differently, the included regressors may not capture all the dynamics in y, which we need to model one way or another. Thus, for example, a preferable model includes lags of the dependent variable,

$$y_t = \beta_0 + \sum_{i=1}^{N_y} \alpha_i y_{t-i} + \sum_{j=1}^{N_x} \delta_j x_{t-j} + \varepsilon_t .$$

This model, a **distributed lag regression model with lagged dependent variables,** is closely related to, but not exactly the same as, the rational distributed lag model introduced earlier. (Why?) You can think of it as arising by beginning with a univariate autoregressive model for y and then introducing additional explanatory variables. If the lagged y's don't play a role, as assessed with the usual tests, we can always delete them, but we never want to eliminate from the outset the possibility that lagged dependent variables play a role. Lagged dependent variables absorb residual serial correlation and can *dramatically* enhance forecasting performance.

Alternatively, we can capture own-variable dynamics in distributed lag regression models by using a **distributed lag regression model with ARMA disturbances.** Recall that our ARMA(p, q) models are equivalent to regression models, with only a constant regressor and with ARMA(p, q) disturbances:

$$\begin{aligned} y_t &= \beta_0 + \varepsilon_t \\ \varepsilon_t &= \frac{\Theta(L)}{\Phi(L)} v_t \\ v_t &\sim WN(0, \sigma^2)\,. \end{aligned}$$

We want to begin with the univariate model as a baseline and then generalize it to allow for multivariate interaction, resulting in models such as

$$\begin{aligned} y_t &= \beta_0 + \sum_{i=1}^{N_x} \delta_i x_{t-i} + \varepsilon_t \\ \varepsilon_t &= \frac{\Theta(L)}{\Phi(L)} v_t \\ v_t &\sim WN(0, \sigma^2)\,. \end{aligned}$$

Regressions with ARMA disturbances make clear that regression (a statistical and econometric tool with a long tradition) and the ARMA model of time series dynamics (a more recent innovation) are not at all competitors; rather, when used appropriately, they can be highly complementary.

It turns out that the distributed lag regression model with autoregressive disturbances—a great workhorse in econometrics—is a special case of the more general model with lags of both y and x and white noise disturbances. To see this, let's take the simple example of an unconditional (1-step-ahead) regression forecasting model with AR(1) disturbances:

$$\begin{aligned} y_t &= \beta_0 + \beta_1 x_{t-1} + \varepsilon_t \\ \varepsilon_t &= \varphi \varepsilon_{t-1} + v_t \\ v_t &\sim WN(0, \sigma^2)\,. \end{aligned}$$

In lag operator notation, we write the AR(1) regression disturbance as

$$(1 - \varphi L)\varepsilon_t = v_t\,,$$

or

$$\varepsilon_t = \frac{1}{(1 - \varphi L)} v_t\,.$$

Thus, we can rewrite the regression model as

$$y_t = \beta_0 + \beta_1 x_{t-1} + \frac{1}{(1 - \varphi L)} v_t\,.$$

Now multiply both sides by $(1 - \varphi L)$ to get

$$(1 - \varphi L)y_t = (1 - \varphi)\beta_0 + \beta_1(1 - \varphi L)x_{t-1} + v_t\,,$$

or

$$y_t = \varphi y_{t-1} + (1 - \varphi)\beta_0 + \beta_1 x_{t-1} - \varphi\beta_1 x_{t-2} + v_t\,.$$

Thus, a model with one lag of x on the right and AR(1) disturbances is equivalent to a model with y_{t-1}, x_{t-1}, and x_{t-2} on the right-hand side and white noise errors, *subject to the restriction* that the coefficient on the second lag of x_{t-2} is the negative of the product of the coefficients on y_{t-1} and x_{t-1}.

Distributed lag regressions with lagged dependent variables are more general than distributed lag regressions with dynamic disturbances. **Transfer function models** are more general still and include both as special cases.[4] The basic idea is to exploit the power and parsimony of rational distributed lags in modeling both own-variable and cross-variable dynamics. Imagine beginning with a univariate ARMA model,

$$y_t = \frac{C(L)}{D(L)}\varepsilon_t\,,$$

TABLE 11.1 *The Transfer Function Model and Various Special Cases*

Name	Model	Restrictions
Transfer function	$y_t = \frac{A(L)}{B(L)}x_t + \frac{C(L)}{D(L)}\varepsilon_t$	None
Standard distributed lag	$y_t = A(L)x_t + \varepsilon_t$	$B(L) = C(L) = D(L) = 1$
Rational distributed lag	$y_t = \frac{A(L)}{B(L)}x_t + \varepsilon_t$	$C(L) = D(L) = 1$
Univariate AR	$y_t = \frac{1}{D(L)}\varepsilon_t$	$A(L) = 0,\ C(L) = 1$
Univariate MA	$y_t = C(L)\varepsilon_t$	$A(L) = 0,\ D(L) = 1$
Univariate ARMA	$y_t = \frac{C(L)}{D(L)}\varepsilon_t$	$A(L) = 0$
Distributed lag with lagged dependent variables	$B(L)y_t = A(L)x_t + \varepsilon_t$, or $y_t = \frac{A(L)}{B(L)}x_t + \frac{1}{B(L)}\varepsilon_t$	$C(L) = 1,\ D(L) = B(L)$
Distributed lag with ARMA disturbances	$y_t = A(L)x_t + \frac{C(L)}{D(L)}\varepsilon_t$	$B(L) = 1$
Distributed lag with AR disturbances	$y_t = A(L)x_t + \frac{1}{D(L)}\varepsilon_t$	$B(L) = C(L) = 1$

---

[4] Table 11.1 displays a variety of important forecasting models, all of which are special cases of the transfer function model.

which captures own-variable dynamics using a rational distributed lag. Now extend the model to capture cross-variable dynamics using a rational distributed lag of the other variable, which yields the general transfer function model,

$$y_t = \frac{A(L)}{B(L)} x_t + \frac{C(L)}{D(L)} \varepsilon_t \,.$$

Distributed lag regression with lagged dependent variables is a potentially restrictive special case, which emerges when $C(L) = 1$ and $B(L) = D(L)$. (Verify this for yourself.) Distributed lag regression with ARMA disturbances is also a special case, which emerges when $B(L) = 1$. (Verify this, too.)

In practice, the important thing is to allow for own-variable dynamics *somehow*, in order to account for dynamics in y not explained by the right-hand-side variables. Whether we do so by including lagged dependent variables, or by allowing for ARMA disturbances, or by estimating general transfer function models, can occasionally be important, but usually it's a comparatively minor issue.

6. Vector Autoregressions

A univariate autoregression involves one variable. In a univariate autoregression of order p, we regress a variable on p lags of itself. In contrast, a multivariate autoregression—that is, a vector autoregression, or VAR—involves N variables. In an N-variable **vector autoregression of order p**, or VAR(p), we estimate N different equations. In each equation, we regress the relevant left-hand-side variable on p lags of itself *and p lags of every other variable.*[5] Thus, the right-hand-side variables are the same in every equation—p lags of every variable.

The key point is that, in contrast to the univariate case, vector autoregressions allow for **cross-variable dynamics.** Each variable is related not only to its own past but also to the past of all the other variables in the system. In a two-variable VAR(1), for example, we have two equations, one for each variable (y_1 and y_2). We write

$$y_{1,t} = \varphi_{11} y_{1,t-1} + \varphi_{12} y_{2,t-1} + \varepsilon_{1,t}$$

$$y_{2,t} = \varphi_{21} y_{1,t-1} + \varphi_{22} y_{2,t-1} + \varepsilon_{2,t} \,.$$

Each variable depends on one lag of the other variable in addition to one lag of itself; that's one obvious source of multivariate interaction captured by the VAR that may be useful for forecasting. In addition, the disturbances may be correlated, so that when one equation is shocked, the other will typically be shocked as well, which is another type of multivariate interaction that univariate models miss. We summarize the disturbance variance-covariance structure as

---

[5] Trends, seasonals, and other exogenous variables may also be included, as long as they're all included in every equation.

$$\varepsilon_{1,t} \sim WN(0, \sigma_1^2)$$
$$\varepsilon_{2,t} \sim WN(0, \sigma_2^2)$$
$$\text{cov}(\varepsilon_{1,t}, \varepsilon_{2,t}) = \sigma_{12} \,.$$

The innovations *could* be uncorrelated, which occurs when $\sigma_{12} = 0$, but they needn't be.

You might guess that VARs would be hard to estimate. After all, they're fairly complicated models, with potentially many equations and many right-hand-side variables in each equation. In fact, precisely the opposite is true. VARs are very easy to estimate, because we need only run N linear regressions. That's one reason why VARs are so popular—OLS estimation of autoregressive models is simple and stable, in contrast to the numerical estimation required for models with moving average components.[6] Equation-by-equation OLS estimation also turns out to have very good statistical properties when each equation has the same regressors, as is the case in standard VARs. Otherwise, a more complicated estimation procedure called *seemingly unrelated regression,* which explicitly accounts for correlation across equation disturbances, would be required to obtain estimates with good statistical properties.[7]

When fitting VARs to data, we use the Schwarz and Akaike criteria, just as in the univariate case. The formulas differ, however, because we're now working with a multivariate system of equations rather than a single equation. To get an AIC or SIC value for a VAR system, we could add up the equation-by-equation AICs or SICs, but, unfortunately, doing so is appropriate only if the innovations are uncorrelated across equations, which is a very special and unusual situation. Instead, explicitly multivariate versions of the AIC and SIC—and more advanced formulas—are required that account for cross-equation innovation correlation. It's beyond the scope of this book to derive and present those formulas, because they involve unavoidable use of matrix algebra, but fortunately we don't need to. They're preprogrammed in many computer packages, and we interpret the AIC and SIC values computed for VARs of various orders in exactly the same way as in the univariate case: We select that order p such that the AIC or SIC is minimized.

We construct VAR forecasts in a way that precisely parallels the univariate case. We can construct 1-step-ahead point forecasts immediately, because all variables on the right-hand side are lagged by one period. Armed with the 1-step-ahead forecasts, we can construct the 2-step-ahead forecasts, from which we can construct the 3-step-ahead forecasts, and so on in the usual way, following the chain rule of forecasting. We construct interval and density forecasts in ways that also parallel the univariate case. The multivariate nature of VARs makes the derivations more tedious, however, so we bypass them. As always, to construct practical forecasts we replace unknown parameters by estimates.

---

[6] Estimation of MA and ARMA models is stable enough in the univariate case but rapidly becomes unwieldy in multivariate situations. Hence, multivariate ARMA models are used infrequently in practice, in spite of the potential they hold for providing parsimonious approximations to the Wold representation.

[7] For an exposition of seemingly unrelated regression, see Pindyck and Rubinfeld (1991).

7. Predictive Causality

There's an important statistical notion of causality that's intimately related to forecasting and naturally introduced in the context of VARs. It is based on two key principles: First, cause should occur before effect; second, a causal series should contain information useful for forecasting that is not available in the other series (including the past history of the variable being forecast). In the unrestricted VARs that we've studied thus far, *everything* causes everything else, because lags of every variable appear on the right of every equation. Cause precedes effect because the right-hand-side variables are lagged, and each variable is useful in forecasting every other variable.

We stress from the outset that the notion of **predictive causality** contains little if any information about causality in the philosophical sense. Rather, the statement "y_i causes y_j" is just shorthand for the more precise but long-winded statement "y_i contains useful information for predicting y_j (in the linear least-squares sense), over and above the past histories of the other variables in the system." To save space, we simply say that y_i causes y_j.

To understand what predictive causality means in the context of a VAR(p), consider the jth equation of the N-equation system, which has y_j on the left and p lags of each of the N variables on the right. If y_i causes y_j, then at least one of the lags of y_i that appear on the right side of the y_j equation must have a nonzero coefficient.

It's also useful to consider the opposite situation, in which y_i does not cause y_j. In that case, all of the lags of that y_i that appear on the right side of the y_j equation must have zero coefficients.[8] Statistical causality tests are based on this formulation of noncausality. We use an F-test to assess whether all coefficients on lags of y_i are jointly 0.

Note that we've defined noncausality in terms of 1-step-ahead prediction errors. In the bivariate VAR, this implies noncausality in terms of h-step-ahead prediction errors, for all h. (Why?) In higher dimensional cases, things are trickier; 1-step-ahead noncausality does not necessarily imply noncausality at other horizons. For example, variable i may 1-step-cause variable j, and variable j may 1-step-cause variable k. Thus, variable i 2-step-causes variable k but does not 1-step-cause variable k.

---

[8] Note that in such a situation the error variance in forecasting y_j using lags of all variables in the system will be the same as the error variance in forecasting y_j using lags of all variables in the system *except* y_i.

Causality tests are often used when building and assessing forecasting models, because they can inform us about those parts of the workings of complicated multivariate models that are particularly relevant for forecasting. Just staring at the coefficients of an estimated VAR (and in complicated systems there are *many* coefficients) rarely yields insights into its workings. Thus, we need tools that help us to see through to the practical forecasting properties of the model that concern us. And we often have keen interest in the answers to questions such as "Does y_i contribute toward improving forecasts of y_j?" and "Does y_j contribute toward improving forecasts of y_i?" If the results violate intuition or theory, then we might scrutinize the model more closely. In a situation in which we can't reject a certain noncausality hypothesis, and neither intuition nor theory makes us uncomfortable with it, we might want to *impose* it, by omitting certain lags of certain variables from certain equations.

Various types of causality hypotheses are sometimes entertained. In any equation (the jth, say), we've already discussed testing the simple noncausality hypothesis that

a. No lags of variable i aid in 1-step-ahead prediction of variable j.

We can broaden the idea, however. Sometimes we test stronger noncausality hypotheses such as

b. No lags of a *set* of other variables aid in 1-step-ahead prediction of variable j.
c. No lags of *any other variables* aid in 1-step-ahead prediction of variable j.

All of hypotheses a, b, and c amount to assertions that various coefficients are 0. Finally, sometimes we test noncausality hypotheses that involve more than one equation, such as

d. No variable in a set A causes any variable in a set B, in which case we say that the variables in A are block noncausal for those in B.

This particular noncausality hypothesis corresponds to exclusion restrictions that hold simultaneously in a number of equations. Again, however, standard test procedures are applicable.

8. Impulse-Response Functions and Variance Decompositions

The **impulse-response function** is another device that helps us to learn about the dynamic properties of vector autoregressions of interest to forecasters. We'll introduce it first in the *univariate* context, and then we'll move to VARs. The question of interest is simple and direct: How does a unit innovation to a series affect it, now and in the future? To answer the question, we simply read off the coefficients in the moving average representation of the process.

We're used to normalizing the coefficient on ε_t to unity in moving average representations, but we don't have to do so; more generally, we can write

$$y_t = b_0\varepsilon_t + b_1\varepsilon_{t-1} + b_2\varepsilon_{t-2} + \cdots$$
$$\varepsilon_t \sim WN(0, \sigma^2)\,.$$

The additional generality introduces ambiguity, however, because we can always multiply and divide every ε_t by an arbitrary constant m, yielding an equivalent model but with different parameters and innovations,

$$y_t = (b_0 m)\left(\frac{1}{m}\varepsilon_t\right) + (b_1 m)\left(\frac{1}{m}\varepsilon_{t-1}\right) + (b_2 m)\left(\frac{1}{m}\varepsilon_{t-2}\right) + \cdots$$

$$\varepsilon_t \sim WN(0, \sigma^2)$$

or

$$y_t = b'_0\varepsilon'_t + b'_1\varepsilon'_{t-1} + b'_2\varepsilon'_{t-2} + \cdots$$

$$\varepsilon'_t \sim WN\left(0, \frac{\sigma^2}{m^2}\right)\,,$$

where $b'_i = b_i m$ and $\varepsilon'_t = \frac{\varepsilon_t}{m}$.

To remove the ambiguity, we must set a value of m. Typically we set $m = 1$, which yields the standard form of the moving average representation. For impulse-response analysis, however, a different normalization turns out to be particularly convenient; we choose $m = \sigma$, which yields

$$y_t = (b_0\sigma)\left(\frac{1}{\sigma}\varepsilon_t\right) + (b_1\sigma)\left(\frac{1}{\sigma}\varepsilon_{t-1}\right) + (b_2\sigma)\left(\frac{1}{\sigma}\varepsilon_{t-2}\right) + \cdots$$
$$\varepsilon_t \sim WN(0, \sigma^2)\,,$$

or

$$y_t = b'_0\varepsilon'_t + b'_1\varepsilon'_{t-1} + b'_2\varepsilon'_{t-2} + \cdots$$

$$\varepsilon'_t \sim WN(0, 1)\,,$$

where $b'_i = b_i\sigma$ and $\varepsilon'_t = \frac{\varepsilon_t}{\sigma}$. Taking $m = \sigma$ converts shocks to "standard deviation units," because a unit shock to ε'_t corresponds to a one-standard-deviation shock to ε_t.

To make matters concrete, consider the univariate AR(1) process,

$$y_t = \varphi y_{t-1} + \varepsilon_t$$
$$\varepsilon_t \sim WN(0, \sigma^2)\,.$$

The standard moving average form is

$$y_t = \varepsilon_t + \varphi\varepsilon_{t-1} + \varphi^2\varepsilon_{t-2} + \cdots$$

$$\varepsilon_t \sim WN(0, \sigma^2)\,,$$

and the equivalent representation in standard deviation units is

$$y_t = b_0\varepsilon'_t + b_1\varepsilon'_{t-1} + b_2\varepsilon'_{t-2} + \cdots$$
$$\varepsilon'_t \sim WN(0,1),$$

where $b_i = \varphi'\sigma$ and $\varepsilon'_t = \frac{\varepsilon_t}{\sigma}$. The impulse-response function is $\{b_0, b_1, \ldots\}$. The parameter b_0 is the contemporaneous effect of a unit shock to ε'_t or, equivalently, a one-standard-deviation shock to ε_t; as must be the case, then, $b_0 = \sigma$. Note well that b_0 gives the immediate effect of the shock at time t, when it hits. The parameter b_1, which multiplies ε'_{t-1}, gives the effect of the shock one period later, and so on. The full set of impulse-response coefficients, $\{b_0, b_1, \ldots\}$, tracks the complete dynamic response of y to the shock.

Now we consider the multivariate case. The idea is the same, but there are more shocks to track. The key question is, "How does a unit shock to ε_i affect y_j, now and in the future, for all the various combinations of i and j?" Consider, for example, the bivariate VAR(1),

$$y_{1t} = \varphi_{11}y_{1,t-1} + \varphi_{12}y_{2,t-1} + \varepsilon_{1t}$$
$$y_{2t} = \varphi_{21}y_{1,t-1} + \varphi_{22}y_{2,t-1} + \varepsilon_{2t}$$
$$\varepsilon_{1,t} \sim WN(0, \sigma_1^2)$$
$$\varepsilon_{2,t} \sim WN(0, \sigma_2^2)$$
$$\text{cov}(\varepsilon_1, \varepsilon_2) = \sigma_{12}\,.$$

The standard moving average representation, obtained by back substitution, is

$$y_{1t} = \varepsilon_{1t} + \varphi_{11}\varepsilon_{1,t-1} + \varphi_{12}\varepsilon_{2,t-1} + \cdots$$
$$y_{2t} = \varepsilon_{2t} + \varphi_{21}\varepsilon_{1,t-1} + \varphi_{22}\varepsilon_{2,t-1} + \cdots$$
$$\varepsilon_{1,t} \sim WN(0, \sigma_1^2)$$
$$\varepsilon_{2,t} \sim WN(0, \sigma_2^2)$$
$$\text{cov}(\varepsilon_1, \varepsilon_2) = \sigma_{12}\,.$$

Just as in the univariate case, it proves fruitful to adopt a different normalization of the moving average representation for impulse-response analysis. The multivariate analog of our univariate normalization by σ is called *normalization by the Cholesky factor.*[9] The resulting VAR moving average representation has a number of useful properties that parallel the univariate case precisely. First, the innovations of the transformed system are in standard deviation units. Second, although the current innovations in the standard representation have unit coefficients, the current innovations in the normalized representation have nonunit coefficients. In fact, the first equation has only one current innovation, ε_{1t}. (The other has a 0 coefficient.) The second equation has both current innovations. Thus, the ordering of the variables can matter.[10]

---

[9] For detailed discussion and derivation of this advanced topic, see Hamilton (1994).

If y_1 is ordered first, the normalized representation is

$$y_{1,t} = b^0_{11}\varepsilon'_{1,t} + b^1_{11}\varepsilon'_{1,t-1} + b^1_{12}\varepsilon'_{2,t-1} + \cdots$$

$$y_{2,t} = b^0_{21}\varepsilon'_{1,t} + b^0_{22}\varepsilon'_{2,t} + b^1_{21}\varepsilon'_{1,t-1} + b^1_{22}\varepsilon'_{2,t-1} + \cdots$$

$$\varepsilon'_{1,t} \sim WN(0, 1)$$

$$\varepsilon'_{2,t} \sim WN(0, 1)$$

$$\operatorname{cov}(\varepsilon'_1, \varepsilon'_2) = 0\,.$$

Alternatively, if y_2 ordered first, the normalized representation is

$$y_{2,t} = b^0_{22}\varepsilon'_{2,t} + b^1_{21}\varepsilon'_{1,t-1} + b^1_{22}\varepsilon'_{2,t-1} + \cdots$$

$$y_{1,t} = b^0_{11}\varepsilon'_{1,t} + b^0_{12}\varepsilon'_{2,t} + b^1_{11}\varepsilon'_{1,t-1} + b^1_{12}\varepsilon'_{2,t-1} + \cdots$$

$$\varepsilon'_{1,t} \sim WN(0, 1)$$

$$\varepsilon'_{2,t} \sim WN(0, 1)$$

$$\operatorname{cov}(\varepsilon'_1, \varepsilon'_2) = 0\,.$$

Finally, the normalization adopted yields a 0 covariance between the disturbances of the transformed system. This is crucial, because it lets us perform the experiment of interest—shocking one variable in isolation of the others, which we can do if the innovations are uncorrelated but can't do if they're correlated, as in the original unnormalized representation.

After normalizing the system, for a given ordering—say, y_1 first—we compute four sets of impulse-response functions for the bivariate model: response of y_1 to a unit normalized innovation to y_1, $\{b^0_{11}, b^1_{11}, b^2_{11}, \ldots\}$, response of y_1 to a unit normalized innovation to y_2, $\{b^1_{12}, b^2_{12}, \ldots\}$, response of y_2 to a unit normalized innovation to y_2, $\{b^0_{22}, b^1_{22}, b^2_{22}, \ldots\}$, and response of y_2 to a unit normalized innovation to y_1, $\{b^0_{21}, b^1_{21}, b^2_{21}, \ldots\}$. Typically we examine the set of impulse-response functions graphically. Often it turns out that impulse-response functions aren't sensitive to ordering, but the only way to be sure is to check.[11]

In practical applications of impulse-response analysis, we simply replace unknown parameters by estimates, which immediately yields point estimates of the impulse-response functions. Getting confidence intervals for impulse-response functions is trickier, however, and adequate procedures are still under development.

---

[10] In higher-dimensional VARs, the equation that's first in the ordering has only one current innovation, ε'_{1t}. The equation that's second has only current innovations ε'_{1t} and ε'_{2t}; the equation that's third has only current innovations ε'_{1t}, ε'_{2t}, and ε'_{3t}; and so on.

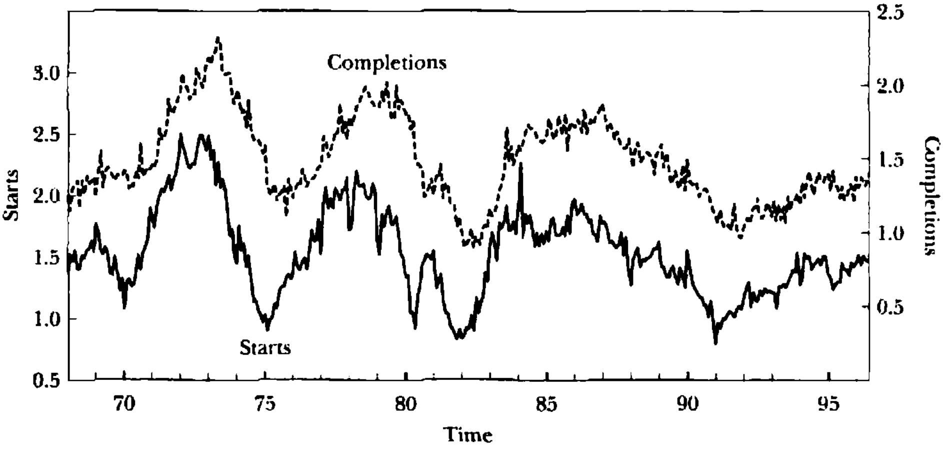
[11] Note well that the issues of normalization and ordering only affect impulse-response analysis; for forecasting, we only need the unnormalized model.

Another way of characterizing the dynamics associated with VARs, closely related to impulse-response functions, is the **variance decomposition.** Variance decompositions have an immediate link to forecasting—they answer the question "How much of the h-step-ahead forecast error variance of variable i is explained by innovations to variable j, for $h = 1, 2, \ldots$?" As with impulse-response functions, we typically make a separate graph for every (i, j) pair. Impulse-response functions and the variance decompositions present the same information (although they do so in different ways). For that reason, it's not strictly necessary to present both, and impulse-response analysis has gained greater popularity. Hence, we offer only this brief discussion of variance decomposition. In the application to housing starts and completions that follows, however, we examine both impulse-response functions and variance decompositions. The two are highly complementary, as with information criteria and correlograms for model selection, and the variance decompositions have a nice forecasting motivation.

9. Application: Housing Starts and Completions

We estimate a bivariate VAR for U.S. seasonally adjusted housing starts and completions, two widely watched business cycle indicators, 1968.01–1996.06. We use the VAR to produce point extrapolation forecasts. We show housing starts and completions in Figure 11.2. Both are highly cyclical, increasing during business cycle expansions and decreasing during contractions. Moreover,

FIGURE 11.2 *U.S. Housing Starts and Completions, 1968.01–1996.06*



completions tend to lag behind starts, which makes sense because a house takes time to complete.

We split the data into an estimation sample, 1968.01–1991.12, and a holdout sample, 1992.01–1996.06, for forecasting. We therefore perform all model specification analysis and estimation, to which we now turn, on the 1968.01–1991.12 data. We show the starts correlogram in Table 11.2 and Figure 11.3. The sample autocorrelation function decays slowly, whereas the sample partial autocorrelation function appears to cut off at displacement 2. The patterns in the sample autocorrelations and partial autocorrelations are highly statistically significant, as evidenced by both the Bartlett standard errors and the Ljung-Box Q-statistics. The completions correlogram, in Table 11.3 and Figure 11.4, behaves similarly.

TABLE 11.2
Starts Correlogram

Sample: 1968:01 1991:12
Included observations: 288

	Acorr.	P. Acorr.	Std. Error	Ljung-Box	*p*-value
1	0.937	0.937	0.059	255.24	0.000
2	0.907	0.244	0.059	495.53	0.000
3	0.877	0.054	0.059	720.95	0.000
4	0.838	−0.077	0.059	927.39	0.000
5	0.795	−0.096	0.059	1113.7	0.000
6	0.751	−0.058	0.059	1280.9	0.000
7	0.704	−0.067	0.059	1428.2	0.000
8	0.650	−0.098	0.059	1554.4	0.000
9	0.604	0.004	0.059	1663.8	0.000
10	0.544	−0.129	0.059	1752.6	0.000
11	0.496	0.029	0.059	1826.7	0.000
12	0.446	−0.008	0.059	1886.8	0.000
13	0.405	0.076	0.059	1936.8	0.000
14	0.346	−0.144	0.059	1973.3	0.000
15	0.292	−0.079	0.059	1999.4	0.000
16	0.233	−0.111	0.059	2016.1	0.000
17	0.175	−0.050	0.059	2025.6	0.000
18	0.122	−0.018	0.059	2030.2	0.000
19	0.070	0.002	0.059	2031.7	0.000
20	0.019	−0.025	0.059	2031.8	0.000
21	−0.034	−0.032	0.059	2032.2	0.000
22	−0.074	0.036	0.059	2033.9	0.000
23	−0.123	−0.028	0.059	2038.7	0.000
24	−0.167	−0.048	0.059	2047.4	0.000

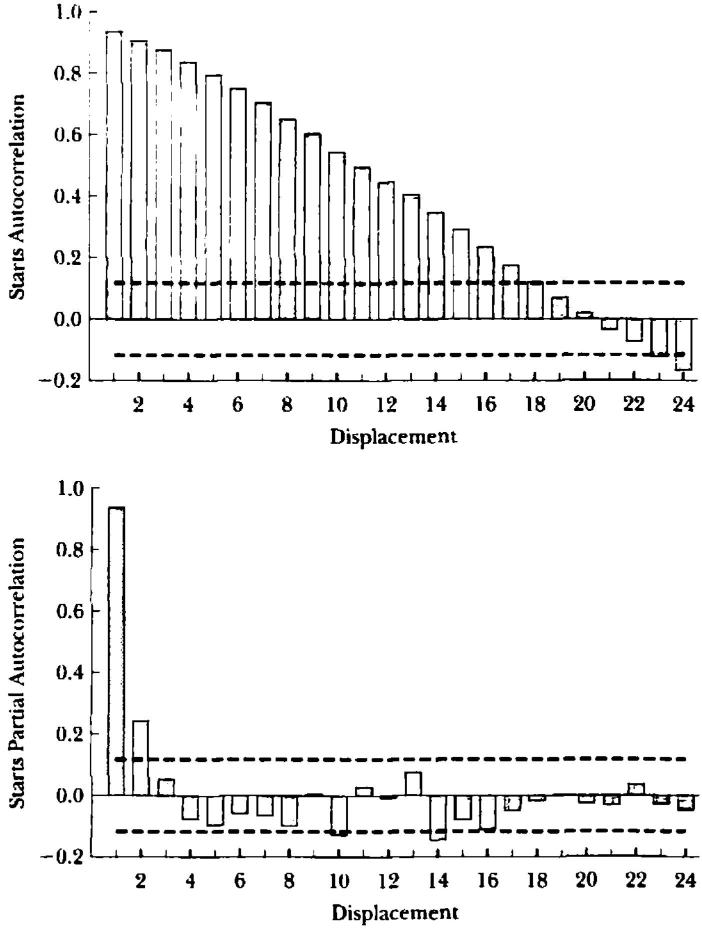


FIGURE 11.3
Starts, Sample Autocorrelations and Partial Autocorrelations

We've not yet introduced the **cross-correlation function.** There's been no need, because it's not relevant for univariate modeling. It provides important information, however, in the multivariate environments that now concern us. Recall that the autocorrelation function is the correlation between a variable and lags of itself. The cross-correlation function is a natural multivariate analog; it's simply the correlation between a variable and lags of *another* variable. We estimate those correlations using the usual estimator and graph them as a function of displacement along with the Bartlett two-standard-error bands, which apply just as in the univariate case.

The cross-correlation function (Figure 11.5) for housing starts and completions is very revealing. Starts and completions are highly correlated at all

TABLE 11.3
Completions Correlogram

Sample: 1968:01 1991:12
Included observations: 288

	Acorr.	P. Acorr.	Std. Error	Ljung-Box	*p*-value
1	0.939	0.939	0.059	256.61	0.000
2	0.920	0.328	0.059	504.05	0.000
3	0.896	0.066	0.059	739.19	0.000
4	0.874	0.023	0.059	963.73	0.000
5	0.834	−0.165	0.059	1168.9	0.000
6	0.802	−0.067	0.059	1359.2	0.000
7	0.761	−0.100	0.059	1531.2	0.000
8	0.721	−0.070	0.059	1686.1	0.000
9	0.677	−0.055	0.059	1823.2	0.000
10	0.633	−0.047	0.059	1943.7	0.000
11	0.583	−0.080	0.059	2046.3	0.000
12	0.533	−0.073	0.059	2132.2	0.000
13	0.483	−0.038	0.059	2203.2	0.000
14	0.434	−0.020	0.059	2260.6	0.000
15	0.390	0.041	0.059	2307.0	0.000
16	0.337	−0.057	0.059	2341.9	0.000
17	0.290	−0.008	0.059	2367.9	0.000
18	0.234	−0.109	0.059	2384.8	0.000
19	0.181	−0.082	0.059	2395.0	0.000
20	0.128	−0.047	0.059	2400.1	0.000
21	0.068	−0.133	0.059	2401.6	0.000
22	0.020	0.037	0.059	2401.7	0.000
23	−0.038	−0.092	0.059	2402.2	0.000
24	−0.087	−0.003	0.059	2404.6	0.000

displacements, and a clear pattern emerges as well: Although the contemporaneous correlation is high (0.78), completions are maximally correlated with starts lagged by roughly 6–12 months (around 0.90). Again, this makes good sense in light of the time it takes to build a house.

Now we proceed to model starts and completions. We need to select the order, p, of our VAR(p). Based on exploration using multivariate versions of SIC and AIC, we adopt a VAR(4).

First consider the starts equation (Table 11.4), residual plot (Figure 11.6), and residual correlogram (Table 11.5 and Figure 11.7). The explanatory power of the model is good, as judged by the R^2 as well as the plots of actual and fitted values, and the residuals appear white, as judged by the residual sample autocorrelations, partial autocorrelations, and Ljung-Box statistics. Note as well that no lag of completions has a significant effect on starts, which makes sense—we obviously expect starts to cause completions, but not conversely. The

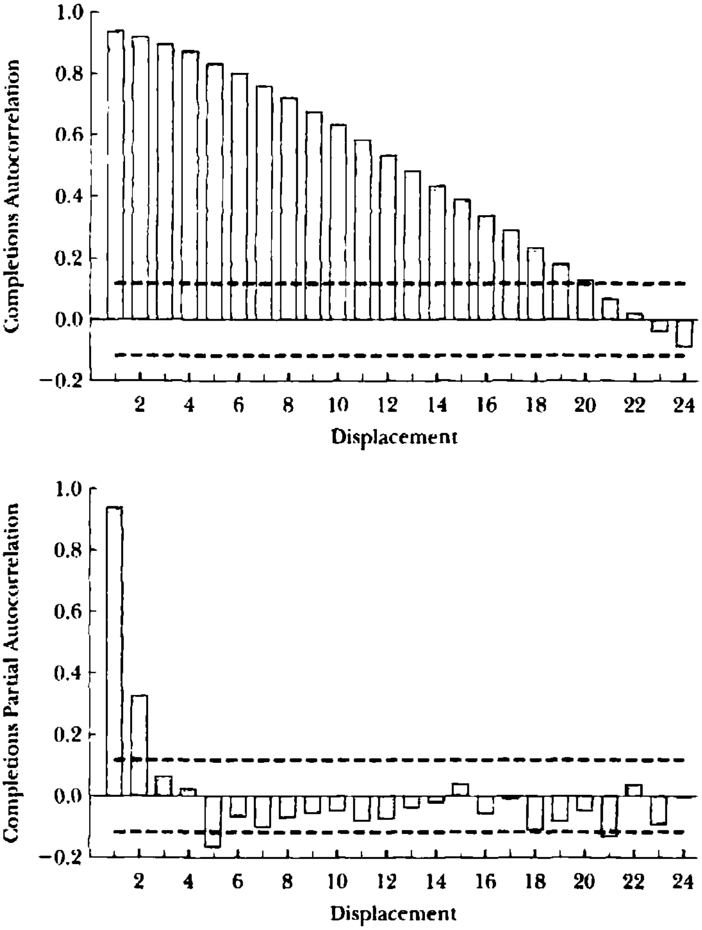


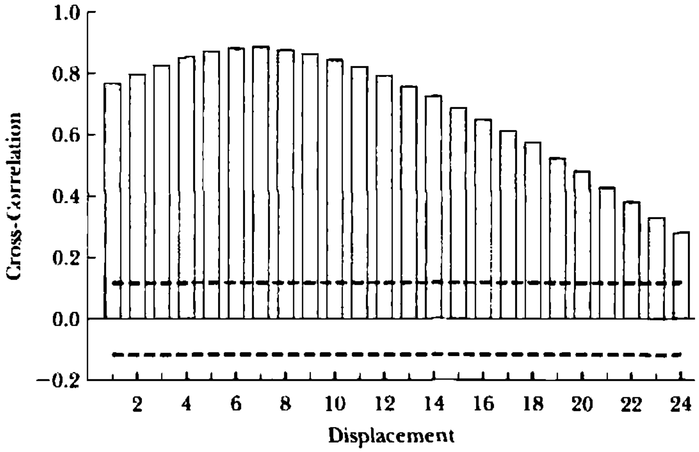
FIGURE 11.4
Completions, Sample Autocorrelations and Partial Autocorrelations

completions equation (Table 11.6), residual plot (Figure 11.8), and residual correlogram (Table 11.7 and Figure 11.9) appear similarly good. Lagged starts, moreover, most definitely have a significant effect on completions.

Table 11.8 shows the results of formal causality tests. The hypothesis that starts don't cause completions is simply that the coefficients on the four lags of starts in the completions equation are all 0. The *F*-statistic is overwhelmingly significant, which is not surprising in light of the previously noticed highly significant *t*-statistics. Thus, we reject noncausality from starts to completions at any reasonable level. Perhaps more surprising is the fact that we also reject noncausality from completions to starts at roughly the 5% level. The causality appears bidirectional, in which case we say there is **feedback.**

To get a feel for the dynamics of the estimated VAR before producing forecasts, we compute impulse-response functions and variance decompositions.

FIGURE 11.5
Starts and Completions, Sample Cross-Correlations



Note: We graph the sample correlation between completions at time t and starts at time $t-i$, $i = 1, 2, \ldots, 24$.

TABLE 11.4
VAR Starts Equation

LS // Dependent variable is STARTS.
Sample(adjusted): 1968:05 1991:12
Included observations: 284 after adjusting endpoints

Variable	Coefficient	Std. Error	t-Statistic	Prob.
C	0.146871	0.044235	3.320264	0.0010
STARTS(−1)	0.659939	0.061242	10.77587	0.0000
STARTS(−2)	0.229632	0.072724	3.157587	0.0018
STARTS(−3)	0.142859	0.072655	1.966281	0.0503
STARTS(−4)	0.007806	0.066032	0.118217	0.9060
COMPS(−1)	0.031611	0.102712	0.307759	0.7585
COMPS(−2)	−0.120781	0.103847	−1.163069	0.2458
COMPS(−3)	−0.020601	0.100946	−0.204078	0.8384
COMPS(−4)	−0.027404	0.094569	−0.289779	0.7722

R^2	0.895566	Mean dependent var.	1.574771
Adjusted R^2	0.892528	SD dependent var.	0.382362
SE of regression	0.125350	Akaike info criterion	−4.122118
Sum squared resid.	4.320952	Schwarz criterion	−4.006482
Log likelihood	191.3622	F-statistic	294.7796
Durbin-Watson stat.	1.991908	Prob(F-statistic)	0.000000

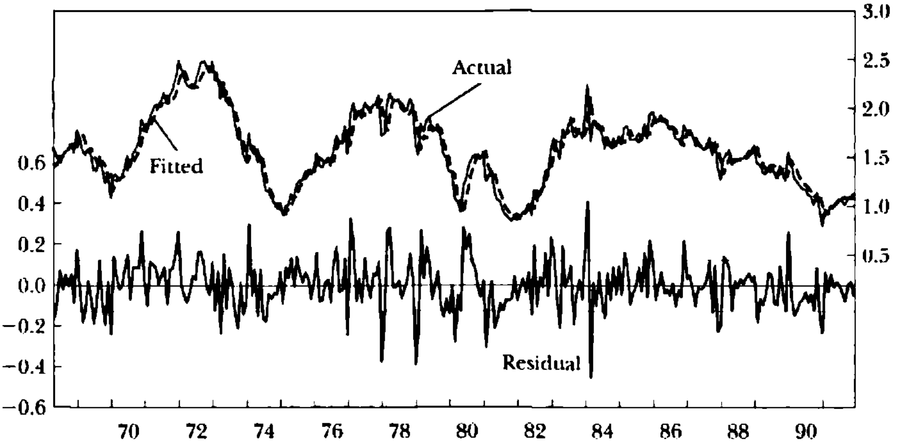


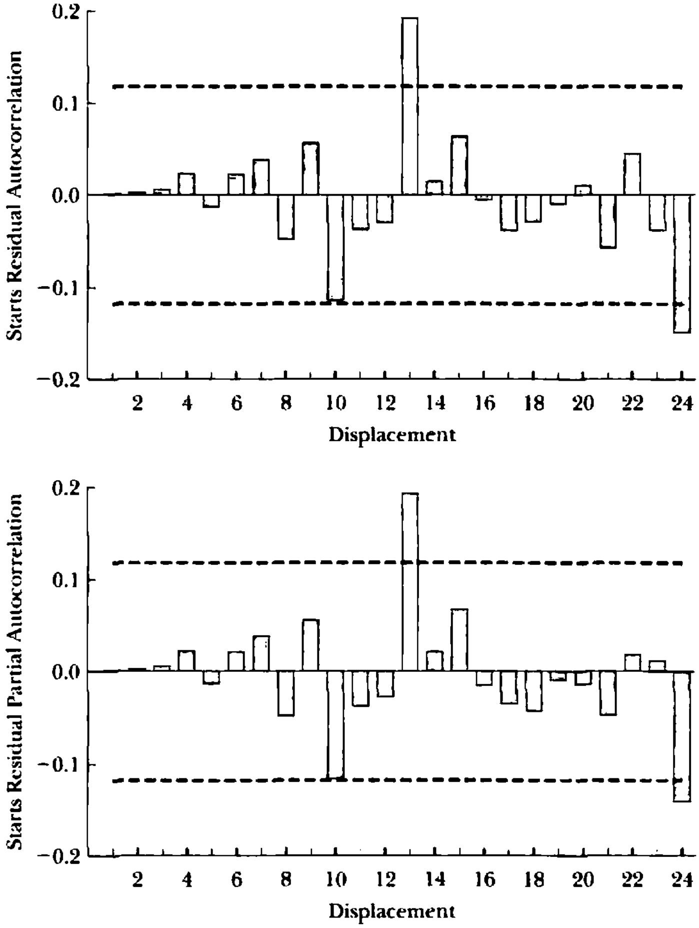
FIGURE 11.6
VAR Starts Equation, Residual Plot

TABLE 11.5
VAR Starts Equation, Residual Correlogram

Sample: 1968:01 1991:12
Included observations: 284

	Acorr.	P. Acorr.	Std. Error	Ljung-Box	*p*-value
1	0.001	0.001	0.059	0.0004	0.985
2	0.003	0.003	0.059	0.0029	0.999
3	0.006	0.006	0.059	0.0119	1.000
4	0.023	0.023	0.059	0.1650	0.997
5	−0.013	−0.013	0.059	0.2108	0.999
6	0.022	0.021	0.059	0.3463	0.999
7	0.038	0.038	0.059	0.7646	0.998
8	−0.048	−0.048	0.059	1.4362	0.994
9	0.056	0.056	0.059	2.3528	0.985
10	−0.114	−0.116	0.059	6.1868	0.799
11	−0.038	−0.038	0.059	6.6096	0.830
12	−0.030	−0.028	0.059	6.8763	0.866
13	0.192	0.193	0.059	17.947	0.160
14	0.014	0.021	0.059	18.010	0.206
15	0.063	0.067	0.059	19.199	0.205
16	−0.006	−0.015	0.059	19.208	0.258
17	−0.039	−0.035	0.059	19.664	0.292
18	−0.029	−0.043	0.059	19.927	0.337
19	−0.010	−0.009	0.059	19.959	0.397
20	0.010	−0.014	0.059	19.993	0.458
21	−0.057	−0.047	0.059	21.003	0.459
22	0.045	0.018	0.059	21.644	0.481
23	−0.038	0.011	0.059	22.088	0.515
24	−0.149	−0.141	0.059	29.064	0.218

FIGURE 11.7 *VAR Starts Equation, Residual Sample Autocorrelations and Partial Autocorrelations*



We present results for starts first in the ordering, so that a current innovation to starts affects only current starts, but the results are robust to reversal of the ordering.

Figure 11.10 displays the impulse-response functions. First let's consider the own-variable impulse responses—that is, the effects of a starts innovation on subsequent starts or a completions innovation on subsequent completions. The effects are similar. In each case, the impulse response is large and decays in a slow, approximately monotonic fashion. In contrast, the cross-variable impulse responses are very different. An innovation to starts produces no movement in completions at first, but the effect gradually builds and becomes large, peaking at about 14 months. (It takes time to build houses.) An innovation to completions, however, produces little movement in starts at any time.

TABLE 11.6
VAR Completions Equation

LS // Dependent variable is COMPS.
Sample(adjusted): 1968:05 1991:12
Included observations: 284 after adjusting endpoints

Variable	Coefficient	Std. Error	*t*-Statistic	Prob.
C	0.045347	0.025794	1.758045	0.0799
STARTS(−1)	0.074724	0.035711	2.092461	0.0373
STARTS(−2)	0.040047	0.042406	0.944377	0.3458
STARTS(−3)	0.047145	0.042366	1.112805	0.2668
STARTS(−4)	0.082331	0.038504	2.138238	0.0334
COMPS(−1)	0.236774	0.059893	3.953313	0.0001
COMPS(−2)	0.206172	0.060554	3.404742	0.0008
COMPS(−3)	0.120998	0.058863	2.055593	0.0408
COMPS(−4)	0.156729	0.055144	2.842160	0.0048

R^2	0.936835	Mean dependent var.	1.547958
Adjusted R^2	0.934998	SD dependent var.	0.286689
SE of regression	0.073093	Akaike info criterion	−5.200872
Sum squared resid.	1.469205	Schwarz criterion	−5.085236
Log likelihood	344.5453	*F*-statistic	509.8375
Durbin-Watson stat.	2.013370	Prob(*F*-statistic)	0.000000

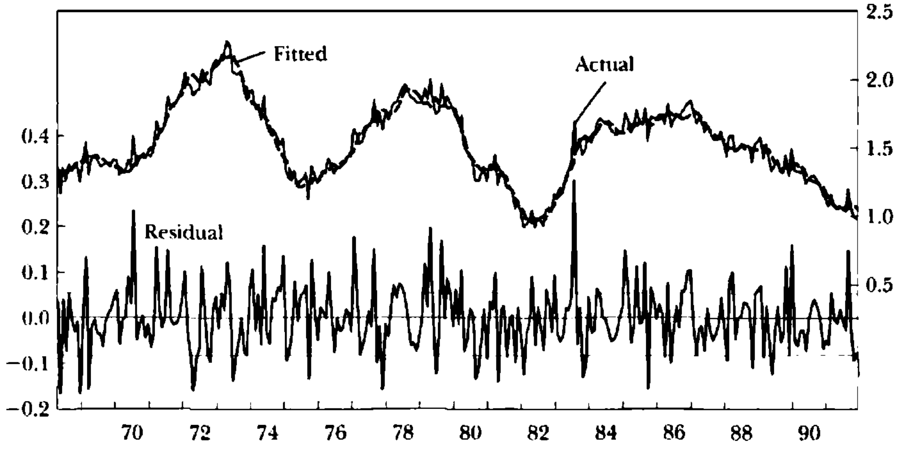


FIGURE 11.8
VAR Completions Equation, Residual Plot

TABLE 11.7
VAR Completions Equation, Residual Correlogram

Sample: 1968:01 1991:12
Included observations: 284

	Acorr.	P. Acorr.	Std. Error	Ljung-Box	*p*-value
1	−0.009	−0.009	0.059	0.0238	0.877
2	−0.035	−0.035	0.059	0.3744	0.829
3	−0.037	−0.037	0.059	0.7640	0.858
4	−0.088	−0.090	0.059	3.0059	0.557
5	−0.105	−0.111	0.059	6.1873	0.288
6	0.012	0.000	0.059	6.2291	0.398
7	−0.024	−0.041	0.059	6.4047	0.493
8	0.041	0.024	0.059	6.9026	0.547
9	0.048	0.029	0.059	7.5927	0.576
10	0.045	0.037	0.059	8.1918	0.610
11	−0.009	−0.005	0.059	8.2160	0.694
12	−0.050	−0.046	0.059	8.9767	0.705
13	−0.038	−0.024	0.059	9.4057	0.742
14	−0.055	−0.049	0.059	10.318	0.739
15	0.027	0.028	0.059	10.545	0.784
16	−0.005	−0.020	0.059	10.553	0.836
17	0.096	0.082	0.059	13.369	0.711
18	0.011	−0.002	0.059	13.405	0.767
19	0.041	0.040	0.059	13.929	0.788
20	0.046	0.061	0.059	14.569	0.801
21	−0.096	−0.079	0.059	17.402	0.686
22	0.039	0.077	0.059	17.875	0.713
23	−0.113	−0.114	0.059	21.824	0.531
24	−0.136	−0.125	0.059	27.622	0.276

Figure 11.11 shows the variance decompositions. The fraction of the error variance in forecasting starts due to innovations in starts is close to 100% at all horizons. In contrast, the fraction of the error variance in forecasting completions due to innovations in starts is near 0 at short horizons, but it rises steadily and is near 100% at long horizons, again reflecting time-to-build effects.

Finally, we construct forecasts for the out-of-sample period, 1992.01–1996.06. The starts forecast appears in Figure 11.12. Starts begin their recovery before 1992.01, and the VAR projects continuation of the recovery. The VAR forecasts captures the general pattern quite well, but it forecasts quicker mean

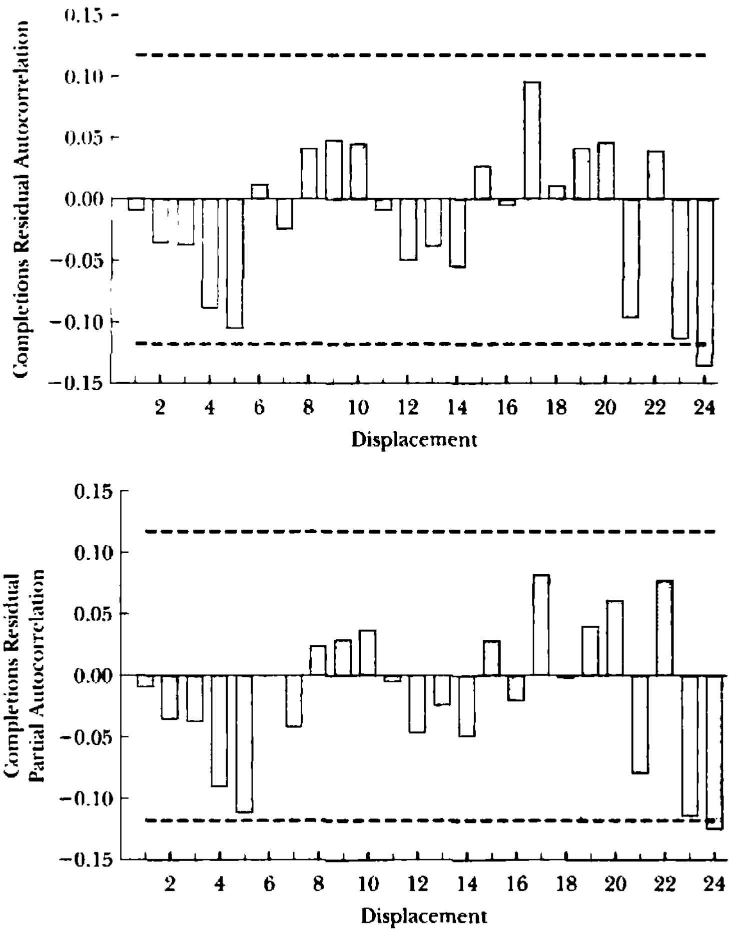


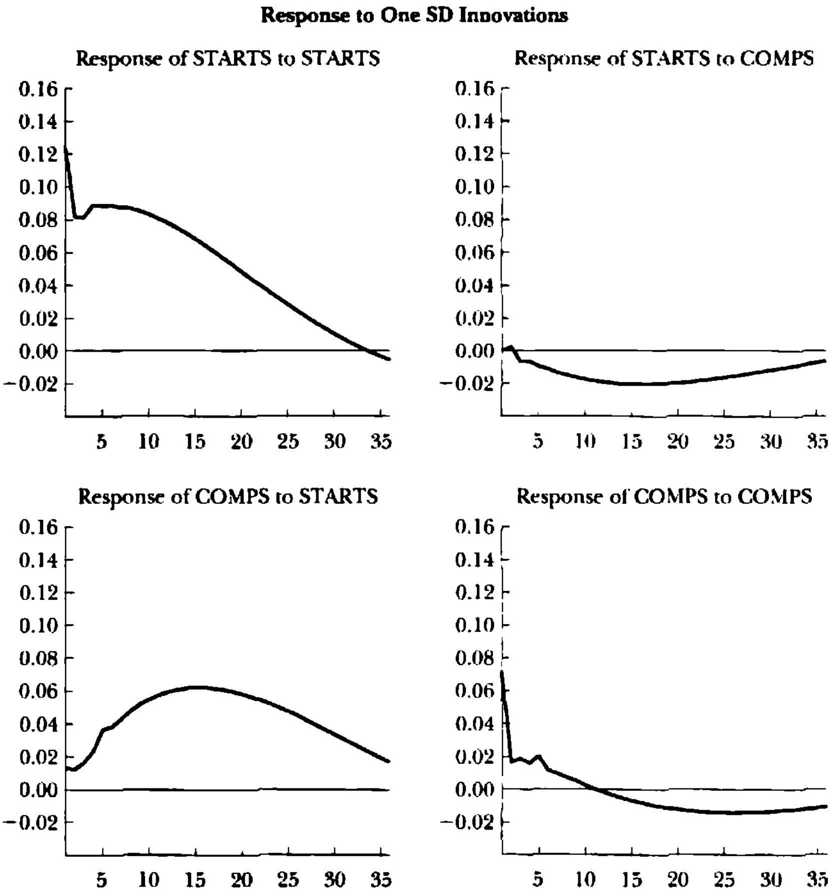
FIGURE 11.13
VAR Completions Equation, Residual Sample Autocorrelations and Partial Autocorrelations

TABLE 11.8
Housing Starts and Completions, Causality Tests

Sample: 1968:01 1991:12
Lags: 4
Obs: 284

Null Hypothesis:	***F*-Statistic**	**Probability**
STARTS does not cause COMPS	26.2658	0.00000
COMPS does not cause STARTS	2.23876	0.06511

FIGURE 11.10
Housing Starts and Completions, VAR Impulse-Response Functions



reversion than actually occurs, as is clear when comparing the forecast and realization in Figure 11.13. The figure also makes clear that the recovery of housing starts from the recession of 1990 was slower than the previous recoveries in the sample, which naturally makes for difficult forecasting. The completions forecast suffers the same fate, as shown in Figures 11.14 and 11.15. Interestingly, however, completions had not yet turned by 1991.12, but the forecast nevertheless correctly predicts the turning point. (Why?)

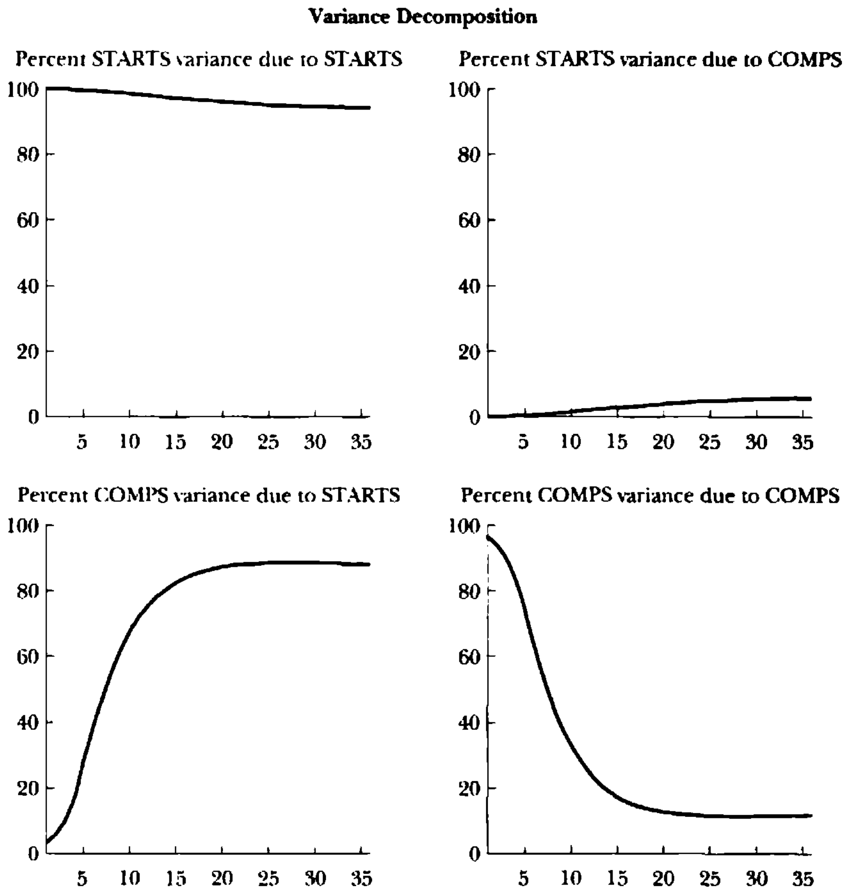


FIGURE 11.11 *Housing Starts and Completions, VAR Variance Decompositions*

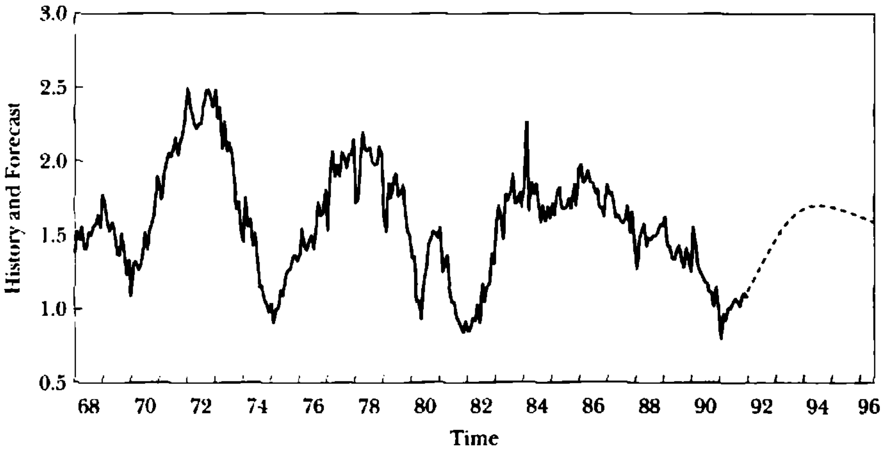


FIGURE 11.12 *Starts: History, 1968.01–1991.12, and Forecast, 1992.01–1996.06*

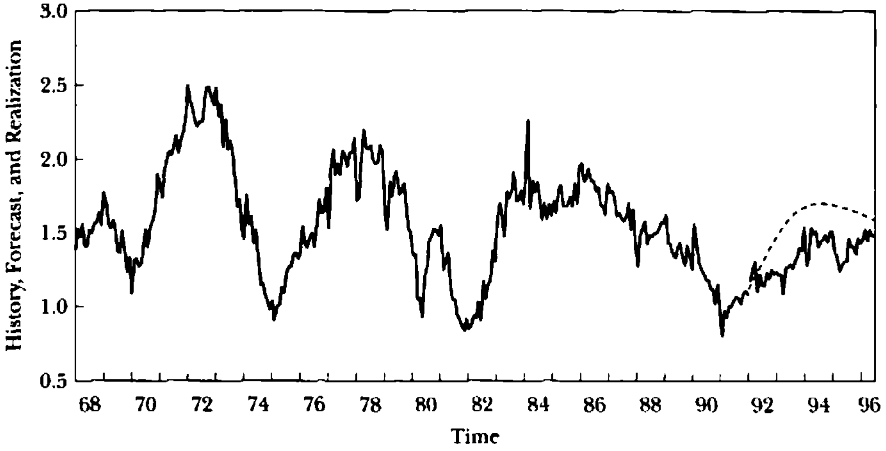


FIGURE 11.13
Starts: History,
1968.01–1991.12,
Forecast, and
Realization,
1992.01–1996.06

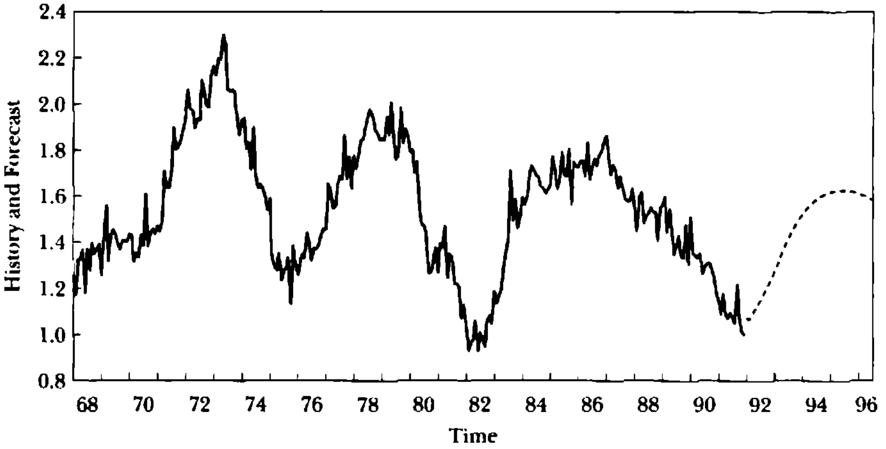


FIGURE 11.14
Completions:
History,
1968.01–1991.12,
and Forecast,
1992.01–1996.06

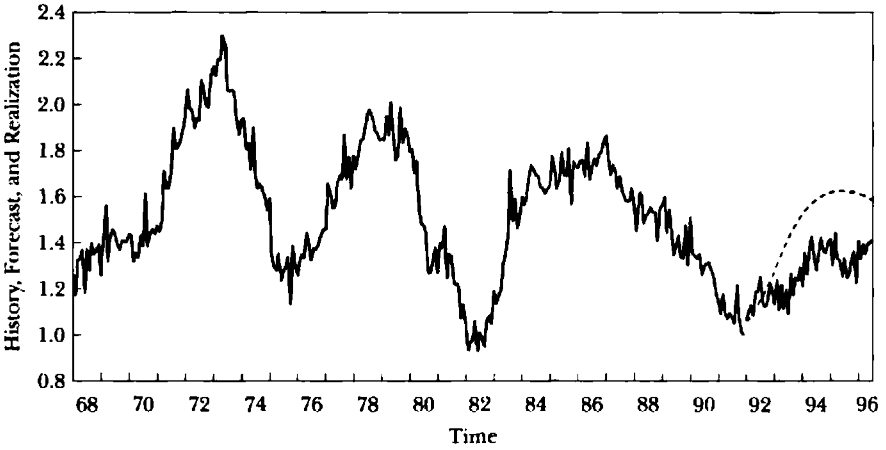


FIGURE 11.15
Completions:
History,
1968.01–1991.12,
Forecast, and
Realization,
1992.01–1996.06

Exercises, Problems, and Complements

1. (Econometrics, time series analysis, and forecasting) As recently as the early 1970s, time series analysis was mostly univariate and made little use of economic theory. Econometrics, in contrast, stressed the cross-variable dynamics associated with economic theory, with equations estimated using multiple regression. Econometrics, moreover, made use of simultaneous systems of such equations, requiring complicated estimation methods. Thus, the econometric and time series approaches to forecasting were very different.[12]

   As Klein (1981) notes, however, the complicated econometric system estimation methods had little payoff for practical forecasting and were therefore largely abandoned, whereas the rational distributed lag patterns associated with time series models led to large improvements in practical forecast accuracy.[13] Thus, in more recent times, the distinction between econometrics and time series analysis has largely vanished, with the union incorporating the best of both. In many respects, the VAR is a modern embodiment of both econometric and time series traditions. VARs use economic considerations to determine which variables to include and which (if any) restrictions should be imposed, allow for rich multivariate dynamics, typically require only simple estimation techniques, and are explicit forecasting models.

2. (Forecasting crop yields) Consider the following dilemma in agricultural crop yield forecasting:

The possibility of forecasting crop yields several years in advance would, of course, be of great value in the planning of agricultural production. However, the success of long-range crop forecasts is contingent not only on our knowledge of the weather factors determining yield, but also on our ability to predict the weather. Despite an abundant literature in this field, no firm basis for reliable long-range weather forecasts has yet been found. (Sanderson, 1953, p. 3)

   a. How is the situation related to our concerns in this chapter and, specifically, to the issue of conditional versus unconditional forecasting?
   b. What variables other than weather might be useful for predicting crop yield?
   c. How would you suggest that the forecaster should proceed?

3. (Regression forecasting models with expectations, or anticipatory, data) A number of surveys exist of anticipated market conditions, investment intentions, buying plans, advance commitments, consumer sentiment, and so on.
   a. Search the Internet for such series, and report your results. A good place to start is the Resources for Economists page mentioned in Chapter 1.

---

[12] Klein and Young (1980) and Klein (1983) provide good discussions of the traditional econometric simultaneous equations paradigm, as well as the link between structural simultaneous equations models and reduced-form time series models. Wallis (1995) provides a good summary of modern large-scale macroeconometric modeling and forecasting, and Pagan and Robertson (2002) provide an intriguing discussion of the variety of macroeconomic forecasting approaches currently employed in central banks around the world.

[13] For an acerbic assessment circa the mid-1970s, see Jenkins (1979).

b. How might you use the series you found in an unconditional regression forecasting model of GDP? Are the implicit forecast horizons known for all the anticipatory series you found? If not, how might you decide how to lag them in your regression forecasting model?
c. How would you test whether the anticipatory series you found provide incremental forecast enhancement, relative to the own past history of GDP?

4. (Business cycle analysis and forecasting: expansions, contractions, turning points, and leading indicators[14]) The use of anticipatory data is linked to business cycle analysis in general and to leading indicators in particular. During the first half of the 20th century, much research was devoted to obtaining an empirical characterization of the business cycle. The most prominent example of this work is Burns and Mitchell (1946), whose summary empirical definition is as follows:

Business cycles are a type of fluctuation found in the aggregate economic activity of nations that organize their work mainly in business enterprises: a cycle consists of expansions occurring at about the same time in many economic activities, followed by similarly general recessions, contractions, and revivals which merge into the expansion phase of the next cycle. (p. 3)

The comovement among individual economic variables is a key feature of Burns and Mitchell's definition of business cycles. Indeed, the comovement among series, taking into account possible leads and lags in timing, is the centerpiece of Burns and Mitchell's methodology. In their analysis, Burns and Mitchell consider the historical concordance of hundreds of series, including those measuring commodity output, income, prices, interest rates, banking transactions, and transportation services, and they classify series as leading, lagging, or coincident. One way to define a leading indicator is to say that a series x is a leading indicator for a series y if x causes y in the predictive sense. According to that definition, for example, our analysis of housing starts and completions indicates that starts are a leading indicator for completions.

Leading indicators have the potential to be used in forecasting equations in the same way as anticipatory variables. Inclusion of a leading indicator, appropriately lagged, can improve forecasts. Zellner and Hong (1989) and Zellner, Hong, and Min (1991), for example, make good use of that idea in their ARLI (autoregressive leading-indicator) models for forecasting aggregate output growth. In those models, Zellner et al. build forecasting models by regressing output on lagged output and lagged leading indicators; they also use shrinkage techniques to coax the forecasted growth rates toward the international average, which improves forecast performance.

Burns and Mitchell use the clusters of turning points in individual series to determine the monthly dates of the turning points in the overall business cycle and to construct composite indexes of leading, coincident, and lagging indicators. Such indexes have been produced by the National Bureau of Economic Research (a think tank in Cambridge, Massachusetts), the Department of Commerce (a U.S. government agency in Washington, D.C.), and the Conference Board (a business membership organization based in New York).[15]

---

[14] This complement draws in part on Diebold and Rudebusch (1996).

[15] The indexes build on very early work, such as the Harvard "Index of General Business Conditions." For a fascinating discussion of the early work, see Hardy (1923), Chapter 7.

Composite indexes of leading indicators are often used to gauge likely future economic developments, but their usefulness is by no means uncontroversial and remains the subject of ongoing research. For example, leading indexes apparently cause aggregate output in analyses of ex post historical data (Auerbach, 1982), but they appear much less useful in real-time forecasting, which is what's relevant (Diebold and Rudebusch, 1991).

5. (Subjective information, Bayesian VARs, and the Minnesota prior) When building and using forecasting models, we frequently have hard-to-quantify subjective information, such as a reasonable range in which we expect a parameter to be. We can incorporate such subjective information in a number of ways. One way is informal judgmental adjustment of estimates. Based on a variety of factors, for example, we might feel that an estimate of a certain parameter in a forecasting model is too high, so we might reduce it a bit.

   Bayesian analysis allows us to incorporate subjective information in a rigorous and replicable way. We summarize subjective information about parameters with a probability distribution called the *prior distribution*, and as always we summarize the information in the data with the likelihood function. The centerpiece of Bayesian analysis is a mathematical formula called *Bayes' rule*, which tells us how to combine the information in the prior and the likelihood to form the posterior distribution of model parameters, which then feed their way into forecasts.

   The Minnesota prior (introduced and popularized by Robert Litterman and Christopher Sims at the University of Minnesota) is commonly used for Bayesian estimation of VAR forecasting models, called *Bayesian VARs*, or BVARs. The Minnesota prior is centered on a parameterization called a **random walk**, in which the current value of each variable is equal to its lagged value plus a white noise error term. Thus, the parameter estimates in BVARs are coaxed, but not forced, in the direction of univariate random walks. This sort of stochastic restriction has an immediate shrinkage interpretation, which suggests that it's likely to improve forecast accuracy.[16] This hunch is verified by Doan, Litterman, and Sims (1984), who study forecasting with standard and Bayesian VARs. Ingram and Whiteman (1994) replace the Minnesota prior with a prior derived from macroeconomic theory, and they obtain even better forecasting performance.

6. (Housing starts and completions, continued) Our VAR analysis of housing starts and completions, as always, involved many judgment calls. Using the starts and completions data, assess the adequacy of our models and forecasts. Among other things, you may want to consider the following questions:
   a. Should we allow for a trend in the forecasting model?
   b. How do the results change if, in light of the results of the causality tests, we exclude lags of completions from the starts equation, reestimate by seemingly unrelated regression, and forecast?
   c. Are the VAR forecasts of starts and completions more accurate than univariate forecasts?

7. (Nonlinear regression models I: **functional form** and Ramsey's test) The idea of using powers of a right-hand-side variable to pick up nonlinearity in a regression can also be used to test for linearity of functional form, following Ramsey (1969).

---

[16] Effectively, the shrinkage allows us to recover a large number of degrees of freedom.

If we were concerned that we'd missed some important nonlinearity, an obvious strategy to capture it, based on the idea of a Taylor series expansion of a function, would be to include powers and cross products of the various x variables in the regression. Such a strategy would be wasteful of degrees of freedom, however, particularly if there were more than just one or two right-hand-side variables in the regression and/or if the nonlinearity were severe, so that fairly high powers and interactions would be necessary to capture it. In light of this, Ramsey suggests first fitting a linear regression and obtaining the fitted values, $\hat{y}_t$, $t = 1, \ldots, T$. Then, to test for nonlinearity, we run the regression again with powers of $\hat{y}_t$ included. There is no need to include the first power of $\hat{y}_t$, because that would be redundant with the included x variables. Instead, we include powers $\hat{y}_t^2$, $\hat{y}_t^3, \ldots,$ $\hat{y}_t^m$, where m is a maximum power determined in advance. Note that the powers of $\hat{y}_t$ are linear combinations of powers and cross products of the x variables—just what the doctor ordered. Significance of the included set of powers of $\hat{y}_t$ can be checked using an F-test or an asymptotic likelihood ratio test.

8. (Nonlinear regression models 2: **logarithmic regression models**) We've already seen the use of logarithms in our studies of trend and seasonality. In those setups, however, we had occasion only to take logs of the left-hand-side variable. In more general regression models, such as those that we're studying now, with variables other than trend or seasonals on the right-hand side, it's sometimes useful to take logs of *both* the left- and right-hand-side variables. Doing so allows us to pick up multiplicative nonlinearity. To see this, consider the regression model,

$$y_t = \beta_0 x_t^{\beta_1} e^{\varepsilon_t} .$$

This model is clearly nonlinear due to the multiplicative interactions. Direct estimation of its parameters would require special techniques. Taking natural logs, however, yields the model

$$\ln y_t = \ln \beta_0 + \beta_1 \ln x_t + \varepsilon_t .$$

This transformed model can be immediately estimated by ordinary least squares, by regressing log y on an intercept and log x. Such "log-log regressions" often capture nonlinearities relevant for forecasting, while maintaining the convenience of ordinary least squares.

9. (Nonlinear regression models 3: neural networks) Neural networks amount to a particular nonlinear functional form associated with repeatedly running linear combinations of inputs through nonlinear "squashing" functions. The 0-1 squashing function is useful in classification, and the logistic function is useful for regression.

The neural net literature is full of biological jargon, which serves to obfuscate rather than clarify. We speak, for example, of a "single-output feed-forward neural network with n inputs and 1 hidden layer with q neurons." But the idea is simple. If the output is y and the inputs are x's, we write

$$y_t = \Phi\left(\beta_0 + \sum_{i=1}^{q} \beta_i h_{it}\right) ,$$

where

$$h_{it} = \Psi\left(\gamma_{i0} + \sum_{j=1}^{n} \gamma_{ij} x_{jt}\right), \quad i = 1, \ldots, q$$

are the "neurons" ("hidden units"), and the "activation functions" Ψ and Φ are arbitrary, except that Ψ (the squashing function) is generally restricted to be bounded. (Commonly $\Phi(x) = x$.) Assembling it all, we write

$$y_t = \Phi\left(\beta_0 + \sum_{i=1}^{q} \beta_i \Psi\left(\gamma_{i0} + \sum_{j=1}^{n} \gamma_{ij} x_{jt}\right)\right) = f(x_t; \theta),$$

which makes clear that a neural net is just a particular nonlinear functional form for a regression model.

To incorporate dynamics, we can allow for autoregressive effects in the hidden units. A dynamic ("recurrent") neural network is

$$y_t = \Phi\left(\beta_0 + \sum_{i=1}^{q} \beta_i h_{it}\right),$$

where

$$h_{it} = \Psi\left(\gamma_{i0} + \sum_{j=1}^{n} \gamma_{ij} x_{jt} + \sum_{r=1}^{q} \delta_{ir} h_{r,t-1}\right), i = 1, \ldots, q.$$

Compactly,

$$y_t = \Phi\left(\beta_0 + \sum_{i=1}^{q} \beta_i \Psi\left(\gamma_{i0} + \sum_{j=1}^{n} \gamma_{ij} x_{jt} + \sum_{r=1}^{q} \delta_{ir} h_{r,t-1}\right)\right).$$

Recursive back substitution reveals that y is a nonlinear function of the history of the x's.

$$y_t = g(x^t; \theta),$$

where $x^t = (x_t, \ldots, x_1)$ and $x_t = (x_{1t}, \ldots, x_{nt})$.

The Matlab Neural Network Toolbox implements a variety of networks. The toolbox manual is itself a useful guide to the literature on the practical aspects of constructing and forecasting with neural nets. Kuan and Liu (1995) use a dynamic neural network to predict foreign exchange rates, and Faraway and Chatfield (1995) provide an insightful case study of the efficacy of neural networks in applied forecasting. Ripley (1996) provides a fine and statistically informed (in contrast to much of the neural net literature) survey of the use of neural nets in a variety of fields.

10. **(Spurious regression)** Consider two variables y and x, both of which are highly serially correlated, as are most series in business, finance, and economics. Suppose in addition that y and x are completely unrelated but that we don't know they're unrelated, and we regress y on x using ordinary least squares.
    a. If the usual regression diagnostics (e.g., R^2, t-statistics, F-statistic) were reliable, we'd expect to see small values of all of them. Why?
    b. In fact the opposite occurs; we tend to see large R^2, t-, and F-statistics, and *a very low Durbin-Watson statistic*. Why the low Durbin-Watson? Why, given the low Durbin-Watson, might you *expect* misleading R^2, t-, and F-statistics?
    c. This situation, in which highly persistent series that are in fact unrelated nevertheless appear highly related, is called *spurious regression*. Study of the

phenomenon dates to the early 20th century, and a key study by Granger and Newbold (1974) drove home the prevalence and potential severity of the problem. How might you insure yourself against the spurious regression problem? (*Hint:* Consider allowing for lagged dependent variables, or dynamics in the regression disturbances, as we've advocated repeatedly.)

11. (Comparative forecasting performance of VAR and univariate models) Using the housing starts and completions data on the book's web page, compare the forecasting performance of the VAR used in this chapter with that of the obvious competitor: univariate autoregressions. Use the same in-sample and out-of-sample periods as in the chapter. Why might the forecasting performance of the VAR and univariate methods differ? Why might you expect the VAR completions forecast to outperform the univariate autoregression, but the VAR starts forecast to be no better than the univariate autoregression? Do your results support your conjectures?

Bibliographical and Computational Notes

Some software, such as EViews, automatically accounts for parameter uncertainty when forming conditional regression forecast intervals by using variants of the techniques we introduced in Section 2. Similar but advanced techniques are sometimes used to produce unconditional forecast intervals for dynamic models, such as autoregressions (see Lütkepohl, 1991), but bootstrap simulation techniques are becoming increasingly popular (Efron and Tibshirani, 1993).

Chatfield (1993) argues that innovation uncertainty and parameter estimation uncertainty are likely of minor importance compared to specification uncertainty. We rarely acknowledge specification uncertainty, because we don't know how to quantify "what we don't know we don't know." Quantifying it is a major challenge for future research, and useful recent work in that direction includes Chatfield (1995).

The idea that regression models with serially correlated disturbances are more restrictive than other sorts of transfer function models has a long history in econometrics and engineering and is highlighted in a memorably titled paper, "Serial Correlation as a Convenient Simplification, Not a Nuisance," by Hendry and Mizon (1978). Engineers have scolded econometricians for not using more general transfer function models, as, for example, in Jenkins (1979). But the fact is, as we've seen repeatedly, that generality for generality's sake in business and economic forecasting is not necessarily helpful and can be positively harmful. The shrinkage principle asserts that the imposition of restrictions—even false restrictions—can be helpful in forecasting.

Sims (1980) is an influential paper arguing the virtues of VARs. The idea of predictive causality and associated tests in VARs is due to Granger (1969) and Sims (1972), who build on earlier work by the mathematician Norbert Weiner. Lütkepohl (1991) is a good reference on VAR analysis and forecasting.

Gershenfeld and Weigend (1993) provide a perspective on time series forecasting from the computer science/engineering/nonlinear/neural net perspective, and Swanson and White (1995) compare and contrast a variety of linear and nonlinear forecasting methods.

Concepts for Review

Conditional forecasting model
Scenario analysis
Contingency analysis
Specification uncertainty
Innovation uncertainty
Parameter uncertainty
Unconditional forecast model
Forecasting the right-hand-side variables problem
Distributed lag model
Polynomial distributed lags
Rational distributed lags
Distributed lag regression model with lagged dependent variables
Distributed lag regression model with ARMA disturbances
Transfer function model
Vector autoregression of order *p*
Cross-variable dynamics
Predictive causality
Impulse-response function
Variance decomposition
Cross-correlation function
Feedback
Bayesian analysis
Random walk
Functional form
Logarithmic regression models
Spurious regression

References and Additional Readings

Auerbach, A. J. (1982). "The Index of Leading Indicators: 'Measurement without Theory' Thirty-Five Years Later." *Review of Economics and Statistics*, 64, 589–595.

Burns, A. F., and Mitchell, W. C. (1946). *Measuring Business Cycles.* New York: National Bureau of Economic Research.

Chatfield, C. (1993). "Calculating Interval Forecasts." *Journal of Business and Economic Statistics*, 11, 121–135.

Chatfield, C. (1995). "Model Uncertainty, Data Mining and Statistical Inference (with Discussion)." *Journal of the Royal Statistical Society* A, 158, 419–466.

Diebold, F. X., and Rudebusch, G. D. (1991). "Forecasting Output with the Composite Leading Index: An Ex Ante Analysis." *Journal of the American Statistical Association*, 86, 603–610. Reprinted in Diebold and Rudebusch (1999).

Diebold, F. X., and Rudebusch, G. D. (1996). "Measuring Business Cycles: A Modern Perspective." *Review of Economics and Statistics*. 78, 67–77. Reprinted in Diebold and Rudebusch (1999).

Diebold, F. X., and Rudebusch, G. D. (1999). *Business Cycles: Durations, Dynamics, and Forecasting.* Princeton, N.J.: Princeton University Press.

Doan, T., Litterman, R., and Sims, C. (1984). "Forecasting and Conditional Prediction Using Realistic Prior Distributions." *Econometric Reviews*, 3, 1–144.

Efron, B., and Tibshirani, R. J. (1993). *An Introduction to the Bootstrap.* New York: Chapman and Hall.

Engle, R. F., and Granger, C. W. J. (1987). "Co-Integration and Error Correction: Representation, Estimation and Testing." *Econometrica*, 55, 251–276.

Engle, R. F., and Yoo, B. S. (1987). "Forecasting and Testing in Cointegrated Systems." *Journal of Econometrics*, 35, 143–159.

Faraway, J., and Chatfield, C. (1995). "Time Series Forecasting with Neural Networks: A Case Study." Research Report 95–06. Statistics Group, University of Bath, UK.

Gershenfeld, N. A., and Weigend, A. S. (1993). "The Future of Time Series." In A. S. Weigend and N. A. Gershenfeld (eds.), *Time Series Prediction: Forecasting the Future and Understanding the Past*, 1–70. Reading, Mass.: Addison-Wesley.

Granger, C. W. J. (1969). "Investigating Causal Relations by Econometric Models and Cross-Spectral Methods." *Econometrica*, 37, 424–438.

Granger, C. W. J., and Newbold, P. (1974). "Spurious Regressions in Econometrics." *Journal of Econometrics*, 2, 111–120.

Hamilton, J. D. (1994). *Time Series Analysis*. Princeton, N.J.: Princeton University Press.

Hardy, C. O. (1923). *Risk and Risk Bearing*. Chicago: University of Chicago Press. (Reissued in the *Risk Classics Library*, London: Risk Publications, 1999.)

Hendry, D. F., and Mizon, G. E. (1978). "Serial Correlation as a Convenient Simplification, Not a Nuisance: A Comment on a Study of the Demand for Money by the Bank of England." *Economic Journal*, 88, 549–563.

Ingram, B., and Whiteman, C. (1994). "Supplanting the 'Minnesota' Prior: Forecasting Macroeconomic Time Series Using Real Business Cycle Model Priors." *Journal of Monetary Economics*, 34, 497–510.

Jenkins, G. M. (1979). "Practical Experiences with Modelling and Forecasting Time Series." In O. D. Anderson (ed.), *Forecasting*. Amsterdam: North-Holland.

Johansen, S. (1995). *Likelihood-Based Inference in Cointegrated Vector Autoregressive Models*. Oxford: Oxford University Press.

Klein, L. R. (1981). *Econometric Models as Guides for Decision Making*. New York: Free Press.

Klein, L. R. (1983). *Lectures in Econometrics*. Amsterdam: North-Holland.

Klein, L. R., and Young, R. M. (1980). *An Introduction to Econometric Forecasting and Forecasting Models*. Lexington, Mass.: Heath.

Kuan, C. M., and Liu, Y. (1995). "Forecasting Exchange Rates Using Feedforward and Recurrent Neural Networks." *Journal of Applied Econometrics*, 10, 347–364.

Lütkepohl, H. (1991). *Introduction to Multiple Time Series Analysis*. New York: Springer.

Pagan, A. R., and Robertson, J. (2002). "Forecasting for Policy." In M. P. Clements and D. F. Hendry (eds.), *A Companion to Economic Forecasting*. Oxford: Blackwell.

Pindyck, R. S., and Rubinfeld D. L. (1991). *Econometric Models and Economic Forecasts*. 3rd ed. New York: McGraw-Hill.

Ramsey, J. (1969). "Tests for Specification Errors in Classical Linear Least Squares Regression Analysis." *Journal of the Royal Statistical Society* B, 31, 350–371.

Ripley, B. D. (1996). *Pattern Recognition and Neural Networks*. Oxford: Oxford University Press.

Sanderson, F. H. (1953). *Methods of Crop Forecasting*. Cambridge, Mass.: Harvard University Press.

Sims, C. A. (1972). "Money, Income and Causality." *American Economic Review*, 62, 540–552.

Sims, C. A. (1980). "Macroeconomics and Reality." *Econometrica*, 48, 1–48.

Stock, J. H., and Watson, M. W. (1988). "Variable Trends in Economic Time Series." *Journal of Economic Perspectives*, 2, 147–174.

Swanson, N. R., and White, H. (1995). "A Model-Selection Approach to Assessing the Information in the Term Structure Using Linear-Models and Artificial Neural Networks." *Journal of Business and Economic Statistics*, 13, 265–275.

Wallis, K. F. (1995). "Large-Scale Macroeconometric Modeling." In M. H. Pesaran and M. R. Wickens (eds.), *Handbook of Applied Econometrics*. Oxford: Blackwell.

Zellner, A., and Hong, C. (1989), "Forecasting International Growth Rates Using Bayesian Shrinkage and Other Procedures." *Journal of Econometrics*, 40, 183–202.

Zellner, A., Hong, C., and Min, C.-K. (1991). "Forecasting Turning Points in International Output Growth Rates Using Bayesian Exponentially Weighted Autoregression, Time-Varying Parameter, and Pooling Techniques." *Journal of Econometrics*, 49, 275–304.

Evaluating and Combining Forecasts

As we've stressed repeatedly, good forecasts lead to good decisions. The importance of forecast evaluation and combination techniques follows immediately. Given a track record of forecasts, $y_{t+h,t}$, and corresponding realizations, y_{t+h}, we naturally want to monitor and improve forecast performance. In this chapter we show how to do so. First we discuss evaluation of a single forecast. Second, we discuss the **evaluation and comparison of forecast accuracy.** Third, we discuss whether and how a set of forecasts may be combined to produce a superior composite forecast.

1. Evaluating a Single Forecast

Evaluating a single forecast amounts to checking whether it has the properties expected of an optimal forecast. Denote by y_t the covariance stationary time series to be forecast. The Wold representation is

$$y_t = \mu + \varepsilon_t + b_1\varepsilon_{t-1} + b_2\varepsilon_{t-2} + \cdots$$

$$\varepsilon_t \sim WN(0, \sigma^2) .$$

Thus, the h-step-ahead linear least-squares forecast is

$$y_{t+h,t} = \mu + b_h \varepsilon_t + b_{h+1}\varepsilon_{t-1} + \cdots ,$$

and the corresponding h-step-ahead forecast error is

$$e_{t+h,t} = y_{t+h} - y_{t+h,t} = \varepsilon_{t+h} + b_1\varepsilon_{t+h-1} + \cdots + b_{h-1}\varepsilon_{t+1} ,$$

with variance

$$\sigma_h^2 = \sigma^2 \left(1 + \sum_{i=1}^{h-1} b_i^2\right) .$$

Four key properties of optimal forecasts, which we can easily check, are as follows:

a. Optimal forecasts are unbiased.
b. Optimal forecasts have 1-step-ahead errors that are white noise.
c. Optimal forecasts have h-step-ahead errors that are at most MA($h - 1$).
d. Optimal forecasts have h-step-ahead errors with variances that are nondecreasing in h and that converge to the unconditional variance of the process.

TESTING PROPERTIES OF OPTIMAL FORECASTS

Optimal Forecasts Are Unbiased

If the forecast is unbiased, then the forecast error has a zero mean. A variety of tests of the zero-mean hypothesis can be performed, depending on the assumptions we're willing to maintain. For example, if $e_{t+h,t}$ is Gaussian white noise (as might be reasonably the case for 1-step-ahead errors), then the standard t-test is the obvious choice. We would simply regress the forecast error series on a constant and use the reported t-statistic to test the hypothesis that the population mean is 0. If the errors are non-Gaussian but remain independent and identically distributed (iid), then the t-test is still applicable in large samples.

If the forecast errors are dependent, then more sophisticated procedures are required. Serial correlation in forecast errors can arise for many reasons. Multi-step-ahead forecast errors will be serially correlated, even if the forecasts are optimal, because of the forecast period overlap associated with multi-step-ahead forecasts. More generally, serial correlation in forecast errors may indicate that the forecasts are suboptimal. The upshot is simply that when regressing forecast errors on an intercept, we need to be sure that any serial correlation in the disturbance is appropriately modeled. A reasonable starting point for a regression involving h-step-ahead forecast errors is MA($h - 1$) disturbances, which we'd expect if the forecast were optimal. The forecast may, of course, *not* be optimal, so we don't adopt MA($h - 1$) disturbances uncritically; instead, we try a variety of models using the AIC and SIC to guide selection in the usual way.

Optimal Forecasts Have 1-Step-Ahead Errors That Are White Noise

Under various sets of maintained assumptions, we can use standard tests of the white noise hypothesis. For example, the sample autocorrelation and partial autocorrelation functions, together with Bartlett asymptotic standard errors, are often useful in that regard. Tests based on the first autocorrelation (e.g., the Durbin-Watson test), as well as more general tests, such as the Box-Pierce and Ljung-Box statistics, are useful as well. We implement all of these tests by regression on a constant term.

Optimal Forecasts Have h-Step-Ahead Errors That Are at Most MA($h-1$)

The MA($h-1$) structure implies a cutoff in the forecast error's autocorrelation function beyond displacement $h-1$. This immediately suggests examining the statistical significance of the sample autocorrelations beyond displacement $h-1$ using the Bartlett standard errors. In addition, we can regress the errors on a constant, allowing for MA(q) disturbances with $q > (h-1)$, and test whether the moving-average parameters beyond lag $h-1$ are 0.

Optimal Forecasts Have h-Step-Ahead Errors with Variances That Are Non-Decreasing in h

It's often useful to examine the sample h-step-ahead forecast error variances as a function of h, both to be sure they're nondecreasing in h and to see their *pattern, which often conveys useful information.*

ASSESSING OPTIMALITY WITH RESPECT TO AN INFORMATION SET

The key property of optimal forecast errors, from which all others follow (including those cataloged earlier), is that they should be unforecastable on the basis of information available at the time the forecast was made. This **unforecastability principle** is valid in great generality; it holds, for example, regardless of whether linear-projection optimality or conditional-mean optimality is of interest, regardless of whether the relevant loss function is quadratic, and regardless of whether the series being forecast is stationary.

Many of the tests of properties of optimal forecasts introduced above are based on the unforecastability principle. For example, 1-step-ahead errors had better be white noise, because otherwise we could forecast the errors using information readily available when the forecast is made. Those tests, however, make incomplete use of the unforecastability principle, insofar as they assess only the *univariate* properties of the errors.

We can make a more complete assessment by broadening the information set and assessing *optimality with respect to various sets of information,* by estimating regressions of the form

$$e_{t+h,t} = \alpha_0 + \sum_{i=1}^{k-1} \alpha_i x_{it} + u_t \, .$$

The hypothesis of interest is that all the α's are 0, which is a necessary condition for forecast optimality with respect to the information contained in the x's. The particular case of testing optimality with respect to $y_{t+h,t}$ is very important in practice. The relevant regression is

$$e_{t+h,t} = \alpha_0 + \alpha_1 y_{t+h,t} + u_t \, ,$$

and optimality corresponds to $(\alpha_0, \alpha_1) = (0, 0)$. Keep in mind that the disturbances may be serially correlated, especially if the forecast errors are multistep-ahead, in which case they should be modeled accordingly.

If this regression seems a little strange to you, consider what may seem like a more natural approach to testing optimality, regression of the realization on the forecast:

$$y_{t+h} = \beta_0 + \beta_1 y_{t+h,t} + u_t \, .$$

This is called a *Mincer-Zarnowitz regression*. If the forecast is optimal with respect to the information used to construct it, then we'd expect $(\beta_0, \beta_1) = (0, 1)$, in which case

$$y_{t+h} = y_{t+h,t} + u_t \, .$$

Note, however, that if we start with the regression

$$y_{t+h} = \beta_0 + \beta_1 y_{t+h,t} + u_t \, ,$$

and then subtract $y_{t+h,t}$ from each side, we obtain

$$e_{t+h,t} = \alpha_0 + \alpha_1 y_{t+h,t} + u_t \, ,$$

where $(\alpha_0, \alpha_1) = (0, 0)$ when $(\beta_0, \beta_1) = (0, 1)$. Thus, the two approaches are identical.

2. Evaluating Two or More Forecasts: Comparing Forecast Accuracy

MEASURES OF FORECAST ACCURACY

In practice, it is unlikely that we'll ever stumble upon a fully optimal forecast; instead, situations often arise in which a number of forecasts (all of them suboptimal) are compared and possibly combined. Even for very good forecasts, the actual and forecasted values may be very different. To take an extreme example, note that the linear least-squares forecast for a zero-mean white noise process is simply 0—the paths of forecasts and realizations will look very different, yet there does not exist a better linear forecast under quadratic loss. This highlights the inherent limits to forecastability, which depends on the process being forecast; some processes are inherently easy to forecast, while others are hard to forecast. In other words, sometimes the information on which the forecaster conditions is very valuable, and sometimes it isn't.

The crucial object in measuring forecast accuracy is the loss function, $L(y_{t+h}, y_{t+h,t})$, often restricted to $L(e_{t+h,t})$, which charts the "loss," "cost," or "disutility" associated with various pairs of forecasts and realizations.[1] In addition to the shape of the loss function, the forecast horizon h is of crucial importance. Rankings of forecast accuracy may of course be very different across different loss functions and different horizons.

Let's discuss a few accuracy measures that are important and popular. Accuracy measures are usually defined on the forecast errors, $e_{t+h,t} = y_{t+h} - y_{t+h,t}$, or percent errors, $p_{t+h,t} = (y_{t+h} - y_{t+h,t})/y_{t+h}$. **Mean error,**

$$\text{ME} = \frac{1}{T}\sum_{t=1}^{T} e_{t+h,t} ,$$

measures **bias,** which is one component of accuracy. Other things the same, we prefer a forecast with a small bias. **Error variance,**

$$\text{EV} = \frac{1}{T}\sum_{t=1}^{T} (e_{t+h,t} - \text{ME})^2 ,$$

measures dispersion of the forecast errors. Other things the same, we prefer a forecast whose errors have small variance. Although the mean error and the error variance are components of accuracy, neither provides an overall accuracy measure. For example, one forecast might have a small ME but a large EV, and another might have a large ME and a small EV. Hence, we would like an accuracy measure that somehow incorporates *both* ME and EV. The mean squared error, to which we now turn, does just that.

The most common overall accuracy measures, by far, are **mean squared error,**

$$\text{MSE} = \frac{1}{T}\sum_{t=1}^{T} e_{t+h,t}^2 ,$$

and mean squared percent error,

$$\text{MSPE} = \frac{1}{T}\sum_{t=1}^{T} p_{t+h,t}^2 .$$

Often the square roots of these measures are used to preserve units, yielding the **root mean squared error,**

$$\text{RMSE} = \sqrt{\frac{1}{T}\sum_{t=1}^{T} e_{t+h,t}^2} ,$$

and the root mean squared percent error,

$$\text{RMSPE} = \sqrt{\frac{1}{T}\sum_{t=1}^{T} p_{t+h,t}^2} .$$

---

[1] Because in many applications the loss function will be a direct function of the forecast error, $L(y_t, y_{t+h,t}) = L(e_{t+h,t})$, we write $L(e_{t+h,t})$ from this point on to economize on notation, while recognizing that certain loss functions (such as direction of change) don't collapse to the $L(e_{t+h,t})$ form.

To understand the meaning of "preserving units," and why it's sometimes helpful to do so, suppose that the forecast errors are measured in dollars. Then the mean squared error, which is built up from *squared* errors, is measured in dollars *squared*. Taking square roots—that is, moving from MSE to RMSE—brings the units back to dollars.

MSE can be decomposed into bias and variance components, reflecting the trade-off between bias (ME) and variance (EV) in forecast accuracy under quadratic loss. In particular, MSE can be decomposed into the sum of variance and squared bias (you should verify this),

$$\text{MSE} = \text{EV} + \text{ME}^2 .$$

Somewhat less popular, but nevertheless common, accuracy measures are **mean absolute error,**

$$\text{MAE} = \frac{1}{T}\sum_{t=1}^{T} |e_{t+h,t}| ,$$

and mean absolute percent error,

$$\text{MAPE} = \frac{1}{T}\sum_{t=1}^{T} |p_{t+h,t}| .$$

When using MAE or MAPE, we don't have to take square roots to preserve units. Why?

STATISTICAL COMPARISON OF FORECAST ACCURACY

All the accuracy measures we've discussed are actually *sample estimates* of *population* accuracy. Population MSE, for example, is defined as the *expected* squared error,

$$\text{MSE}_{pop} = E\left(e^2_{t+h,t}\right) ,$$

which we estimate by replacing the expectation with a sample average,

$$\text{MSE} = \frac{1}{T}\sum_{t=1}^{T} e^2_{t+h,t} ,$$

yielding the sample MSE.

Once we've decided on a loss function, it is often of interest to know whether one forecast is more accurate than another. In hypothesis-testing terms, we might want to test the equal accuracy hypothesis,

$$E\left(L(e^a_{t+h,t})\right) = E\left(L(e^b_{t+h,t})\right) ,$$

against the alternative hypothesis that one or the other is better. Equivalently, we might want to test the hypothesis that the expected loss differential is 0,

$$E(d_t) = E\left(L(e^a_{t+h,t})\right) - E\left(L(e^b_{t+h,t})\right) = 0 .$$

The hypothesis concerns population expected loss; we test it using sample average loss.

In fact, we can show that if d_t is a covariance stationary series, then the large-sample distribution of the sample mean loss differential is[2]

$$\sqrt{T}(\bar{d} - \mu) \sim N(0, f),$$

where $\bar{d} = \frac{1}{T}\sum_{t=1}^{T} \left(L\left(e^{a}_{t+h,t}\right) - \left(e^{b}_{t+h,t}\right)\right)$ is the sample mean loss differential, f is the variance of the sample mean loss differential, and μ is the population mean loss differential. This implies that in large samples, under the null hypothesis of a zero population mean loss differential, the standardized sample mean loss differential has a standard normal distribution,

$$B = \frac{\bar{d}}{\sqrt{\frac{\hat{f}}{T}}} \sim N(0, 1),$$

where $\hat{f}$ is a consistent estimator of f. In practice, using $\hat{f} = \sum_{\tau=-M}^{M} \hat{\gamma}_d(\tau)$, where $M = T^{1/3}$ and $\hat{\gamma}_d(\tau)$ denotes the sample autocovariance of the loss differential at displacement τ, provides an adequate estimator in many cases.

Note that the statistic B is just a t-statistic for the hypothesis of a zero population mean loss differential, adjusted to reflect the fact that the loss differential series is not necessarily white noise. We can compute it by regressing the loss differential series on an intercept, taking care to correct the equation for serial correlation. The procedure outlined here amounts to a "nonparametric" way of doing so. It's called *nonparametric* because instead of assuming a particular model for the serial correlation, we use the sample autocorrelations of the loss differential directly.

The nonparametric serial correlation correction is a bit tedious, however, and it involves the rather arbitrary selection of the truncation lag, M. Alternatively, and perhaps preferably, we can proceed by regressing the loss differential on an intercept, allowing for ARMA(p, q) disturbances, and using information criteria to select p and q. This model-based parametric serial correlation correction is easy to do, economizes on degrees of freedom, and makes use of convenient model selection procedures.

3. Forecast Encompassing and Forecast Combination

In forecast accuracy comparison, we ask which forecast is best with respect to a particular loss function. Such "horse races" arise constantly in practical work. Regardless of whether one forecast is significantly better than the others, however, the question arises as to whether competing forecasts may be fruitfully combined to produce a composite forecast superior to all the original forecasts. Thus, forecast combination, although obviously related to forecast accuracy comparison, is logically distinct and of independent interest.

---

[2] We simply assert the result, a proof of which is beyond the scope of this book.

FORECAST ENCOMPASSING

We use **forecast encompassing** tests to determine whether one forecast incorporates (or encompasses) all the relevant information in competing forecasts. If one forecast incorporates all the relevant information, nothing can be gained by combining forecasts. For simplicity, let's focus on the case of two forecasts, $y^a_{t+h,t}$ and $y^b_{t+h,t}$. Consider the regression

$$y_{t+h} = \beta_a y^a_{t+h,t} + \beta_b y^b_{t+h,t} + \varepsilon_{t+h,t} .$$

If $(\beta_a, \beta_b) = (1, 0)$, we'll say that model *a* forecast-encompasses model *b*, and if $(\beta_a, \beta_b) = (0, 1)$, we'll say that model *b* forecast-encompasses model *a*. For other (β_a, β_b) values, neither model encompasses the other, and both forecasts contain useful information about y_{t+h}. In covariance stationary environments, encompassing hypotheses can be tested using standard methods.[3] If neither forecast encompasses the other, forecast combination is potentially desirable.

FORECAST COMBINATION

Failure of each model's forecasts to encompass other model's forecasts indicates that both models are misspecified and that there may be gains from **forecast combination.** It should come as no surprise that such situations are typical in practice, because forecasting models are *likely* to be misspecified—they are intentional abstractions of a much more complex reality.

Many combining methods have been proposed, and they fall roughly into two groups, "variance-covariance" methods and "regression" methods. As we'll see, the **variance-covariance forecast combination method** is in fact a special case of the **regression-based forecast combination method,** so there's really only one method. However, for historical reasons—and, more important, to build valuable intuition—it's important to understand the variance-covariance forecast combination, so let's begin with it. Suppose we have two unbiased forecasts from which we form a composite as

$$y^c_{t+h,t} = \omega y^a_{t+h,t} + (1 - \omega) y^b_{t+h,t} .$$

Because the weights sum to unity, the composite forecast will necessarily be unbiased. Moreover, the combined forecast error will satisfy the same relation as the combined forecast; that is,

$$e^c_{t+h,t} = \omega e^a_{t+h,t} + (1 - \omega) e^b_{t+h,t} ,$$

with variance $\sigma_c^2 = \omega^2 \sigma_{aa}^2 + (1 - \omega)^2 \sigma_{bb}^2 + 2\omega(1 - \omega)\sigma_{ab}^2$, where σ_{aa}^2 and σ_{bb}^2 are the forecast error variances and σ_{ab}^2 is their covariance. We find the optimal combining weight by minimizing the variance of the combined forecast error with respect to ω, which yields

---

[3] Note that $\varepsilon_{t+h,t}$ may be serially correlated, particularly if $h > 1$, and any such serial correlation should be accounted for.

$$\omega^* = \frac{\sigma_{bb}^2 - \sigma_{ab}^2}{\sigma_{bb}^2 + \sigma_{aa}^2 - 2\sigma_{ab}^2}.$$

The optimal combining weight is a simple function of the variances and covariances of the underlying forecast errors. The forecast error variance associated with the optimally combined forecast is less than or equal to the smaller of σ_{aa}^2 and σ_{bb}^2; thus, in population, we have nothing to lose by combining forecasts and potentially much to gain. In practical applications, the unknown variances and covariances that underlie the optimal combining weights are unknown, so we replace them with consistent estimates. That is, we estimate ω^* by replacing σ_{ij}^2 with $\hat{\sigma}_{ij}^2 = \frac{1}{T}\sum_{t=1}^{T} e_{t+h,t}^i e_{t+h,t}^j$, yielding the combining weight estimates,

$$\hat{\omega}^* = \frac{\hat{\sigma}_{bb}^2 - \hat{\sigma}_{ab}^2}{\hat{\sigma}_{bb}^2 + \hat{\sigma}_{aa}^2 - 2\hat{\sigma}_{ab}^2}.$$

To gain intuition for the formula that defines the optimal combining weight, consider the special case in which the forecast errors are uncorrelated, so that $\sigma_{ab}^2 = 0$. Then

$$\omega^* = \frac{\sigma_{bb}^2}{\sigma_{bb}^2 + \sigma_{aa}^2}.$$

As σ_{aa}^2 approaches 0, forecast *a* becomes progressively more accurate. The formula for ω^* indicates that as σ_{aa}^2 approaches 0, ω^* approaches 1, so that all weight is put on forecast *a*, which is desirable. Similarly, as σ_{bb}^2 approaches 0, forecast *b* becomes progressively more accurate. The formula for ω^* indicates that as σ_{bb}^2 approaches 0, ω^* approaches 0, so that all weight is put on forecast *b*, which is also desirable. In general, the forecast with the smaller error variance receives the higher weight, with the precise size of the weight depending on the disparity between variances.

The full formula for the optimal combining weight indicates that the variances *and* the covariance are relevant, but the basic intuition remains valid. Effectively, we're forming a portfolio of forecasts, and as we know from standard results in finance, the optimal shares in a portfolio depend on the variances *and* covariances of the underlying assets.

Now consider the regression method of forecast combination. The form of forecast-encompassing regressions immediately suggests combining forecasts by simply regressing realizations on forecasts. This intuition proves accurate, and in fact the optimal variance-covariance combining weights have a regression interpretation as the coefficients of a linear projection of y_{t+h} onto the forecasts, subject to two constraints: the weights sum to unity, and the intercept is excluded.

In practice, of course, population linear projection is impossible, so we simply run the regression on the available data. Moreover, it's usually preferable *not* to force the weights to add to unity, or to exclude an intercept. Inclusion of an intercept, for example, facilitates bias correction and allows biased forecasts to be combined. Typically, then, we simply estimate the regression,

$$y_{t+h} = \beta_0 + \beta_1 y^a_{t+h,t} + \beta_2 y^b_{t+h,t} + \varepsilon_{t+h,t} \,.$$

Extension to the fully general case of more than two forecasts is immediate.

In general, the regression method is simple and flexible. There are many variations and extensions, because any regression tool is potentially applicable. The key is to use generalizations with sound motivation. We'll give four examples in an attempt to build an intuitive feel for the sorts of extensions that are possible: time-varying combining weights, dynamic combining regressions, shrinkage of combining weights toward equality, and nonlinear combining regressions.

Time-Varying Combining Weights

Relative accuracies of different forecasts may change, and if they do, we naturally want to weight the improving forecasts progressively more heavily and the worsening forecasts less heavily. Relative accuracies can change for a number of reasons. For example, the design of a particular forecasting model may make it likely to perform well in some situations, but poorly in others. Alternatively, people's decision rules and firms' strategies may change over time, and certain forecasting techniques may be relatively more vulnerable to such change.

We allow for time-varying combining weights in the regression framework by using weighted or rolling estimation of combining regressions, or by allowing for explicitly time-varying parameters. If, for example, we suspect that the combining weights are evolving over time in a trendlike fashion, we might use the combining regression

$$y_{t+h} = (\beta_0^0 + \beta_0^1 \mathrm{TIME}) + (\beta_a^0 + \beta_a^1 \mathrm{TIME}) y^a_{t+h,t} + (\beta_b^0 + \beta_b^1 \mathrm{TIME}) y^b_{t+h,t} + \varepsilon_{t+h,t} \,,$$

which we estimate by regressing the realization on an intercept, time, each of the two forecasts, the product of time and the first forecast, and the product of time and the second forecast. We assess the importance of time variation by examining the size and statistical significance of the estimates of β_0^1, β_a^1, and β_b^1.

Serial Correlation

It's a good idea to allow for serial correlation in combining regressions, for two reasons. First, as always, even in the best of conditions we need to allow for the usual serial correlation induced by overlap when forecasts are more than 1-step-ahead. This suggests that instead of treating the disturbance in the combining regression as white noise, we should allow for MA($h - 1$) serial correlation,

$$y_{t+h} = \beta_0 + \beta_a y^a_{t+h,t} + \beta_b y^b_{t+h,t} + \varepsilon_{t+h,t}$$

$$\varepsilon_{t+h,t} \sim \mathrm{MA}(h-1) \,.$$

Second, and very important, the MA($h - 1$) error structure is associated with forecasts that are optimal with respect to their information sets, of which there's no guarantee. That is, although the primary forecasts were designed to capture the dynamics in y, there's no guarantee that they do so. Thus, just as in standard regressions, it's important in combining regressions that we allow either for serially correlated disturbances or for lagged dependent variables, to capture any dynamics in y not captured by the various forecasts. A combining regression with ARMA(p, q) disturbances,

$$y_{t+h} = \beta_0 + \beta_a y^a_{t+h,t} + \beta_b y^b_{t+h,t} + \varepsilon_{t+h,t}$$

$$\varepsilon_{t+h,t} \sim \text{ARMA}(p, q)\,,$$

with p and q selected using information criteria in conjunction with other diagnostics, is usually adequate.

Shrinkage of Combining Weights toward Equality

Simple arithmetic averages of forecasts—that is, combinations in which the weights are constrained to be equal—sometimes perform very well in out-of-sample forecast competitions, even relative to "optimal" combinations. The equal-weights constraint eliminates sampling variation in the combining weights at the cost of possibly introducing bias. Sometimes the benefits of imposing equal weights exceed the cost, so that the MSE of the combined forecast is reduced.

The equal-weights constraint associated with the arithmetic average is an example of extreme shrinkage; regardless of the information contained in the data, the weights are forced into equality. We've seen before that shrinkage can produce forecast improvements, but typically we want to *coax* estimates in a particular direction, rather than to force them. In that way we guide our parameter estimates toward reasonable values when the data are uninformative, while nevertheless paying a great deal of attention to the data when they are informative.

Thus, instead of imposing a *deterministic* equal-weights constraint, we might like to impose a *stochastic* constraint. With this in mind, we sometimes coax the combining weights toward equality without forcing equality. A simple way to do so is to take a weighted average of the simple average combination and the least-squares combination. Let the shrinkage parameter γ be the weight put on the simple average combination, and let $(1 - \gamma)$ be the weight put on the least-squares combination, where γ is chosen by the user. The larger is γ, the more the combining weights are shrunken toward equality. Thus, the combining weights are coaxed toward the arithmetic mean, but the data are still allowed to speak, when they have something important to say.

Nonlinear Combining Regressions

There is no reason to force linearity of combining regressions, and various of the nonlinear techniques that we've already introduced may be used. We might, for example, regress realizations not only on forecasts but also on squares and cross products of the various forecasts, to capture quadratic deviations from linearity,

$$y_{t+h} = \beta_0 + \beta_a y^a_{t+h,t} + \beta_b y^b_{t+h,t} + \beta_{aa}\left(y^a_{t+h,t}\right)^2 + \beta_{bb}\left(y^b_{t+h,t}\right)^2$$

$$+ \beta_{ab} y^a_{t+h,t} y^b_{t+h,t} + \varepsilon_{t+h,t}\,.$$

We assess the importance of nonlinearity by examining the size and statistical significance of estimates of β_{aa}, β_{bb}, and β_{ab}: if the linear combining regression is adequate, those estimates should differ significantly from 0. If, on the other hand, the nonlinear terms are found to be important, then the full nonlinear combining regression should be used.

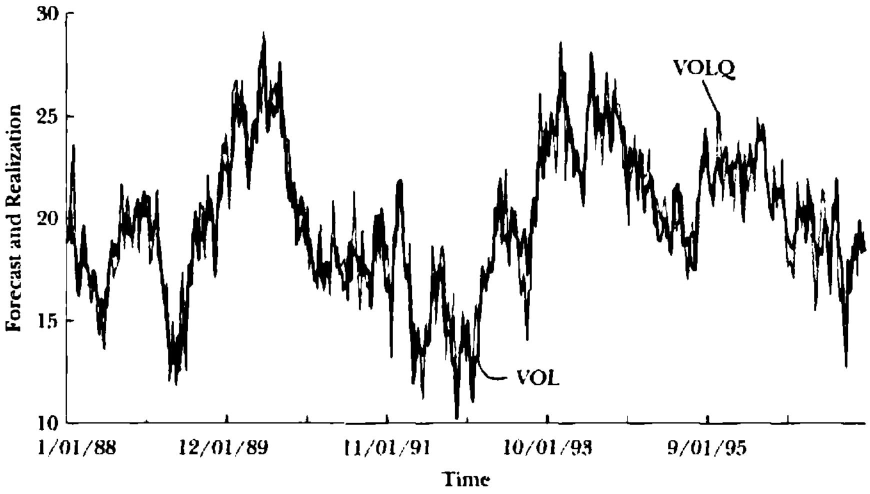
4. Application: OverSea Shipping Volume on the Atlantic East Trade Lane

OverSea Services, Inc., is a major international cargo shipper. To help guide fleet allocation decisions, each week OverSea makes forecasts of volume shipped over each of its major trade lanes, at horizons ranging from 1 week ahead through 16 weeks ahead. In fact, OverSea produces two sets of forecasts: a quantitative forecast is produced using modern quantitative techniques, and a judgmental forecast is produced by soliciting the opinion of the sales representatives, many of whom have years of valuable experience.

Here we'll examine the realizations and 2-week-ahead forecasts of volume on the Atlantic East trade lane (North America to Europe). We have nearly 10 years of data on weekly realized volume (VOL) and weekly 2-week-ahead forecasts (the quantitative forecast VOLQ, and the judgmental forecast VOLJ), from January 1988 through mid-July 1997, for a total of 499 weeks.

In Figure 12.1, we plot realized volume versus the quantitative forecast; in Figure 12.2, we show realized volume versus the judgmental forecast. The two

FIGURE 12.1
Shipping Volume, Quantitative Forecast and Realization



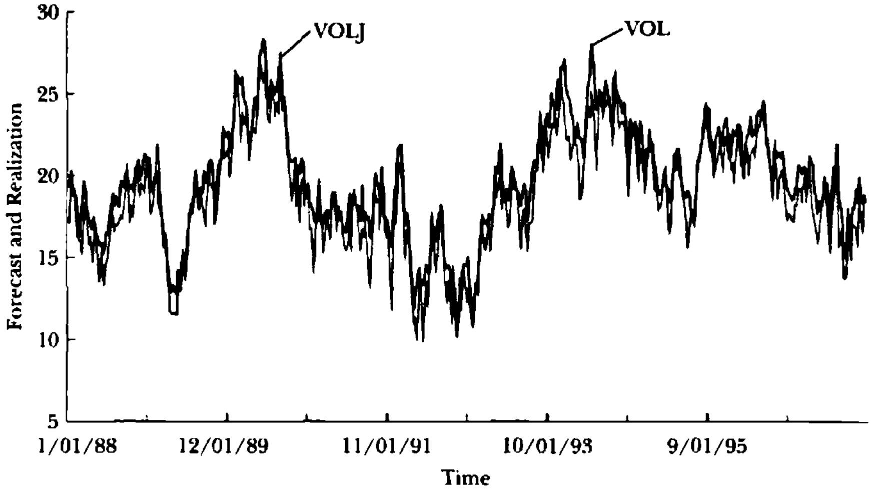


FIGURE 12.2
Shipping Volume, Judgmental Forecast and Realization

plots look similar, and both forecasts appear quite accurate; it's not too hard to forecast shipping volume just two weeks ahead.

In Figures 12.3 and 12.4, we plot the errors from the quantitative and judgmental forecasts, which are more revealing. The quantitative error, in particular, appears roughly centered on 0, whereas the judgmental error seems to be a bit higher than 0 on average. That is, the judgmental forecast appears biased in a

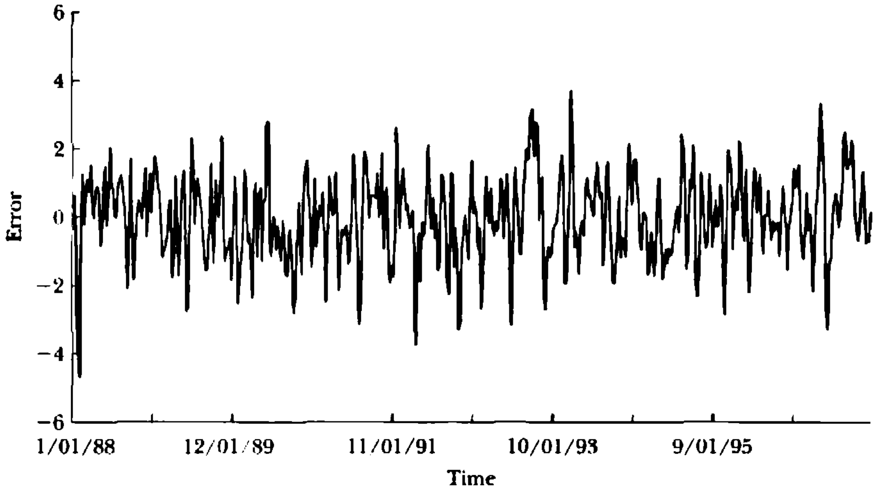
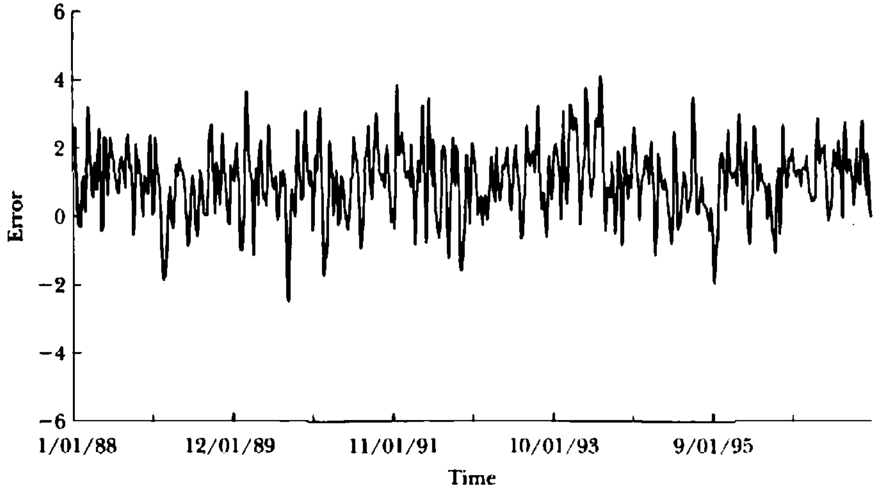


FIGURE 12.3
Quantitative Forecast Error

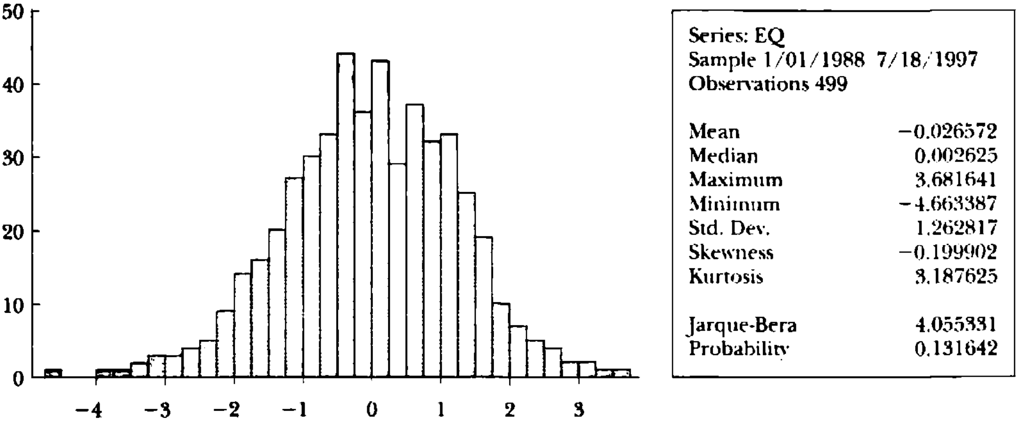
FIGURE 12.4 *Judgmental Forecast Error*



pessimistic way—on average, actual realized volume is a bit higher than forecasted volume.

Figures 12.5 and 12.6 present histograms and related statistics for the quantitative and judgmental forecast errors. The histograms confirm our earlier suspicions based on the error plots; the histogram for the quantitative error is centered on a mean of –0.03, whereas that for the judgmental error is centered on 1.02. The error standard deviations, however, reveal that the

FIGURE 12.5 *Histogram and Related Statistics, Quantitative Forecast Error*

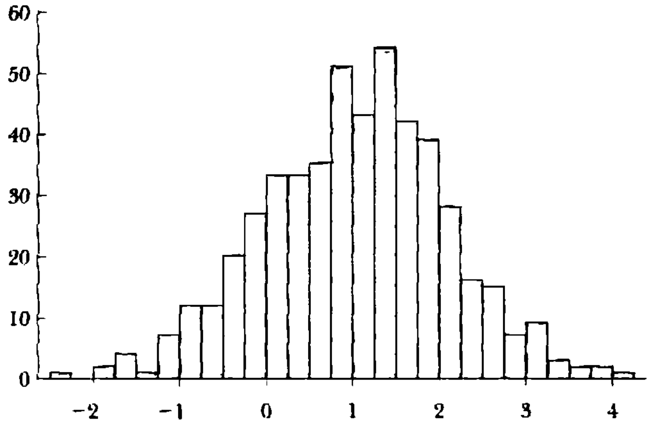


```
Series: EQ
Sample 1/01/1988 7/18/1997
Observations 499

Mean            -0.026572
Median           0.002625
Maximum          3.681641
Minimum         -4.663387
Std. Dev.        1.262817
Skewness        -0.199902
Kurtosis         3.187625

Jarque-Bera      4.055331
Probability      0.131642
```

FIGURE 12.[illegible] *Histogram and Related Statistics, Judgmental Forecast Error*



Series: EJ
Sample 1/01/1988 7/18/1997
Observations 499

Mean	1.023744
Median	1.060523
Maximum	4.100623
Minimum	−2.481030
Std. Dev.	1.063681
Skewness	−0.106148
Kurtosis	3.078680
Jarque-Bera	1.065789
Probability	0.586904

judgmental forecast errors vary a bit less around their mean than do the quantitative errors. Finally, the Jarque-Bera test can't reject the hypothesis that the errors are normally distributed.

In Tables 12.1 and 12.2 and Figures 12.7 and 12.8, we show the correlograms of the quantitative and judgmental forecast errors. In each case, the errors appear to have MA(1) structure; the sample autocorrelations cut off at displacement 1, whereas the sample partial autocorrelations display damped oscillation, which is reasonable for 2-step-ahead forecast errors.

To test for the statistical significance of bias, we need to account for the MA(1) serial correlation. To do so, we regress the forecast errors on a constant, allowing for MA(1) disturbances. We show the results for the quantitative forecast errors in Table 12.3 and those for the judgmental forecast errors

TABLE 12.1 *Correlogram, Quantitative Forecast Error*

Sample: 1/01/1988 7/18/1997
Included observations: 499

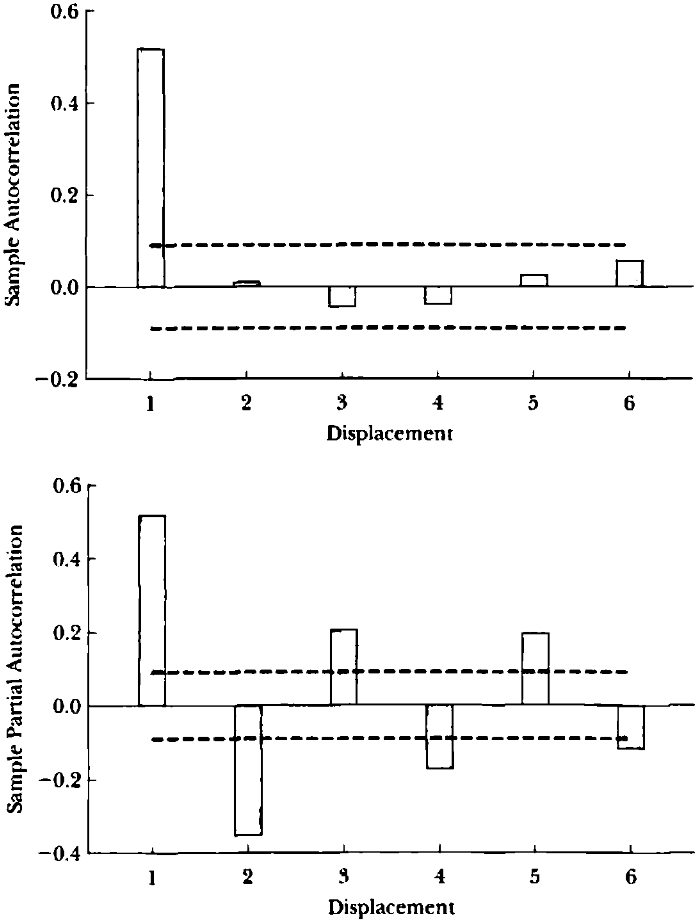
	Acorr.	P. Acorr.	Std. Error	Ljung-Box	*p*-value
1	0.518	0.518	.045	134.62	0.000
2	0.010	−0.353	.045	134.67	0.000
3	−0.044	0.205	.045	135.65	0.000
4	−0.039	−0.172	.045	136.40	0.000
5	0.025	0.195	.045	136.73	0.000
6	0.057	−0.117	.045	138.36	0.000

TABLE 12.2
Correlogram, Judgmental Forecast Error

Sample: 1/01/1988 7/18/1997
Included observations: 499

	Acorr.	P. Acorr.	Std. Error	Ljung-Box	*p*-value
1	0.495	0.495	.045	122.90	0.000
2	−0.027	−0.360	.045	123.26	0.000
3	−0.045	0.229	.045	124.30	0.000
4	−0.056	−0.238	.045	125.87	0.000
5	−0.033	0.191	.045	126.41	0.000
6	0.087	−0.011	.045	130.22	0.000

FIGURE 12.7
Sample Autocorrelations and Partial Autocorrelations, Quantitative Forecast Error



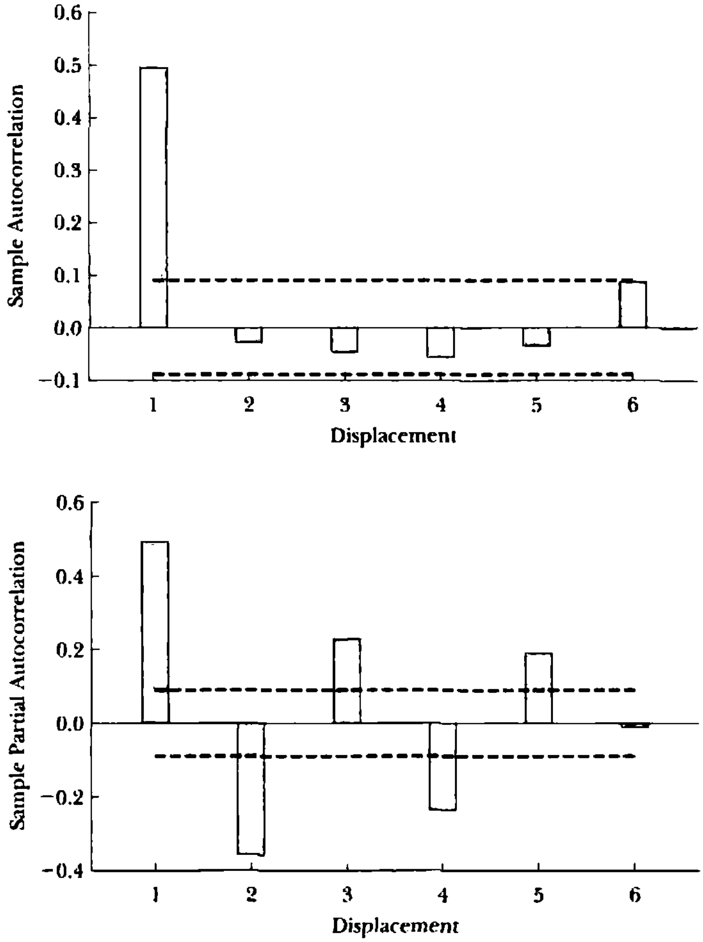


FIGURE 12.8 *Sample Autocorrelations and Partial Autocorrelations, Judgmental Forecast Error*

in Table 12.4. The *t*-statistic indicates no bias in the quantitative forecasts but sizeable and highly statistically significant bias in the judgmental forecasts.

In Tables 12.5 and 12.6, we show the results of Mincer-Zarnowitz regressions; both forecasts fail miserably. We expected the judgmental forecast to fail, because it's biased, but until now no defects were found in the quantitative forecast.

Now let's compare forecast accuracy. We show the histogram and descriptive statistics for the squared quantitative and judgmental errors in Figures 12.9 and 12.10. The histogram for the squared judgmental error is pushed rightward relative to that of the quantitative error, due to bias. The RMSE of the quantitative forecast is 1.26, while that of the judgmental forecast is 1.48.

TABLE 12.3
Quantitative Forecast Error, Regression on Intercept, MA(1) Disturbances

LS // Dependent variable is EQ.
Sample: 1/01/1988 7/18/1997
Included observations: 499
Convergence achieved after 6 iterations

Variable	Coefficient	Std. Error	*t*-Statistic	Prob.
C	−0.024770	0.079851	−0.310200	0.7565
MA(1)	0.935393	0.015850	59.01554	0.0000

R^2	0.468347	Mean dependent var.	−0.026572
Adjusted R^2	0.467277	SD dependent var.	1.262817
SE of regression	0.921703	Akaike info criterion	−0.159064
Sum squared resid.	422.2198	Schwarz criterion	−0.142180
Log likelihood	−666.3639	*F*-statistic	437.8201
Durbin-Watson stat.	1.988237	Prob(*F*-statistic)	0.000000
Inverted MA roots	−.94		

Figure 12.11 shows the (quadratic) loss differential; it's fairly small but looks a little negative. Figure 12.12 shows the histogram of the loss differential; the mean is −0.58, which is small relative to the standard deviation of the loss differential, but remember that we have not yet corrected for serial correlation. In Table 12.7 we show the correlogram of the loss differential, which strongly suggests MA(1) structure. The sample autocorrelations and partial

TABLE 12.4
Judgmental Forecast Error, Regression on Intercept, MA(1) Disturbances

LS // Dependent variable is EJ.
Sample: 1/01/1988 7/18/1997
Included observations: 499
Convergence achieved after 7 iterations

Variable	Coefficient	Std. Error	*t*-Statistic	Prob.
C	1.026372	0.067191	15.27535	0.0000
MA(1)	0.961524	0.012470	77.10450	0.0000

R^2	0.483514	Mean dependent var.	1.023744
Adjusted R^2	0.482475	SD dependent var.	1.063681
SE of regression	0.765204	Akaike info criterion	−0.531226
Sum squared resid.	291.0118	Schwarz criterion	−0.514342
Log likelihood	−573.5094	*F*-statistic	465.2721
Durbin-Watson stat.	1.968750	Prob(*F*-statistic)	0.000000
Inverted MA roots	−.96		

TABLE 12.5
Mincer-Zarnowitz Regression, Quantitative Forecast Error

LS // Dependent variable is VOL.
Sample: 1/01/1988 7/18/1997
Included observations: 499
Convergence achieved after 10 iterations

Variable	Coefficient	Std. Error	t-Statistic	Prob.
C	2.958191	0.341841	8.653696	0.0000
VOLQ	0.849559	0.016839	50.45317	0.0000
MA(1)	0.912559	0.018638	48.96181	0.0000

R^2	0.936972	Mean dependent var.	19.80609
Adjusted R^2	0.936718	SD dependent var.	3.403283
SE of regression	0.856125	Akaike info criterion	−0.304685
Sum squared resid.	363.5429	Schwarz criterion	−0.279358
Log likelihood	−629.0315	F-statistic	3686.790
Durbin-Watson stat.	1.815577	Prob(F-statistic)	0.000000
Inverted MA roots	−.91		

Wald test:
Null hypothesis: $C(1) = 0$ $C(2) = 1$

F-statistic	39.96862	Probability	0.000000
Chi-square	79.93723	Probability	0.000000

TABLE 12.6
Mincer-Zarnowitz Regression, Judgmental Forecast Error

LS // Dependent variable is VOL.
Sample: 1/01/1988 7/18/1997
Included observations: 499
Convergence achieved after 11 iterations

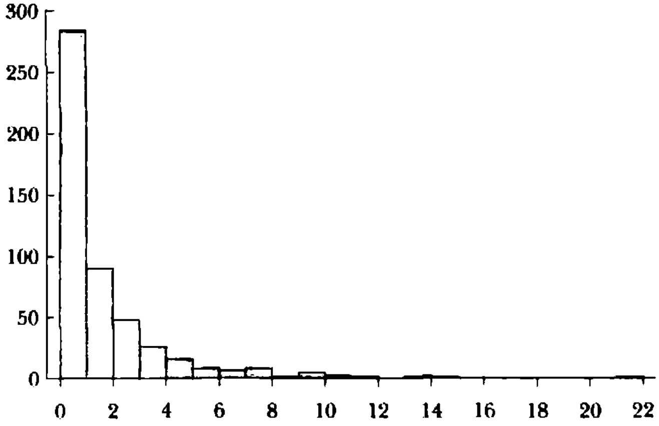
Variable	Coefficient	Std. Error	t-Statistic	Prob.
C	2.592648	0.271740	9.540928	0.0000
VOLJ	0.916576	0.014058	65.20021	0.0000
MA(1)	0.949690	0.014621	64.95242	0.0000

R^2	0.952896	Mean dependent var.	19.80609
Adjusted R^2	0.952706	SD dependent var.	3.403283
SE of regression	0.740114	Akaike info criterion	−0.595907
Sum squared resid.	271.6936	Schwarz criterion	−0.570581
Log likelihood	−556.3715	F-statistic	5016.993
Durbin-Watson stat.	1.917179	Prob(F-statistic)	0.000000
Inverted MA roots	−.95		

Wald test:
Null hypothesis: $C(1) = 0$ $C(2) = 1$

F-statistic	143.8323	Probability	0.000000
Chi-square	287.6647	Probability	0.000000

FIGURE 12.9 *Histogram and Related Statistics, Squared Quantitative Forecast Error*



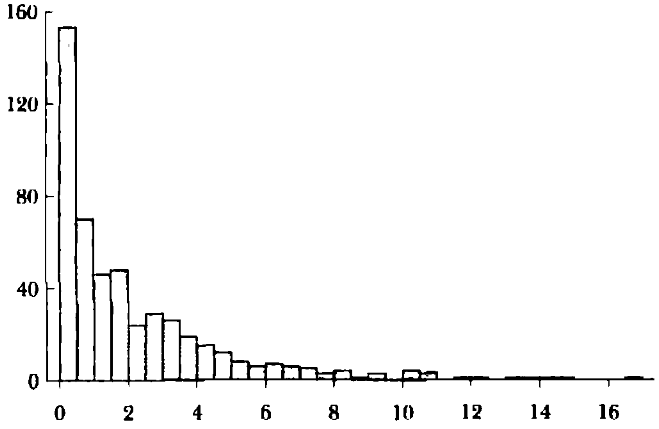
Series: EQSQ
Sample 1/01/1988 7/18/1997
Observations 499

Mean	1.592217
Median	0.763750
Maximum	21.74718
Minimum	5.61E-06
Std. Dev.	2.369751
Skewness	3.293315
Kurtosis	18.88079
Jarque-Bera	6145.666
Probability	0.000000

autocorrelations, presented in Figure 12.13, confirm that impression. Thus, to test for significance of the loss differential, we regress it on a constant and allow for MA(1) disturbances; the results appear in Table 12.8. The mean loss differential is highly statistically significant, with a *p*-value less than 0.01; we conclude that the quantitative forecast is more accurate than the judgmental forecast under quadratic loss.

Now let's combine the forecasts. Both failed Mincer-Zarnowitz tests, which suggests that there may be scope for combining. The correlation between the

FIGURE 12.10 *Histogram and Related Statistics, Squared Judgmental Forecast Error*



Series: EJSQ
Sample 1/01/1988 7/18/1997
Observations 499

Mean	2.177201
Median	1.308296
Maximum	16.81511
Minimum	4.63E-05
Std. Dev.	2.623644
Skewness	2.134551
Kurtosis	8.646748
Jarque-Bera	1041.891
Probability	0.000000

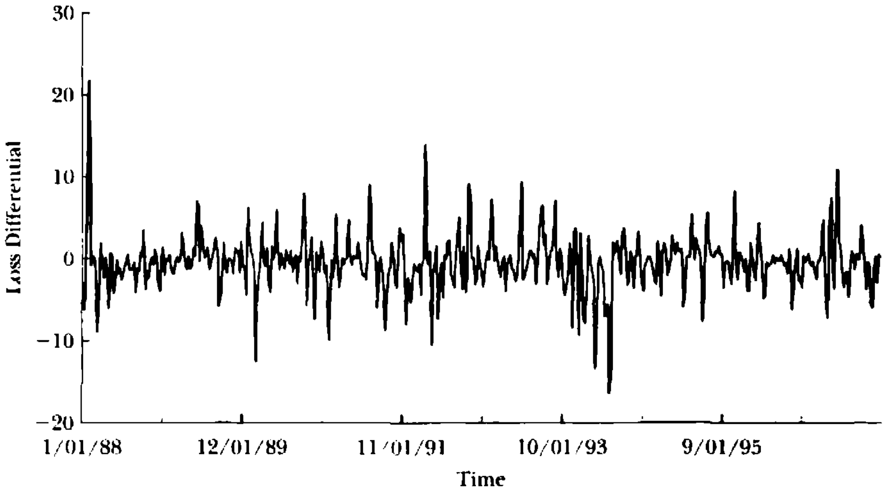
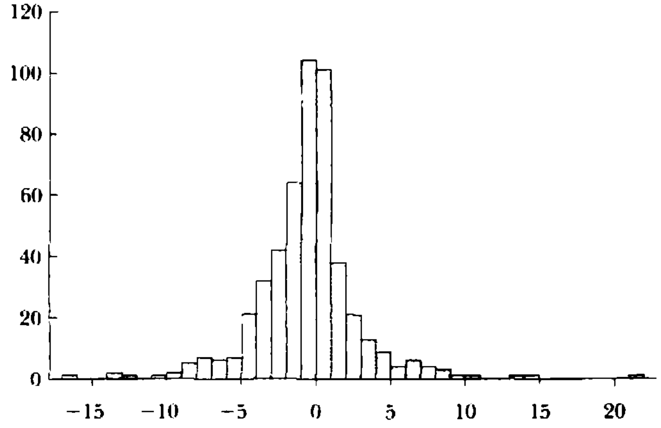


FIGURE 12.11
Loss Differential

two forecast errors is 0.54, positive but not too high. In Table 12.9, we show the results of estimating the unrestricted combining regression with MA(1) errors (equivalently, a forecast encompassing test). Neither forecast encompasses the other; both combining weights, as well as the intercept, are highly statistically significantly different from 0. Interestingly, the judgmental forecast actually gets *more* weight than the quantitative forecast in the combination, in spite of the fact that its RMSE was higher. That's because, after correcting for bias, the judgmental forecast appears a bit more accurate.

FIGURE 12.12 *Histogram and Related Statistics, Loss Differential*



Series: DD
Sample 1/01/1988 7/18/1997
Observations 499

Mean	−0.584984
Median	−0.395646
Maximum	21.65003
Minimum	−16.50010
Std. Dev.	3.416190
Skewness	0.421513
Kurtosis	9.472586
Jarque-Bera	885.8303
Probability	0.000000

TABLE 12.7
Loss Differential Correlogram

Sample: 1/01/1988 7/18/1997
Included observations: 499

	Acorr.	P. Acorr.	Std. Error	Ljung-Box	*p*-value
1	0.357	0.357	.045	64.113	0.000
2	−0.069	−0.226	.045	66.519	0.000
3	−0.050	0.074	.045	67.761	0.000
4	−0.044	−0.080	.045	68.746	0.000
5	−0.078	−0.043	.045	71.840	0.000
6	0.017	0.070	.045	71.989	0.000

FIGURE 12.13
Sample Autocorrelations and Partial Autocorrelations, Loss Differential

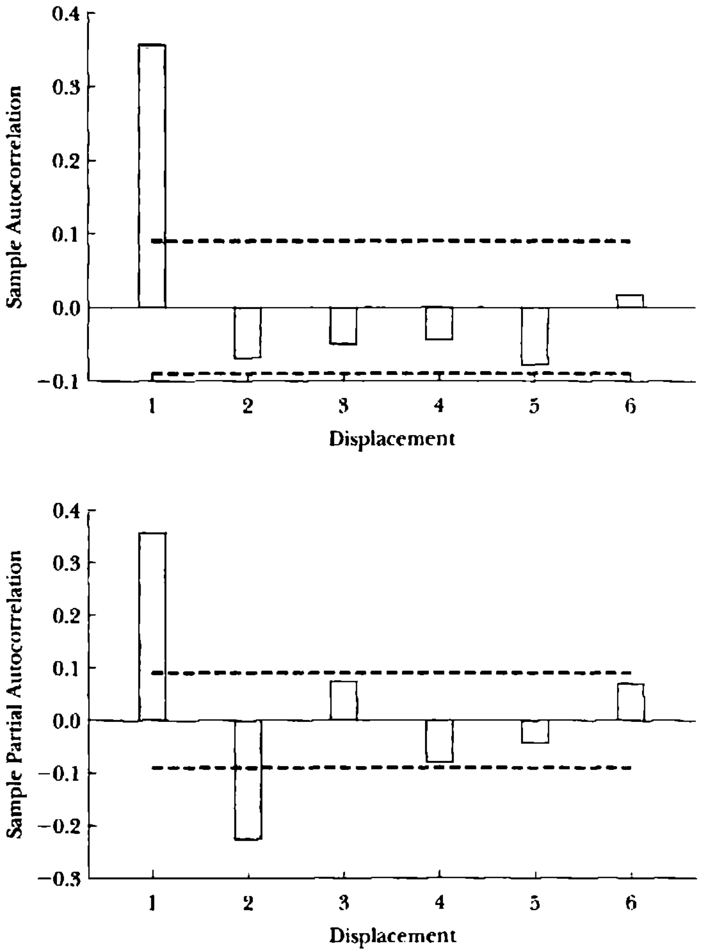


TABLE 12.8
Loss Differential, Regression on Intercept with MA(1) Disturbances

LS // Dependent variable is DD.
Sample: 1/01/1988 7/18/1997
Included observations: 499
Convergence achieved after 4 iterations

Variable	Coefficient	Std. Error	*t*-Statistic	Prob.
C	−0.585333	0.204737	−2.858945	0.0044
MA(1)	0.472901	0.039526	11.96433	0.0000

R^2	0.174750	Mean dependent var.	−0.584984
Adjusted R^2	0.173089	SD dependent var.	3.416190
SE of regression	3.106500	Akaike info criterion	2.270994
Sum squared resid.	4796.222	Schwarz criterion	2.287878
Log likelihood	−1272.663	*F*-statistic	105.2414
Durbin-Watson stat.	2.023606	Prob(*F*-statistic)	0.000000
Inverted MA roots	−.47		

It's interesting to track the RMSEs as we progress from the original forecasts to the combined forecast. The RMSE of the quantitative forecast is 1.26 and that of the judgmental forecast is 1.48. The RMSE associated with using the modified quantitative forecast that we obtain using the weights

TABLE 12.9
Shipping Volume Combining Regression

LS // Dependent variable is VOL.
Sample: 1/01/1988 7/18/1997
Included observations: 499
Convergence achieved after 11 iterations

Variable	Coefficient	Std. Error	*t*-Statistic	Prob.
C	2.181977	0.259774	8.399524	0.0000
VOLQ	0.291577	0.038346	7.603919	0.0000
VOLJ	0.630551	0.039935	15.78944	0.0000
MA(1)	0.951107	0.014174	67.10327	0.0000

R^2	0.957823	Mean dependent var.	19.80609
Adjusted R^2	0.957567	SD dependent var.	3.403283
SE of regression	0.701049	Akaike info criterion	−0.702371
Sum squared resid.	243.2776	Schwarz criterion	−0.668603
Log likelihood	−528.8088	*F*-statistic	3747.077
Durbin-Watson stat.	1.925091	Prob(*F*-statistic)	0.000000
Inverted MA roots	−.95		

estimated in the Mincer-Zarnowitz regression is 0.85, and that of the modified judgmental forecast is 0.74. Finally, the RMSE of the combined forecast is 0.70. In this case, we get a big improvement in forecast accuracy from using the modifications associated with the Mincer-Zarnowitz regressions and a smaller, but nonnegligible, additional improvement from using the full combining regression.[4]

Exercises, Problems, and Complements

1. (Forecast evaluation in action) Discuss in detail how you would use forecast evaluation techniques to address each of the following questions.
   a. Are asset returns (e.g., stocks, bonds, exchange rates) forecastable over long horizons?
   b. Do forward exchange rates provide unbiased forecasts of future spot exchange rates at all horizons?
   c. Are government budget projections systematically too optimistic, perhaps for strategic reasons?
   d. Can interest rates be used to provide good forecasts of future inflation?

2. (Forecast error analysis) You are working for a London-based hedge fund, Thompson Energy Investors, and your boss has assigned you to assess a model used to forecast U.S. crude oil imports. On the last day of each quarter, the model is used to forecast oil imports at 1-quarter-ahead through 4-quarter-ahead horizons. Thompson has done this for each of the past 80 quarters and has kept the corresponding four forecast error series, which appear on the book's web page.
   a. Based on a correlogram analysis, assess whether the 1-quarter-ahead forecast errors are white noise. (Be sure to discuss all parts of the correlogram: sample autocorrelations, sample partial autocorrelations, Bartlett standard errors, and Ljung-Box statistics.) Why care?
   b. Regress each of the four forecast error series on constants, in each case allowing for a MA(5) disturbances. Comment on the significance of the MA coefficients in each of the four cases, and use the results to assess the optimality of the forecasts at each of the four horizons. Does your 1-step-ahead MA(5)-based assessment match the correlogram-based assessment obtained in part *a*? Do the multistep forecasts appear optimal?
   c. Overall, what do your results suggest about the model's ability to predict U.S. crude oil imports?

3. (Combining forecasts) You are a managing director at Paramex, a boutique investment bank in Paris. Each day during the summer your two interns, Alex and Betsy, give you a 1-day-ahead forecast of the euro/dollar exchange rate. At the

---

[4] The RMSEs associated with forecasts from the partial optimality regressions as well as from the full combining regression are of course in-sample RMSEs. It remains to be seen how they'll perform out of sample, but all indications look good.

end of the summer, you calculate each intern's series of daily forecast errors. You find that the mean errors are zero, and the error variances and covariances are $\hat{\sigma}^2_{AA} = 153.76$, $\hat{\sigma}^2_{BB} = 92.16$ and $\hat{\sigma}^2_{AB} = 0.2$.

a. If you were forced to choose between Alex's forecast and Betsy's forecast, which would you choose? Why?

b. If instead you had the opportunity to combine the two forecasts by forming a weighted average, what would be the optimal weights according to the variance-covariance method? Why?

c. Is it guaranteed that a combined forecast formed using the "optimal" weights calculated in part *b* will have lower mean squared prediction error? Why or why not?

4. (Quantitative forecasting, judgmental forecasting, forecast combination, and shrinkage) Interpretation of the modern quantitative approach to forecasting as eschewing judgment is most definitely misguided. How is judgment used routinely and informally to modify quantitative forecasts? How can judgment be formally used to modify quantitative forecasts via forecast combination? How can judgment be formally used to modify quantitative forecasts via shrinkage? Discuss the comparative merits of each approach. Klein (1981) provides insightful discussion of the interaction between judgment and models, as well as the comparative track record of judgmental versus model-based forecasts.

5. (The algebra of forecast combination) Consider the combined forecast,

$$y^c_{t+h,t} = \omega y^a_{t+h,t} + (1-\omega) y^b_{t+h,t} \ .$$

Verify the following claims made in the text:

a. The combined forecast error will satisfy the same relation as the combined forecast; that is,

$$e^c_{t+h,t} = \omega e^a_{t+h,t} + (1-\omega) e^b_{t+h,t} \, .$$

b. Because the weights sum to unity, if the primary forecasts are unbiased, then so, too, is the combined forecast.

c. The variance of the combined forecast error is

$$\sigma_c^2 = \omega^2 \sigma^2_{aa} + (1-\omega)^2 \sigma^2_{bb} + 2\omega(1-\omega)\sigma^2_{ab} \, ,$$

where σ^2_{11} and σ^2_{22} are unconditional forecast error variances and σ^2_{12} is their covariance.

d. The combining weight that minimizes the combined forecast error variance (and hence the combined forecast error MSE, by unbiasedness) is

$$\omega^* = \frac{\sigma^2_{bb} - \sigma^2_{ab}}{\sigma^2_{bb} + \sigma^2_{aa} - 2\sigma^2_{ab}} \ .$$

e. If neither forecast encompasses the other, then

$$\sigma_c^2 < \min\left(\sigma^2_{aa}, \sigma^2_{bb}\right) \ .$$

f. If one forecast encompasses the other, then

$$\sigma_c^2 = \min\left(\sigma^2_{aa}, \sigma^2_{bb}\right) \ .$$

6. (The mechanics of practical forecast evaluation and combination) On the book's web page, you'll find the time series of shipping volume, quantitative forecasts, and judgmental forecasts used in this chapter.
   a. Replicate the empirical results reported in this chapter. Explore and discuss any variations or extensions that you find interesting.
   b. Using the first 250 weeks of shipping volume data, specify and estimate a univariate autoregressive model of shipping volume (with trend and seasonality if necessary), and provide evidence to support the adequacy of your chosen specification.
   c. Use your model each week to forecast 2 weeks ahead, each week estimating the model using all available data, producing forecasts for observations 252 through 499, made using information available at times 250 through 497. Calculate the corresponding series of 248 2-step-ahead recursive forecast errors.
   d. Using the methods of this chapter, evaluate the quality of your forecasts, both in isolation and relative to the original quantitative and judgmental forecasts. Discuss.
   e. Using the methods of this chapter, assess whether your forecasting model can usefully be combined with the original quantitative and judgmental models. Discuss.

7. (What are we forecasting? Preliminary series, revised series, and the limits to forecast accuracy) Many economic series are revised as underlying source data increase in quantity and quality. For example, a typical quarterly series might be issued as follows. First, shortly after the end of the relevant quarter, a "preliminary" value for the current quarter is issued. A few months later, a "revised" value is issued, and a year or so later the "final revised" value is issued. For extensive discussion, see Croushore and Stark (2001).
   a. If you're evaluating the accuracy of a forecast or forecasting technique, you've got to decide on what to use for the "actual" values, or realizations, to which the forecasts will be compared. Should you use the preliminary value? The final revised value? Something else? Be sure to weigh as many relevant issues as possible in defending your answer.
   b. Morgenstern (1963) assesses the accuracy of economic data and reports that the great mathematician Norbert Wiener, after reading an early version of Morgenstern's book, remarked that "economics is a one- or two-digit science." What might Wiener have meant?
   c. Theil (1966) is well aware of the measurement error in economic data; he speaks of "predicting the future and *estimating* the past." Klein (1981) notes that, in addition to the usual innovation uncertainty, measurement error in economic data—even "final revised" data—provides additional limits to measured forecast accuracy. That is, even if a forecast were perfect, so that forecast errors were consistently 0, *measured* forecast errors would be nonzero due to measurement error. The larger the measurement error, the more severe the inflation of measured forecast error. Evaluate.
   d. When assessing improvements (or lack thereof) in forecast accuracy over time, how might you guard against the possibility of spurious assessed improvements due not to true forecast improvement but rather to structural change toward a more "forcastable" process? (On forecastability, see Diebold and Kilian, 2001).

8. (Ex post versus real-time forecast evaluation) If you're evaluating a forecasting model, you've also got to take a stand on precisely what information is available to the forecaster, and when. Suppose, for example, that you're evaluating the forecasting accuracy of a particular regression model.
    a. Do you prefer to estimate and forecast recursively or to simply estimate once using the full sample of data?
    b. Do you prefer to estimate using final-revised values of the left- and right-hand-side variables, or do you prefer to use the preliminary, revised, and final-revised data as it became available in real time?
    c. If the model is explanatory rather than causal, do you prefer to substitute the true realized values of right-hand-side variables or to substitute forecasts of the right-hand-side variables that could actually be constructed in real time?

    These sorts of timing issues can make large differences in conclusions. For an application to using the composite index of leading indicators to forecast industrial production, see Diebold and Rudebusch (1991).

9. (What do we know about the accuracy of macroeconomic forecasts?) Zarnowitz and Braun (1993) provide a fine assessment of the track record of economic forecasts since the late 1960s. Read their article, and try to assess just what we really know about
    a. comparative forecast accuracy at business cycle turning points versus other times;
    b. comparative accuracy of judgmental versus model-based forecasts;
    c. improvements in forecast accuracy over time;
    d. the comparative forecastability of various series;
    e. the comparative accuracy of linear versus nonlinear forecasting models.

    Other well-known and useful comparative assessments of U.S. macroeconomic forecasts have been published over the years by Stephen K. McNees, a private consultant formerly with the Federal Reserve Bank of Boston. McNees (1988) is a good example. Similarly useful studies for the U.K., with particular attention to decomposing forecast error into its various possible sources, have recently been produced by Kenneth F. Wallis and his coworkers at the ESRC Macroeconomic Modelling Bureau at the University of Warwick. Wallis and Whitley (1991) is a good example. Finally, the Model Comparison Seminar, founded by Lawrence R. Klein of the University of Pennsylvania and now led by Michael Donihue of Colby College, is dedicated to the ongoing comparative assessment of macroeconomic forecasting models. Klein (1991) provides a good survey of some of the group's recent work, and more recent information can be found at http://www.colby.edu/economics/faculty/mrdonihu/mcs/.

10. (Forecast evaluation when realizations are unobserved) Sometimes we never see the realization of the variable being forecast. Pesaran and Samiei (1995), for example, develop models for forecasting ultimate resource recovery, such as the total amount of oil in an underground reserve. The actual value, however, won't be known until the reserve is depleted, which may be decades away. Such situations obviously make for difficult accuracy evaluation! How would you evaluate such forecasting models?

11. (Forecast error variances in models with estimated parameters) As we've seen, computing forecast error variances that acknowledge parameter estimation

uncertainty is very difficult; that's one reason why we've ignored it. We've learned a number of lessons about optimal forecasts while ignoring parameter estimation uncertainty, such as that

a. forecast error variance grows as the forecast horizon lengthens, and
b. in covariance stationary environments, the forecast error variance approaches the (finite) unconditional variance as the horizon grows.

Such lessons provide valuable insight and intuition regarding the workings of forecasting models and provide a useful benchmark for assessing actual forecasts. They sometimes need modification, however, when parameter estimation uncertainty is acknowledged. For example, in models with estimated parameters, keep in mind the following points:

a. Forecast error variance needn't grow monotonically with horizon. Typically we *expect* forecast error variance to increase monotonically with horizon, but it doesn't *have* to.
b. Even in covariance stationary environments, the forecast error variance needn't converge to the unconditional variance as the forecast horizon lengthens; instead, it may grow without bound. Consider, for example, forecasting a series that's just a stationary AR(1) process around a linear trend. With known parameters, the point forecast will converge to the trend as the horizon grows, and the forecast error variance will converge to the unconditional variance of the AR(1) process. With estimated parameters, however, if the estimated trend parameters are even the slightest bit different from the true values (as they almost surely will be, due to sampling variation), that error will be magnified as the horizon grows, so the forecast error variance will grow.

Thus, results derived under the assumption of known parameters should be viewed as a benchmark to guide our intuition, rather than as precise rules.

12. (The empirical success of forecast combination) In the text we mentioned that we have nothing to lose by forecast combination and potentially much to gain. That's certainly true in population, with optimal combining weights. However, in finite samples of the size typically available, sampling error contaminates the combining weight estimates, and the problem of sampling error may be exacerbated by the collinearity that typically exists between $y^a_{t+h,t}$ and $y^b_{t+h,t}$. Thus, although we hope to reduce out-of-sample forecast MSE by combining, there is no guarantee. Fortunately, however, in practice forecast combination often leads to very good results. The efficacy of forecast combination is well documented in Clemen's (1989) review of the vast literature, and it emerges clearly in the landmark study by Stock and Watson (1999).

13. (Forecast combination and the Box-Jenkins paradigm) In an influential book, Box and Jenkins (latest edition is Box, Jenkins, and Reinsel, 1994) envision an ongoing, iterative process of model selection and estimation, forecasting, and forecast evaluation. What is the role of forecast combination in that paradigm? In a world in which information sets can be instantaneously and costlessly combined, there is no role; it is always optimal to combine information sets rather than forecasts. That is, if no model forecast-encompasses the others, we might hope to eventually figure out what's gone wrong, learn from our mistakes, and come up with a model based on a combined information set that *does* forecast-encompass the others. But in the short run—particularly when deadlines must be met and

timely forecasts produced—pooling of information sets typically is either impossible or prohibitively costly. This simple insight motivates the pragmatic idea of forecast combination, in which forecasts rather than models are the basic object of analysis, due to an assumed inability to combine information sets. Thus, forecast combination can be viewed as a key link between the short-run, real-time forecast production process, and the longer-run, ongoing process of model development.

14. (Consensus forecasts) A number of services, some commercial and some nonprofit, regularly survey economic and financial forecasters and publish "consensus" forecasts, typically the mean or median of the forecasters surveyed. The consensus forecasts often perform very well relative to the individual forecasts. The Survey of Professional Forecasters is a leading consensus forecast that has been produced each quarter since the late 1960s; currently it's produced by the Federal Reserve Bank of Philadelphia. See Zarnowitz and Braun (1993) and Croushore (1993).

15. (The Delphi method for combining experts' forecasts) The "Delphi method" is a structured judgmental forecasting technique that sometimes proves useful in very difficult forecasting situations not amenable to quantification, such as new-technology forecasting. The basic idea is to survey a panel of experts anonymously, reveal the distribution of opinions to the experts so they can revise their opinions, repeat the survey, and so on. Typically the diversity of opinion is reduced as the iterations proceed.
    a. Delphi and related techniques are fraught with difficulties and pitfalls. Discuss them.
    b. At the same time, it's not at all clear that we should dispense with such techniques; they may be of real value. Why?

Bibliographical and Computational Notes

This chapter draws on Diebold and Lopez (1996) and Diebold (1989).

Mincer-Zarnowitz regressions are due to Mincer and Zarnowitz (1969).

The test for a zero expected loss differential, due to Diebold and Mariano (1995), builds on earlier work by Granger and Newbold (1986) and has been improved and extended by Harvey, Leybourne, and Newbold (1997); West (1996); White (2000); and Hansen (2001).

The idea of forecast encompassing dates at least to Nelson (1972) and was formalized and extended by Chong and Hendry (1986) and Fair and Shiller (1990).

The variance-covariance method of forecast combination is due to Bates and Granger (1969), and the regression interpretation is due to Granger and Ramanathan (1984).

Winkler and Makridakis (1983) document the frequent good performance of simple averages. In large part motivated by that finding, Clemen and Winkler (1986) and Diebold and Pauly (1990) develop forecast combination techniques that feature shrinkage toward the mean, and Stock and Watson (1998) arrive at a similar end via a very different route. See also Elliott and Timmermann (2002).

Concepts for Review

Evaluation and comparison of forecast accuracy
Unforecastability principle
Mean error
Bias
Error variance
Mean squared error
Root mean squared error
Mean absolute error
Forecast encompassing
Forecast combination
Variance-covariance forecast combination method
Regression-based forecast combination method

References and Additional Readings

Bates, J. M., and Granger, C. W. J. (1969). "The Combination of Forecasts." *Operations Research Quarterly*, 20, 451–468.

Box, G. E. P., Jenkins, G. W., and Reinsel, G. (1994). *Time Series Analysis, Forecasting and Control.* 3rd ed. Englewood Cliffs, N.J.: Prentice Hall.

Chong, Y. Y., and Hendry, D. F. (1986). "Econometric Evaluation of Linear Macroeconomic Models." *Review of Economic Studies*, 53, 671–690.

Clemen, R. T. (1989). "Combining Forecasts: A Review and Annotated Bibliography." *International Journal of Forecasting*, 5, 559–581.

Clemen, R. T., and Winkler, R. L. (1986). "Combining Economic Forecasts." *Journal of Business and Economic Statistics*, 4, 39–46.

Croushore, D. (1993). "The Survey of Professional Forecasters." *Business Review*, Federal Reserve Bank of Philadelphia, November–December.

Croushore, D., and Stark, T. (2001). "A Real-Time Dataset for Macroeconomists." *Journal of Econometrics*, 105, 111–130.

Diebold, F. X. (1989). "Forecast Combination and Encompassing: Reconciling Two Divergent Literatures." *International Journal of Forecasting*, 5, 589–592.

Diebold, F. X., and Kilian, L. (2001). "Measuring Predictability: Theory and Macroeconomic Applications." *Journal of Applied Econometrics*, 16, 657–669.

Diebold, F., X., and Lopez, J. (1996). "Forecast Evaluation and Combination." In G. S. Maddala and C. R. Rao (eds.). *Handbook of Statistics*. Amsterdam: North-Holland, 241–268.

Diebold, F. X., and Mariano, R. (1995). "Comparing Predictive Accuracy." *Journal of Business and Economic Statistics*, 13, 253–265. Reprinted in Diebold and Rudebusch (1999).

Diebold, F. X., and Pauly, P. (1990). "The Use of Prior Information in Forecast Combination." *International Journal of Forecasting*, 6, 503–508.

Diebold, F. X., and Rudebusch, G. D. (1989). "Scoring the Leading Indicators." *Journal of Business*, 62, 369–391. Reprinted in Diebold and Rudebusch (1999).

Diebold, F. X., and Rudebusch, G. D. (1991). "Forecasting Output with the Composite Leading Index: An Ex Ante Analysis." *Journal of the American Statistical Association*, 86, 603–610. Reprinted in Diebold and Rudebusch (1999).

Diebold, F. X., and Rudebusch, G. D. (1999). *Business Cycles: Durations, Dynamics, and Forecasting*. Princeton, N.J.: Princeton University Press.

Elliott, G., and Timmermann, A. (2002). "Optimal Forecast Combination under Regime Switching." Manuscript, Department of Economics, University of California, San Diego.

Fair, R. C., and Shiller, R. J. (1990). "Comparing Information in Forecasts from Econometric Models." *American Economic Review*, 80, 375–389.

Granger, C. W. J. and Newbold, P. (1986). *Forecasting Economic Time Series*. 2nd ed. San Diego: Academic Press.

Granger, C. W. J., and Ramanathan, R. (1984). "Improved Methods of Forecasting." *Journal of Forecasting*, 3, 197–204.

Hansen, P. R. (2001). "An Unbiased and Powerful Test for Superior Predictive Ability." Working Paper 01-06, Department of Economics, Brown University.

Harvey, D. I., Leybourne, S. J., and Newbold, P. (1997). "Testing the Equality of Prediction Mean Squared Errors." *International Journal of Forecasting*, 13, 281–291.

Klein, L. R. (1981). *Econometric Models as Guides for Decision Making*. New York: Free Press.

Klein, L. R., ed. (1991). *Comparative Performance of U.S. Econometric Models*. Oxford: Oxford University Press.

McNees, S. K. (1988). "How Accurate Are Macroeconomic Forecasts?" *New England Economic Review*, July/August, 15–36.

Mincer, J., and Zarnowitz, V. (1969). "The Evaluation of Economic Forecasts." In J. Mincer (ed.), *Economic Forecasts and Expectations*. New York: National Bureau of Economic Research.

Morgenstern, O. (1963). *On the Accuracy of Economic Observations*. Princeton, N.J.: Princeton University Press.

Nelson, C. R. (1972). "The Prediction Performance of the F.R.B.-M.I.T.-Penn Model of the U.S. Economy." *American Economic Review*, 62, 902–917.

Pesaran, M. H., and Samiei, H. (1995). "Forecasting Ultimate Resource Recovery." *International Journal of Forecasting*, 11, 543–555.

Stock, J. H., and Watson, M. W. (1998). "A Dynamic Factor Model Framework for Forecast Combination." Manuscript, Kennedy School, Harvard University, and Woodrow Wilson School, Princeton University.

Stock, J. H., and Watson, M. W. (1999). "A Comparison of Linear and Nonlinear Univariate Models for Forecasting Macroeconomic Time Series." In R. Engle and H. White (eds.), *Cointegration, Causality, and Forecasting: A Festschrift in Honor of Clive W. J. Granger*, 1–44. Oxford: Oxford University Press.

Theil, H. (1966). *Applied Economic Forecasting*. Amsterdam: North-Holland.

Wallis, K. F. and Whitley, J. D. (1991). "Sources of Error in Forecasts and Expectations: UK Economic Models, 1984–88." *Journal of Forecasting*, 10, 231–253.

West, K. D. (1996). "Asymptotic Inference about Predictive Ability." *Econometrica*, 64, 1067–1084.

White, H. (2000). "A Reality Check for Data Snooping." *Econometrica*, 68, 1097–1126.

Winkler, R. L., and Makridakis, S. (1983). "The Combination of Forecasts." *Journal of the Royal Statistical Society* A, 146, 150–157.

Zarnowitz, V., and Braun, P. (1993). "Twenty-Two Years of the N.B.E.R.-A.S.A. Quarterly Economic Outlook Surveys: Aspects and Comparisons of Forecasting Performance." In J. H. Stock and M. W. Watson (eds.), *Business Cycles, Indicators and Forecasting*. Chicago: University of Chicago Press for NBER.

Unit Roots, Stochastic Trends, ARIMA Forecasting Models, and Smoothing

Thus far we've handled nonstationarities, such as trend, using deterministic components. Now we consider an alternative, stochastic approach. Stochastic trend is important insofar as it sometimes provides a good description of certain business, economic, and financial time series, and it has a number of special properties and implications. As we'll see, for example, if we knew for sure that a series had a stochastic trend, then we'd want to difference the series and then fit a stationary model to the difference.[1] The strategy of differencing to achieve stationarity contrasts with the approach of earlier chapters, in which we worked in levels and included deterministic trends. In practice, it's sometimes very difficult to decide whether trend is best modeled as deterministic or stochastic, and the decision is an important part of the science—and art—of building forecasting models.

1. Stochastic Trends and Forecasting

Consider an ARMA(p, q) process,

$$\Phi(L)y_t = \Theta(L)\varepsilon_t ,$$

---

[1] We speak of modeling in "differences," as opposed to "levels." We also use *differences* and *changes* interchangeably.

with all the autoregressive roots on or outside the unit circle, at most one autoregressive root on the unit circle, and all moving average roots outside the unit circle. We say that y has a **unit autoregressive root,** or simply a **unit root,** if one of the p roots of its autoregressive lag operator polynomial is 1, in which case we can factor the autoregressive lag operator polynomial as

$$\Phi(L) = \Phi'(L)(1 - L),$$

where $\Phi'(L)$ is of degree $p - 1$. Thus, y is really an ARMA($p - 1$, q) process in differences, because

$$\Phi'(L)(1 - L)y_t = \Theta(L)\varepsilon_t$$

is simply

$$\Phi'(L)\Delta y_t = \Theta(L)\varepsilon_t .$$

Note that y is not covariance stationary, because one of the roots of its autoregressive lag operator polynomial is on the unit circle, whereas covariance stationarity requires all roots to be outside the unit circle. However, Δy is a covariance stationary and invertible ARMA($p - 1$, q) process.

You may recall from calculus that we can "undo" an integral by taking a derivative. By analogy, we say that a nonstationary series is integrated if its nonstationarity is appropriately "undone" by differencing. If only one difference is required (as with the series y earlier), we say that the series is integrated of order 1, or $I(1)$ (pronounced "eye-one") for short. More generally, if d differences are required, the series is $I(d)$. The order of integration equals the number of autoregressive unit roots. In practice, $I(0)$ and $I(1)$ processes are by far the most important cases, which is why we restricted our discussion to allow for at most one unit root.[2] To get a feel for the behavior of $I(1)$ processes, let's take a simple and very important example, the **random walk,** which is nothing more than an AR(1) process with a unit coefficient,

$$y_t = y_{t-1} + \varepsilon_t$$
$$\varepsilon_t \sim WN(0, \sigma^2) .$$

The random walk is not covariance stationary, because the AR(1) coefficient is not less than 1. In particular, it doesn't display **mean reversion;** in contrast to a stationary AR(1), it wanders up and down randomly, as its name suggests, with no tendency to return to any particular point. Although the random walk is somewhat ill behaved, its first difference is the ultimate well-behaved series: zero-mean white noise.

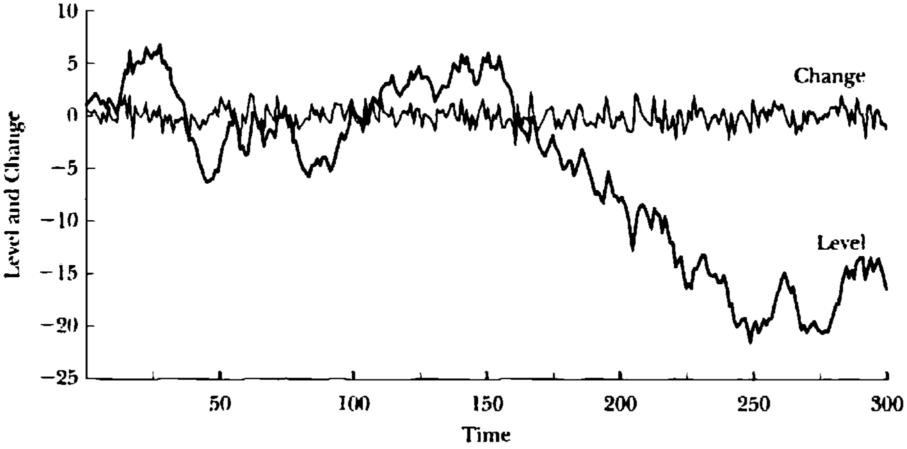
As an illustration, we show a random walk realization of length 300, as well as its first difference, in Figure 13.1.[3] The difference of the random walk is white noise, which vibrates randomly. In contrast, the level of the random

---

[2] $I(2)$ series sometimes, but rarely, arise, and orders of integration greater than 2 are almost unheard of.

[3] The random walk was simulated on a computer with $y_1 = 1$ and $N(0, 1)$ innovations.

FIGURE 13.1
Random Walk Level and Change



walk, which is the cumulative sum of the white noise changes, wanders aimlessly and persistently.

Now let's consider a **random walk with drift,**

$$y_t = \delta + y_{t-1} + \varepsilon_t$$

$$\varepsilon_t \sim WN(0, \sigma^2) .$$

Note that the random walk with drift is effectively a model of trend, because on average it grows each period by the drift, δ. Thus, the drift parameter plays the same role as the slope parameter in our earlier model of linear deterministic trend. We call the random walk with drift (and of course also the **random walk without drift**) a model of **stochastic trend,** because the trend is driven by stochastic shocks, in contrast to the **deterministic trends** considered in Chapter 5.

Just as the random walk has no particular level to which it returns, so too the random walk with drift has no particular trend to which it returns. If a shock lowers the value of a random walk, for example, there is no tendency for it to necessarily rise again—we expect it to stay permanently lower. Similarly, if a shock moves the value of a random walk with drift below the currently projected trend, there's no tendency for it to return—the trend simply begins anew from the series' new location. Thus, shocks to random walks have completely permanent effects; a unit shock forever moves the expected future path of the series by one unit, regardless of the presence of drift.

For illustration, we show in Figure 13.2 a realization of a random walk with drift, in levels and differences. As before, the sample size is 300 and $y_1 = 1$. The innovations are $N(0, 1)$ white noise, and the drift is $\delta = 0.3$ per period, so the differences are white noise with a mean of 0.3. It's hard to notice the nonzero mean in the difference, because the stochastic trend in the level, which is the cumulative sum of $N(0.3, 1)$ white noise, dominates the scale.

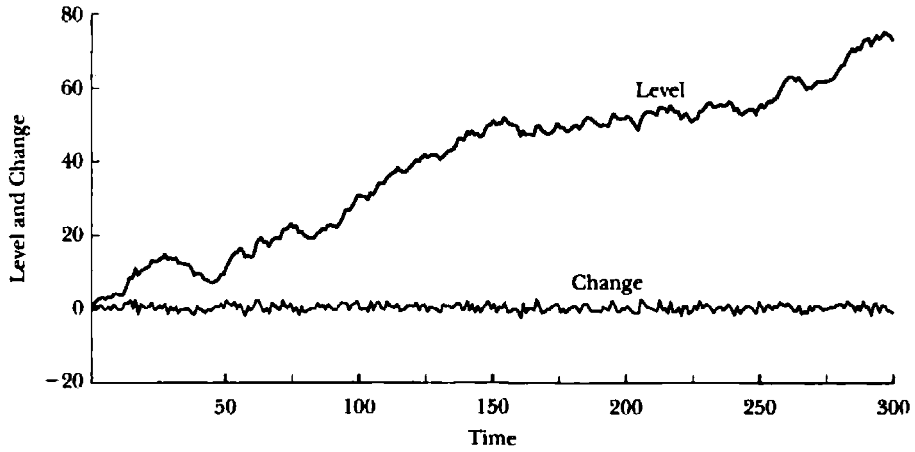


FIGURE 13.2
Random Walk with Drift Level and Change

Let's study the properties of random walks in greater detail. The random walk is

$$y_t = y_{t-1} + \varepsilon_t$$
$$\varepsilon_t \sim WN(0, \sigma^2)\,.$$

Assuming the process started at some time 0 with value y_0, we can write it as

$$y_t = y_0 + \sum_{i=1}^{t} \varepsilon_i\,.$$

Immediately,

$$E(y_t) = y_0$$

and

$$\operatorname{var}(y_t) = t\sigma^2\,.$$

In particular, note that

$$\lim_{t\to\infty} \operatorname{var}(y_t) = \infty\,,$$

so that the variance grows continuously rather than converging to some finite unconditional variance.

Now consider the random walk with drift. The process is

$$y_t = \delta + y_{t-1} + \varepsilon_t$$
$$\varepsilon_t \sim WN(0, \sigma^2)\,.$$

Assuming the process started at some time 0 with value y_0, we have

$$y_t = t\delta + y_0 + \sum_{i=1}^{t} \varepsilon_i\,.$$

Immediately

$$E(y_t) = y_0 + t\delta$$

and

$$\text{var}(y_t) = t\sigma^2 .$$

As with the simple random walk, then, the random walk with drift also has the property that

$$\lim_{t \to \infty} \text{var}(y_t) = \infty .$$

Just as white noise is the simplest $I(0)$ process, the random walk is the simplest $I(1)$ process. And just as $I(0)$ processes with richer dynamics than white noise can be constructed by transforming white noise, so too can $I(1)$ processes with richer dynamics than the random walk be obtained by transforming the random walk. We're led immediately to the **ARIMA(p, 1, q) model,**

$$\Phi(L)\,(1 - L)y_t = c + \Theta(L)\varepsilon_t$$

or

$$(1 - L)y_t = c\Phi^{-1}(1) + \Phi^{-1}(L)\Theta(L)\varepsilon_t ,$$

where

$$\Phi(L) = 1 - \Phi_1 L - \cdots - \Phi_p L^p$$

$$\Theta(L) = 1 - \Theta_1 L - \cdots - \Theta_q L^q ,$$

and all the roots of both lag operator polynomials are outside the unit circle. ARIMA stands for autoregressive *integrated* moving average. The ARIMA(p, 1, q) process is just a stationary and invertible ARMA(p, q) process in first differences.

More generally, we can work with the **ARIMA(p, d, q) model,**

$$\Phi(L)\,(1 - L)^d y_t = c + \Theta(L)\varepsilon_t$$

or

$$(1 - L)^d y_t = c\Phi^{-1}(1) + \Phi^{-1}(L)\Theta(L)\varepsilon_t ,$$

where

$$\Phi(L) = 1 - \Phi_1 L - \cdots - \Phi_p L^p$$

$$\Theta(L) = 1 - \Theta_1 L - \cdots - \Theta_q L^q ,$$

and all the roots of both lag operator polynomials are outside the unit circle. The ARIMA(p, d, q) process is a stationary and invertible ARMA(p, q) after differencing d times. In practice, $d = 0$ and $d = 1$ are by far the most important cases. When $d = 0$, y is covariance stationary, or $I(0)$, with mean $c\Phi^{-1}(1)$. When $d = 1$, y is $I(1)$ with drift, or stochastic linear trend, of $c\Phi^{-1}(1)$ per period.

It turns out that more complicated ARIMA(p, 1, q) processes behave like random walks in certain key respects. First, ARIMA(p, 1, q) processes are appropriately made stationary by differencing. Second, shocks to ARIMA(p, 1, q) processes have permanent effects.[4] Third, the variance of an ARIMA(p, 1, q) process grows without bound as time progresses. The special properties of $I(1)$ series, associated with the fact that innovations have permanent effects, have important implications for forecasting. In regard to point forecasting, the **shock persistence** means that optimal forecasts, even at very long horizons, don't completely revert to a mean or a trend. And in regard to interval and density forecasting, the fact that the variance of an $I(1)$ process approaches infinity as time progresses means that the uncertainty associated with our forecasts, which translates into the width of interval forecasts and the spread of density forecasts, increases without bound as the forecast horizon grows.[5]

Let's see how all this works in the context of a simple random walk, which is an AR(1) process with a unit coefficient. Recall that for the AR(1) process,

$$y_t = \varphi y_{t-1} + \varepsilon_t$$

$$\varepsilon_t \sim WN(0, \sigma^2),$$

the optimal forecast is

$$y_{T+h,T} = \varphi^h y_T .$$

Thus, in the random walk case of $\varphi = 1$, the optimal forecast is simply the current value, regardless of horizon. This makes clear the way that the permanence of shocks to random walk processes affects forecasts: Any shock that moves the series up or down today also moves the optimal forecast up or down, at all horizons. In particular, the effects of shocks don't wash out as the forecast horizon lengthens, because the series does not revert to a mean.

In Figure 13.3, we illustrate the important differences in forecasts from deterministic trend and stochastic trend models for U.S. GNP per capita.[6] We show GNP per capita 1869–1933, followed by the forecasts from the best-fitting deterministic and stochastic trend models, 1934–1993, made in 1933. The best-fitting deterministic trend model is an AR(2) in levels with linear trend, and the best-fitting stochastic trend model is an AR(1) in differences (i.e., an ARIMA(1, 1, 0)) with a drift.[7] Because 1932 and 1933 were years of severe recession, the forecasts are made from a position well below trend. The forecast from the deterministic trend model reverts to trend quickly, in sharp contrast to that from the stochastic trend model, which remains permanently lower. As

---

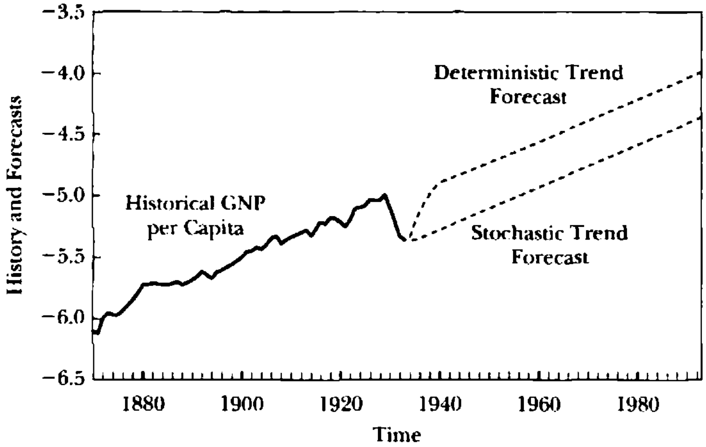
[4] In contrast to random walks, however, the long-run effect of a unit shock to an ARIMA(p, 1, q) process may be greater or less than unity, depending on the parameters of the process.

[5] This is true even if we ignore parameter estimation uncertainty.

[6] The GNP per capita data are in logarithms. See Diebold and Senhadji (1996) for details.

[7] Note well that the two dashed lines are two different point extrapolation forecasts, not an interval forecast.

FIGURE 13.3
U.S. per Capita GNP History and Two Forecasts



it happens, the forecast from the deterministic trend model turns out to be distinctly better in this case, as shown in Figure 13.4, which includes the realization.

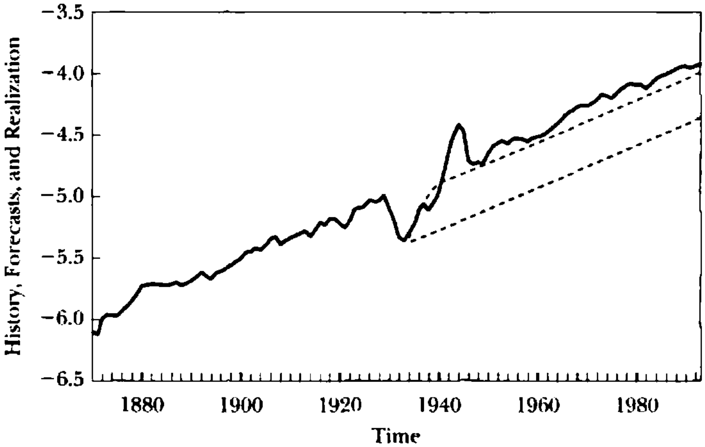
Now let's consider interval and density forecasts from $I(1)$ models. Again, it's instructive to consider a simple random walk. Recall that the error associated with the optimal forecast of an AR(1) process is

$$e_{T+h,T} = (y_{T+h} - y_{T+h,T}) = \varepsilon_{T+h} + \varphi\varepsilon_{T+h-1} + \cdots + \varphi^{h-1}\varepsilon_{T+1},$$

with variance

$$\sigma_h^2 = \sigma^2 \sum_{i=0}^{h-1} \varphi^{2i}.$$

FIGURE 13.4
U.S. per Capita GNP History, Two Forecasts, and Realization



Thus, in the random walk case, the error is the sum of h white noise innovations,

$$e_{T+h,T} = \sum_{i=0}^{h-1} \varepsilon_{T+h-i} \,,$$

with variance $h\sigma^2$. The forecast error variance is proportional to h and therefore grows without bound as h grows. An h-step-ahead 95% interval forecast for any future horizon is then $y_T \pm 1.96\sigma\sqrt{h}$, and an h-step-ahead density forecast is $N(y_T, h\sigma^2)$.

Thus far, we've explicitly illustrated the construction of point, interval, and density forecasts for a simple random walk. Forecasts from more complicated $I(1)$ models are constructed similarly. Point forecasts of levels of ARIMA(p, 1, q) processes, for example, are obtained by recognizing that ARIMA processes are ARMA processes in differences, and we know how to forecast ARMA processes. Thus, we forecast the changes, cumulate the forecasts of changes, and add them to the current level, yielding

$$y_{T+h,T} = y_T + (\Delta y)_{T+1,T} + \cdots + (\Delta y)_{T+h,T} \,.$$

2. Unit Roots: Estimation and Testing

LEAST-SQUARES REGRESSION WITH UNIT ROOTS

The properties of least-squares estimators in models with unit roots are of interest to us, because they have implications for forecasting. We'll use a random walk for illustration, but the results carry over to general ARIMA(p, 1, q) processes. Suppose that y is a random walk, so that

$$y_t = y_{t-1} + \varepsilon_t \,,$$

but we don't know that the autoregressive coefficient is 1, so we estimate the AR(1) model,

$$y_t = \varphi y_{t-1} + \varepsilon_t \,.$$

Two key and offsetting properties of the least squares estimator emerge: superconsistency and bias.

First we consider **superconsistency.** In the unit root case of $\varphi = 1$, the difference between $\hat{\varphi}_{LS}$ and 1 vanishes quickly as the sample size (T) grows; in fact, it shrinks like $\frac{1}{T}$. Thus, $T(\hat{\varphi}_{LS} - 1)$ converges to a nondegenerate random variable. In contrast, in the covariance stationary case of $|\varphi| < 1$, the difference between $\hat{\varphi}_{LS}$ and φ shrinks more slowly, like $\frac{1}{\sqrt{T}}$, so that $\sqrt{T}(\hat{\varphi}_{LS} - \varphi)$ converges to a nondegenerate random variable. We call the extrafast convergence in the unit root case *superconsistency*; we say that the least-squares estimator of a unit root is superconsistent.

Now we consider bias. It can be shown that the least-squares estimator, $\hat{\varphi}_{LS}$, is biased downward, so that if the true value of φ is φ^*, the expected value of $\hat{\varphi}_{LS}$ is less than φ^*.[8] Other things the same, the larger is the true value of φ, the larger the bias, so the bias is worst in the unit root case. The bias is also larger if an intercept is included in the regression and larger still if a trend is included. The bias vanishes as the sample size grows, as the estimate converges to the true population value, but the bias can be sizeable in samples of the size that concern us.

Superconsistency and bias have offsetting effects as regards forecasting. Superconsistency is helpful; it means that the sampling uncertainty in our parameter estimates vanishes unusually quickly as sample size grows. Bias, in contrast, is harmful, because badly biased parameter estimates can translate into poor forecasts. The superconsistency associated with unit roots guarantees that bias vanishes quickly as sample size grows, but it may nevertheless be highly relevant in small samples.

EFFECTS OF UNIT ROOTS ON THE SAMPLE AUTOCORRELATION AND PARTIAL AUTOCORRELATION FUNCTIONS

If a series has a unit root, its autocorrelation function isn't well defined in population, because its variance is infinite. But the *sample* autocorrelation function can of course be mechanically computed in the usual way, because the computer software doesn't know or care whether the data being fed into it have a unit root. The sample autocorrelation function will tend to damp extremely slowly; loosely speaking, we say that it fails to damp. The reason is that, because a random walk fails to revert to any population mean, any given sample path will tend to wander above and below its sample mean for long periods of time, leading to very large positive sample autocorrelations, even at long displacements. The sample partial autocorrelation function of a unit root process, in contrast, will damp quickly: It will tend to be very large and close to one at displacement 1 but will tend to be smaller and decay quickly thereafter.

If the properties of the sample autocorrelations and partial autocorrelations of unit root processes appear rather exotic, the properties of the sample autocorrelations and partial autocorrelations of *differences* of unit root processes are much more familiar. That's because the first difference of an $I(1)$ process, by definition, is covariance stationary and invertible.

We illustrate the properties of sample autocorrelations and partial autocorrelations of levels and differences of unit root processes in Figures 13.5 and 13.6. Figure 13.5 shows the correlogram of our simulated random walk.

---

[8] The bias in the least-squares estimator in the unit root and near–unit root cases was studied by Dickey (1976) and Fuller (1976), and it is sometimes called the *Dickey-Fuller bias*.

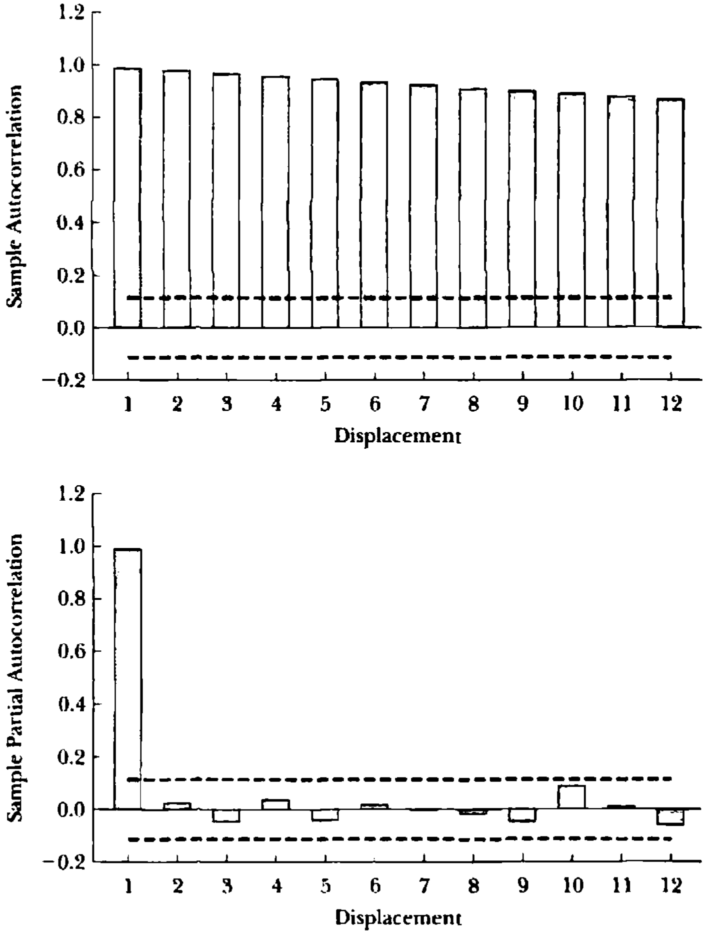


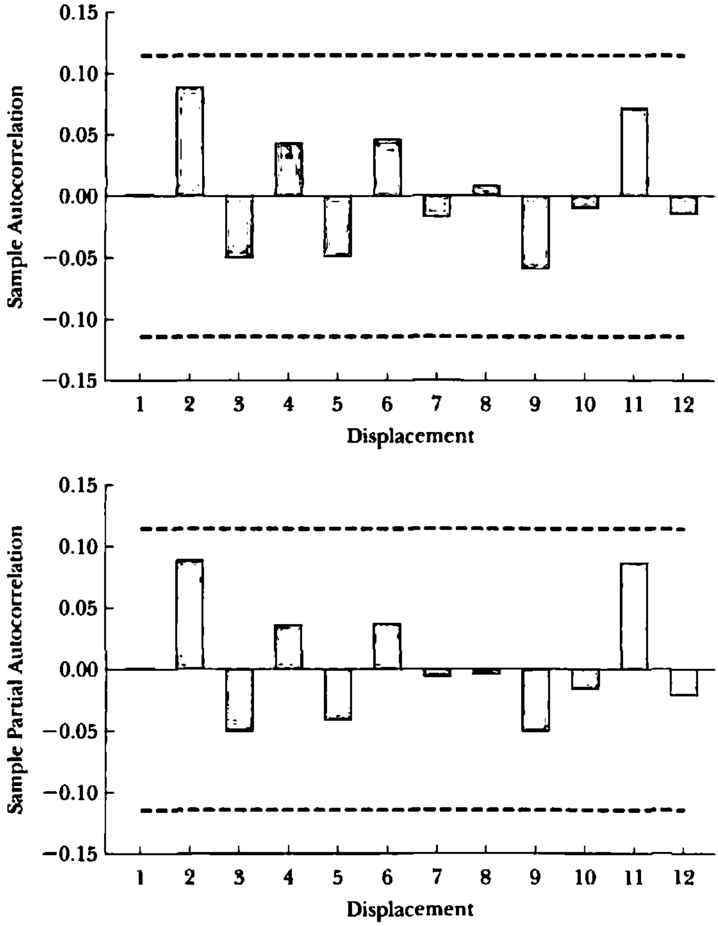
FIGURE 13.5 *Random Walk, Levels: Sample Autocorrelation Function (Top Panel) and Sample Partial Autocorrelation Function (Bottom Panel)*

The sample autocorrelations fail to damp, and the sample partial autocorrelation is huge at displacement 1, but tiny thereafter. Figure 13.6 shows the correlogram of the first difference of the random walk. All the sample autocorrelations and partial autocorrelations are insignificantly different from 0, as expected, because the first difference of a random walk is white noise.

UNIT ROOT TESTS

In light of the special properties of series with unit roots, it's sometimes of interest to test for their presence, with an eye toward the desirability of imposing

FIGURE 13.6
Random Walk, First Differences: Sample Autocorrelation Function (Top Panel) and Sample Partial Autocorrelation Function (Bottom Panel)



them, by differencing the data, if they seem to be present. Let's start with the simple AR(1) process,

$$y_t = \varphi y_{t-1} + \varepsilon_t$$

$$\varepsilon_t \overset{\text{iid}}{\sim} N(0, \sigma^2) .$$

We can regress y_t on y_{t-1}, and then use the standard t-test for testing $\varphi = 1$,

$$\hat{\tau} = \frac{\hat{\varphi} - 1}{s\sqrt{\dfrac{1}{\sum_{t=2}^{T} y_{t-1}^2}}} ,$$

where s is the standard error of the regression. Note that the $\hat{\tau}$-statistic is *not* the t-statistic computed automatically by regression packages; the standard

t-statistic is for the null of a 0 coefficient, whereas $\hat{\tau}$ is the t-statistic for a *unit* coefficient. A simple trick, however, coaxes standard software into printing $\hat{\tau}$ automatically. Simply rewrite the first-order autoregression as

$$y_t - y_{t-1} = (\varphi - 1)y_{t-1} + \varepsilon_t .$$

Thus, $\hat{\tau}$ is the usual t-statistic in a regression of the first *difference* of y on the first *lag* of y.

A key result is that, in the unit root case, $\hat{\tau}$ does *not* have the t distribution. Instead it has a special distribution now called the **Dickey-Fuller distribution,** named for two statisticians who studied it extensively in the 1970s and 1980s. Fuller (1976) presents tables of the percentage points of the distribution of $\hat{\tau}$, which we'll call the *Dickey-Fuller statistic,* under the null hypothesis of a unit root. Because we're only allowing for roots on or outside the unit circle, a one-sided test is appropriate.

Thus far, we've shown how to test the null hypothesis of a random walk with no drift against the alternative of a zero-mean, covariance-stationary AR(1). Now we allow for a **nonzero mean, μ, under the alternative hypothesis,** which is of potential importance because business and economic data can rarely be assumed to have zero mean. Under the alternative hypothesis, the process becomes a covariance stationary AR(1) process *in deviations from the mean,*

$$(y_t - \mu) = \varphi(y_{t-1} - \mu) + \varepsilon_t ,$$

which we can rewrite as

$$y_t = \alpha + \varphi y_{t-1} + \varepsilon_t ,$$

where $\alpha = \mu(1 - \varphi)$. If we knew μ, we could simply center the data and proceed as before. In practice, of course, μ must be estimated along with the other parameters. Although α vanishes under the unit root null hypothesis of $\varphi = 1$, it is nevertheless present under the alternative hypothesis, and so we include an intercept in the regression. The distribution of the corresponding Dickey-Fuller statistic, $\hat{\tau}_\mu$, has been tabulated under the null hypothesis of $(\alpha, \varphi) = (0, 1)$; tables appear in Fuller (1976).

Finally, let's allow for **deterministic linear trend under the alternative hypothesis,** by writing the AR(1) in deviations from a linear trend,

$$(y_t - a - b\,\mathrm{TIME}_t) = \varphi(y_{t-1} - a - b\,\mathrm{TIME}_{t-1}) + \varepsilon_t ,$$

or

$$y_t = \alpha + \beta\mathrm{TIME}_t + \varphi y_{t-1} + \varepsilon_t ,$$

where $\alpha = a(1 - \varphi) + b\varphi$ and $\beta = b(1 - \varphi)$. Under the unit root hypothesis that $\varphi = 1$, we have a random walk with drift,

$$y_t = b + y_{t-1} + \varepsilon_t ,$$

which is a stochastic trend, but under the deterministic trend alternative hypothesis, both the intercept and the trend enter, so they must be included in the regression. The random walk with drift is a null hypothesis that frequently arises in economic applications; stationary deviations from linear trend are a natural alternative. The distribution of the Dickey-Fuller statistic $\hat{\tau}_\tau$, which allows for linear trend under the alternative hypothesis, has been tabulated under the unit root null hypothesis by Fuller (1976).

Now we generalize the test to allow for higher-order autoregressive dynamics. Consider the AR(p) process

$$y_t + \sum_{j=1}^{p} \varphi_j y_{t-j} = \varepsilon_t ,$$

which we rewrite as

$$y_t = \rho_1 y_{t-1} + \sum_{j=2}^{p} \rho_j (y_{t-j+1} - y_{t-j}) + \varepsilon_t ,$$

where $p \geq 2$, $\rho_1 = -\sum_{j=1}^{p} \varphi_j$, and $\rho_i = \sum_{j=i}^{p} \varphi_j$, $i = 2, \ldots, p$. If there is a unit root, then $\rho_1 = 1$, and y is simply an AR($p - 1$) in first differences. The Dickey-Fuller statistic for the null hypothesis of $\rho_1 = 1$ has the same asymptotic distribution as $\hat{\tau}$. Thus, the results for the AR(1) process generalize (asymptotically) in a straightforward manner to higher-order processes.

To allow for a nonzero mean in the AR(p) case, write

$$(y_t - \mu) + \sum_{j=1}^{p} \varphi_j (y_{t-j} - \mu) = \varepsilon_t ,$$

or

$$y_t = \alpha + \rho_1 y_{t-1} + \sum_{j=2}^{p} \rho_j (y_{t-j+1} - y_{t-j}) + \varepsilon_t ,$$

where $\alpha = \mu (1 + \sum_{j=1}^{p} \varphi_j)$, and the other parameters are as noted earlier. Under the null hypothesis of a unit root, the intercept vanishes, because in that case $\sum_{j=1}^{p} \varphi_j = -1$. The distribution of the Dickey-Fuller statistic for testing $\rho_1 = 1$ in this regression is asymptotically identical to that of $\hat{\tau}_\mu$.

Finally, to allow for linear trend under the alternative hypothesis, write

$$(y_t - a - b \cdot \mathrm{TIME}_t) + \sum_{j=1}^{p} \varphi_j (y_{t-j} - a - b \cdot \mathrm{TIME}_{t-j}) = \varepsilon_t ,$$

which we rewrite as

$$y_t = k_1 + k_2 \,\mathrm{TIME}_t + \rho_1 y_{t-1} + \sum_{j=2}^{p} \rho_j (y_{t-j+1} - y_{t-j}) + \varepsilon_t ,$$

where

$$k_1 = a\left(1 + \sum_{i=1}^{p} \varphi_i\right) - b \sum_{i=1}^{p} i\varphi_i$$

and

$$k_2 = b\left(1 + \sum_{i=1}^{p} \varphi_i\right) .$$

Under the null hypothesis, $k_1 = -b \sum_{i=1}^{p} i\varphi_i$ and $k_2 = 0$. The Dickey-Fuller statistic for the hypothesis that $\rho_1 = 1$ has the $\hat{\tau}_\tau$ distribution asymptotically.

Now we consider general ARMA representations. We've seen that the original Dickey-Fuller test for a unit root in AR(1) models is easily generalized to test for a unit root in the AR(p) case, $p < \infty$; we simply augment the test regression with lagged first differences, which is called an **augmented Dickey-Fuller test** or augmented Dickey-Fuller regression. Matters are more complex in the ARMA(p, q) case, however, because the corresponding autoregression is of infinite order. A number of tests have been suggested, and the most popular is to approximate the infinite autoregression with a finite-order augmented Dickey-Fuller regression. We let the number of augmentation lags increase with the sample size, but at a slower rate. Hall (1994) shows that, under certain conditions, the asymptotic null distribution of the Dickey-Fuller statistic with augmentation lag order selected by SIC is the same as if the true order were known, so that the SIC provides a useful guide to augmentation lag order selection in Dickey-Fuller regressions. Ng and Perron (1995), however, argue that standard t-testing provides more reliable inference. Additional research is needed, but it does appear that, unlike when selecting lag orders for forecasting models, it may be better to use less harsh degrees-of-freedom penalties, such as those associated with t-testing or the AIC, when selecting augmentation lag orders in Dickey-Fuller regressions.

Depending on whether a zero mean, a nonzero mean, or a linear trend is allowed under the alternative hypothesis, we write either

$$y_t = \rho_1 y_{t-1} + \sum_{j=2}^{k-1} \rho_j (y_{t-j+1} - y_{t-j}) + \varepsilon_t$$

$$y_t = \alpha + \rho_1 y_{t-1} + \sum_{j=2}^{k-1} \rho_j (y_{t-j+1} - y_{t-j}) + \varepsilon_t$$

or

$$y_t = k_1 + k_2 \text{TIME}_t + \rho_1 y_{t-1} + \sum_{j=2}^{k-1} \rho_j (y_{t-j+1} - y_{t-j}) + \varepsilon_t ,$$

where $k - 1$ augmentation lags have been included. The Dickey-Fuller statistics on y_{t-1} continue to have the $\hat{\tau}$, $\hat{\tau}_\mu$, and $\hat{\tau}_\tau$ asymptotic distributions under the null hypothesis of a unit root. For selecting the number of augmentation lags, $k - 1$, we can use the SIC or AIC, as well as the t-statistics on the various lags of Δy, which have the standard normal distribution in large samples, regardless of whether the unit root hypothesis is true or false.

New tests, with better power than the Dickey-Fuller tests in certain situations, have been proposed recently.[9] But power and size problems will always plague unit root tests—power problems, because the relevant alternative hypotheses are typically very close to the null hypothesis; and size problems, because we should include infinitely many augmentation lags in principle but we can't in practice.

---

[9] See Elliott, Rothenberg, and Stock (1996); Dickey and Gonzalez-Farias (1992); and the comparisons in Pantula, Gonzalez-Farias, and Fuller (1994).

Thus, although unit root tests are sometimes useful, don't be fooled into thinking they're the end of the story regarding the decision of whether to specify models in levels or differences. For example, the fact that we can't reject a unit root doesn't necessarily mean that we should impose it—the power of unit root tests against alternative hypotheses near the null hypothesis, which are the relevant alternatives, is likely to be low. At the same time, it may sometimes be desirable to impose a unit root even when the true root is less than 1, if the true root is nevertheless very close to 1, because the Dickey-Fuller bias plagues estimation in levels. We need to use introspection and theory, in addition to formal tests, to guide the difficult decision of whether to impose unit roots, and we need to compare the forecasting performance of different models with and without unit roots imposed.

In certain respects, the most important part of unit root theory for forecasting concerns estimation, not testing. It's important for forecasters to understand the effects of unit roots on consistency and small-sample bias. Such understanding on the one hand leads to the insight that at least *asymptotically* we're probably better off estimating forecasting models in levels with trends included, because then we'll get an accurate approximation to the dynamics in the data regardless of the true state of the world, unit root or no unit root. If there's no unit root, then of course it's desirable to work in levels; if there is a unit root, then the estimated largest root will converge appropriately to unity, and at a fast rate. On the other hand, differencing is appropriate only in the unit root case, and inappropriate differencing can be harmful, even asymptotically.

3. Application: Modeling and Forecasting the Yen/Dollar Exchange Rate

Let's apply and illustrate what we've learned by modeling and forecasting the Japanese yen/U.S. dollar (JPY/USD) exchange rate. For convenience, we call the yen/dollar series y, the log level $\ln y$, and the change in the log level $\Delta \ln y$. We have end-of-month data from 1973.01 through 1996.07; we plot $\ln y$ in the top panel of Figure 13.7 and $\Delta \ln y$ in the bottom panel.[10] Note that $\ln y$ looks very highly persistent; perhaps it has a unit root. Conversely, $\Delta \ln y$ looks

---

[10] Throughout, we work with the log of the exchange rate, because the change in the log has the convenient interpretation of approximate percentage change. Thus, when we refer to the level of the exchange rate, we mean the log of the level ($\ln y$), and when we refer to the change, we mean the change of the log exchange rate ($\Delta \ln y$).

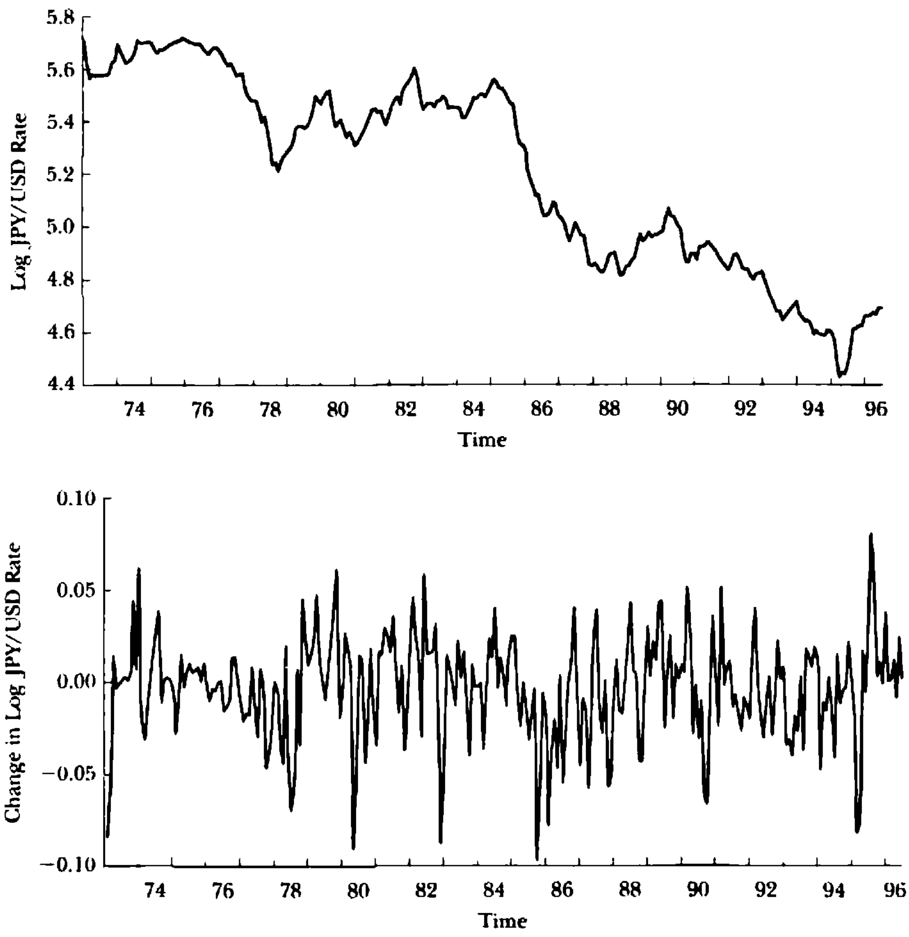
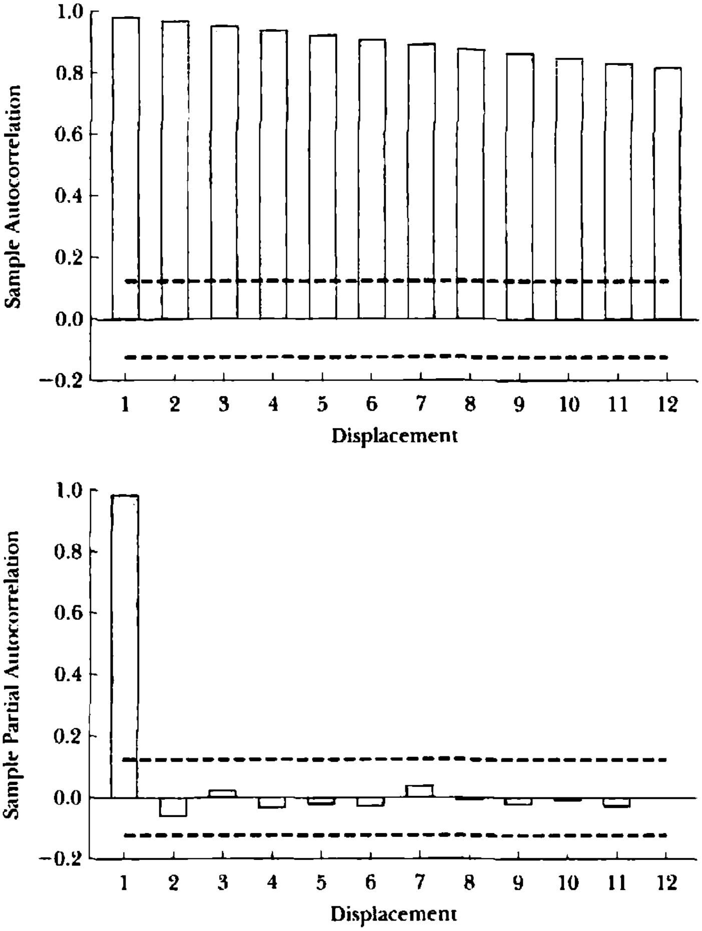


FIGURE 13.7 *Log JPY/USD Exchange Rate (Top Panel) and Change in Log JPY/USD Exchange Rate (Bottom Panel)*

thoroughly stationary and in fact rather close to white noise. Figure 13.8, which shows the correlogram for $\Delta \ln y$, and Figure 13.9, which shows the correlogram for the $\Delta \ln y$, confirm the impression we gleaned from the plots. The sample autocorrelations of $\ln y$ are all very large and fail to damp, and the first sample partial autocorrelation is huge while all the others are insignificantly different from 0. The correlogram of $\Delta \ln y$, however, looks very different. Both the sample autocorrelation and partial autocorrelation functions damp quickly; in fact, beyond displacement 1, they're all insignificantly different from 0. All of this suggests that $\ln y$ is $I(1)$.

Now we fit forecasting models. We base all analysis and modeling on $\ln y$, 1973.01–1994.12, and we reserve 1995.01–1996.07 for out-of-sample forecasting. We begin by fitting deterministic trend models to $\ln y$; we regress $\ln y$ on an

FIGURE 13.8 *Log JPY/USD Exchange Rate Sample Autocorrelations (Top Panel) and Sample Partial Autocorrelations (Bottom Panel)*



intercept and a time trend, allowing for up to ARMA(3, 3) dynamics in the disturbances. In Tables 13.1 and 13.2 we show the AIC and SIC values for all the ARMA(p, q) combinations. The AIC selects an ARMA(3, 1) model, while the SIC selects an AR(2). We proceed with the more parsimonious model selected by the SIC. The estimation results appear in Table 13.3 and the residual plot in Figure 13.10; note in particular that the dominant inverse root is very close to 1 (0.96), while the second inverse root is positive but much smaller (0.35).

Out-of-sample forecasts appear in Figures 13.11–13.13. Figure 13.11 shows the history, 1990.01–1994.12, and point and interval forecasts, 1995.01–1996.07. Although the estimated highly persistent dynamics imply very slow reversion to trend, it happens that the end-of-sample values of ln y in 1994 are very close to the estimated trend. Thus, to a good approximation, the forecast simply extrapolates the fitted trend. In Figure 13.12, we show the

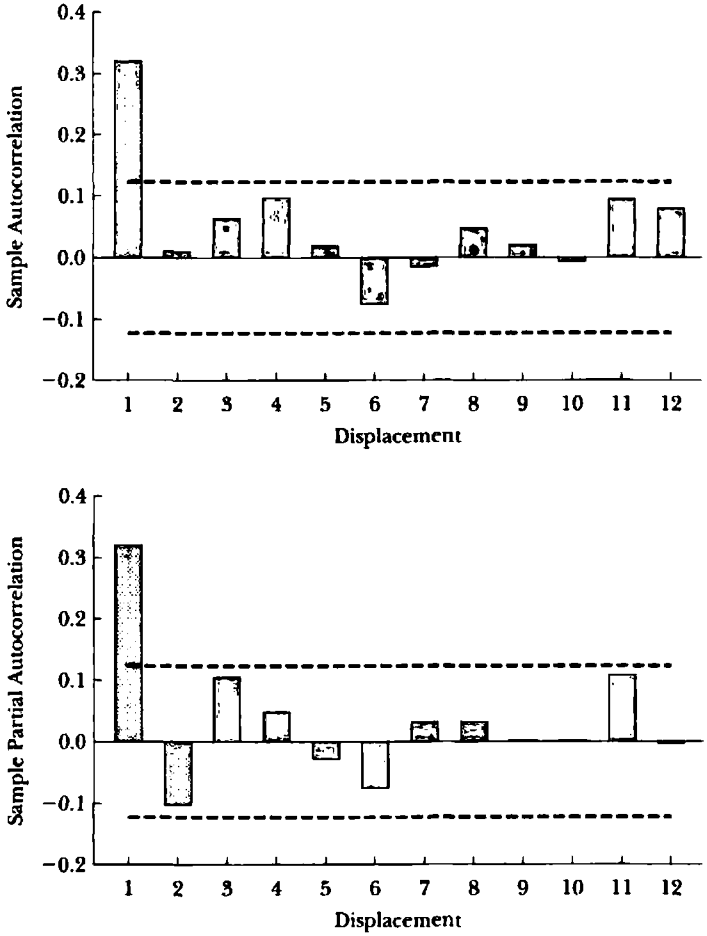


FIGURE 13.9
Log JPY/USD Exchange Rate, First Differences: Sample Autocorrelations (Top Panel) and Sample Partial Autocorrelations (Bottom Panel)

TABLE 13.1
Log JPY/USD Rate, Levels: AIC Values, Various ARMA Models

		MA Order			
		0	1	2	3
AR Order	0		−5.171	−5.953	−6.428
	1	−7.171	−7.300	−7.293	−7.287
	2	−7.319	−7.314	−7.320	−7.317
	3	−7.322	−7.323	−7.316	−7.308

TABLE 13.2

Log JPY/USD Rate, Levels: SIC Values, Various ARMA Models

		MA Order			
		0	1	2	3
	0		−5.130	−5.899	−6.360
AR Order	1	−7.131	−7.211	−7.225	−7.205
	2	−7.265	−7.246	−7.238	−7.221
	3	−7.253	−7.241	−7.220	−7.199

history together with a very long-horizon forecast (through 2020.12), in order to illustrate the fact that the confidence intervals eventually flatten at plus or minus two standard errors. Finally, Figure 13.13 displays the history and forecast together with the realization. Most of the realization is inside the 95% confidence intervals.

In light of the suggestive nature of the correlograms, we now perform a formal unit root test, with trend allowed under the alternative hypothesis. Table 13.4 presents the results with three augmentation lags.[11] There's no evidence whatsoever against the unit root; thus, we consider modeling $\Delta \ln y$. We regress $\Delta \ln y$ on an intercept and allow for up to ARMA(3, 3) dynamics in the disturbance.

TABLE 13.3

Log Exchange Rate, Best-Fitting Deterministic Trend Model

LS // Dependent variable is LYEN.
Sample (adjusted): 1973:03 1994:12
Included observations: 262 after adjusting endpoints
Convergence achieved after 3 iterations

Variable	**Coefficient**	**Std. Error**	***t*-Statistic**	**Prob.**
C	5.904705	0.136665	43.20570	0.0000
TIME	−0.004732	0.000781	−6.057722	0.0000
AR(1)	1.305829	0.057587	22.67561	0.0000
AR(2)	−0.334210	0.057656	−5.796676	0.0000

R^2	0.994468	Mean dependent var.	5.253984
Adjusted R^2	0.994404	SD dependent var.	0.341563
SE of regression	0.025551	Akaike info criterion	−7.319015
Sum squared resid.	0.168435	Schwarz criterion	−7.264536
Log likelihood	591.0291	F-statistic	15461.07
Durbin-Watson stat.	1.964687	Prob(F-statistic)	0.000000
Inverted AR roots	.96	.35	

---

[11] We considered a variety of augmentation lag orders, and the results were robust—the unit root hypothesis can't be rejected. For the record, the SIC selected one augmentation lag, while the AIC and F-testing selected three augmentation lags.

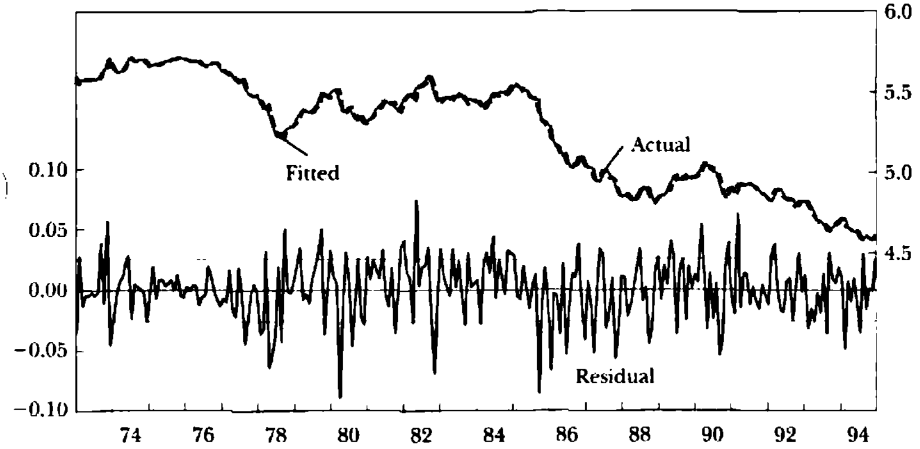


FIGURE 13.10 *Log JPY/USD Exchange Rate, Best-Fitting Deterministic Trend Model, Residual Plot*

The AIC values appear in Table 13.5 and the SIC values in Table 13.6. AIC selects an ARMA(3, 2), and SIC selects an AR(1). Note that the models for $\ln y$ and $\Delta \ln y$ selected by the SIC are consistent with each other under the unit root hypothesis—an AR(2) with a unit root in levels is equivalent to an AR(1) in differences—in contrast to the models selected by the AIC. For this reason and of course for the usual parsimony considerations, we proceed with the AR(1) selected by SIC. We show the regression results in Table 13.7 and Figure 13.14; note the small but nevertheless significant coefficient of 0.32.[12]

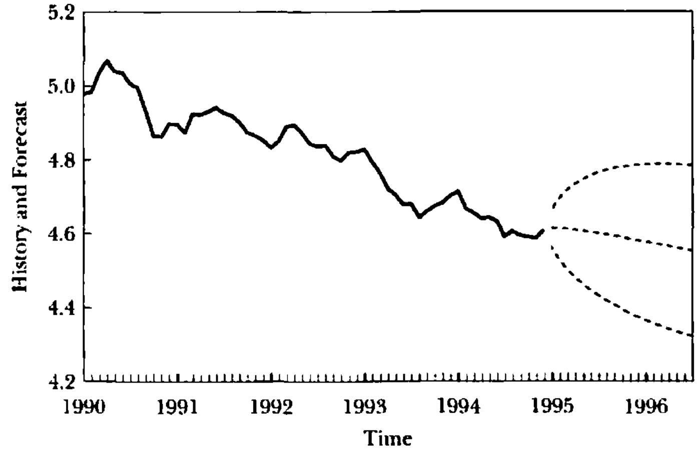


FIGURE 13.11 *Log JPY/USD Rate History and Forecast, AR(2) in Levels with Linear Trend*

---

[12] The ARMA(3, 2) selected by the AIC is in fact very close to an AR(1), because the two estimated MA roots nearly cancel with two of the estimated AR roots, which would leave an AR(1).

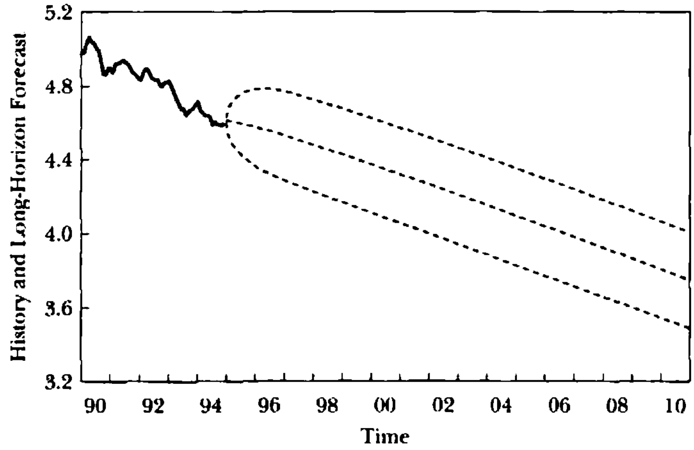


FIGURE 13.12 *Log JPY/USD Rate History and Long-Horizon Forecast, AR(2) in Levels with Linear Trend*

Out-of-sample forecasting results appear in Figures 13.15–13.17. Figure 13.15 shows the history and forecast. The forecast looks very similar—in fact, almost identical—to the forecast from the deterministic trend model examined earlier. That's because the stochastic trend and deterministic trend models are in fact extremely close to one another in this case; even when we don't impose a unit root, we get an estimated dominant root that's very close to unity. In Figure 13.16 we show the history and a very long-horizon forecast. The long-horizon forecast reveals one minor and one major difference between the forecasts from the deterministic trend and stochastic trend models. The minor difference is that, by the time we're out to 2010, the point forecast from the deterministic trend model is a little lower, reflecting the fact that the estimated trend slope is a bit more negative for the deterministic trend model than for the stochastic trend model. Statistically speaking, however, the

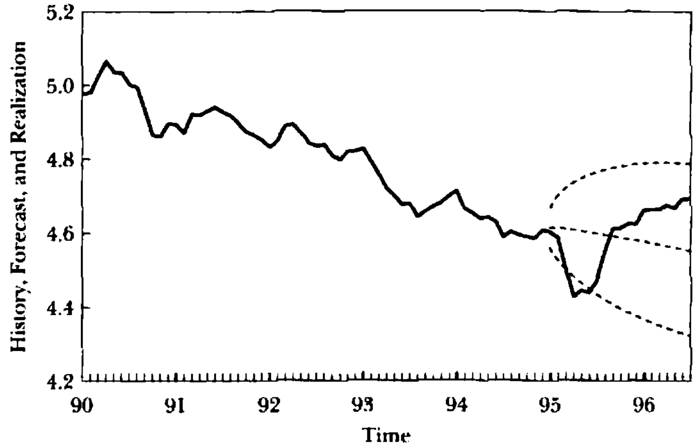


FIGURE 13.13 *Log JPY/USD Rate History, Forecast, and Realization, AR(2) in Levels with Linear Trend*

TABLE 13.4
Log JPY/USD Exchange Rate Augmented Dickey-Fuller Unit Root Test

Augmented Dickey-Fuller test statistic	−2.498863	1%	Critical Value	−3.9966
		5%	Critical Value	−3.4284
		10%	Critical Value	−3.1373

Augmented Dickey-Fuller Test Equation
LS // Dependent variable is *D*(LYEN).
Sample (adjusted): 1973:05 1994:12
Included observations: 260 after adjusting endpoints

Variable	**Coefficient**	**Std. Error**	***t*-Statistic**	**Prob.**
LYEN(−1)	−0.029423	0.011775	−2.498863	0.0131
D(LYEN(−1))	0.362319	0.061785	5.864226	0.0000
D(LYEN(−2))	−0.114269	0.064897	−1.760781	0.0795
D(LYEN(−3))	0.118386	0.061020	1.940116	0.0535
C	0.170875	0.068474	2.495486	0.0132
@TREND(1973:01)	−0.000139	5.27E-05	−2.639758	0.0088

R^2	0.142362	Mean dependent var.	−0.003749
Adjusted R^2	0.125479	SD dependent var.	0.027103
SE of regression	0.025345	Akaike info criterion	−7.327517
Sum squared resid.	0.163166	Schwarz criterion	−7.245348
Log likelihood	589.6532	*F*-statistic	8.432417
Durbin-Watson stat.	2.010829	Prob(*F*-statistic)	0.000000

point forecasts are indistinguishable. The major difference concerns the interval forecasts: The interval forecasts from the stochastic trend model widen continuously as the horizon lengthens, whereas the interval forecasts from the deterministic trend model don't. Finally, in Figure 13.17 we show the history and forecast together with the realization 1995.01–1996.07.

Comparing the AR(2) with trend in levels (the levels model selected by the SIC) and the AR(1) in differences (the differences model selected by the SIC), it appears that the differences model is favored in that it has a lower SIC value. The AR(1) in differences fits only slightly worse than the AR(2) in levels—recall that the AR(2) in levels had a near unit root—and saves 1 degree of freedom.[13] Moreover, economic and financial considerations suggest that exchange rates should be close to random walks, because if the change were predictable, one could make a lot of money with very little effort, and the very act of doing so would eliminate the opportunity.[14]

---

[13] A word of caution: In a sense, the AR(1) model in differences may not save the degree of freedom, insofar as the decision to impose a unit root was itself based on an earlier estimation (the augmented Dickey-Fuller test), which is not acknowledged when computing the SIC for the AR(1) in differences.

[14] As for the trend (drift), it may help as a local approximation, but be wary of too long an extrapolation. See the Exercises, Problems, and Complements at the end of this chapter.

TABLE 13.5
Log JPY/USD Rate, Changes: AIC Values, Various ARMA Models

		MA Order			
		0	1	2	3
	0		−7.298	−7.290	−7.283
AR Order	1	−7.308	−7.307	−7.307	−7.302
	2	−7.312	−7.314	−7.307	−7.299
	3	−7.316	−7.309	−7.340	−7.336

TABLE 13.6
Log JPY/USD Rate, Changes: SIC Values, Various ARMA Models

		MA Order			
		0	1	2	3
	0		−7.270	−7.249	−7.228
AR Order	1	−7.281	−7.266	−7.252	−7.234
	2	−7.271	−7.259	−7.238	−7.217
	3	−7.261	−7.241	−7.258	−7.240

TABLE 13.7
Log JPY/USD Exchange Rate, Best-Fitting Stochastic Trend Model

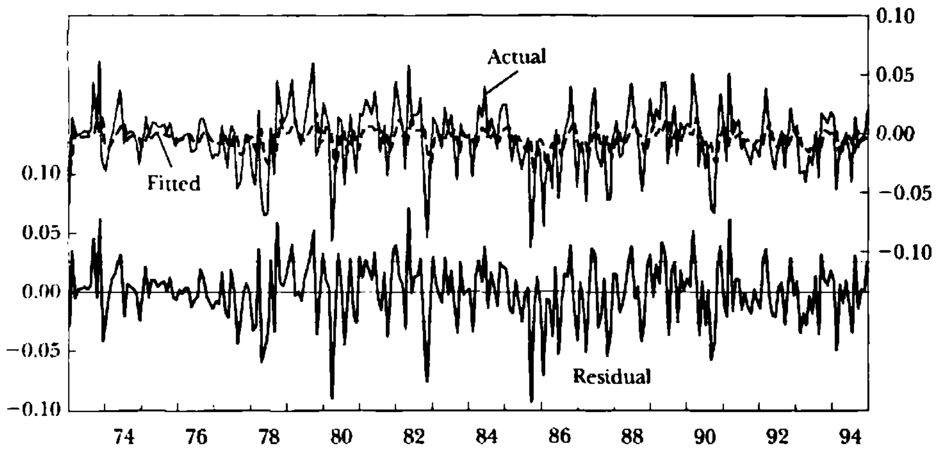
LS // Dependent variable is DLYEN.
Sample (adjusted): 1973:03 1994:12
Included observations: 262 after adjusting endpoints
Convergence achieved after 3 iterations

Variable	**Coefficient**	**Std. Error**	***t*-Statistic**	**Prob.**
C	−0.003697	0.002350	−1.573440	0.1168
AR(1)	0.321870	0.057767	5.571863	0.0000

R^2	0.106669	Mean dependent var.	−0.003888
Adjusted R^2	0.103233	SD dependent var.	0.027227
SE of regression	0.025784	Akaike info criterion	−7.308418
Sum squared resid.	0.172848	Schwarz criterion	−7.281179
Log likelihood	587.6409	*F*-statistic	31.04566
Durbin-Watson stat.	1.948933	Prob(*F*-statistic)	0.000000

Inverted AR roots .32

FIGURE 13.14 *Log JPY/USD Exchange Rate, Best-Fitting Stochastic Trend Model Residual Plot*



Ironically enough, in spite of the arguments in favor of the stochastic-trend model for ln *y*, the deterministic trend model does slightly better in out-of-sample forecasting on this particular dataset. The mean-squared forecast error from the deterministic trend model is 0.0107, while that from the stochastic trend model is 0.0109. The difference, however, is likely statistically insignificant.

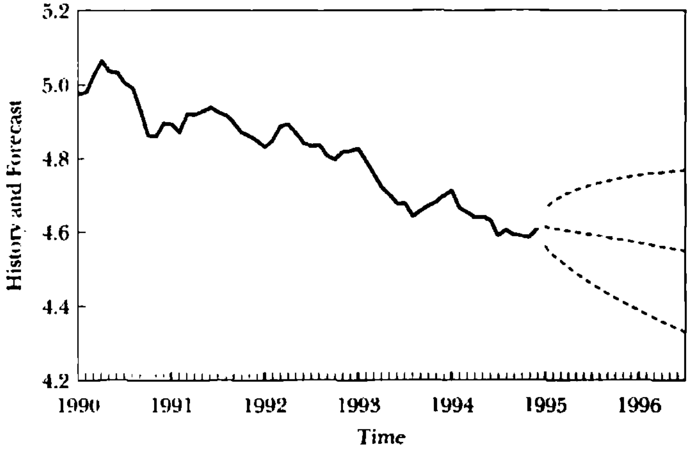
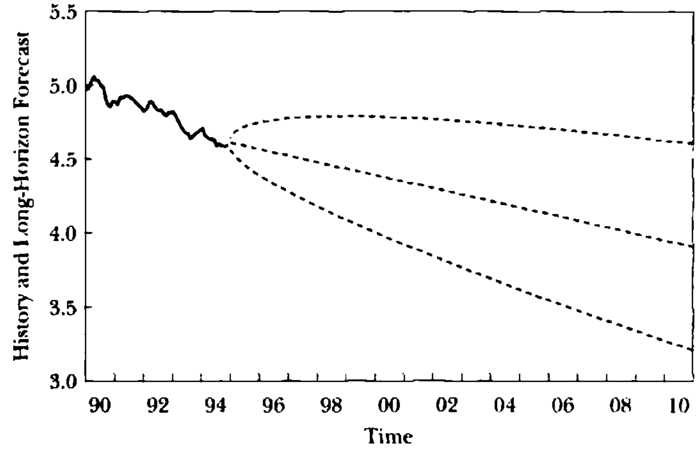


FIGURE 13.15 *Log JPY/USD Rate History and Forecast, AR(1) in Differences with Intercept*

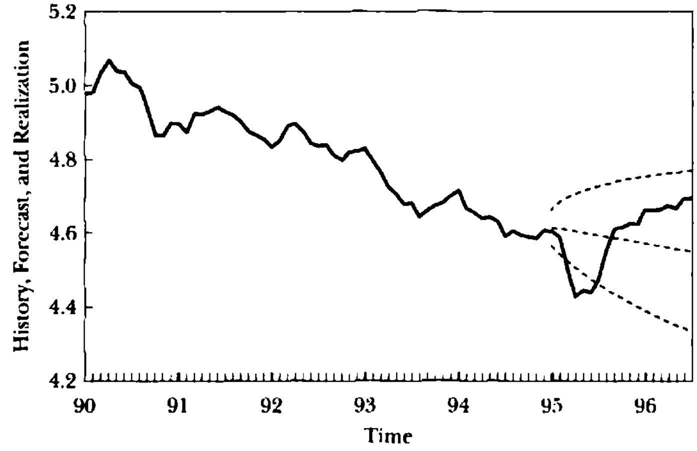
FIGURE 13.16
Log JPY/USD Rate History and Long-Horizon Forecast, AR(1) in Differences with Intercept



4. Smoothing

We bumped into the idea of time series smoothing early on, when we introduced simple moving-average smoothers as ways of estimating trend.[15] Now we introduce additional smoothing techniques and show how they can be used to produce forecasts.

FIGURE 13.17
Log JPY/USD Rate History, Forecast, and Realization, AR(1) in Differences with Intercept



---

[15] See the Exercises, Problems, and Complements of Chapter 5.

Smoothing techniques, as traditionally implemented, have a different flavor than the modern model-based methods that we've used in this book. Smoothing techniques, for example, don't require "best-fitting models," and they don't generally produce "optimal forecasts." Rather, they're simply a way to tell a computer to draw a smooth line through data, just as we'd do with a pencil, and to extrapolate the smooth line in a reasonable and replicable way.

When using smoothing techniques, we make no attempt to find the model that best fits the data: rather, we force a prespecified model on the data. Some academics turn their nose at smoothing techniques for that reason, but such behavior reflects a shallow understanding of key aspects of applied forecasting. Smoothing techniques have been used productively for many years—and for good reason. They're most useful in situations when model-based methods can't, or shouldn't, be used. First, available samples of data are sometimes very small. Suppose, for example, that we must produce a forecast based on a sample of historical data containing only four observations. This scenario sounds extreme, and it is, but such scenarios arise occasionally in certain important applications, as when forecasting the sales of a newly introduced product. In such cases, available degrees of freedom are so limited as to render any estimated model of dubious value. Smoothing techniques, in contrast, require no estimation, or minimal estimation.

Second, the forecasting task is sometimes immense. Suppose, for example, that each week we must forecast the prices of 10,000 inputs to a manufacturing process. Again, such situations are extreme, but they do occur in practice (e.g., think of how many parts there are in a large airplane). In such situations, even if historical data are plentiful (and, of course, they might not be), there is simply no way to provide the tender loving care required for estimation and maintenance of 10,000 different forecasting models. Smoothing techniques, in contrast, require little attention. They're one example of what are sometimes called "automatic" forecasting methods and are often useful for forecasting voluminous, high-frequency data.

Finally, smoothing techniques *do* produce optimal forecasts under certain conditions, which turn out to be intimately related to the presence of unit roots in the series being forecast. That's why we waited until now to introduce them. Moreover, fancier approaches produce optimal forecasts only under certain conditions as well, such as correct specification of the forecasting model. As we've stressed throughout, all our models are approximations, and all are surely false. Any procedure with a successful track record in practice is worthy of serious consideration, and smoothing techniques do have successful track records in the situations sketched here.

MOVING AVERAGE SMOOTHING, REVISITED

As a precursor to the more sophisticated smoothing techniques that we'll soon introduce, recall the workings of simple moving average smoothers. Denote the original data by $\{y_t\}_{t=1}^{T}$ and the smoothed data by $\{\bar{y}_t\}$. Then the two-sided moving average is $\bar{y}_t = (2m+1)^{-1} \sum_{i=-m}^{m} y_{t-i}$, the one-sided moving average is $\bar{y}_t = (m+1)^{-1} \sum_{i=0}^{m} y_{t-i}$, and the one-sided weighted moving average is $\bar{y}_t = \sum_{i=0}^{m} w_i y_{t-i}$. The standard one-sided moving average corresponds to a one-sided weighted moving average with all weights equal to $(m+1)^{-1}$. The user must choose the smoothing parameter, m; the larger m is, the more smoothing is done.

One-sided weighted moving averages turn out to be very useful in practice. The one-sided structure means that at any time t, we need only current and past data for computation of the time-t smoothed value, which means that it can be implemented in real time. The weighting, moreover, enables flexibility in the way that we discount the past. Often, for example, we want to discount the distant past more heavily than the recent past. Exponential smoothing, to which we now turn, is a particularly useful and convenient way of implementing such a moving average.

EXPONENTIAL SMOOTHING

Exponential smoothing, also called **simple exponential smoothing** or **single exponential smoothing,** is what's called an **exponentially weighted moving average,** for reasons that will be apparent soon. The basic framework is simple. Imagine that a series c_0 is a random walk,

$$c_{0,t} = c_{0,t-1} + \eta_t$$
$$\eta_t \sim WN(0, \sigma_\eta^2) \ ,$$

in which case the level of c_0 wanders randomly up and down, and the best forecast of any future value is simply the current value. Suppose, however, that we don't see c_0; instead, we see y, which is c_0 plus white noise,[16]

$$y_t = c_{0,t} + \varepsilon_t \ ,$$

where ε is uncorrelated with η at all leads and lags. Then our optimal forecast of any future y is just our optimal forecast of future c_0, which is current c_0, plus our optimal forecast of future ε, which is 0. The problem, of course, is that we don't know current c_0, the current "local level." We do know current and past y, however, which should contain information about current c_0. When the data-generating process is as written here, exponential smoothing constructs the optimal estimate of c_0—and hence the optimal forecast of any future value of y—on the basis of current and past y. When the data-generating process is not as written here, the exponential smoothing forecast may not be optimal, but recent work suggests that exponential smoothing remains optimal or nearly optimal under surprisingly broad circumstances.[17]

As is common, we state the exponential smoothing procedure as an algorithm for converting the observed series, $\{y_t\}_{t=1}^{T}$, into a smoothed series,

---

[16] We can think of the added white noise as measurement error.

[17] See, in particular, Chatfield et al. (2001).

$\{\bar{y}_t\}_{t=1}^{T}$, and forecasts, $\hat{y}_{T+h,T}$:

1. Initialize at $t = 1$: $\bar{y}_1 = y_1$.
2. Update: $\bar{y}_t = \alpha y_t + (1-\alpha)\bar{y}_{t-1}$, $t = 2, \ldots, T$.
3. Forecast: $\hat{y}_{T+h,T} = \bar{y}_T$.

Referring to the level of c_0, we call $\bar{y}_t$ the estimate of the *level* at time t. The smoothing parameter α is in the unit interval, $\alpha \in [0, 1]$. The smaller is α the smoother the estimated level. As α approaches 0, the smoothed series approaches constancy, and as α approaches 1, the smoothed series approaches point-by-point interpolation. Typically, the more observations we have per unit of calendar time, the more smoothing we need; thus, we'd smooth weekly data (52 observations per year) more than quarterly data (4 observations per year). There is no substitute, however, for a trial-and-error approach involving a variety of values of the smoothing parameter.

It's not obvious at first that the algorithm we just described delivers a one-sided moving average with exponentially declining weights. To convince yourself, start with the basic recursion,

$$\bar{y}_t = \alpha y_t + (1-\alpha)\bar{y}_{t-1} ,$$

and substitute backward for $\bar{y}$, which yields

$$\bar{y}_t = \sum_{j=0}^{t-1} w_j y_{t-j} ,$$

where

$$w_j = \alpha(1-\alpha)^j .$$

Suppose, for example, that $\alpha = 0.5$. Then

$$w_0 = 0.5(1-0.5)^0 = 0.5$$
$$w_1 = 0.5(1-0.5) = 0.25$$
$$w_2 = 0.5(1-0.5)^2 = 0.125 ,$$

and so forth. Thus, moving average weights decline exponentially, as claimed.

Notice that exponential smoothing has a recursive structure, which can be very convenient when data are voluminous. At any time t, the new time-t estimate of the level, $\bar{y}_t$, is a function only of the previously computed estimate, $\bar{y}_{t-1}$, and the new observation, y_t. Thus, there's no need to resmooth the entire dataset as new data arrive.

HOLT-WINTERS SMOOTHING

Now imagine that we have not only a slowly evolving local level but also a trend with a slowly evolving local slope,

$$y_t = c_{0,t} + c_{1,t}\,\mathrm{TIME}_t + \varepsilon_t$$
$$c_{0,t} = c_{0,t-1} + \eta_t$$
$$c_{1,t} = c_{1,t-1} + \nu_t ,$$

where all the disturbances are orthogonal at all leads and lags. Then the optimal smoothing algorithm, named **Holt-Winters smoothing** after the researchers who worked it out in the 1950s and 1960s, is

1. Initialize at $t = 2$:

$$\bar{y}_2 = y_2$$

$$F_2 = y_2 - y_1 .$$

2. Update:

$$\bar{y}_t = \alpha y_t + (1-\alpha)(\bar{y}_{t-1} + F_{t-1}), \quad 0 < \alpha < 1$$

$$F_t = \beta(\bar{y}_t - \bar{y}_{t-1}) + (1-\beta)F_{t-1}, \quad 0 < \beta < 1$$

$$t = 3, 4, \ldots, T.$$

3. Forecast:

$$\hat{y}_{T+h,T} = \bar{y}_T + hF_T .$$

$\bar{y}_t$ is the estimated, or smoothed, level at time t, and F_t is the estimated slope at time t. The parameter α controls smoothing of the level, and β controls smoothing of the slope. The h-step-ahead forecast simply takes the estimated level at time T and augments it with h times the estimated slope at time T.

Again, note that although we've displayed the data-generating process for which Holt-Winters smoothing produces optimal forecasts, when we apply Holt-Winters, we don't assume that the data are actually generated by that process. We hope, however, that the actual data-generating process is close to the one for which Holt-Winters is optimal, in which case the Holt-Winters forecasts may be close to optimal.

HOLT-WINTERS SMOOTHING TO ALLOW FOR SEASONALITY

We can apply **Holt-Winters smoothing with seasonality** with period s. The algorithm becomes

1. Initialize at $t = s$:

$$\bar{y}_s = \frac{1}{s}\sum_{t=1}^{s} y_t$$

$$F_s = 0$$

$$G_j = \frac{y_j}{\frac{1}{s}\sum_{t=1}^{s} y_t}, \quad j = 1, 2, \ldots, s .$$

2. Update:

$$\begin{aligned}
\bar{y}_t &= \alpha(y_t - G_{t-s}) + (1-\alpha)(\bar{y}_{t-1} + F_{t-1}), && 0 < \alpha < 1\\
F_t &= \beta(\bar{y}_t - \bar{y}_{t-1}) + (1-\beta)\, F_{t-1}, && 0 < \beta < 1\\
G_t &= \gamma(y_t - \bar{y}_t) + (1-\gamma)\, G_{t-s}, && 0 < \gamma < 1
\end{aligned}$$

$$t = s+1, \ldots, T\,.$$

3. Forecast:

$$\begin{aligned}
\hat{y}_{T+h,T} &= \bar{y}_T + hF_T + G_{T+h-s}, && h = 1, 2, \ldots, s\,,\\
\hat{y}_{T+h,T} &= \bar{y}_T + hF_T + G_{T+h-2s}, && h = s+1, s+2, \ldots, 2s\,,
\end{aligned}$$

and so forth.

The only thing new is the recursion for the seasonal, with smoothing parameter γ.

FORECASTING WITH SMOOTHING TECHNIQUES

Regardless of which smoothing technique we use, the basic paradigm is the same. We plug data into an algorithm that smooths the data and lets us generate point forecasts. The resulting point forecasts are optimal for certain data-generating processes, as we indicated for simple exponential smoothing and Holt-Winters smoothing without seasonality. In practice, of course, we don't know if the actual data-generating process is close to the one for which the adopted smoothing technique is optimal; instead, we just swallow hard and proceed. That's the main contrast with the model-based approach, in which we typically spend a lot of time trying to find a "good" specification.

The "one size fits all" flavor of the smoothing approach has its costs, because surely one size does *not* fit all, but it also has benefits in that no, or just a few, parameters need be estimated. Sometimes we simply set the smoothing parameter values based on our knowledge of the properties of the series being considered, and sometimes we select parameter values that provide the best *h*-step-ahead forecasts under the relevant loss function. For example, under 1-step-ahead squared-error loss, if the sample size is large enough so that we're willing to entertain estimation of the smoothing parameters, we can estimate them as

$$\hat{\theta} = \underset{\theta}{\operatorname{argmin}} \sum_{t=m+1}^{T} (y_t - \hat{y}_{t-1,t})^2\,,$$

where *m* is an integer large enough such that the start-up values of the smoothing algorithm have little effect.

In closing this section, we note that smoothing techniques, as typically implemented, produce point forecasts only. They may produce optimal point forecasts for certain special data-generating processes, but typically we don't assume that those special data-generating processes are the truth. Instead, the smoothing techniques are used as "black boxes" to produce point forecasts, with no attempt to exploit the stochastic structure of the data to find a best-fitting model, which could be used to produce interval or density forecasts in addition to point forecasts.

5. Exchange Rates, Continued

Now we forecast the JPY/USD exchange rate using a smoothing procedure. In the ARIMA(p, d, q) models considered earlier, we always allowed for a trend (whether deterministic or stochastic). To maintain comparability, we'll use a Holt-Winters smoother, which allows for locally linear trend. We present the estimation results in Table 13.8. The estimate of α is large, so the estimated local level moves closely with the series. The estimate of β is small, so the local slope of the trend is much less adaptive.

The Holt-Winters forecast is simply the trend line beginning at the estimated end-of-period level, with the estimated end-of-period slope. Because the estimated slope of the trend at the end of the sample is larger in absolute value than the corresponding trend slopes in the deterministic trend and stochastic trend models studied earlier, we expect the Holt-Winters point forecasts to decrease a bit more quickly than those from the ARIMA models. In Figure 13.18, we show the history and out-of-sample forecast. No confidence intervals appear with the forecast because the smoothing techniques don't produce them. The forecast looks similar to those of the ARIMA models, except that it drops a bit more quickly, as is made clear by the very long horizon forecast that we show in Figure 13.19. Finally, in Figure 13.20, we show the realization as well. For out-of-sample forecasting, Holt-Winters fares the worst of all the forecasting methods tried in this chapter; the mean-squared forecast error is 0.0135.

TABLE 13.8 *Log JPY/USD Exchange Rate, Holt-Winters Smoothing*

Sample: 1973:01 1994:12
Included observations: 264
Method: Holt-Winters, no seasonal
Original series: LYEN
Forecast series: LYENSM

Parameters:	Alpha	1.000000
	Beta	0.090000
Sum of squared residuals		0.202421
Root mean-squared error		0.027690
End-of-period levels:	Mean	4.606969
	Trend	−0.005193

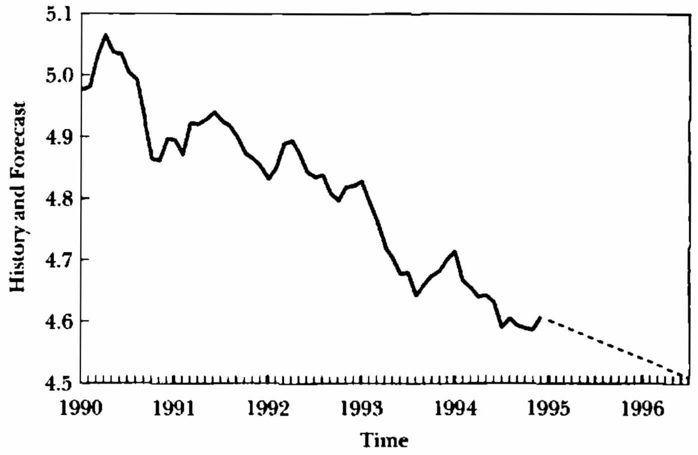


FIGURE 13.18
Log JPY/USD Rate History and Forecast, Holt-Winters Smoothing

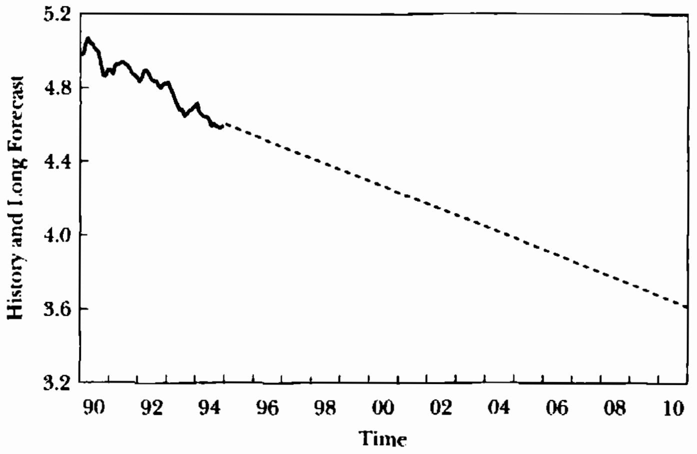


FIGURE 13.19
Log JPY/USD Rate History and Long-Horizon Forecast, Holt-Winters Smoothing

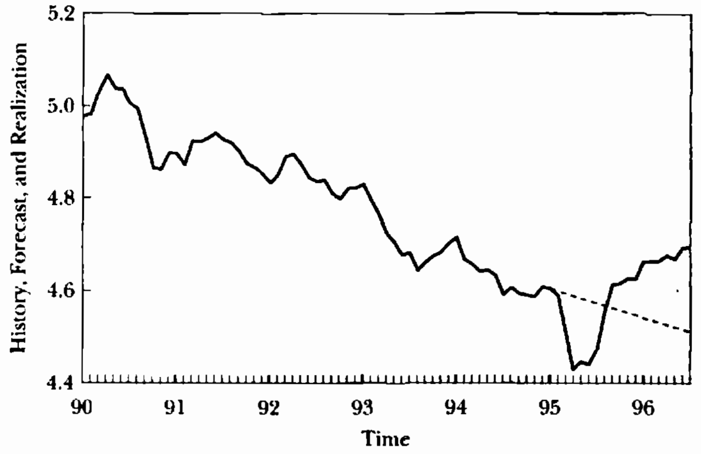


FIGURE 13.20
Log JPY/USD Rate History, Forecast, and Realization, Holt-Winters Smoothing

Exercises, Problems, and Complements

1. (Modeling and forecasting the deutschemark/dollar (DEM/USD) exchange rate) On the book's web page, you'll find monthly data on the DEM/USD exchange rate for the same sample period as the JPY/USD data studied in the text.
   a. Model and forecast the DEM/USD rate, in parallel with the analysis in the text, and discuss your results in detail.
   b. Redo your analysis using forecasting approaches without trends—a levels model without trend, a first-differenced model without drift, and simple exponential smoothing.
   c. Compare the forecasting ability of the approaches with and without trend.
   d. Do you feel comfortable with the inclusion of trend in an exchange rate forecasting model? Why or why not?

2. (Housing starts and completions, continued) As always, our Chapter 11 VAR analysis of housing starts and completions involved many judgment calls. Using the starts and completions data, assess the adequacy of our models and forecasts. Among other things, you may want to consider the following questions:
   a. How would you choose the number of augmentation lags? How sensitive are the results of the augmented Dickey-Fuller tests to the number of augmentation lags?
   b. When performing augmented Dickey-Fuller tests, is it adequate to allow only for an intercept under the alternative hypothesis, or should we allow for both intercept and trend?
   c. Should we allow for a trend in the forecasting model?
   d. Does it make sense to allow a large number of lags in the augmented Dickey-Fuller tests, but not in the actual forecasting model?
   e. How do the results change if, in light of the results of the causality tests, we exclude lags of completions from the starts equation, reestimate by seemingly unrelated regression, and forecast?
   f. Are the VAR forecasts of starts and completions more accurate than univariate forecasts?

3. (ARIMA models, smoothers, and shrinkage) From the vantage point of the shrinkage principle, discuss the trade-offs associated with "optimal" forecasts from fitted ARIMA models versus "ad hoc" forecasts from smoothers.

4. (Using stochastic trend unobserved-components models to implement smoothing techniques in a probabilistic framework) We noted that smoothing techniques, as typically implemented, are used as "black boxes" to produce point forecasts. There is no attempt to exploit stochastic structure to produce interval or density forecasts in addition to point forecasts. Recall, however, that the various smoothers produce optimal forecasts for specific data-generating processes specified as unobserved-components models.
   a. For what data-generating process is exponential smoothing optimal?
   b. For what data-generating process is Holt-Winters smoothing optimal?
   c. Under the assumption that the data-generating process for which exponential smoothing produces optimal forecasts is in fact the true data-generating process, how might you estimate the unobserved-components model and use

it to produce optimal interval and density forecasts? (*Hint:* Browse through Harvey, 1989.)

d. How would you interpret the interval and density forecasts produced by the method of part *c*, if we no longer assume a particular model for the true data-generating process?

5. (Automatic ARIMA modeling) "Automatic" forecasting software exists for implementing the ARIMA and exponential smoothing techniques of this and previous chapters without any human intervention.
   a. What are do you think are the benefits of such software?
   b. What do you think are the costs?
   c. When do you think it would be most useful?
   d. Read Ord and Lowe (1996), who review most of the automatic forecasting software, and report what you learned. After reading Ord and Lowe, how, if at all, would you revise your answers to parts *a*, *b*, and *c*?
6. (The multiplicative seasonal ARIMA$(p, d, q) \times (P, D, Q)$ model) Consider the forecasting model,

$$\Phi_s(L^s)\Phi(L)(1-L)^d(1-L^s)^D y_t = \Theta_s(L^s)\Theta(L)\varepsilon_t$$

$$\Phi_s(L^s) = 1 - \varphi_1^s L^s - \cdots - \varphi_P^s L^{Ps}$$

$$\Phi(L) = 1 - \varphi_1 L - \cdots - \varphi_p L^p$$

$$\Theta_s(L^s) = 1 - \theta_1^s L^s - \cdots - \theta_Q^s L^{Qs}$$

$$\Theta(L) = 1 - \theta_1 L - \cdots - \theta_q L^q .$$

   a. The standard ARIMA(p, d, q) model is a special case of this more general model. In what situation does it emerge? What is the meaning of the ARIMA $(p, d, q) \times (P, D, Q)$ notation?
   b. The operator $(1 - L^s)$ is called the *seasonal difference operator.* What does it do when it operates on y_t? Why might it routinely appear in models for seasonal data?
   c. The appearance of $(1 - L^s)$ in the autoregressive lag operator polynomial moves us into the realm of **stochastic seasonality,** in contrast to the deterministic seasonality of Chapter 6, just as the appearance of $(1 - L)$ produces stochastic as opposed to deterministic trend. Comment.
   d. Can you provide some intuitive motivation for the model? Hint: Consider a purely seasonal ARIMA(P, D, Q) model, shocked by serially correlated disturbances. Why might the disturbances be serially correlated? What, in particular, happens if an ARIMA(P, D, Q) model has ARIMA(p, d, q) disturbances?
   e. The multiplicative structure implies restrictions. What, for example, do you get when you multiply $\Phi_s(L)$ and $\Phi(L)$?
   f. What do you think are the costs and benefits of forecasting with the multiplicative ARIMA model versus the "standard" ARIMA model?
   g. Recall that in Chapter 10 we analyzed and forecasted liquor sales using an ARMA model with deterministic trend. Instead, analyze and forecast liquor sales using an ARIMA $(p, d, q) \times (P, D, Q)$ model, and compare the results.
7. (The Dickey-Fuller regression in the AR(2) case) Consider the AR(2) process,

$$y_t + \varphi_1 y_{t-1} + \varphi_2 y_{t-2} = \varepsilon_t .$$

a. Show that it can be written as

$$y_t = \rho_1 y_{t-1} + \rho_2(y_{t-1} - y_{t-2}) + \varepsilon_t \, ,$$

where

$$\rho_1 = -(\varphi_1 + \varphi_2)$$

$$\rho_2 = \varphi_2 \, .$$

b. Show that it can also be written as a regression of Δy_t on y_{t-1} and Δy_{t-1}.

c. Show that if $\rho_1 = 1$, the AR(2) process is really an AR(1) process in first differences; that is, the AR(2) process has a unit root.

8. (Holt-Winters smoothing with multiplicative seasonality) Consider a seasonal Holt-Winters smoother, written as

1. Initialize at $t = s$:

$$\bar{y}_s = \frac{1}{s}\sum_{t=1}^{s} y_t$$

$$T_s = 0$$

$$F_j = \frac{y_j}{\left(\frac{1}{s}\sum_{t=1}^{s} y_t\right)}, \quad j = 1, 2, \ldots, s$$

2. Update:

$$\bar{y}_t = \alpha\left(\frac{y_t}{F_{t-s}}\right) + (1-\alpha)(\bar{y}_{t-1} + T_{t-1}), \quad 0 < \alpha < 1$$

$$T_t = \beta(\bar{y}_t - \bar{y}_{t-1}) + (1-\beta)T_{t-1}, \quad 0 < \beta < 1$$

$$F_t = \gamma\left(\frac{y_t}{\bar{y}_t}\right) + (1-\gamma)F_{t-s}, \quad 0 < \gamma < 1$$

$$t = s+1, \ldots, T \, .$$

3. Forecast:

$$\hat{y}_{T+h,T} = (\bar{y}_T + hT_T)F_{T+h-s}, \quad h = 1, 2, \ldots, s,$$

$$\hat{y}_{T+h,T} = (\bar{y}_T + hT_T)F_{T+h-2s}, \quad h = s+1, s+2, \ldots, 2s,$$

and so forth.

a. The Holt-Winters seasonal smoothing algorithm given in the text is more precisely called *Holt-Winters seasonal smoothing with additive seasonality*. The algorithm given here, in contrast, is called *Holt-Winters seasonal smoothing with multiplicative seasonality*. How does this algorithm differ from the one given in the text, and what, if anything, is the significance of the difference?

b. Assess the claim that Holt-Winters with multiplicative seasonality is appropriate when the seasonal pattern exhibits increasing variation.

c. How does Holt-Winters with multiplicative seasonality compare with the use of Holt-Winters with additive seasonality applied to logarithms of the original data?

9. (Cointegration) Consider two series, x and y, both of which are $I(1)$. In general, there is no way to form a weighted average of x and y to produce an $I(0)$ series, but in the very special case where such a weighting does exist, we say that x and y are cointegrated. **Cointegration** corresponds to situations in which variables tend to cling to one another, in the sense that the cointegrating combination is stationary, even though each variable is nonstationary. Such situations arise frequently in business, economics, and finance. To take a business example, it's often the case that both inventories and sales of a product appear $I(1)$, yet their ratio (or, when working in logs, their difference) appears $I(0)$, a natural by-product of various schemes that adjust inventories to sales. Engle and Granger (1987) is the key early research paper on cointegration; Johansen (1995) surveys most of the more recent developments, with an emphasis on maximum likelihood estimation.
   a. Consider the bivariate system

$$x_t = x_{t-1} + v_t, \qquad v_t \sim WN(0, \sigma^2)$$
$$y_t = x_t + \varepsilon_t, \qquad \varepsilon_t \sim WN(0, \sigma^2)$$

   Both x and y are $I(1)$. Why? Show, in addition, that x and y are cointegrated. What is the cointegrating combination?
   b. Engle and Yoo (1987) show that optimal long-run forecasts of cointegrated variables obey the cointegrating relationship exactly. Verify their result for the system at hand.

10. (Error correction) In an **error correction** model, we take a long-run model relating $I(1)$ variables, and we augment it with short-run dynamics. Suppose, for example, that in long-run equilibrium, y and x are related by $y = bx$. Then the deviation from equilibrium is $z = y - bx$, and the deviation from equilibrium at any time may influence the future evolution of the variables, which we acknowledge by modeling Δx as a function of lagged values of itself, lagged values of Δy, *and the lagged value of* z, *the error correction term*. For example, allowing for one lag of Δx and one lag of Δy on the right side, we write equation for x as

$$\Delta x_t = \alpha_x \Delta x_{t-1} + \beta_x \Delta y_{t-1} + \gamma_x z_{t-1} + \varepsilon_{xt} \, .$$

   Similarly, the y equation is

$$\Delta y_t = \alpha_y \Delta x_{t-1} + \beta_y \Delta y_{t-1} + \gamma_y z_{t-1} + \varepsilon_{yt} \, .$$

   So long as one or both of γ_x and γ_y are nonzero, the system is very different from a VAR in first differences; the key feature that distinguishes the error correction system from a simple VAR in first differences is the inclusion of the error correction term, so that the deviation from equilibrium affects the evolution of the system.
   a. Engle and Granger (1987) establish the key result that existence of cointegration in a VAR and existence of error correction are equivalent—a VAR is cointegrated if and only if it has an error correction representation. Try to sketch some intuition as to why the two should be linked. Why, in particular, might cointegration imply error correction?
   b. Why are cointegration and error correction of interest to forecasters in business, finance, economics, and government?
   c. Evaluation of forecasts of cointegrated series poses special challenges, insofar as traditional accuracy measures don't value the preservation of cointegrating

relationships, whereas presumably they *should*. For details and constructive suggestions, see Christoffersen and Diebold (1998).

11. (Forecast encompassing tests for $I(1)$ series) An alternative approach to testing for forecast encompassing, which complements the one presented in Chapter 12, is particularly useful in $I(1)$ environments. It's based on forecasted h-step *changes*. We run the regression

$$(y_{t+h} - y_t) = \beta_a \left(y^a_{t+h,t} - y_t\right) + \beta_b \left(y^b_{t+h,t} - y_t\right) + \varepsilon_{t+h,t} .$$

As before, forecast encompassing corresponds to coefficient values of (1, 0) or (0, 1). Under the null hypothesis of forecast encompassing, the regression based on levels and the regression based on changes are identical.

12. (Evaluating forecasts of integrated series) The unforecastability principle remains intact regardless of whether the series being forecast is stationary or integrated. The errors from optimal forecasts are not predictable on the basis of information available at the time the forecast was made. However, some additional implications of the unforecastability principle emerge in the case of forecasting $I(1)$ series, including these:

a. If the series being forecast is $I(1)$, then so, too, is the optimal forecast.
b. An $I(1)$ series is always cointegrated with its optimal forecast, which means that there exists an $I(0)$ linear combination of the series and its optimal forecast, in spite of the fact that both the series and the forecast are $I(1)$.
c. The cointegrating combination is simply the difference of the actual and forecasted values—the forecast error. Thus, the error corresponding to an optimal forecast of an $I(1)$ series is $I(0)$, in spite of the fact that the series is not.

Cheung and Chinn (1999) make good use of these results in a study of the information content of U.S. macroeconomic forecasts; try to sketch their intuition. (*Hint:* Suppose the error in forecasting an $I(1)$ series were *not* $I(0)$. What would that imply?)

13. (Theil's U-statistic) Sometimes it's informative to compare the accuracy of a forecast to that of a "naive" competitor. A simple and popular such comparison is achieved by the U-statistic, which is the ratio of the 1-step-ahead MSE for a given forecast relative to that of a random walk forecast $y_{t+1,t} = y_t$; that is,

$$U = \frac{\sum_{t=1}^{T} (y_{t+1} - y_{t+1,t})^2}{\sum_{t=1}^{T} (y_{t+1} - y_t)^2} .$$

One must remember, of course, that the random walk is not necessarily a naive competitor, particularly for many economic and financial variables, so that values of U near 1 are not necessarily "bad."

The U-statistic is due to Theil (1966, p. 28) and is often called *Theil's U-statistic*. Several authors, including Armstrong and Fildes (1995), have advocated using the U-statistic and close relatives for comparing the accuracy of various forecasting methods across series.

Bibliographical and Computational Notes

We expect random walks, or near-random walks, to be good models for financial asset prices, and they are. See Malkiel (1999). More general ARIMA(p, 1, q) models have found wide application in business, finance, economics, and government. Beveridge and Nelson (1981) show that $I(1)$ processes can always be decomposed into the sum of a random walk component and a covariance stationary component. Tsay (1984) shows that information criteria such as the SIC remain valid for selecting ARMA model orders, regardless of whether a unit autoregressive root is present.

In parallel to the Nerlove, Grether, and Carvalho (1979) treatment of unobserved-components models with deterministic trend, Harvey (1989) treats specification, estimation, and forecasting with unobserved-components models with stochastic trend, estimated by using state-space representations in conjunction with the Kalman filter.

The forecasts of U.S. GNP per capita that we examine in the text, and the related discussion, draw heavily on Diebold and Senhadji (1996).

Development of methods for removing the Dickey-Fuller bias from the parameters of estimated forecasting models, which might lead to improved forecasts, is currently an active research area. See, among others, Andrews (1993), Rudebusch (1993), and Fair (1996).

In an influential book, Box and Jenkins propose an iterative modeling process that consists of repeated cycles of model specification, estimation, diagnostic checking, and forecasting. (The latest edition is Box, Jenkins, and Reinsel, 1994.) One key element of the Box-Jenkins modeling strategy is the assumption that the data follow an ARIMA model (sometimes called a *Box-Jenkins model*),

$$\Phi(L)(1 - L)^d y_t = \Theta(L)\varepsilon_t \, .$$

Thus, although y_t is nonstationary, it is assumed that its dth difference follows a stationary and invertible ARMA process. The appropriateness of the Box-Jenkins tactic of differencing to achieve stationarity depends on the existence of one or more unit roots in the autoregressive lag operator polynomial, which is partly responsible for the large amount of subsequent research on unit root tests.

Dickey-Fuller tests trace to Dickey (1976) and Fuller (1976). Using simulation techniques, MacKinnon (1991) obtains highly accurate estimates of the percentage points of the various Dickey-Fuller distributions.

Alternatives to Dickey-Fuller unit root tests, called *Phillips-Perron tests*, are proposed in Phillips and Perron (1988). The basic idea of Phillips-Perron tests is to estimate a Dickey-Fuller regression without augmentation,

$$x_t = \varphi x_{t-1} + e_t \, ,$$

and then to correct the Dickey-Fuller statistic for general forms of serial correlation and/or heteroskedasticity that might be present in e_t. See Hamilton (1994) for detailed discussion of the Phillips-Perron tests and comparison with augmented Dickey-Fuller tests.

A key question for forecasters is determination of the comparative costs of misspecifying forecasting models in levels versus differences, as a function of sample size, forecast horizon, true value of the dominant root, and so on. Related, we need to learn more about the efficacy for forecasting of rules such as "Impose a unit root unless a Dickey-Fuller test rejects at the 5% level."[18] Campbell and Perron (1991) make some initial progress in that direction; Diebold and Kilian (2000) explore the issue in detail and argue that such strategies are likely to be successful; and in an extensive forecasting competition, Stock and Watson (1999) show that such strategies are in fact successful.

Smoothing techniques were originally proposed as reasonable, if ad hoc, forecasting strategies; only later were they formalized in terms of optimal forecasts for underlying stochastic trend unobserved-components models. This idea—implementing smoothing techniques in stochastic environments via stochastic trend unobserved-components models—is a key theme of Harvey (1989), which also contains references to important earlier contributions to the smoothing literature, including Holt (1957) and Winters (1960). The impressive Stamp software of Koopman, Harvey, Doornik, and Shephard (1995) can be used to estimate and diagnose stochastic trend unobserved-components models and to produce forecasts.[19] *Stamp* stands for "structural time series analyzer, modeler, and predictor"; unobserved-components models are sometimes called *structural time series models.*

Concepts for Review

- Unit autoregressive root
- Unit root
- Random walk
- Random walk with drift
- Random walk without drift
- Mean reversion
- ARIMA(p, 1, q) model
- ARIMA(p, d, q) model
- Shock persistence
- Superconsistency
- Dickey-Fuller distribution
- Unit root test with nonzero mean allowed under the alternative hypothesis
- Unit root test with deterministic linear trend allowed under the alternative hypothesis
- Augmented Dickey-Fuller test
- Exponential smoothing
- Simple exponential smoothing
- Single exponential smoothing
- Exponentially weighted moving average
- Holt-Winters smoothing
- Holt-Winters smoothing with seasonality
- Cointegration
- Error correction

References and Additional Readings

Armstrong, J. S., and Fildes, R. (1995). "On the Selection of Error Measures for Comparisons among Forecasting Methods." *Journal of Forecasting*, 14, 67–71.

Andrews, D. W. K. (1993). "Exactly Median-Unbiased Estimation of First Order Autoregressive/Unit Root Models." *Econometrica*, 61, 139–165.

---

[18] Notice that, once again, the shrinkage principle appears. Such rules, which impose the unit root unless there's strong evidence against it, may lead to forecasting models that perform better than unrestricted models, even if the unit root restriction is false. And of course if the restriction is true, it's helpful to impose it—all the more so in light of the Dickey-Fuller bias that plagues estimation in levels.

[19] For a review, see Diebold, Giorgianni, and Inoue (1996).

Beveridge, S., and Nelson, C. R. (1981). "A New Approach to the Decomposition of Economic Time Series into Permanent and Transient Components with Particular Attention to Measurement of the Business Cycle." *Journal of Monetary Economics*, 7, 151–174.

Box, G. E. P., Jenkins, G. W., and Reinsel, G. (1994). *Time Series Analysis, Forecasting and Control.* 3rd ed. Englewood Cliffs, N.J.: Prentice Hall.

Campbell, J. Y., and Perron, P. (1991). "Pitfalls and Opportunities: What Macroeconomists Should Know about Unit Roots." In O. J. Blanchard and S. S. Fischer (eds.), *NBER Macroeconomics Annual, 1991.* Cambridge, Mass.: MIT Press.

Chatfield, C., Koehler, A. B., Ord, J. K., and Snyder, R. D. (2001). "A New Look at Models for Exponential Smoothing." *The Statistician*, 50, part 2, 147–159.

Cheung, Y.-W., and Chinn, M. D. (1996). "Are Macroeconomic Forecasts Informative? Cointegration Evidence from the ASA/NBER Surveys." Working Paper #350, University of California, Santa Cruz.

Christoffersen, P. F., and Diebold, F. X. (1998). "Cointegration and Long-Horizon Forecasting." *Journal of Business and Economic Statistics*, 16, 450–458.

Dickey, D. A. (1976). *Estimation and Hypothesis Testing in Nonstationary Time Series.* Ph.D. dissertation, Iowa State University.

Dickey, D. A., and Gonzalez-Farias, G. (1992). "A New Maximum-Likelihood Approach to Testing for Unit Roots." Manuscript, Department of Statistics, North Carolina State University.

Diebold, F. X., Giorgianni, L., and Inoue, A. (1996). "STAMP 5.0: A Review." *International Journal of Forecasting*, 12, 309–315.

Diebold, F. X., and Kilian, L. (2000). "Unit Root Tests Are Useful for Selecting Forecasting Models." *Journal of Business and Economic Statistics*, 18, 265–273.

Diebold, F. X., and Rudebusch, G. D. (1999). *Business Cycles: Durations, Dynamics, and Forecasting.* Princeton, N.J.: Princeton University Press.

Diebold, F. X., and Senhadji, A. (1996). "The Uncertain Unit Root in Real GNP: Comment." *American Economic Review*, 86, 1291–1298. Reprinted in Diebold and Rudebusch (1999).

Elliott, G., Rothenberg, T. J., and Stock, J. H. (1996). "Efficient Tests for an Autoregressive Unit Root." *Econometrica*, 64, 813–836.

Engle, R. F., and Granger, C. W. J. (1987). "Co-Integration and Error Correction: Representation, Estimation and Testing." *Econometrica*, 55, 251–276.

Engle, R. F., and Yoo, B. S. (1987). "Forecasting and Testing in Cointegrated Systems." *Journal of Econometrics*, 35, 143–159.

Fair, R. C. (1996). "Computing Median Unbiased Estimates in Macroeconometric Models." *Journal of Applied Econometrics*, 11, 431–435.

Fuller, W. A. (1976). *Introduction to Statistical Time Series.* New York: Wiley.

Hall, A. (1994). "Testing for a Unit Root in Time Series with Pretest Data-Based Model Selection." *Journal of Business and Economic Statistics*, 12, 461–470.

Hamilton, J. D. (1994). *Time Series Analysis.* Princeton, N.J.: Princeton University Press.

Harvey, A. C. (1989). *Forecasting, Structural Time Series Models and the Kalman Filter.* Cambridge: Cambridge University Press.

Holt, C. C. (1957). "Forecasting Seasonals and Trends by Exponentially Weighted Moving Averages." ONR Research Memorandum No. 52, Carnegie Institute of Technology, Pittsburgh.

Johansen, S. (1995). *Likelihood Based Inference in Cointegrated Vector Autoregressive Models.* Oxford: Oxford University Press.

Koopman, S. J., Harvey, A. C., Doornik, J. A., and Shephard, N. (1995). *Stamp 5.0: Structural Time Series Analyzer, Modeller and Predictor.* London: Chapman and Hall.

MacKinnon, J. G. (1991). "Critical Values for Cointegration Tests." In R. F. Engle and C. W. J. Granger (eds.), *Long-Run Economic Relationships.* Oxford: Oxford University Press.

Malkiel, B. G. (1999). *A Random Walk Down Wall Street.* 7th ed. New York: Norton.

Nerlove, M., Grether, D. M., and Carvalho, J. L. (1979). *Analysis of Economic Time Series: A Synthesis* (2nd ed., 1996). New York: Academic Press.

Ng, S., and Perron, P. (1995). "Unit Root Tests in ARMA Models with Data-Dependent Methods for the Selection of the Truncation Lag." *Journal of the American Statistical Association*, 90, 268–281.

Ord, K. and Lowe, S. (1996). "Automatic Forecasting." *American Statistician*, 50, 88–94.

Pantula, S. G., Gonzalez-Farias, G., and Fuller, W. A. (1994). "A Comparison of Unit Root Test Criteria." *Journal of Business and Economic Statistics*, 12, 449–459.

Phillips, P. C. B., and Perron, P. (1988). "Testing for a Unit Root in Time Series Regression." *Biometrika*, 75, 335–346.

Rudebusch, G. D. (1993). "The Uncertain Unit Root in Real GNP." *American Economic Review*, 83, 264–272. Reprinted in Diebold and Rudebusch (1999).

Stock, J. H., and Watson, M. W. (1999). "A Comparison of Linear and Nonlinear Univariate Models for Forecasting Macroeconomic Time Series." In R. Engle and H. White (eds.), *Cointegration, Causality, and Forecasting: A Festschrift in Honor of Clive W. J. Granger.* Oxford: Oxford University Press, 1–44.

Theil, H. (1966). *Applied Economic Forecasting.* Amsterdam: North-Holland.

Tsay, R. (1984). "Order Selection in Nonstationary Autoregressive Models." *Annals of Statistics*, 12, 1425–1433.

Winters, P. R. (1960). "Forecasting Sales by Exponentially Weighted Moving Averages." *Management Science*, 6, 324–342.

Volatility Measurement, Modeling, and Forecasting

The celebrated Wold decomposition makes clear that every covariance stationary series may be viewed as ultimately driven by underlying weak white noise innovations. Hence, it is no surprise that every forecasting model discussed in this book is driven by underlying white noise. To take a simple example, if the series y_t follows an AR(1) process, then

$$y_t = \varphi y_{t-1} + \varepsilon_t,$$

where ε_t is white noise. In some situations, it is inconsequential whether ε_t is weak or strong white noise—that is, whether ε_t is independent, as opposed to merely serially uncorrelated. Hence, to simplify matters, we sometimes assume strong white noise,

$$\varepsilon_t \overset{\text{iid}}{\sim} (0, \sigma^2).$$

Throughout this book, we have thus far taken that approach, sometimes explicitly and sometimes implicitly.

When ε_t is independent, there is no distinction between the unconditional distribution of ε_t and the distribution of ε_t conditional on its past, by definition of independence. Hence, σ^2 is both the unconditional and conditional variance of ε_t. The Wold decomposition, however, does not require that ε_t be serially independent; rather, it requires only that ε_t be serially uncorrelated.

If ε_t is dependent, then its unconditional and conditional distributions will differ. We denote the unconditional innovation distribution by

$$\varepsilon_t \sim (0, \sigma^2)\,.$$

We are particularly interested in conditional dynamics characterized by **heteroskedasticity,** or time-varying volatility. We denote the conditional distribution by

$$\varepsilon_t \mid \Omega_{t-1} \sim \left(0, \sigma_t^2\right),$$

where $\Omega_{t-1} = \{\varepsilon_{t-1}, \varepsilon_{t-2}, \ldots\}$. The conditional variance σ_t^2 will in general evolve as Ω_{t-1} evolves, which focuses attention on the possibility of time-varying innovation volatility.[1]

Allowing for **time-varying volatility** is crucially important in certain economic and financial contexts. The volatility of financial asset returns, for example, is often time varying. That is, markets are sometimes tranquil and sometimes turbulent, as can readily be seen by examining the time series of stock market returns in Figure 14.1, to which we shall return in detail. Time-varying volatility has important implications for financial risk management, asset allocation, and asset pricing, and it has therefore become a central part of the emerging field of **financial econometrics.** Quite apart from financial applications, however, time-varying volatility also has direct implications for interval and density forecasting in a wide variety of applications: correct confidence intervals and density forecasts in the presence of volatility fluctuations require time-varying confidence interval widths and time-varying density forecast spreads. The forecasting models that we have considered thus far, however, do not allow for that possibility. In this chapter we do so.

1. The Basic ARCH Process

Consider the general linear process,

$$y_t = B(L)\varepsilon_t$$

$$B(L) = \sum_{i=0}^{\infty} b_i L^i \qquad \sum_{i=0}^{\infty} b_i^2 < \infty \qquad b_0 = 1$$

$$\varepsilon_t \sim WN(0, \sigma^2)\,.$$

We will work with various cases of this process.

Suppose first that ε_t is strong white noise. $\varepsilon_t \overset{iid}{\sim} (0, \sigma^2)$. Let us review some results already discussed for the general linear process, which will prove useful in what follows. The *unconditional* mean and variance of y are $E(y_t) = 0$ and $E(y_t^2) = \sigma^2 \sum_{i=0}^{\infty} b_i^2$, which are both time-invariant, as must be the case under covariance stationarity. However, the *conditional* mean of y is time varying: $E(y_t \mid \Omega_{t-1}) = \sum_{i=1}^{\infty} b_i \varepsilon_{t-i}$, where the information set is $\Omega_{t-1} = \{\varepsilon_{t-1}, \varepsilon_{t-2}, \ldots\}$. The ability of the general linear process to capture covariance stationary conditional mean dynamics is the source of its power.

---

[1] In principle, aspects of the conditional distribution other than the variance, such as conditional skewness, could also fluctuate. Conditional variance fluctuations are by far the most important in practice, however, so we assume that fluctuations in the conditional distribution of ε are due exclusively to fluctuations in σ_t^2.

Because the volatility of many economic time series varies, one would hope that the general linear process could capture conditional variance dynamics as well, but such is not the case for the model as presently specified: The conditional variance of y is constant at $E((y_t - E(y_t \mid \Omega_{t-1}))^2 \mid \Omega_{t-1}) = \sigma^2$. This potentially unfortunate restriction manifests itself in the properties of the h-step-ahead conditional prediction error variance. The minimum mean squared error forecast is the conditional mean,

$$E(y_{t+h} \mid \Omega_t) = \sum_{i=0}^{\infty} b_{h+i} \varepsilon_{t-i} \, ,$$

and so the associated prediction error is

$$y_{t+h} - E(y_{t+h} \mid \Omega_t) = \sum_{i=0}^{h-1} b_i \varepsilon_{t+h-i} \, ,$$

which has a conditional prediction error variance of

$$E((y_{t+h} - E(y_{t+h} \mid \Omega_t))^2 \mid \Omega_t) = \sigma^2 \sum_{i=0}^{h-1} b_i^2 \, .$$

The conditional prediction error variance is different from the unconditional variance, but it is not time varying. It depends only on h, not on the conditioning information Ω_t. In the process as presently specified, the conditional variance is not allowed to adapt to readily available and potentially useful conditioning information.

So much for the general linear process with iid innovations. Now we extend it by allowing ε_t to be weak rather than strong white noise, *with a particular nonlinear dependence structure.* In particular, suppose that, as before,

$$y_t = B(L)\varepsilon_t$$

$$B(L) = \sum_{i=0}^{\infty} b_i L^i \qquad \sum_{i=0}^{\infty} b_i^2 < \infty \qquad b_0 = 1 \, ,$$

but now suppose as well that

$$\varepsilon_t \mid \Omega_{t-1} \sim N(0, \sigma_t^2)$$

$$\sigma_t^2 = \omega + \gamma(L)\varepsilon_t^2$$

$$\omega > 0 \quad \gamma(L) = \sum_{i=1}^{p} \gamma_i L^i \quad \gamma_i \geq 0 \text{ for all } i \quad \sum \gamma_i < 1 \, .$$

Note that we parameterize the innovation process in terms of its conditional density, $\varepsilon_t \mid \Omega_{t-1}$, which we assume to be normal with a zero conditional mean and a conditional variance that depends linearly on p past squared innovations.

ε_t is serially uncorrelated but not serially independent, because the current conditional variance σ_t^2 depends on the history of ε_t.[2] The stated regularity conditions are sufficient to ensure that the conditional and unconditional variances are positive and finite and that y_t is covariance stationary.

The unconditional moments of ε_t are constant and are given by $E(\varepsilon_t) = 0$ and $E((\varepsilon_t - E(\varepsilon_t))^2) = \frac{\omega}{1 - \sum \gamma_i}$. The important result is not the particular formulas for the unconditional mean and variance but the fact that they are fixed, as required for covariance stationarity. As for the conditional moments of ε_t, its conditional variance is time varying,

$$E((\varepsilon_t - E(\varepsilon_t \mid \Omega_{t-1}))^2 \mid \Omega_{t-1}) = \omega + \gamma(L)\varepsilon_t^2 ,$$

and of course its conditional mean is zero by construction.

Assembling the results to move to the unconditional and conditional moments of y as opposed to ε_t, it is easy to see that both the unconditional mean and variance of y are constant (again, as required by covariance stationarity) but that both the conditional mean and variance are time varying:

$$E(y_t \mid \Omega_{t-1}) = \sum_{i=1}^{\infty} b_i \varepsilon_{t-i}$$

$$E((y_t - E(y_t \mid \Omega_{t-1}))^2 \mid \Omega_{t-1}) = \omega + \gamma(L)\varepsilon_t^2 .$$

Thus, we now treat conditional mean and variance dynamics in a symmetric fashion by allowing for movement in each, as determined by the evolving information set Ω_{t-1}.

In the development just described, ε_t is called an **ARCH(p) process,** and the full model sketched is an infinite-ordered moving average with ARCH(p) innovations, where ARCH stands for autoregressive conditional heteroskedasticity. Clearly ε_t is conditionally heteroskedastic, because its conditional variance fluctuates. There are many models of conditional heteroskedasticity, but most are designed for cross-sectional contexts, such as when the variance of a cross-sectional regression disturbance depends on one or more of the regressors.[3] However, heteroskedasticity is often present as well in the time series contexts relevant for forecasting, particularly in financial markets. The particular conditional variance function associated with the ARCH process,

$$\sigma_t^2 = \omega + \gamma(L)\varepsilon_t^2 ,$$

is tailor-made for time series environments, in which one often sees **volatility clustering,** such that large changes tend to be followed by large changes and small by small, *of either sign*. That is, one may see persistence, or serial correlation, in **volatility dynamics** (conditional variance dynamics), quite apart from persistence (or lack thereof) in conditional mean dynamics. The ARCH process approximates volatility dynamics in an autoregressive fashion—hence the name *autoregressive* conditional heteroskedasticity. To understand why, note that the ARCH conditional variance function links today's conditional variance positively to earlier lagged ε_t^2's, so that large ε_t^2's in the recent past produce a large conditional variance today, thereby increasing the likelihood of a large ε_t^2 today. Hence, ARCH processes are to conditional variance dynamics precisely as standard autoregressive processes are to conditional mean dynamics.

---

[2] In particular, σ_t^2 depends on the previous p values of ε_t via the distributed lag $\gamma(L)\varepsilon_t^2$.

[3] The variance of the disturbance in a model of household expenditure, for example, may depend on income.

The ARCH process may be viewed as a model for the disturbance in a broader model, as was the case when we introduced it earlier as a model for the innovation in a general linear process. Alternatively, if there are no conditional mean dynamics of interest, the ARCH process may be used for an observed series. It turns out that financial asset returns often have negligible conditional mean dynamics but strong conditional variance dynamics; hence, in much of what follows, we will view the ARCH process as a model for an observed series, which for convenience we will sometimes call a "return."

2. The GARCH Process

Thus far, we have used an ARCH(p) process to model conditional variance dynamics. We now introduce the **GARCH(p, q) process** (GARCH stands for generalized ARCH), which we shall subsequently use almost exclusively. As we shall see, GARCH is to ARCH (for conditional variance dynamics) as ARMA is to AR (for conditional mean dynamics).

The pure GARCH(p, q) process is given by[4]

$$
\begin{gathered}
y_t = \varepsilon_t \\
\varepsilon_t \mid \Omega_{t-1} \sim N(0, \sigma_t^2) \\
\sigma_t^2 = \omega + \alpha(L)\varepsilon_t^2 + \beta(L)\sigma_t^2 \\
\alpha(L) = \sum_{i=1}^{p} \alpha_i L^i, \quad \beta(L) = \sum_{i=1}^{q} \beta_i L^i \\
\omega > 0, \quad \alpha_i \geq 0, \quad \beta_i \geq 0, \quad \sum \alpha_i + \sum \beta_i < 1 .
\end{gathered}
$$

The stated conditions ensure that the conditional variance is positive and that y_t is covariance stationary.

Back substitution on σ_t^2 reveals that the GARCH(p, q) process can be represented as a restricted infinite-ordered ARCH process,

$$
\sigma_t^2 = \frac{\omega}{1 - \sum \beta_i} + \frac{\alpha(L)}{1 - \beta(L)}\varepsilon_t^2 = \frac{\omega}{1 - \sum \beta_i} + \sum_{i=1}^{\infty} \delta_i \varepsilon_{t-i}^2 ,
$$

which precisely parallels writing an ARMA process as a restricted infinite-ordered AR. Thus, the GARCH(p, q) process is a parsimonious approximation to what may truly be infinite-ordered ARCH volatility dynamics.

---

[4] By "pure" we mean that we have allowed only for conditional variance dynamics, by setting $y_t = \varepsilon_t$. We could of course also introduce conditional mean dynamics, but doing so would only clutter the discussion while adding nothing new.

It is important to note a number of special cases of the GARCH(p, q) process. First, of course, the ARCH(p) process emerges when $\beta(L) = 0$. Second, if *both* $\alpha(L)$ and $\beta(L)$ are 0, then the process is simply iid Gaussian noise with variance ω. Hence, although ARCH and GARCH processes may at first appear unfamiliar and potentially ad hoc, they are in fact much more general than standard iid white noise, which emerges as a potentially highly restrictive special case.

Here we highlight some important properties of GARCH processes. All of the discussion of course applies as well to ARCH processes, which are special cases of GARCH processes. First, consider the second-order moment structure of GARCH processes. The first two unconditional moments of the pure GARCH process are constant and given by $E(\varepsilon_t) = 0$ and

$$E((\varepsilon_t - E(\varepsilon_t))^2) = \frac{\omega}{1 - \sum \alpha_i - \sum \beta_i},$$

while the conditional moments are $E(\varepsilon_t \mid \Omega_{t-1}) = 0$ and of course

$$E((\varepsilon_t - E(\varepsilon_t \mid \Omega_{t-1}))^2 \mid \Omega_{t-1}) = \omega + \alpha(L)\varepsilon_t^2 + \beta(L)\sigma_t^2 .$$

In particular, the unconditional variance is fixed, as must be the case under covariance stationarity, while the conditional variance is time varying. It is no *surprise* that the conditional variance is time varying—the GARCH process was of course *designed* to allow for a time-varying conditional variance—but it is certainly worth emphasizing: The conditional variance is itself a serially correlated time series process.

Second, consider the unconditional higher-order (third and fourth) moment structure of GARCH processes. Real-world financial asset returns, which are often modeled as GARCH processes, are typically unconditionally symmetric but leptokurtic (i.e., more peaked in the center and with fatter tails than a normal distribution). It turns out that the implied unconditional distribution of the conditionally Gaussian GARCH process introduced earlier is also symmetric and leptokurtic. The unconditional leptokurtosis of GARCH processes follows from the persistence in conditional variance, which produces clusters of "low-volatility" and "high-volatility" episodes associated with observations in the center and in the tails of the unconditional distribution, respectively. Both the unconditional symmetry and unconditional leptokurtosis agree nicely with a variety of financial market data.

Third, consider the conditional prediction error variance of a GARCH process and its dependence on the conditioning information set. Because the conditional variance of a GARCH process is a serially correlated random variable, it is of interest to examine the optimal h-step-ahead prediction, prediction error, and conditional prediction error variance. Immediately, the h-step-ahead prediction is $E(\varepsilon_{t+h} \mid \Omega_t) = 0$, and the corresponding prediction error is

$$\varepsilon_{t+h} - E(\varepsilon_{t+h} \mid \Omega_t) = \varepsilon_{t+h} .$$

This implies that the conditional variance of the prediction error,

$$E((\varepsilon_{t+h} - E(\varepsilon_{t+h} \mid \Omega_t))^2 \mid \Omega_t) = E(\varepsilon_{t+h}^2 \mid \Omega_t),$$

depends on both h *and* Ω_t, because of the dynamics in the conditional variance. Simple calculations reveal that the expression for the GARCH(p, q) process is given by

$$E(\varepsilon_{t+h}^2 \mid \Omega_t) = \omega\left(\sum_{i=0}^{h-2}(\alpha(1)+\beta(1))^i\right) + (\alpha(1)+\beta(1))^{h-1}\sigma_{t+1}^2 .$$

In the limit, this conditional variance reduces to the unconditional variance of the process,

$$\lim_{h\to\infty} E(\varepsilon_{t+h}^2 \mid \Omega_t) = \frac{\omega}{1-\alpha(1)-\beta(1)} .$$

For finite h, the dependence of the prediction error variance on the current information set Ω_t can be exploited to improve interval and density forecasts.

Fourth, consider the relationship between ε_t^2 and σ_t^2. The relationship is important: GARCH dynamics in σ_t^2 turn out to introduce ARMA dynamics in ε_t^2.[5] More precisely, if ε_t is a GARCH(p, q) process, then ε_t^2 has the ARMA representation

$$\varepsilon_t^2 = \omega + (\alpha(L)+\beta(L))\varepsilon_t^2 - \beta(L)v_t + v_t ,$$

where $v_t = \varepsilon_t^2 - \sigma_t^2$ is the difference between the squared innovation and the conditional variance at time t. To see this, note that if ε_t is GARCH(p, q), then $\sigma_t^2 = \omega + \alpha(L)\varepsilon_t^2 + \beta(L)\sigma_t^2$. Adding and subtracting $\beta(L)\varepsilon_t^2$ from the right side gives

$$\begin{aligned}\sigma_t^2 &= \omega + \alpha(L)\varepsilon_t^2 + \beta(L)\varepsilon_t^2 - \beta(L)\varepsilon_t^2 + \beta(L)\sigma_t^2\\ &= \omega + (\alpha(L)+\beta(L))\varepsilon_t^2 - \beta(L)(\varepsilon_t^2 - \sigma_t^2) .\end{aligned}$$

Adding ε_t^2 to each side then gives

$$\sigma_t^2 + \varepsilon_t^2 = \omega + (\alpha(L)+\beta(L))\varepsilon_t^2 - \beta(L)(\varepsilon_t^2 - \sigma_t^2) + \varepsilon_t^2 ,$$

so that

$$\begin{aligned}\varepsilon_t^2 &= \omega + (\alpha(L)+\beta(L))\varepsilon_t^2 - \beta(L)(\varepsilon_t^2 - \sigma_t^2) + (\varepsilon_t^2 - \sigma_t^2) ,\\ &= \omega + (\alpha(L)+\beta(L))\varepsilon_t^2 - \beta(L)v_t + v_t .\end{aligned}$$

Thus, ε_t^2 is an ARMA(max(p, q), p) process with innovation v_t, where $v_t \in [-\sigma_t^2, \infty)$. ε_t^2 is covariance stationary if the roots of $\alpha(L)+\beta(L)=1$ are outside the unit circle.

Fifth, consider in greater depth the similarities and differences between σ_t^2 and ε_t^2. It is worth studying closely the key expression $v_t = \varepsilon_t^2 - \sigma_t^2$, which makes clear that ε_t^2 is effectively a "proxy" for σ_t^2, behaving similarly but not identically, with v_t being the difference, or error. In particular, ε_t^2 is a *noisy* proxy: ε_t^2 is an unbiased estimator of σ_t^2, but it is more volatile. It seems reasonable, then, that reconciling the noisy proxy ε_t^2 and the true underlying σ_t^2 should involve some sort of smoothing of ε_t^2. Indeed, in the GARCH(1, 1) case σ_t^2 is precisely obtained by exponentially smoothing ε_t^2. To see why, consider the exponential smoothing recursion, which gives the current smoothed value as a convex combination of the current unsmoothed value and the lagged smoothed value,

---

[5] Put differently, the GARCH process approximates conditional variance dynamics in the same way that an ARMA process approximates conditional mean dynamics.

$$\bar{\varepsilon}_t^2 = \gamma\varepsilon_t^2 + (1-\gamma)\bar{\varepsilon}_{t-1}^2 .$$

Back substitution yields an expression for the current smoothed value as an exponentially weighted moving average of past actual values:

$$\bar{\varepsilon}_t^2 = \sum w_j \varepsilon_{t-j}^2 ,$$

where

$$w_j = \gamma(1-\gamma)^j .$$

Now compare this result to the GARCH(1, 1) model, which gives the current volatility as a linear combination of lagged volatility and the lagged squared return,

$$\sigma_t^2 = \omega + \alpha\varepsilon_{t-1}^2 + \beta\sigma_{t-1}^2 .$$

Back substitution yields

$$\sigma_t^2 = \frac{\omega}{1-\beta} + \alpha \sum \beta^{j-1}\varepsilon_{t-j}^2 ,$$

so that the GARCH(1, 1) process gives current volatility as an exponentially weighted moving average of past squared returns.

Sixth, consider the temporal aggregation of GARCH processes. By temporal aggregation, we mean aggregation over time, as, for example, when we convert a series of daily returns to weekly returns, and then to monthly returns, then quarterly, and so on. It turns out that convergence toward normality under temporal aggregation is a feature of real-world financial asset returns. That is, although high-frequency (e.g., daily) returns tend to be fat tailed relative to the normal, the fat tails tend to get thinner under temporal aggregation, and normality is approached. Convergence to normality under temporal aggregation is also a property of covariance stationary GARCH processes. The key insight is that a low-frequency change is simply the sum of the corresponding high-frequency changes. For example, an annual change is the sum of the internal quarterly changes, each of which is the sum of its internal monthly changes, and so on. Thus, if a Gaussian central limit theorem can be invoked for sums of GARCH processes, convergence to normality under temporal aggregation is assured. Such theorems can be invoked if the process is covariance stationary.

In closing this section, it is worth noting that the symmetry and leptokurtosis of the unconditional distribution of the GARCH process, as well as the disappearance of the leptokurtosis under temporal aggregation, provide nice independent confirmation of the accuracy of GARCH approximations to asset return volatility dynamics, insofar as GARCH was certainly not invented with the intent of explaining those features of financial asset return data. On the contrary, the unconditional distributional results emerged as unanticipated byproducts of allowing for conditional variance dynamics, thereby providing a unified explanation of phenomena that were previously believed unrelated.

3. Extensions of ARCH and GARCH Models

There are numerous extensions of the basic GARCH model. In this section, we highlight several of the most important. One important class of extensions allows for **asymmetric response**; that is, it allows for last period's squared return to have different effects on today's volatility, depending on its sign.[6] Asymmetric response is often present, for example, in stock returns.

ASYMMETRIC RESPONSE

The simplest GARCH model allowing for asymmetric response is the **threshold GARCH,** or TGARCH, model.[7] We replace the standard GARCH conditional variance function,

$$\sigma_t^2 = \omega + \alpha\varepsilon_{t-1}^2 + \beta\sigma_{t-1}^2 ,$$

with

$$\sigma_t^2 = \omega + \alpha\varepsilon_{t-1}^2 + \gamma\varepsilon_{t-1}^2 D_{t-1} + \beta\sigma_{t-1}^2 ,$$

where

$$D_t = \begin{cases} 1, & \text{if } \varepsilon_t < 0 \\ 0 & \text{otherwise} . \end{cases}$$

The dummy variable D keeps track of whether the lagged return is positive or negative. When the lagged return is positive (good news yesterday), $D = 0$, so the effect of the lagged squared return on the current conditional variance is simply α. In contrast, when the lagged return is negative (bad news yesterday), $D = 1$, so the effect of the lagged squared return on the current conditional variance is $\alpha + \gamma$. If $\gamma = 0$, the response is symmetric, and we have a standard GARCH model; but if $\gamma \neq 0$, we have asymmetric response of volatility to news. Allowance for asymmetric response has proved useful for modeling "leverage effects" in stock returns, which occur when $\gamma < 0$.[8]

---

[6] In the GARCH model studied thus far, only the *square* of last period's return affects the current conditional variance; hence, its sign is irrelevant.

[7] For expositional convenience, we will introduce all GARCH extensions in the context of GARCH(1, 1), which is by far the most important case for practical applications. Extensions to the GARCH(p, q) case are immediate but notationally cumbersome.

[8] Negative shocks appear to contribute more to stock market volatility than do positive shocks. This is called the *leverage effect*, because a negative shock to the market value of equity increases the aggregate debt/equity ratio (other things the same), thereby increasing leverage.

Asymmetric response may also be introduced via the **exponential GARCH** (EGARCH) model,

$$\ln(\sigma_t^2) = \omega + \alpha \left|\frac{\varepsilon_{t-1}}{\sigma_{t-1}}\right| + \gamma \frac{\varepsilon_{t-1}}{\sigma_{t-1}} + \beta \ln(\sigma_{t-1}^2) .$$

Note that volatility is driven by both size and sign of shocks; hence, the model allows for an asymmetric response depending on the sign of news.[9] The log specification also ensures that the conditional variance is automatically positive, because σ_t^2 is obtained by exponentiating $\ln(\sigma_t^2)$—hence the name "exponential GARCH."

EXOGENOUS VARIABLES IN THE VOLATILITY FUNCTION

Just as ARMA models of conditional mean dynamics can be augmented to include the effects of exogenous variables, so too can GARCH models of conditional variance dynamics. We simply modify the standard GARCH volatility function in the obvious way, writing

$$\sigma_t^2 = \omega + \alpha\varepsilon_{t-1}^2 + \beta\sigma_{t-1}^2 + \gamma x_t ,$$

where γ is a parameter and x is a positive exogenous variable.[10] Allowance for exogenous variables in the conditional variance function is sometimes useful. Financial market volume, for example, often helps to explain market volatility.

REGRESSION WITH GARCH DISTURBANCES AND GARCH-M

Just as ARMA models may be viewed as models for disturbances in regressions, so too may GARCH models. We write

$$\begin{aligned} y_t &= \beta_0 + \beta_1 x_t + \varepsilon_t \\ \varepsilon_t \mid \Omega_{t-1} &\sim N(0, \sigma_t^2) \\ \sigma_t^2 &= \omega + \alpha\varepsilon_{t-1}^2 + \beta\sigma_{t-1}^2 . \end{aligned}$$

Consider now a regression model with GARCH disturbances of the usual sort, with one additional twist: The conditional variance enters as a regressor, thereby affecting the conditional mean. We write

$$\begin{aligned} y_t &= \beta_0 + \beta_1 x_t + \gamma\sigma_t^2 + \varepsilon_t \\ \varepsilon_t \mid \Omega_{t-1} &\sim N(0, \sigma_t^2) \\ \sigma_t^2 &= \omega + \alpha\varepsilon_{t-1}^2 + \beta\sigma_{t-1}^2 . \end{aligned}$$

---

[9] The absolute "size" of news is captured by $|r_{t-1}/\sigma_{t-1}|$, and the sign is captured by r_{t-1}/σ_{t-1}.

[10] Extension to allow multiple exogenous variables is straightforward.

This model, which is a special case of the general regression model with GARCH disturbances, is called **GARCH-in-mean** (GARCH-M). It is sometimes useful in modeling the relationship between risks and returns on financial assets when risk, as measured by the conditional variance, varies.[11]

COMPONENT GARCH

Note that the standard GARCH(1, 1) process may be written as

$$(\sigma_t^2 - \bar{\omega}) = \alpha(\varepsilon_{t-1}^2 - \bar{\omega}) + \beta(\sigma_{t-1}^2 - \bar{\omega}) ,$$

where $\bar{\omega} = \frac{\omega}{1-\alpha-\beta}$ is the unconditional variance.[12] This is precisely the GARCH(1, 1) model introduced earlier, rewritten it in a slightly different but equivalent form. In this model, short-run volatility dynamics are governed by the parameters α and β, and there are no long-run volatility dynamics, because $\bar{\omega}$ is constant.

Sometimes we might want to allow for both long-run and short-run, or persistent and transient, volatility dynamics in addition to the short-run volatility dynamics already incorporated. To do this, we replace $\bar{\omega}$ with a time-varying process, yielding

$$(\sigma_t^2 - q_t) = \alpha(\varepsilon_{t-1}^2 - q_{t-1}) + \beta(\sigma_{t-1}^2 - q_{t-1}) ,$$

where the time-varying long-run volatility, q_t, is given by

$$q_t = \omega + \rho(q_{t-1} - \omega) + \varphi(\varepsilon_{t-1}^2 - \sigma_{t-1}^2) .$$

This **"component GARCH"** model effectively lets us decompose volatility dynamics into long-run (persistent) and short-run (transitory) components, which sometimes yields useful insights. The persistent dynamics are governed by ρ, and the transitory dynamics are governed by α and β.[13]

MIXING AND MATCHING

In closing this section, we note that the different variations and extensions of the GARCH process may of course be mixed. Consider the following conditional variance function as an example:

$$(\sigma_t^2 - q_t) = \alpha(\varepsilon_{t-1}^2 - q_{t-1}) + \gamma(\varepsilon_{t-1}^2 - q_{t-1})D_{t-1} + \beta(\sigma_t^2 - q_{t-1}) + \theta x_t .$$

---

[11] One may also allow the conditional standard deviation, rather than the conditional variance, to enter the regression.

[12] $\bar{\omega}$ is sometimes called the "long-run" variance, referring to the fact that the unconditional variance is the long-run average of the conditional variance.

[13] It turns out, moreover, that under suitable conditions, the component GARCH model introduced here is covariance stationary and equivalent to a GARCH(2, 2) process subject to certain nonlinear restrictions on its parameters.

This is a component GARCH specification, generalized to allow for asymmetric response of volatility to news via the sign dummy *D*, as well as effects from the exogenous variable *x*.

4. Estimating, Forecasting, and Diagnosing GARCH Models

Recall that the likelihood function is the joint density function of the data, viewed as a function of the model parameters, and that maximum likelihood estimation finds the parameter values that maximize the likelihood function. This makes good sense: We choose those parameter values that maximize the likelihood of obtaining the data that were actually obtained. It turns out that construction and evaluation of the likelihood function is easily done for GARCH models, and maximum likelihood has emerged as the estimation method of choice.[14] No closed-form expression exists for the GARCH maximum likelihood estimator, so we must maximize the likelihood numerically.[15]

Construction of optimal forecasts of GARCH processes is simple. In fact, we derived the key formula earlier but did not comment extensively on it. Recall, in particular, that

$$\sigma^2_{t+h,t} = E\left(\varepsilon^2_{t+h} \mid \Omega_t\right) = \omega\left(\sum_{i=1}^{h-1}(\alpha(1)+\beta(1))^i\right) + (\alpha(1)+\beta(1))^{h-1}\sigma^2_{t+1}\,.$$

In words, the optimal *h*-step-ahead forecast is proportional to the optimal 1-step-ahead forecast. The optimal 1-step-ahead forecast, moreover, is easily calculated: All of the determinants of σ^2_{t+1} are lagged by at least one period, so that there is no problem of forecasting the right-hand side variables. In practice, of course, the underlying GARCH parameters α and β are unknown and so must be estimated, resulting in the feasible forecast $\hat{\sigma}^2_{t+h,t}$ formed in the obvious way.

In financial applications, volatility forecasts are often of direct interest, and the GARCH model delivers the optimal *h*-step-ahead point forecast, $\sigma^2_{t+h,t}$. Alternatively, and more generally, we might not be intrinsically interested in volatility; rather, we may simply want to use GARCH volatility forecasts to improve *h*-step-ahead interval or density forecasts of ε_t, which are crucially dependent on the *h*-step-ahead prediction error variance, $\sigma^2_{t+h,t}$. Consider, for example, the case of interval forecasting. In the case of constant volatility, we earlier worked with Gaussian 95% interval forecasts of the form

$$y_{t+h,t} \pm 1.96\,\sigma_h\,,$$

---

[14] The precise form of the likelihood is complicated, and we will not give an explicit expression here, but it may be found in various of the surveys mentioned in the Bibliographical and Computational Notes at the end of the chapter.

[15] Routines for maximizing the GARCH likelihood are available in a number of modern software packages such as EViews. As with any numerical optimization, care must be taken with startup values and convergence criteria to help ensure convergence to a global, as opposed to merely local, maximum.

where σ_h denotes the unconditional h-step-ahead standard deviation (which also equals the conditional h-step-ahead standard deviation in the absence of volatility dynamics). Now, however, in the presence of volatility dynamics we use

$$y_{t+h,t} \pm 1.96\,\sigma_{t+h,t}\,.$$

The ability of the conditional prediction interval to adapt to changes in volatility is natural and desirable: when volatility is low, the intervals are naturally tighter, and conversely. In the presence of volatility dynamics, the unconditional interval forecast is correct on average but likely incorrect at any given time, whereas the conditional interval forecast is correct at all times.

The issue arises as to how to detect GARCH effects in observed returns and, related, how to assess the adequacy of a fitted GARCH model. A key and simple device is the correlogram of squared returns, ε_t^2. As discussed earlier, ε_t^2 is a proxy for the latent conditional variance; if the conditional variance displays persistence, so, too, will ε_t^2.[16] One can of course also fit a GARCH model and assess significance of the GARCH coefficients in the usual way.

Note that we can write the GARCH process for returns as

$$\varepsilon_t = \sigma_t v_t\,,$$

where

$$v_t \overset{\text{iid}}{\sim} N(0,1)$$

$$\sigma_t^2 = \omega + \alpha\varepsilon_{t-1}^2 + \beta\sigma_{t-1}^2\,.$$

Equivalently, the *standardized* return, v, is iid,

$$\frac{\varepsilon_t}{\sigma_t} = v_t \overset{\text{iid}}{\sim} N(0,1)\,.$$

This observation suggests a way to evaluate the adequacy of a fitted GARCH model: Standardize returns by the conditional standard deviation from the fitted GARCH model, $\hat{\sigma}$, and then check for volatility dynamics missed by the fitted model by examining the correlogram of the squared *standardized* return, $(\varepsilon_t/\hat{\sigma}_t)^2$. This is routinely done in practice.

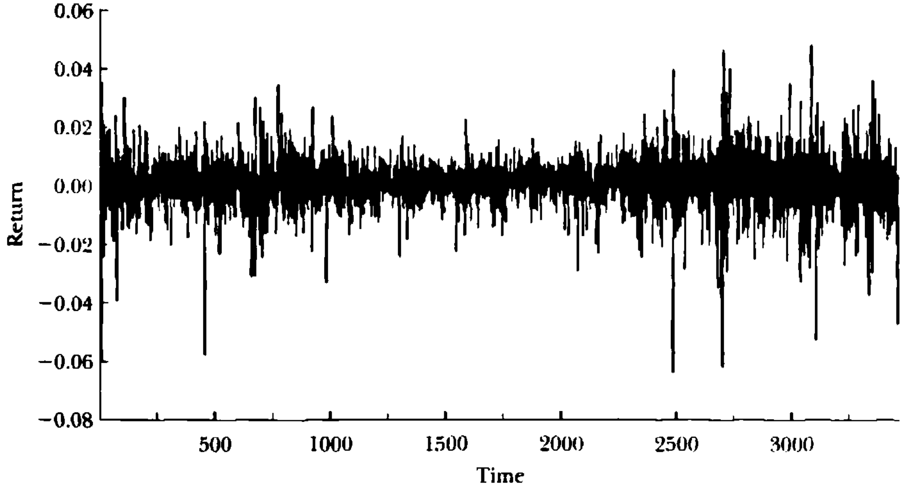
5. Application: Stock Market Volatility

We model and forecast the volatility of daily returns on the New York Stock Exchange (NYSE) from January 1, 1988, through December 31, 2001, excluding holidays, for a total of 3531 observations. We estimate using observations 1–3461, and then we forecast observations 3462–3531.

---

[16] Note well, however, that the converse is not true. That is, if ε_t^2 displays persistence, it does not necessarily follow that the conditional variance displays persistence. In particular, neglected serial correlation associated with conditional mean dynamics may cause serial correlation in ε_t and hence also in ε_t^2. Thus, before proceeding to examine and interpret the correlogram of ε_t^2 as a check for volatility dynamics, it is important that any conditional mean effects be appropriately modeled, in which case ε_t should be interpreted as the disturbance in an appropriate conditional mean model.

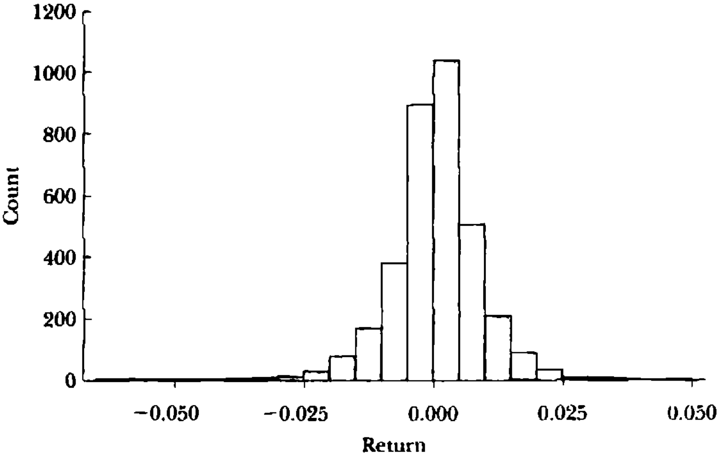
FIGURE 14.1
Time Series Plot, NYSE Returns



In Figure 14.1 we plot the daily returns, r_t. There is no visual evidence of serial correlation in the returns, but there *is* evidence of serial correlation in the *amplitude* of the returns. That is, volatility appears to cluster: Large changes tend to be followed by large changes and small by small, *of either sign.*

Figure 14.2 presents the histogram and related statistics for r_t. The mean daily return is slightly positive. Moreover, the returns are approximately symmetric (only slightly left skewed) but highly leptokurtic. The Jarque-Bera statistic indicates decisive rejection of normality.

FIGURE 14.2 *Histogram and Related Diagnostic Statistics, NYSE Returns*



```
Series: R
Sample 13461
Observations 3461

Mean            0.000522
Median          0.000640
Maximum         0.047840
Minimum        −0.063910
Std. Dev.       0.008541
Skewness       −0.505540
Kurtosis        8.535016

Jarque-Bera     4565.446
Probability     0.000000
```

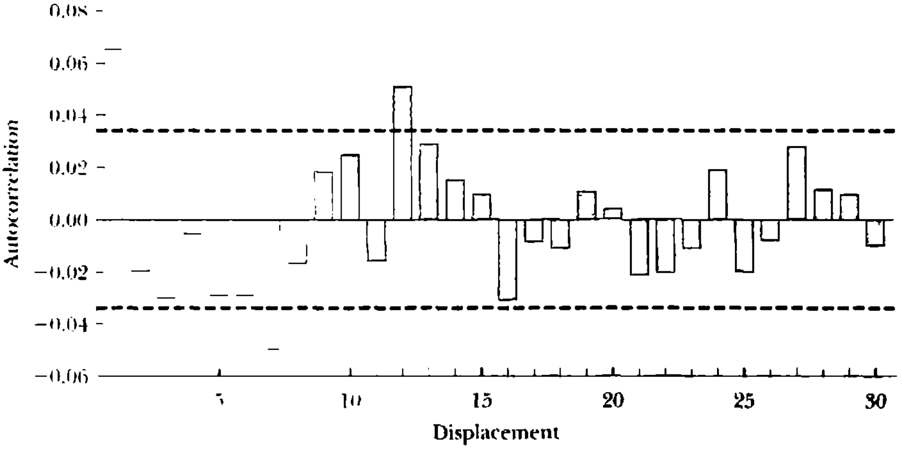


FIGURE 14.3
Correlogram, NYSE Returns

The correlogram for r_t appears in Figure 14.3. The sample autocorrelations are tiny and usually insignificant relative to the Bartlett standard errors, yet the autocorrelation function shows some evidence of a systematic cyclical pattern, and the Q-statistics (not shown), which cumulate the information across all displacements, reject the null of weak white noise. Despite the weak serial correlation evidently present in the returns, we will proceed for now as if returns were weak white noise, which is approximately, if not exactly, the case.[17]

In Figure 14.4 we plot r_t^2. The volatility clustering is even more evident than it was in the time series plot of returns. Perhaps the strongest evidence of

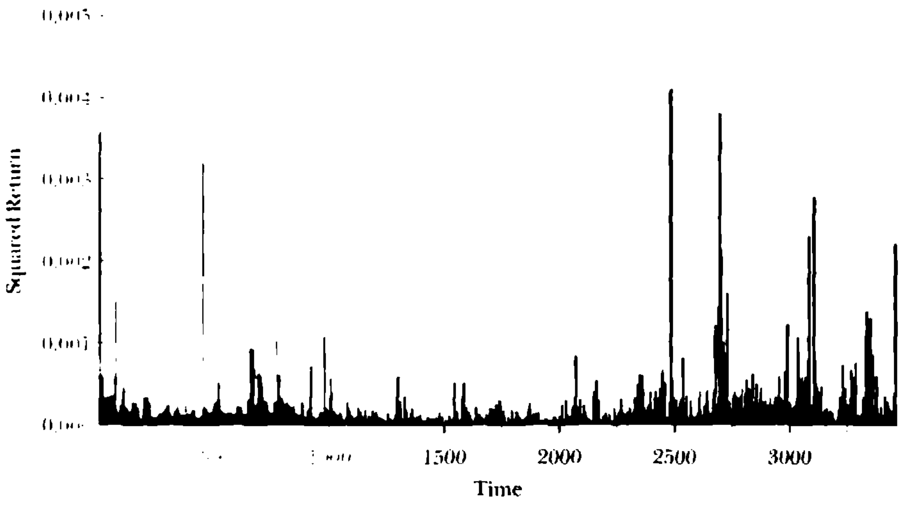
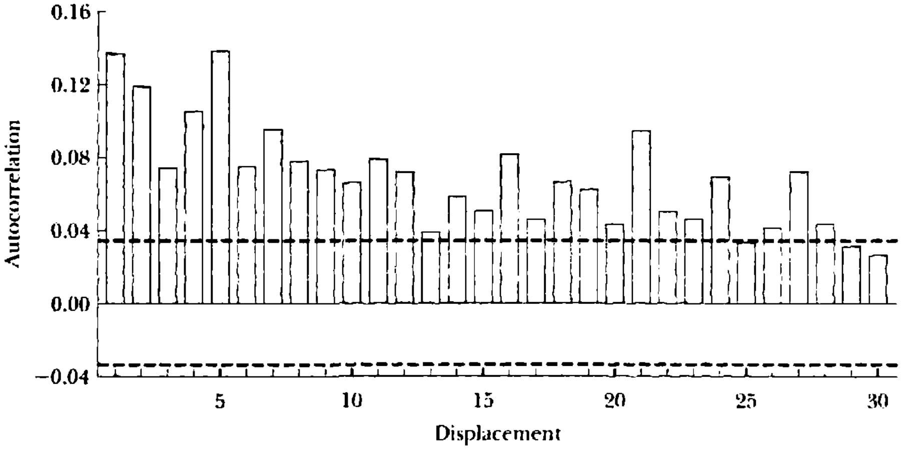


FIGURE 14.4
Time Series Plot, Squared NYSE Returns

---

[17] In the Exercises, Problems, and Complements at the end of this chapter, we model the conditional mean, as well as the conditional variance, of returns.

FIGURE 14.5
Correlogram, Squared NYSE Returns



all comes from the correlogram of r_t^2, which we show in Figure 14.5: All sample autocorrelations of r_t^2 are positive, overwhelmingly larger than those of the returns themselves, and statistically significant.

As a crude first pass at modeling the stock market volatility, we fit an AR(5) model directly to r_t^2; the results appear in Table 14.1. It is interesting to note

TABLE 14.1
AR(5) Model, Squared NYSE Returns

Dependent Variable: *R*2
Method: Least squares

Sample(adjusted): 6 3461
Included observations: 3456 after adjusting endpoints

Variable	**Coefficient**	**Std. Error**	***t*-Statistic**	**Prob.**
C	4.40E-05	3.78E-06	11.62473	0.0000
*R*2(−1)	0.107900	0.016137	6.686547	0.0000
*R*2(−2)	0.091840	0.016186	5.674167	0.0000
*R*2(−3)	0.028981	0.016250	1.783389	0.0746
*R*2(−4)	0.039312	0.016481	2.385241	0.0171
*R*2(−5)	0.116436	0.016338	7.126828	0.0000

R^2	0.052268	Mean dependent var.	7.19E-05
Adjusted R^2	0.050894	SD dependent var.	0.000189
SE of regression	0.000184	Akaike info criterion	−14.36434
Sum squared resid.	0.000116	Schwarz criterion	−14.35366
Log likelihood	24827.58	*F*-statistic	38.05372
Durbin-Watson stat.	1.975672	Prob(*F*-statistic)	0.000000

TABLE 14.2
ARCH(5) Model, NYSE Returns

Dependent variable: *R*
Method: ML-ARCH (Marquardt)

Sample: 1 3461
Included observations: 3461
Convergence achieved after 13 iterations
Variance backcast: ON

Coefficient	**Std. Error**	**z-Statistic**	**Prob.**	
C	0.000689	0.000127	5.437097	0.0000
Variance Equation				
C	3.16E-05	1.08E-06	29.28536	0.0000
ARCH(1)	0.128948	0.013847	9.312344	0.0000
ARCH(2)	0.166852	0.015055	11.08281	0.0000
ARCH(3)	0.072551	0.014345	5.057526	0.0000
ARCH(4)	0.143778	0.015363	9.358870	0.0000
ARCH(5)	0.089254	0.018480	4.829789	0.0000

R^2	−0.000381	Mean dependent var.	0.000522
Adjusted R^2	−0.002118	SD dependent var.	0.008541
SE of regression	0.008550	Akaike info criterion	−6.821461
Sum squared resid.	0.252519	Schwarz criterion	−6.809024
Log likelihood	11811.54	Durbin-Watson stat.	1.861036

that the *t*-statistics on the lagged squared returns are often significant, even at long lags, yet the R^2 of the regression is low, reflecting the fact that r_t^2 is a very noisy volatility proxy.

As a more sophisticated second pass at modeling NYSE volatility, we fit an ARCH(5) model to r_t: the results appear in Table 14.2. The lagged squared returns appear significant even at long lags. The correlogram of squared standardized residuals shown in Figure 14.6, however, displays some remaining systematic behavior, indicating that the ARCH(5) model fails to capture all of the volatility dynamics, potentially because even longer lags are needed.[18]

Table 14.3 shows the results of fitting a GARCH(1, 1) model. All of the parameter estimates are highly statistically significant, and the "ARCH coefficient" (α) and "GARCH coefficient" (β) sum to a value near unity (0.987), with β substantially larger than α, as is commonly found for financial asset returns. We show the correlogram of squared standardized GARCH(1, 1) residuals in

---

[18] In the Exercises, Problems, and Complements at the end of this chapter, we also examine ARCH(p) models with $p > 5$.

FIGURE 14.6
Correlogram, Squared Standardized ARCH(5) Residuals, NYSE Returns

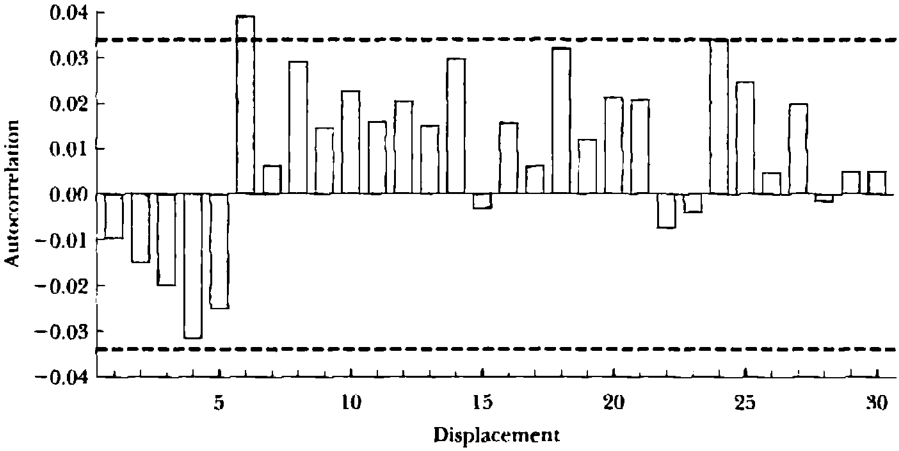


Figure 14.7. All sample autocorrelations are tiny and inside the Bartlett bands, and they display noticeably less evidence of any systematic pattern than for the squared standardized ARCH(5) residuals.

In Figure 14.8 we show the time series of estimated conditional standard deviations implied by the estimated GARCH(1, 1) model. Clearly, volatility

TABLE 14.3
GARCH(1, 1) Model, NYSE Returns

Dependent variable: *R*
Method: ML-ARCH (Marquardt)

Sample: 1 3461
Included observations: 3461
Convergence achieved after 19 iterations
Variance backcast: ON

Coefficient	**Std. Error**	**z-Statistic**	**Prob.**	
C	0.000640	0.000127	5.036942	0.0000
Variance Equation				
C	1.06E-06	1.49E-07	7.136840	0.0000
ARCH(1)	0.067410	0.004955	13.60315	0.0000
GARCH(1)	0.919714	0.006122	150.2195	0.0000

R^2	−0.000191	Mean dependent var.	0.000522
Adjusted R^2	−0.001059	SD dependent var.	0.008541
SE of regression	0.008546	Akaike info criterion	−6.868008
Sum squared resid.	0.252471	Schwarz criterion	−6.860901
Log likelihood	11889.09	Durbin-Watson stat.	1.861389

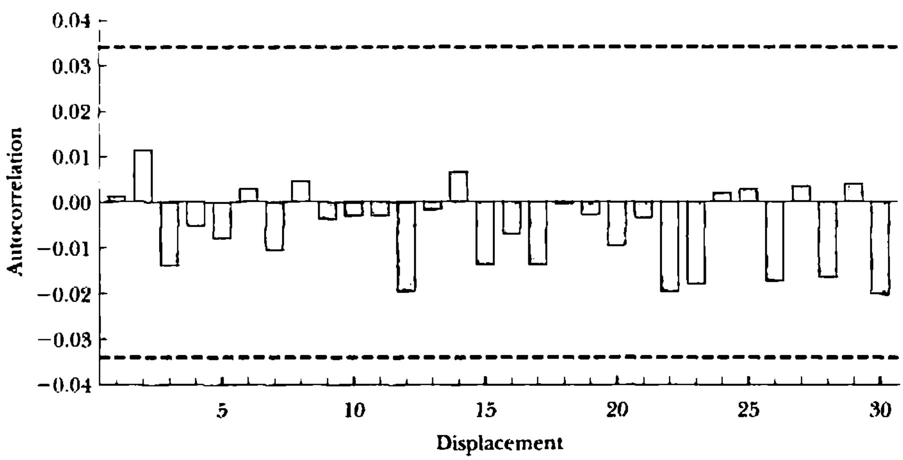


FIGURE 14.7 *Correlogram, Squared Standardized GARCH(1, 1) Residuals, NYSE Returns*

fluctuates a great deal and is highly persistent. For comparison we show in Figure 14.9 the series of exponentially smoothed r_t^2, computed using a standard smoothing parameter of 0.05.[19] Clearly the GARCH and exponential smoothing volatility estimates behave similarly, although not at all identically. The difference reflects the fact that the GARCH smoothing parameter is

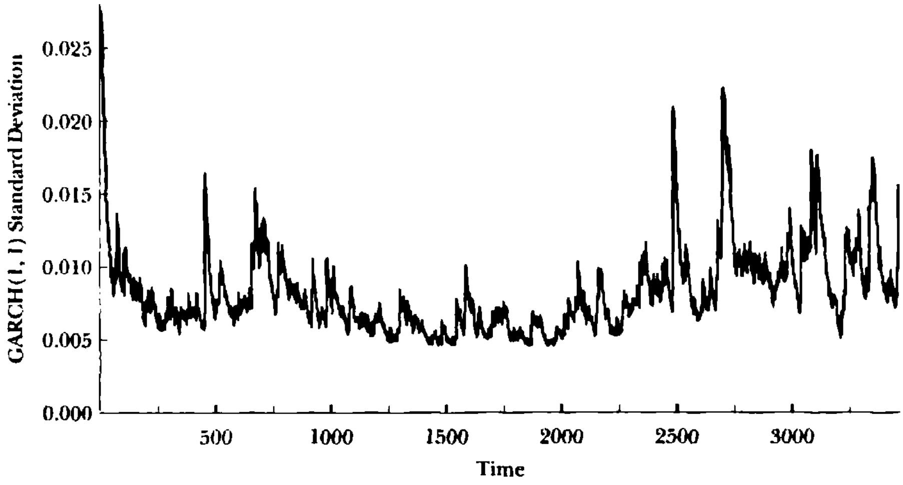
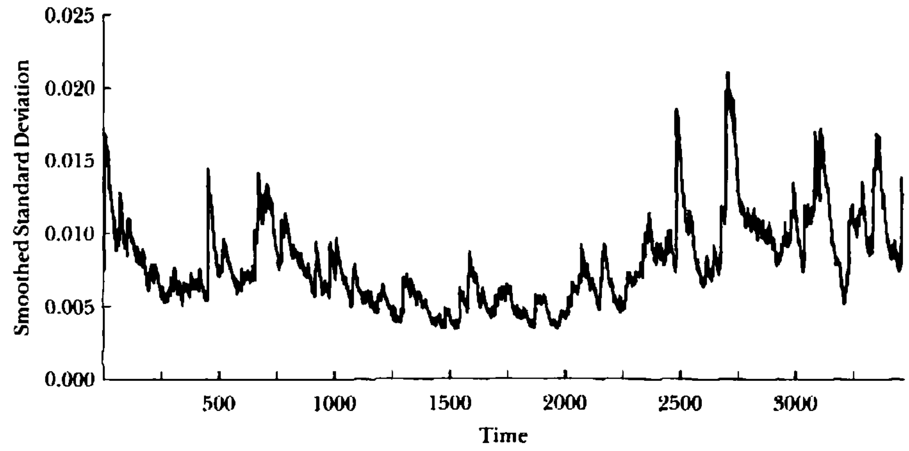


FIGURE 14.8 *Estimated Conditional Standard Deviation, GARCH(1, 1) Model, NYSE Returns*

---

[19] For comparability with the earlier-computed GARCH estimated conditional standard deviation, we actually show the square root of exponentially smoothed r_t^2.

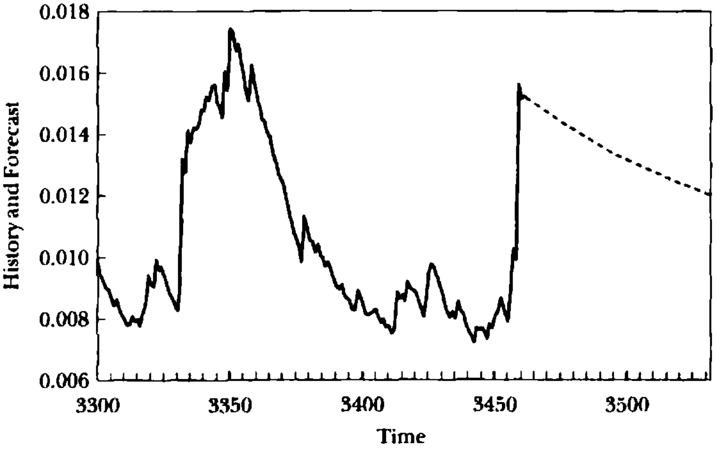
FIGURE 14.8
Estimated Conditional Standard Deviation, Exponential Smoothing, NYSE Returns



effectively estimated by the method of maximum likelihood, whereas the exponential smoothing parameter is set rather arbitrarily.

Now, using the model estimated using observations 1–3461, we generate a forecast of the conditional standard deviation for the out-of-sample observations 3462–3531. We show the results in Figure 14.10. The forecast period begins just following a volatility burst, so it is not surprising that the forecast calls for gradual volatility reduction. For greater understanding, Figure 14.11 presents both a longer history and a longer forecast. Clearly the forecast conditional standard deviation is reverting exponentially to the unconditional standard deviation (0.009), per the formula discussed earlier.

FIGURE 14.10
Conditional Standard Deviation, History and Forecast, GARCH(1, 1) Model



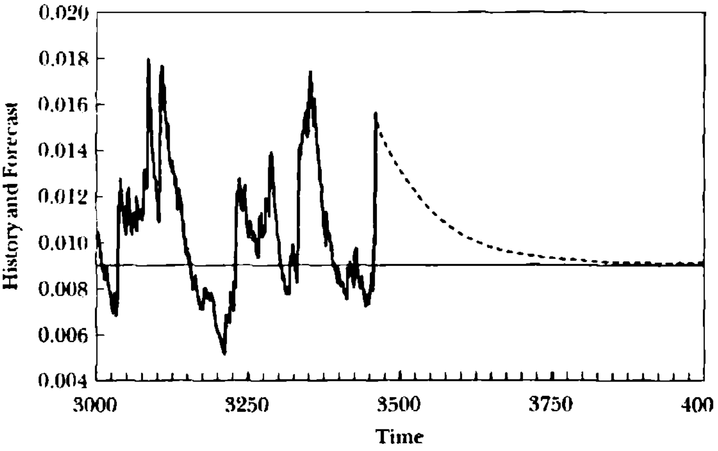


FIGURE 14.11 *Conditional Standard Deviation, Extended History and Extended Forecast, GARCH(1, 1) Model*

Exercises, Problems, and Complements

1. (Removing conditional mean dynamics before modeling volatility dynamics) In the application in the text, we noted that NYSE stock returns appeared to have some weak conditional mean dynamics, yet we ignored them and proceeded directly to model volatility.
   a. Instead, first fit autoregressive models using the SIC to guide order selection, and then fit GARCH models to the residuals. Redo the entire empirical analysis reported in the text in this way, and discuss any important differences in the results.
   b. Consider instead the simultaneous estimation of all parameters of AR(p)-GARCH models. That is, estimate regression models where the regressors are lagged dependent variables and the disturbances display GARCH. Redo the entire empirical analysis reported in the text in this way, and discuss any important differences in the results relative to those in the text and those obtained in part *a*.

2. (Variations on the basic ARCH and GARCH models). Using the stock return data, consider richer models than the pure ARCH and GARCH models discussed in the text.
   a. Estimate, diagnose, and discuss a threshold GARCH(1, 1) model.
   b. Estimate, diagnose, and discuss an EGARCH(1, 1) model.
   c. Estimate, diagnose, and discuss a component GARCH(1, 1) model.
   d. Estimate, diagnose, and discuss a GARCH-M model.

3. (Empirical performance of pure ARCH models as approximations to volatility dynamics) Here we will fit pure ARCH(p) models to the stock return data, including values of p larger than $p = 5$ as done in the text, and contrast the results with those from fitting GARCH(p, q) models.
   a. When fitting pure ARCH(p) models, what value of p seems adequate?

b. When fitting GARCH(p, q) models, what values of p and q seem adequate?
c. Which approach appears more parsimonious?

4. (Direct modeling of volatility proxies) In the text, we fit an AR(5) directly to a subset of the squared NYSE stock returns. In this exercise, use the *entire* NYSE dataset.
   a. Construct, display, and discuss the fitted volatility series from the AR(5) model.
   b. Construct, display, and discuss an alternative fitted volatility series obtained by exponential smoothing, using a smoothing parameter of 0.10, corresponding to a large amount of smoothing, but less than that done in the text.
   c. Construct, display, and discuss the volatility series obtained by fitting an appropriate GARCH model.
   d. Contrast the results of parts *a*, *b*, and *c*.
   e. Why is fitting of a GARCH model preferable in principle to the AR(5) or exponential smoothing approaches?

5. (GARCH volatility forecasting) You work for Xanadu, a luxury resort in the tropics. The daily temperature in the region is beautiful year-round, with a mean around 76 (Fahrenheit!) and no conditional mean dynamics. Occasional pressure systems, however, can cause bursts of temperature volatility. Such volatility bursts generally don't last long enough to drive away guests, but the resort still loses revenue from fees on activities that are less popular when the weather isn't perfect. In the middle of such a period of high temperature volatility, your boss gets worried and asks you make a forecast of volatility over the next 10 days. After some experimentation, you find that daily temperature y_t follows

$$y_t \mid \Omega_{t-1} \sim N\left(\mu, \sigma_t^2\right),$$

where σ_t^2 follows a GARCH(1, 1) process,

$$\sigma_t^2 = \omega + \alpha\varepsilon_{t-1}^2 + \beta\sigma_{t-1}^2 .$$

   a. Estimation of your model using historical daily temperature data yields $\hat{\mu} = 76$, $\hat{\omega} = 3$, $\hat{\alpha} = 0.6$, and $\hat{\beta} = 0$. If yesterday's temperature was 92 degrees, generate point forecasts for each of the next 10 days' conditional variance.
   b. According to your volatility forecasts, how many days will it take until volatility drops enough such that there is at least a 90% probability that the temperature will be within 4 degrees of 76?
   c. Your boss is impressed by your knowledge of forecasting and asks you whether your model can predict the next spell of bad weather. How would you answer him?

6. (Assessing volatility dynamics in observed returns and in standardized returns) In the text, we sketched the use of correlograms of squared observed returns for the detection of GARCH and squared standardized returns for diagnosing the adequacy of a fitted GARCH model. Examination of Ljung-Box statistics is an important part of a correlogram analysis. McLeod and Li (1983) show that the Ljung-Box statistics may be legitimately used on squared observed returns, in which case it will have the usual χ_m^2 distribution under the null hypothesis of independence. Bollerslev and Mikkelson (1996) argue that one may also use the Ljung-Box statistic on the squared standardized returns but that a better distributional approximation is obtained in that case by using a χ_{m-k}^2 distribution,

where k is the number of estimated GARCH parameters, to account for degrees of freedom used in model fitting.

7. (Allowing for leptokurtic conditional densities) Thus far, we have worked exclusively with conditionally Gaussian GARCH models, which correspond to

$$\varepsilon_t = \sigma_t v_t$$
$$v_t \overset{\text{iid}}{\sim} N(0, 1)$$

or, equivalently, to normality of the standardized return, ε_t/σ_t.

a. The conditional normality assumption may sometimes be violated. However, Bollerslev and Wooldridge (1992) show that GARCH parameters are consistently estimated by Gaussian maximum likelihood even when the normality assumption is incorrect. Sketch some intuition for this result.

b. Fit an appropriate conditionally Gaussian GARCH model to the stock return data. How might you use the histogram of the standardized returns to assess the validity of the conditional normality assumption? Do so and discuss your results.

c. Sometimes the conditionally Gaussian GARCH model does indeed fail to explain all of the leptokurtosis in returns; that is, especially with very high-frequency data, we sometimes find that the conditional density is leptokurtic. Fortunately, leptokurtic conditional densities are easily incorporated into the GARCH model. For example, in Bollerslev's (1987) conditionally **Student's *t* GARCH** model, the conditional density is assumed to be Student's *t*, with the degrees-of-freedom d treated as another parameter to be estimated. More precisely, we write

$$\varepsilon_t = \sigma_t v_t$$
$$v_t \overset{\text{iid}}{\sim} \frac{t_d}{\text{std}(t_d)}.$$

What is the reason for dividing the Student's t variable, t_d, by its standard deviation, $\text{std}(t_d)$? How might such a model be estimated?

8. (Optimal prediction under asymmetric loss) In the text, we stressed GARCH modeling for improved interval and density forecasting, implicitly working under a symmetric loss function. Less obvious but equally true is the fact that, under *asymmetric* loss, volatility dynamics can be exploited to produce improved *point* forecasts, as shown by Christoffersen and Diebold (1996, 1997). The optimal predictor under asymmetric loss is not the conditional mean but rather the conditional mean shifted by a time-varying adjustment that depends on the conditional variance. The intuition for the bias in the optimal predictor is simple: When errors of one sign are more costly than errors of the other sign, it is desirable to bias the forecasts in such a way as to reduce the chance of making an error of the more damaging type. The optimal amount of bias depends on the conditional prediction error variance of the process because, as the conditional variance grows, so, too, does the optimal amount of bias needed to avoid large prediction errors of the more damaging type.

9. (**Multivariate GARCH** models) In the multivariate case, such as when modeling a *set* of returns rather than a single return, we need to model not only conditional variances but also conditional *covariances*.

a. Is the GARCH conditional variance specification introduced earlier, say for the *i*th return,

$$\sigma^2_{it} = \omega + \alpha\varepsilon^2_{i,t-1} + \beta\sigma^2_{i,t-1} \,,$$

still appealing in the multivariate case? Why or why not?

b. Consider the following specification for the conditional covariance between *i*th and *j*th returns:

$$\sigma_{ij,t} = \omega + \alpha\varepsilon_{i,t-1}\varepsilon_{j,t-1} + \beta\sigma_{ij,t-1} \,.$$

Is it appealing? Why or why not?

c. Consider a fully general multivariate volatility model, in which every conditional variance and covariance may depend on lags of every conditional variance and covariance, as well as lags of every squared return and cross product of returns. What are the strengths and weaknesses of such a model? Would it be useful for modeling, say, a set of 500 returns? If not, how might you proceed?

Bibliographical and Computational Notes

This chapter draws on the survey by Diebold and Lopez (1995), which may be consulted for additional details. Other broad surveys include Bollerslev, Chou, and Kroner (1992); Bollerslev, Engle, and Nelson (1994); Taylor (2005); and Andersen et al. (2007).

Engle (1982) is the original development of the ARCH model. Bollerslev (1986) provides the important GARCH extension, and Engle (1995) contains many others. Diebold (1988) shows convergence to normality under temporal aggregation.

TGARCH traces to Glosten, Jagannathan, and Runkle (1993) and EGARCH to Nelson (1991). Engle, Lilien, and Robins (1987) introduce the GARCH-M model, and Engle and Lee (1999) introduce component GARCH.

Recently, methods of volatility measurement, modeling, and forecasting have been developed that exploit the increasing availability of high-frequency financial asset return data. For a fine overview, see Dacorogna et al. (2001); for more recent developments, see Andersen, Bollerslev, Diebold, and Labys (2003) and Andersen, Bollerslev, and Diebold (2006). For insights into the emerging field of financial econometrics, see Diebold (2001) and many of the other essays in the same collection.

Concepts for Review

Heteroskedasticity
Time-varying volatility
Financial econometrics
ARCH(*p*) process
Volatility clustering
Volatility dynamics
GARCH(*p*, *q*) process
Asymmetric response
Threshold GARCH
Exponential GARCH
GARCH-in-mean
Component GARCH
Student's *t* GARCH
Multivariate GARCH

References and Additional Readings

Andersen, T. G., Bollerslev, T., Christoffersen, P. F., and Diebold, F. X. (2007). *Volatility: Practical Methods for Financial Applications.* Princeton, N.J.: Princeton University Press.

Andersen, T. G., Bollerslev, T., and Diebold, F. X. (2006). "Parametric and Nonparametric Volatility Measurement." In L. P. Hansen and Y. Ait-Sahalia (eds.), *Handbook of Financial Econometrics.* Amsterdam: North-Holland.

Andersen, T. G., Bollerslev, T., Diebold, F. X., and Labys, P. (2003). "Modeling and Forecasting Realized Volatility." *Econometrica,* 71, 579–626.

Bollerslev, T. (1986). "Generalized Autoregressive Conditional Heteroskedasticity." *Journal of Econometrics,* 31, 307–327.

Bollerslev, T. (1987). "A Conditional Heteroskedastic Time Series Model for Speculative Prices and Rates of Return." *Review of Economics and Statistics,* 69, 542–547.

Bollerslev, T., Chou, R. Y., and Kroner, K. F. (1992). "ARCH Modeling in Finance: A Selective Review of the Theory and Empirical Evidence." *Journal of Econometrics,* 52, 5–59.

Bollerslev, T., Engle, R. F., and Nelson, D. B. (1994). "ARCH Models." In R. F. Engle and D. McFadden (eds.), *Handbook of Econometrics, Vol. IV.* Amsterdam: North-Holland.

Bollerslev, T., and Mikkelsen, H. O. (1996). "Modeling and Pricing Long Memory in Stock Market Volatility." *Journal of Econometrics,* 73, 151–184.

Bollerslev, T., and Wooldridge, J. M. (1992). "Quasi-Maximum Likelihood Estimation and Inference in Dynamic Models with Time-Varying Covariances." *Econometric Reviews,* 11, 143–179.

Christoffersen, P. F., and Diebold, F. X. (1996). "Further Results on Forecasting and Model Selection under Asymmetric Loss." *Journal of Applied Econometrics,* 11, 561–572.

Christoffersen, P. F., and Diebold, F. X. (1997). "Optimal Prediction under Asymmetric Loss." *Econometric Theory,* 13, 808–817.

Dacorogna, M. M., et al. (2001). *An Introduction to High-Frequency Finance.* New York: Academic Press.

Diebold, F. X. (1988). *Empirical Modeling of Exchange Rate Dynamics.* New York: Springer.

Diebold, F. X. (2001). "Econometrics: Retrospect and Prospect." *Journal of Econometrics,* 100, 73–75.

Diebold, F. X., and Lopez, J. (1995). "Modeling Volatility Dynamics." In K. Hoover (ed.), *Macroeconometrics: Developments, Tensions and Prospects.* Boston: Kluwer Academic Press, 427–472.

Diebold, F. X., and Nerlove, M. (1989). "The Dynamics of Exchange Rate Volatility: A Multivariate Latent-Factor ARCH Model." *Journal of Applied Econometrics,* 4, 1–22.

Engle, R. F. (1982). "Autoregressive Conditional Heteroskedasticity with Estimates of the Variance of U.K. Inflation." *Econometrica,* 50, 987–1008.

Engle, R. F., ed. (1995). *ARCH: Selected Readings.* Oxford: Oxford University Press.

Engle, R. F., and Lee, G. (1999). "A Permanent and Transitory Model of Stock Return Volatility." In R. Engle and H. White (eds.), *Cointegration, Causality, and Forecasting: A Festschrift in Honor of Clive W. J. Granger.* Oxford: Oxford University Press.

Engle, R. F., Lilien, D. M., and Robins, R. P. (1987). "Estimating Time Varying Risk Premia in the Term Structure: The ARCH-M Model." *Econometrica,* 55, 391–407.

Glosten, L. R., Jagannathan, R., and Runkle, D. E. (1993). "On the Relation between the Expected Value and the Volatility of the Nominal Excess Return on Stocks." *Journal of Finance,* 48, 1779–1801.

McLeod, A. I., and Li, W. K. (1983). "Diagnostic Checking of ARMA Time Series Models Using Squared Residual Autocorrelations." *Journal of Time Series Analysis,* 4, 269–273.

Nelson, D. B. (1991). "Conditional Heteroskedasticity in Asset Returns: A New Approach." *Econometrica,* 59, 347–370.

Taylor, S. (2005). *Asset Price Dynamics, Volatility and Prediction.* Princeton, N.J.: Princeton University Press.

Akaike, H. (1974). "A New Look at the Statistical Model Identification." *IEEE Transactions on Automatic Control*, AC-19, 716–723.

Andersen, T. G., Bollerslev, T., Christoffersen, P. F., and Diebold, F. X. (2007). *Volatility: Practical Methods for Financial Applications*. Princeton, N.J.: Princeton University Press.

Andersen, T. G., Bollerslev, T., and Diebold, F. X. (2004). "Parametric and Nonparametric Volatility Measurement." In L. P. Hansen and Y. Ait-Sahalia (eds.), *Handbook of Financial Econometrics*. Amsterdam: North-Holland.

Andersen, T. G., Bollerslev, T., Diebold, F. X., and Labys, P. (2003). "Modeling and Forecasting Realized Volatility." *Econometrica*, 71, 579–626.

Anderson, D. R., Sweeney, D. J., and Williams, T. A. (2006). *Statistics for Business and Economics*. 4th ed. Cincinnati: South-Western.

Andrews, D. W. K. (1993). "Exactly Median-Unbiased Estimation of First Order Autoregressive/Unit Root Models." *Econometrica*, 61, 139–65.

Anscombe, F. J. (1973). "Graphs in Statistical Analysis." *American Statistician*, 27, 17–21.

Armstrong, J. S. (1978). *Long Run Forecasting: From Crystal Ball to Computer*. New York: Wiley.

Armstrong, J. S., ed. (2001). *The Principles of Forecasting*. Norwell, Mass.: Kluwer Academic Forecasting.

Armstrong, J. S., and Fildes, R. (1995). "On the Selection of Error Measures for Comparisons among Forecasting Methods." *Journal of Forecasting*, 14, 67–71.

Auerbach, A. J. (1982). "The Index of Leading Indicators: 'Measurement without Theory' Thirty-Five Years Later." *Review of Economics and Statistics*, 64, 589–595.

Bails, D. G., and Peppers, L. C. (1997). *Business Fluctuations*. 2nd ed. Englewood Cliffs, N.J.: Prentice Hall.

Bartlett, M. (1946). "On the Theoretical Specification of Sampling Properties of Autocorrelated Time Series." *Journal of the Royal Statistical Society* B, 8, 27–41.

Bates, J. M., and Granger, C. W. J. (1969). "The Combination of Forecasts." *Operations Research Quarterly*, 20, 451–468.

Baumohl, B. (2005). *Secrets of Economic Indicators: The Hidden Clues to Future Economic Trends and Investment Opportunities*. Philadelphia: Wharton School Publishing.

Beveridge, S., and Nelson, C. R. (1981). "A New Approach to the Decomposition of Economic Time Series into Permanent and Transient Components with Particular Attention to Measurement of the Business Cycle." *Journal of Monetary Economics*, 7, 151–174.

Bhansali, R. J. (2002). "Multi-Step Forecasting." In M. P. Clements and D. H. Hendry (eds.), *A Companion to Economic Forecasting*. Oxford: Blackwell, 206–221.

Bollerslev, T. (1986). "Generalized Autoregressive Conditional Heteroskedasticity." *Journal of Econometrics*, 31, 307–327.

Bollerslev, T. (1987). "A Conditional Heteroskedastic Time Series Model for Speculative Prices and Rates of Return." *Review of Economics and Statistics*, 69, 542–547.

Bollerslev, T., Chou, R. Y., and Kroner, K. F. (1992). "ARCH Modeling in Finance: A Selective Review of the Theory and Empirical Evidence." *Journal of Econometrics*, 52, 5–59.

Bollerslev, T., Engle, R. F., and Nelson, D. B. (1994). "ARCH Models." In R. F. Engle and D. McFadden (eds.), *Handbook of Econometrics, Vol. IV.* Amsterdam: North-Holland.

Bollerslev, T., and Mikkelsen, H. O. (1996). "Modeling and Pricing Long Memory in Stock Market Volatility." *Journal of Econometrics,* 73, 151–184.

Bollerslev, T., and Wooldridge, J. M. (1992). "Quasi-Maximum Likelihood Estimation and Inference in Dynamic Models with Time-Varying Covariances." *Econometric Reviews,* 11, 143–179.

Box, G. E. P., Jenkins, G. W., and Reinsel, G. (1994). *Time Series Analysis, Forecasting and Control.* 3rd ed. Englewood Cliffs, N.J.: Prentice Hall.

Box, G. E. P., and Pierce, D. A. (1970). "Distribution of Residual Autocorrelations in ARIMA Time-Series Models." *Journal of the American Statistical Association,* 65, 1509–1526.

Breidt, F. J., Davis, R. A., and Dunsmuir, W. T. M. (1995). "Improved Bootstrap Prediction Intervals for Autoregressions." *Journal of Time Series Analysis,* 16, 177–200.

Brown, R. L., Durbin, J., and Evans, J. M. (1975). "Techniques for Testing the Constance of Regression Relationships over Time." *Journal of the Royal Statistical Society* B, 37, 149–163.

Burns, A. F., and Mitchell, W. C. (1946). *Measuring Business Cycles.* New York: National Bureau of Economic Research.

Campbell, J. Y., and Perron, P. (1991). "Pitfalls and Opportunities: What Macroeconomists Should Know about Unit Roots." In O. J. Blanchard and S. S. Fischer (eds.), *NBER Macroeconomics Annual, 1991.* Cambridge, Mass.: MIT Press.

Chatfield, C. (1993). "Calculating Interval Forecasts (with Discussion)." *Journal of Business and Economic Statistics,* 11, 121–144.

Chatfield, C. (1995). "Model Uncertainty, Data Mining and Statistical Inference (with Discussion)." *Journal of the Royal Statistical Society* A, 158, part 3, 419–466.

Chatfield, C. (1996). *The Analysis of Time Series: An Introduction.* 5th ed. London: Chapman and Hall.

Chatfield, C., Koehler, A. B., Ord, J. K., and Snyder, R. D. (2001). "A New Look at Models for Exponential Smoothing." *The Statistician,* 50, part 2, 147–159.

Cheung, Y.-W., and Chinn, M. D. (1999). "Are Macroeconomic Forecasts Informative? Cointegration Evidence from the ASA/NBER Surveys." NBER Working Paper No. 6926.

Chong, Y. Y., and Hendry, D. F. (1986). "Econometric Evaluation of Linear Macroeconomic Models." *Review of Economic Studies,* 53, 671–690.

Christoffersen, P. F., and Diebold, F. X. (1996). "Further Results on Forecasting and Model Selection under Asymmetric Loss." *Journal of Applied Econometrics,* 11, 561–572.

Christoffersen, P. F., and Diebold, F. X. (1997). "Optimal Prediction under Asymmetric Loss." *Econometric Theory,* 13, 808–817.

Christoffersen, P. F., and Diebold, F. X. (1998). "Cointegration and Long-Horizon Forecasting." *Journal of Business and Economic Statistics,* 16, 450–458.

Clemen, R. T. (1989). "Combining Forecasts: A Review and Annotated Bibliography." *International Journal of Forecasting,* 5, 559–581.

Clemen, R. T., and Winkler, R. L. (1986). "Combining Economic Forecasts." *Journal of Business and Economic Statistics,* 4, 39–46.

Clements, M. P., and Hendry, D. F. (1994). "Towards a Theory of Economic Forecasting." In C. P. Hargreaves (ed.), *Nonstationary Times Series Analysis and Cointegration.* Oxford: Oxford University Press.

Clements, M. P., and Hendry, D. F. (1998). *Forecasting Economic Time Series.* The Marshall Lectures in Economic Forecasting. Cambridge: Cambridge University Press.

Clements, M. P., and Hendry, D. F., eds. (2002). *A Companion to Economic Forecasting.* Oxford: Blackwell.

Cleveland, W. S. (1993). *Visualizing Data.* Summit, N.J.: Hobart.

Cleveland, W. S. (1994). *The Elements of Graphing Data.* 2nd ed. Monterey Park, Calif.: Wadsworth.

Cook, R. D., and Weisberg, S. (1994). *An Introduction to Regression Graphics.* New York: Wiley.

Croushore, D. (1993). "The Survey of Professional Forecasters." *Business Review,* November–December, Federal Reserve Bank of Philadelphia (http://www.phil.frb.org/econ/br).

Croushore, D., and Stark, T. (2001). "A Real-Time Dataset for Macroeconomists." *Journal of Econometrics,* 105, 111–130.

Dacorogna, M. M., et al. (2001). *An Introduction to High-Frequency Finance.* New York: Academic Press.

Dickey, D. A. (1976). *Estimation and Hypothesis Testing in Nonstationary Time Series.* Ph.D. dissertation, Iowa State University.

Dickey, D. A., and Gonzalez-Farias, G. (1992). "A New Maximum-Likelihood Approach to Testing for Unit Roots." Manuscript, Department of Statistics, North Carolina State University.

Diebold, F. X. (1988). *Empirical Modeling of Exchange Rate Dynamics.* New York: Springer.

Diebold, F. X. (1989). "Forecast Combination and Encompassing: Reconciling Two Divergent Literatures." *International Journal of Forecasting,* 5, 589–592.

Diebold, F. X. (2001). "Econometrics: Retrospect and Prospect." *Journal of Econometrics,* 100, 73–75.

Diebold, F. X, Engle, R. F., Favero, C., Gallo, G., and Schorfheide, F. (2005). *The Econometrics of Macroeconomics, Finance, and the Interface.* Special issue of *Journal of Econometrics.*

Diebold, F. X., Giorgianni, L., and Inoue, A. (1996). "STAMP 5.0: A Review." *International Journal of Forecasting*, 12, 309–315.

Diebold, F. X., and Kilian, L. (2000). "Unit Root Tests are Useful for Selecting Forecasting Models." *Journal of Business and Economic Statistics*, 18, 265–273.

Diebold, F. X., and Kilian, L. (2001). "Measuring Predictability: Theory and Macroeconomic Applications." *Journal of Applied Econometrics*, 16, 657–669.

Diebold, F. X., Lee, J.-H., and Weinbach, G. (1994). "Regime Switching with Time-Varying Transition Probabilities." In C. Hargreaves (ed.), *Nonstationary Time Series Analysis and Cointegration*. Oxford: Oxford University Press, 283–302. Reprinted in Diebold and Rudebusch (1999).

Diebold, F. X., and Lopez, J. (1995). "Modeling Volatility Dynamics." In Kevin Hoover (ed.), *Macroeconometrics: Developments, Tensions and Prospects*. Boston: Kluwer Academic Press, 427–472.

Diebold, F. X., and Lopez, J. (1996). "Forecast Evaluation and Combination." In G. S. Maddala and C. R. Rao (eds.), *Handbook of Statistics*. Amsterdam: North-Holland, 241–268.

Diebold, F. X., and Mariano, R. (1995). "Comparing Predictive Accuracy." *Journal of Business and Economic Statistics*, 13, 253–265. Reprinted in Diebold and Rudebusch (1999).

Diebold, F. X., and Nerlove, M. (1989). "The Dynamics of Exchange Rate Volatility: A Multivariate Latent-Factor ARCH Model." *Journal of Applied Econometrics*, 4, 1–22.

Diebold, F. X., and Pauly, P. (1990). "The Use of Prior Information in Forecast Combination." *International Journal of Forecasting*, 6, 503–508.

Diebold, F. X., and Rudebusch, G. D. (1989). "Scoring the Leading Indicators." *Journal of Business*, 62, 369–391. Reprinted in Diebold and Rudebusch (1999).

Diebold, F. X., and Rudebusch, G. D. (1991). "Forecasting Output with the Composite Leading Index: An Ex Ante Analysis." *Journal of the American Statistical Association*, 86, 603–610. Reprinted in Diebold and Rudebusch (1999).

Diebold, F. X., and Rudebusch, G. D. (1996). "Measuring Business Cycles: A Modern Perspective." *Review of Economics and Statistics*, 78, 67–77. Reprinted in Diebold and Rudebusch (1999).

Diebold, F. X., and Rudebusch, G. D. (1999). *Business Cycles: Durations, Dynamics, and Forecasting*. Princeton, N. J.: Princeton University Press.

Diebold, F. X., and Senhadji, A. (1996). "The Uncertain Unit Root in Real GNP: Comment." *American Economic Review*, 86, 1291–1298. Reprinted in Diebold and Rudebusch (1999).

Diebold, F. X., Stock, J. H., and West, K. D., eds. (1999). *Forecasting and Empirical Methods in Macroeconomics and Finance, II*. Special issue of *Review of Economics and Statistics*, 81, 553–673.

Diebold, F. X., and Watson, M. W., eds. (1996). *New Developments in Economic Forecasting*. Special issue of *Journal of Applied Econometrics*, 11, 453–594.

Diebold, F. X., and West, K. D., eds. (1998). *Forecasting and Empirical Methods in Macroeconomics and Finance*. Special issue of *International Economic Review*, 39, 811–1144.

Diebold, F. X., and West, K. D., eds. (2001). *Forecasting and Empirical Methods in Macroeconomics and Finance III*. Special issue of *Journal of Econometrics*, 105, 1–308.

Doan, T., Litterman, R., and Sims, C. (1984). "Forecasting and Conditional Prediction Using Realistic Prior Distributions." *Econometric Reviews*, 3, 1–144.

Efron, B., and Tibshirani, R. J. (1993). *An Introduction to the Bootstrap*. New York: Chapman and Hall.

Elliott, G., Granger, C. W. J., and Timmermann, A., eds. (2005). *Handbook of Economic Forecasting*. Amsterdam: North-Holland.

Elliott, G., Rothenberg, T. J., and Stock, J. H. (1996). "Efficient Tests for an Autoregressive Unit Root." *Econometrica*, 64, 813–836.

Elliott, G., and Timmermann, A. (2002). *Optimal Forecast Combination under Regime Switching*. Unpublished manuscript, Department of Economics, University of California, San Diego.

Engle, R. F. (1982). "Autoregressive Conditional Heteroskedasticity with Estimates of the Variance of U. K. Inflation." *Econometrica*, 50, 987–1008.

Engle, R. F., ed. (1995). *ARCH: Selected Readings*. Oxford: Oxford University Press.

Engle, R. F., and Brown, S. J. (1986). "Model Selection for Forecasting." *Applied Mathematics and Computation*, 20, 313–327.

Engle, R. F., and Granger, C. W. J. (1987). "Co-Integration and Error Correction: Representation, Estimation and Testing." *Econometrica*, 55, 251–276.

Engle, R. F., and Lee, G. (1999). "A Permanent and Transitory Model of Stock Return Volatility." In R. Engle and H. White (eds.), *Cointegration, Causality, and Forecasting: A Festschrift in Honor of Clive W. J. Granger*. Oxford: Oxford University Press.

Engle, R. F., Lilien, D. M., and Robins, R. P. (1987). "Estimating Time Varying Risk Premia in the Term Structure: The Arch-M Model." *Econometrica*, 55, 391–407.

Engle, R. F., and Yoo, B. S. (1987). "Forecasting and Testing in Cointegrated Systems." *Journal of Econometrics*, 35, 143–159.

Fair, R. C. (1996). "Computing Median Unbiased Estimates in Macroeconometric Models." *Journal of Applied Econometrics*, 11, 431–435.

Fair, R. C., and Shiller, R. J. (1990). "Comparing Information in Forecasts from Econometric Models." *American Economic Review*, 80, 375–389.

Faraway, J., and Chatfield, C. (1995). *Time Series Forecasting with Neural Networks: A Case Study.* Research Report 95-06, Statistics Group, University of Bath, U.K.

Ferrall, C. (1994). "A Review of Stata 3.1." *Journal of Applied Econometrics*, 9, 469–478.

Fildes, R., and Stekler, H. (2000). "The State of Macroeconomic Forecasting." Manuscript.

Findley, D. F. (1983). "On the Use of Multiple Models for Multi-Period Forecasting." *Proceedings of the American Statistical Association, Business and Economic Statistics Section*, 1983, 528–531.

Findley, D. F. (1985). "Model Selection for Multi-Step-Ahead Forecasting." In *Identification and System Parameter Estimation*, 7th IFAC/FORS Symposium, 1039–1044.

Franses, P. H., and Paap, R. (2004). *Periodic Time Series Models.* Oxford: Oxford University Press.

Frumkin, N. (2004). *Tracking America's Economy.* 4th ed. Armonk, N.Y.: Sharpe.

Fuller, W. A. (1976). *Introduction to Statistical Time Series.* New York: Wiley.

Gershenfeld, N. A., and Weigend, A. S. (1993). "The Future of Time Series." In A. S. Weigend and N. A. Gershenfeld (eds.), *Time Series Prediction: Forecasting the Future and Understanding the Past.* Reading, Mass.: Addison-Wesley, 1–70.

Ghysels, E., and Osborne, D. R. (2001). *The Econometric Analysis of Seasonal Time Series.* Cambridge: Cambridge University Press.

Glosten, L. R., Jagannathan, R., and Runkle, D. E. (1993). "On the Relation between the Expected Value and the Volatility of the Nominal Excess Return on Stocks." *Journal of Finance*, 48, 1779–1801.

Gourieroux, C., and Jasiak, J. (2001). *Financial Econometrics.* Princeton, N.J.: Princeton University Press.

Granger, C. W. J. (1969). "Investigating Causal Relations by Econometric Models and Cross-Spectral Methods." *Econometrica*, 37, 424–438.

Granger, C. W. J. (1990). "Aggreagtion of Time Series Variables: A Survey." In T. Barker and M. H. Pesaran (eds.), *Disaggregation in Econometric Modelling.* London: Routledge.

Granger, C. W. J., King, M. L., and White, H. (1995). "Comments on the Testing of Economic Theories and the use of Model Selection Criteria." *Journal of Econometrics*, 67, 173–187.

Granger, C. W. J., and Newbold, P. (1974). "Spurious Regressions in Econometrics." *Journal of Econometrics*, 2, 111–120.

Granger, C. W. J., and Newbold, P. (1986). *Forecasting Economic Time Series.* 2nd ed. Orlando, Fla.: Academic Press.

Granger, C. W. J., and Ramanathan, R. (1984). "Improved Methods of Forecasting." *Journal of Forecasting*, 3, 197–204.

Granger, C. W. J., and Teräsvirta, Y. (1993). *Modelling Nonlinear Economic Relationships.* Oxford: Oxford University Press.

Hall, A. (1994). "Testing for a Unit Root in Time Series with Pretest Data-Based Model Selection." *Journal of Business and Economic Statistics*, 12, 461–470.

Hallman, J. (1993). "Review of S+." *Journal of Applied Econometrics*, 8, 213–220.

Hamilton, J. D. (1989). "A New Approach to the Economic Analysis of Nonstationary Time Series and the Business Cycle." *Econometrica*, 57, 357–384.

Hamilton, J. D. (1994). *Time Series Analysis.* Princeton, N.J.: Princeton University Press.

Hansen, P. R. (2001). "An Unbiased and Powerful Test for Superior Predictive Ability." Working Paper 01-06, Department of Economics, Brown University.

Hardy, C. O. (1923). *Risk and Risk Bearing.* Chicago: University of Chicago Press. (Reissued in the *Risk Classics Library.* London: Risk Publications, 1999.)

Harvey, A. C. (1989). *Forecasting, Structural Time Series Models and the Kalman Filter.* Cambridge: Cambridge University Press.

Harvey, A. C. (1990). *The Econometric Analysis of Time Series.* 2nd ed. Cambridge, Mass.: MIT Press.

Harvey, A. C. (1993). *Time Series Models.* 2nd ed. Cambridge, Mass.: MIT Press.

Harvey, D. I., Leybourne, S. J., and Newbold, P. (1997). "Testing the Equality of Prediction Mean Squared Errors." *International Journal of Forecasting*, 13, 281–291.

Hendry, D. F., and Mizon, G. E. (1978). "Serial Correlation as a Convenient Simplification, Not a Nuisance: A Comment on a Study of the Demand for Money by the Bank of England." *Economic Journal*, 88, 549–563.

Holt, C. C. (1957). "Forecasting Seasonals and Trends by Exponentially Weighted Moving Averages." ONR Research Memorandum No. 52, Carnegie Institute of Technology.

Hylleberg, S. (1986). *Seasonality in Regression.* Orlando, Fla.: Academic Press.

Ingram, B., and Whiteman, C. (1994). "Supplanting the 'Minnesota' Prior: Forecasting Macroeconomic Time Series Using Real Business Cycle Model Priors." *Journal of Monetary Economics*, 34, 497–510.

Jarque, C. M., and Bera, A. K. (1987). "A Test for Normality of Observations and Regression Residuals." *International Statistical Review*, 55, 163–172.

Jenkins, G. M. (1979). "Practical Experiences with Modelling and Forecasting Time Series." In O. D. Anderson (ed.), *Forecasting.* Amsterdam: North-Holland.

Johansen, S. (1995). *Likelihood Based Inference in Cointegrated Vector Autoregressive Models.* Oxford: Oxford University Press.

Jorgenson, D. (1966). "Rational Distributed Lag Functions." *Econometrica*, 34, 135–149.

Kennedy, P. (1998). *A Guide to Econometrics.* 4th ed. Cambridge, Mass.: MIT Press.

Kim, J., and Trivedi, P. (1995). "Econometric Time Series Analysis Software: A Review." *American Statistician,* 48, 336–346.

Klein, L. R. (1971). An Essay on the Theory of Economic Prediction. Chicago: Markham.

Klein, L. R. (1981). *Econometric Models as Guides for Decision Making.* New York: Free Press.

Klein, L. R. (1983). Lectures in Econometrics. Amsterdam: North-Holland.

Klein, L. R., ed. (1991). *Comparative Performance of U.S. Econometric Models.* Oxford: Oxford University Press.

Klein, L. R., and Young, R. M. (1980). *An Introduction to Econometric Forecasting and Forecasting Models.* Lexington, Mass.: Heath.

Koopman, S. J., Harvey, A. C., Doornik, J. A., and Shephard, N. (1995). *Stamp 5.0: Structural Time Series Analyzer, Modeller and Predictor.* London: Chapman and Hall.

Krämer, W., Ploberger, W., and Alt, R. (1988). "Testing for Structural Change in Dynamic Models." *Econometrica,* 56, 1355–1369.

Kuan, C. M., and Liu, Y. (1995). "Forecasting Exchange Rates Using Feedforward and Recurrent Neural Networks." *Journal of Applied Econometrics,* 10, 347–364.

Levenbach, H., and Cleary, J. P. (1984). *The Modern Forecaster.* Belmont, Calif.: Lifetime Learning Publications.

Ljung, G. M., and Box, G. E. P. (1978). "On a Measure of Lack of Fit in Time-Series Models." *Biometrika,* 65, 297–303.

Lütkepohl, H. (1991). *Introduction to Multiple Time Series Analysis.* New York: Springer.

MacKinnon, J. G. (1991). "Critical Values for Cointegration Tests." In R. F. Engle and C. W. J. Granger (eds.), *Long-Run Economic Relationships.* Oxford: Oxford University Press.

Maddala, G. S. (2001). *Introduction to Econometrics.* 3rd ed. New York: Macmillan.

Makridakis, S., and Wheelwright, S. (1987). *The Handbook of Forecasting: A Manager's Guide.* 2nd ed. New York: Wiley.

Makridakis, S., and Wheelwright, S. (1997). *Forecasting: Methods and Applications.* 3rd ed. New York: Wiley.

Malkiel, B. G. (1999). *A Random Walk Down Wall Street.* 7th ed. New York: Norton.

McCullough, B. D., and Vinod, H. D. (1999). "The Numerical Reliability of Econometric Software." *Journal of Economic Literature,* 37, 633–665.

McLeod, A. I., and Li, W. K. (1983). "Diagnostic Checking of ARMA Time Series Models Using Squared Residual Autocorrelations." *Journal of Time Series Analysis,* 4, 269–273.

McNees, S. K. (1988). "How Accurate Are Macroeconomic Forecasts?" *New England Economic Review,* July–August, 15–36.

Mincer, J., and Zarnowitz, V. (1969). "The Evaluation of Economic Forecasts." In J. Mincer (ed.), *Economic Forecasts and Expectations.* New York: National Bureau of Economic Research.

Morgenstern, O. (1963). *On the Accuracy of Economic Observations.* Princeton, N.J.: Princeton University Press.

Nelson, C. R. (1972). "The Prediction Performance of the F.R.B.-M.I.T.-Penn Model of the U.S. Economy." *American Economic Review,* 62, 902–917.

Nelson, D. B. (1991). "Conditional Heteroskedasticity in Asset Returns: A New Approach." *Econometrica,* 59, 347–370.

Nerlove, M., Grether, D. M., and Carvalho, J. L. (1979). *Analysis of Economic Time Series: A Synthesis* (2nd ed., 1996). New York: Academic Press.

Newbold, P., Agiakloglou, C., and Miller, J. P. (1994). "Adventures with ARIMA Software." *International Journal of Forecasting,* 10, 573–581.

Ng, S., and Perron, P. (1995). "Unit Root Tests in ARMA Models with Data-Dependent Methods for the Selection of the Truncation Lag." *Journal of the American Statistical Association,* 90, 268–281.

Ord, K., and Lowe, S. (1996). "Automatic Forecasting." *American Statistician,* 50, 88–94.

Pagan, A. R., and Robertson, J. (2002). "Forecasting for Policy." In M. P. Clements and D. F. Hendry (eds.), *A Companion to Economic Forecasting.* Oxford: Blackwell.

Pantula, S. G., Gonzalez-Farias, G., and Fuller, W. A. (1994). "A Comparison of Unit Root Test Criteria." *Journal of Business and Economic Statistics,* 12, 449–459.

Pesaran, M. H., Pierse, R. G., and Kumar, M. S. (1989). "Econometric Analysis of Aggregation in the Context of Linear Prediction Models." *Econometrica,* 57, 861–888.

Pesaran, M. H., and Samiei, H. (1995). "Forecasting Ultimate Resource Recovery." *International Journal of Forecasting,* 11, 543–555.

Phillips, P. C. B., and Perron, P. (1988). "Testing for a Unit Root in Time Series Regression." *Biometrika,* 75, 335–346.

Pindyck, R. S., and Rubinfeld, D. L. (1997). *Econometric Models and Economic Forecasts.* 4th ed. New York: McGraw-Hill.

Press, W. H., Teukolsky, S. A., Vetterling, W. T., and Flannery, B. P. (1992). *Numerical Recipes: The Art of Scientific Computing.* Cambridge: Cambridge University Press.

Ramsey, J. (1969). "Tests for Specification Errors in Classical Linear Least Squares Regression Analysis." *Journal of the Royal Statistical Society* B, 31, 350–371.

Ripley, B. D. (1996). *Pattern Recognition and Neural Networks.* Oxford: Oxford University Press.

Rudebusch, G. D. (1993). "The Uncertain Unit Root in Real GNP." *American Economic Review,* 83, 264–272. Reprinted in Diebold and Rudebusch (1999).

Rust, J. (1993). "Gauss and Matlab: A Comparison." *Journal of Applied Econometrics*, 8, 307–324.

Rycroft, R. S. (1993). "Microcomputer Software of Interest to Forecasters in Comparative Review: An Update." *International Journal of Forecasting*, 9, 531–575.

Sanderson, F. H. (1953). *Methods of Crop Forecasting.* Cambridge, Mass.: Harvard University Press.

Schwarz, G. (1978). "Estimating the Dimension of a Model." *Annals of Statistics*, 6, 461–464.

Shibata, R. (1980). "Asymptotically Efficient Selection of the Order of the Model for Estimating the Parameters of a Linear Process." *Annals of Statistics*, 8, 147–164.

Sims, C. A. (1972). "Money, Income and Causality." *American Economic Review*, 62, 540–552.

Sims, C. A. (1980). "Macroeconomics and Reality." *Econometrica*, 48, 1–48.

Slutsky, E. (1927). "The Summation of Random Causes as the Source of Cyclic Processes." *Econometrica*, 5, 105–146.

Stine, R. A. (1987). "Estimating Properties of Autoregressive Forecasts." *Journal of the American Statistical Association*, 82, 1072–1078.

Stock, J. H., and Watson, M. W. (1988). "Variable Trends in Economic Time Series." *Journal of Economic Perspectives*, 2, 147–174.

Stock, J. H., and Watson, M. W. (1999). "A Comparison of Linear and Nonlinear Univariate Models for Forecasting Macroeconomic Time Series." In R. Engle and H. White (eds.), *Cointegration, Causality, and Forecasting: A Festschrift in Honor of Clive W. J. Granger.* Oxford: Oxford University Press, 1–44.

Stock, J. H., and Watson, M. W. (1998). A Dynamic Factor Model Framework for Forecast Combination. Unpublished manuscript, Kennedy School, Harvard University, and Woodrow Wilson School, Princeton University.

Stock, J. H., and Watson, M. W. (2003). "Macroeconomic Forecasting in the Euro Area: Country-Specific versus Area-Wide Information." *European Economic Review*, 47, 1–18.

Swanson, N. R., and White, H. (1995). "A Model-Selection Approach to Assessing the Information in the Term Structure Using Linear-Models and Artificial Neural Networks." *Journal of Business and Economic Statistics*, 13, 265–275.

Taylor, S. (1996). *Modeling Financial Time Series.* 2nd ed. New York: Wiley.

Taylor, S. (2005). *Asset Price Dynamics, Volatility and Prediction.* Princeton, N. J.: Princeton University Press.

Theil, H. (1966). *Applied Economic Forecasting.* Amsterdam: North-Holland.

Tiao, G. C., and Tsay, R. S. (1994). "Some Advances in Non-Linear and Adaptive Modeling in Time Series." *Journal of Forecasting*, 13, 109–131.

Tong, H. (1990). *Non-linear Time Series.* Oxford: Clarendon.

Tsay, R. (1984). "Order Selection in Nonstationary Autoregressive Models." *Annals of Statistics*, 12, 1425–1433.

Tufte, E. R. (1983). *The Visual Display of Quantitative Information.* Cheshire, Conn.: Graphics Press.

Tukey, J. W. (1977). *Exploratory Data Analysis.* Reading Mass.: Addison-Wesley.

Varian, H. (1974). "A Bayesian Approach to Real Estate Assessment." In S. E. Feinberg and A. Zellner (eds.), *Studies in Bayesian Econometrics and Statistics in Honor of L. J. Savage.* Amsterdam: North-Holland.

Wallis, K. F. (1995). "Large-Scale Macroeconometric Modeling." In M. H. Pesaran and M. R. Wickens (eds.), *Handbook of Applied Econometrics.* Oxford: Blackwell.

Wallis, K. F., and Whitley, J. D. (1991). "Sources of Error in Forecasts and Expectations: UK Economic Models, 1984–88." *Journal of Forecasting*, 10, 231–253.

West, K. D. (1996). "Asymptotic Inference About Predictive Ability." *Econometrica*, 64, 1067–1084.

White, H. (2000). "A Reality Check for Data Snooping." *Econometrica*, 68, 1097–1126.

Winkler, R. L., and Makridakis, S. (1983). "The Combination of Forecasts." *Journal of the Royal Statistical Society* A, 146, 150–157.

Winters, P. R. (1960). "Forecasting Sales by Exponentially Weighted Moving Averages." *Management Science*, 6, 324–342.

Wold, H. O. (1954). *A Study in the Analysis of Stationary Time Series.* 2nd ed. Uppsala, Sweden: Almquist and Wicksell.

Wonnacott, T. H., and Wonnacott, R. J. (1990). *Introductory Statistics.* 5th ed. New York: Wiley.

Zarnowitz, V., and Braun, P. (1993). "Twenty-Two Years of the N.B.E.R.-A.S.A. Quarterly Economic Outlook Surveys: Aspects and Comparisons of Forecasting Performance." In J. H. Stock and M. W. Watson (eds.), *Business Cycles, Indicators and Forecasting.* Chicago: University of Chicago Press for NBER.

Zellner, A. (1986). "Bayesian Estimation and Prediction Using Asymmetric Loss Functions." *Journal of the American Statistical Association*, 81, 446–451.

Zellner, A. (1992). "Statistics, Science and Public Policy." *Journal of the American Statistical Association*, 87, 1–6.

Zellner, A., and Hong, C. (1989). "Forecasting International Growth Rates Using Bayesian Shrinkage and Other Procedures." *Journal of Econometrics*, 40, 183–202.

Zellner, A., Hong, C., and Min, C.-K. (1991). "Forecasting Turning Points in International Output Growth Rates Using Bayesian Exponentially Weighted Autoregression, Time-Varying Parameter, and Pooling Techniques." *Journal of Econometrics*, 49, 275–304.

A

Agiakloglou, C., 166
Akaike, H., 97
Alt, R., 217
Andersen, T. G., 352
Anderson, D. R., 6
Andrews, D. W. K., 325
Anscombe, F. J., 70
Armstrong, J. S., 7, 49, 324
Auerbach, A. J., 251

B

Bails, D. G., 7
Bartlett, M., 135
Bates, J. M., 285
Baumohl, B., 10
Bera, A. K., 32, 70
Beveridge, S., 325
Bollerslev, T., 350, 351, 352
Box, G. E. P., 49, 135, 167, 284, 325
Braun, P., 283, 285
Breidt, F. J., 189
Brown, R. L., 217
Brown, S. J., 97
Burns, A. F., 250

C

Campbell, J. Y., 326
Carvalho, J. L., 111, 217, 325
Chatfield, C., 6, 189, 253, 254, 314*n*
Cheung, Y.-W., 324
Chinn, M. D., 324
Chong, Y. Y., 285
Chou, R. Y., 352
Christoffersen, P. F., 49, 324, 351
Cleary, J. P., 49
Clemen, R. T., 284, 285
Clements, M. P., 7, 189
Cleveland, W. S., 70
Cook, R. D., 70
Croushore, D., 282, 285

D

Dacorogna, M. M., 352
Davis, R. A., 189
Dickey, D. A., 296*n*, 301*n*, 325
Diebold, F. X., 7, 49, 168, 250*n*, 251, 282, 283, 285, 293*n*, 324, 325, 326, 326*n*, 351, 352
Doan, T., 251
Donihue, M., 283
Doornik, J. A., 326
Dunsmuir, W. T. M., 189
Durbin, J., 217

E

Efron, B., 189, 217, 254
Elliott, G., 7, 285, 301*n*
Engle, R. F., 7, 97, 323, 352
Evans, J. M., 217

F

Fair, R. C., 285, 325
Faraway, J., 253
Favero, C., 7
Ferrall, C., 8*n*
Fildes, R., 165, 324
Findley, D. F., 98
Franses, P. H., 111
Frumkin, N., 10
Fuller, W. A., 296*n*, 299, 300, 301*n*, 325

G

Gallo, G., 7
Gershenfeld, N. A., 254
Ghysels, E., 111
Giorgianni, L., 326*n*
Glosten, L. R., 352
Gonzalez-Farias, G., 301, 301*n*
Granger, C. W. J., 6, 7, 49, 97, 152*n*, 167, 168, 254, 285, 323
Grether, D. M., 111, 217, 325

H

Hallman, J., 8*n*
Hamilton, J. D., 6, 168, 189, 233*n*, 325
Hansen, P. R., 285
Hardy, C. O., 250*n*
Harvey, A. C., 6, 189, 217, 321, 325, 326
Harvey, D. I., 285
Hendry, D. F., 7, 189, 254, 285
Holt, C. C., 326
Hong, C., 250
Hylleberg, S., 111

I

Ingram, B., 251
Inoue, A., 326*n*

J

Jagannathan, R., 352
Jarque, C. M., 32, 70
Jenkins, G. M., 249*n*, 254, 325
Jenkins, G. W., 49, 167, 284
Johansen, S., 323
Jorgenson, D., 135

K

Kennedy, P., 6, 98
Kilian, L., 282, 326
Kim, J., 8*n*
King, M. L., 97
Klein, L. R., 49, 249*n*, 282, 283
Kolmogorov, A., 189
Koopman, S. J., 326
Krämer, W., 217
Kroner, K. F., 352
Kuan, C. M., 217, 253
Kumar, M. S., 168

L

Labys, P., 352
Lee, G., 352
Lee, J.-H., 168
Levenbach, H., 49
Leybourne, S. J., 285
Li, W. K., 350
Lilien, D. M., 352
Litterman, R., 251
Liu, Y., 217, 253
Ljung, G. M., 135
Lopez, J., 285, 352
Lowe, S., 321
Lütkepohl, H., 254

M

MacKinnon, J. G., 325
Maddala, G. S., 6, 32
Makridakis, S., 7, 49, 285
Malkiel, B. G., 325
Mariano, R., 285
McCullough, B. D., 166
McLeod, A. I., 350
McNees, S. K., 283
Michener, R., 214*n*
Mikkelson, H. O., 350
Miller, J. P., 166
Min, C.-K., 250
Mincer, J., 285
Mitchell, W. C., 250
Mizon, G. E., 254
Morgenstern, O., 282

N

Nelson, C. R., 285, 325
Nelson, D. B., 352
Nerlove, M., 111, 217, 325
Newbold, P., 6, 49, 152*n*, 166, 167, 254, 285
Ng, S., 301

O

Ord, K., 321
Osborne, D. R., 111

P

Paap, R., 111
Pagan, A. R., 249*n*
Pantula, S. G., 301*n*
Pauly, P., 285
Peppers, L. C., 7
Perron, P., 301, 325, 326
Pesaran, M. H., 168, 283
Phillips, P. C. B., 325
Pierce, D. A., 135
Pierse, R. G., 168
Pindyck, R. S., 6, 32, 229*n*
Ploberger, W., 217
Press, W. H., 135

R

Ramanathan, R., 285
Ramsey, J., 251
Reinsel, G., 49, 167, 284, 325
Ripley, B. D., 253
Rissanen, J., 217
Robertson, J., 249*n*
Robins, R. P., 352
Rothenberg, T. J., 301*n*
Rubinfeld, D. L., 6, 32, 229*n*
Rudebusch, G. D., 168, 250*n*, 251, 283, 325
Runkle, D. E., 352
Rust, J., 8*n*
Rycroft, R. S., 7*n*

S

Samiei, H., 283
Sanderson, F. H., 249
Schorfheide, F., 7
Schwarz, G., 97
Senhadji, A., 293*n*, 325
Shephard, N., 326
Shibata, R., 97
Shiller, R. J., 285
Sims, C., 251, 254
Slutsky, E., 167
Stark, T., 282
Steckler, H., 165
Stine, R. A., 189
Stock, J. H., 7, 168, 284, 285, 301*n*, 326
Swanson, N. R., 254
Sweeney, D. J., 6

T

Taylor, S., 7*n*, 352
Teräsvirta, Y., 168
Theil, H., 282, 324
Tiao, G. C., 98
Tibshirani, R. J., 189, 217, 254
Timmerman, A., 7, 285
Tong, H., 168
Trivedi, P., 8*n*
Tsay, R., 98, 325
Tufte, E. R., 70
Tukey, J. W., 70

V

Varian, H., 189
Vinod, H. D., 166

W

Wallis, K. F., 249*n*, 283
Watson, M. W., 7, 49, 168, 284, 285, 326
Weigend, A. S., 254
Weinbach, G., 168
Weisberg, S., 70
West, K. D., 7, 285
Wheelwright, S., 7, 49
White, H., 97, 254, 285
Whiteman, C., 251
Whitley, J. D., 283
Wiener, N., 189, 254, 282
Williams, T. A., 6
Winkler, R. L., 285
Winters, P. R., 326
Wold, H. O., 135, 189
Wonnacott, R. J., 6, 32, 122*n*
Wonnacott, T. H., 6, 32, 122*n*
Wooldridge, J. M., 351

Y

Yoo, B. S., 323
Young, R. M., 249*n*

Z

Zarnowitz, V., 283, 285
Zellner, A., 49, 189, 250

A

absolute error loss, 38
absolute loss, 38
additive unobserved-components decomposition, 217
adjusted R^2, 27
aggregation, 48, 165
Akaike information criterion, 28, 82–86
analog principle, 127
Anscombe's quartet, 51
approximation of the Wold representation, 127
ARCH(*p*) process, 330–333, 337–340
argmin, 80*n*
ARIMA(*p*, 1, *q*) model, 292
ARIMA(*p*, *d*, *q*) model, 292
ARMA (autoregressive moving average) models, 137, 152–153
ARMA(*p*, *q*) process, 152, 178–179, 226–227
AR (autoregressive) models, 137, 145–152
AR(1) process, 145–149, 178
AR(*p*) process, 150–152, 177–178
aspect ratio, 61
asymmetric loss, 36, 186
asymmetric response, 337–338
asymmetry, 15
asymptotic efficiency, 86
augmented Dickey-Fuller test, 301
autocorrelation function, 115, 140
autocovariance function, 114
autoregression, 116
autoregressive representation, 141

B

banking to 45 degrees, 62
Bartlett bands, 135
Bayesian analysis, 251
Bayesian VARs, 251
bias, 261
bias corrections, 96
bivariate scatterplots, 52
bootstrapping, 188
bottom-up forecasting model, 165
Box-Jenkins model, 167, 325
Box-Pierce Q-statistic, 129
Breusch-Godfrey test, 164

C

calendar effects, 102
causal models, 219
central tendency, 14
chain rule of forecasting, 177–180
chartjunk, 60
Cholesky factor, normalization by, 233
cointegration, 323
combining weights, 266–267
common factors, 163
common scales, 69
complex roots, 144
component GARCH, 339
conditional distribution, 16
conditional expectation, 20
conditional forecasting models, 220–223
conditional mean, 16, 122, 172
conditional moments, 16
conditional variance, 16, 122
condition for covariance stationarity, 150*n*
condition for invertibility of the MA(*q*) process, 144
confidence interval, 41*n*
consistency, 85
constant term, 22*n*
contingency analysis, 220
continuous random variables, 14
correlation, 16
correlogram analysis, 130*n*
covariance, 15, 114, 264
covariance stationarity, 113–117, 150*n*
cross-correlation function, 237
cross validation, 210
cross-variable dynamics, 228
CUSUM analysis, 209
CUSUM plot, 209
CUSUM statistic, 209
cutoff in the autocorrelation function, 140
cycles
  autocorrelation, 127–129
  Canadian employment dynamics, 130–132
  covariance stationary time series, 113–117
  defined, 112
  estimation and inference for the mean, 127
  general linear process, 125–126
  lag operator, 123–124
  rational distributed lags, 126–127

  - sample partial autocorrelations, 129–130
  - white noise, 117–123
  - Wold's theorem, 124–125

D

- data-generating process (DGP), 85–86
- data mining, 83
- decision environment, 34–39
- degrees of freedom, 84*n*
- Delphi method, 285
- density forecast, 41, 175–176, 179–180
- dependent variables, 24
- deterministic seasonality, 99
- deterministic trends
  - defined, 5, 72
  - stochastic trends and, 290
  - unit root tests, 299
- detrending, 96
- DGP (data-generating process), 85–86
- Dickey-Fuller bias, 296*n*
- Dickey-Fuller distribution, 299
- Dickey-Fuller statistic, 299–300
- direction-of-change forecast, 39
- disaggregation, 48, 165
- discrete probability distributions, 14
- discrete random variables, 14
- dispersion, 14
- distributed lag regression model, 224–228
- distributed lags, 123, 227
- distributions, 14–15
- disturbance, 20
- Durbin's *h*-test, 164
- Durbin-Watson statistic, 28–29

E

- econometric models, 6, 330
- endogenous variable, 219
- equal-weights constraint, 267–268
- error correction, 323
- error variances, 83, 174, 261
- estimator, 16*n*
- evaluation and comparison of forecast accuracy, 260–262
- event outcome forecasts, 40
- event timing forecasts, 40
- Eviews software package, 7–8, 57*n*
- exogenous variables, 219, 338
- expected loss, 172
- expected value, 14
- explanatory models, 219
- explanatory variable, 219
- exploratory data analysis, 70
- exponential GARCH, 338
- exponentially weighted moving average, 314
- exponential smoothing, 314
- exponential trend, 78–79
- ex post (off-line) smoothing, 96

F

- *F* distribution, 18
- feedback, 239
- financial econometrics, 330
- first-order serial correlation, 164
- fitted values, 20
- forecast combination, 263–268
- forecast encompassing, 263–264
- forecast errors, 37, 173
- forecast error variances, 174
- forecast horizon, 35, 43–45
- forecasting
  - chain rule of, 177–180
  - defined, 1
  - employment, 180–184
  - evaluating multiple, 260–263
  - evaluating single, 257–260
  - GARCH models, 340–341
  - housing starts, 104–108
  - making operational, 176–177
  - moving average processes, 172–176
  - optimal, 171–172
  - with regression models, 219
  - resources, 6–8
  - retail sales, 87–94
  - the right-hand-side variables problem, 223
  - shipping volume, 268–280
  - with smoothing techniques, 317–318
  - stochastic trends and, 288–295
  - stock market volatility, 341–349
  - uses for, 1–3
  - yen/dollar exchange rate, 302–312, 318–319
- forecasting models
  - conditional, 220–223
  - liquor sales, 193–206, 212–214
  - overview, 3–5, 81–82, 191–193
  - recursive estimation procedures, 207–212
  - selecting, 82–86
  - specifying for employment forecasting, 154–163
  - top-down/bottom-up, 165
  - unconditional, 223
- forecast object, 34, 39–40
- forecast statement, 34–35, 40–43
- formal models of unobserved components, 216
- *F*-statistic, 25
- functional form, 251

G

- GARCH-in-mean, 339
- GARCH(p, q) process
  - extensions of, 337–340
  - forecasting, 340–341
  - overview, 333–337
- Gauss, Karl Friedrich, 119*n*
- general linear process, 125–126
- golden ratio, 61
- goodness of fit, 27
- Guassian distribution, 15
- Guassian white noise, 119

H

- heteroskedasticity, 330
- histogram, 57
- holiday variation, 102
- Holt-Winters smoothing, 315–317, 321
- *h*-step-ahead errors, 259
- *h*-step-ahead extrapolation forecasts, 44
- *h*-step-ahead forecast, 44

I

- impulse-response functions, 231–235
- independent white noise, 119
- independent zero-mean random noise, 81*n*
- information set
  - assessing optimality, 259–260
  - defined, 35, 171
  - overview, 45
- innovation distribution uncertainty, 188
- innovation outliers, 67
- innovations, 124
- innovation uncertainty, 220
- in-sample forecast errors, 20
- in-sample forecasts, 20
- in-sample overfitting, 83
- *International Journal of Forecasting*, 7
- interval forecast, 41, 175–176, 179–180
- invertibility, 141, 144

J

- Jarque-Bera test, 18
- joint distribution, 15
- *Journal of Applied Econometrics*, 7
- *Journal of Business and Economic Statistics*, 7
- *Journal of Business Forecasting*, 7
- *Journal of Forecasting*, 7

K

KISS principle, 46
Kolmogorov-Smirnov test, [illegible]
kurtosis, 15

L

lag operator, 123–124
least squares, 19
least-squares regression, [illegible]
leptokurtosis, 15
leverage effect, 337*n*
likelihood function, 24
linear forecasts, 172
linear least-squares forecast, 17[illegible]
linear models, 165
linear projection, 30, 172
linear trends, 73–75, 79
linex loss function, 186
linlin loss function, 186
Ljung-Box Q-statistic, 12[illegible]
location, 14
logarithmic regression [illegible]
logistic trend, 95
log-linear trend, 78
long-run variance, [illegible]
loss function, 34–39

M

MA(1) process, 138–14[illegible]
MA(*q*) process, 143–14[illegible]
marginal distributions, [illegible]
maximum likelihood [illegible]
mean, 14
mean absolute error, [illegible]
mean error, 261
mean reversion, 28[illegible]
mean squared error, [illegible]
measurement error, [illegible]
measurement outliers, [illegible]
methods and [illegible]
Mincer-Zarnowitz regression, [illegible]
missing observations, [illegible]
model selection, [illegible]
moments, 14–16, [illegible]
moving average (MA) [illegible] process
  ARMA models, [illegible]
    178–179, [illegible]
  exponentially weighted, 314
  forecasting, 172–176
  one-sided, 95
  one-sided weighted, 95
  overview, 137–145
  smoothing, 313–314
  two-sided, 95
multiple comparisons, 54
multiple linear regression models, 21
multiplicative unobserved-components decomposition, 217
multivariate GARCH, 351
multivariate model, 10–11, 58–59
multivariate random variables, 15–16
multiway scatterplot, 59

N

noise, 217
nondata link, 60
nonlinear least squares, 32
nonlinear model, 165
nonlinear regressions, 267–268
nonseasonal fluctuations, 101
nonzero mean, 299
normal distribution, 15
normalization by the Cholesky factor, 233
normal white noise, 119

O

one-step-ahead errors, 83, 259
one-sided moving averages, 95
one-sided weighted moving averages, 95
optimal forecasts
  defined, 39
  moving averages, 172–175
  overview, 171–172
  testing properties, 258–259
outliers, 48, 67
out-of-sample 1-step-ahead prediction error variance, 83
overfitting, 217

P

pairwise scatterplots, 52
parameter estimation uncertainty, 81–82, 188
parameter instability, 208
parameters, 20, 110
parameter uncertainty, 220–223
parsimonious, 127
parsimony principle, 35, 46
partial autocorrelation function, 116, 296–297
partial correlation, 67
periodic regression model, 110
Phillips-Perron tests, 325
point forecast, 41
polynomial distributed lags, 224–225
polynomial in the lag operator, 123
polynomial trend, 95
population distributions, 16
population regression, 116*n*
positive serial correlation, 29
predictive causality, 230–231
predictive stochastic complexity, 218
Prob(*F*-statistic), 25
probability density function, 14
probability forecast, 43
probability value, 22–23

Q

quadratic loss, 37
quadratic trends, 76–77

R

R^2, 27
ragged edges, 48
random number generators, 134
random variables, 14–15
random walks
  defined, 209*n*, 251, 289
  with drift, 290
  without drift, 290
rational distributed lags, 126–127, 224–225, 227
rational polynomials, 126
realization, 113
real-time (on-line) smoothing, 96
recursive cross validation, 210–212
recursive estimation, 207–212
recursive residuals, 207–212
recursive updating, 207*n*
regime switching, 166
regression analysis, 18–29
regression-based forecast combination method, 264
regression function, 20
regression intercept, 20*n*, 73
regression models
  forecasting with, 219
  housing starts/completions, 235–248
  logarithmic, 252
regression on seasonal dummies, 101
regression slope, 20*n*, 73
relational graphics, 58
residual plot, 29
residuals, 20
residual serial correlation, 87
Review of Economics and Statistics, 7
root mean squared error, 261–262

S

sample, 16
sample autocorrelation function, 127–129, 296–297
sample kurtosis, 17
sample mean
  defined, 17, 127
  of the dependent variable, 24
sample partial autocorrelations, 129–130
sample path, 113
sample skewness, 17
sample standard deviation
  defined, 17
  of the dependent variable, 24
sample statistic, 16*n*
sample variance, 17
scale, 14
scatterplot matrix, 59
scenario analysis, 220
Schwarz information criterion, 28, 82–86
seasonal adjustment, 87*n*, 101*n*
seasonal difference operator, 321
seasonal dummy variables, 102
seasonality
  defined, 99
  forecasting housing starts, 104–108
  forecasting series, 103–104
  Holt-Winters smoothing, 316–317, 321
  modeling, 101–103
  nature and sources of, 99–101
  stochastic, 321
seasonally adjusted time series, 101
second-order stationarity, 115*n*
serial correlation, 28, 266–267
serially uncorrelated, 118
shock persistence, 293
shrinkage principle, 35, 46
signal, 217
simple correlation, 67
simple exponential smoothing, 314
simple linear regressions, 21
simulation of a time series process, 134
single exponential smoothing, 314
skewness, 15
smoothing
  defined, 95
  exponential, 314
  ex post, 96
  forecasting with, 317–318
  overview, 312–318
  real-time, 96
specification uncertainty, 220
spurious regression, 253
stabilization of variance, 108
Stamp software, 326
standard deviation, 15
standard error
  defined, 22
  of the regression, 26
standardized recursive residuals, 208
state-space representations, 187
statistical graphics
  elements of style, 59–62
  graphing real GDP components, 63–66
  power of, 51–54
  simple techniques, 55–59
statistical models, 3
statistics, 16–18
stochastic processes, 137*n*
stochastic seasonality, 99, 321
stochastic trends, 5, 72*n*, 288–295
strong white noise, 119*n*
student's *t* GARCH, 351
sum of squared residuals, 24
superconsistency, 295
symmetric loss, 36, 186

T

t distribution, 18
Theil's *U*-statistic, 324
threshold GARCH, 337
threshold models, 166
time dummy, 73
time series
  covariance stationarity, 113–117
  defined, 4*n*, 113
  properties of, 117
  simulation of, 134
time series analysis, 6*n*
time series forecasts, 40
time series plot, 55
time trend, 73
time-varying combining weights, 266
time-varying parameters, 110, 215
time-varying volatility, 330
top-down forecasting model, 165
trading-day variation, 103
transfer function model, 227
trend, 72
trend models
  estimating, 80
  forecasting retail sales, 87–94
  overview, 72–79
t-statistic, 22
two-sided moving averages, 95

U

unconditional forecast models, 223
unconditional mean, 122
unconditional variance, 122
unforecastability principle, 259
unit (autoregressive) roots, 289, 295–302
unit root tests, 297–302
univariate models
  defined, 10
  restrictions, 227
  statistical graphics, 55–58
U statistic, 324

V

variance
  conditional, 16, 122
  defined, 14
  error, 83, 174, 261
  long-run, 339*n*
  long-run variance, 339*n*
  stabilization of, 108
  unconditional, 122
variance-covariance forecast combination method, 264
variance decomposition, 231–235
vector autoregressions, 219, 228–230
volatility clustering, 332
volatility dynamics, 332

W

weak stationarity, 115*n*
weak white noise, 119*n*
white noise, 117–123
Wold representation, 124–125, 127
Wold's representation theorem, 124–125

X

χ^2 distribution, 18

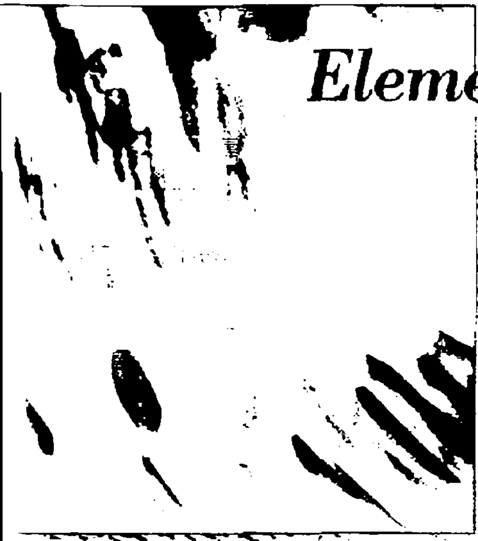
Y

Yule-Walker equation, 147

Z

zero-mean white noise, 119

Francis X. Diebold



Elements of FORECASTING

Fourth Edition

STUDY WITH ECONAPPS
Your purchase of this new textbook includes complimentary access to Thomson South-Western's Economic Applications Web site.

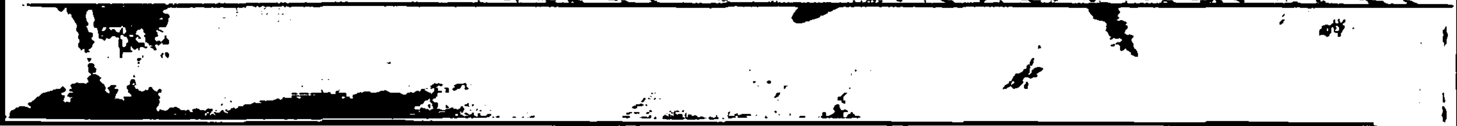
The Economic Applications Web site includes a suite of highly-acclaimed and content-rich dynamic Web features developed specifically for economics students:

- EconNews Online
- EconDebate Online
- EconData Online

Organized and searchable by key economic topic, these regularly updated features are designed to deepen your understanding of theoretical concepts through hands-on exploration and analysis of the latest economic news stories, policy debates, and data. They also provide an excellent starting point for research you may need to do for assignments in your economics classes — such as papers, projects, and presentations.

Get Started Today at http://www.thomsonedu.com. If an access card came with this book, you can start using many of these resources right away by following the directions on the card.

ABOUT THOMSON SOUTH-WESTERN
For over 100 years, Thomson South-Western has been a leader in business education, producing educational and professional products for all areas of business. Visit www.thomsonedu.com for business education resources whether you are in the classroom or in the workplace.



ISBN-13: 978-0-324-35904-6
ISBN-10: 0-324-35904-7
90000

THOMSON
SOUTH-WESTERN

Join us on the Internet
South-Western – **www.thomsonedu.com**
Economics – **www.thomsonedu.com/economics**

For your learning solutions:
www.thomson.com/learning