Time Series Analysis

James D. Hamilton

PRINCETON UNIVERSITY PRESS
PRINCETON, NEW JERSEY

Copyright © 1994 by Princeton University Press
Published by Princeton University Press, 41 William St.,
Princeton, New Jersey 08540
In the United Kingdom: Princeton University Press,
Chichester, West Sussex

All Rights Reserved

Library of Congress Cataloging-in-Publication Data
Hamilton, James D. (James Douglas), (1954–)
Time series analysis / James D. Hamilton.
p. cm.
Includes bibliographical references and indexes.
ISBN 0-691-04289-6
1. Time-series analysis. I. Title.
QA280.H264 1994
519.5'5—dc20 93-4958
CIP

This book has been composed in Times Roman.

Princeton University Press books are printed on acid-free paper and meet the guidelines for permanence and durability of the Committee on Production Guidelines for Book Longevity of the Council on Library Resources.

Printed in the United States of America

10 9 8 7 6 5 4 3 2 1

Contents

PREFACE xiii

1 *Difference Equations* *1*

1.1. First-Order Difference Equations 1
1.2. *p*th-Order Difference Equations 7

APPENDIX 1.A. ***Proofs of Chapter 1 Propositions 21***
References 24

2 *Lag Operators* *25*

2.1. Introduction 25
2.2. First-Order Difference Equations 27
2.3. Second-Order Difference Equations 29
2.4. *p*th-Order Difference Equations 33
2.5. Initial Conditions and Unbounded Sequences 36

References 42

3 *Stationary* ARMA *Processes* *43*

3.1. Expectations, Stationarity, and Ergodicity 43
3.2. White Noise 47
3.3. Moving Average Processes 48
3.4. Autoregressive Processes 53
3.5. Mixed Autoregressive Moving Average Processes 59

3.6. The Autocovariance-Generating Function 61
3.7. Invertibility 64

APPENDIX 3.A. ***Convergence Results for Infinite-Order Moving Average Processes 69***
Exercises 70 ***References 71***

4 *Forecasting* 72

4.1. Principles of Forecasting 72
4.2. Forecasts Based on an Infinite Number of Observations 77
4.3. Forecasts Based on a Finite Number of Observations 85
4.4. The Triangular Factorization of a Positive Definite Symmetric Matrix 87
4.5. Updating a Linear Projection 92
4.6. Optimal Forecasts for Gaussian Processes 100
4.7. Sums of *ARMA* Processes 102
4.8. Wold's Decomposition and the Box-Jenkins Modeling Philosophy 108

APPENDIX 4.A. ***Parallel Between* OLS *Regression and Linear Projection 113***
APPENDIX 4.B. ***Triangular Factorization of the Covariance Matrix for an* MA(1) *Process 114***
Exercises 115 ***References 116***

5 *Maximum Likelihood Estimation* 117

5.1. Introduction 117
5.2. The Likelihood Function for a Gaussian $AR(1)$ Process 118
5.3. The Likelihood Function for a Gaussian $AR(p)$ Process 123
5.4. The Likelihood Function for a Gaussian $MA(1)$ Process 127
5.5. The Likelihood Function for a Gaussian $MA(q)$ Process 130
5.6. The Likelihood Function for a Gaussian $ARMA(p, q)$ Process 132
5.7. Numerical Optimization 133

5.8. Statistical Inference with Maximum Likelihood Estimation 142
5.9. Inequality Constraints 146

APPENDIX 5.A. ***Proofs of Chapter 5 Propositions 148***
Exercises 150 ***References 150***

6 *Spectral Analysis* 152

6.1. The Population Spectrum 152
6.2. The Sample Periodogram 158
6.3. Estimating the Population Spectrum 163
6.4. Uses of Spectral Analysis 167

APPENDIX 6.A. ***Proofs of Chapter 6 Propositions 172***
Exercises 178 ***References 178***

7 *Asymptotic Distribution Theory* 180

7.1. Review of Asymptotic Distribution Theory 180
7.2. Limit Theorems for Serially Dependent Observations 186

APPENDIX 7.A. ***Proofs of Chapter 7 Propositions 195***
Exercises 198 ***References 199***

8 *Linear Regression Models* 200

8.1. Review of Ordinary Least Squares with Deterministic Regressors and i.i.d. Gaussian Disturbances 200
8.2. Ordinary Least Squares Under More General Conditions 207
8.3. Generalized Least Squares 220

APPENDIX 8.A. ***Proofs of Chapter 8 Propositions 228***
Exercises 230 ***References 231***

9 *Linear Systems of Simultaneous Equations* 233

9.1. Simultaneous Equations Bias 233
9.2. Instrumental Variables and Two-Stage Least Squares 238

9.3. Identification 243
9.4. Full-Information Maximum Likelihood Estimation 247
9.5. Estimation Based on the Reduced Form 250
9.6. Overview of Simultaneous Equations Bias 252

APPENDIX 9.A. *Proofs of Chapter 9 Proposition 253*
Exercise 255 *References 256*

10 *Covariance-Stationary Vector Processes* 257

10.1. Introduction to Vector Autoregressions 257
10.2. Autocovariances and Convergence Results for Vector Processes 261
10.3. The Autocovariance-Generating Function for Vector Processes 266
10.4. The Spectrum for Vector Processes 268
10.5. The Sample Mean of a Vector Process 279

APPENDIX 10.A. *Proofs of Chapter 10 Propositions 285*
Exercises 290 *References 290*

11 *Vector Autoregressions* 291

11.1. Maximum Likelihood Estimation and Hypothesis Testing for an Unrestricted Vector Autoregression 291
11.2. Bivariate Granger Causality Tests 302
11.3. Maximum Likelihood Estimation of Restricted Vector Autoregressions 309
11.4. The Impulse-Response Function 318
11.5. Variance Decomposition 323
11.6. Vector Autoregressions and Structural Econometric Models 324
11.7. Standard Errors for Impulse-Response Functions 336

APPENDIX 11.A. *Proofs of Chapter 11 Propositions 340*
APPENDIX 11.B. *Calculation of Analytic Derivatives 344*
Exercises 348 *References 349*

12 *Bayesian Analysis* 351

12.1. Introduction to Bayesian Analysis 351
12.2. Bayesian Analysis of Vector Autoregressions 360
12.3. Numerical Bayesian Methods 362

APPENDIX 12.A. ***Proofs of Chapter 12 Propositions 366***
Exercise 370 ***References 370***

13 *The Kalman Filter* 372

13.1. The State-Space Representation of a Dynamic System 372
13.2. Derivation of the Kalman Filter 377
13.3. Forecasts Based on the State-Space Representation 381
13.4. Maximum Likelihood Estimation of Parameters 385
13.5. The Steady-State Kalman Filter 389
13.6. Smoothing 394
13.7. Statistical Inference with the Kalman Filter 397
13.8. Time-Varying Parameters 399

APPENDIX 13.A. ***Proofs of Chapter 13 Propositions 403***
Exercises 406 ***References 407***

14 *Generalized Method of Moments* 409

14.1. Estimation by the Generalized Method of Moments 409
14.2. Examples 415
14.3. Extensions 424
14.4. *GMM* and Maximum Likelihood Estimation 427

APPENDIX 14.A. ***Proofs of Chapter 14 Propositions 431***
Exercise 432 ***References 433***

15 *Models of Nonstationary Time Series* 435

15.1. Introduction 435
15.2. Why Linear Time Trends and Unit Roots? 438

15.3. Comparison of Trend-Stationary and Unit Root Processes 438
15.4. The Meaning of Tests for Unit Roots 444
15.5. Other Approaches to Trended Time Series 447

APPENDIX 15.A. ***Derivation of Selected Equations for Chapter 15 451***
References 452

16 *Processes with Deterministic Time Trends* 454

16.1. Asymptotic Distribution of *OLS* Estimates of the Simple Time Trend Model 454
16.2. Hypothesis Testing for the Simple Time Trend Model 461
16.3. Asymptotic Inference for an Autoregressive Process Around a Deterministic Time Trend 463

APPENDIX 16.A. ***Derivation of Selected Equations for Chapter 16 472***
Exercises 474 ***References 474***

17 *Univariate Processes with Unit Roots* 475

17.1. Introduction 475
17.2. Brownian Motion 477
17.3. The Functional Central Limit Theorem 479
17.4. Asymptotic Properties of a First-Order Autoregression when the True Coefficient Is Unity 486
17.5. Asymptotic Results for Unit Root Processes with General Serial Correlation 504
17.6. Phillips-Perron Tests for Unit Roots 506
17.7. Asymptotic Properties of a *p*th-Order Autoregression and the Augmented Dickey-Fuller Tests for Unit Roots 516
17.8. Other Approaches to Testing for Unit Roots 531
17.9. Bayesian Analysis and Unit Roots 532

APPENDIX 17.A. ***Proofs of Chapter 17 Propositions 534***
Exercises 537 ***References 541***

18 *Unit Roots in Multivariate Time Series* *544*

18.1. Asymptotic Results for Nonstationary Vector Processes 544
18.2. Vector Autoregressions Containing Unit Roots 549
18.3. Spurious Regressions 557

APPENDIX 18.A. ***Proofs of Chapter 18 Propositions 562***
Exercises 568 ***References 569***

19 *Cointegration* *571*

19.1. Introduction 571
19.2. Testing the Null Hypothesis of No Cointegration 582
19.3. Testing Hypotheses About the Cointegrating Vector 601

APPENDIX 19.A. ***Proofs of Chapter 19 Propositions 618***
Exercises 625 ***References 627***

20 *Full-Information Maximum Likelihood Analysis of Cointegrated Systems* *630*

20.1. Canonical Correlation 630
20.2. Maximum Likelihood Estimation 635
20.3. Hypothesis Testing 645
20.4. Overview of Unit Roots—To Difference or Not to Difference? 651

APPENDIX 20.A. ***Proofs of Chapter 20 Propositions 653***
Exercises 655 ***References 655***

21 *Time Series Models of Heteroskedasticity* *657*

21.1. Autoregressive Conditional Heteroskedasticity (*ARCH*) 657
21.2. Extensions 665

APPENDIX 21.A. ***Derivation of Selected Equations for Chapter 21 673***
References 674

22 *Modeling Time Series with Changes in Regime* *677*

22.1. Introduction 677
22.2. Markov Chains 678
22.3. Statistical Analysis of i.i.d. Mixture Distributions 685
22.4. Time Series Models of Changes in Regime 690

APPENDIX 22.A. ***Derivation of Selected Equations for Chapter 22 699***
Exercise 702 ***References 702***

A *Mathematical Review* *704*

A.1. Trigonometry 704
A.2. Complex Numbers 708
A.3. Calculus 711
A.4. Matrix Algebra 721
A.5. Probability and Statistics 739

References 750

B *Statistical Tables* *751*

C *Answers to Selected Exercises* *769*

D *Greek Letters and Mathematical Symbols Used in the Text* *786*

AUTHOR INDEX 789

SUBJECT INDEX 792

Preface

Much of economics is concerned with modeling dynamics. There has been an explosion of research in this area in the last decade, as "time series econometrics" has practically come to be synonymous with "empirical macroeconomics."

Several texts provide good coverage of the advances in the economic analysis of dynamic systems, while others summarize the earlier literature on statistical inference for time series data. There seemed a use for a text that could integrate the theoretical and empirical issues as well as incorporate the many advances of the last decade, such as the analysis of vector autoregressions, estimation by generalized method of moments, and statistical inference for nonstationary data. This is the goal of *Time Series Analysis*.

A principal anticipated use of the book would be as a textbook for a graduate econometrics course in time series analysis. The book aims for maximum flexibility through what might be described as an integrated modular structure. As an example of this, the first three sections of Chapter 13 on the Kalman filter could be covered right after Chapter 4, if desired. Alternatively, Chapter 13 could be skipped altogether without loss of comprehension. Despite this flexibility, state-space ideas are fully integrated into the text beginning with Chapter 1, where a state-space representation is used (without any jargon or formalism) to introduce the key results concerning difference equations. Thus, when the reader encounters the formal development of the state-space framework and the Kalman filter in Chapter 13, the notation and key ideas should already be quite familiar.

Spectral analysis (Chapter 6) is another topic that could be covered at a point of the reader's choosing or skipped altogether. In this case, the integrated modular structure is achieved by the early introduction and use of autocovariance-generating functions and filters. Wherever possible, results are described in terms of these rather than the spectrum.

Although the book is designed with an econometrics couse in time series methods in mind, the book should be useful for several other purposes. It is completely self-contained, starting from basic principles accessible to first-year graduate students and including an extensive math review appendix. Thus the book would be quite suitable for a first-year graduate course in macroeconomics or dynamic methods that has no econometric content. Such a course might use Chapters 1 and 2, Sections 3.1 through 3.5, and Sections 4.1 and 4.2.

Yet another intended use for the book would be in a conventional econometrics course without an explicit time series focus. The popular econometrics texts do not have much discussion of such topics as numerical methods; asymptotic results for serially dependent, heterogeneously distributed observations; estimation of models with distributed lags; autocorrelation- and heteroskedasticity-consistent

standard errors; Bayesian analysis; or generalized method of moments. All of these topics receive extensive treatment in *Time Series Analysis*. Thus, an econometrics course without an explicit focus on time series might make use of Sections 3.1 through 3.5, Chapters 7 through 9, and Chapter 14, and perhaps any of Chapters 5, 11, and 12 as well. Again, the text is self-contained, with a fairly complete discussion of conventional simultaneous equations methods in Chapter 9. Indeed, a very important goal of the text is to develop the parallels between (1) the traditional econometric approach to simultaneous equations and (2) the current popularity of vector autoregressions and generalized method of moments estimation.

Finally, the book attempts to provide a rigorous motivation for the methods and yet still be accessible for researchers with purely applied interests. This is achieved by relegation of many details to mathematical appendixes at the ends of chapters, and by inclusion of numerous examples that illustrate exactly how the theoretical results are used and applied in practice.

The book developed out of my lectures at the University of Virginia. I am grateful first and foremost to my many students over the years whose questions and comments have shaped the course of the manuscript. I also have an enormous debt to numerous colleagues who have kindly offered many useful suggestions, and would like to thank in particular Donald W. K. Andrews, Stephen R. Blough, John Cochrane, George Davis, Michael Dotsey, Robert Engle, T. Wake Epps, Marjorie Flavin, John Geweke, Eric Ghysels, Carlo Giannini, Clive W. J. Granger, Alastair Hall, Bruce E. Hansen, Kevin Hassett, Tomoo Inoue, Ravi Jagannathan, Kenneth F. Kroner, Rocco Mosconi, Masao Ogaki, Adrian Pagan, Peter C. B. Phillips, Peter Rappoport, Glenn Rudebusch, Raul Susmel, Mark Watson, Kenneth D. West, Halbert White, and Jeffrey M. Wooldridge. I would also like to thank Pok-sang Lam and John Rogers for graciously sharing their data. Thanks also go to Keith Sill and Christopher Stomberg for assistance with the figures, to Rita Chen for assistance with the statistical tables in Appendix B, and to Richard Mickey for a superb job of copy editing.

James D. Hamilton

Time Series Analysis

1 Difference Equations

1.1. *First-Order Difference Equations*

This book is concerned with the dynamic consequences of events over time. Let's say we are studying a variable whose value at date t is denoted y_t. Suppose we are given a dynamic equation relating the value y takes on at date t to another variable w_t and to the value y took on in the previous period:

$$y_t = \phi y_{t-1} + w_t. \qquad [1.1.1]$$

Equation [1.1.1] is a *linear first-order difference equation.* A *difference equation* is an expression relating a variable y_t to its previous values. This is a *first-order* difference equation because only the first lag of the variable (y_{t-1}) appears in the equation. Note that it expresses y_t as a linear function of y_{t-1} and w_t.

An example of [1.1.1] is Goldfeld's (1973) estimated money demand function for the United States. Goldfeld's model related the log of the real money holdings of the public (m_t) to the log of aggregate real income (I_t), the log of the interest rate on bank accounts (r_{bt}), and the log of the interest rate on commercial paper (r_{ct}):

$$m_t = 0.27 + 0.72m_{t-1} + 0.19I_t - 0.045r_{bt} - 0.019r_{ct}. \qquad [1.1.2]$$

This is a special case of [1.1.1] with $y_t = m_t$, $\phi = 0.72$, and

$$w_t = 0.27 + 0.19I_t - 0.045r_{bt} - 0.019r_{ct}.$$

For purposes of analyzing the dynamics of such a system, it simplifies the algebra a little to summarize the effects of all the input variables (I_t, r_{bt}, and r_{ct}) in terms of a scalar w_t as here.

In Chapter 3 the input variable w_t will be regarded as a random variable, and the implications of [1.1.1] for the statistical properties of the output series y_t will be explored. In preparation for this discussion, it is necessary first to understand the mechanics of difference equations. For the discussion in Chapters 1 and 2, the values for the input variable $\{w_1, w_2, \ldots\}$ will simply be regarded as a sequence of deterministic numbers. Our goal is to answer the following question: If a dynamic system is described by [1.1.1], what are the effects on y of changes in the value of w?

Solving a Difference Equation by Recursive Substitution

The presumption is that the dynamic equation [1.1.1] governs the behavior of y for all dates t. Thus, for each date we have an equation relating the value of

y for that date to its previous value and the current value of w:

Date	*Equation*	
0	$y_0 = \phi y_{-1} + w_0$	[1.1.3]
1	$y_1 = \phi y_0 + w_1$	[1.1.4]
2	$y_2 = \phi y_1 + w_2$	[1.1.5]
$\vdots$	$\vdots$	
t	$y_t = \phi y_{t-1} + w_t.$	[1.1.6]

If we know the starting value of y for date $t = -1$ and the value of w for dates $t = 0, 1, 2, \ldots$, then it is possible to simulate this dynamic system to find the value of y for any date. For example, if we know the value of y for $t = -1$ and the value of w for $t = 0$, we can calculate the value of y for $t = 0$ directly from [1.1.3]. Given this value of y_0 and the value of w for $t = 1$, we can calculate the value of y for $t = 1$ from [1.1.4]:

$$y_1 = \phi y_0 + w_1 = \phi(\phi y_{-1} + w_0) + w_1,$$

or

$$y_1 = \phi^2 y_{-1} + \phi w_0 + w_1.$$

Given this value of y_1 and the value of w for $t = 2$, we can calculate the value of y for $t = 2$ from [1.1.5]:

$$y_2 = \phi y_1 + w_2 = \phi(\phi^2 y_{-1} + \phi w_0 + w_1) + w_2,$$

or

$$y_2 = \phi^3 y_{-1} + \phi^2 w_0 + \phi w_1 + w_2.$$

Continuing recursively in this fashion, the value that y takes on at date t can be described as a function of its initial value y_{-1} and the history of w between date 0 and date t:

$$y_t = \phi^{t+1} y_{-1} + \phi^t w_0 + \phi^{t-1} w_1 + \phi^{t-2} w_2 + \cdots + \phi w_{t-1} + w_t. \quad [1.1.7]$$

This procedure is known as solving the difference equation [1.1.1] by *recursive substitution*.

Dynamic Multipliers

Note that [1.1.7] expresses y_t as a linear function of the initial value y_{-1} and the historical values of w. This makes it very easy to calculate the effect of w_0 on y_t. If w_0 were to change with y_{-1} and $w_1, w_2, \ldots, w_t$ taken as unaffected, the effect on y_t would be given by

$$\frac{\partial y_t}{\partial w_0} = \phi^t. \quad [1.1.8]$$

Note that the calculations would be exactly the same if the dynamic simulation were started at date t (taking y_{t-1} as given); then y_{t+j} could be described as a

function of y_{t-1} and $w_t, w_{t+1}, \ldots, w_{t+j}$:

$$y_{t+j} = \phi^{j+1}y_{t-1} + \phi^j w_t + \phi^{j-1}w_{t+1} + \phi^{j-2}w_{t+2} + \cdots + \phi w_{t+j-1} + w_{t+j}. \qquad [1.1.9]$$

The effect of w_t on y_{t+j} is given by

$$\frac{\partial y_{t+j}}{\partial w_t} = \phi^j. \qquad [1.1.10]$$

Thus the *dynamic multiplier* [1.1.10] depends only on j, the length of time separating the disturbance to the input (w_t) and the observed value of the output (y_{t+j}). The multiplier does not depend on t; that is, it does not depend on the dates of the observations themselves. This is true of any linear difference equation.

As an example of calculating a dynamic multiplier, consider again Goldfeld's money demand specification [1.1.2]. Suppose we want to know what will happen to money demand two quarters from now if current income I_t were to increase by one unit today with future income I_{t+1} and I_{t+2} unaffected:

$$\frac{\partial m_{t+2}}{\partial I_t} = \frac{\partial m_{t+2}}{\partial w_t} \times \frac{\partial w_t}{\partial I_t} = \phi^2 \times \frac{\partial w_t}{\partial I_t}.$$

From [1.1.2], a one-unit increase in I_t will increase w_t by 0.19 units, meaning that $\partial w_t/\partial I_t = 0.19$. Since $\phi = 0.72$, we calculate

$$\frac{\partial m_{t+2}}{\partial I_t} = (0.72)^2(0.19) = 0.098.$$

Because I_t is the log of income, an increase in I_t of 0.01 units corresponds to a 1% increase in income. An increase in m_t of $(0.01)\cdot(0.098) \approx 0.001$ corresponds to a 0.1% increase in money holdings. Thus the public would be expected to increase its money holdings by a little less than 0.1% two quarters following a 1% increase in income.

Different values of ϕ in [1.1.1] can produce a variety of dynamic responses of y to w. If $0 < \phi < 1$, the multiplier $\partial y_{t+j}/\partial w_t$ in [1.1.10] decays geometrically toward zero. Panel (a) of Figure 1.1 plots ϕ^j as a function of j for $\phi = 0.8$. If $-1 < \phi < 0$, the multiplier $\partial y_{t+j}/\partial w_t$ will alternate in sign as in panel (b). In this case an increase in w_t will cause y_t to be higher, y_{t+1} to be lower, y_{t+2} to be higher, and so on. Again the absolute value of the effect decays geometrically toward zero. If $\phi > 1$, the dynamic multiplier increases exponentially over time as in panel (c). A given increase in w_t has a larger effect the farther into the future one goes. For $\phi < -1$, the system [1.1.1] exhibits explosive oscillation as in panel (d).

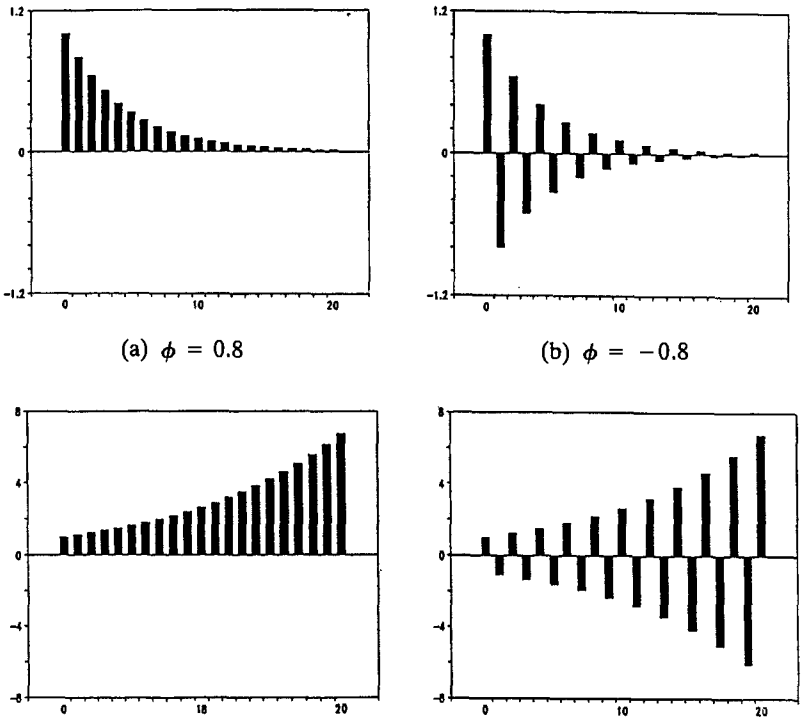
Thus, if $|\phi| < 1$, the system is stable; the consequences of a given change in w_t will eventually die out. If $|\phi| > 1$, the system is explosive. An interesting possibility is the borderline case, $\phi = 1$. In this case, the solution [1.1.9] becomes

$$y_{t+j} = y_{t-1} + w_t + w_{t+1} + w_{t+2} + \cdots + w_{t+j-1} + w_{t+j}. \qquad [1.1.11]$$

Here the output variable y is the sum of the historical inputs w. A one-unit increase in w will cause a permanent one-unit increase in y:

$$\frac{\partial y_{t+j}}{\partial w_t} = 1 \qquad \text{for } j = 0, 1, \ldots.$$

We might also be interested in the effect of w on the present value of the stream of future realizations of y. For a given stream of future values $y_t, y_{t+1},$



(a) $\phi = 0.8$ (b) $\phi = -0.8$

(c) $\phi = 1.1$ (d) $\phi = -1.1$

FIGURE 1.1 Dynamic multiplier for first-order difference equation for different values of ϕ (plot of $\partial y_{t+j}/\partial w_t = \phi^j$ as a function of the lag j).

$y_{t+2}, \ldots$ and a constant interest rate[1] $r > 0$, the *present value* of the stream at time t is given by

$$y_t + \frac{y_{t+1}}{1 + r} + \frac{y_{t+2}}{(1 + r)^2} + \frac{y_{t+3}}{(1 + r)^3} + \cdots. \qquad [1.1.12]$$

Let β denote the discount factor:

$$\beta \equiv 1/(1 + r).$$

Note that $0 < \beta < 1$. Then the present value [1.1.12] can be written as

$$\sum_{j=0}^{\infty} \beta^j y_{t+j}. \qquad [1.1.13]$$

Consider what would happen if there were a one-unit increase in w_t with $w_{t+1}, w_{t+2}, \ldots$ unaffected. The consequences of this change for the present value of y are found by differentiating [1.1.13] with respect to w_t and then using [1.1.10]

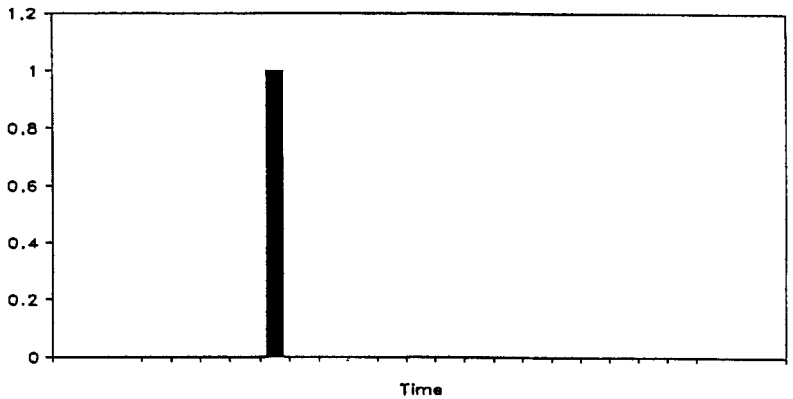
[1]The interest rate is measured here as a fraction of 1; thus $r = 0.1$ corresponds to a 10% interest rate.

to evaluate each derivative:

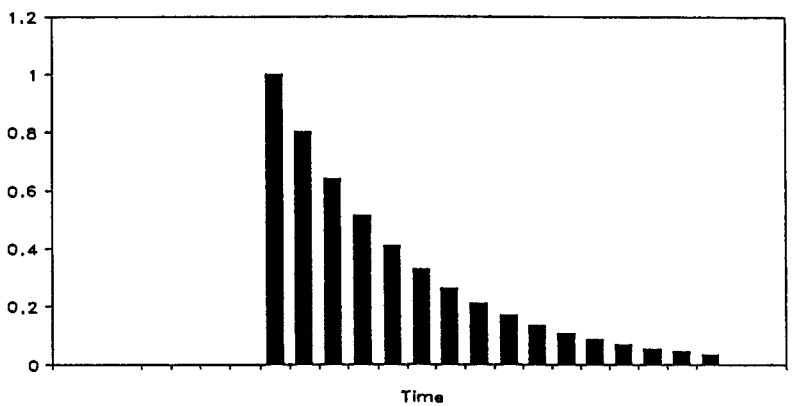
$$\sum_{j=0}^{\infty} \beta^j \frac{\partial y_{t+j}}{\partial w_t} = \sum_{j=0}^{\infty} \beta^j \phi^j = 1/(1 - \beta\phi), \qquad [1.1.14]$$

provided that $|\beta\phi| < 1$.

In calculating the dynamic multipliers [1.1.10] or [1.1.14], we were asking what would happen if w_t were to increase by one unit with $w_{t+1}, w_{t+2}, \ldots, w_{t+j}$ unaffected. We were thus finding the effect of a purely transitory change in w. Panel (a) of Figure 1.2 shows the time path of w associated with this question, and panel (b) shows the implied path for y. Because the dynamic multiplier [1.1.10] calculates the response of y to a single impulse in w, it is also referred to as the *impulse-response function*.



(a) Value of w



(b) Value of y

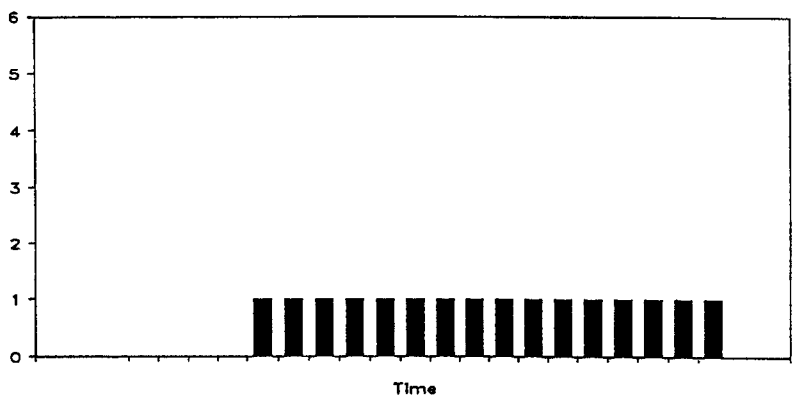
FIGURE 1.2 Paths of input variable (w_t) and output variable (y_t) assumed for dynamic multiplier and present-value calculations.

Sometimes we might instead be interested in the consequences of a permanent change in w. A permanent change in w means that $w_t, w_{t+1}, \ldots$, and w_{t+j} would all increase by one unit, as in Figure 1.3. From formula [1.1.10], the effect on y_{t+j} of a permanent change in w beginning in period t is given by

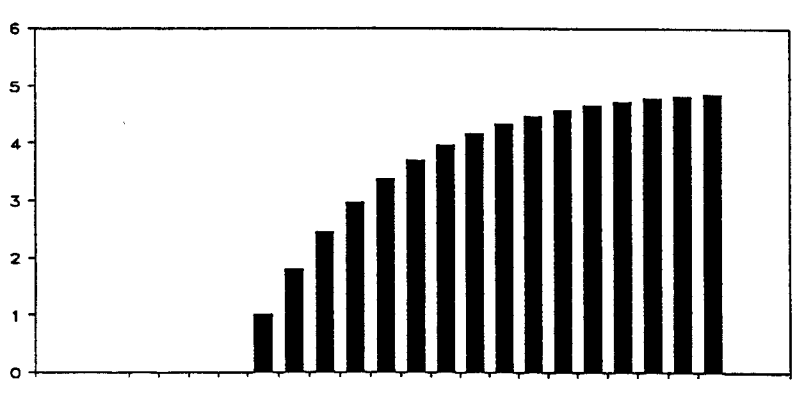
$$\frac{\partial y_{t+j}}{\partial w_t} + \frac{\partial y_{t+j}}{\partial w_{t+1}} + \frac{\partial y_{t+j}}{\partial w_{t+2}} + \cdots + \frac{\partial y_{t+j}}{\partial w_{t+j}} = \phi^j + \phi^{j-1} + \phi^{j-2} + \cdots + \phi + 1.$$

When $|\phi| < 1$, the limit of this expression as j goes to infinity is sometimes described as the "long-run" effect of w on y:

$$\begin{aligned}\lim_{j\to\infty}\left[\frac{\partial y_{t+j}}{\partial w_t} + \frac{\partial y_{t+j}}{\partial w_{t+1}} + \frac{\partial y_{t+j}}{\partial w_{t+2}} + \cdots + \frac{\partial y_{t+j}}{\partial w_{t+j}}\right] &= 1 + \phi + \phi^2 + \cdots \\ &= 1/(1 - \phi).\end{aligned} \qquad [1.1.15]$$



(a) Value of w



(b) Value of y

FIGURE 1.3 Paths of input variable (w_t) and output variable (y_t) assumed for long-run effect calculations.

For example, the long-run income elasticity of money demand in the system [1.1.2] is given by

$$\frac{0.19}{1 - 0.72} = 0.68.$$

A permanent 1% increase in income will eventually lead to a 0.68% increase in money demand.

Another related question concerns the cumulative consequences for y of a one-time change in w. Here we consider a transitory disturbance to w as in panel (a) of Figure 1.2, but wish to calculate the sum of the consequences for all future values of y. Another way to think of this is as the effect on the present value of y [1.1.13] with the discount rate $\beta = 1$. Setting $\beta = 1$ in [1.1.14] shows this cumulative effect to be equal to

$$\sum_{j=0}^{\infty} \frac{\partial y_{t+j}}{\partial w_t} = 1/(1 - \phi), \qquad [1.1.16]$$

provided that $|\phi| < 1$. Note that the cumulative effect on y of a transitory change in w (expression [1.1.16]) is the same as the long-run effect on y of a permanent change in w (expression [1.1.15]).

1.2. *pth-Order Difference Equations*

Let us now generalize the dynamic system [1.1.1] by allowing the value of y at date t to depend on p of its own lags along with the current value of the input variable w_t:

$$y_t = \phi_1 y_{t-1} + \phi_2 y_{t-2} + \cdots + \phi_p y_{t-p} + w_t. \qquad [1.2.1]$$

Equation [1.2.1] is a linear pth-order difference equation.

It is often convenient to rewrite the pth-order difference equation [1.2.1] in the scalar y_t as a first-order difference equation in a vector $\boldsymbol{\xi}_t$. Define the $(p \times 1)$ vector $\boldsymbol{\xi}_t$ by

$$\boldsymbol{\xi}_t \equiv \begin{bmatrix} y_t \\ y_{t-1} \\ y_{t-2} \\ \vdots \\ y_{t-p+1} \end{bmatrix}. \qquad [1.2.2]$$

That is, the first element of the vector $\boldsymbol{\xi}$ at date t is the value y took on at date t. The second element of $\boldsymbol{\xi}_t$ is the value y took on at date $t - 1$, and so on. Define the $(p \times p)$ matrix $\mathbf{F}$ by

$$\mathbf{F} \equiv \begin{bmatrix} \phi_1 & \phi_2 & \phi_3 & \cdots & \phi_{p-1} & \phi_p \\ 1 & 0 & 0 & \cdots & 0 & 0 \\ 0 & 1 & 0 & \cdots & 0 & 0 \\ \vdots & \vdots & \vdots & \cdots & \vdots & \vdots \\ 0 & 0 & 0 & \cdots & 1 & 0 \end{bmatrix}. \qquad [1.2.3]$$

For example, for $p = 4$, $\mathbf{F}$ refers to the following 4×4 matrix:

$$\mathbf{F} = \begin{bmatrix} \phi_1 & \phi_2 & \phi_3 & \phi_4 \\ 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \end{bmatrix}.$$

For $p = 1$ (the first-order difference equation [1.1.1]), $\mathbf{F}$ is just the scalar ϕ. Finally, define the $(p \times 1)$ vector $\mathbf{v}_t$ by

$$\mathbf{v}_t \equiv \begin{bmatrix} w_t \\ 0 \\ 0 \\ \vdots \\ 0 \end{bmatrix}. \qquad [1.2.4]$$

Consider the following first-order vector difference equation:

$$\boldsymbol{\xi}_t = \mathbf{F}\boldsymbol{\xi}_{t-1} + \mathbf{v}_t, \qquad [1.2.5]$$

or

$$\begin{bmatrix} y_t \\ y_{t-1} \\ y_{t-2} \\ \vdots \\ y_{t-p+1} \end{bmatrix} = \begin{bmatrix} \phi_1 & \phi_2 & \phi_3 & \cdots & \phi_{p-1} & \phi_p \\ 1 & 0 & 0 & \cdots & 0 & 0 \\ 0 & 1 & 0 & \cdots & 0 & 0 \\ \vdots & \vdots & \vdots & \cdots & \vdots & \vdots \\ 0 & 0 & 0 & \cdots & 1 & 0 \end{bmatrix} \begin{bmatrix} y_{t-1} \\ y_{t-2} \\ y_{t-3} \\ \vdots \\ y_{t-p} \end{bmatrix} + \begin{bmatrix} w_t \\ 0 \\ 0 \\ \vdots \\ 0 \end{bmatrix}.$$

This is a system of p equations. The first equation in this system is identical to equation [1.2.1]. The second equation is simply the identity

$$y_{t-1} = y_{t-1},$$

owing to the fact that the second element of $\boldsymbol{\xi}_t$ is the same as the first element of $\boldsymbol{\xi}_{t-1}$. The third equation in [1.2.5] states that $y_{t-2} = y_{t-2}$; the pth equation states that $y_{t-p+1} = y_{t-p+1}$.

Thus, the first-order vector system [1.2.5] is simply an alternative representation of the pth-order scalar system [1.2.1]. The advantage of rewriting the pth-order system [1.2.1] in the form of a first-order system [1.2.5] is that first-order systems are often easier to work with than pth-order systems.

A dynamic multiplier for [1.2.5] can be found in exactly the same way as was done for the first-order scalar system of Section 1.1. If we knew the value of the vector $\boldsymbol{\xi}$ for date $t = -1$ and of $\mathbf{v}$ for date $t = 0$, we could find the value of $\boldsymbol{\xi}$ for date 0 from

$$\boldsymbol{\xi}_0 = \mathbf{F}\boldsymbol{\xi}_{-1} + \mathbf{v}_0.$$

The value of $\boldsymbol{\xi}$ for date 1 is

$$\boldsymbol{\xi}_1 = \mathbf{F}\boldsymbol{\xi}_0 + \mathbf{v}_1 = \mathbf{F}(\mathbf{F}\boldsymbol{\xi}_{-1} + \mathbf{v}_0) + \mathbf{v}_1 = \mathbf{F}^2\boldsymbol{\xi}_{-1} + \mathbf{F}\mathbf{v}_0 + \mathbf{v}_1.$$

Proceeding recursively in this fashion produces a generalization of [1.1.7]:

$$\boldsymbol{\xi}_t = \mathbf{F}^{t+1}\boldsymbol{\xi}_{-1} + \mathbf{F}^t\mathbf{v}_0 + \mathbf{F}^{t-1}\mathbf{v}_1 + \mathbf{F}^{t-2}\mathbf{v}_2 + \cdots + \mathbf{F}\mathbf{v}_{t-1} + \mathbf{v}_t. \qquad [1.2.6]$$

Writing this out in terms of the definitions of ξ and v,

$$\begin{bmatrix} y_t \\ y_{t-1} \\ y_{t-2} \\ \vdots \\ y_{t-p+1} \end{bmatrix} = \mathbf{F}^{t+1}\begin{bmatrix} y_{-1} \\ y_{-2} \\ y_{-3} \\ \vdots \\ y_{-p} \end{bmatrix} + \mathbf{F}^{t}\begin{bmatrix} w_0 \\ 0 \\ 0 \\ \vdots \\ 0 \end{bmatrix} + \mathbf{F}^{t-1}\begin{bmatrix} w_1 \\ 0 \\ 0 \\ \vdots \\ 0 \end{bmatrix} + \cdots$$
$$+ \mathbf{F}^{1}\begin{bmatrix} w_{t-1} \\ 0 \\ 0 \\ \vdots \\ 0 \end{bmatrix} + \begin{bmatrix} w_t \\ 0 \\ 0 \\ \vdots \\ 0 \end{bmatrix}. \qquad [1.2.7]$$

Consider the first equation of this system, which characterizes the value of y_t. Let $f_{11}^{(t)}$ denote the (1, 1) element of $\mathbf{F}^t$, $f_{12}^{(t)}$ the (1, 2) element of $\mathbf{F}^t$, and so on. Then the first equation of [1.2.7] states that

$$y_t = f_{11}^{(t+1)}y_{-1} + f_{12}^{(t+1)}y_{-2} + \cdots + f_{1p}^{(t+1)}y_{-p} + f_{11}^{(t)}w_0 + f_{11}^{(t-1)}w_1 + \cdots + f_{11}^{(1)}w_{t-1} + w_t. \qquad [1.2.8]$$

This describes the value of y at date t as a linear function of p initial values of y ($y_{-1}, y_{-2}, \ldots, y_{-p}$) and the history of the input variable w since time 0 ($w_0, w_1, \ldots, w_t$). Note that whereas only one initial value for y (the value y_{-1}) was needed in the case of a first-order difference equation, p initial values for y (the values $y_{-1}, y_{-2}, \ldots, y_{-p}$) are needed in the case of a pth-order difference equation.

The obvious generalization of [1.1.9] is

$$\xi_{t+j} = \mathbf{F}^{j+1}\xi_{t-1} + \mathbf{F}^{j}\mathbf{v}_t + \mathbf{F}^{j-1}\mathbf{v}_{t+1} + \mathbf{F}^{j-2}\mathbf{v}_{t+2} + \cdots + \mathbf{F}\mathbf{v}_{t+j-1} + \mathbf{v}_{t+j} \qquad [1.2.9]$$

from which

$$y_{t+j} = f_{11}^{(j+1)}y_{t-1} + f_{12}^{(j+1)}y_{t-2} + \cdots + f_{1p}^{(j+1)}y_{t-p} + f_{11}^{(j)}w_t + f_{11}^{(j-1)}w_{t+1} + f_{11}^{(j-2)}w_{t+2} + \cdots + f_{11}^{(1)}w_{t+j-1} + w_{t+j}. \qquad [1.2.10]$$

Thus, for a pth-order difference equation, the dynamic multiplier is given by

$$\frac{\partial y_{t+j}}{\partial w_t} = f_{11}^{(j)} \qquad [1.2.11]$$

where $f_{11}^{(j)}$ denotes the (1, 1) element of $\mathbf{F}^j$. For $j = 1$, this is simply the (1, 1) element of $\mathbf{F}$, or the parameter ϕ_1. Thus, for any pth-order system, the effect on y_{t+1} of a one-unit increase in w_t is given by the coefficient relating y_t to y_{t-1} in equation [1.2.1]:

$$\frac{\partial y_{t+1}}{\partial w_t} = \phi_1.$$

Direct multiplication of [1.2.3] reveals that the (1, 1) element of $\mathbf{F}^2$ is $(\phi_1^2 + \phi_2)$, so

$$\frac{\partial y_{t+2}}{\partial w_t} = \phi_1^2 + \phi_2$$

in a pth-order system.

For larger values of j, an easy way to obtain a numerical value for the dynamic multiplier $\partial y_{t+j}/\partial w_t$ is to simulate the system. This is done as follows. Set $y_{-1} = y_{-2} = \cdots = y_{-p} = 0$, $w_0 = 1$, and set the value of w for all other dates to 0. Then use [1.2.1] to calculate the value of y_t for $t = 0$ (namely, $y_0 = 1$). Next substitute this value along with $y_{t-1}, y_{t-2}, \ldots, y_{t-p+1}$ back into [1.2.1] to calculate y_{t+1}, and continue recursively in this fashion. The value of y at step t gives the effect of a one-unit change in w_0 on y_t.

Although numerical simulation may be adequate for many circumstances, it is also useful to have a simple analytical characterization of $\partial y_{t+j}/\partial w_t$, which, we know from [1.2.11], is given by the (1, 1) element of $\mathbf{F}^j$. This is fairly easy to obtain in terms of the eigenvalues of the matrix $\mathbf{F}$. Recall that the eigenvalues of a matrix $\mathbf{F}$ are those numbers λ for which

$$|\mathbf{F} - \lambda \mathbf{I}_p| = 0. \qquad [1.2.12]$$

For example, for $p = 2$ the eigenvalues are the solutions to

$$\left| \begin{bmatrix} \phi_1 & \phi_2 \\ 1 & 0 \end{bmatrix} - \begin{bmatrix} \lambda & 0 \\ 0 & \lambda \end{bmatrix} \right| = 0$$

or

$$\begin{vmatrix} (\phi_1 - \lambda) & \phi_2 \\ 1 & -\lambda \end{vmatrix} = \lambda^2 - \phi_1\lambda - \phi_2 = 0. \qquad [1.2.13]$$

The two eigenvalues of $\mathbf{F}$ for a second-order difference equation are thus given by

$$\lambda_1 = \frac{\phi_1 + \sqrt{\phi_1^2 + 4\phi_2}}{2} \qquad [1.2.14]$$

$$\lambda_2 = \frac{\phi_1 - \sqrt{\phi_1^2 + 4\phi_2}}{2}. \qquad [1.2.15]$$

For a general pth-order system, the determinant in [1.2.12] is a pth-order polynomial in λ whose p solutions characterize the p eigenvalues of $\mathbf{F}$. This polynomial turns out to take a very similar form to [1.2.13]. The following result is proved in Appendix 1.A at the end of this chapter.

Proposition 1.1: *The eigenvalues of the matrix* $\mathbf{F}$ *defined in equation* [*1.2.3*] *are the values of* λ *that satisfy*

$$\lambda^p - \phi_1\lambda^{p-1} - \phi_2\lambda^{p-2} - \cdots - \phi_{p-1}\lambda - \phi_p = 0. \qquad [1.2.16]$$

Once we know the eigenvalues, it is straightforward to characterize the dynamic behavior of the system. First we consider the case when the eigenvalues of $\mathbf{F}$ are distinct; for example, we require that λ_1 and λ_2 in [1.2.14] and [1.2.15] be different numbers.

General Solution of a pth-Order Difference Equation with Distinct Eigenvalues

Recall[2] that if the eigenvalues of a $(p \times p)$ matrix $\mathbf{F}$ are distinct, there exists a nonsingular $(p \times p)$ matrix $\mathbf{T}$ such that

$$\mathbf{F} = \mathbf{T}\boldsymbol{\Lambda}\mathbf{T}^{-1} \qquad [1.2.17]$$

where $\boldsymbol{\Lambda}$ is a $(p \times p)$ matrix with the eigenvalues of $\mathbf{F}$ along the principal diagonal and zeros elsewhere:

$$\boldsymbol{\Lambda} = \begin{bmatrix} \lambda_1 & 0 & 0 & \cdots & 0 \\ 0 & \lambda_2 & 0 & \cdots & 0 \\ \vdots & \vdots & \vdots & \cdots & \vdots \\ 0 & 0 & 0 & \cdots & \lambda_p \end{bmatrix}. \qquad [1.2.18]$$

This enables us to characterize the dynamic multiplier (the (1, 1) element of $\mathbf{F}^j$ in [1.2.11]) very easily. For example, from [1.2.17] we can write $\mathbf{F}^2$ as

$$\begin{aligned} \mathbf{F}^2 &= \mathbf{T}\boldsymbol{\Lambda}\mathbf{T}^{-1} \times \mathbf{T}\boldsymbol{\Lambda}\mathbf{T}^{-1} \\ &= \mathbf{T} \times \boldsymbol{\Lambda} \times (\mathbf{T}^{-1}\mathbf{T}) \times \boldsymbol{\Lambda} \times \mathbf{T}^{-1} \\ &= \mathbf{T} \times \boldsymbol{\Lambda} \times \mathbf{I}_p \times \boldsymbol{\Lambda} \times \mathbf{T}^{-1} \\ &= \mathbf{T}\boldsymbol{\Lambda}^2\mathbf{T}^{-1}. \end{aligned}$$

The diagonal structure of $\boldsymbol{\Lambda}$ implies that $\boldsymbol{\Lambda}^2$ is also a diagonal matrix whose elements are the squares of the eigenvalues of $\mathbf{F}$:

$$\boldsymbol{\Lambda}^2 = \begin{bmatrix} \lambda_1^2 & 0 & 0 & \cdots & 0 \\ 0 & \lambda_2^2 & 0 & \cdots & 0 \\ \vdots & \vdots & \vdots & \cdots & \vdots \\ 0 & 0 & 0 & \cdots & \lambda_p^2 \end{bmatrix}.$$

More generally, we can characterize $\mathbf{F}^j$ in terms of the eigenvalues of $\mathbf{F}$ as

$$\begin{aligned} \mathbf{F}^j &= \underbrace{\mathbf{T}\boldsymbol{\Lambda}\mathbf{T}^{-1} \times \mathbf{T}\boldsymbol{\Lambda}\mathbf{T}^{-1} \times \cdots \times \mathbf{T}\boldsymbol{\Lambda}\mathbf{T}^{-1}}_{j \text{ terms}} \\ &= \mathbf{T} \times \boldsymbol{\Lambda} \times (\mathbf{T}^{-1}\mathbf{T}) \times \boldsymbol{\Lambda} \times (\mathbf{T}^{-1}\mathbf{T}) \times \cdots \times \boldsymbol{\Lambda} \times \mathbf{T}^{-1}, \end{aligned}$$

which simplifies to

$$\mathbf{F}^j = \mathbf{T}\boldsymbol{\Lambda}^j\mathbf{T}^{-1} \qquad [1.2.19]$$

where

$$\boldsymbol{\Lambda}^j = \begin{bmatrix} \lambda_1^j & 0 & 0 & \cdots & 0 \\ 0 & \lambda_2^j & 0 & \cdots & 0 \\ \vdots & \vdots & \vdots & \cdots & \vdots \\ 0 & 0 & 0 & \cdots & \lambda_p^j \end{bmatrix}.$$

[2]See equation [A.4.24] in the Mathematical Review (Appendix A) at the end of the book.

Let t_{ij} denote the row i, column j element of $\mathbf{T}$ and let t^{ij} denote the row i, column j element of $\mathbf{T}^{-1}$. Equation [1.2.19] written out explicitly becomes

$$\mathbf{F}^j = \begin{bmatrix} t_{11} & t_{12} & \cdots & t_{1p} \\ t_{21} & t_{22} & \cdots & t_{2p} \\ \vdots & \vdots & \cdots & \vdots \\ t_{p1} & t_{p2} & \cdots & t_{pp} \end{bmatrix} \begin{bmatrix} \lambda_1^j & 0 & 0 & \cdots & 0 \\ 0 & \lambda_2^j & 0 & \cdots & 0 \\ \vdots & \vdots & \vdots & \cdots & \vdots \\ 0 & 0 & 0 & \cdots & \lambda_p^j \end{bmatrix} \begin{bmatrix} t^{11} & t^{12} & \cdots & t^{1p} \\ t^{21} & t^{22} & \cdots & t^{2p} \\ \vdots & \vdots & \cdots & \vdots \\ t^{p1} & t^{p2} & \cdots & t^{pp} \end{bmatrix}$$

$$= \begin{bmatrix} t_{11}\lambda_1^j & t_{12}\lambda_2^j & \cdots & t_{1p}\lambda_p^j \\ t_{21}\lambda_1^j & t_{22}\lambda_2^j & \cdots & t_{2p}\lambda_p^j \\ \vdots & \vdots & \cdots & \vdots \\ t_{p1}\lambda_1^j & t_{p2}\lambda_2^j & \cdots & t_{pp}\lambda_p^j \end{bmatrix} \begin{bmatrix} t^{11} & t^{12} & \cdots & t^{1p} \\ t^{21} & t^{22} & \cdots & t^{2p} \\ \vdots & \vdots & \cdots & \vdots \\ t^{p1} & t^{p2} & \cdots & t^{pp} \end{bmatrix}$$

from which the (1, 1) element of $\mathbf{F}^j$ is given by

$$f_{11}^{(j)} = [t_{11}t^{11}]\lambda_1^j + [t_{12}t^{21}]\lambda_2^j + \cdots + [t_{1p}t^{p1}]\lambda_p^j$$

or

$$f_{11}^{(j)} = c_1\lambda_1^j + c_2\lambda_2^j + \cdots + c_p\lambda_p^j \qquad [1.2.20]$$

where

$$c_i = [t_{1i}t^{i1}]. \qquad [1.2.21]$$

Note that the sum of the c_i terms has the following interpretation:

$$c_1 + c_2 + \cdots + c_p = [t_{11}t^{11}] + [t_{12}t^{21}] + \cdots + [t_{1p}t^{p1}], \qquad [1.2.22]$$

which is the (1, 1) element of $\mathbf{T}\cdot\mathbf{T}^{-1}$. Since $\mathbf{T}\cdot\mathbf{T}^{-1}$ is just the $(p \times p)$ identity matrix, [1.2.22] implies that the c_i terms sum to unity:

$$c_1 + c_2 + \cdots + c_p = 1. \qquad [1.2.23]$$

Substituting [1.2.20] into [1.2.11] gives the form of the dynamic multiplier for a pth-order difference equation:

$$\frac{\partial y_{t+j}}{\partial w_t} = c_1\lambda_1^j + c_2\lambda_2^j + \cdots + c_p\lambda_p^j. \qquad [1.2.24]$$

Equation [1.2.24] characterizes the dynamic multiplier as a weighted average of each of the p eigenvalues raised to the jth power.

The following result provides a closed-form expression for the constants $(c_1, c_2, \ldots, c_p)$.

Proposition 1.2: *If the eigenvalues $(\lambda_1, \lambda_2, \ldots, \lambda_p)$ of the matrix $\mathbf{F}$ in [1.2.3] are distinct, then the magnitude c_i in [1.2.21] can be written*

$$c_i = \frac{\lambda_i^{p-1}}{\prod_{\substack{k=1 \\ k \neq i}}^{p} (\lambda_i - \lambda_k)}. \qquad [1.2.25]$$

To summarize, the pth-order difference equation [1.2.1] implies that

$$\begin{aligned} y_{t+j} = {} & f_{11}^{(j+1)}y_{t-1} + f_{12}^{(j+1)}y_{t-2} + \cdots + f_{1p}^{(j+1)}y_{t-p} \\ & + w_{t+j} + \psi_1 w_{t+j-1} + \psi_2 w_{t+j-2} + \cdots + \psi_{j-1}w_{t+1} + \psi_j w_t. \end{aligned} \qquad [1.2.26]$$

The dynamic multiplier

$$\frac{\partial y_{t+j}}{\partial w_t} = \psi_j \qquad [1.2.27]$$

is given by the (1, 1) element of $\mathbf{F}^j$:

$$\psi_j = f_{11}^{(j)}. \qquad [1.2.28]$$

A closed-form expression for ψ_j can be obtained by finding the eigenvalues of $\mathbf{F}$, or the values of λ satisfying [1.2.16]. Denoting these p values by $(\lambda_1, \lambda_2, \ldots, \lambda_p)$ and assuming them to be distinct, the dynamic multiplier is given by

$$\psi_j = c_1\lambda_1^j + c_2\lambda_2^j + \cdots + c_p\lambda_p^j \qquad [1.2.29]$$

where $(c_1, c_2, \ldots, c_p)$ is a set of constants summing to unity given by expression [1.2.25].

For a first-order system ($p = 1$), this rule would have us solve [1.2.16],

$$\lambda - \phi_1 = 0,$$

which has the single solution

$$\lambda_1 = \phi_1. \qquad [1.2.30]$$

According to [1.2.29], the dynamic multiplier is given by

$$\frac{\partial y_{t+j}}{\partial w_t} = c_1\lambda_1^j. \qquad [1.2.31]$$

From [1.2.23], $c_1 = 1$. Substituting this and [1.2.30] into [1.2.31] gives

$$\frac{\partial y_{t+j}}{\partial w_t} = \phi_1^j,$$

or the same result found in Section 1.1.

For higher-order systems, [1.2.29] allows a variety of more complicated dynamics. Suppose first that all the eigenvalues of $\mathbf{F}$ (or solutions to [1.2.16]) are real. This would be the case, for example, if $p = 2$ and $\phi_1^2 + 4\phi_2 > 0$ in the solutions [1.2.14] and [1.2.15] for the second-order system. If, furthermore, all of the eigenvalues are less than 1 in absolute value, then the system is stable, and its dynamics are represented as a weighted average of decaying exponentials or decaying exponentials oscillating in sign. For example, consider the following second-order difference equation:

$$y_t = 0.6y_{t-1} + 0.2y_{t-2} + w_t.$$

From equations [1.2.14] and [1.2.15], the eigenvalues of this system are given by

$$\lambda_1 = \frac{0.6 + \sqrt{(0.6)^2 + 4(0.2)}}{2} = 0.84$$

$$\lambda_2 = \frac{0.6 - \sqrt{(0.6)^2 + 4(0.2)}}{2} = -0.24.$$

From [1.2.25], we have

$$c_1 = \lambda_1/(\lambda_1 - \lambda_2) = 0.778$$

$$c_2 = \lambda_2/(\lambda_2 - \lambda_1) = 0.222.$$

The dynamic multiplier for this system,

$$\frac{\partial y_{t+j}}{\partial w_t} = c_1\lambda_1^j + c_2\lambda_2^j,$$

is plotted as a function of j in panel (a) of Figure 1.4.[3] Note that as j becomes larger, the pattern is dominated by the larger eigenvalue (λ_1), approximating a simple geometric decay at rate λ_1.

If the eigenvalues (the solutions to [1.2.16]) are real but at least one is greater than unity in absolute value, the system is explosive. If λ_1 denotes the eigenvalue that is largest in absolute value, the dynamic multiplier is eventually dominated by an exponential function of that eigenvalue:

$$\lim_{j\to\infty} \frac{\partial y_{t+j}}{\partial w_t} \cdot \frac{1}{\lambda_1^j} = c_1.$$

Other interesting possibilities arise if some of the eigenvalues are complex. Whenever this is the case, they appear as complex conjugates. For example, if $p = 2$ and $\phi_1^2 + 4\phi_2 < 0$, then the solutions λ_1 and λ_2 in [1.2.14] and [1.2.15] are complex conjugates. Suppose that λ_1 and λ_2 are complex conjugates, written as

$$\lambda_1 = a + bi \qquad [1.2.32]$$

$$\lambda_2 = a - bi. \qquad [1.2.33]$$

For the $p = 2$ case of [1.2.14] and [1.2.15], we would have

$$a = \phi_1/2 \qquad [1.2.34]$$

$$b = (1/2)\sqrt{-\phi_1^2 - 4\phi_2}. \qquad [1.2.35]$$

Our goal is to characterize the contribution to the dynamic multiplier $c_1\lambda_1^j$ when λ_1 is a complex number as in [1.2.32]. Recall that to raise a complex number to a power, we rewrite [1.2.32] in polar coordinate form:

$$\lambda_1 = R\cdot[\cos(\theta) + i\cdot\sin(\theta)], \qquad [1.2.36]$$

where θ and R are defined in terms of a and b by the following equations:

$$R = \sqrt{a^2 + b^2}$$

$$\cos(\theta) = a/R$$

$$\sin(\theta) = b/R.$$

Note that R is equal to the modulus of the complex number λ_1.

The eigenvalue λ_1 in [1.2.36] can be written as[4]

$$\lambda_1 = R[e^{i\theta}],$$

and so

$$\lambda_1^j = R^j[e^{i\theta j}] = R^j[\cos(\theta j) + i\cdot\sin(\theta j)]. \qquad [1.2.37]$$

Analogously, if λ_2 is the complex conjugate of λ_1, then

$$\lambda_2 = R[\cos(\theta) - i\cdot\sin(\theta)],$$

which can be written[5]

$$\lambda_2 = R[e^{-i\theta}].$$

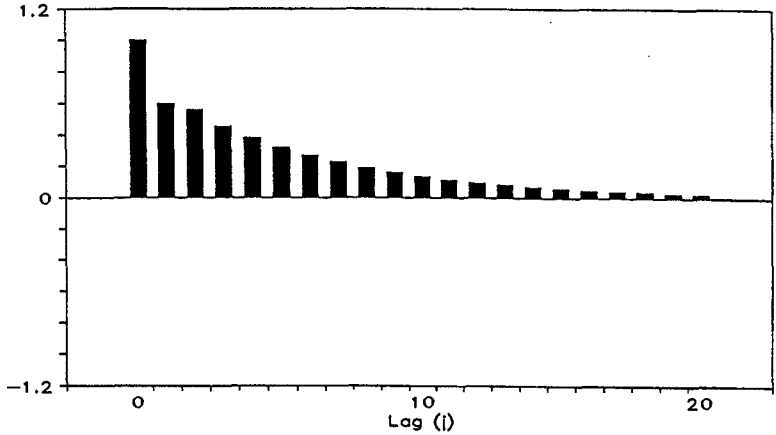
Thus

$$\lambda_2^j = R^j[e^{-i\theta j}] = R^j[\cos(\theta j) - i\cdot\sin(\theta j)]. \qquad [1.2.38]$$

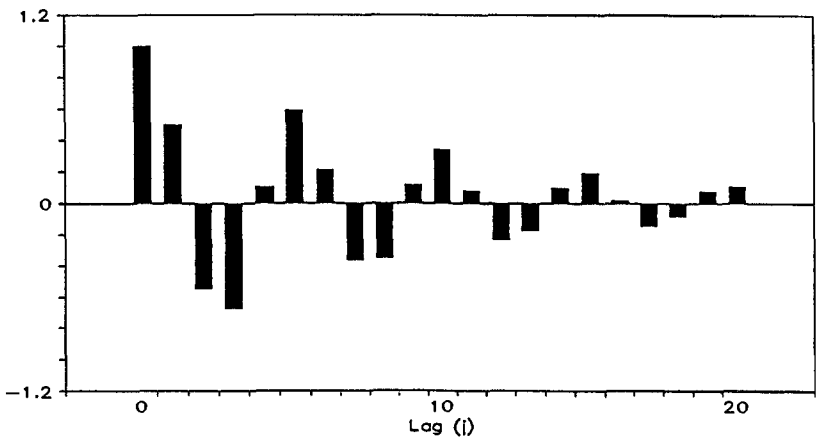
[3]Again, if one's purpose is solely to generate a numerical plot as in Figure 1.4, the easiest approach is numerical simulation of the system.

[4]See equation [A.3.25] in the Mathematical Review (Appendix A) at the end of the book.

[5]See equation [A.3.26].



(a) $\phi_1 = 0.6$, $\phi_2 = 0.2$



(b) $\phi_1 = 0.5$, $\phi_2 = -0.8$

FIGURE 1.4 Dynamic multiplier for second-order difference equation for different values of ϕ_1 and ϕ_2 (plot of $\partial y_{t+j}/\partial w_t$ as a function of the lag j).

Substituting [1.2.37] and [1.2.38] into [1.2.29] gives the contribution of the complex conjugates to the dynamic multiplier $\partial y_{t+j}/\partial w_t$:

$$\begin{aligned} c_1\lambda_1^j + c_2\lambda_2^j &= c_1R^j[\cos(\theta j) + i\cdot\sin(\theta j)] + c_2R^j[\cos(\theta j) - i\cdot\sin(\theta j)] \\ &= [c_1 + c_2]\cdot R^j\cdot\cos(\theta j) + i\cdot[c_1 - c_2]\cdot R^j\cdot\sin(\theta j). \end{aligned} \qquad [1.2.39]$$

The appearance of the imaginary number i in [1.2.39] may seem a little troubling. After all, this calculation was intended to give the effect of a change in the real-valued variable w_t on the real-valued variable y_{t+j} as predicted by the real-valued system [1.2.1], and it would be odd indeed if the correct answer involved the imaginary number i! Fortunately, it turns out from [1.2.25] that if λ_1 and λ_2 are complex conjugates, then c_1 and c_2 are complex conjugates; that is, they can

be written as

$$c_1 = \alpha + \beta i$$
$$c_2 = \alpha - \beta i$$

for some real numbers α and β. Substituting these expressions into [1.2.39] yields

$$\begin{aligned} c_1\lambda_1^j + c_2\lambda_2^j &= [(\alpha + \beta i) + (\alpha - \beta i)]\cdot R^j\cos(\theta j) + i\cdot[(\alpha + \beta i) - (\alpha - \beta i)]\cdot R^j\sin(\theta j) \\ &= [2\alpha]\cdot R^j\cos(\theta j) + i\cdot[2\beta i]\cdot R^j\sin(\theta j) \\ &= 2\alpha R^j\cos(\theta j) - 2\beta R^j\sin(\theta j), \end{aligned}$$

which is strictly real.

Thus, when some of the eigenvalues are complex, they contribute terms proportional to $R^j\cos(\theta j)$ and $R^j\sin(\theta j)$ to the dynamic multiplier $\partial y_{t+j}/\partial w_t$. Note that if $R = 1$—that is, if the complex eigenvalues have unit modulus—the multipliers are periodic sine and cosine functions of j. A given increase in w_t increases y_{t+j} for some ranges of j and decreases y_{t+j} over other ranges, with the impulse never dying out as $j \to \infty$. If the complex eigenvalues are less than 1 in modulus ($R < 1$), the impulse again follows a sinusoidal pattern though its amplitude decays at the rate R^j. If the complex eigenvalues are greater than 1 in modulus ($R > 1$), the amplitude of the sinusoids explodes at the rate R^j.

For an example of dynamic behavior characterized by decaying sinusoids, consider the second-order system

$$y_t = 0.5y_{t-1} - 0.8y_{t-2} + w_t.$$

The eigenvalues for this system are given from [1.2.14] and [1.2.15]:

$$\lambda_1 = \frac{0.5 + \sqrt{(0.5)^2 - 4(0.8)}}{2} = 0.25 + 0.86i$$
$$\lambda_2 = \frac{0.5 - \sqrt{(0.5)^2 - 4(0.8)}}{2} = 0.25 - 0.86i,$$

with modulus

$$R = \sqrt{(0.25)^2 + (0.86)^2} = 0.9.$$

Since $R < 1$, the dynamic multiplier follows a pattern of damped oscillation plotted in panel (b) of Figure 1.4. The frequency[6] of these oscillations is given by the parameter θ in [1.2.39], which was defined implicitly by

$$\cos(\theta) = a/R = (0.25)/(0.9) = 0.28$$

or

$$\theta = 1.29.$$

The cycles associated with the dynamic multiplier function [1.2.39] thus have a period of

$$\frac{2\pi}{\theta} = \frac{(2)(3.14159)}{1.29} = 4.9;$$

that is, the peaks in the pattern in panel (b) of Figure 1.4 appear about five periods apart.

[6]See Section A.1 of the Mathematical Review (Appendix A) at the end of the book for a discussion of the frequency and period of a sinusoidal function.

Solution of a Second-Order Difference Equation with Distinct Eigenvalues

The second-order difference equation ($p = 2$) comes up sufficiently often that it is useful to summarize the properties of the solution as a general function of ϕ_1 and ϕ_2, which we now do.[7]

The eigenvalues λ_1 and λ_2 in [1.2.14] and [1.2.15] are complex whenever

$$\phi_1^2 + 4\phi_2 < 0,$$

or whenever (ϕ_1, ϕ_2) lies below the parabola indicated in Figure 1.5. For the case of complex eigenvalues, the modulus R satisfies

$$R^2 = a^2 + b^2,$$

or, from [1.2.34] and [1.2.35],

$$R^2 = (\phi_1/2)^2 - (\phi_1^2 + 4\phi_2)/4 = -\phi_2.$$

Thus, a system with complex eigenvalues is explosive whenever $\phi_2 < -1$. Also, when the eigenvalues are complex, the frequency of oscillations is given by

$$\theta = \cos^{-1}(a/R) = \cos^{-1}[\phi_1/(2\sqrt{-\phi_2})],$$

where "$\cos^{-1}(x)$" denotes the inverse of the cosine function, or the radian measure of an angle whose cosine is x.

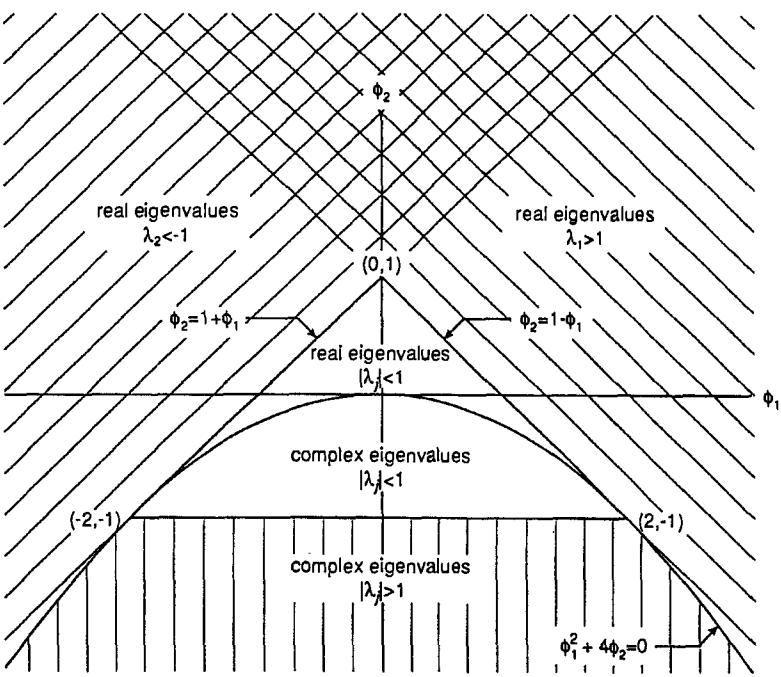


FIGURE 1.5 Summary of dynamics for a second-order difference equation.

[7]This discussion closely follows Sargent (1987, pp. 188–89).

For the case of real eigenvalues, the arithmetically larger eigenvalue (λ_1) will be greater than unity whenever

$$\frac{\phi_1 + \sqrt{\phi_1^2 + 4\phi_2}}{2} > 1$$

or

$$\sqrt{\phi_1^2 + 4\phi_2} > 2 - \phi_1.$$

Assuming that λ_1 is real, the left side of this expression is a positive number and the inequality would be satisfied for any value of $\phi_1 > 2$. If, on the other hand, $\phi_1 < 2$, we can square both sides to conclude that λ_1 will exceed unity whenever

$$\phi_1^2 + 4\phi_2 > 4 - 4\phi_1 + \phi_1^2$$

or

$$\phi_2 > 1 - \phi_1.$$

Thus, in the real region, λ_1 will be greater than unity either if $\phi_1 > 2$ or if (ϕ_1, ϕ_2) lies northeast of the line $\phi_2 = 1 - \phi_1$ in Figure 1.5. Similarly, with real eigenvalues, the arithmetically smaller eigenvalue (λ_2) will be less than -1 whenever

$$\frac{\phi_1 - \sqrt{\phi_1^2 + 4\phi_2}}{2} < -1$$

$$-\sqrt{\phi_1^2 + 4\phi_2} < -2 - \phi_1$$

$$\sqrt{\phi_1^2 + 4\phi_2} > 2 + \phi_1.$$

Again, if $\phi_1 < -2$, this must be satisfied, and in the case when $\phi_1 > -2$, we can square both sides:

$$\phi_1^2 + 4\phi_2 > 4 + 4\phi_1 + \phi_1^2$$

$$\phi_2 > 1 + \phi_1.$$

Thus, in the real region, λ_2 will be less than -1 if either $\phi_1 < -2$ or (ϕ_1, ϕ_2) lies to the northwest of the line $\phi_2 = 1 + \phi_1$ in Figure 1.5.

The system is thus stable whenever (ϕ_1, ϕ_2) lies within the triangular region of Figure 1.5.

General Solution of a pth-Order Difference Equation with Repeated Eigenvalues

In the more general case of a difference equation for which **F** has repeated eigenvalues and $s < p$ linearly independent eigenvectors, result [1.2.17] is generalized by using the Jordan decomposition,

$$\mathbf{F} = \mathbf{MJM}^{-1} \qquad [1.2.40]$$

where **M** is a $(p \times p)$ matrix and **J** takes the form

$$\mathbf{J} = \begin{bmatrix} \mathbf{J}_1 & \mathbf{0} & \cdots & \mathbf{0} \\ \mathbf{0} & \mathbf{J}_2 & \cdots & \mathbf{0} \\ \vdots & \vdots & \cdots & \vdots \\ \mathbf{0} & \mathbf{0} & \cdots & \mathbf{J}_s \end{bmatrix}$$

with

$$\mathbf{J}_i = \begin{bmatrix} \lambda_i & 1 & 0 & \cdots & 0 & 0 \\ 0 & \lambda_i & 1 & \cdots & 0 & 0 \\ 0 & 0 & \lambda_i & \cdots & 0 & 0 \\ \vdots & \vdots & \vdots & \cdots & \vdots & \vdots \\ 0 & 0 & 0 & \cdots & \lambda_i & 1 \\ 0 & 0 & 0 & \cdots & 0 & \lambda_i \end{bmatrix} \qquad [1.2.41]$$

for λ_i an eigenvalue of $\mathbf{F}$. If [1.2.17] is replaced by [1.2.40], then equation [1.2.19] generalizes to

$$\mathbf{F}^j = \mathbf{M}\mathbf{J}^j\mathbf{M}^{-1} \qquad [1.2.42]$$

where

$$\mathbf{J}^j = \begin{bmatrix} \mathbf{J}_1^j & \mathbf{0} & \cdots & \mathbf{0} \\ \mathbf{0} & \mathbf{J}_2^j & \cdots & \mathbf{0} \\ \vdots & \vdots & \cdots & \vdots \\ \mathbf{0} & \mathbf{0} & \cdots & \mathbf{J}_s^j \end{bmatrix}.$$

Moreover, from [1.2.41], if $\mathbf{J}_i$ is of dimension $(n_i \times n_i)$, then[8]

$$\mathbf{J}_i^j = \begin{bmatrix} \lambda_i^j & \binom{j}{1}\lambda_i^{j-1} & \binom{j}{2}\lambda_i^{j-2} & \cdots & \binom{j}{n_i-1}\lambda_i^{j-n_i+1} \\ 0 & \lambda_i^j & \binom{j}{1}\lambda_i^{j-1} & \cdots & \binom{j}{n_i-2}\lambda_i^{j-n_i+2} \\ \vdots & \vdots & \vdots & \cdots & \vdots \\ 0 & 0 & 0 & \cdots & \lambda_i^j \end{bmatrix} \qquad [1.2.43]$$

where

$$\binom{j}{n} \equiv \begin{cases} \dfrac{j(j-1)(j-2)\cdots(j-n+1)}{n(n-1)\cdots 3\cdot 2\cdot 1} & \text{for } j \geq n \\ 0 & \text{otherwise.} \end{cases}$$

Equation [1.2.43] may be verified by induction by multiplying [1.2.41] by [1.2.43] and noticing that $\binom{j}{n} + \binom{j}{n-1} = \binom{j+1}{n}$.

For example, consider again the second-order difference equation, this time with repeated roots. Then

$$\mathbf{F}^j = \mathbf{M}\begin{bmatrix} \lambda^j & j\lambda^{j-1} \\ 0 & \lambda^j \end{bmatrix}\mathbf{M}^{-1},$$

so that the dynamic multiplier takes the form

$$\frac{\partial y_{t+j}}{\partial w_t} = f_{11}^{(j)} = k_1\lambda^j + k_2 j\lambda^{j-1}.$$

Long-Run and Present-Value Calculations

If the eigenvalues are all less than 1 in modulus, then $\mathbf{F}^j$ in [1.2.9] goes to zero as j becomes large. If all values of w and y are taken to be bounded, we can

[8]This expression is taken from Chiang (1980, p. 444).

think of a "solution" of y_t in terms of the infinite history of w,

$$y_t = w_t + \psi_1 w_{t-1} + \psi_2 w_{t-2} + \psi_3 w_{t-3} + \cdots, \qquad [1.2.44]$$

where ψ_j is given by the (1, 1) element of $\mathbf{F}^j$ and takes the particular form of [1.2.29] in the case of distinct eigenvalues.

It is also straightforward to calculate the effect on the present value of y of a transitory increase in w. This is simplest to find if we first consider the slightly more general problem of the hypothetical consequences of a change in any element of the vector $\mathbf{v}_t$ on any element of $\boldsymbol{\xi}_{t+j}$ in a general system of the form of [1.2.5]. The answer to this more general problem can be inferred immediately from [1.2.9]:

$$\frac{\partial \boldsymbol{\xi}_{t+j}}{\partial \mathbf{v}_t'} = \mathbf{F}^j. \qquad [1.2.45]$$

The true dynamic multiplier of interest, $\partial y_{t+j}/\partial w_t$, is just the (1, 1) element of the $(p \times p)$ matrix in [1.2.45]. The effect on the present value of $\boldsymbol{\xi}$ of a change in $\mathbf{v}$ is given by

$$\frac{\partial \sum_{j=0}^{\infty} \beta^j \boldsymbol{\xi}_{t+j}}{\partial \mathbf{v}_t'} = \sum_{j=0}^{\infty} \beta^j \mathbf{F}^j = (\mathbf{I}_p - \beta\mathbf{F})^{-1}, \qquad [1.2.46]$$

provided that the eigenvalues of $\mathbf{F}$ are all less than β^{-1} in modulus. The effect on the present value of y of a change in w,

$$\frac{\partial \sum_{j=0}^{\infty} \beta^j y_{t+j}}{\partial w_t},$$

is thus the (1, 1) element of the $(p \times p)$ matrix in [1.2.46]. This value is given by the following proposition.

Proposition 1.3: If the eigenvalues of the $(p \times p)$ matrix $\mathbf{F}$ defined in [1.2.3] are all less than β^{-1} in modulus, then the matrix $(\mathbf{I}_p - \beta\mathbf{F})^{-1}$ exists and the effect of w on the present value of y is given by its (1, 1) element:

$$1/(1 - \phi_1\beta - \phi_2\beta^2 - \cdots - \phi_{p-1}\beta^{p-1} - \phi_p\beta^p).$$

Note that Proposition 1.3 includes the earlier result for a first-order system (equation [1.1.14]) as a special case.

The cumulative effect of a one-time change in w_t on $y_t, y_{t+1}, \ldots$ can be considered a special case of Proposition 1.3 with no discounting. Setting $\beta = 1$ in Proposition 1.3 shows that, provided the eigenvalues of $\mathbf{F}$ are all less than 1 in modulus, the cumulative effect of a one-time change in w on y is given by

$$\sum_{j=0}^{\infty} \frac{\partial y_{t+j}}{\partial w_t} = 1/(1 - \phi_1 - \phi_2 - \cdots - \phi_p). \qquad [1.2.47]$$

Notice again that [1.2.47] can alternatively be interpreted as giving the eventual long-run effect on y of a permanent change in w:

$$\lim_{j\to\infty} \frac{\partial y_{t+j}}{\partial w_t} + \frac{\partial y_{t+j}}{\partial w_{t+1}} + \frac{\partial y_{t+j}}{\partial w_{t+2}} + \cdots + \frac{\partial y_{t+j}}{\partial w_{t+j}} = 1/(1 - \phi_1 - \phi_2 - \cdots - \phi_p).$$

APPENDIX 1.A. *Proofs of Chapter 1 Propositions*

■ **Proof of Proposition 1.1.** The eigenvalues of $\mathbf{F}$ satisfy

$$|\mathbf{F} - \lambda \mathbf{I}_p| = 0. \qquad [1.A.1]$$

For the matrix $\mathbf{F}$ defined in equation [1.2.3], this determinant would be

$$\left| \begin{bmatrix} \phi_1 & \phi_2 & \phi_3 & \cdots & \phi_{p-1} & \phi_p \\ 1 & 0 & 0 & \cdots & 0 & 0 \\ 0 & 1 & 0 & \cdots & 0 & 0 \\ \vdots & \vdots & \vdots & \cdots & \vdots & \vdots \\ 0 & 0 & 0 & \cdots & 1 & 0 \end{bmatrix} - \begin{bmatrix} \lambda & 0 & 0 & \cdots & 0 & 0 \\ 0 & \lambda & 0 & \cdots & 0 & 0 \\ 0 & 0 & \lambda & \cdots & 0 & 0 \\ \vdots & \vdots & \vdots & \cdots & \vdots & \vdots \\ 0 & 0 & 0 & \cdots & 0 & \lambda \end{bmatrix} \right|$$

$$= \begin{vmatrix} (\phi_1 - \lambda) & \phi_2 & \phi_3 & \cdots & \phi_{p-1} & \phi_p \\ 1 & -\lambda & 0 & \cdots & 0 & 0 \\ 0 & 1 & -\lambda & \cdots & 0 & 0 \\ \vdots & \vdots & \vdots & \cdots & \vdots & \vdots \\ 0 & 0 & 0 & \cdots & 1 & -\lambda \end{vmatrix}. \qquad [1.A.2]$$

Recall that if we multiply a column of a matrix by a constant and add the result to another column, the determinant of the matrix is unchanged. If we multiply the pth column of the matrix in [1.A.2] by $(1/\lambda)$ and add the result to the $(p-1)$th column, the result is a matrix with the same determinant as that in [1.A.2]:

$$|\mathbf{F} - \lambda \mathbf{I}_p| = \begin{vmatrix} \phi_1 - \lambda & \phi_2 & \phi_3 & \cdots & \phi_{p-2} & \phi_{p-1} + (\phi_p/\lambda) & \phi_p \\ 1 & -\lambda & 0 & \cdots & 0 & 0 & 0 \\ 0 & 1 & -\lambda & \cdots & 0 & 0 & 0 \\ \vdots & \vdots & \vdots & \cdots & \vdots & \vdots & \vdots \\ 0 & 0 & 0 & \cdots & 1 & -\lambda & 0 \\ 0 & 0 & 0 & \cdots & 0 & 0 & -\lambda \end{vmatrix}.$$

Next, multiply the $(p-1)$th column by $(1/\lambda)$ and add the result to the $(p-2)$th column:

$$|\mathbf{F} - \lambda \mathbf{I}_p|$$

$$= \begin{vmatrix} \phi_1 - \lambda & \phi_2 & \phi_3 & \cdots & \phi_{p-2} + \phi_{p-1}/\lambda + \phi_p/\lambda^2 & \phi_{p-1} + \phi_p/\lambda & \phi_p \\ 1 & -\lambda & 0 & \cdots & 0 & 0 & 0 \\ 0 & 1 & -\lambda & \cdots & 0 & 0 & 0 \\ \vdots & \vdots & \vdots & \cdots & \vdots & \vdots & \vdots \\ 0 & 0 & 0 & \cdots & 0 & -\lambda & 0 \\ 0 & 0 & 0 & \cdots & 0 & 0 & -\lambda \end{vmatrix}.$$

Continuing in this fashion shows [1.A.1] to be equivalent to the determinant of the following upper triangular matrix:

$$|\mathbf{F} - \lambda \mathbf{I}_p|$$

$$= \begin{vmatrix} \phi_1 - \lambda + \phi_2/\lambda + \phi_3/\lambda^2 + \cdots + \phi_p/\lambda^{p-1} & \phi_2 + \phi_3/\lambda + \phi_4/\lambda^2 + \cdots + \phi_p/\lambda^{p-2} & \cdots & \phi_{p-1} + \phi_p/\lambda & \phi_p \\ 0 & -\lambda & \cdots & 0 & 0 \\ 0 & 0 & \cdots & 0 & 0 \\ \vdots & \vdots & \cdots & \vdots & \vdots \\ 0 & 0 & \cdots & -\lambda & 0 \\ 0 & 0 & \cdots & 0 & -\lambda \end{vmatrix}.$$

But the determinant of an upper triangular matrix is simply the product of the terms along the principal diagonal:

$$\begin{aligned} |\mathbf{F} - \lambda \mathbf{I}_p| &= [\phi_1 - \lambda + \phi_2/\lambda + \phi_3/\lambda^2 + \cdots + \phi_p/\lambda^{p-1}] \cdot [-\lambda]^{p-1} \\ &= (-1)^p \cdot [\lambda^p - \phi_1\lambda^{p-1} - \phi_2\lambda^{p-2} - \cdots - \phi_p]. \end{aligned} \qquad [1.A.3]$$

The eigenvalues of $\mathbf{F}$ are thus the values of λ for which [1.A.3] is zero, or for which

$$\lambda^p - \phi_1\lambda^{p-1} - \phi_2\lambda^{p-2} - \cdots - \phi_p = 0,$$

as asserted in Proposition 1.1. ■

■ **Proof of Proposition 1.2.** Assuming that the eigenvalues $(\lambda_1, \lambda_2, \ldots, \lambda_p)$ are distinct, the matrix $\mathbf{T}$ in equation [1.2.17] can be constructed from the eigenvectors of $\mathbf{F}$. Let $\mathbf{t}_i$ denote the following $(p \times 1)$ vector,

$$\mathbf{t}_i = \begin{bmatrix} \lambda_i^{p-1} \\ \lambda_i^{p-2} \\ \lambda_i^{p-3} \\ \vdots \\ \lambda_i^1 \\ 1 \end{bmatrix}, \qquad [1.A.4]$$

where λ_i denotes the ith eigenvalue of $\mathbf{F}$. Notice

$$\begin{aligned} \mathbf{F}\mathbf{t}_i &= \begin{bmatrix} \phi_1 & \phi_2 & \phi_3 & \cdots & \phi_{p-1} & \phi_p \\ 1 & 0 & 0 & \cdots & 0 & 0 \\ 0 & 1 & 0 & \cdots & 0 & 0 \\ \vdots & \vdots & \vdots & \cdots & \vdots & \vdots \\ 0 & 0 & 0 & \cdots & 1 & 0 \end{bmatrix} \begin{bmatrix} \lambda_i^{p-1} \\ \lambda_i^{p-2} \\ \lambda_i^{p-3} \\ \vdots \\ \lambda_i^1 \\ 1 \end{bmatrix} \\ &= \begin{bmatrix} \phi_1\lambda_i^{p-1} + \phi_2\lambda_i^{p-2} + \phi_3\lambda_i^{p-3} + \cdots + \phi_{p-1}\lambda_i + \phi_p \\ \lambda_i^{p-1} \\ \lambda_i^{p-2} \\ \vdots \\ \lambda_i^2 \\ \lambda_i \end{bmatrix}. \end{aligned} \qquad [1.A.5]$$

Since λ_i is an eigenvalue of $\mathbf{F}$, it satisfies [1.2.16]:

$$\lambda_i^p - \phi_1\lambda_i^{p-1} - \phi_2\lambda_i^{p-2} - \cdots - \phi_{p-1}\lambda_i - \phi_p = 0. \qquad [1.A.6]$$

Substituting [1.A.6] into [1.A.5] reveals

$$\mathbf{F}\mathbf{t}_i = \begin{bmatrix} \lambda_i^p \\ \lambda_i^{p-1} \\ \lambda_i^{p-2} \\ \vdots \\ \lambda_i^2 \\ \lambda_i \end{bmatrix} = \lambda_i \begin{bmatrix} \lambda_i^{p-1} \\ \lambda_i^{p-2} \\ \lambda_i^{p-3} \\ \vdots \\ \lambda_i^1 \\ 1 \end{bmatrix}$$

or

$$\mathbf{F}\mathbf{t}_i = \lambda_i\mathbf{t}_i. \qquad [1.A.7]$$

Thus $\mathbf{t}_i$ is an eigenvector of $\mathbf{F}$ associated with the eigenvalue λ_i.

We can calculate the matrix $\mathbf{T}$ by combining the eigenvectors $(\mathbf{t}_1, \mathbf{t}_2, \ldots, \mathbf{t}_p)$ into a $(p \times p)$ matrix

$$\mathbf{T} = [\mathbf{t}_1 \quad \mathbf{t}_2 \quad \cdots \quad \mathbf{t}_p]. \qquad [1.A.8]$$

To calculate the particular values for c_i in equation [1.2.21], recall that $\mathbf{T}^{-1}$ is characterized by

$$\mathbf{T}\mathbf{T}^{-1} = \mathbf{I}_p, \qquad [1.A.9]$$

where $\mathbf{T}$ is given by [1.A.4] and [1.A.8]. Writing out the first column of the matrix system of equations [1.A.9] explicitly, we have

$$\begin{bmatrix} \lambda_1^{p-1} & \lambda_2^{p-1} & \cdots & \lambda_p^{p-1} \\ \lambda_1^{p-2} & \lambda_2^{p-2} & \cdots & \lambda_p^{p-2} \\ \lambda_1^{p-3} & \lambda_2^{p-3} & \cdots & \lambda_p^{p-3} \\ \vdots & \vdots & \cdots & \vdots \\ \lambda_1^1 & \lambda_2^1 & \cdots & \lambda_p^1 \\ 1 & 1 & \cdots & 1 \end{bmatrix} \begin{bmatrix} t^{11} \\ t^{21} \\ t^{31} \\ \vdots \\ t^{p-1,1} \\ t^{p1} \end{bmatrix} = \begin{bmatrix} 1 \\ 0 \\ 0 \\ \vdots \\ 0 \\ 0 \end{bmatrix}.$$

This gives a system of p linear equations in the p unknowns $(t^{11}, t^{21}, \ldots, t^{p1})$. Provided that the λ_i are all distinct, the solution can be shown to be[9]

$$t^{11} = \frac{1}{(\lambda_1 - \lambda_2)(\lambda_1 - \lambda_3)\cdots(\lambda_1 - \lambda_p)}$$

$$t^{21} = \frac{1}{(\lambda_2 - \lambda_1)(\lambda_2 - \lambda_3)\cdots(\lambda_2 - \lambda_p)}$$

$$\vdots$$

$$t^{p1} = \frac{1}{(\lambda_p - \lambda_1)(\lambda_p - \lambda_2)\cdots(\lambda_p - \lambda_{p-1})}.$$

Substituting these values into [1.2.21] gives equation [1.2.25]. ■

■ **Proof of Proposition 1.3.** The first claim in this proposition is that if the eigenvalues of $\mathbf{F}$ are less than β^{-1} in modulus, then the inverse of $(\mathbf{I}_p - \beta\mathbf{F})$ exists. Suppose the inverse of $(\mathbf{I}_p - \beta\mathbf{F})$ did not exist. Then the determinant $|\mathbf{I}_p - \beta\mathbf{F}|$ would have to be zero. But

$$|\mathbf{I}_p - \beta\mathbf{F}| = |-\beta \cdot [\mathbf{F} - \beta^{-1}\mathbf{I}_p]| = (-\beta)^p|\mathbf{F} - \beta^{-1}\mathbf{I}_p|,$$

so that $|\mathbf{F} - \beta^{-1}\mathbf{I}_p|$ would have to be zero whenever the inverse of $(\mathbf{I}_p - \beta\mathbf{F})$ fails to exist. But this would mean that β^{-1} is an eigenvalue of $\mathbf{F}$, which is ruled out by the assumption that all eigenvalues of $\mathbf{F}$ are strictly less than β^{-1} in modulus. Thus, the matrix $\mathbf{I}_p - \beta\mathbf{F}$ must be nonsingular.

Since $[\mathbf{I}_p - \beta\mathbf{F}]^{-1}$ exists, it satisfies the equation

$$[\mathbf{I}_p - \beta\mathbf{F}]^{-1}[\mathbf{I}_p - \beta\mathbf{F}] = \mathbf{I}_p. \qquad [1.A.10]$$

Let x_{ij} denote the row i, column j element of $[\mathbf{I}_p - \beta\mathbf{F}]^{-1}$, and write [1.A.10] as

$$\begin{bmatrix} x_{11} & x_{12} & \cdots & x_{1p} \\ x_{21} & x_{22} & \cdots & x_{2p} \\ \vdots & \vdots & \cdots & \vdots \\ x_{p1} & x_{p2} & \cdots & x_{pp} \end{bmatrix} \begin{bmatrix} 1 - \beta\phi_1 & -\beta\phi_2 & \cdots & -\beta\phi_{p-1} & -\beta\phi_p \\ -\beta & 1 & \cdots & 0 & 0 \\ \vdots & \vdots & \cdots & \vdots & \vdots \\ 0 & 0 & \cdots & -\beta & 1 \end{bmatrix} = \begin{bmatrix} 1 & 0 & \cdots & 0 \\ 0 & 1 & \cdots & 0 \\ \vdots & \vdots & \cdots & \vdots \\ 0 & 0 & \cdots & 1 \end{bmatrix}. \qquad [1.A.11]$$

The task is then to find the (1, 1) element of $[\mathbf{I}_p - \beta\mathbf{F}]^{-1}$, that is, to find the value of x_{11}. To do this we need only consider the first row of equations in [1.A.11]:

$$[x_{11} \;\; x_{12} \;\; \cdots \;\; x_{1p}] \begin{bmatrix} 1 - \beta\phi_1 & -\beta\phi_2 & \cdots & -\beta\phi_{p-1} & -\beta\phi_p \\ -\beta & 1 & \cdots & 0 & 0 \\ \vdots & \vdots & \cdots & \vdots & \vdots \\ 0 & 0 & \cdots & -\beta & 1 \end{bmatrix} = [1 \;\; 0 \;\; \cdots \;\; 0 \;\; 0]. \qquad [1.A.12]$$

[9]See Lemma 2 of Chiang (1980, p. 144).

Consider postmultiplying this system of equations by a matrix with 1s along the principal diagonal, β in the row p, column $p - 1$ position, and 0s elsewhere:

$$\begin{bmatrix} 1 & 0 & \cdots & 0 & 0 \\ 0 & 1 & \cdots & 0 & 0 \\ \vdots & \vdots & \cdots & \vdots & \vdots \\ 0 & 0 & \cdots & \beta & 1 \end{bmatrix}.$$

The effect of this operation is to multiply the pth column of a matrix by β and add the result to the $(p - 1)$th column:

$$[x_{11} \quad x_{12} \cdots x_{1p}] \begin{bmatrix} 1 - \beta\phi_1 & -\beta\phi_2 & \cdots & -\beta\phi_{p-1} - \beta^2\phi_p & -\beta\phi_p \\ -\beta & 1 & \cdots & 0 & 0 \\ \vdots & \vdots & \cdots & \vdots & \vdots \\ 0 & 0 & \cdots & 0 & 1 \end{bmatrix} = [1 \quad 0 \cdots 0 \quad 0].$$

Next multiply the $(p - 1)$th column by β and add the result to the $(p - 2)$th column. Proceeding in this fashion, we arrive at

$$[x_{11} \quad x_{12} \quad \cdots \quad x_{1p}] \times$$

$$\begin{bmatrix} 1 - \beta\phi_1 - \beta^2\phi_2 - \cdots - \beta^{p-1}\phi_{p-1} - \beta^p\phi_p & -\beta\phi_2 - \beta^2\phi_3 - \cdots - \beta^{p-1}\phi_p & \cdots & -\beta\phi_{p-1} - \beta^2\phi_p & -\beta\phi_p \\ 0 & 1 & \cdots & 0 & 0 \\ \vdots & \vdots & \cdots & \vdots & \vdots \\ 0 & 0 & \cdots & 0 & 1 \end{bmatrix}$$

$$= [1 \quad 0 \quad \cdots \quad 0 \quad 0]. \qquad [1.A.13]$$

The first equation in [1.A.13] states that

$$x_{11} \cdot (1 - \beta\phi_1 - \beta^2\phi_2 - \cdots - \beta^{p-1}\phi_{p-1} - \beta^p\phi_p) = 1$$

or

$$x_{11} = 1/(1 - \beta\phi_1 - \beta^2\phi_2 - \cdots - \beta^p\phi_p),$$

as claimed in Proposition 1.3. ■

Chapter 1 References

Chiang, Chin Long. 1980. *An Introduction to Stochastic Processes and Their Applications.* Huntington, N.Y.: Krieger.

Goldfeld, Stephen M. 1973. "The Demand for Money Revisited," *Brookings Papers on Economic Activity* 3:577–638.

Sargent, Thomas J. 1987. *Macroeconomic Theory*, 2d ed. Boston: Academic Press.

2 Lag Operators

2.1. *Introduction*

The previous chapter analyzed the dynamics of linear difference equations using matrix algebra. This chapter develops some of the same results using time series operators. We begin with some introductory remarks on some useful time series operators.

A time series is a collection of observations indexed by the date of each observation. Usually we have collected data beginning at some particular date (say, $t = 1$) and ending at another (say, $t = T$):

$$(y_1, y_2, \ldots, y_T).$$

We often imagine that we could have obtained earlier observations (y_0, y_{-1}, $y_{-2}, \ldots$) or later observations ($y_{T+1}, y_{T+2}, \ldots$) had the process been observed for more time. The observed sample ($y_1, y_2, \ldots, y_T$) could then be viewed as a finite segment of a doubly infinite sequence, denoted $\{y_t\}_{t=-\infty}^{\infty}$:

$$\{y_t\}_{t=-\infty}^{\infty} = \{\ldots, y_{-1}, y_0, \underbrace{y_1, y_2, \ldots, y_T}_{\text{observed sample}}, y_{T+1}, y_{T+2}, \ldots\}.$$

Typically, a time series $\{y_t\}_{t=-\infty}^{\infty}$ is identified by describing the tth element. For example, a *time trend* is a series whose value at date t is simply the date of the observation:

$$y_t = t.$$

We could also consider a time series in which each element is equal to a constant c, regardless of the date of the observation t:

$$y_t = c.$$

Another important time series is a *Gaussian white noise process*, denoted

$$y_t = \varepsilon_t,$$

where $\{\varepsilon_t\}_{t=-\infty}^{\infty}$ is a sequence of independent random variables each of which has a $N(0, \sigma^2)$ distribution.

We are used to thinking of a function such as $y = f(x)$ or $y = g(x, w)$ as an operation that accepts as input a number (x) or group of numbers (x, w) and produces the output (y). A time series *operator* transforms one time series or group

of time series into a new time series. It accepts as input a sequence such as $\{x_t\}_{t=-\infty}^{\infty}$ or a group of sequences such as $(\{x_t\}_{t=-\infty}^{\infty}, \{w_t\}_{t=-\infty}^{\infty})$ and has as output a new sequence $\{y_t\}_{t=-\infty}^{\infty}$. Again, the operator is summarized by describing the value of a typical element of $\{y_t\}_{t=-\infty}^{\infty}$ in terms of the corresponding elements of $\{x_t\}_{t=-\infty}^{\infty}$.

An example of a time series operator is the multiplication operator, represented as

$$y_t = \beta x_t. \qquad [2.1.1]$$

Although it is written exactly the same way as simple scalar multiplication, equation [2.1.1] is actually shorthand for an infinite sequence of multiplications, one for each date t. The operator multiplies the value x takes on at any date t by some constant β to generate the value of y for that date.

Another example of a time series operator is the addition operator:

$$y_t = x_t + w_t.$$

Here the value of y at any date t is the sum of the values that x and w take on for that date.

Since the multiplication or addition operators amount to element-by-element multiplication or addition, they obey all the standard rules of algebra. For example, if we multiply each observation of $\{x_t\}_{t=-\infty}^{\infty}$ by β and each observation of $\{w_t\}_{t=-\infty}^{\infty}$ by β and add the results,

$$\beta x_t + \beta w_t,$$

the outcome is the same as if we had first added $\{x_t\}_{t=-\infty}^{\infty}$ to $\{w_t\}_{t=-\infty}^{\infty}$ and then multiplied each element of the resulting series by β:

$$\beta(x_t + w_t).$$

A highly useful operator is the lag operator. Suppose that we start with a sequence $\{x_t\}_{t=-\infty}^{\infty}$ and generate a new sequence $\{y_t\}_{t=-\infty}^{\infty}$, where the value of y for date t is equal to the value x took on at date $t - 1$:

$$y_t = x_{t-1}. \qquad [2.1.2]$$

This is described as applying the *lag operator* to $\{x_t\}_{t=-\infty}^{\infty}$. The operation is represented by the symbol L:

$$Lx_t \equiv x_{t-1}. \qquad [2.1.3]$$

Consider the result of applying the lag operator twice to a series:

$$L(Lx_t) = L(x_{t-1}) = x_{t-2}.$$

Such a double application of the lag operator is indicated by "L^2":

$$L^2x_t = x_{t-2}.$$

In general, for any integer k,

$$L^kx_t = x_{t-k}. \qquad [2.1.4]$$

Notice that if we first apply the multiplication operator and then the lag operator, as in

$$x_t \rightarrow \beta x_t \rightarrow \beta x_{t-1},$$

the result will be exactly the same as if we had applied the lag operator first and then the multiplication operator:

$$x_t \rightarrow x_{t-1} \rightarrow \beta x_{t-1}.$$

Thus the lag operator and multiplication operator are commutative:

$$L(\beta x_t) = \beta \cdot L x_t.$$

Similarly, if we first add two series and then apply the lag operator to the result,

$$(x_t, w_t) \to x_t + w_t \to x_{t-1} + w_{t-1},$$

the result is the same as if we had applied the lag operator before adding:

$$(x_t, w_t) \to (x_{t-1}, w_{t-1}) \to x_{t-1} + w_{t-1}.$$

Thus, the lag operator is distributive over the addition operator:

$$L(x_t + w_t) = Lx_t + Lw_t.$$

We thus see that the lag operator follows exactly the same algebraic rules as the multiplication operator. For this reason, it is tempting to use the expression "multiply y_t by L" rather than "operate on $\{y_t\}_{t=-\infty}^{\infty}$ by L." Although the latter expression is technically more correct, this text will often use the former shorthand expression to facilitate the exposition.

Faced with a time series defined in terms of compound operators, we are free to use the standard commutative, associative, and distributive algebraic laws for multiplication and addition to express the compound operator in an alternative form. For example, the process defined by

$$y_t = (a + bL)Lx_t$$

is exactly the same as

$$y_t = (aL + bL^2)x_t = ax_{t-1} + bx_{t-2}.$$

To take another example,

$$\begin{aligned}(1 - \lambda_1 L)(1 - \lambda_2 L)x_t &= (1 - \lambda_1 L - \lambda_2 L + \lambda_1\lambda_2 L^2)x_t \\ &= (1 - [\lambda_1 + \lambda_2]L + \lambda_1\lambda_2 L^2)x_t \qquad [2.1.5] \\ &= x_t - (\lambda_1 + \lambda_2)x_{t-1} + (\lambda_1\lambda_2)x_{t-2}.\end{aligned}$$

An expression such as $(aL + bL^2)$ is referred to as a *polynomial in the lag operator*. It is algebraically similar to a simple polynomial $(az + bz^2)$ where z is a scalar. The difference is that the simple polynomial $(az + bz^2)$ refers to a particular number, whereas a polynomial in the lag operator $(aL + bL^2)$ refers to an operator that would be applied to one time series $\{x_t\}_{t=-\infty}^{\infty}$ to produce a new time series $\{y_t\}_{t=-\infty}^{\infty}$.

Notice that if $\{x_t\}_{t=-\infty}^{\infty}$ is just a series of constants,

$$x_t = c \qquad \text{for all } t,$$

then the lag operator applied to x_t produces the same series of constants:

$$Lx_t = x_{t-1} = c.$$

Thus, for example,

$$(\alpha L + \beta L^2 + \gamma L^3)c = (\alpha + \beta + \gamma) \cdot c. \qquad [2.1.6]$$

2.2. *First-Order Difference Equations*

Let us now return to the first-order difference equation analyzed in Section 1.1:

$$y_t = \phi y_{t-1} + w_t. \qquad [2.2.1]$$

Equation [2.2.1] can be rewritten using the lag operator [2.1.3] as

$$y_t = \phi L y_t + w_t.$$

This equation, in turn, can be rearranged using standard algebra,

$$y_t - \phi L y_t = w_t,$$

or

$$(1 - \phi L)y_t = w_t. \qquad [2.2.2]$$

Next consider "multiplying" both sides of [2.2.2] by the following operator:

$$(1 + \phi L + \phi^2 L^2 + \phi^3 L^3 + \cdots + \phi^t L^t). \qquad [2.2.3]$$

The result would be

$$\begin{aligned}(1 + \phi L + \phi^2 L^2 + \phi^3 L^3 + \cdots + \phi^t L^t)(1 - \phi L)y_t \\ = (1 + \phi L + \phi^2 L^2 + \phi^3 L^3 + \cdots + \phi^t L^t)w_t.\end{aligned} \qquad [2.2.4]$$

Expanding out the compound operator on the left side of [2.2.4] results in

$$\begin{aligned}&(1 + \phi L + \phi^2 L^2 + \phi^3 L^3 + \cdots + \phi^t L^t)(1 - \phi L) \\ &= (1 + \phi L + \phi^2 L^2 + \phi^3 L^3 + \cdots + \phi^t L^t) \\ &\quad - (1 + \phi L + \phi^2 L^2 + \phi^3 L^3 + \cdots + \phi^t L^t)\phi L \\ &= (1 + \phi L + \phi^2 L^2 + \phi^3 L^3 + \cdots + \phi^t L^t) \\ &\quad - (\phi L + \phi^2 L^2 + \phi^3 L^3 + \cdots + \phi^t L^t + \phi^{t+1} L^{t+1}) \\ &= (1 - \phi^{t+1} L^{t+1}).\end{aligned} \qquad [2.2.5]$$

Substituting [2.2.5] into [2.2.4] yields

$$(1 - \phi^{t+1} L^{t+1})y_t = (1 + \phi L + \phi^2 L^2 + \phi^3 L^3 + \cdots + \phi^t L^t)w_t. \qquad [2.2.6]$$

Writing [2.2.6] out explicitly using [2.1.4] produces

$$y_t - \phi^{t+1} y_{t-(t+1)} = w_t + \phi w_{t-1} + \phi^2 w_{t-2} + \phi^3 w_{t-3} + \cdots + \phi^t w_{t-t}$$

or

$$y_t = \phi^{t+1} y_{-1} + w_t + \phi w_{t-1} + \phi^2 w_{t-2} + \phi^3 w_{t-3} + \cdots + \phi^t w_0. \qquad [2.2.7]$$

Notice that equation [2.2.7] is identical to equation [1.1.7]. Applying the operator [2.2.3] is performing exactly the same set of recursive substitutions that were employed in the previous chapter to arrive at [1.1.7].

It is interesting to reflect on the nature of the operator [2.2.3] as t becomes large. We saw in [2.2.5] that

$$(1 + \phi L + \phi^2 L^2 + \phi^3 L^3 + \cdots + \phi^t L^t)(1 - \phi L)y_t = y_t - \phi^{t+1} y_{-1}.$$

That is, $(1 + \phi L + \phi^2 L^2 + \phi^3 L^3 + \cdots + \phi^t L^t)(1 - \phi L)y_t$ differs from y_t by the term $\phi^{t+1} y_{-1}$. If $|\phi| < 1$ and if y_{-1} is a finite number, this residual $\phi^{t+1} y_{-1}$ will become negligible as t becomes large:

$$(1 + \phi L + \phi^2 L^2 + \phi^3 L^3 + \cdots + \phi^t L^t)(1 - \phi L)y_t \cong y_t \qquad \text{for } t \text{ large.}$$

A sequence $\{y_t\}_{t=-\infty}^{\infty}$ is said to be *bounded* if there exists a finite number $\bar{y}$ such that

$$|y_t| < \bar{y} \qquad \text{for all } t.$$

Thus, when $|\phi| < 1$ and when we are considering applying an operator to a bounded sequence, we can think of

$$(1 + \phi L + \phi^2 L^2 + \phi^3 L^3 + \cdots + \phi^j L^j)$$

as approximating the inverse of the operator $(1 - \phi L)$, with this approximation made arbitrarily accurate by choosing j sufficiently large:

$$(1 - \phi L)^{-1} = \lim_{j \to \infty} (1 + \phi L + \phi^2 L^2 + \phi^3 L^3 + \cdots + \phi^j L^j). \quad [2.2.8]$$

This operator $(1 - \phi L)^{-1}$ has the property

$$(1 - \phi L)^{-1}(1 - \phi L) = 1,$$

where "1" denotes the identity operator:

$$1 y_t = y_t.$$

The following chapter discusses stochastic sequences rather than the deterministic sequences studied here. There we will speak of mean square convergence and stationary stochastic processes in place of limits of bounded deterministic sequences, though the practical meaning of [2.2.8] will be little changed.

Provided that $|\phi| < 1$ and we restrict ourselves to bounded sequences or stationary stochastic processes, both sides of [2.2.2] can be "divided" by $(1 - \phi L)$ to obtain

$$y_t = (1 - \phi L)^{-1} w_t$$

or

$$y_t = w_t + \phi w_{t-1} + \phi^2 w_{t-2} + \phi^3 w_{t-3} + \cdots. \quad [2.2.9]$$

It should be emphasized that if we were not restricted to considering bounded sequences or stationary stochastic processes $\{w_t\}_{t=-\infty}^{\infty}$ and $\{y_t\}_{t=-\infty}^{\infty}$, then expression [2.2.9] would not be a necessary implication of [2.2.1]. Equation [2.2.9] is consistent with [2.2.1], but adding a term $a_o \phi^t$,

$$y_t = a_o \phi^t + w_t + \phi w_{t-1} + \phi^2 w_{t-2} + \phi^3 w_{t-3} + \cdots, \quad [2.2.10]$$

produces another series consistent with [2.2.1] for any constant a_o. To verify that [2.2.10] is consistent with [2.2.1], multiply [2.2.10] by $(1 - \phi L)$:

$$\begin{aligned}(1 - \phi L) y_t &= (1 - \phi L) a_o \phi^t + (1 - \phi L)(1 - \phi L)^{-1} w_t \\ &= a_o \phi^t - \phi \cdot a_o \phi^{t-1} + w_t \\ &= w_t,\end{aligned}$$

so that [2.2.10] is consistent with [2.2.1] for any constant a_o.

Although any process of the form of [2.2.10] is consistent with the difference equation [2.2.1], notice that since $|\phi| < 1$,

$$|a_o \phi^t| \to \infty \quad \text{as} \quad t \to -\infty.$$

Thus, even if $\{w_t\}_{t=-\infty}^{\infty}$ is a bounded sequence, the solution $\{y_t\}_{t=-\infty}^{\infty}$ given by [2.2.10] is unbounded unless $a_o = 0$ in [2.2.10]. Thus, there was a particular reason for defining the operator [2.2.8] to be the inverse of $(1 - \phi L)$—namely, $(1 - \phi L)^{-1}$ defined in [2.2.8] is the unique operator satisfying

$$(1 - \phi L)^{-1}(1 - \phi L) = 1$$

that maps a bounded sequence $\{w_t\}_{t=-\infty}^{\infty}$ into a bounded sequence $\{y_t\}_{t=-\infty}^{\infty}$.

The nature of $(1 - \phi L)^{-1}$ when $|\phi| \geq 1$ will be discussed in Section 2.5.

2.3. *Second-Order Difference Equations*

Consider next a second-order difference equation:

$$y_t = \phi_1 y_{t-1} + \phi_2 y_{t-2} + w_t. \quad [2.3.1]$$

Rewriting this in lag operator form produces

$$(1 - \phi_1 L - \phi_2 L^2)y_t = w_t. \qquad [2.3.2]$$

The left side of [2.3.2] contains a second-order polynomial in the lag operator L. Suppose we factor this polynomial, that is, find numbers λ_1 and λ_2 such that

$$(1 - \phi_1 L - \phi_2 L^2) = (1 - \lambda_1 L)(1 - \lambda_2 L) = (1 - [\lambda_1 + \lambda_2]L + \lambda_1\lambda_2 L^2). \qquad [2.3.3]$$

This is just the operation in [2.1.5] in reverse. Given values for ϕ_1 and ϕ_2, we seek numbers λ_1 and λ_2 with the properties that

$$\lambda_1 + \lambda_2 = \phi_1$$

and

$$\lambda_1\lambda_2 = -\phi_2.$$

For example, if $\phi_1 = 0.6$ and $\phi_2 = -0.08$, then we should choose $\lambda_1 = 0.4$ and $\lambda_2 = 0.2$:

$$(1 - 0.6L + 0.08L^2) = (1 - 0.4L)(1 - 0.2L). \qquad [2.3.4]$$

It is easy enough to see that these values of λ_1 and λ_2 work for this numerical example, but how are λ_1 and λ_2 found in general? The task is to choose λ_1 and λ_2 so as to make sure that the operator on the right side of [2.3.3] is identical to that on the left side. This will be true whenever the following represent the identical functions of z:

$$(1 - \phi_1 z - \phi_2 z^2) = (1 - \lambda_1 z)(1 - \lambda_2 z). \qquad [2.3.5]$$

This equation simply replaces the lag operator L in [2.3.3] with a scalar z. What is the point of doing so? With [2.3.5], we can now ask, For what values of z is the right side of [2.3.5] equal to zero? The answer is, if either $z = \lambda_1^{-1}$ or $z = \lambda_2^{-1}$, then the right side of [2.3.5] would be zero. It would not have made sense to ask an analogous question of [2.3.3]—L denotes a particular operator, not a number, and $L = \lambda_1^{-1}$ is not a sensible statement.

Why should we care that the right side of [2.3.5] is zero if $z = \lambda_1^{-1}$ or if $z = \lambda_2^{-1}$? Recall that the goal was to choose λ_1 and λ_2 so that the two sides of [2.3.5] represented the identical polynomial in z. This means that for any particular value z the two functions must produce the same number. If we find a value of z that sets the right side to zero, that same value of z must set the left side to zero as well. But the values of z that set the left side to zero,

$$(1 - \phi_1 z - \phi_2 z^2) = 0, \qquad [2.3.6]$$

are given by the quadratic formula:

$$z_1 = \frac{\phi_1 - \sqrt{\phi_1^2 + 4\phi_2}}{-2\phi_2} \qquad [2.3.7]$$

$$z_2 = \frac{\phi_1 + \sqrt{\phi_1^2 + 4\phi_2}}{-2\phi_2}. \qquad [2.3.8]$$

Setting $z = z_1$ or z_2 makes the left side of [2.3.5] zero, while $z = \lambda_1^{-1}$ or λ_2^{-1} sets the right side of [2.3.5] to zero. Thus

$$\lambda_1^{-1} = z_1 \qquad [2.3.9]$$

$$\lambda_2^{-1} = z_2. \qquad [2.3.10]$$

Returning to the numerical example [2.3.4] in which $\phi_1 = 0.6$ and $\phi_2 = -0.08$, we would calculate

$$z_1 = \frac{0.6 - \sqrt{(0.6)^2 - 4(0.08)}}{2(0.08)} = 2.5$$

$$z_2 = \frac{0.6 + \sqrt{(0.6)^2 - 4(0.08)}}{2(0.08)} = 5.0,$$

and so

$$\lambda_1 = 1/(2.5) = 0.4$$

$$\lambda_2 = 1/(5.0) = 0.2,$$

as was found in [2.3.4].

When $\phi_1^2 + 4\phi_2 < 0$, the values z_1 and z_2 are complex conjugates, and their reciprocals λ_1 and λ_2 can be found by first writing the complex number in polar coordinate form. Specifically, write

$$z_1 = a + bi$$

as

$$z_1 = R\cdot[\cos(\theta) + i\cdot\sin(\theta)] = R\cdot e^{i\theta}.$$

Then

$$z_1^{-1} = R^{-1}\cdot e^{-i\theta} = R^{-1}\cdot[\cos(\theta) - i\cdot\sin(\theta)].$$

Actually, there is a more direct method for calculating the values of λ_1 and λ_2 from ϕ_1 and ϕ_2. Divide both sides of [2.3.5] by z^2:

$$(z^{-2} - \phi_1 z^{-1} - \phi_2) = (z^{-1} - \lambda_1)(z^{-1} - \lambda_2) \qquad [2.3.11]$$

and define λ to be the variable z^{-1}:

$$\lambda \equiv z^{-1}. \qquad [2.3.12]$$

Substituting [2.3.12] into [2.3.11] produces

$$(\lambda^2 - \phi_1\lambda - \phi_2) = (\lambda - \lambda_1)(\lambda - \lambda_2). \qquad [2.3.13]$$

Again, [2.3.13] must hold for all values of λ in order for the two sides of [2.3.5] to represent the same polynomial. The values of λ that set the right side to zero are $\lambda = \lambda_1$ and $\lambda = \lambda_2$. These same values must set the left side of [2.3.13] to zero as well:

$$(\lambda^2 - \phi_1\lambda - \phi_2) = 0. \qquad [2.3.14]$$

Thus, to calculate the values of λ_1 and λ_2 that factor the polynomial in [2.3.3], we can find the roots of [2.3.14] directly from the quadratic formula:

$$\lambda_1 = \frac{\phi_1 + \sqrt{\phi_1^2 + 4\phi_2}}{2} \qquad [2.3.15]$$

$$\lambda_2 = \frac{\phi_1 - \sqrt{\phi_1^2 + 4\phi_2}}{2}. \qquad [2.3.16]$$

For the example of [2.3.4], we would thus calculate

$$\lambda_1 = \frac{0.6 + \sqrt{(0.6)^2 - 4(0.08)}}{2} = 0.4$$

$$\lambda_2 = \frac{0.6 - \sqrt{(0.6)^2 - 4(0.08)}}{2} = 0.2.$$

It is instructive to compare these results with those in Chapter 1. There the dynamics of the second-order difference equation [2.3.1] were summarized by calculating the eigenvalues of the matrix **F** given by

$$\mathbf{F} = \begin{bmatrix} \phi_1 & \phi_2 \\ 1 & 0 \end{bmatrix}. \qquad [2.3.17]$$

The eigenvalues of **F** were seen to be the two values of λ that satisfy equation [1.2.13]:

$$(\lambda^2 - \phi_1\lambda - \phi_2) = 0.$$

But this is the same calculation as in [2.3.14]. This finding is summarized in the following proposition.

Proposition 2.1: *Factoring the polynomial* $(1 - \phi_1 L - \phi_2 L^2)$ *as*

$$(1 - \phi_1 L - \phi_2 L^2) = (1 - \lambda_1 L)(1 - \lambda_2 L) \qquad [2.3.18]$$

is the same calculation as finding the eigenvalues of the matrix **F** *in* [*2.3.17*]. *The eigenvalues* λ_1 *and* λ_2 *of* **F** *are the same as the parameters* λ_1 *and* λ_2 *in* [*2.3.18*], *and are given by equations* [*2.3.15*] *and* [*2.3.16*].

The correspondence between calculating the eigenvalues of a matrix and factoring a polynomial in the lag operator is very instructive. However, it introduces one minor source of possible semantic confusion about which we have to be careful. Recall from Chapter 1 that the system [2.3.1] is stable if both λ_1 and λ_2 are less than 1 in modulus and explosive if either λ_1 or λ_2 is greater than 1 in modulus. Sometimes this is described as the requirement that the roots of

$$(\lambda^2 - \phi_1\lambda - \phi_2) = 0 \qquad [2.3.19]$$

lie inside the unit circle. The possible confusion is that it is often convenient to work directly with the polynomial in the form in which it appears in [2.3.2],

$$(1 - \phi_1 z - \phi_2 z^2) = 0, \qquad [2.3.20]$$

whose roots, we have seen, are the reciprocals of those of [2.3.19]. Thus, we could say with equal accuracy that "the difference equation [2.3.1] is stable whenever the roots of [2.3.19] lie *inside* the unit circle" or that "the difference equation [2.3.1] is stable whenever the roots of [2.3.20] lie *outside* the unit circle." The two statements mean exactly the same thing. Some scholars refer simply to the "roots of the difference equation [2.3.1]," though this raises the possibility of confusion between [2.3.19] and [2.3.20]. This book will follow the convention of using the term "eigenvalues" to refer to the roots of [2.3.19]. Wherever the term "roots" is used, we will indicate explicitly the equation whose roots are being described.

From here on in this section, it is assumed that the second-order difference equation is stable, with the eigenvalues λ_1 and λ_2 distinct and both inside the unit circle. Where this is the case, the inverses

$$(1 - \lambda_1 L)^{-1} = 1 + \lambda_1^1 L + \lambda_1^2 L^2 + \lambda_1^3 L^3 + \cdots$$
$$(1 - \lambda_2 L)^{-1} = 1 + \lambda_2^1 L + \lambda_2^2 L^2 + \lambda_2^3 L^3 + \cdots$$

are well defined for bounded sequences. Write [2.3.2] in factored form:

$$(1 - \lambda_1 L)(1 - \lambda_2 L)y_t = w_t$$

and operate on both sides by $(1 - \lambda_1 L)^{-1}(1 - \lambda_2 L)^{-1}$:

$$y_t = (1 - \lambda_1 L)^{-1}(1 - \lambda_2 L)^{-1} w_t. \qquad [2.3.21]$$

Following Sargent (1987, p. 184), when $\lambda_1 \neq \lambda_2$, we can use the following operator:

$$(\lambda_1 - \lambda_2)^{-1}\left\{\frac{\lambda_1}{1 - \lambda_1 L} - \frac{\lambda_2}{1 - \lambda_2 L}\right\}. \qquad [2.3.22]$$

Notice that this is simply another way of writing the operator in [2.3.21]:

$$(\lambda_1 - \lambda_2)^{-1}\left\{\frac{\lambda_1}{1 - \lambda_1 L} - \frac{\lambda_2}{1 - \lambda_2 L}\right\}$$

$$= (\lambda_1 - \lambda_2)^{-1}\left\{\frac{\lambda_1(1 - \lambda_2 L) - \lambda_2(1 - \lambda_1 L)}{(1 - \lambda_1 L)\cdot(1 - \lambda_2 L)}\right\}$$

$$= \frac{1}{(1 - \lambda_1 L)\cdot(1 - \lambda_2 L)}.$$

Thus, [2.3.21] can be written as

$$y_t = (\lambda_1 - \lambda_2)^{-1}\left\{\frac{\lambda_1}{1 - \lambda_1 L} - \frac{\lambda_2}{1 - \lambda_2 L}\right\}w_t$$

$$= \left\{\frac{\lambda_1}{\lambda_1 - \lambda_2}[1 + \lambda_1 L + \lambda_1^2 L^2 + \lambda_1^3 L^3 + \cdots]\right.$$

$$\left. - \frac{\lambda_2}{\lambda_1 - \lambda_2}[1 + \lambda_2 L + \lambda_2^2 L^2 + \lambda_2^3 L^3 + \cdots]\right\}w_t$$

or

$$y_t = [c_1 + c_2]w_t + [c_1\lambda_1 + c_2\lambda_2]w_{t-1} + [c_1\lambda_1^2 + c_2\lambda_2^2]w_{t-2} + [c_1\lambda_1^3 + c_2\lambda_2^3]w_{t-3} + \cdots, \qquad [2.3.23]$$

where

$$c_1 = \lambda_1/(\lambda_1 - \lambda_2) \qquad [2.3.24]$$

$$c_2 = -\lambda_2/(\lambda_1 - \lambda_2). \qquad [2.3.25]$$

From [2.3.23] the dynamic multiplier can be read off directly as

$$\frac{\partial y_{t+j}}{\partial w_t} = c_1\lambda_1^j + c_2\lambda_2^j,$$

the same result arrived at in equations [1.2.24] and [1.2.25].

2.4. *pth-Order Difference Equations*

These techniques generalize in a straightforward way to a pth-order difference equation of the form

$$y_t = \phi_1 y_{t-1} + \phi_2 y_{t-2} + \cdots + \phi_p y_{t-p} + w_t. \qquad [2.4.1]$$

Write [2.4.1] in terms of lag operators as

$$(1 - \phi_1 L - \phi_2 L^2 - \cdots - \phi_p L^p)y_t = w_t. \qquad [2.4.2]$$

Factor the operator on the left side of [2.4.2] as

$$(1 - \phi_1 L - \phi_2 L^2 - \cdots - \phi_p L^p) = (1 - \lambda_1 L)(1 - \lambda_2 L)\cdots(1 - \lambda_p L). \qquad [2.4.3]$$

This is the same as finding the values of $(\lambda_1, \lambda_2, \ldots, \lambda_p)$ such that the following polynomials are the same for all z:

$$(1 - \phi_1 z - \phi_2 z^2 - \cdots - \phi_p z^p) = (1 - \lambda_1 z)(1 - \lambda_2 z)\cdots(1 - \lambda_p z).$$

As in the second-order system, we multiply both sides of this equation by z^{-p} and define $\lambda \equiv z^{-1}$:

$$(\lambda^p - \phi_1\lambda^{p-1} - \phi_2\lambda^{p-2} - \cdots - \phi_{p-1}\lambda - \phi_p) = (\lambda - \lambda_1)(\lambda - \lambda_2)\cdots(\lambda - \lambda_p). \qquad [2.4.4]$$

Clearly, setting $\lambda = \lambda_i$ for $i = 1, 2, \ldots,$ or p causes the right side of [2.4.4] to equal zero. Thus the values $(\lambda_1, \lambda_2, \ldots, \lambda_p)$ must be the numbers that set the left side of expression [2.4.4] to zero as well:

$$\lambda_p - \phi_1\lambda^{p-1} - \phi_2\lambda^{p-2} - \cdots - \phi_{p-1}\lambda - \phi_p = 0. \qquad [2.4.5]$$

This expression again is identical to that given in Proposition 1.1, which characterized the eigenvalues $(\lambda_1, \lambda_2, \ldots, \lambda_p)$ of the matrix **F** defined in equation [1.2.3]. Thus, Proposition 2.1 readily generalizes.

Proposition 2.2: *Factoring a pth-order polynomial in the lag operator,*

$$(1 - \phi_1 L - \phi_2 L^2 - \cdots - \phi_p L^p) = (1 - \lambda_1 L)(1 - \lambda_2 L)\cdots(1 - \lambda_p L),$$

is the same calculation as finding the eigenvalues of the matrix **F** *defined in [1.2.3]. The eigenvalues* $(\lambda_1, \lambda_2, \ldots, \lambda_p)$ *of* **F** *are the same as the parameters* $(\lambda_1, \lambda_2, \ldots, \lambda_p)$ *in [2.4.3] and are given by the solutions to equation [2.4.5].*

The difference equation [2.4.1] is stable if the eigenvalues (the roots of [2.4.5]) lie inside the unit circle, or equivalently if the roots of

$$1 - \phi_1 z - \phi_2 z^2 - \cdots - \phi_p z^p = 0 \qquad [2.4.6]$$

lie outside the unit circle.

Assuming that the eigenvalues are inside the unit circle and that we are restricting ourselves to considering bounded sequences, the inverses $(1 - \lambda_1 L)^{-1}$, $(1 - \lambda_2 L)^{-1}, \ldots, (1 - \lambda_p L)^{-1}$ all exist, permitting the difference equation

$$(1 - \lambda_1 L)(1 - \lambda_2 L)\cdots(1 - \lambda_p L)y_t = w_t$$

to be written as

$$y_t = (1 - \lambda_1 L)^{-1}(1 - \lambda_2 L)^{-1}\cdots(1 - \lambda_p L)^{-1}w_t. \qquad [2.4.7]$$

Provided further that the eigenvalues $(\lambda_1, \lambda_2, \ldots, \lambda_p)$ are all distinct, the polynomial associated with the operator on the right side of [2.4.7] can again be expanded with partial fractions:

$$\frac{1}{(1 - \lambda_1 z)(1 - \lambda_2 z)\cdots(1 - \lambda_p z)} = \frac{c_1}{(1 - \lambda_1 z)} + \frac{c_2}{(1 - \lambda_2 z)} + \cdots + \frac{c_p}{(1 - \lambda_p z)}. \qquad [2.4.8]$$

Following Sargent (1987, pp. 192–93), the values of $(c_1, c_2, \ldots, c_p)$ that make [2.4.8] true can be found by multiplying both sides by $(1 - \lambda_1 z)(1 - \lambda_2 z)\cdots(1 - \lambda_p z)$:

$$\begin{aligned} 1 = {} & c_1(1 - \lambda_2 z)(1 - \lambda_3 z)\cdots(1 - \lambda_p z) \\ & + c_2(1 - \lambda_1 z)(1 - \lambda_3 z)\cdots(1 - \lambda_p z) + \cdots \\ & + c_p(1 - \lambda_1 z)(1 - \lambda_2 z)\cdots(1 - \lambda_{p-1} z). \end{aligned} \qquad [2.4.9]$$

Equation [2.4.9] has to hold for all values of z. Since it is a $(p - 1)$th-order polynomial, if $(c_1, c_2, \ldots, c_p)$ are chosen so that [2.4.9] holds for p particular

distinct values of z, then [2.4.9] must hold for all z. To ensure that [2.4.9] holds at $z = \lambda_1^{-1}$ requires that

$$1 = c_1(1 - \lambda_2\lambda_1^{-1})(1 - \lambda_3\lambda_1^{-1}) \cdots (1 - \lambda_p\lambda_1^{-1})$$

or

$$c_1 = \frac{\lambda_1^{p-1}}{(\lambda_1 - \lambda_2)(\lambda_1 - \lambda_3) \cdots (\lambda_1 - \lambda_p)}. \qquad [2.4.10]$$

For [2.4.9] to hold for $z = \lambda_2^{-1}, \lambda_3^{-1}, \ldots, \lambda_p^{-1}$ requires

$$c_2 = \frac{\lambda_2^{p-1}}{(\lambda_2 - \lambda_1)(\lambda_2 - \lambda_3) \cdots (\lambda_2 - \lambda_p)} \qquad [2.4.11]$$

$$\vdots$$

$$c_p = \frac{\lambda_p^{p-1}}{(\lambda_p - \lambda_1)(\lambda_p - \lambda_2) \cdots (\lambda_p - \lambda_{p-1})}. \qquad [2.4.12]$$

Note again that these are identical to expression [1.2.25] in Chapter 1. Recall from the discussion there that $c_1 + c_2 + \cdots + c_p = 1$.

To conclude, [2.4.7] can be written

$$\begin{aligned} y_t &= \frac{c_1}{(1 - \lambda_1 L)} w_t + \frac{c_2}{(1 - \lambda_2 L)} w_t + \cdots + \frac{c_p}{(1 - \lambda_p L)} w_t \\ &= c_1(1 + \lambda_1 L + \lambda_1^2 L^2 + \lambda_1^3 L^3 + \cdots)w_t + c_2(1 + \lambda_2 L + \lambda_2^2 L^2 + \lambda_2^3 L^3 + \cdots)w_t \\ &\quad + \cdots + c_p(1 + \lambda_p L + \lambda_p^2 L^2 + \lambda_p^3 L^3 + \cdots)w_t \end{aligned}$$

or

$$\begin{aligned} y_t = {} & [c_1 + c_2 + \cdots + c_p]w_t + [c_1\lambda_1 + c_2\lambda_2 + \cdots + c_p\lambda_p]w_{t-1} \\ & + [c_1\lambda_1^2 + c_2\lambda_2^2 + \cdots + c_p\lambda_p^2]w_{t-2} \\ & + [c_1\lambda_1^3 + c_2\lambda_2^3 + \cdots + c_p\lambda_p^3]w_{t-3} + \cdots \end{aligned} \qquad [2.4.13]$$

where $(c_1, c_2, \ldots, c_p)$ are given by equations [2.4.10] through [2.4.12]. Again, the dynamic multiplier can be read directly off [2.4.13]:

$$\frac{\partial y_{t+j}}{\partial w_t} = [c_1\lambda_1^j + c_2\lambda_2^j + \cdots + c_p\lambda_p^j], \qquad [2.4.14]$$

reproducing the result from Chapter 1.

There is a very convenient way to calculate the effect of w on the present value of y using the lag operator representation. Write [2.4.13] as

$$y_t = \psi_0 w_t + \psi_1 w_{t-1} + \psi_2 w_{t-2} + \psi_3 w_{t-3} + \cdots \qquad [2.4.15]$$

where

$$\psi_j = [c_1\lambda_1^j + c_2\lambda_2^j + \cdots + c_p\lambda_p^j]. \qquad [2.4.16]$$

Next rewrite [2.4.15] in lag operator notation as

$$y_t = \psi(L)w_t, \qquad [2.4.17]$$

where $\psi(L)$ denotes an infinite-order polynomial in the lag operator:

$$\psi(L) = \psi_0 + \psi_1 L + \psi_2 L^2 + \psi_3 L^3 + \cdots.$$

Notice that ψ_j is the dynamic multiplier [2.4.14]. The effect of w_t on the present value of y is given by

$$\begin{aligned}\frac{\partial \sum_{j=0}^{\infty} \beta^j y_{t+j}}{\partial w_t} &= \sum_{j=0}^{\infty} \beta^j \frac{\partial y_{t+j}}{\partial w_t} \\ &= \sum_{j=0}^{\infty} \beta^j \psi_j.\end{aligned} \tag{2.4.18}$$

Thinking of $\psi(z)$ as a polynomial in a real number z,

$$\psi(z) = \psi_0 + \psi_1 z + \psi_2 z^2 + \psi_3 z^3 + \cdots,$$

it appears that the multiplier [2.4.18] is simply this polynomial evaluated at $z = \beta$:

$$\frac{\partial \sum_{j=0}^{\infty} \beta^j y_{t+j}}{\partial w_t} = \psi(\beta) = \psi_0 + \psi_1\beta + \psi_2\beta^2 + \psi_3\beta^3 + \cdots. \tag{2.4.19}$$

But comparing [2.4.17] with [2.4.7], it is apparent that

$$\psi(L) = [(1 - \lambda_1 L)(1 - \lambda_2 L) \cdots (1 - \lambda_p L)]^{-1},$$

and from [2.4.3] this means that

$$\psi(L) = [1 - \phi_1 L - \phi_2 L^2 - \cdots - \phi_p L^p]^{-1}.$$

We conclude that

$$\psi(z) = [1 - \phi_1 z - \phi_2 z^2 - \cdots - \phi_p z^p]^{-1}$$

for any value of z, so, in particular,

$$\psi(\beta) = [1 - \phi_1\beta - \phi_2\beta^2 - \cdots - \phi_p\beta^p]^{-1}. \tag{2.4.20}$$

Substituting [2.4.20] into [2.4.19] reveals that

$$\frac{\partial \sum_{j=0}^{\infty} \beta^j y_{t+j}}{\partial w_t} = \frac{1}{1 - \phi_1\beta - \phi_2\beta^2 - \cdots - \phi_p\beta^p}, \tag{2.4.21}$$

reproducing the claim in Proposition 1.3. Again, the long-run multiplier obtains as the special case of [2.4.21] with $\beta = 1$:

$$\lim_{j\to\infty}\left[\frac{\partial y_{t+j}}{\partial w_t} + \frac{\partial y_{t+j}}{\partial w_{t+1}} + \cdots + \frac{\partial y_{t+j}}{\partial w_{t+j}}\right] = \frac{1}{1 - \phi_1 - \phi_2 - \cdots - \phi_p}.$$

2.5. *Initial Conditions and Unbounded Sequences*

Section 1.2 analyzed the following problem. Given a pth-order difference equation

$$y_t = \phi_1 y_{t-1} + \phi_2 y_{t-2} + \cdots + \phi_p y_{t-p} + w_t, \tag{2.5.1}$$

p initial values of y,

$$y_{-1}, y_{-2}, \ldots, y_{-p}, \tag{2.5.2}$$

and a sequence of values for the input variable w,

$$\{w_0, w_1, \ldots, w_t\}, \tag{2.5.3}$$

we sought to calculate the sequence of values for the output variable y:

$$\{y_0, y_1, \ldots, y_t\}.$$

Certainly there are systems where the question is posed in precisely this form. We may know the equation of motion for the system [2.5.1] and its current state [2.5.2] and wish to characterize the values that $\{y_0, y_1, \ldots, y_t\}$ might take on for different specifications of $\{w_0, w_1, \ldots, w_t\}$.

However, there are many examples in economics and finance in which a theory specifies just the equation of motion [2.5.1] and a sequence of driving variables [2.5.3]. Clearly, these two pieces of information alone are insufficient to determine the sequence $\{y_0, y_1, \ldots, y_t\}$, and some additional theory beyond that contained in the difference equation [2.5.1] is needed to describe fully the dependence of y on w. These additional restrictions can be of interest in their own right and also help give some insight into some of the technical details of manipulating difference equations. For these reasons, this section discusses in some depth an example of the role of initial conditions and their implications for solving difference equations.

Let P_t denote the price of a stock and D_t its dividend payment. If an investor buys the stock at date t and sells it at $t + 1$, the investor will earn a yield of D_t/P_t from the dividend and a yield of $(P_{t+1} - P_t)/P_t$ in capital gains. The investor's total return (r_{t+1}) is thus

$$r_{t+1} = (P_{t+1} - P_t)/P_t + D_t/P_t.$$

A very simple model of the stock market posits that the return investors earn on stocks is constant across time periods:

$$r = (P_{t+1} - P_t)/P_t + D_t/P_t \qquad r > 0. \qquad [2.5.4]$$

Equation [2.5.4] may seem too simplistic to be of much practical interest; it assumes among other things that investors have perfect foresight about future stock prices and dividends. However, a slightly more realistic model in which *expected* stock returns are constant involves a very similar set of technical issues. The advantage of the perfect-foresight model [2.5.4] is that it can be discussed using the tools already in hand to gain some further insight into using lag operators to solve difference equations.

Multiply [2.5.4] by P_t to arrive at

$$rP_t = P_{t+1} - P_t + D_t$$

or

$$P_{t+1} = (1 + r)P_t - D_t. \qquad [2.5.5]$$

Equation [2.5.5] will be recognized as a first-order difference equation of the form of [1.1.1] with $y_t = P_{t+1}$, $\phi = (1 + r)$, and $w_t = -D_t$. From [1.1.7], we know that [2.5.5] implies that

$$\begin{aligned} P_{t+1} = {} & (1 + r)^{t+1}P_0 - (1 + r)^t D_0 - (1 + r)^{t-1}D_1 - (1 + r)^{t-2}D_2 \\ & - \cdots - (1 + r)D_{t-1} - D_t. \end{aligned} \qquad [2.5.6]$$

If the sequence $\{D_0, D_1, \ldots, D_t\}$ *and* the value of P_0 were given, then [2.5.6] could determine the values of $\{P_1, P_2, \ldots, P_{t+1}\}$. But if only the values $\{D_0, D_1, \ldots, D_t\}$ are given, then equation [2.5.6] would not be enough to pin down $\{P_1, P_2, \ldots, P_{t+1}\}$. There are an infinite number of possible sequences $\{P_1, P_2, \ldots, P_{t+1}\}$ consistent with [2.5.5] and with a given $\{D_0, D_1, \ldots, D_t\}$. This infinite number of possibilities is indexed by the initial value P_0.

A further simplifying assumption helps clarify the nature of these different paths for $\{P_1, P_2, \ldots, P_{t+1}\}$. Suppose that dividends are constant over time:

$$D_t = D \qquad \text{for all } t.$$

Then [2.5.6] becomes

$$\begin{aligned} P_{t+1} &= (1 + r)^{t+1}P_0 - [(1 + r)^t + (1 + r)^{t-1} \\ &\quad + \cdots + (1 + r) + 1]D \\ &= (1 + r)^{t+1}P_0 - \frac{1 - (1 + r)^{t+1}}{1 - (1 + r)} D \\ &= (1 + r)^{t+1}[P_0 - (D/r)] + (D/r). \end{aligned} \tag{2.5.7}$$

Consider first the solution in which $P_0 = D/r$. If the initial stock price should happen to take this value, then [2.5.7] implies that

$$P_t = D/r \tag{2.5.8}$$

for all t. In this solution, dividends are constant at D and the stock price is constant at D/r. With no change in stock prices, investors never have any capital gains or losses, and their return is solely the dividend yield $D/P = r$. In a world with no changes in dividends this seems to be a sensible expression of the theory represented by [2.5.4]. Equation [2.5.8] is sometimes described as the "market fundamentals" solution to [2.5.4] for the case of constant dividends.

However, even with constant dividends, equation [2.5.8] is not the only result consistent with [2.5.4]. Suppose that the initial price exceeded D/r:

$$P_0 > D/r.$$

Investors seem to be valuing the stock beyond the potential of its constant dividend stream. From [2.5.7] this could be consistent with the asset pricing theory [2.5.4] provided that P_1 exceeds D/r by an even larger amount. As long as investors all believe that prices will continue to rise over time, each will earn the required return r from the realized capital gain and [2.5.4] will be satisfied. This scenario has reminded many economists of a speculative bubble in stock prices.

If such bubbles are to be ruled out, additional knowledge about the process for $\{P_t\}_{t=-\infty}^{\infty}$ is required beyond that contained in the theory of [2.5.4]. For example, we might argue that finite world resources put an upper limit on feasible stock prices, as in

$$|P_t| < \overline{P} \qquad \text{for all } t. \tag{2.5.9}$$

Then the only sequence for $\{P_t\}_{t=-\infty}^{\infty}$ consistent with both [2.5.4] and [2.5.9] would be the market fundamentals solution [2.5.8].

Let us now relax the assumption that dividends are constant and replace it with the assumption that $\{D_t\}_{t=-\infty}^{\infty}$ is a bounded sequence. What path for $\{P_t\}_{t=-\infty}^{\infty}$ in [2.5.6] is consistent with [2.5.9] in this case? The answer can be found by returning to the difference equation [2.5.5]. We arrived at the form [2.5.6] by recursively substituting this equation backward. That is, we used the fact that [2.5.5] held for dates $t, t - 1, t - 2, \ldots, 0$ and recursively substituted to arrive at [2.5.6] as a logical implication of [2.5.5]. Equation [2.5.5] could equally well be solved recursively *forward*. To do so, equation [2.5.5] is written as

$$P_t = \frac{1}{1 + r}[P_{t+1} + D_t]. \tag{2.5.10}$$

An analogous equation must hold for date $t + 1$:

$$P_{t+1} = \frac{1}{1 + r}[P_{t+2} + D_{t+1}]. \qquad [2.5.11]$$

Substitute [2.5.11] into [2.5.10] to deduce

$$\begin{aligned} P_t &= \frac{1}{1 + r}\left[\frac{1}{1 + r}[P_{t+2} + D_{t+1}] + D_t\right] \\ &= \left[\frac{1}{1 + r}\right]^2 P_{t+2} + \left[\frac{1}{1 + r}\right]^2 D_{t+1} + \left[\frac{1}{1 + r}\right] D_t. \end{aligned} \qquad [2.5.12]$$

Using [2.5.10] for date $t + 2$,

$$P_{t+2} = \frac{1}{1 + r}[P_{t+3} + D_{t+2}],$$

and substituting into [2.5.12] gives

$$P_t = \left[\frac{1}{1 + r}\right]^3 P_{t+3} + \left[\frac{1}{1 + r}\right]^3 D_{t+2} + \left[\frac{1}{1 + r}\right]^2 D_{t+1} + \left[\frac{1}{1 + r}\right] D_t.$$

Continuing in this fashion T periods into the future produces

$$\begin{aligned} P_t = &\left[\frac{1}{1 + r}\right]^T P_{t+T} + \left[\frac{1}{1 + r}\right]^T D_{t+T-1} + \left[\frac{1}{1 + r}\right]^{T-1} D_{t+T-2} \\ &+ \cdots + \left[\frac{1}{1 + r}\right]^2 D_{t+1} + \left[\frac{1}{1 + r}\right] D_t. \end{aligned} \qquad [2.5.13]$$

If the sequence $\{P_t\}_{t=-\infty}^{\infty}$ is to satisfy [2.5.9], then

$$\lim_{T\to\infty}\left[\frac{1}{1 + r}\right]^T P_{t+T} = 0.$$

If $\{D_t\}_{t=-\infty}^{\infty}$ is likewise a bounded sequence, then the following limit exists:

$$\lim_{T\to\infty}\sum_{j=0}^{T}\left[\frac{1}{1 + r}\right]^{j+1} D_{t+j}.$$

Thus, if $\{P_t\}_{t=-\infty}^{\infty}$ is to be a bounded sequence, then we can take the limit of [2.5.13] as $T \to \infty$ to conclude

$$P_t = \sum_{j=0}^{\infty}\left[\frac{1}{1 + r}\right]^{j+1} D_{t+j}, \qquad [2.5.14]$$

which is referred to as the "market fundamentals" solution of [2.5.5] for the general case of time-varying dividends. Notice that [2.5.14] produces [2.5.8] as a special case when $D_t = D$ for all t.

Describing the value of a variable at time t as a function of future realizations of another variable as in [2.5.14] may seem an artifact of assuming a perfect-foresight model of stock prices. However, an analogous set of operations turns out to be appropriate in a system similar to [2.5.4] in which expected returns are constant.[1] In such systems [2.5.14] generalizes to

$$P_t = \sum_{j=0}^{\infty}\left[\frac{1}{1 + r}\right]^{j+1} E_t D_{t+j},$$

[1]See Sargent (1987) and Whiteman (1983) for an introduction to the manipulation of difference equations involving expectations.

where E_t denotes an expectation of an unknown future quantity based on information available to investors at date t.

Expression [2.5.14] determines the particular value for the initial price P_0 that is consistent with the boundedness condition [2.5.9]. Setting $t = 0$ in [2.5.14] and substituting into [2.5.6] produces

$$\begin{aligned} P_{t+1} &= (1 + r)^{t+1}\left\{\left[\frac{1}{1 + r}\right]D_0 + \left[\frac{1}{1 + r}\right]^2 D_1 + \left[\frac{1}{1 + r}\right]^3 D_2 \right. \\ &\quad \left. + \cdots + \left[\frac{1}{1 + r}\right]^{t+1} D_t + \left[\frac{1}{1 + r}\right]^{t+2} D_{t+1} + \cdots\right\} - (1 + r)^t D_0 \\ &\quad - (1 + r)^{t-1}D_1 - (1 + r)^{t-2}D_2 - \cdots - (1 + r)D_{t-1} - D_t \\ &= \left[\frac{1}{1 + r}\right]D_{t+1} + \left[\frac{1}{1 + r}\right]^2 D_{t+2} + \left[\frac{1}{1 + r}\right]^3 D_{t+3} + \cdots . \end{aligned}$$

Thus, setting the initial condition P_0 to satisfy [2.5.14] is sufficient to ensure that it holds for all t. Choosing P_0 equal to any other value would cause the consequences of each period's dividends to accumulate over time so as to lead to a violation of [2.5.9] eventually.

It is useful to discuss these same calculations from the perspective of lag operators. In Section 2.2 the recursive substitution backward that led from [2.5.5] to [2.5.6] was represented by writing [2.5.5] in terms of lag operators as

$$[1 - (1 + r)L]P_{t+1} = -D_t \qquad [2.5.15]$$

and multiplying both sides of [2.5.15] by the following operator:

$$[1 + (1 + r)L + (1 + r)^2L^2 + \cdots + (1 + r)^tL^t]. \qquad [2.5.16]$$

If $(1 + r)$ were less than unity, it would be natural to consider the limit of [2.5.16] as $t \to \infty$:

$$[1 - (1 + r)L]^{-1} = 1 + (1 + r)L + (1 + r)^2L^2 + \cdots .$$

In the case of the theory of stock returns discussed here, however, $r > 0$ and this operator is not defined. In this case, a lag operator representation can be sought for the recursive substitution forward that led from [2.5.5] to [2.5.13]. This is accomplished using the inverse of the lag operator,

$$L^{-1}w_t = w_{t+1},$$

which extends result [2.1.4] to negative values of k. Note that L^{-1} is indeed the inverse of the operator L:

$$L^{-1}(Lw_t) = L^{-1}w_{t-1} = w_t.$$

In general,

$$L^{-k}L^j = L^{j-k},$$

with L^0 defined as the identity operator:

$$L^0w_t = w_t.$$

Now consider multiplying [2.5.15] by

$$[1 + (1 + r)^{-1}L^{-1} + (1 + r)^{-2}L^{-2} + \cdots + (1 + r)^{-(T-1)}L^{-(T-1)}] \times [-(1 + r)^{-1}L^{-1}] \qquad [2.5.17]$$

to obtain

$$\begin{aligned}&[1 + (1 + r)^{-1}L^{-1} + (1 + r)^{-2}L^{-2} + \cdots + (1 + r)^{-(T-1)}L^{-(T-1)}]\\ &\times [1 - (1 + r)^{-1}L^{-1}]P_{t+1}\\ &= [1 + (1 + r)^{-1}L^{-1} + (1 + r)^{-2}L^{-2} + \cdots\\ &\quad + (1 + r)^{-(T-1)}L^{-(T-1)}] \times (1 + r)^{-1}D_{t+1}\end{aligned}$$

or

$$\begin{aligned}[1 - (1 + r)^{-T}L^{-T}]P_{t+1} &= \left[\frac{1}{1 + r}\right]D_{t+1} + \left[\frac{1}{1 + r}\right]^2 D_{t+2}\\ &\quad + \left[\frac{1}{1 + r}\right]^3 D_{t+3} + \cdots + \left[\frac{1}{1 + r}\right]^T D_{t+T},\end{aligned}$$

which is identical to [2.5.13] with t in [2.5.13] replaced with $t + 1$.

When $r > 0$ and $\{P_t\}_{t=-\infty}^{\infty}$ is a bounded sequence, the left side of the preceding equation will approach P_{t+1} as T becomes large. Thus, when $r > 0$ and $\{P_t\}_{t=-\infty}^{\infty}$ and $\{D_t\}_{t=-\infty}^{\infty}$ are bounded sequences, the limit of the operator in [2.5.17] exists and could be viewed as the inverse of the operator on the left side of [2.5.15]:

$$\begin{aligned}[1 - (1 + r)L]^{-1} &= -(1 + r)^{-1}L^{-1}\\ &\quad \times [1 + (1 + r)^{-1}L^{-1} + (1 + r)^{-2}L^{-2} + \cdots].\end{aligned}$$

Applying this limiting operator to [2.5.15] amounts to solving the difference equation forward as in [2.5.14] and selecting the market fundamentals solution among the set of possible time paths for $\{P_t\}_{t=-\infty}^{\infty}$ given a particular time path for dividends $\{D_t\}_{t=-\infty}^{\infty}$.

Thus, given a first-order difference equation of the form

$$(1 - \phi L)y_t = w_t, \qquad [2.5.18]$$

Sargent's (1987) advice was to solve the equation "backward" when $|\phi| < 1$ by multiplying by

$$[1 - \phi L]^{-1} = [1 + \phi L + \phi^2L^2 + \phi^3L^3 + \cdots] \qquad [2.5.19]$$

and to solve the equation "forward" when $|\phi| > 1$ by multiplying by

$$\begin{aligned}[1 - \phi L]^{-1} &= \frac{-\phi^{-1}L^{-1}}{1 - \phi^{-1}L^{-1}}\\ &= -\phi^{-1}L^{-1}[1 + \phi^{-1}L^{-1} + \phi^{-2}L^{-2} + \phi^{-3}L^{-3} + \cdots].\end{aligned} \qquad [2.5.20]$$

Defining the inverse of $[1 - \phi L]$ in this way amounts to selecting an operator $[1 - \phi L]^{-1}$ with the properties that

$$[1 - \phi L]^{-1} \times [1 - \phi L] = 1 \qquad \text{(the identity operator)}$$

and that, when it is applied to a bounded sequence $\{w_t\}_{t=-\infty}^{\infty}$,

$$[1 - \phi L]^{-1}w_t,$$

the result is another bounded sequence.

The conclusion from this discussion is that in applying an operator such as $[1 - \phi L]^{-1}$, we are implicitly imposing a boundedness assumption that rules out

phenomena such as the speculative bubbles of equation [2.5.7] a priori. Where that is our intention, so much the better, though we should not apply the rules [2.5.19] or [2.5.20] without some reflection on their economic content.

Chapter 2 References

Sargent, Thomas J. 1987. *Macroeconomic Theory*, 2d ed. Boston: Academic Press.

Whiteman, Charles H. 1983. *Linear Rational Expectations Models: A User's Guide*. Minneapolis: University of Minnesota Press.

3 Stationary ARMA Processes

This chapter introduces univariate *ARMA* processes, which provide a very useful class of models for describing the dynamics of an individual time series. The chapter begins with definitions of some of the key concepts used in time series analysis. Sections 3.2 through 3.5 then investigate the properties of various *ARMA* processes. Section 3.6 introduces the autocovariance-generating function, which is useful for analyzing the consequences of combining different time series and for an understanding of the population spectrum. The chapter concludes with a discussion of invertibility (Section 3.7), which can be important for selecting the *ARMA* representation of an observed time series that is appropriate given the uses to be made of the model.

3.1. *Expectations, Stationarity, and Ergodicity*

Expectations and Stochastic Processes

Suppose we have observed a sample of size T of some random variable Y_t:

$$\{y_1, y_2, \ldots, y_T\}. \qquad [3.1.1]$$

For example, consider a collection of T independent and identically distributed (i.i.d.) variables ε_t,

$$\{\varepsilon_1, \varepsilon_2, \ldots, \varepsilon_T\}, \qquad [3.1.2]$$

with

$$\varepsilon_t \sim N(0, \sigma^2).$$

This is referred to as a sample of size T from a *Gaussian white noise* process.

The observed sample [3.1.1] represents T particular numbers, but this set of T numbers is only one possible outcome of the underlying stochastic process that generated the data. Indeed, even if we were to imagine having observed the process for an infinite period of time, arriving at the sequence

$$\{y_t\}_{t=-\infty}^{\infty} = \{\ldots, y_{-1}, y_0, y_1, y_2, \ldots, y_T, y_{T+1}, y_{T+2}, \ldots\},$$

the infinite sequence $\{y_t\}_{t=-\infty}^{\infty}$ would still be viewed as a single realization from a time series process. For example, we might set one computer to work generating an infinite sequence of i.i.d. $N(0, \sigma^2)$ variates, $\{\varepsilon_t^{(1)}\}_{t=-\infty}^{\infty}$, and a second computer generating a separate sequence, $\{\varepsilon_t^{(2)}\}_{t=-\infty}^{\infty}$. We would then view these as two independent realizations of a Gaussian white noise process.

Imagine a battery of I such computers generating sequences $\{y_t^{(1)}\}_{t=-\infty}^{\infty}$, $\{y_t^{(2)}\}_{t=-\infty}^{\infty}, \ldots, \{y_t^{(I)}\}_{t=-\infty}^{\infty}$, and consider selecting the observation associated with date t from each sequence:

$$\{y_t^{(1)}, y_t^{(2)}, \ldots, y_t^{(I)}\}.$$

This would be described as a sample of I realizations of the random variable Y_t. This random variable has some density, denoted $f_{Y_t}(y_t)$, which is called the *unconditional density* of Y_t. For example, for the Gaussian white noise process, this density is given by

$$f_{Y_t}(y_t) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left[\frac{-y_t^2}{2\sigma^2}\right].$$

The *expectation* of the tth observation of a time series refers to the mean of this probability distribution, provided it exists:

$$E(Y_t) \equiv \int_{-\infty}^{\infty} y_t f_{Y_t}(y_t)\, dy_t. \qquad [3.1.3]$$

We might view this as the probability limit of the ensemble average:

$$E(Y_t) = \operatorname*{plim}_{I\to\infty} (1/I) \sum_{i=1}^{I} Y_t^{(i)}. \qquad [3.1.4]$$

For example, if $\{Y_t\}_{t=-\infty}^{\infty}$ represents the sum of a constant μ plus a Gaussian white noise process $\{\varepsilon_t\}_{t=-\infty}^{\infty}$,

$$Y_t = \mu + \varepsilon_t, \qquad [3.1.5]$$

then its mean is

$$E(Y_t) = \mu + E(\varepsilon_t) = \mu. \qquad [3.1.6]$$

If Y_t is a time trend plus Gaussian white noise,

$$Y_t = \beta t + \varepsilon_t, \qquad [3.1.7]$$

then its mean is

$$E(Y_t) = \beta t. \qquad [3.1.8]$$

Sometimes for emphasis the expectation $E(Y_t)$ is called the *unconditional mean* of Y_t. The unconditional mean is denoted μ_t:

$$E(Y_t) = \mu_t.$$

Note that this notation allows the general possibility that the mean can be a function of the date of the observation t. For the process [3.1.7] involving the time trend, the mean [3.1.8] is a function of time, whereas for the constant plus Gaussian white noise, the mean [3.1.6] is not a function of time.

The *variance* of the random variable Y_t (denoted γ_{0t}) is similarly defined as

$$\gamma_{0t} \equiv E(Y_t - \mu_t)^2 = \int_{-\infty}^{\infty} (y_t - \mu_t)^2 f_{Y_t}(y_t)\, dy_t. \qquad [3.1.9]$$

For example, for the process [3.1.7], the variance is

$$\gamma_{0t} = E(Y_t - \beta t)^2 = E(\varepsilon_t^2) = \sigma^2.$$

Autocovariance

Given a particular realization such as $\{y_t^{(1)}\}_{t=-\infty}^{\infty}$ on a time series process, consider constructing a vector $\mathbf{x}_t^{(1)}$ associated with date t. This vector consists of the $[j + 1]$ most recent observations on y as of date t for that realization:

$$\mathbf{x}_t^{(1)} \equiv \begin{bmatrix} y_t^{(1)} \\ y_{t-1}^{(1)} \\ \vdots \\ y_{t-j}^{(1)} \end{bmatrix}.$$

We think of each realization $\{y_t\}_{t=-\infty}^{\infty}$ as generating one particular value of the vector $\mathbf{x}_t$ and want to calculate the probability distribution of this vector $\mathbf{x}_t^{(i)}$ across realizations i. This distribution is called the *joint distribution* of $(Y_t, Y_{t-1}, \ldots, Y_{t-j})$. From this distribution we can calculate the jth *autocovariance* of Y_t (denoted γ_{jt}):

$$\begin{aligned} \gamma_{jt} &= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty}\cdots\int_{-\infty}^{\infty} (y_t - \mu_t)(y_{t-j} - \mu_{t-j}) \\ &\quad \times f_{Y_t, Y_{t-1}, \ldots, Y_{t-j}}(y_t, y_{t-1}, \ldots, y_{t-j})\, dy_t\, dy_{t-1} \cdots dy_{t-j} \qquad [3.1.10] \\ &= E(Y_t - \mu_t)(Y_{t-j} - \mu_{t-j}). \end{aligned}$$

Note that [3.1.10] has the form of a covariance between two variables X and Y:

$$\mathrm{Cov}(X, Y) = E(X - \mu_X)(Y - \mu_Y).$$

Thus [3.1.10] could be described as the covariance of Y_t with its own lagged value; hence, the term "autocovariance." Notice further from [3.1.10] that the 0th autocovariance is just the variance of Y_t, as anticipated by the notation γ_{0t} in [3.1.9].

The autocovariance γ_{jt} can be viewed as the $(1, j + 1)$ element of the variance-covariance matrix of the vector $\mathbf{x}_t$. For this reason, the autocovariances are described as the second moments of the process for Y_t.

Again it may be helpful to think of the jth autocovariance as the probability limit of an ensemble average:

$$\gamma_{jt} = \operatorname*{plim}_{I\to\infty} (1/I) \sum_{i=1}^{I} [Y_t^{(i)} - \mu_t] \cdot [Y_{t-j}^{(i)} - \mu_{t-j}]. \qquad [3.1.11]$$

As an example of calculating autocovariances, note that for the process in [3.1.5] the autocovariances are all zero for $j \neq 0$:

$$\gamma_{jt} = E(Y_t - \mu)(Y_{t-j} - \mu) = E(\varepsilon_t \varepsilon_{t-j}) = 0 \qquad \text{for } j \neq 0.$$

Stationarity

If neither the mean μ_t nor the autocovariances γ_{jt} depend on the date t, then the process for Y_t is said to be *covariance-stationary* or *weakly stationary*:

$$\begin{aligned} E(Y_t) &= \mu && \text{for all } t \\ E(Y_t - \mu)(Y_{t-j} - \mu) &= \gamma_j && \text{for all } t \text{ and any } j. \end{aligned}$$

For example, the process in [3.1.5] is covariance-stationary:

$$E(Y_t) = \mu$$

$$E(Y_t - \mu)(Y_{t-j} - \mu) = \begin{cases} \sigma^2 & \text{for } j = 0 \\ 0 & \text{for } j \neq 0. \end{cases}$$

By contrast, the process of [3.1.7] is not covariance-stationary, because its mean, βt, is a function of time.

Notice that if a process is covariance-stationary, the covariance between Y_t and Y_{t-j} depends only on j, the length of time separating the observations, and not on t, the date of the observation. It follows that for a covariance-stationary process, γ_j and γ_{-j} would represent the same magnitude. To see this, recall the definition

$$\gamma_j = E(Y_t - \mu)(Y_{t-j} - \mu). \qquad [3.1.12]$$

If the process is covariance-stationary, then this magnitude is the same for any value of t we might have chosen; for example, we can replace t with $t + j$:

$$\gamma_j = E(Y_{t+j} - \mu)(Y_{[t+j]-j} - \mu) = E(Y_{t+j} - \mu)(Y_t - \mu) = E(Y_t - \mu)(Y_{t+j} - \mu).$$

But referring again to the definition [3.1.12], this last expression is just the definition of γ_{-j}. Thus, for any covariance-stationary process,

$$\gamma_j = \gamma_{-j} \qquad \text{for all integers } j. \qquad [3.1.13]$$

A different concept is that of *strict stationarity*. A process is said to be strictly stationary if, for any values of $j_1, j_2, \ldots, j_n$, the joint distribution of $(Y_t, Y_{t+j_1}, Y_{t+j_2}, \ldots, Y_{t+j_n})$ depends only on the intervals separating the dates $(j_1, j_2, \ldots, j_n)$ and not on the date itself (t). Notice that if a process is strictly stationary with finite second moments, then it must be covariance-stationary—if the densities over which we are integrating in [3.1.3] and [3.1.10] do not depend on time, then the moments μ_t and γ_{jt} will not depend on time. However, it is possible to imagine a process that is covariance-stationary but not strictly stationary; the mean and autocovariances could not be functions of time, but perhaps higher moments such as $E(Y_t^3)$ are.

In this text the term "stationary" by itself is taken to mean "covariance-stationary."

A process $\{Y_t\}$ is said to be *Gaussian* if the joint density

$$f_{Y_t, Y_{t+j_1}, \ldots, Y_{t+j_n}}(y_t, y_{t+j_1}, \ldots, y_{t+j_n})$$

is Gaussian for any $j_1, j_2, \ldots, j_n$. Since the mean and variance are all that are needed to parameterize a multivariate Gaussian distribution completely, a covariance-stationary Gaussian process is strictly stationary.

Ergodicity

We have viewed expectations of a time series in terms of ensemble averages such as [3.1.4] and [3.1.11]. These definitions may seem a bit contrived, since usually all one has available is a single realization of size T from the process, which we earlier denoted $\{y_1^{(1)}, y_2^{(1)}, \ldots, y_T^{(1)}\}$. From these observations we would calculate the sample mean $\bar{y}$. This, of course, is not an ensemble average but rather a time average:

$$\bar{y} \equiv (1/T) \sum_{t=1}^{T} y_t^{(1)}. \qquad [3.1.14]$$

Whether time averages such as [3.1.14] eventually converge to the ensemble concept $E(Y_t)$ for a stationary process has to do with *ergodicity*. A covariance-stationary process is said to be *ergodic for the mean* if [3.1.14] converges in probability to $E(Y_t)$ as $T \to \infty$.[1] A process will be ergodic for the mean provided that the autocovariance γ_j goes to zero sufficiently quickly as j becomes large. In Chapter 7 we will see that if the autocovariances for a covariance-stationary process satisfy

$$\sum_{j=0}^{\infty} |\gamma_j| < \infty, \qquad [3.1.15]$$

then $\{Y_t\}$ is ergodic for the mean.

Similarly, a covariance-stationary process is said to be ergodic for second moments if

$$[1/(T-j)] \sum_{t=j+1}^{T} (Y_t - \mu)(Y_{t-j} - \mu) \xrightarrow{p} \gamma_j$$

for all j. Sufficient conditions for second-moment ergodicity will be presented in Chapter 7. In the special case where $\{Y_t\}$ is a stationary Gaussian process, condition [3.1.15] is sufficient to ensure ergodicity for all moments.

For many applications, stationarity and ergodicity turn out to amount to the same requirements. For purposes of clarifying the concepts of stationarity and ergodicity, however, it may be helpful to consider an example of a process that is stationary but not ergodic. Suppose the mean $\mu^{(i)}$ for the ith realization $\{y_t^{(i)}\}_{t=-\infty}^{\infty}$ is generated from a $N(0, \lambda^2)$ distribution, say

$$Y_t^{(i)} = \mu^{(i)} + \varepsilon_t. \qquad [3.1.16]$$

Here $\{\varepsilon_t\}$ is a Gaussian white noise process with mean zero and variance σ^2 that is independent of $\mu^{(i)}$. Notice that

$$\mu_t = E(\mu^{(i)}) + E(\varepsilon_t) = 0.$$

Also,

$$\gamma_{0t} = E(\mu^{(i)} + \varepsilon_t)^2 = \lambda^2 + \sigma^2$$

and

$$\gamma_{jt} = E(\mu^{(i)} + \varepsilon_t)(\mu^{(i)} + \varepsilon_{t-j}) = \lambda^2 \qquad \text{for } j \neq 0.$$

Thus the process of [3.1.16] is covariance-stationary. It does not satisfy the sufficient condition [3.1.15] for ergodicity for the mean, however, and indeed, the time average

$$(1/T) \sum_{t=1}^{T} Y_t^{(i)} = (1/T) \sum_{t=1}^{T} (\mu^{(i)} + \varepsilon_t) = \mu^{(i)} + (1/T) \sum_{t=1}^{T} \varepsilon_t$$

converges to $\mu^{(i)}$ rather than to zero, the mean of Y_t.

3.2. *White Noise*

The basic building block for all the processes considered in this chapter is a sequence $\{\varepsilon_t\}_{t=-\infty}^{\infty}$ whose elements have mean zero and variance σ^2,

$$E(\varepsilon_t) = 0 \qquad [3.2.1]$$

$$E(\varepsilon_t^2) = \sigma^2, \qquad [3.2.2]$$

and for which the ε's are uncorrelated across time:

[1]Often "ergodicity" is used in a more general sense; see Anderson and Moore (1979, p. 319) or Hannan (1970, pp. 201–20).

$$E(\varepsilon_t \varepsilon_\tau) = 0 \qquad \text{for } t \neq \tau. \qquad [3.2.3]$$

A process satisfying [3.2.1] through [3.2.3] is described as a *white noise process*.

We shall on occasion wish to replace [3.2.3] with the slightly stronger condition that the ε's are independent across time:

$$\varepsilon_t,\ \varepsilon_\tau \text{ independent for } t \neq \tau. \qquad [3.2.4]$$

Notice that [3.2.4] implies [3.2.3] but [3.2.3] does not imply [3.2.4]. A process satisfying [3.2.1] through [3.2.4] is called an *independent white noise process*.

Finally, if [3.2.1] through [3.2.4] hold along with

$$\varepsilon_t \sim N(0, \sigma^2), \qquad [3.2.5]$$

then we have the *Gaussian white noise process*.

3.3. *Moving Average Processes*

The First-Order Moving Average Process

Let $\{\varepsilon_t\}$ be white noise as in [3.2.1] through [3.2.3], and consider the process

$$Y_t = \mu + \varepsilon_t + \theta\varepsilon_{t-1}, \qquad [3.3.1]$$

where μ and θ could be any constants. This time series is called a *first-order moving average process*, denoted $MA(1)$. The term "moving average" comes from the fact that Y_t is constructed from a weighted sum, akin to an average, of the two most recent values of ε.

The expectation of Y_t is given by

$$E(Y_t) = E(\mu + \varepsilon_t + \theta\varepsilon_{t-1}) = \mu + E(\varepsilon_t) + \theta \cdot E(\varepsilon_{t-1}) = \mu. \qquad [3.3.2]$$

We used the symbol μ for the constant term in [3.3.1] in anticipation of the result that this constant term turns out to be the mean of the process.

The variance of Y_t is

$$\begin{aligned} E(Y_t - \mu)^2 &= E(\varepsilon_t + \theta\varepsilon_{t-1})^2 \\ &= E(\varepsilon_t^2 + 2\theta\varepsilon_t\varepsilon_{t-1} + \theta^2\varepsilon_{t-1}^2) \\ &= \sigma^2 + 0 + \theta^2\sigma^2 \\ &= (1 + \theta^2)\sigma^2. \end{aligned} \qquad [3.3.3]$$

The first autocovariance is

$$\begin{aligned} E(Y_t - \mu)(Y_{t-1} - \mu) &= E(\varepsilon_t + \theta\varepsilon_{t-1})(\varepsilon_{t-1} + \theta\varepsilon_{t-2}) \\ &= E(\varepsilon_t\varepsilon_{t-1} + \theta\varepsilon_{t-1}^2 + \theta\varepsilon_t\varepsilon_{t-2} + \theta^2\varepsilon_{t-1}\varepsilon_{t-2}) \\ &= 0 + \theta\sigma^2 + 0 + 0. \end{aligned} \qquad [3.3.4]$$

Higher autocovariances are all zero:

$$E(Y_t - \mu)(Y_{t-j} - \mu) = E(\varepsilon_t + \theta\varepsilon_{t-1})(\varepsilon_{t-j} + \theta\varepsilon_{t-j-1}) = 0 \qquad \text{for } j > 1. \qquad [3.3.5]$$

Since the mean and autocovariances are not functions of time, an $MA(1)$ process is covariance-stationary regardless of the value of θ. Furthermore, [3.1.15] is clearly satisfied:

$$\sum_{j=0}^{\infty} |\gamma_j| = (1 + \theta^2)\sigma^2 + |\theta\sigma^2|.$$

Thus, if $\{\varepsilon_t\}$ is Gaussian white noise, then the $MA(1)$ process [3.3.1] is ergodic for all moments.

The jth *autocorrelation* of a covariance-stationary process (denoted ρ_j) is defined as its jth autocovariance divided by the variance:

$$\rho_j \equiv \gamma_j/\gamma_0. \qquad [3.3.6]$$

Again the terminology arises from the fact that ρ_j is the correlation between Y_t and Y_{t-j}:

$$\text{Corr}(Y_t, Y_{t-j}) = \frac{\text{Cov}(Y_t, Y_{t-j})}{\sqrt{\text{Var}(Y_t)}\sqrt{\text{Var}(Y_{t-j})}} = \frac{\gamma_j}{\sqrt{\gamma_0}\sqrt{\gamma_0}} = \rho_j.$$

Since ρ_j is a correlation, $|\rho_j| \leq 1$ for all j, by the Cauchy-Schwarz inequality. Notice also that the 0th autocorrelation ρ_0 is equal to unity for any covariance-stationary process by definition.

From [3.3.3] and [3.3.4], the first autocorrelation for an $MA(1)$ process is given by

$$\rho_1 = \frac{\theta\sigma^2}{(1+\theta^2)\sigma^2} = \frac{\theta}{(1+\theta^2)}. \qquad [3.3.7]$$

Higher autocorrelations are all zero.

The autocorrelation ρ_j can be plotted as a function of j as in Figure 3.1. Panel (a) shows the autocorrelation function for white noise, while panel (b) gives the autocorrelation function for the $MA(1)$ process:

$$Y_t = \varepsilon_t + 0.8\varepsilon_{t-1}.$$

For different specifications of θ we would obtain different values for the first autocorrelation ρ_1 in [3.3.7]. Positive values of θ induce positive autocorrelation in the series. In this case, an unusually large value of Y_t is likely to be followed by a larger-than-average value for Y_{t+1}, just as a smaller-than-average Y_t may well be followed by a smaller-than-average Y_{t+1}. By contrast, negative values of θ imply negative autocorrelation—a large Y_t might be expected to be followed by a small value for Y_{t+1}.

The values for ρ_1 implied by different specifications of θ are plotted in Figure 3.2. Notice that the largest possible value for ρ_1 is 0.5; this occurs if $\theta = 1$. The smallest value for ρ_1 is -0.5, which occurs if $\theta = -1$. For any value of ρ_1 between -0.5 and 0.5, there are two different values of θ that could produce that autocorrelation. This is because the value of $\theta/(1+\theta^2)$ is unchanged if θ is replaced by $1/\theta$:

$$\rho_1 = \frac{(1/\theta)}{1+(1/\theta)^2} = \frac{\theta^2\cdot(1/\theta)}{\theta^2[1+(1/\theta)^2]} = \frac{\theta}{\theta^2+1}.$$

For example, the processes

$$Y_t = \varepsilon_t + 0.5\varepsilon_{t-1}$$

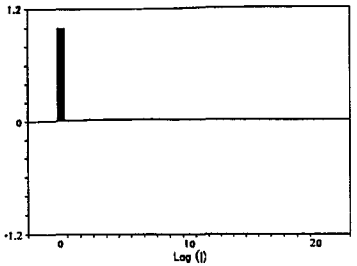
and

$$Y_t = \varepsilon_t + 2\varepsilon_{t-1}$$

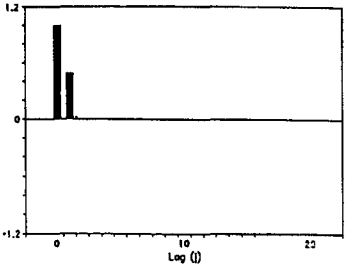
would have the same autocorrelation function:

$$\rho_1 = \frac{2}{(1+2^2)} = \frac{0.5}{(1+0.5^2)} = 0.4.$$

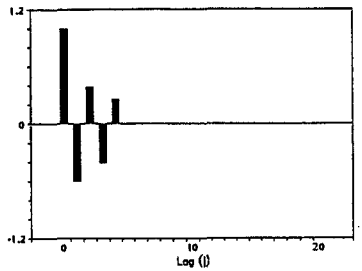
We will have more to say about the relation between two $MA(1)$ processes that share the same autocorrelation function in Section 3.7.



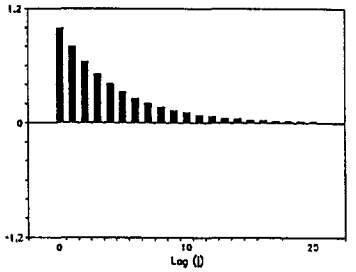
(a) White noise: $Y_t = \varepsilon_t$



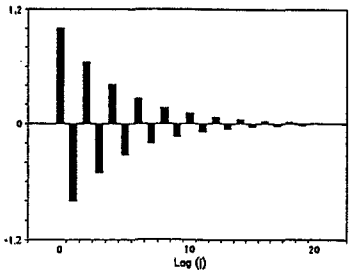
(b) $MA(1)$: $Y_t = \varepsilon_t + 0.8\varepsilon_{t-1}$



(c) $MA(4)$: $Y_t = \varepsilon_t - 0.6\varepsilon_{t-1} + 0.3\varepsilon_{t-2} - 0.5\varepsilon_{t-3} + 0.5\varepsilon_{t-4}$



(d) $AR(1)$: $Y_t = 0.8Y_{t-1} + \varepsilon_t$



(e) $AR(1)$: $Y_t = -0.8Y_{t-1} + \varepsilon_t$

FIGURE 3.1 Autocorrelation functions for assorted *ARMA* processes.

The qth-Order Moving Average Process

A *qth-order moving average process*, denoted $MA(q)$, is characterized by

$$Y_t = \mu + \varepsilon_t + \theta_1\varepsilon_{t-1} + \theta_2\varepsilon_{t-2} + \cdots + \theta_q\varepsilon_{t-q}, \qquad [3.3.8]$$

where $\{\varepsilon_t\}$ satisfies [3.2.1] through [3.2.3] and $(\theta_1, \theta_2, \ldots, \theta_q)$ could be any real numbers. The mean of [3.3.8] is again given by μ:

$$E(Y_t) = \mu + E(\varepsilon_t) + \theta_1 \cdot E(\varepsilon_{t-1}) + \theta_2 \cdot E(\varepsilon_{t-2}) + \cdots + \theta_q \cdot E(\varepsilon_{t-q}) = \mu.$$

The variance of an $MA(q)$ process is

$$\gamma_0 = E(Y_t - \mu)^2 = E(\varepsilon_t + \theta_1\varepsilon_{t-1} + \theta_2\varepsilon_{t-2} + \cdots + \theta_q\varepsilon_{t-q})^2. \qquad [3.3.9]$$

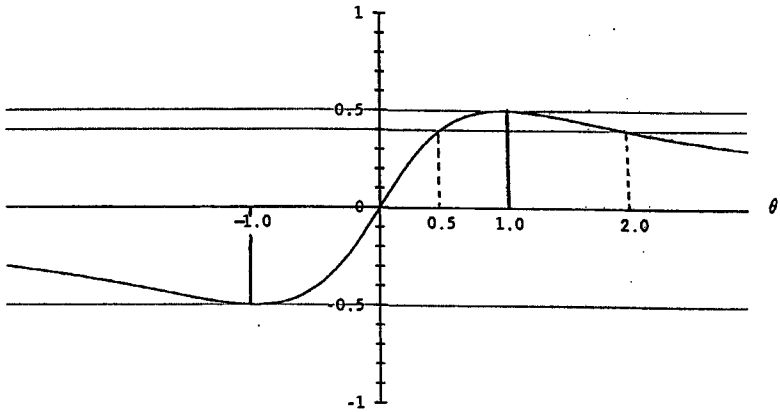


FIGURE 3.2 The first autocorrelation (ρ_1) for an $MA(1)$ process possible for different values of θ.

Since the ε's are uncorrelated, the variance [3.3.9] is[2]

$$\gamma_0 = \sigma^2 + \theta_1^2\sigma^2 + \theta_2^2\sigma^2 + \cdots + \theta_q^2\sigma^2 = (1 + \theta_1^2 + \theta_2^2 + \cdots + \theta_q^2)\sigma^2. \quad [3.3.10]$$

For $j = 1, 2, \ldots, q$,

$$\begin{aligned}\gamma_j &= E[(\varepsilon_t + \theta_1\varepsilon_{t-1} + \theta_2\varepsilon_{t-2} + \cdots + \theta_q\varepsilon_{t-q}) \\ &\quad \times (\varepsilon_{t-j} + \theta_1\varepsilon_{t-j-1} + \theta_2\varepsilon_{t-j-2} + \cdots + \theta_q\varepsilon_{t-j-q})] \qquad [3.3.11] \\ &= E[\theta_j\varepsilon_{t-j}^2 + \theta_{j+1}\theta_1\varepsilon_{t-j-1}^2 + \theta_{j+2}\theta_2\varepsilon_{t-j-2}^2 + \cdots + \theta_q\theta_{q-j}\varepsilon_{t-q}^2].\end{aligned}$$

Terms involving ε's at different dates have been dropped because their product has expectation zero, and θ_0 is defined to be unity. For $j > q$, there are no ε's with common dates in the definition of γ_j, and so the expectation is zero. Thus,

$$\gamma_j = \begin{cases}[\theta_j + \theta_{j+1}\theta_1 + \theta_{j+2}\theta_2 + \cdots + \theta_q\theta_{q-j}]\cdot\sigma^2 & \text{for } j = 1, 2, \ldots, q \\ 0 & \text{for } j > q.\end{cases} \quad [3.3.12]$$

For example, for an $MA(2)$ process,

$$\begin{aligned}\gamma_0 &= [1 + \theta_1^2 + \theta_2^2]\cdot\sigma^2 \\ \gamma_1 &= [\theta_1 + \theta_2\theta_1]\cdot\sigma^2 \\ \gamma_2 &= [\theta_2]\cdot\sigma^2 \\ \gamma_3 &= \gamma_4 = \cdots = 0.\end{aligned}$$

For any values of $(\theta_1, \theta_2, \ldots, \theta_q)$, the $MA(q)$ process is thus covariance-stationary. Condition [3.1.15] is satisfied, so for Gaussian ε_t the $MA(q)$ process is also ergodic for all moments. The autocorrelation function is zero after q lags, as in panel (c) of Figure 3.1.

The Infinite-Order Moving Average Process

The $MA(q)$ process can be written

$$Y_t = \mu + \sum_{j=0}^{q} \theta_j\varepsilon_{t-j}$$

[2]See equation [A.5.18] in Appendix A at the end of the book.

with $\theta_0 \equiv 1$. Consider the process that results as $q \to \infty$:

$$Y_t = \mu + \sum_{j=0}^{\infty} \psi_j \varepsilon_{t-j} = \mu + \psi_0 \varepsilon_t + \psi_1 \varepsilon_{t-1} + \psi_2 \varepsilon_{t-2} + \cdots . \qquad [3.3.13]$$

This could be described as an $MA(\infty)$ process. To preserve notational flexibility later, we will use ψ's for the coefficients of an infinite-order moving average process and θ's for the coefficients of a finite-order moving average process.

Appendix 3.A to this chapter shows that the infinite sequence in [3.3.13] generates a well defined covariance-stationary process provided that

$$\sum_{j=0}^{\infty} \psi_j^2 < \infty. \qquad [3.3.14]$$

It is often convenient to work with a slightly stronger condition than [3.3.14]:

$$\sum_{j=0}^{\infty} |\psi_j| < \infty. \qquad [3.3.15]$$

A sequence of numbers $\{\psi_j\}_{j=0}^{\infty}$ satisfying [3.3.14] is said to be *square summable*, whereas a sequence satisfying [3.3.15] is said to be *absolutely summable*. Absolute summability implies square-summability, but the converse does not hold—there are examples of square-summable sequences that are not absolutely summable (again, see Appendix 3.A).

The mean and autocovariances of an $MA(\infty)$ process with absolutely summable coefficients can be calculated from a simple extrapolation of the results for an $MA(q)$ process:[3]

$$\begin{aligned} E(Y_t) &= \lim_{T\to\infty} E(\mu + \psi_0\varepsilon_t + \psi_1\varepsilon_{t-1} + \psi_2\varepsilon_{t-2} + \cdots + \psi_T\varepsilon_{t-T}) \qquad [3.3.16] \\ &= \mu \\ \gamma_0 &= E(Y_t - \mu)^2 \\ &= \lim_{T\to\infty} E(\psi_0\varepsilon_t + \psi_1\varepsilon_{t-1} + \psi_2\varepsilon_{t-2} + \cdots + \psi_T\varepsilon_{t-T})^2 \qquad [3.3.17] \\ &= \lim_{T\to\infty} (\psi_0^2 + \psi_1^2 + \psi_2^2 + \cdots + \psi_T^2)\cdot\sigma^2 \\ \gamma_j &= E(Y_t - \mu)(Y_{t-j} - \mu) \qquad [3.3.18] \\ &= \sigma^2(\psi_j\psi_0 + \psi_{j+1}\psi_1 + \psi_{j+2}\psi_2 + \psi_{j+3}\psi_3 + \cdots). \end{aligned}$$

Moreover, an $MA(\infty)$ process with absolutely summable coefficients has absolutely summable autocovariances:

$$\sum_{j=0}^{\infty} |\gamma_j| < \infty. \qquad [3.3.19]$$

Hence, an $MA(\infty)$ process satisfying [3.3.15] is ergodic for the mean (see Appendix 3.A). If the ε's are Gaussian, then the process is ergodic for all moments.

[3]Absolute summability of $\{\psi_j\}_{j=0}^{\infty}$ and existence of the second moment $E(\varepsilon_t^2)$ are sufficient conditions to permit interchanging the order of integration and summation. Specifically, if $\{X_T\}_{T=1}^{\infty}$ is a sequence of random variables such that

$$\sum_{T=1}^{\infty} E|X_T| < \infty,$$

then

$$E\left\{\sum_{T=1}^{\infty} X_T\right\} = \sum_{T=1}^{\infty} E(X_T).$$

See Rao (1973, p. 111).

3.4. *Autoregressive Processes*

The First-Order Autoregressive Process

A *first-order autoregression*, denoted $AR(1)$, satisfies the following difference equation:

$$Y_t = c + \phi Y_{t-1} + \varepsilon_t. \qquad [3.4.1]$$

Again, $\{\varepsilon_t\}$ is a white noise sequence satisfying [3.2.1] through [3.2.3]. Notice that [3.4.1] takes the form of the first-order difference equation [1.1.1] or [2.2.1] in which the input variable w_t is given by $w_t = c + \varepsilon_t$. We know from the analysis of first-order difference equations that if $|\phi| \geq 1$, the consequences of the ε's for Y accumulate rather than die out over time. It is thus perhaps not surprising that when $|\phi| \geq 1$, there does not exist a covariance-stationary process for Y_t with finite variance that satisfies [3.4.1]. In the case when $|\phi| < 1$, there is a covariance-stationary process for Y_t satisfying [3.4.1]. It is given by the stable solution to [3.4.1] characterized in [2.2.9]:

$$\begin{aligned} Y_t &= (c + \varepsilon_t) + \phi\cdot(c + \varepsilon_{t-1}) + \phi^2\cdot(c + \varepsilon_{t-2}) + \phi^3\cdot(c + \varepsilon_{t-3}) + \cdots \\ &= [c/(1 - \phi)] + \varepsilon_t + \phi\varepsilon_{t-1} + \phi^2\varepsilon_{t-2} + \phi^3\varepsilon_{t-3} + \cdots. \end{aligned} \qquad [3.4.2]$$

This can be viewed as an $MA(\infty)$ process as in [3.3.13] with ψ_j given by ϕ^j. When $|\phi| < 1$, condition [3.3.15] is satisfied:

$$\sum_{j=0}^{\infty} |\psi_j| = \sum_{j=0}^{\infty} |\phi|^j,$$

which equals $1/(1 - |\phi|)$ provided that $|\phi| < 1$. The remainder of this discussion of first-order autoregressive processes assumes that $|\phi| < 1$. This ensures that the $MA(\infty)$ representation exists and can be manipulated in the obvious way, and that the $AR(1)$ process is ergodic for the mean.

Taking expectations of [3.4.2], we see that

$$E(Y_t) = [c/(1 - \phi)] + 0 + 0 + \cdots,$$

so that the mean of a stationary $AR(1)$ process is

$$\mu = c/(1 - \phi). \qquad [3.4.3]$$

The variance is

$$\begin{aligned} \gamma_0 &= E(Y_t - \mu)^2 \\ &= E(\varepsilon_t + \phi\varepsilon_{t-1} + \phi^2\varepsilon_{t-2} + \phi^3\varepsilon_{t-3} + \cdots)^2 \\ &= (1 + \phi^2 + \phi^4 + \phi^6 + \cdots)\cdot\sigma^2 \\ &= \sigma^2/(1 - \phi^2), \end{aligned} \qquad [3.4.4]$$

while the jth autocovariance is

$$\begin{aligned} \gamma_j &= E(Y_t - \mu)(Y_{t-j} - \mu) \\ &= E[\varepsilon_t + \phi\varepsilon_{t-1} + \phi^2\varepsilon_{t-2} + \cdots + \phi^j\varepsilon_{t-j} + \phi^{j+1}\varepsilon_{t-j-1} \\ &\quad + \phi^{j+2}\varepsilon_{t-j-2} + \cdots] \times [\varepsilon_{t-j} + \phi\varepsilon_{t-j-1} + \phi^2\varepsilon_{t-j-2} + \cdots] \\ &= [\phi^j + \phi^{j+2} + \phi^{j+4} + \cdots]\cdot\sigma^2 \\ &= \phi^j[1 + \phi^2 + \phi^4 + \cdots]\cdot\sigma^2 \\ &= [\phi^j/(1 - \phi^2)]\cdot\sigma^2. \end{aligned} \qquad [3.4.5]$$

It follows from [3.4.4] and [3.4.5] that the autocorrelation function,

$$\rho_j = \gamma_j/\gamma_0 = \phi^j, \qquad [3.4.6]$$

follows a pattern of geometric decay as in panel (d) of Figure 3.1. Indeed, the autocorrelation function [3.4.6] for a stationary $AR(1)$ process is identical to the dynamic multiplier or impulse-response function [1.1.10]; the effect of a one-unit increase in ε_t on Y_{t+j} is equal to the correlation between Y_t and Y_{t+j}. A positive value of ϕ, like a positive value of θ for an $MA(1)$ process, implies positive correlation between Y_t and Y_{t+1}. A negative value of ϕ implies negative first-order but positive second-order autocorrelation, as in panel (e) of Figure 3.1.

Figure 3.3 shows the effect on the appearance of the time series $\{y_t\}$ of varying the parameter ϕ. The panels show realizations of the process in [3.4.1] with $c = 0$ and $\varepsilon_t \sim N(0, 1)$ for different values of the autoregressive parameter ϕ. Panel (a) displays white noise ($\phi = 0$). A series with no autocorrelation looks choppy and patternless to the eye; the value of one observation gives no information about the value of the next observation. For $\phi = 0.5$ (panel (b)), the series seems smoother, with observations above or below the mean often appearing in clusters of modest duration. For $\phi = 0.9$ (panel (c)), departures from the mean can be quite prolonged; strong shocks take considerable time to die out.

The moments for a stationary $AR(1)$ were derived above by viewing it as an $MA(\infty)$ process. A second way to arrive at the same results is to assume that the process is covariance-stationary and calculate the moments directly from the difference equation [3.4.1]. Taking expectations of both sides of [3.4.1],

$$E(Y_t) = c + \phi \cdot E(Y_{t-1}) + E(\varepsilon_t). \qquad [3.4.7]$$

Assuming that the process is covariance-stationary,

$$E(Y_t) = E(Y_{t-1}) = \mu. \qquad [3.4.8]$$

Substituting [3.4.8] into [3.4.7],

$$\mu = c + \phi\mu + 0$$

or

$$\mu = c/(1 - \phi), \qquad [3.4.9]$$

reproducing the earlier result [3.4.3].

Notice that formula [3.4.9] is clearly not generating a sensible statement if $|\phi| \geq 1$. For example, if $c > 0$ and $\phi > 1$, then Y_t in [3.4.1] is equal to a positive constant plus a positive number times its lagged value plus a mean-zero random variable. Yet [3.4.9] seems to assert that Y_t would be negative on average for such a process! The reason that formula [3.4.9] is not valid when $|\phi| \geq 1$ is that we assumed in [3.4.8] that Y_t is covariance-stationary, an assumption which is not correct when $|\phi| \geq 1$.

To find the second moments of Y_t in an analogous manner, use [3.4.3] to rewrite [3.4.1] as

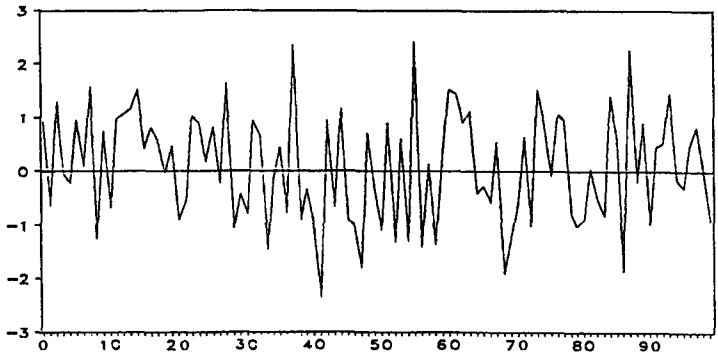
$$Y_t = \mu(1 - \phi) + \phi Y_{t-1} + \varepsilon_t$$

or

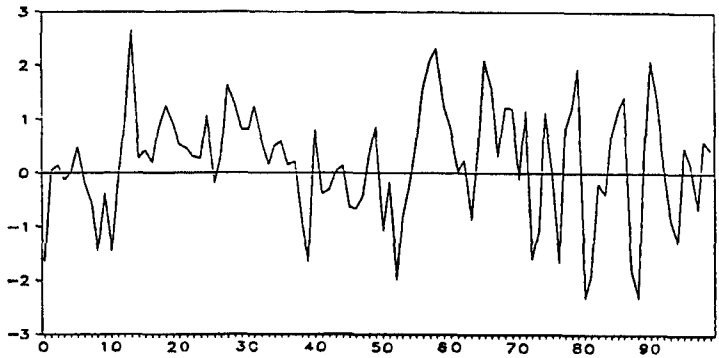
$$(Y_t - \mu) = \phi(Y_{t-1} - \mu) + \varepsilon_t. \qquad [3.4.10]$$

Now square both sides of [3.4.10] and take expectations:

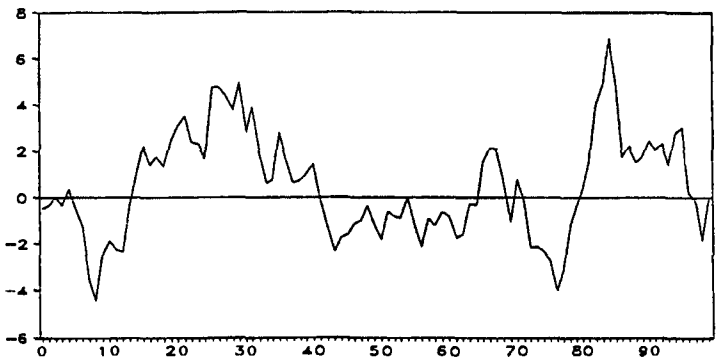
$$E(Y_t - \mu)^2 = \phi^2 E(Y_{t-1} - \mu)^2 + 2\phi E[(Y_{t-1} - \mu)\varepsilon_t] + E(\varepsilon_t^2). \qquad [3.4.11]$$



(a) $\phi = 0$ (white noise)



(b) $\phi = 0.5$



(c) $\phi = 0.9$

FIGURE 3.3 Realizations of an $AR(1)$ process, $Y_t = \phi Y_{t-1} + \varepsilon_t$, for alternative values of ϕ.

Recall from [3.4.2] that $(Y_{t-1} - \mu)$ is a linear function of $\varepsilon_{t-1}, \varepsilon_{t-2}, \ldots$:

$$(Y_{t-1} - \mu) = \varepsilon_{t-1} + \phi\varepsilon_{t-2} + \phi^2\varepsilon_{t-3} + \cdots.$$

But ε_t is uncorrelated with $\varepsilon_{t-1}, \varepsilon_{t-2}, \ldots$, so ε_t must be uncorrelated with $(Y_{t-1} - \mu)$. Thus the middle term on the right side of [3.4.11] is zero:

$$E[(Y_{t-1} - \mu)\varepsilon_t] = 0. \qquad [3.4.12]$$

Again, assuming covariance-stationarity, we have

$$E(Y_t - \mu)^2 = E(Y_{t-1} - \mu)^2 = \gamma_0. \qquad [3.4.13]$$

Substituting [3.4.13] and [3.4.12] into [3.4.11],

$$\gamma_0 = \phi^2\gamma_0 + 0 + \sigma^2$$

or

$$\gamma_0 = \sigma^2/(1 - \phi^2),$$

reproducing [3.4.4].

Similarly, we could multiply [3.4.10] by $(Y_{t-j} - \mu)$ and take expectations:

$$\begin{aligned} &E[(Y_t - \mu)(Y_{t-j} - \mu)] \\ &\quad = \phi \cdot E[(Y_{t-1} - \mu)(Y_{t-j} - \mu)] + E[\varepsilon_t(Y_{t-j} - \mu)]. \end{aligned} \qquad [3.4.14]$$

But the term $(Y_{t-j} - \mu)$ will be a linear function of $\varepsilon_{t-j}, \varepsilon_{t-j-1}, \varepsilon_{t-j-2}, \ldots$, which, for $j > 0$, will be uncorrelated with ε_t. Thus, for $j > 0$, the last term on the right side in [3.4.14] is zero. Notice, moreover, that the expression appearing in the first term on the right side of [3.4.14],

$$E[(Y_{t-1} - \mu)(Y_{t-j} - \mu)],$$

is the autocovariance of observations on Y separated by $j - 1$ periods:

$$E[(Y_{t-1} - \mu)(Y_{|t-1|-|j-1|} - \mu)] = \gamma_{j-1}.$$

Thus, for $j > 0$, [3.4.14] becomes

$$\gamma_j = \phi\gamma_{j-1}. \qquad [3.4.15]$$

Equation [3.4.15] takes the form of a first-order difference equation,

$$y_t = \phi y_{t-1} + w_t,$$

in which the autocovariance γ takes the place of the variable y and in which the subscript j (which indexes the order of the autocovariance) replaces t (which indexes time). The input w_t in [3.4.15] is identically equal to zero. It is easy to see that the difference equation [3.4.15] has the solution

$$\gamma_j = \phi^j\gamma_0,$$

which reproduces [3.4.6]. We now see why the impulse-response function and autocorrelation function for an $AR(1)$ process coincide—they both represent the solution to a first-order difference equation with autoregressive parameter ϕ, an initial value of unity, and no subsequent shocks.

The Second-Order Autoregressive Process

A *second-order autoregression*, denoted $AR(2)$, satisfies

$$Y_t = c + \phi_1 Y_{t-1} + \phi_2 Y_{t-2} + \varepsilon_t, \qquad [3.4.16]$$

or, in lag operator notation,

$$(1 - \phi_1 L - \phi_2 L^2)Y_t = c + \varepsilon_t. \qquad [3.4.17]$$

The difference equation [3.4.16] is stable provided that the roots of

$$(1 - \phi_1 z - \phi_2 z^2) = 0 \qquad [3.4.18]$$

lie outside the unit circle. When this condition is satisfied, the $AR(2)$ process turns out to be covariance-stationary, and the inverse of the autoregressive operator in [3.4.17] is given by

$$\psi(L) = (1 - \phi_1 L - \phi_2 L^2)^{-1} = \psi_0 + \psi_1 L + \psi_2 L^2 + \psi_3 L^3 + \cdots. \qquad [3.4.19]$$

Recalling [1.2.44], the value of ψ_j can be found from the (1, 1) element of the matrix $\mathbf{F}$ raised to the jth power, as in expression [1.2.28]. Where the roots of [3.4.18] are distinct, a closed-form expression for ψ_j is given by [1.2.29] and [1.2.25]. Exercise 3.3 at the end of this chapter discusses alternative algorithms for calculating ψ_j.

Multiplying both sides of [3.4.17] by $\psi(L)$ gives

$$Y_t = \psi(L)c + \psi(L)\varepsilon_t. \qquad [3.4.20]$$

It is straightforward to show that

$$\psi(L)c = c/(1 - \phi_1 - \phi_2) \qquad [3.4.21]$$

and

$$\sum_{j=0}^{\infty} |\psi_j| < \infty; \qquad [3.4.22]$$

the reader is invited to prove these claims in Exercises 3.4 and 3.5. Since [3.4.20] is an absolutely summable $MA(\infty)$ process, its mean is given by the constant term:

$$\mu = c/(1 - \phi_1 - \phi_2). \qquad [3.4.23]$$

An alternative method for calculating the mean is to assume that the process is covariance-stationary and take expectations of [3.4.16] directly:

$$E(Y_t) = c + \phi_1 E(Y_{t-1}) + \phi_2 E(Y_{t-2}) + E(\varepsilon_t),$$

implying

$$\mu = c + \phi_1\mu + \phi_2\mu + 0,$$

reproducing [3.4.23].

To find second moments, write [3.4.16] as

$$Y_t = \mu\cdot(1 - \phi_1 - \phi_2) + \phi_1 Y_{t-1} + \phi_2 Y_{t-2} + \varepsilon_t$$

or

$$(Y_t - \mu) = \phi_1(Y_{t-1} - \mu) + \phi_2(Y_{t-2} - \mu) + \varepsilon_t. \qquad [3.4.24]$$

Multiplying both sides of [3.4.24] by $(Y_{t-j} - \mu)$ and taking expectations produces

$$\gamma_j = \phi_1\gamma_{j-1} + \phi_2\gamma_{j-2} \qquad \text{for } j = 1, 2, \ldots. \qquad [3.4.25]$$

Thus, the autocovariances follow the same second-order difference equation as does the process for Y_t, with the difference equation for γ_j indexed by the lag j. The autocovariances therefore behave just as the solutions to the second-order difference equation analyzed in Section 1.2. An $AR(2)$ process is covariance-stationary provided that ϕ_1 and ϕ_2 lie within the triangular region of Figure 1.5.

When ϕ_1 and ϕ_2 lie within the triangular region but above the parabola in that figure, the autocovariance function γ_j is the sum of two decaying exponential functions of j. When ϕ_1 and ϕ_2 fall within the triangular region but below the parabola, γ_j is a damped sinusoidal function.

The autocorrelations are found by dividing both sides of [3.4.25] by γ_0:

$$\rho_j = \phi_1\rho_{j-1} + \phi_2\rho_{j-2} \qquad \text{for } j = 1, 2, \ldots. \qquad [3.4.26]$$

In particular, setting $j = 1$ produces

$$\rho_1 = \phi_1 + \phi_2\rho_1$$

or

$$\rho_1 = \phi_1/(1 - \phi_2). \qquad [3.4.27]$$

For $j = 2$,

$$\rho_2 = \phi_1\rho_1 + \phi_2. \qquad [3.4.28]$$

The variance of a covariance-stationary second-order autoregression can be found by multiplying both sides of [3.4.24] by $(Y_t - \mu)$ and taking expectations:

$$E(Y_t - \mu)^2 = \phi_1 \cdot E(Y_{t-1} - \mu)(Y_t - \mu) + \phi_2 \cdot E(Y_{t-2} - \mu)(Y_t - \mu) + E(\varepsilon_t)(Y_t - \mu),$$

or

$$\gamma_0 = \phi_1\gamma_1 + \phi_2\gamma_2 + \sigma^2. \qquad [3.4.29]$$

The last term (σ^2) in [3.4.29] comes from noticing that

$$\begin{aligned} E(\varepsilon_t)(Y_t - \mu) &= E(\varepsilon_t)[\phi_1(Y_{t-1} - \mu) + \phi_2(Y_{t-2} - \mu) + \varepsilon_t] \\ &= \phi_1 \cdot 0 + \phi_2 \cdot 0 + \sigma^2. \end{aligned}$$

Equation [3.4.29] can be written

$$\gamma_0 = \phi_1\rho_1\gamma_0 + \phi_2\rho_2\gamma_0 + \sigma^2. \qquad [3.4.30]$$

Substituting [3.4.27] and [3.4.28] into [3.4.30] gives

$$\gamma_0 = \left[\frac{\phi_1^2}{(1 - \phi_2)} + \frac{\phi_2\phi_1^2}{(1 - \phi_2)} + \phi_2^2\right]\gamma_0 + \sigma^2$$

or

$$\gamma_0 = \frac{(1 - \phi_2)\sigma^2}{(1 + \phi_2)[(1 - \phi_2)^2 - \phi_1^2]}.$$

The pth-Order Autoregressive Process

A *pth-order autoregression*, denoted $AR(p)$, satisfies

$$Y_t = c + \phi_1 Y_{t-1} + \phi_2 Y_{t-2} + \cdots + \phi_p Y_{t-p} + \varepsilon_t. \qquad [3.4.31]$$

Provided that the roots of

$$1 - \phi_1 z - \phi_2 z^2 - \cdots - \phi_p z^p = 0 \qquad [3.4.32]$$

all lie outside the unit circle, it is straightforward to verify that a covariance-stationary representation of the form

$$Y_t = \mu + \psi(L)\varepsilon_t \qquad [3.4.33]$$

exists where

$$\psi(L) = (1 - \phi_1 L - \phi_2 L^2 - \cdots - \phi_p L^p)^{-1}$$

and $\Sigma_{j=0}^{\infty} |\psi_j| < \infty$. Assuming that the stationarity condition is satisfied, one way to find the mean is to take expectations of [3.4.31]:

$$\mu = c + \phi_1\mu + \phi_2\mu + \cdots + \phi_p\mu,$$

or

$$\mu = c/(1 - \phi_1 - \phi_2 - \cdots - \phi_p). \qquad [3.4.34]$$

Using [3.4.34], equation [3.4.31] can be written

$$Y_t - \mu = \phi_1(Y_{t-1} - \mu) + \phi_2(Y_{t-2} - \mu) + \cdots + \phi_p(Y_{t-p} - \mu) + \varepsilon_t. \qquad [3.4.35]$$

Autocovariances are found by multiplying both sides of [3.4.35] by $(Y_{t-j} - \mu)$ and taking expectations:

$$\gamma_j = \begin{cases} \phi_1\gamma_{j-1} + \phi_2\gamma_{j-2} + \cdots + \phi_p\gamma_{j-p} & \text{for } j = 1, 2, \ldots \\ \phi_1\gamma_1 + \phi_2\gamma_2 + \cdots + \phi_p\gamma_p + \sigma^2 & \text{for } j = 0. \end{cases} \qquad [3.4.36]$$

Using the fact that $\gamma_{-j} = \gamma_j$, the system of equations in [3.4.36] for $j = 0, 1, \ldots, p$ can be solved for $\gamma_0, \gamma_1, \ldots, \gamma_p$ as functions of $\sigma^2, \phi_1, \phi_2, \ldots, \phi_p$. It can be shown[4] that the $(p \times 1)$ vector $(\gamma_0, \gamma_1, \ldots, \gamma_{p-1})'$ is given by the first p elements of the first column of the $(p^2 \times p^2)$ matrix $\sigma^2[\mathbf{I}_{p^2} - (\mathbf{F} \otimes \mathbf{F})]^{-1}$ where $\mathbf{F}$ is the $(p \times p)$ matrix defined in equation [1.2.3] and $\otimes$ indicates the Kronecker product.

Dividing [3.4.36] by γ_0 produces the *Yule-Walker equations*:

$$\rho_j = \phi_1\rho_{j-1} + \phi_2\rho_{j-2} + \cdots + \phi_p\rho_{j-p} \qquad \text{for } j = 1, 2, \ldots. \qquad [3.4.37]$$

Thus, the autocovariances and autocorrelations follow the same pth-order difference equation as does the process itself [3.4.31]. For distinct roots, their solutions take the form

$$\gamma_j = g_1\lambda_1^j + g_2\lambda_2^j + \cdots + g_p\lambda_p^j, \qquad [3.4.38]$$

where the eigenvalues $(\lambda_1, \ldots, \lambda_p)$ are the solutions to

$$\lambda^p - \phi_1\lambda^{p-1} - \phi_2\lambda^{p-2} - \cdots - \phi_p = 0.$$

3.5. *Mixed Autoregressive Moving Average Processes*

An *ARMA(p, q) process* includes both autoregressive and moving average terms:

$$Y_t = c + \phi_1 Y_{t-1} + \phi_2 Y_{t-2} + \cdots + \phi_p Y_{t-p} + \varepsilon_t + \theta_1\varepsilon_{t-1} + \theta_2\varepsilon_{t-2} + \cdots + \theta_q\varepsilon_{t-q}, \qquad [3.5.1]$$

or, in lag operator form,

$$(1 - \phi_1 L - \phi_2 L^2 - \cdots - \phi_p L^p)Y_t = c + (1 + \theta_1 L + \theta_2 L^2 + \cdots + \theta_q L^q)\varepsilon_t. \qquad [3.5.2]$$

Provided that the roots of

$$1 - \phi_1 z - \phi_2 z^2 - \cdots - \phi_p z^p = 0 \qquad [3.5.3]$$

[4]The reader will be invited to prove this in Exercise 10.1 in Chapter 10.

lie outside the unit circle, both sides of [3.5.2] can be divided by $(1 - \phi_1 L - \phi_2 L^2 - \cdots - \phi_p L^p)$ to obtain

$$Y_t = \mu + \psi(L)\varepsilon_t$$

where

$$\psi(L) = \frac{(1 + \theta_1 L + \theta_2 L^2 + \cdots + \theta_q L^q)}{(1 - \phi_1 L - \phi_2 L^2 - \cdots - \phi_p L^p)}$$

$$\sum_{j=0}^{\infty} |\psi_j| < \infty$$

$$\mu = c/(1 - \phi_1 - \phi_2 - \cdots - \phi_p).$$

Thus, stationarity of an *ARMA* process depends entirely on the autoregressive parameters $(\phi_1, \phi_2, \ldots, \phi_p)$ and not on the moving average parameters $(\theta_1, \theta_2, \ldots, \theta_q)$.

It is often convenient to write the *ARMA* process [3.5.1] in terms of deviations from the mean:

$$\begin{aligned} Y_t - \mu = {} & \phi_1(Y_{t-1} - \mu) + \phi_2(Y_{t-2} - \mu) + \cdots \\ & + \phi_p(Y_{t-p} - \mu) + \varepsilon_t + \theta_1\varepsilon_{t-1} + \theta_2\varepsilon_{t-2} + \cdots + \theta_q\varepsilon_{t-q}. \end{aligned} \qquad [3.5.4]$$

Autocovariances are found by multiplying both sides of [3.5.4] by $(Y_{t-j} - \mu)$ and taking expectations. For $j > q$, the resulting equations take the form

$$\gamma_j = \phi_1\gamma_{j-1} + \phi_2\gamma_{j-2} + \cdots + \phi_p\gamma_{j-p} \qquad \text{for } j = q+1, q+2, \ldots. \qquad [3.5.5]$$

Thus, after q lags the autocovariance function γ_j (and the autocorrelation function ρ_j) follow the pth-order difference equation governed by the autoregressive parameters.

Note that [3.5.5] does not hold for $j \leq q$, owing to correlation between $\theta_j\varepsilon_{t-j}$ and Y_{t-j}. Hence, an *ARMA*(p, q) process will have more complicated autocovariances for lags 1 through q than would the corresponding *AR*(p) process. For $j > q$ with distinct autoregressive roots, the autocovariances will be given by

$$\gamma_j = h_1\lambda_1^j + h_2\lambda_2^j + \cdots + h_p\lambda_p^j. \qquad [3.5.6]$$

This takes the same form as the autocovariances for an *AR*(p) process [3.4.38], though because the initial conditions $(\gamma_0, \gamma_1, \ldots, \gamma_q)$ differ for the *ARMA* and *AR* processes, the parameters h_k in [3.5.6] will not be the same as the parameters g_k in [3.4.38].

There is a potential for redundant parameterization with *ARMA* processes. Consider, for example, a simple white noise process,

$$Y_t = \varepsilon_t. \qquad [3.5.7]$$

Suppose both sides of [3.5.7] are multiplied by $(1 - \rho L)$:

$$(1 - \rho L)Y_t = (1 - \rho L)\varepsilon_t. \qquad [3.5.8]$$

Clearly, if [3.5.7] is a valid representation, then so is [3.5.8] for any value of ρ. Thus, [3.5.8] might be described as an *ARMA*(1, 1) process, with $\phi_1 = \rho$ and $\theta_1 = -\rho$. It is important to avoid such a parameterization. Since any value of ρ in [3.5.8] describes the data equally well, we will obviously get into trouble trying to estimate the parameter ρ in [3.5.8] by maximum likelihood. Moreover, theoretical manipulations based on a representation such as [3.5.8] may overlook key cancellations. If we are using an *ARMA*(1, 1) model in which θ_1 is close to $-\phi_1$, then the data might better be modeled as simple white noise.

A related overparameterization can arise with an $ARMA(p, q)$ model. Consider factoring the lag polynomial operators in [3.5.2] as in [2.4.3]:

$$(1 - \lambda_1 L)(1 - \lambda_2 L) \cdots (1 - \lambda_p L)(Y_t - \mu) = (1 - \eta_1 L)(1 - \eta_2 L) \cdots (1 - \eta_q L)\varepsilon_t. \qquad [3.5.9]$$

We assume that $|\lambda_i| < 1$ for all i, so that the process is covariance-stationary. If the autoregressive operator $(1 - \phi_1 L - \phi_2 L^2 - \cdots - \phi_p L^p)$ and the moving average operator $(1 + \theta_1 L + \theta_2 L^2 + \cdots + \theta_q L^q)$ have any roots in common, say, $\lambda_i = \eta_j$ for some i and j, then both sides of [3.5.9] can be divided by $(1 - \lambda_i L)$:

$$\prod_{\substack{k=1 \\ k \neq i}}^{p} (1 - \lambda_k L)(Y_t - \mu) = \prod_{\substack{k=1 \\ k \neq j}}^{q} (1 - \eta_k L)\varepsilon_t,$$

or

$$(1 - \phi_1^* L - \phi_2^* L^2 - \cdots - \phi_{p-1}^* L^{p-1})(Y_t - \mu) = (1 + \theta_1^* L + \theta_2^* L^2 + \cdots + \theta_{q-1}^* L^{q-1})\varepsilon_t, \qquad [3.5.10]$$

where

$$(1 - \phi_1^* L - \phi_2^* L^2 - \cdots - \phi_{p-1}^* L^{p-1}) \equiv (1 - \lambda_1 L)(1 - \lambda_2 L) \cdots (1 - \lambda_{i-1} L)(1 - \lambda_{i+1} L) \cdots (1 - \lambda_p L)$$

$$(1 + \theta_1^* L + \theta_2^* L^2 + \cdots + \theta_{q-1}^* L^{q-1}) \equiv (1 - \eta_1 L)(1 - \eta_2 L) \cdots (1 - \eta_{j-1} L)(1 - \eta_{j+1} L) \cdots (1 - \eta_q L).$$

The stationary $ARMA(p, q)$ process satisfying [3.5.2] is clearly identical to the stationary $ARMA(p - 1, q - 1)$ process satisfying [3.5.10].

3.6. *The Autocovariance-Generating Function*

For each of the covariance-stationary processes for Y_t considered so far, we calculated the sequence of autocovariances $\{\gamma_j\}_{j=-\infty}^{\infty}$. If this sequence is absolutely summable, then one way of summarizing the autocovariances is through a scalar-valued function called the *autocovariance-generating function*:

$$g_Y(z) = \sum_{j=-\infty}^{\infty} \gamma_j z^j. \qquad [3.6.1]$$

This function is constructed by taking the jth autocovariance and multiplying it by some number z raised to the jth power, and then summing over all the possible values of j. The argument of this function (z) is taken to be a complex scalar.

Of particular interest as an argument for the autocovariance-generating function is any value of z that lies on the complex unit circle,

$$z = \cos(\omega) - i \sin(\omega) = e^{-i\omega},$$

where $i = \sqrt{-1}$ and ω is the radian angle that z makes with the real axis. If the autocovariance-generating function is evaluated at $z = e^{-i\omega}$ and divided by 2π, the resulting function of ω,

$$s_Y(\omega) = \frac{1}{2\pi} g_Y(e^{-i\omega}) = \frac{1}{2\pi} \sum_{j=-\infty}^{\infty} \gamma_j e^{-i\omega j},$$

is called the *population spectrum* of Y. The population spectrum will be discussed

in detail in Chapter 6. There it will be shown that for a process with absolutely summable autocovariances, the function $s_Y(\omega)$ exists and can be used to calculate all of the autocovariances. This means that if two different processes share the same autocovariance-generating function, then the two processes exhibit the identical sequence of autocovariances.

As an example of calculating an autocovariance-generating function, consider the $MA(1)$ process. From equations [3.3.3] to [3.3.5], its autocovariance-generating function is

$$g_Y(z) = [\theta\sigma^2]z^{-1} + [(1 + \theta^2)\sigma^2]z^0 + [\theta\sigma^2]z^1 = \sigma^2\cdot[\theta z^{-1} + (1 + \theta^2) + \theta z].$$

Notice that this expression could alternatively be written

$$g_Y(z) = \sigma^2(1 + \theta z)(1 + \theta z^{-1}). \qquad [3.6.2]$$

The form of expression [3.6.2] suggests that for the $MA(q)$ process,

$$Y_t = \mu + (1 + \theta_1 L + \theta_2 L^2 + \cdots + \theta_q L^q)\varepsilon_t,$$

the autocovariance-generating function might be calculated as

$$\begin{aligned} g_Y(z) = \sigma^2(&1 + \theta_1 z + \theta_2 z^2 + \cdots + \theta_q z^q) \\ &\times (1 + \theta_1 z^{-1} + \theta_2 z^{-2} + \cdots + \theta_q z^{-q}). \end{aligned} \qquad [3.6.3]$$

This conjecture can be verified by carrying out the multiplication in [3.6.3] and collecting terms by powers of z:

$$\begin{aligned} &(1 + \theta_1 z + \theta_2 z^2 + \cdots + \theta_q z^q) \times (1 + \theta_1 z^{-1} + \theta_2 z^{-2} + \cdots + \theta_q z^{-q}) \\ &= (\theta_q)z^q + (\theta_{q-1} + \theta_q\theta_1)z^{(q-1)} + (\theta_{q-2} + \theta_{q-1}\theta_1 + \theta_q\theta_2)z^{(q-2)} \\ &\quad + \cdots + (\theta_1 + \theta_2\theta_1 + \theta_3\theta_2 + \cdots + \theta_q\theta_{q-1})z^1 \\ &\quad + (1 + \theta_1^2 + \theta_2^2 + \cdots + \theta_q^2)z^0 \\ &\quad + (\theta_1 + \theta_2\theta_1 + \theta_3\theta_2 + \cdots + \theta_q\theta_{q-1})z^{-1} + \cdots + (\theta_q)z^{-q}. \end{aligned} \qquad [3.6.4]$$

Comparison of [3.6.4] with [3.3.10] or [3.3.12] confirms that the coefficient on z^j in [3.6.3] is indeed the jth autocovariance.

This method for finding $g_Y(z)$ extends to the $MA(\infty)$ case. If

$$Y_t = \mu + \psi(L)\varepsilon_t \qquad [3.6.5]$$

with

$$\psi(L) = \psi_0 + \psi_1 L + \psi_2 L^2 + \cdots \qquad [3.6.6]$$

and

$$\sum_{j=0}^{\infty} |\psi_j| < \infty, \qquad [3.6.7]$$

then

$$g_Y(z) = \sigma^2\psi(z)\psi(z^{-1}). \qquad [3.6.8]$$

For example, the stationary $AR(1)$ process can be written as

$$Y_t - \mu = (1 - \phi L)^{-1}\varepsilon_t,$$

which is in the form of [3.6.5] with $\psi(L) = 1/(1 - \phi L)$. The autocovariance-generating function for an $AR(1)$ process could therefore be calculated from

$$g_Y(z) = \frac{\sigma^2}{(1 - \phi z)(1 - \phi z^{-1})}. \qquad [3.6.9]$$

To verify this claim directly, expand out the terms in [3.6.9]:

$$\frac{\sigma^2}{(1-\phi z)(1-\phi z^{-1})} = \sigma^2(1+\phi z+\phi^2 z^2+\phi^3 z^3+\cdots)$$
$$\times(1+\phi z^{-1}+\phi^2 z^{-2}+\phi^3 z^{-3}+\cdots),$$

from which the coefficient on z^j is

$$\sigma^2(\phi^j+\phi^{j+1}\phi+\phi^{j+2}\phi^2+\cdots) = \sigma^2\phi^j/(1-\phi^2).$$

This indeed yields the jth autocovariance as earlier calculated in equation [3.4.5].

The autocovariance-generating function for a stationary $ARMA(p, q)$ process can be written

$$g_Y(z) = \frac{\sigma^2(1+\theta_1 z+\theta_2 z^2+\cdots+\theta_q z^q)(1+\theta_1 z^{-1}+\theta_2 z^{-2}+\cdots+\theta_q z^{-q})}{(1-\phi_1 z-\phi_2 z^2-\cdots-\phi_p z^p)(1-\phi_1 z^{-1}-\phi_2 z^{-2}-\cdots-\phi_p z^{-p})}. \qquad [3.6.10]$$

Filters

Sometimes the data are *filtered*, or treated in a particular way before they are analyzed, and we would like to summarize the effects of this treatment on the autocovariances. This calculation is particularly simple using the autocovariance-generating function. For example, suppose that the original data Y_t were generated from an $MA(1)$ process,

$$Y_t = (1+\theta L)\varepsilon_t, \qquad [3.6.11]$$

with autocovariance-generating function given by [3.6.2]. Let's say that the data as actually analyzed, X_t, represent the change in Y_t over its value the previous period:

$$X_t = Y_t - Y_{t-1} = (1-L)Y_t. \qquad [3.6.12]$$

Substituting [3.6.11] into [3.6.12], the observed data can be characterized as the following $MA(2)$ process,

$$X_t = (1-L)(1+\theta L)\varepsilon_t = [1+(\theta-1)L-\theta L^2]\varepsilon_t \equiv [1+\theta_1 L+\theta_2 L^2]\varepsilon_t, \qquad [3.6.13]$$

with $\theta_1 \equiv (\theta-1)$ and $\theta_2 \equiv -\theta$. The autocovariance-generating function of the observed data X_t can be calculated by direct application of [3.6.3]:

$$g_X(z) = \sigma^2(1+\theta_1 z+\theta_2 z^2)(1+\theta_1 z^{-1}+\theta_2 z^{-2}). \qquad [3.6.14]$$

It is often instructive, however, to keep the polynomial $(1+\theta_1 z+\theta_2 z^2)$ in its factored form of the first line of [3.6.13],

$$(1+\theta_1 z+\theta_2 z^2) = (1-z)(1+\theta z),$$

in which case [3.6.14] could be written

$$\begin{aligned} g_X(z) &= \sigma^2(1-z)(1+\theta z)(1-z^{-1})(1+\theta z^{-1}) \\ &= (1-z)(1-z^{-1})\cdot g_Y(z). \end{aligned} \qquad [3.6.15]$$

Of course, [3.6.14] and [3.6.15] represent the identical function of z, and which way we choose to write it is simply a matter of convenience. Applying the filter

$(1 - L)$ to Y_t thus results in multiplying its autocovariance-generating function by $(1 - z)(1 - z^{-1})$.

This principle readily generalizes. Suppose that the original data series $\{Y_t\}$ satisfies [3.6.5] through [3.6.7]. Let's say the data are filtered according to

$$X_t = h(L)Y_t \qquad [3.6.16]$$

with

$$h(L) = \sum_{j=-\infty}^{\infty} h_j L^j$$

$$\sum_{j=-\infty}^{\infty} |h_j| < \infty.$$

Substituting [3.6.5] into [3.6.16], the observed data X_t are then generated by

$$X_t = h(1)\mu + h(L)\psi(L)\varepsilon_t \equiv \mu^* + \psi^*(L)\varepsilon_t,$$

where $\mu^* \equiv h(1)\mu$ and $\psi^*(L) \equiv h(L)\psi(L)$. The sequence of coefficients associated with the compound operator $\{\psi_j^*\}_{j=-\infty}^{\infty}$ turns out to be absolutely summable,[5] and the autocovariance-generating function of X_t can accordingly be calculated as

$$g_X(z) = \sigma^2\psi^*(z)\psi^*(z^{-1}) = \sigma^2 h(z)\psi(z)\psi(z^{-1})h(z^{-1}) = h(z)h(z^{-1})g_Y(z). \qquad [3.6.17]$$

Applying the filter $h(L)$ to a series thus results in multiplying its autocovariance-generating function by $h(z)h(z^{-1})$.

3.7. *Invertibility*

Invertibility for the MA*(1) Process*

Consider an $MA(1)$ process,

$$Y_t - \mu = (1 + \theta L)\varepsilon_t, \qquad [3.7.1]$$

with

$$E(\varepsilon_t\varepsilon_\tau) = \begin{cases} \sigma^2 & \text{for } t = \tau \\ 0 & \text{otherwise.} \end{cases}$$

[5]Specifically,

$$\begin{aligned}\psi^*(z) &= \left(\sum_{j=-\infty}^{\infty} h_j z^j\right)\left(\sum_{k=0}^{\infty} \psi_k z^k\right) \\ &= (\cdots + h_{-j}z^{-j} + h_{-j+1}z^{-j+1} + \cdots + h_{-1}z^{-1} + h_0 z^0 + h_1 z^1 + \cdots \\ &\quad + h_j z^j + h_{j+1}z^{j+1} + \cdots)(\psi_0 z^0 + \psi_1 z^1 + \psi_2 z^2 + \cdots),\end{aligned}$$

from which the coefficient on z^j is

$$\psi_j^* = h_j\psi_0 + h_{j-1}\psi_1 + h_{j-2}\psi_2 + \cdots = \sum_{v=0}^{\infty} h_{j-v}\psi_v.$$

Then

$$\sum_{j=-\infty}^{\infty} |\psi_j^*| = \sum_{j=-\infty}^{\infty} \left|\sum_{v=0}^{\infty} h_{j-v}\psi_v\right| \le \sum_{j=-\infty}^{\infty}\sum_{v=0}^{\infty} |h_{j-v}\psi_v| = \sum_{v=0}^{\infty} |\psi_v| \sum_{j=-\infty}^{\infty} |h_{j-v}| = \sum_{v=0}^{\infty} |\psi_v| \sum_{j=-\infty}^{\infty} |h_j| < \infty.$$

Provided that $|\theta| < 1$, both sides of [3.7.1] can be multiplied by $(1 + \theta L)^{-1}$ to obtain[6]

$$(1 - \theta L + \theta^2 L^2 - \theta^3 L^3 + \cdots)(Y_t - \mu) = \varepsilon_t, \qquad [3.7.2]$$

which could be viewed as an $AR(\infty)$ representation. If a moving average representation such as [3.7.1] can be rewritten as an $AR(\infty)$ representation such as [3.7.2] simply by inverting the moving average operator $(1 + \theta L)$, then the moving average representation is said to be *invertible*. For an $MA(1)$ process, invertibility requires $|\theta| < 1$; if $|\theta| \geq 1$, then the infinite sequence in [3.7.2] would not be well defined.

Let us investigate what invertibility means in terms of the first and second moments of the process. Recall that the $MA(1)$ process [3.7.1] has mean μ and autocovariance-generating function

$$g_Y(z) = \sigma^2(1 + \theta z)(1 + \theta z^{-1}). \qquad [3.7.3]$$

Now consider a seemingly different $MA(1)$ process,

$$\tilde{Y}_t - \mu = (1 + \tilde{\theta} L)\tilde{\varepsilon}_t, \qquad [3.7.4]$$

with

$$E(\tilde{\varepsilon}_t \tilde{\varepsilon}_\tau) = \begin{cases} \tilde{\sigma}^2 & \text{for } t = \tau \\ 0 & \text{otherwise.} \end{cases}$$

Note that $\tilde{Y}_t$ has the same mean (μ) as Y_t. Its autocovariance-generating function is

$$\begin{aligned} g_{\tilde{Y}}(z) &= \tilde{\sigma}^2(1 + \tilde{\theta} z)(1 + \tilde{\theta} z^{-1}) \\ &= \tilde{\sigma}^2\{(\tilde{\theta}^{-1} z^{-1} + 1)(\tilde{\theta} z)\}\{(\tilde{\theta}^{-1} z + 1)(\tilde{\theta} z^{-1})\} \\ &= (\tilde{\sigma}^2 \tilde{\theta}^2)(1 + \tilde{\theta}^{-1} z)(1 + \tilde{\theta}^{-1} z^{-1}). \end{aligned} \qquad [3.7.5]$$

Suppose that the parameters of [3.7.4], $(\tilde{\theta}, \tilde{\sigma}^2)$, are related to those of [3.7.1] by the following equations:

$$\theta = \tilde{\theta}^{-1} \qquad [3.7.6]$$

$$\sigma^2 = \tilde{\theta}^2 \tilde{\sigma}^2. \qquad [3.7.7]$$

Then the autocovariance-generating functions [3.7.3] and [3.7.5] would be the same, meaning that Y_t and $\tilde{Y}_t$ would have identical first and second moments.

Notice from [3.7.6] that if $|\theta| < 1$, then $|\tilde{\theta}| > 1$. In other words, for any invertible $MA(1)$ representation [3.7.1], we have found a noninvertible $MA(1)$ representation [3.7.4] with the same first and second moments as the invertible representation. Conversely, given any noninvertible representation with $|\tilde{\theta}| > 1$, there exists an invertible representation with $\theta = (1/\tilde{\theta})$ that has the same first and second moments as the noninvertible representation. In the borderline case where $\theta = \pm 1$, there is only one representation of the process, and it is noninvertible.

Not only do the invertible and noninvertible representations share the same moments, either representation [3.7.1] or [3.7.4] could be used as an equally valid description of any given $MA(1)$ process! Suppose a computer generated an infinite sequence of $\tilde{Y}$'s according to [3.7.4] with $\tilde{\theta} > 1$. Thus we know for a fact that the data were generated from an $MA(1)$ process expressed in terms of a noninvertible representation. In what sense could these same data be associated with an invertible $MA(1)$ representation?

[6]Note from [2.2.8] that

$$(1 + \theta L)^{-1} = [1 - (-\theta)L]^{-1} = 1 + (-\theta)L + (-\theta)^2 L^2 + (-\theta)^3 L^3 + \cdots.$$

Imagine calculating a series $\{\varepsilon_t\}_{t=-\infty}^{\infty}$ defined by

$$\begin{aligned}\varepsilon_t &\equiv (1 + \theta L)^{-1}(\tilde{Y}_t - \mu) \\ &= (\tilde{Y}_t - \mu) - \theta(\tilde{Y}_{t-1} - \mu) + \theta^2(\tilde{Y}_{t-2} - \mu) - \theta^3(\tilde{Y}_{t-3} - \mu) + \cdots,\end{aligned} \qquad [3.7.8]$$

where $\theta = (1/\tilde{\theta})$ is the moving average parameter associated with the invertible $MA(1)$ representation that shares the same moments as [3.7.4]. Note that since $|\theta| < 1$, this produces a well-defined, mean square convergent series $\{\varepsilon_t\}$.

Furthermore, the sequence $\{\varepsilon_t\}$ so generated is white noise. The simplest way to verify this is to calculate the autocovariance-generating function of ε_t and confirm that the coefficient on z^j (the jth autocovariance) is equal to zero for any $j \neq 0$. From [3.7.8] and [3.6.17], the autocovariance-generating function for ε_t is given by

$$g_\varepsilon(z) = (1 + \theta z)^{-1}(1 + \theta z^{-1})^{-1} g_{\tilde{Y}}(z). \qquad [3.7.9]$$

Substituting [3.7.5] into [3.7.9],

$$\begin{aligned}g_\varepsilon(z) &= (1 + \theta z)^{-1}(1 + \theta z^{-1})^{-1}(\tilde{\sigma}^2\tilde{\theta}^2)(1 + \tilde{\theta}^{-1}z)(1 + \tilde{\theta}^{-1}z^{-1}) \\ &= \tilde{\sigma}^2\tilde{\theta}^2,\end{aligned} \qquad [3.7.10]$$

where the last equality follows from the fact that $\tilde{\theta}^{-1} = \theta$. Since the autocovariance-generating function is a constant, it follows that ε_t is a white noise process with variance $\tilde{\theta}^2\tilde{\sigma}^2$.

Multiplying both sides of [3.7.8] by $(1 + \theta L)$,

$$\tilde{Y}_t - \mu = (1 + \theta L)\varepsilon_t$$

is a perfectly valid invertible $MA(1)$ representation of data that were actually generated from the noninvertible representation [3.7.4].

The converse proposition is also true—suppose that the data were really generated from [3.7.1] with $|\theta| < 1$, an invertible representation. Then there exists a noninvertible representation with $\tilde{\theta} = 1/\theta$ that describes these data with equal validity. To characterize this noninvertible representation, consider the operator proposed in [2.5.20] as the appropriate inverse of $(1 + \tilde{\theta}L)$:

$$\begin{aligned}&(\tilde{\theta})^{-1}L^{-1}[1 - (\tilde{\theta}^{-1})L^{-1} + (\tilde{\theta}^{-2})L^{-2} - (\tilde{\theta}^{-3})L^{-3} + \cdots] \\ &\quad = \theta L^{-1}[1 - \theta L^{-1} + \theta^2 L^{-2} - \theta^3 L^{-3} + \cdots].\end{aligned}$$

Define $\tilde{\varepsilon}_t$ to be the series that results from applying this operator to $(Y_t - \mu)$,

$$\tilde{\varepsilon}_t \equiv \theta(Y_{t+1} - \mu) - \theta^2(Y_{t+2} - \mu) + \theta^3(Y_{t+3} - \mu) - \cdots, \qquad [3.7.11]$$

noting that this series converges for $|\theta| < 1$. Again this series is white noise:

$$\begin{aligned}g_{\tilde{\varepsilon}}(z) &= \{\theta z^{-1}[1 - \theta z^{-1} + \theta^2 z^{-2} - \theta^3 z^{-3} + \cdots]\} \\ &\quad \times \{\theta z[1 - \theta z^1 + \theta^2 z^2 - \theta^3 z^3 + \cdots]\}\sigma^2(1 + \theta z)(1 + \theta z^{-1}) \\ &= \theta^2\sigma^2.\end{aligned}$$

The coefficient on z^j is zero for $j \neq 0$, so $\tilde{\varepsilon}_t$ is white noise as claimed. Furthermore, by construction,

$$Y_t - \mu = (1 + \tilde{\theta}L)\tilde{\varepsilon}_t,$$

so that we have found a noninvertible $MA(1)$ representation of data that were actually generated by the invertible $MA(1)$ representation [3.7.1].

Either the invertible or the noninvertible representation could characterize any given data equally well, though there is a practical reason for preferring the

invertible representation. To find the value of ε for date t associated with the invertible representation as in [3.7.8], we need to know current and past values of Y. By contrast, to find the value of $\tilde{\varepsilon}$ for date t associated with the noninvertible representation as in [3.7.11], we need to use all of the future values of Y! If the intention is to calculate the current value of ε_t using real-world data, it will be feasible only to work with the invertible representation. Also, as will be noted in Chapters 4 and 5, some convenient algorithms for estimating parameters and forecasting are valid only if the invertible representation is used.

The value of ε_t associated with the invertible representation is sometimes called the *fundamental innovation* for Y_t. For the borderline case when $|\theta| = 1$, the process is noninvertible, but the innovation ε_t for such a process will still be described as the fundamental innovation for Y_t.

Invertibility for the MA(q) *Process*

Consider now the $MA(q)$ process,

$$(Y_t - \mu) = (1 + \theta_1 L + \theta_2 L^2 + \cdots + \theta_q L^q)\varepsilon_t \qquad [3.7.12]$$

$$E(\varepsilon_t \varepsilon_\tau) = \begin{cases} \sigma^2 & \text{for } t = \tau \\ 0 & \text{otherwise.} \end{cases}$$

Provided that the roots of

$$(1 + \theta_1 z + \theta_2 z^2 + \cdots + \theta_q z^q) = 0 \qquad [3.7.13]$$

lie outside the unit circle, [3.7.12] can be written as an $AR(\infty)$ simply by inverting the MA operator,

$$(1 + \eta_1 L + \eta_2 L^2 + \eta_3 L^3 + \cdots)(Y_t - \mu) = \varepsilon_t,$$

where

$$(1 + \eta_1 L + \eta_2 L^2 + \eta_3 L^3 + \cdots) = (1 + \theta_1 L + \theta_2 L^2 + \cdots + \theta_q L^q)^{-1}.$$

Where this is the case, the $MA(q)$ representation [3.7.12] is invertible.

Factor the moving average operator as

$$(1 + \theta_1 L + \theta_2 L^2 + \cdots + \theta_q L^q) = (1 - \lambda_1 L)(1 - \lambda_2 L) \cdots (1 - \lambda_q L). \qquad [3.7.14]$$

If $|\lambda_i| < 1$ for all i, then the roots of [3.7.13] are all outside the unit circle and the representation [3.7.12] is invertible. If instead some of the λ_i are outside (but not on) the unit circle, Hansen and Sargent (1981, p. 102) suggested the following procedure for finding an invertible representation. The autocovariance-generating function of Y_t can be written

$$\begin{aligned} g_Y(z) = \sigma^2 \cdot \{&(1 - \lambda_1 z)(1 - \lambda_2 z) \cdots (1 - \lambda_q z)\} \\ &\times \{(1 - \lambda_1 z^{-1})(1 - \lambda_2 z^{-1}) \cdots (1 - \lambda_q z^{-1})\}. \end{aligned} \qquad [3.7.15]$$

Order the λ's so that $(\lambda_1, \lambda_2, \ldots, \lambda_n)$ are inside the unit circle and $(\lambda_{n+1}, \lambda_{n+2}, \ldots, \lambda_q)$ are outside the unit circle. Suppose σ^2 in [3.7.15] is replaced by $\sigma^2 \cdot \lambda_{n+1}^2 \cdot \lambda_{n+2}^2 \cdots \lambda_q^2$; since complex λ_i appear as conjugate pairs, this is a positive real number. Suppose further that $(\lambda_{n+1}, \lambda_{n+2}, \ldots, \lambda_q)$ are replaced with their

reciprocals, $(\lambda_{n+1}^{-1}, \lambda_{n+2}^{-1}, \ldots, \lambda_q^{-1})$. The resulting function would be

$$\begin{aligned}
&\sigma^2\lambda_{n+1}^2\lambda_{n+2}^2 \cdots \lambda_q^2\left\{\prod_{i=1}^{n}(1-\lambda_i z)\right\}\left\{\prod_{i=n+1}^{q}(1-\lambda_i^{-1}z)\right\} \\
&\quad\times\left\{\prod_{i=1}^{n}(1-\lambda_i z^{-1})\right\}\left\{\prod_{i=n+1}^{q}(1-\lambda_i^{-1}z^{-1})\right\} \\
&= \sigma^2\left\{\prod_{i=1}^{n}(1-\lambda_i z)\right\}\left\{\prod_{i=n+1}^{q}[(\lambda_i z^{-1})(1-\lambda_i^{-1}z)]\right\} \\
&\quad\times\left\{\prod_{i=1}^{n}(1-\lambda_i z^{-1})\right\}\left\{\prod_{i=n+1}^{q}[(\lambda_i z)(1-\lambda_i^{-1}z^{-1})]\right\} \\
&= \sigma^2\left\{\prod_{i=1}^{n}(1-\lambda_i z)\right\}\left\{\prod_{i=n+1}^{q}(\lambda_i z^{-1}-1)\right\} \\
&\quad\times\left\{\prod_{i=1}^{n}(1-\lambda_i z^{-1})\right\}\left\{\prod_{i=n+1}^{q}(\lambda_i z-1)\right\} \\
&= \sigma^2\left\{\prod_{i=1}^{q}(1-\lambda_i z)\right\}\left\{\prod_{i=1}^{q}(1-\lambda_i z^{-1})\right\},
\end{aligned}$$

which is identical to [3.7.15].

The implication is as follows. Suppose a noninvertible representation for an $MA(q)$ process is written in the form

$$Y_t = \mu + \prod_{i=1}^{q}(1-\lambda_i L)\tilde{\varepsilon}_t, \qquad [3.7.16]$$

where

$$\begin{aligned}
|\lambda_i| &< 1 \qquad \text{for } i = 1, 2, \ldots, n \\
|\lambda_i| &> 1 \qquad \text{for } i = n+1, n+2, \ldots, q
\end{aligned}$$

and

$$E(\tilde{\varepsilon}_t\tilde{\varepsilon}_\tau) = \begin{cases} \tilde{\sigma}^2 & \text{for } t = \tau \\ 0 & \text{otherwise.} \end{cases}$$

Then the invertible representation is given by

$$Y_t = \mu + \left\{\prod_{i=1}^{n}(1-\lambda_i L)\right\}\left\{\prod_{i=n+1}^{q}(1-\lambda_i^{-1}L)\right\}\varepsilon_t, \qquad [3.7.17]$$

where

$$E(\varepsilon_t\varepsilon_\tau) = \begin{cases} \tilde{\sigma}^2\lambda_{n+1}^2\lambda_{n+2}^2 \cdots \lambda_q^2 & \text{for } t = \tau \\ 0 & \text{otherwise.} \end{cases}$$

Then [3.7.16] and [3.7.17] have the identical autocovariance-generating function, though only [3.7.17] satisfies the invertibility condition.

From the structure of the preceding argument, it is clear that there are a number of alternative $MA(q)$ representations of the data Y_t associated with all the possible "flips" between λ_i and λ_i^{-1}. Only one of these has all of the λ_i on or inside the unit circle. The innovations associated with this representation are said to be the fundamental innovations for Y_t.

APPENDIX 3.A. *Convergence Results for Infinite-Order Moving Average Processes*

This appendix proves the statements made in the text about convergence for the $MA(\infty)$ process [3.3.13].

First we show that absolute summability of the moving average coefficients implies square-summability. Suppose that $\{\psi_j\}_{j=0}^{\infty}$ is absolutely summable. Then there exists an $N < \infty$ such that $|\psi_j| < 1$ for all $j \geq N$, implying $\psi_j^2 < |\psi_j|$ for all $j \geq N$. Then

$$\sum_{j=0}^{\infty} \psi_j^2 = \sum_{j=0}^{N-1} \psi_j^2 + \sum_{j=N}^{\infty} \psi_j^2 < \sum_{j=0}^{N-1} \psi_j^2 + \sum_{j=N}^{\infty} |\psi_j|.$$

But $\Sigma_{j=0}^{N-1} \psi_j^2$ is finite, since N is finite, and $\Sigma_{j=N}^{\infty} |\psi_j|$ is finite, since $\{\psi_j\}$ is absolutely summable. Hence $\Sigma_{j=0}^{\infty} \psi_j^2 < \infty$, establishing that [3.3.15] implies [3.3.14].

Next we show that square-summability does not imply absolute summability. For an example of a series that is square-summable but not absolutely summable, consider $\psi_j = 1/j$ for $j = 1, 2, \ldots$. Notice that $1/j > 1/x$ for all $x > j$, meaning that

$$1/j > \int_j^{j+1} (1/x)\, dx$$

and so

$$\sum_{j=1}^{N} 1/j > \int_1^{N+1} (1/x)\, dx = \log(N + 1) - \log(1) = \log(N + 1),$$

which diverges to ∞ as $N \to \infty$. Hence $\{\psi_j\}_{j=1}^{\infty}$ is not absolutely summable. It is, however, square-summable, since $1/j^2 < 1/x^2$ for all $x < j$, meaning that

$$1/j^2 < \int_{j-1}^{j} (1/x^2)\, dx$$

and so

$$\sum_{j=1}^{N} 1/j^2 < 1 + \int_1^N (1/x^2)\, dx = 1 + (-1/x)|_{x=1}^{N} = 2 - (1/N),$$

which converges to 2 as $N \to \infty$. Hence $\{\psi_j\}_{j=1}^{\infty}$ is square-summable.

Next we show that square-summability of the moving average coefficients implies that the $MA(\infty)$ representation in [3.3.13] generates a mean square convergent random variable. First recall what is meant by convergence of a deterministic sum such as $\Sigma_{j=0}^{\infty} a_j$ where $\{a_j\}$ is just a sequence of numbers. One criterion for determining whether $\Sigma_{j=0}^{T} a_j$ converges to some finite number as $T \to \infty$ is the *Cauchy criterion*. The Cauchy criterion states that $\Sigma_{j=0}^{\infty} a_j$ converges if and only if, for any $\varepsilon > 0$, there exists a suitably large integer N such that, for any integer $M > N$,

$$\left| \sum_{j=0}^{M} a_j - \sum_{j=0}^{N} a_j \right| < \varepsilon.$$

In words, once we have summed N terms, calculating the sum out to a larger number M does not change the total by any more than an arbitrarily small number ε.

For a stochastic process such as [3.3.13], the comparable question is whether $\Sigma_{j=0}^{T} \psi_j \varepsilon_{t-j}$ converges in mean square to some random variable Y_t as $T \to \infty$. In this case the Cauchy criterion states that $\Sigma_{j=0}^{\infty} \psi_j \varepsilon_{t-j}$ converges if and only if, for any $\varepsilon > 0$, there exists a suitably large integer N such that for any integer $M > N$

$$E\left[\sum_{j=0}^{M} \psi_j \varepsilon_{t-j} - \sum_{j=0}^{N} \psi_j \varepsilon_{t-j} \right]^2 < \varepsilon. \qquad [3.A.1]$$

In words, once N terms have been summed, the difference between that sum and the one obtained from summing to M is a random variable whose mean and variance are both arbitrarily close to zero.

Now, the left side of [3.A.1] is simply

$$\begin{aligned} &E[\psi_M\varepsilon_{t-M} + \psi_{M-1}\varepsilon_{t-M+1} + \cdots + \psi_{N+1}\varepsilon_{t-N-1}]^2 \\ &= (\psi_M^2 + \psi_{M-1}^2 + \cdots + \psi_{N+1}^2)\cdot\sigma^2 \\ &= \left[\sum_{j=0}^{M}\psi_j^2 - \sum_{j=0}^{N}\psi_j^2\right]\cdot\sigma^2. \end{aligned} \qquad [3.A.2]$$

But if $\Sigma_{j=0}^{\infty}\psi_j^2$ converges as required by [3.3.14], then by the Cauchy criterion the right side of [3.A.2] may be made as small as desired by choice of a suitably large N. Thus the infinite series in [3.3.13] converges in mean square provided that [3.3.14] is satisfied.

Finally, we show that absolute summability of the moving average coefficients implies that the process is ergodic for the mean. Write [3.3.18] as

$$\gamma_j = \sigma^2\sum_{k=0}^{\infty}\psi_{j+k}\psi_k.$$

Then

$$|\gamma_j| = \sigma^2\left|\sum_{k=0}^{\infty}\psi_{j+k}\psi_k\right|.$$

A key property of the absolute value operator is that

$$|a + b + c| \le |a| + |b| + |c|.$$

Hence

$$|\gamma_j| \le \sigma^2\sum_{k=0}^{\infty}|\psi_{j+k}\psi_k|$$

and

$$\sum_{j=0}^{\infty}|\gamma_j| \le \sigma^2\sum_{j=0}^{\infty}\sum_{k=0}^{\infty}|\psi_{j+k}\psi_k| = \sigma^2\sum_{j=0}^{\infty}\sum_{k=0}^{\infty}|\psi_{j+k}|\cdot|\psi_k| = \sigma^2\sum_{k=0}^{\infty}|\psi_k|\sum_{j=0}^{\infty}|\psi_{j+k}|.$$

But there exists an $M < \infty$ such that $\Sigma_{j=0}^{\infty}|\psi_j| < M$, and therefore $\Sigma_{j=0}^{\infty}|\psi_{j+k}| < M$ for $k = 0, 1, 2, \ldots$, meaning that

$$\sum_{j=0}^{\infty}|\gamma_j| < \sigma^2\sum_{k=0}^{\infty}|\psi_k|\cdot M < \sigma^2M^2 < \infty.$$

Hence [3.1.15] holds and the process is ergodic for the mean.

Chapter 3 Exercises

3.1. Is the following $MA(2)$ process covariance-stationary?

$$Y_t = (1 + 2.4L + 0.8L^2)\varepsilon_t$$
$$E(\varepsilon_t\varepsilon_\tau) = \begin{cases} 1 & \text{for } t = \tau \\ 0 & \text{otherwise.} \end{cases}$$

If so, calculate its autocovariances.

3.2. Is the following $AR(2)$ process covariance-stationary?

$$(1 - 1.1L + 0.18L^2)Y_t = \varepsilon_t$$
$$E(\varepsilon_t\varepsilon_\tau) = \begin{cases} 1 & \text{for } t = \tau \\ 0 & \text{otherwise.} \end{cases}$$

If so, calculate its autocovariances.

3.3. A covariance-stationary $AR(p)$ process,

$$(1 - \phi_1L - \phi_2L^2 - \cdots - \phi_pL^p)(Y_t - \mu) = \varepsilon_t,$$

has an $MA(\infty)$ representation given by

$$(Y_t - \mu) = \psi(L)\varepsilon_t$$

with

$$\psi(L) = 1/[1 - \phi_1 L - \phi_2 L^2 - \cdots - \phi_p L^p]$$

or

$$[1 - \phi_1 L - \phi_2 L^2 - \cdots - \phi_p L^p]\,[\psi_0 + \psi_1 L + \psi_2 L^2 + \cdots] = 1.$$

In order for this equation to be true, the implied coefficient on L^0 must be unity and the coefficients on $L^1, L^2, L^3, \ldots$ must be zero. Write out these conditions explicitly and show that they imply a recursive algorithm for generating the $MA(\infty)$ weights $\psi_0, \psi_1, \ldots$. Show that this recursion is algebraically equivalent to setting ψ_j equal to the (1, 1) element of the matrix **F** raised to the jth power as in equation [1.2.28].

3.4. Derive [3.4.21].

3.5. Verify [3.4.22].

3.6. Suggest a recursive algorithm for calculating the $AR(\infty)$ weights,

$$(1 + \eta_1 L + \eta_2 L^2 + \cdots)(Y_t - \mu) = \varepsilon_t$$

associated with an invertible $MA(q)$ process,

$$(Y_t - \mu) = (1 + \theta_1 L + \theta_2 L^2 + \cdots + \theta_q L^q)\varepsilon_t.$$

Give a closed-form expression for η_j as a function of the roots of

$$(1 + \theta_1 z + \theta_2 z^2 + \cdots + \theta_q z^q) = 0,$$

assuming that these roots are all distinct.

3.7. Repeat Exercise 3.6 for a noninvertible $MA(q)$ process. (HINT: Recall equation [3.7.17].)

3.8. Show that the $MA(2)$ process in Exercise 3.1 is not invertible. Find the invertible representation for the process. Calculate the autocovariances of the invertible representation using equation [3.3.12] and verify that these are the same as obtained in Exercise 3.1.

Chapter 3 References

Anderson, Brian D. O., and John B. Moore. 1979. *Optimal Filtering*. Englewood Cliffs, N.J.: Prentice-Hall.

Hannan, E. J. 1970. *Multiple Time Series*. New York: Wiley.

Hansen, Lars P., and Thomas J. Sargent. 1981. "Formulating and Estimating Dynamic Linear Rational Expectations Models," in Robert E. Lucas, Jr., and Thomas J. Sargent, eds., *Rational Expectations and Econometric Practice*, Vol. I. Minneapolis: University of Minnesota Press.

Rao, C. Radhakrishna. 1973. *Linear Statistical Inference and Its Applications*, 2d ed. New York: Wiley.

Sargent, Thomas J. 1987. *Macroeconomic Theory*, 2d ed. Boston: Academic Press.

4 Forecasting

This chapter discusses how to forecast time series. Section 4.1 reviews the theory of forecasting and introduces the idea of a linear projection, which is a forecast formed from a linear function of past observations. Section 4.2 describes the forecasts one would use for *ARMA* models if an infinite number of past observations were available. These results are useful in theoretical manipulations and in understanding the formulas in Section 4.3 for approximate optimal forecasts when only a finite number of observations are available.

Section 4.4 describes how to achieve a triangular factorization and Cholesky factorization of a variance-covariance matrix. These results are used in that section to calculate exact optimal forecasts based on a finite number of observations. They will also be used in Chapter 11 to interpret vector autoregressions, in Chapter 13 to derive the Kalman filter, and in a number of other theoretical calculations and numerical methods appearing throughout the text. The triangular factorization is used to derive a formula for updating a forecast in Section 4.5 and to establish in Section 4.6 that for Gaussian processes the linear projection is better than any nonlinear forecast.

Section 4.7 analyzes what kind of process results when two different *ARMA* processes are added together. Section 4.8 states Wold's decomposition, which provides a basis for using an $MA(\infty)$ representation to characterize the linear forecast rule for any covariance-stationary process. The section also describes a popular empirical approach for finding a reasonable approximation to this representation that was developed by Box and Jenkins (1976).

4.1. *Principles of Forecasting*

Forecasts Based on Conditional Expectation

Suppose we are interested in forecasting the value of a variable Y_{t+1} based on a set of variables $\mathbf{X}_t$ observed at date t. For example, we might want to forecast Y_{t+1} based on its m most recent values. In this case, $\mathbf{X}_t$ would consist of a constant plus $Y_t, Y_{t-1}, \ldots,$ and Y_{t-m+1}.

Let $Y^*_{t+1|t}$ denote a forecast of Y_{t+1} based on $\mathbf{X}_t$. To evaluate the usefulness of this forecast, we need to specify a *loss function*, or a summary of how concerned we are if our forecast is off by a particular amount. Very convenient results are obtained from assuming a quadratic loss function. A quadratic loss function means choosing the forecast $Y^*_{t+1|t}$ so as to minimize

$$E(Y_{t+1} - Y^*_{t+1|t})^2. \qquad [4.1.1]$$

Expression [4.1.1] is known as the *mean squared error* associated with the forecast $Y^*_{t+1|t}$, denoted

$$MSE(Y^*_{t+1|t}) \equiv E(Y_{t+1} - Y^*_{t+1|t})^2.$$

The forecast with the smallest mean squared error turns out to be the expectation of Y_{t+1} conditional on $\mathbf{X}_t$:

$$Y^*_{t+1|t} = E(Y_{t+1}|\mathbf{X}_t). \qquad [4.1.2]$$

To verify this claim, consider basing $Y^*_{t+1|t}$ on any function $g(\mathbf{X}_t)$ other than the conditional expectation,

$$Y^*_{t+1|t} = g(\mathbf{X}_t). \qquad [4.1.3]$$

For this candidate forecasting rule, the *MSE* would be

$$\begin{aligned} E[Y_{t+1} - g(\mathbf{X}_t)]^2 &= E[Y_{t+1} - E(Y_{t+1}|\mathbf{X}_t) + E(Y_{t+1}|\mathbf{X}_t) - g(\mathbf{X}_t)]^2 \\ &= E[Y_{t+1} - E(Y_{t+1}|\mathbf{X}_t)]^2 \\ &\quad + 2E\{[Y_{t+1} - E(Y_{t+1}|\mathbf{X}_t)][E(Y_{t+1}|\mathbf{X}_t) - g(\mathbf{X}_t)]\} \\ &\quad + E\{[E(Y_{t+1}|\mathbf{X}_t) - g(\mathbf{X}_t)]^2\}. \end{aligned} \qquad [4.1.4]$$

Write the middle term on the right side of [4.1.4] as

$$2E[\eta_{t+1}], \qquad [4.1.5]$$

where

$$\eta_{t+1} \equiv \{[Y_{t+1} - E(Y_{t+1}|\mathbf{X}_t)][E(Y_{t+1}|\mathbf{X}_t) - g(\mathbf{X}_t)]\}.$$

Consider first the expectation of η_{t+1} conditional on $\mathbf{X}_t$. Conditional on $\mathbf{X}_t$, the terms $E(Y_{t+1}|\mathbf{X}_t)$ and $g(\mathbf{X}_t)$ are known constants and can be factored out of this expectation:[1]

$$\begin{aligned} E[\eta_{t+1}|\mathbf{X}_t] &= [E(Y_{t+1}|\mathbf{X}_t) - g(\mathbf{X}_t)] \times E([Y_{t+1} - E(Y_{t+1}|\mathbf{X}_t)]|\mathbf{X}_t) \\ &= [E(Y_{t+1}|\mathbf{X}_t) - g(\mathbf{X}_t)] \times 0 \\ &= 0. \end{aligned}$$

By a straightforward application of the law of iterated expectations, equation [A.5.10], it follows that

$$E[\eta_{t+1}] = E_{\mathbf{X}_t}(E[\eta_{t+1}|\mathbf{X}_t]) = 0.$$

Substituting this back into [4.1.4] gives

$$E[Y_{t+1} - g(\mathbf{X}_t)]^2 = E[Y_{t+1} - E(Y_{t+1}|\mathbf{X}_t)]^2 + E([E(Y_{t+1}|\mathbf{X}_t) - g(\mathbf{X}_t)]^2). \qquad [4.1.6]$$

The second term on the right side of [4.1.6] cannot be made smaller than zero, and the first term does not depend on $g(\mathbf{X}_t)$. The function $g(\mathbf{X}_t)$ that makes the mean squared error [4.1.6] as small as possible is the function that sets the second term in [4.1.6] to zero:

$$E(Y_{t+1}|\mathbf{X}_t) = g(\mathbf{X}_t). \qquad [4.1.7]$$

Thus the forecast $g(\mathbf{X}_t)$ that minimizes the mean squared error is the conditional expectation $E(Y_{t+1}|\mathbf{X}_t)$, as claimed.

The *MSE* of this optimal forecast is

$$E[Y_{t+1} - g(\mathbf{X}_t)]^2 = E[Y_{t+1} - E(Y_{t+1}|\mathbf{X}_t)]^2. \qquad [4.1.8]$$

[1]The conditional expectation $E(Y_{t+1}|\mathbf{X}_t)$ represents the conditional population moment of the random variable Y_{t+1} and is not a function of the random variable Y_{t+1} itself. For example, if $Y_{t+1}|\mathbf{X}_t \sim N(\boldsymbol{\alpha}'\mathbf{X}_t, \boldsymbol{\Omega})$, then $E(Y_{t+1}|\mathbf{X}_t) = \boldsymbol{\alpha}'\mathbf{X}_t$, which does not depend on Y_{t+1}.

Forecasts Based on Linear Projection

We now restrict the class of forecasts considered by requiring the forecast $Y^*_{t+1|t}$ to be a linear function of $\mathbf{X}_t$:

$$Y^*_{t+1|t} = \boldsymbol{\alpha}'\mathbf{X}_t. \qquad [4.1.9]$$

Suppose we were to find a value for $\boldsymbol{\alpha}$ such that the forecast error $(Y_{t+1} - \boldsymbol{\alpha}'\mathbf{X}_t)$ is uncorrelated with $\mathbf{X}_t$:

$$E[(Y_{t+1} - \boldsymbol{\alpha}'\mathbf{X}_t)\mathbf{X}_t'] = \mathbf{0}'. \qquad [4.1.10]$$

If [4.1.10] holds, then the forecast $\boldsymbol{\alpha}'\mathbf{X}_t$ is called the *linear projection* of Y_{t+1} on $\mathbf{X}_t$.

The linear projection turns out to produce the smallest mean squared error among the class of linear forecasting rules. The proof of this claim closely parallels the demonstration of the optimality of the conditional expectation among the set of all possible forecasts. Let $\mathbf{g}'\mathbf{X}_t$ denote any arbitrary linear forecasting rule. Note that its *MSE* is

$$\begin{aligned} &E[Y_{t+1} - \mathbf{g}'\mathbf{X}_t]^2 \\ &\quad = E[Y_{t+1} - \boldsymbol{\alpha}'\mathbf{X}_t + \boldsymbol{\alpha}'\mathbf{X}_t - \mathbf{g}'\mathbf{X}_t]^2 \\ &\quad = E[Y_{t+1} - \boldsymbol{\alpha}'\mathbf{X}_t]^2 + 2E\{[Y_{t+1} - \boldsymbol{\alpha}'\mathbf{X}_t][\boldsymbol{\alpha}'\mathbf{X}_t - \mathbf{g}'\mathbf{X}_t]\} \\ &\qquad + E[\boldsymbol{\alpha}'\mathbf{X}_t - \mathbf{g}'\mathbf{X}_t]^2. \end{aligned} \qquad [4.1.11]$$

As in the case of [4.1.4], the middle term on the right side of [4.1.11] is zero:

$$E([Y_{t+1} - \boldsymbol{\alpha}'\mathbf{X}_t][\boldsymbol{\alpha}'\mathbf{X}_t - \mathbf{g}'\mathbf{X}_t]) = (E[Y_{t+1} - \boldsymbol{\alpha}'\mathbf{X}_t]\mathbf{X}_t')[\boldsymbol{\alpha} - \mathbf{g}] = \mathbf{0}'[\boldsymbol{\alpha} - \mathbf{g}],$$

by virtue of [4.1.10]. Thus [4.1.11] simplifies to

$$E[Y_{t+1} - \mathbf{g}'\mathbf{X}_t]^2 = E[Y_{t+1} - \boldsymbol{\alpha}'\mathbf{X}_t]^2 + E[\boldsymbol{\alpha}'\mathbf{X}_t - \mathbf{g}'\mathbf{X}_t]^2. \qquad [4.1.12]$$

The optimal linear forecast $\mathbf{g}'\mathbf{X}_t$ is the value that sets the second term in [4.1.12] equal to zero:

$$\mathbf{g}'\mathbf{X}_t = \boldsymbol{\alpha}'\mathbf{X}_t,$$

where $\boldsymbol{\alpha}'\mathbf{X}_t$ satisfies [4.1.10].

For $\boldsymbol{\alpha}'\mathbf{X}_t$ satisfying [4.1.10], we will use the notation

$$\hat{P}(Y_{t+1}|\mathbf{X}_t) = \boldsymbol{\alpha}'\mathbf{X}_t,$$

or sometimes simply

$$\hat{Y}_{t+1|t} = \boldsymbol{\alpha}'\mathbf{X}_t,$$

to indicate the linear projection of Y_{t+1} on $\mathbf{X}_t$. Notice that

$$MSE[\hat{P}(Y_{t+1}|\mathbf{X}_t)] \geq MSE[E(Y_{t+1}|\mathbf{X}_t)],$$

since the conditional expectation offers the best possible forecast.

For most applications a constant term will be included in the projection. We will use the symbol $\hat{E}$ to indicate a linear projection on a vector of random variables $\mathbf{X}_t$ along with a constant term:

$$\hat{E}(Y_{t+1}|\mathbf{X}_t) \equiv \hat{P}(Y_{t+1}|1, \mathbf{X}_t).$$

Properties of Linear Projection

It is straightforward to use [4.1.10] to calculate the projection coefficient $\boldsymbol{\alpha}$ in terms of the moments of Y_{t+1} and $\mathbf{X}_t$:

$$E(Y_{t+1}\mathbf{X}_t') = \boldsymbol{\alpha}'E(\mathbf{X}_t\mathbf{X}_t'),$$

or

$$\boldsymbol{\alpha}' = E(Y_{t+1}\mathbf{X}_t')[E(\mathbf{X}_t\mathbf{X}_t')]^{-1}, \qquad [4.1.13]$$

assuming that $E(\mathbf{X}_t\mathbf{X}_t')$ is a nonsingular matrix. When $E(\mathbf{X}_t\mathbf{X}_t')$ is singular, the coefficient vector $\boldsymbol{\alpha}$ is not uniquely determined by [4.1.10], though the product of this vector with the explanatory variables, $\boldsymbol{\alpha}'\mathbf{X}_t$, is uniquely determined by [4.1.10].[2]

The *MSE* associated with a linear projection is given by

$$E(Y_{t+1} - \boldsymbol{\alpha}'\mathbf{X}_t)^2 = E(Y_{t+1})^2 - 2E(\boldsymbol{\alpha}'\mathbf{X}_tY_{t+1}) + E(\boldsymbol{\alpha}'\mathbf{X}_t\mathbf{X}_t'\boldsymbol{\alpha}). \qquad [4.1.14]$$

Substituting [4.1.13] into [4.1.14] produces

$$\begin{aligned} E(Y_{t+1} - \boldsymbol{\alpha}'\mathbf{X}_t)^2 &= E(Y_{t+1})^2 - 2E(Y_{t+1}\mathbf{X}_t')[E(\mathbf{X}_t\mathbf{X}_t')]^{-1}E(\mathbf{X}_tY_{t+1}) \\ &\quad + E(Y_{t+1}\mathbf{X}_t')[E(\mathbf{X}_t\mathbf{X}_t')]^{-1} \\ &\qquad \times E(\mathbf{X}_t\mathbf{X}_t')[E(\mathbf{X}_t\mathbf{X}_t')]^{-1}E(\mathbf{X}_tY_{t+1}) \\ &= E(Y_{t+1})^2 - E(Y_{t+1}\mathbf{X}_t')[E(\mathbf{X}_t\mathbf{X}_t')]^{-1}E(\mathbf{X}_tY_{t+1}). \end{aligned} \qquad [4.1.15]$$

Notice that if $\mathbf{X}_t$ includes a constant term, then the projection of $(aY_{t+1} + b)$ on $\mathbf{X}_t$ (where a and b are deterministic constants) is equal to

$$\hat{P}[(aY_{t+1} + b)|\mathbf{X}_t] = a\cdot\hat{P}(Y_{t+1}|\mathbf{X}_t) + b.$$

To see this, observe that $a\cdot\hat{P}(Y_{t+1}|\mathbf{X}_t) + b$ is a linear function of $\mathbf{X}_t$. Moreover, the forecast error,

$$[aY_{t+1} + b] - [a\cdot\hat{P}(Y_{t+1}|\mathbf{X}_t) + b] = a[Y_{t+1} - \hat{P}(Y_{t+1}|\mathbf{X}_t)],$$

is uncorrelated with $\mathbf{X}_t$, as required of a linear projection.

Linear Projection and Ordinary Least Squares Regression

Linear projection is closely related to ordinary least squares regression. This subsection discusses the relationship between the two concepts.

A linear regression model relates an observation on y_{t+1} to $\mathbf{x}_t$:

$$y_{t+1} = \boldsymbol{\beta}'\mathbf{x}_t + u_t. \qquad [4.1.16]$$

Given a sample of T observations on y and $\mathbf{x}$, the sample sum of squared residuals is defined as

$$\sum_{t=1}^{T} (y_{t+1} - \boldsymbol{\beta}'\mathbf{x}_t)^2. \qquad [4.1.17]$$

The value of $\boldsymbol{\beta}$ that minimizes [4.1.17], denoted $\mathbf{b}$, is the *ordinary least squares* (*OLS*) estimate of $\boldsymbol{\beta}$. The formula for $\mathbf{b}$ turns out to be

$$\mathbf{b} = \left[\sum_{t=1}^{T} \mathbf{x}_t\mathbf{x}_t'\right]^{-1}\left[\sum_{t=1}^{T} \mathbf{x}_ty_{t+1}\right], \qquad [4.1.18]$$

[2]If $E(\mathbf{X}_t\mathbf{X}_t')$ is singular, there exists a nonzero vector $\mathbf{c}$ such that $\mathbf{c}'\cdot E(\mathbf{X}_t\mathbf{X}_t')\cdot\mathbf{c} = E(\mathbf{c}'\mathbf{X}_t)^2 = 0$, so that some linear combination $\mathbf{c}'\mathbf{X}_t$ is equal to zero for all realizations. For example, if $\mathbf{X}_t$ consists of two random variables, the second variable must be a rescaled version of the first: $X_{2t} = c\cdot X_{1t}$. One could simply drop the redundant variables from such a system and calculate the linear projection of Y_{t+1} on $\mathbf{X}_t^*$, where $\mathbf{X}_t^*$ is a vector consisting of the nonredundant elements of $\mathbf{X}_t$. This linear projection $\boldsymbol{\alpha}^{*\prime}\mathbf{X}_t^*$ can be uniquely calculated from [4.1.13] with $\mathbf{X}_t$ in [4.1.13] replaced by $\mathbf{X}_t^*$. Any linear combination of the original variables $\boldsymbol{\alpha}'\mathbf{X}_t$ satisfying [4.1.10] represents this same random variable; that is, $\boldsymbol{\alpha}'\mathbf{X}_t = \boldsymbol{\alpha}^{*\prime}\mathbf{X}_t^*$ for all values of $\boldsymbol{\alpha}$ consistent with [4.1.10].

which equivalently can be written

$$\mathbf{b} = \left[(1/T) \sum_{t=1}^{T} \mathbf{x}_t \mathbf{x}_t' \right]^{-1} \left[(1/T) \sum_{t=1}^{T} \mathbf{x}_t y_{t+1} \right]. \qquad [4.1.19]$$

Comparing the *OLS* coefficient estimate $\mathbf{b}$ in equation [4.1.19] with the linear projection coefficient $\boldsymbol{\alpha}$ in equation [4.1.13], we see that $\mathbf{b}$ is constructed from the sample moments $(1/T)\Sigma_{t=1}^{T}\mathbf{x}_t\mathbf{x}_t'$ and $(1/T)\Sigma_{t=1}^{T}\mathbf{x}_t y_{t+1}$ while $\boldsymbol{\alpha}$ is constructed from population moments $E(\mathbf{X}_t\mathbf{X}_t')$ and $E(\mathbf{X}_t Y_{t+1})$. Thus *OLS* regression is a summary of the particular sample observations $(\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_T)$ and $(y_2, y_3, \ldots, y_{T+1})$, whereas linear projection is a summary of the population characteristics of the stochastic process $\{\mathbf{X}_t, Y_{t+1}\}_{t=-\infty}^{\infty}$.

Although linear projection describes population moments and ordinary least squares describes sample moments, there is a formal mathematical sense in which the two operations are the same. Appendix 4.A to this chapter discusses this parallel and shows how the formulas for an *OLS* regression can be viewed as a special case of the formulas for a linear projection.

Notice that if the stochastic process $\{\mathbf{X}_t, Y_{t+1}\}$ is covariance-stationary and ergodic for second moments, then the sample moments will converge to the population moments as the sample size T goes to infinity:

$$(1/T) \sum_{t=1}^{T} \mathbf{X}_t \mathbf{X}_t' \xrightarrow{p} E(\mathbf{X}_t \mathbf{X}_t')$$

$$(1/T) \sum_{t=1}^{T} \mathbf{X}_t Y_{t+1} \xrightarrow{p} E(\mathbf{X}_t Y_{t+1}),$$

implying

$$\mathbf{b} \xrightarrow{p} \boldsymbol{\alpha}. \qquad [4.1.20]$$

Thus *OLS* regression of y_{t+1} on $\mathbf{x}_t$ yields a consistent estimate of the linear projection coefficient. Note that this result requires only that the process be ergodic for second moments. By contrast, structural econometric analysis requires much stronger assumptions about the relation between $\mathbf{X}$ and Y. The difference arises because structural analysis seeks the *effect* of $\mathbf{X}$ on Y. In structural analysis, changes in $\mathbf{X}$ are associated with a particular structural event such as a change in Federal Reserve policy, and the objective is to evaluate the consequences for Y. Where that is the objective, it is very important to consider the nature of the correlation between $\mathbf{X}$ and Y before relying on *OLS* estimates. In the case of linear projection, however, the only concern is forecasting, for which it does not matter whether it is $\mathbf{X}$ that causes Y or Y that causes $\mathbf{X}$. Their observed historical comovements (as summarized by $E(\mathbf{X}_t Y_{t+1})$) are all that is needed for calculating a forecast. Result [4.1.20] shows that ordinary least squares regression provides a sound basis for forecasting under very mild assumptions.

One possible violation of these assumptions should nevertheless be noted. Result [4.1.20] was derived by assuming a covariance-stationary, ergodic process. However, the moments of the data may have changed over time in fundamental ways, or the future environment may be different from that in the past. Where this is the case, ordinary least squares may be undesirable, and better forecasts can emerge from careful structural analysis.

Forecasting Vectors

The preceding results can be extended to forecast an $(n \times 1)$ vector $\mathbf{Y}_{t+1}$ on the basis of a linear function of an $(m \times 1)$ vector $\mathbf{X}_t$:

$$\hat{P}(\mathbf{Y}_{t+1}|\mathbf{X}_t) = \boldsymbol{\alpha}'\mathbf{X}_t \equiv \hat{\mathbf{Y}}_{t+1|t}. \qquad [4.1.21]$$

Then $\boldsymbol{\alpha}'$ would denote an $(n \times m)$ matrix of projection coefficients satisfying

$$E[(\mathbf{Y}_{t+1} - \boldsymbol{\alpha}'\mathbf{X}_t)\mathbf{X}_t'] = \mathbf{0}; \qquad [4.1.22]$$

that is, each of the n elements of $(\mathbf{Y}_{t+1} - \hat{\mathbf{Y}}_{t+1|t})$ is uncorrelated with each of the m elements of $\mathbf{X}_t$. Accordingly, the jth element of the vector $\hat{\mathbf{Y}}_{t+1|t}$ gives the minimum *MSE* forecast of the scalar $Y_{j,t+1}$. Moreover, to forecast any linear combination of the elements of $\mathbf{Y}_{t+1}$, say, $z_{t+1} = \mathbf{h}'\mathbf{Y}_{t+1}$, the minimum *MSE* forecast of z_{t+1} requires $(z_{t+1} - \hat{z}_{t+1|t})$ to be uncorrelated with $\mathbf{X}_t$. But since each of the elements of $(\mathbf{Y}_{t+1} - \hat{\mathbf{Y}}_{t+1|t})$ is uncorrelated with $\mathbf{X}_t$, clearly $\mathbf{h}'(\mathbf{Y}_{t+1} - \hat{\mathbf{Y}}_{t+1|t})$ is also uncorrelated with $\mathbf{X}_t$. Thus when $\hat{\mathbf{Y}}_{t+1|t}$ satisfies [4.1.22], then $\mathbf{h}'\hat{\mathbf{Y}}_{t+1|t}$ is the minimum *MSE* forecast of $\mathbf{h}'\mathbf{Y}_{t+1}$ for any value of $\mathbf{h}$.

From [4.1.22], the matrix of projection coefficients is given by

$$\boldsymbol{\alpha}' = [E(\mathbf{Y}_{t+1}\mathbf{X}_t')] \cdot [E(\mathbf{X}_t\mathbf{X}_t')]^{-1}. \qquad [4.1.23]$$

The matrix generalization of the formula for the mean squared error [4.1.15] is

$$\begin{aligned} MSE(\boldsymbol{\alpha}'\mathbf{X}_t) &\equiv E\{[\mathbf{Y}_{t+1} - \boldsymbol{\alpha}'\mathbf{X}_t] \cdot [\mathbf{Y}_{t+1} - \boldsymbol{\alpha}'\mathbf{X}_t]'\} \\ &= E(\mathbf{Y}_{t+1}\mathbf{Y}_{t+1}') - [E(\mathbf{Y}_{t+1}\mathbf{X}_t')] \cdot [E(\mathbf{X}_t'\mathbf{X}_t)]^{-1} \cdot [E(\mathbf{X}_t\mathbf{Y}_{t+1}')]. \end{aligned} \qquad [4.1.24]$$

4.2. *Forecasts Based on an Infinite Number of Observations*

Forecasting Based on Lagged ε's

Consider a process with an $MA(\infty)$ representation

$$(Y_t - \mu) = \psi(L)\varepsilon_t \qquad [4.2.1]$$

with ε_t white noise and

$$\begin{aligned} \psi(L) &\equiv \sum_{j=0}^{\infty} \psi_j L^j \\ \psi_0 &= 1 \\ \sum_{j=0}^{\infty} |\psi_j| &< \infty. \end{aligned} \qquad [4.2.2]$$

Suppose that we have an infinite number of observations on ε through date t, $\{\varepsilon_t, \varepsilon_{t-1}, \varepsilon_{t-2}, \ldots\}$, and further know the values of μ and $\{\psi_1, \psi_2, \ldots\}$. Say we want to forecast the value of Y_{t+s}, that is, the value that Y will take on s periods from now. Note that [4.2.1] implies

$$\begin{aligned} Y_{t+s} = \mu + \varepsilon_{t+s} + \psi_1\varepsilon_{t+s-1} + \cdots + \psi_{s-1}\varepsilon_{t+1} + \psi_s\varepsilon_t \\ + \psi_{s+1}\varepsilon_{t-1} + \cdots. \end{aligned} \qquad [4.2.3]$$

The optimal linear forecast takes the form

$$\hat{E}[Y_{t+s}|\varepsilon_t, \varepsilon_{t-1}, \ldots] = \mu + \psi_s\varepsilon_t + \psi_{s+1}\varepsilon_{t-1} + \psi_{s+2}\varepsilon_{t-2} + \cdots. \qquad [4.2.4]$$

That is, the unknown future ε's are set to their expected value of zero. The error associated with this forecast is

$$Y_{t+s} - \hat{E}[Y_{t+s}|\varepsilon_t, \varepsilon_{t-1}, \ldots] = \varepsilon_{t+s} + \psi_1\varepsilon_{t+s-1} + \cdots + \psi_{s-1}\varepsilon_{t+1}. \quad [4.2.5]$$

In order for [4.2.4] to be the optimal linear forecast, condition [4.1.10] requires the forecast error to have mean zero and to be uncorrelated with ε_t, ε_{t-1}, ... It is readily confirmed that the error in [4.2.5] has these properties, so [4.2.4] must indeed be the linear projection, as claimed. The mean squared error associated with this forecast is

$$E(Y_{t+s} - \hat{E}[Y_{t+s}|\varepsilon_t, \varepsilon_{t-1}, \ldots])^2 = (1 + \psi_1^2 + \psi_2^2 + \cdots + \psi_{s-1}^2)\sigma^2. \quad [4.2.6]$$

For example, for an $MA(q)$ process,

$$\psi(L) = 1 + \theta_1 L + \theta_2 L^2 + \cdots + \theta_q L^q,$$

the optimal linear forecast is

$$\hat{E}[Y_{t+s}|\varepsilon_t, \varepsilon_{t-1}, \ldots] \quad [4.2.7]$$
$$= \begin{cases} \mu + \theta_s\varepsilon_t + \theta_{s+1}\varepsilon_{t-1} + \cdots + \theta_q\varepsilon_{t-q+s} & \text{for } s = 1, 2, \ldots, q \\ \mu & \text{for } s = q+1, q+2, \ldots \end{cases}$$

The MSE is

$$\begin{array}{ll} \sigma^2 & \text{for } s = 1 \\ (1 + \theta_1^2 + \theta_2^2 + \cdots + \theta_{s-1}^2)\sigma^2 & \text{for } s = 2, 3, \ldots, q \\ (1 + \theta_1^2 + \theta_2^2 + \cdots + \theta_q^2)\sigma^2 & \text{for } s = q+1, q+2, \ldots. \end{array}$$

The MSE increases with the forecast horizon s up until $s = q$. If we try to forecast an $MA(q)$ farther than q periods into the future, the forecast is simply the unconditional mean of the series ($E(Y_t) = \mu$) and the MSE is the unconditional variance of the series ($\text{Var}(Y_t) = (1 + \theta_1^2 + \theta_2^2 + \cdots + \theta_q^2)\sigma^2$).

These properties also characterize the $MA(\infty)$ case as the forecast horizon s goes to infinity. It is straightforward to establish from [4.2.2] that as $s \to \infty$, the forecast in [4.2.4] converges in mean square to μ, the unconditional mean. The MSE [4.2.6] likewise converges to $\sigma^2 \sum_{j=0}^{\infty}\psi_j^2$, which is the unconditional variance of the $MA(\infty)$ process [4.2.1].

A compact lag operator expression for the forecast in [4.2.4] is sometimes used. Consider taking the polynomial $\psi(L)$ and dividing by L^s:

$$\frac{\psi(L)}{L^s} = L^{-s} + \psi_1 L^{1-s} + \psi_2 L^{2-s} + \cdots + \psi_{s-1}L^{-1} + \psi_s L^0$$
$$+ \psi_{s+1}L^1 + \psi_{s+2}L^2 + \cdots.$$

The *annihilation operator*[3] (indicated by $[\cdot]_+$) replaces negative powers of L by zero; for example,

$$\left[\frac{\psi(L)}{L^s}\right]_+ = \psi_s + \psi_{s+1}L^1 + \psi_{s+2}L^2 + \cdots. \quad [4.2.8]$$

Comparing [4.2.8] with [4.2.4], the optimal forecast could be written in lag operator notation as

$$\hat{E}[Y_{t+s}|\varepsilon_t, \varepsilon_{t-1}, \ldots] = \mu + \left[\frac{\psi(L)}{L^s}\right]_+ \varepsilon_t. \quad [4.2.9]$$

[3]This discussion of forecasting based on the annihilation operator is similar to that in Sargent (1987).

Forecasting Based on Lagged Y's

The previous forecasts were based on the assumption that ε_t is observed directly. In the usual forecasting situation, we actually have observations on lagged Y's, not lagged ε's. Suppose that the process [4.2.1] has an $AR(\infty)$ representation given by

$$\eta(L)(Y_t - \mu) = \varepsilon_t, \qquad [4.2.10]$$

where $\eta(L) \equiv \Sigma_{j=0}^{\infty}\eta_j L^j$, $\eta_0 = 1$, and $\Sigma_{j=0}^{\infty}|\eta_j| < \infty$. Suppose further that the AR polynomial $\eta(L)$ and the MA polynomial $\psi(L)$ are related by

$$\eta(L) = [\psi(L)]^{-1}. \qquad [4.2.11]$$

A covariance-stationary $AR(p)$ model of the form

$$(1 - \phi_1 L - \phi_2 L^2 - \cdots - \phi_p L^p)(Y_t - \mu) = \varepsilon_t, \qquad [4.2.12]$$

or, more compactly,

$$\phi(L)(Y_t - \mu) = \varepsilon_t,$$

clearly satisfies these requirements, with $\eta(L) = \phi(L)$ and $\psi(L) = [\phi(L)]^{-1}$. An $MA(q)$ process

$$Y_t - \mu = (1 + \theta_1 L + \theta_2 L^2 + \cdots + \theta_q L^q)\varepsilon_t \qquad [4.2.13]$$

or

$$Y_t - \mu = \theta(L)\varepsilon_t$$

is also of this form, with $\psi(L) = \theta(L)$ and $\eta(L) = [\theta(L)]^{-1}$, provided that [4.2.13] is based on the invertible representation. With a noninvertible $MA(q)$, the roots must first be flipped as described in Section 3.7 before applying the formulas given in this section. An $ARMA(p, q)$ also satisfies [4.2.10] and [4.2.11] with $\psi(L) = \theta(L)/\phi(L)$, provided that the autoregressive operator $\phi(L)$ satisfies the stationarity condition (roots of $\phi(z) = 0$ lie outside the unit circle) and that the moving average operator $\theta(L)$ satisfies the invertibility condition (roots of $\theta(z) = 0$ lie outside the unit circle).

Where the restrictions associated with [4.2.10] and [4.2.11] are satisfied, observations on $\{Y_t, Y_{t-1}, \ldots\}$ will be sufficient to construct $\{\varepsilon_t, \varepsilon_{t-1}, \ldots\}$. For example, for an $AR(1)$ process [4.2.10] would be

$$(1 - \phi L)(Y_t - \mu) = \varepsilon_t. \qquad [4.2.14]$$

Thus, given ϕ and μ and observation of Y_t and Y_{t-1}, the value of ε_t can be constructed from

$$\varepsilon_t = (Y_t - \mu) - \phi(Y_{t-1} - \mu).$$

For an $MA(1)$ process written in invertible form, [4.2.10] would be

$$(1 + \theta L)^{-1}(Y_t - \mu) = \varepsilon_t.$$

Given an infinite number of observations on Y, we could construct ε from

$$\begin{aligned}\varepsilon_t = (Y_t - \mu) &- \theta(Y_{t-1} - \mu) + \theta^2(Y_{t-2} - \mu) \\ &- \theta^3(Y_{t-3} - \mu) + \cdots.\end{aligned} \qquad [4.2.15]$$

Under these conditions, [4.2.10] can be substituted into [4.2.9] to obtain the forecast of Y_{t+s} as a function of lagged Y's:

$$\hat{E}[Y_{t+s}|Y_t, Y_{t-1}, \ldots] = \mu + \left[\frac{\psi(L)}{L^s}\right]_+ \eta(L)(Y_t - \mu);$$

or, using [4.2.11],

$$\hat{E}[Y_{t+s}|Y_t, Y_{t-1}, \ldots] = \mu + \left[\frac{\psi(L)}{L^s}\right]_+ \frac{1}{\psi(L)}(Y_t - \mu). \qquad [4.2.16]$$

Equation [4.2.16] is known as the *Wiener-Kolmogorov prediction formula*. Several examples of using this forecasting rule follow.

Forecasting an AR*(1) Process*

For the covariance-stationary $AR(1)$ process [4.2.14], we have

$$\psi(L) = 1/(1 - \phi L) = 1 + \phi L + \phi^2 L^2 + \phi^3 L^3 + \cdots \qquad [4.2.17]$$

and

$$\left[\frac{\psi(L)}{L^s}\right]_+ = \phi^s + \phi^{s+1}L^1 + \phi^{s+2}L^2 + \cdots = \phi^s/(1 - \phi L). \qquad [4.2.18]$$

Substituting [4.2.18] into [4.2.16] yields the optimal linear s-period-ahead forecast for a stationary $AR(1)$ process:

$$\begin{aligned}\hat{E}[Y_{t+s}|Y_t, Y_{t-1}, \ldots] &= \mu + \frac{\phi^s}{1 - \phi L}(1 - \phi L)(Y_t - \mu) \\ &= \mu + \phi^s(Y_t - \mu).\end{aligned} \qquad [4.2.19]$$

The forecast decays geometrically from $(Y_t - \mu)$ toward μ as the forecast horizon s increases. From [4.2.17], the moving average weight ψ_j is given by ϕ^j, so from [4.2.6], the mean squared s-period-ahead forecast error is

$$[1 + \phi^2 + \phi^4 + \cdots + \phi^{2(s-1)}]\sigma^2.$$

Notice that this grows with s and asymptotically approaches $\sigma^2/(1 - \phi^2)$, the unconditional variance of Y.

Forecasting an AR(p) *Process*

Next consider forecasting the stationary $AR(p)$ process [4.2.12]. The Wiener-Kolmogorov formula in [4.2.16] essentially expresses the value of $(Y_{t+s} - \mu)$ in terms of initial values $\{(Y_t - \mu), (Y_{t-1} - \mu), \ldots\}$ and subsequent values of $\{\varepsilon_{t+1}, \varepsilon_{t+2}, \ldots, \varepsilon_{t+s}\}$ and then drops the terms involving future ε's. An expression of this form was provided by equation [1.2.26], which described the value of a variable subject to a pth-order difference equation in terms of initial conditions and subsequent shocks:

$$\begin{aligned}Y_{t+s} - \mu = {} & f_{11}^{(s)}(Y_t - \mu) + f_{12}^{(s)}(Y_{t-1} - \mu) + \cdots + f_{1p}^{(s)}(Y_{t-p+1} - \mu) \\ & + \varepsilon_{t+s} + \psi_1\varepsilon_{t+s-1} + \psi_2\varepsilon_{t+s-2} + \cdots + \psi_{s-1}\varepsilon_{t+1},\end{aligned} \qquad [4.2.20]$$

where

$$\psi_j = f_{11}^{(j)}. \qquad [4.2.21]$$

Recall that $f_{11}^{(j)}$ denotes the (1, 1) element of $\mathbf{F}^j$, $f_{12}^{(j)}$ denotes the (1, 2) element of $\mathbf{F}^j$, and so on, where $\mathbf{F}$ is the following $(p \times p)$ matrix:

$$\mathbf{F} \equiv \begin{bmatrix} \phi_1 & \phi_2 & \phi_3 & \cdots & \phi_{p-1} & \phi_p \\ 1 & 0 & 0 & \cdots & 0 & 0 \\ 0 & 1 & 0 & \cdots & 0 & 0 \\ \vdots & \vdots & \vdots & \cdots & \vdots & \vdots \\ 0 & 0 & 0 & \cdots & 1 & 0 \end{bmatrix}.$$

The optimal s-period-ahead forecast is thus

$$\begin{aligned} \hat{Y}_{t+s|t} = \mu &+ f_{11}^{(s)}(Y_t - \mu) + f_{12}^{(s)}(Y_{t-1} - \mu) + \cdots \\ &+ f_{1p}^{(s)}(Y_{t-p+1} - \mu). \end{aligned} \qquad [4.2.22]$$

Notice that for any forecast horizon s the optimal forecast is a constant plus a linear function of $\{Y_t, Y_{t-1}, \ldots, Y_{t-p+1}\}$. The associated forecast error is

$$Y_{t+s} - \hat{Y}_{t+s|t} = \varepsilon_{t+s} + \psi_1\varepsilon_{t+s-1} + \psi_2\varepsilon_{t+s-2} + \cdots + \psi_{s-1}\varepsilon_{t+1}. \qquad [4.2.23]$$

The easiest way to calculate the forecast in [4.2.22] is through a simple recursion. This recursion can be deduced independently from a principle known as the *law of iterated projections*, which will be proved formally in Section 4.5. Suppose that at date t we wanted to make a one-period-ahead forecast of Y_{t+1}. The optimal forecast is clearly

$$\begin{aligned} (\hat{Y}_{t+1|t} - \mu) = \phi_1(Y_t - \mu) &+ \phi_2(Y_{t-1} - \mu) + \cdots \\ &+ \phi_p(Y_{t-p+1} - \mu). \end{aligned} \qquad [4.2.24]$$

Consider next a two-period-ahead forecast. Suppose that at date $t + 1$ we were to make a one-period-ahead forecast of Y_{t+2}. Replacing t with $t + 1$ in [4.2.24] gives the optimal forecast as

$$\begin{aligned} (\hat{Y}_{t+2|t+1} - \mu) = \phi_1(Y_{t+1} - \mu) &+ \phi_2(Y_t - \mu) + \cdots \\ &+ \phi_p(Y_{t-p+2} - \mu). \end{aligned} \qquad [4.2.25]$$

The law of iterated projections asserts that if this date $t + 1$ forecast of Y_{t+2} is projected on date t information, the result is the date t forecast of Y_{t+2}. At date t the values $Y_t, Y_{t-1}, \ldots, Y_{t-p+2}$ in [4.2.25] are known. Thus,

$$\begin{aligned} (\hat{Y}_{t+2|t} - \mu) = \phi_1(\hat{Y}_{t+1|t} - \mu) &+ \phi_2(Y_t - \mu) + \cdots \\ &+ \phi_p(Y_{t-p+2} - \mu). \end{aligned} \qquad [4.2.26]$$

Substituting [4.2.24] into [4.2.26] then yields the two-period-ahead forecast for an $AR(p)$ process:

$$\begin{aligned} (\hat{Y}_{t+2|t} - \mu) &= \phi_1[\phi_1(Y_t - \mu) + \phi_2(Y_{t-1} - \mu) + \cdots + \phi_p(Y_{t-p+1} - \mu)] \\ &\quad + \phi_2(Y_t - \mu) + \phi_3(Y_{t-1} - \mu) + \cdots + \phi_p(Y_{t-p+2} - \mu) \\ &= (\phi_1^2 + \phi_2)(Y_t - \mu) + (\phi_1\phi_2 + \phi_3)(Y_{t-1} - \mu) + \cdots \\ &\quad + (\phi_1\phi_{p-1} + \phi_p)(Y_{t-p+2} - \mu) + \phi_1\phi_p(Y_{t-p+1} - \mu). \end{aligned}$$

The s-period-ahead forecasts of an $AR(p)$ process can be obtained by iterating on

$$\begin{aligned} (\hat{Y}_{t+j|t} - \mu) = \phi_1(\hat{Y}_{t+j-1|t} - \mu) &+ \phi_2(\hat{Y}_{t+j-2|t} - \mu) + \cdots \\ &+ \phi_p(\hat{Y}_{t+j-p|t} - \mu) \end{aligned} \qquad [4.2.27]$$

for $j = 1, 2, \ldots, s$ where

$$\hat{Y}_{\tau|t} = Y_\tau \qquad \text{for } \tau \leq t.$$

Forecasting an MA*(1) Process*

Next consider an invertible $MA(1)$ representation,

$$Y_t - \mu = (1 + \theta L)\varepsilon_t \qquad [4.2.28]$$

with $|\theta| < 1$. Replacing $\psi(L)$ in the Wiener-Kolmogorov formula [4.2.16] with $(1 + \theta L)$ gives

$$\hat{Y}_{t+s|t} = \mu + \left[\frac{1 + \theta L}{L^s}\right]_+ \frac{1}{1 + \theta L}(Y_t - \mu). \qquad [4.2.29]$$

To forecast an $MA(1)$ process one period into the future ($s = 1$),

$$\left[\frac{1 + \theta L}{L^1}\right]_+ = \theta,$$

and so

$$\begin{aligned}\hat{Y}_{t+1|t} &= \mu + \frac{\theta}{1 + \theta L}(Y_t - \mu) \\ &= \mu + \theta(Y_t - \mu) - \theta^2(Y_{t-1} - \mu) + \theta^3(Y_{t-2} - \mu) - \cdots.\end{aligned} \qquad [4.2.30]$$

It is sometimes useful to write [4.2.28] as

$$\varepsilon_t = \frac{1}{1 + \theta L}(Y_t - \mu)$$

and view ε_t as the outcome of an infinite recursion,

$$\hat{\varepsilon}_t = (Y_t - \mu) - \theta\hat{\varepsilon}_{t-1}. \qquad [4.2.31]$$

The one-period-ahead forecast [4.2.30] could then be written as

$$\hat{Y}_{t+1|t} = \mu + \theta\hat{\varepsilon}_t. \qquad [4.2.32]$$

Equation [4.2.31] is in fact an exact characterization of ε_t, deduced from simple rearrangement of [4.2.28]. The "hat" notation ($\hat{\varepsilon}_t$) is introduced at this point in anticipation of the approximations to ε_t that will be introduced in the following section and substituted into [4.2.31] and [4.2.32].

To forecast an $MA(1)$ process for $s = 2, 3, \ldots$ periods into the future,

$$\left[\frac{1 + \theta L}{L^s}\right]_+ = 0 \qquad \text{for } s = 2, 3, \ldots;$$

and so, from [4.2.29],

$$\hat{Y}_{t+s|t} = \mu \qquad \text{for } s = 2, 3, \ldots. \qquad [4.2.33]$$

Forecasting an MA(q) *Process*

For an invertible $MA(q)$ process,

$$(Y_t - \mu) = (1 + \theta_1 L + \theta_2 L^2 + \cdots + \theta_q L^q)\varepsilon_t,$$

the forecast [4.2.16] becomes

$$\hat{Y}_{t+s|t} = \mu + \left[\frac{1 + \theta_1 L + \theta_2 L^2 + \cdots + \theta_q L^q}{L^s}\right]_+ \times \frac{1}{1 + \theta_1 L + \theta_2 L^2 + \cdots + \theta_q L^q}(Y_t - \mu). \qquad [4.2.34]$$

Now

$$\left[\frac{1 + \theta_1 L + \theta_2 L^2 + \cdots + \theta_q L^q}{L^s}\right]_+ = \begin{cases} \theta_s + \theta_{s+1}L + \theta_{s+2}L^2 + \cdots + \theta_q L^{q-s} & \text{for } s = 1, 2, \ldots, q \\ 0 & \text{for } s = q + 1, q + 2, \ldots. \end{cases}$$

Thus, for horizons of $s = 1, 2, \ldots, q$, the forecast is given by

$$\hat{Y}_{t+s|t} = \mu + (\theta_s + \theta_{s+1}L + \theta_{s+2}L^2 + \cdots + \theta_q L^{q-s})\hat{\varepsilon}_t, \qquad [4.2.35]$$

where $\hat{\varepsilon}_t$ can be characterized by the recursion

$$\hat{\varepsilon}_t = (Y_t - \mu) - \theta_1\hat{\varepsilon}_{t-1} - \theta_2\hat{\varepsilon}_{t-2} - \cdots - \theta_q\hat{\varepsilon}_{t-q}. \qquad [4.2.36]$$

A forecast farther than q periods into the future is simply the unconditional mean μ.

Forecasting an ARMA*(1, 1) Process*

For an *ARMA*(1, 1) process

$$(1 - \phi L)(Y_t - \mu) = (1 + \theta L)\varepsilon_t$$

that is stationary ($|\phi| < 1$) and invertible ($|\theta| < 1$),

$$\hat{Y}_{t+s|t} = \mu + \left[\frac{1 + \theta L}{(1 - \phi L)L^s}\right]_+ \frac{1 - \phi L}{1 + \theta L}(Y_t - \mu). \qquad [4.2.37]$$

Here

$$\begin{aligned} &\left[\frac{1 + \theta L}{(1 - \phi L)L^s}\right]_+ \\ &= \left[\frac{(1 + \phi L + \phi^2 L^2 + \cdots)}{L^s} + \frac{\theta L(1 + \phi L + \phi^2 L^2 + \cdots)}{L^s}\right]_+ \\ &= (\phi^s + \phi^{s+1}L + \phi^{s+2}L^2 + \cdots) + \theta(\phi^{s-1} + \phi^s L + \phi^{s+1}L^2 + \cdots) \\ &= (\phi^s + \theta\phi^{s-1})(1 + \phi L + \phi^2 L^2 + \cdots) \\ &= \frac{\phi^s + \theta\phi^{s-1}}{1 - \phi L}. \end{aligned} \qquad [4.2.38]$$

Substituting [4.2.38] into [4.2.37] gives

$$\begin{aligned} \hat{Y}_{t+s|t} &= \mu + \left[\frac{\phi^s + \theta\phi^{s-1}}{1 - \phi L}\right]\frac{1 - \phi L}{1 + \theta L}(Y_t - \mu) \\ &= \mu + \frac{\phi^s + \theta\phi^{s-1}}{1 + \theta L}(Y_t - \mu). \end{aligned} \qquad [4.2.39]$$

Note that for $s = 2, 3, \ldots$, the forecast [4.2.39] obeys the recursion

$$(\hat{Y}_{t+s|t} - \mu) = \phi(\hat{Y}_{t+s-1|t} - \mu).$$

Thus, beyond one period, the forecast decays geometrically at the rate ϕ toward the unconditional mean μ. The one-period-ahead forecast ($s = 1$) is given by

$$\hat{Y}_{t+1|t} = \mu + \frac{\phi + \theta}{1 + \theta L}(Y_t - \mu). \qquad [4.2.40]$$

This can equivalently be written

$$(\hat{Y}_{t+1|t} - \mu) = \frac{\phi(1 + \theta L) + \theta(1 - \phi L)}{1 + \theta L}(Y_t - \mu) = \phi(Y_t - \mu) + \theta\hat{\varepsilon}_t \qquad [4.2.41]$$

where

$$\hat{\varepsilon}_t = \frac{(1 - \phi L)}{(1 + \theta L)}(Y_t - \mu)$$

or

$$\hat{\varepsilon}_t = (Y_t - \mu) - \phi(Y_{t-1} - \mu) - \theta\hat{\varepsilon}_{t-1} = Y_t - \hat{Y}_{t|t-1}. \qquad [4.2.42]$$

Forecasting an ARMA(p, q) *Process*

Finally, consider forecasting a stationary and invertible $ARMA(p, q)$ process:

$$(1 - \phi_1 L - \phi_2 L^2 - \cdots - \phi_p L^p)(Y_t - \mu) = (1 + \theta_1 L + \theta_2 L^2 + \cdots + \theta_q L^q)\varepsilon_t.$$

The natural generalizations of [4.2.41] and [4.2.42] are

$$\begin{aligned}(\hat{Y}_{t+1|t} - \mu) = {} & \phi_1(Y_t - \mu) + \phi_2(Y_{t-1} - \mu) + \cdots \\ & + \phi_p(Y_{t-p+1} - \mu) + \theta_1\hat{\varepsilon}_t + \theta_2\hat{\varepsilon}_{t-1} + \cdots + \theta_q\hat{\varepsilon}_{t-q+1},\end{aligned} \qquad [4.2.43]$$

with $\{\hat{\varepsilon}_t\}$ generated recursively from

$$\hat{\varepsilon}_t = Y_t - \hat{Y}_{t|t-1}. \qquad [4.2.44]$$

The s-period-ahead forecasts would be

$$(\hat{Y}_{t+s|t} - \mu) \qquad [4.2.45]$$

$$= \begin{cases} \phi_1(\hat{Y}_{t+s-1|t} - \mu) + \phi_2(\hat{Y}_{t+s-2|t} - \mu) + \cdots + \phi_p(\hat{Y}_{t+s-p|t} - \mu) \\ \quad + \theta_s\hat{\varepsilon}_t + \theta_{s+1}\hat{\varepsilon}_{t-1} + \cdots + \theta_q\hat{\varepsilon}_{t+s-q} \qquad \text{for } s = 1, 2, \ldots, q \\ \phi_1(\hat{Y}_{t+s-1|t} - \mu) + \phi_2(\hat{Y}_{t+s-2|t} - \mu) + \cdots + \phi_p(\hat{Y}_{t+s-p|t} - \mu) \\ \qquad\qquad \text{for } s = q + 1, q + 2, \ldots, \end{cases}$$

where

$$\hat{Y}_{\tau|t} = Y_\tau \qquad \text{for } \tau \leq t.$$

Thus for a forecast horizon s greater than the moving average order q, the forecasts follow a pth-order difference equation governed solely by the autoregressive parameters.

4.3. *Forecasts Based on a Finite Number of Observations*

The formulas in the preceding section assumed that we had an infinite number of past observations on Y, $\{Y_t, Y_{t-1}, \ldots\}$, and knew with certainty population parameters such as μ, ϕ, and θ. This section continues to assume that population parameters are known with certainty, but develops forecasts based on a finite number of observations $\{Y_t, Y_{t-1}, \ldots, Y_{t-m+1}\}$.

For forecasting an $AR(p)$ process, an optimal s-period-ahead linear forecast based on an infinite number of observations $\{Y_t, Y_{t-1}, \ldots\}$ in fact makes use of only the p most recent values $\{Y_t, Y_{t-1}, \ldots, Y_{t-p+1}\}$. For an MA or $ARMA$ process, however, we would in principle require all of the historical values of Y in order to implement the formulas of the preceding section.

Approximations to Optimal Forecasts

One approach to forecasting based on a finite number of observations is to act as if presample ε's were all equal to zero. The idea is thus to use the approximation

$$\begin{aligned}&\hat{E}(Y_{t+s}|Y_t, Y_{t-1}, \ldots)\\ &\quad \cong \hat{E}(Y_{t+s}|Y_t, Y_{t-1}, \ldots, Y_{t-m+1}, \varepsilon_{t-m} = 0, \varepsilon_{t-m-1} = 0, \ldots).\end{aligned} \tag{4.3.1}$$

For example, consider forecasting an $MA(q)$ process. The recursion [4.2.36] can be started by setting

$$\hat{\varepsilon}_{t-m} = \hat{\varepsilon}_{t-m-1} = \cdots = \hat{\varepsilon}_{t-m-q+1} = 0 \tag{4.3.2}$$

and then iterating on [4.2.36] to generate $\hat{\varepsilon}_{t-m+1}, \hat{\varepsilon}_{t-m+2}, \ldots, \hat{\varepsilon}_t$. These calculations produce

$$\begin{aligned}\hat{\varepsilon}_{t-m+1} &= (Y_{t-m+1} - \mu),\\ \hat{\varepsilon}_{t-m+2} &= (Y_{t-m+2} - \mu) - \theta_1\hat{\varepsilon}_{t-m+1},\\ \hat{\varepsilon}_{t-m+3} &= (Y_{t-m+3} - \mu) - \theta_1\hat{\varepsilon}_{t-m+2} - \theta_2\hat{\varepsilon}_{t-m+1},\end{aligned}$$

and so on. The resulting values for $(\hat{\varepsilon}_t, \hat{\varepsilon}_{t-1}, \ldots, \hat{\varepsilon}_{t-q+s})$ are then substituted directly into [4.2.35] to produce the forecast [4.3.1]. For example, for $s = q = 1$, the forecast would be

$$\begin{aligned}\hat{Y}_{t+1|t} = \mu &+ \theta(Y_t - \mu) - \theta^2(Y_{t-1} - \mu)\\ &+ \theta^3(Y_{t-2} - \mu) - \cdots + (-1)^{m-1}\theta^m(Y_{t-m+1} - \mu),\end{aligned} \tag{4.3.3}$$

which is to be used as an approximation to the $AR(\infty)$ forecast,

$$\mu + \theta(Y_t - \mu) - \theta^2(Y_{t-1} - \mu) + \theta^3(Y_{t-2} - \mu) - \cdots. \tag{4.3.4}$$

For m large and $|\theta|$ small, this clearly gives an excellent approximation. For $|\theta|$ closer to unity, the approximation may be poorer. Note that if the moving average operator is noninvertible, the forecast [4.3.1] is inappropriate and should not be used.

Exact Finite-Sample Forecasts

An alternative approach is to calculate the exact projection of Y_{t+1} on its m most recent values. Let

$$\mathbf{X}_t \equiv \begin{bmatrix} 1 \\ Y_t \\ Y_{t-1} \\ \vdots \\ Y_{t-m+1} \end{bmatrix}.$$

We thus seek a linear forecast of the form

$$\boldsymbol{\alpha}^{(m)\prime}\mathbf{X}_t = \alpha_0^{(m)} + \alpha_1^{(m)}Y_t + \alpha_2^{(m)}Y_{t-1} + \cdots + \alpha_m^{(m)}Y_{t-m+1}. \qquad [4.3.5]$$

The coefficient relating Y_{t+1} to Y_t in a projection of Y_{t+1} on the m most recent values of Y is denoted $\alpha_1^{(m)}$ in [4.3.5]. This will in general be different from the coefficient relating Y_{t+1} to Y_t in a projection of Y_{t+1} on the $m + 1$ most recent values of Y; the latter coefficient would be denoted $\alpha_1^{(m+1)}$.

If Y_t is covariance-stationary, then $E(Y_tY_{t-j}) = \gamma_j + \mu^2$. Setting $\mathbf{X}_t = (1, Y_t, Y_{t-1}, \ldots, Y_{t-m+1})'$ in [4.1.13] implies

$$\begin{aligned}\boldsymbol{\alpha}^{(m)\prime} &\equiv [\alpha_0^{(m)} \quad \alpha_1^{(m)} \quad \alpha_2^{(m)} \cdots \alpha_m^{(m)}] \\ &= [\mu \quad (\gamma_1 + \mu^2) \quad (\gamma_2 + \mu^2) \cdots (\gamma_m + \mu^2)] \\ &\quad \times \begin{bmatrix} 1 & \mu & \mu & \cdots & \mu \\ \mu & \gamma_0 + \mu^2 & \gamma_1 + \mu^2 & \cdots & \gamma_{m-1} + \mu^2 \\ \mu & \gamma_1 + \mu^2 & \gamma_0 + \mu^2 & \cdots & \gamma_{m-2} + \mu^2 \\ \vdots & \vdots & \vdots & \cdots & \vdots \\ \mu & \gamma_{m-1} + \mu^2 & \gamma_{m-2} + \mu^2 & \cdots & \gamma_0 + \mu^2 \end{bmatrix}^{-1}. \qquad [4.3.6\end{aligned}$$

When a constant term is included in $\mathbf{X}_t$, it is more convenient to express variables in deviations from the mean. Then we could calculate the projection of $(Y_{t+1} - \mu)$ on $\mathbf{X}_t = [(Y_t - \mu), (Y_{t-1} - \mu), \ldots, (Y_{t-m+1} - \mu)]'$:

$$\begin{aligned}\hat{Y}_{t+1|t} - \mu = \alpha_1^{(m)}(Y_t - \mu) + \alpha_2^{(m)}(Y_{t-1} - \mu) + \cdots \\ + \alpha_m^{(m)}(Y_{t-m+1} - \mu).\end{aligned} \qquad [4.3.7]$$

For this definition of $\mathbf{X}_t$ the coefficients can be calculated directly from [4.1.13] to be

$$\begin{bmatrix} \alpha_1^{(m)} \\ \alpha_2^{(m)} \\ \vdots \\ \alpha_m^{(m)} \end{bmatrix} = \begin{bmatrix} \gamma_0 & \gamma_1 & \cdots & \gamma_{m-1} \\ \gamma_1 & \gamma_0 & \cdots & \gamma_{m-2} \\ \vdots & \vdots & \cdots & \vdots \\ \gamma_{m-1} & \gamma_{m-2} & \cdots & \gamma_0 \end{bmatrix}^{-1} \begin{bmatrix} \gamma_1 \\ \gamma_2 \\ \vdots \\ \gamma_m \end{bmatrix}. \qquad [4.3.8]$$

We will demonstrate in Section 4.5 that the coefficients $(\alpha_1^{(m)}, \alpha_2^{(m)}, \ldots, \alpha_m^{(m)})$ in equations [4.3.8] and [4.3.6] are identical. This is analogous to a familiar result for ordinary least squares regression—slope coefficients would be unchanged if all variables are expressed in deviations from their sample means and the constant term is dropped from the regression.

To generate an s-period-ahead forecast $\hat{Y}_{t+s|t}$, we would use

$$\hat{Y}_{t+s|t} = \mu + \alpha_1^{(m,s)}(Y_t - \mu) + \alpha_2^{(m,s)}(Y_{t-1} - \mu) + \cdots + \alpha_m^{(m,s)}(Y_{t-m+1} - \mu),$$

where

$$\begin{bmatrix} \alpha_1^{(m,s)} \\ \alpha_2^{(m,s)} \\ \vdots \\ \alpha_m^{(m,s)} \end{bmatrix} = \begin{bmatrix} \gamma_0 & \gamma_1 & \cdots & \gamma_{m-1} \\ \gamma_1 & \gamma_0 & \cdots & \gamma_{m-2} \\ \vdots & \vdots & \cdots & \vdots \\ \gamma_{m-1} & \gamma_{m-2} & \cdots & \gamma_0 \end{bmatrix}^{-1} \begin{bmatrix} \gamma_s \\ \gamma_{s+1} \\ \vdots \\ \gamma_{s+m-1} \end{bmatrix}. \qquad [4.3.9]$$

Using expressions such as [4.3.8] requires inverting an $(m \times m)$ matrix. Several algorithms can be used to evaluate [4.3.8] using relatively simple calculations. One approach is based on the Kalman filter discussed in Chapter 13, which can generate exact finite-sample forecasts for a broad class of processes including any *ARMA* specification. A second approach is based on triangular factorization of the matrix in [4.3.8]. This second approach is developed in the next two sections. This approach will prove helpful for the immediate question of calculating finite-sample forecasts and is also a useful device for establishing a number of later results.

4.4. *The Triangular Factorization of a Positive Definite Symmetric Matrix*

Any positive definite symmetric $(n \times n)$ matrix $\mathbf{\Omega}$ has a unique representation of the form

$$\mathbf{\Omega} = \mathbf{ADA}', \qquad [4.4.1]$$

where $\mathbf{A}$ is a lower triangular matrix with 1s along the principal diagonal,

$$\mathbf{A} = \begin{bmatrix} 1 & 0 & 0 & \cdots & 0 \\ a_{21} & 1 & 0 & \cdots & 0 \\ a_{31} & a_{32} & 1 & \cdots & 0 \\ \vdots & \vdots & \vdots & \cdots & \vdots \\ a_{n1} & a_{n2} & a_{n3} & \cdots & 1 \end{bmatrix},$$

and $\mathbf{D}$ is a diagonal matrix,

$$\mathbf{D} = \begin{bmatrix} d_{11} & 0 & 0 & \cdots & 0 \\ 0 & d_{22} & 0 & \cdots & 0 \\ 0 & 0 & d_{33} & \cdots & 0 \\ \vdots & \vdots & \vdots & \cdots & \vdots \\ 0 & 0 & 0 & \cdots & d_{nn} \end{bmatrix},$$

where $d_{ii} > 0$ for all i. This is known as the *triangular factorization* of $\mathbf{\Omega}$.

To see how the triangular factorization can be calculated, consider

$$\mathbf{\Omega} = \begin{bmatrix} \Omega_{11} & \Omega_{12} & \Omega_{13} & \cdots & \Omega_{1n} \\ \Omega_{21} & \Omega_{22} & \Omega_{23} & \cdots & \Omega_{2n} \\ \Omega_{31} & \Omega_{32} & \Omega_{33} & \cdots & \Omega_{3n} \\ \vdots & \vdots & \vdots & \cdots & \vdots \\ \Omega_{n1} & \Omega_{n2} & \Omega_{n3} & \cdots & \Omega_{nn} \end{bmatrix}. \qquad [4.4.2]$$

We assume that $\mathbf{\Omega}$ is positive definite, meaning that $\mathbf{x}'\mathbf{\Omega}\mathbf{x} > 0$ for any nonzero $(n \times 1)$ vector $\mathbf{x}$. We also assume that $\mathbf{\Omega}$ is symmetric, so that $\Omega_{ij} = \Omega_{ji}$.

The matrix $\mathbf{\Omega}$ can be transformed into a matrix with zero in the (2, 1) position by multiplying the first row of $\mathbf{\Omega}$ by $\Omega_{21}\Omega_{11}^{-1}$ and subtracting the resulting row from the second. A zero can be put in the (3, 1) position by multiplying the first row by $\Omega_{31}\Omega_{11}^{-1}$ and subtracting the resulting row from the third. We proceed in this fashion down the first column. This set of operations can be summarized as premultiplying $\mathbf{\Omega}$ by the following matrix:

$$\mathbf{E}_1 = \begin{bmatrix} 1 & 0 & 0 & \cdots & 0 \\ -\Omega_{21}\Omega_{11}^{-1} & 1 & 0 & \cdots & 0 \\ -\Omega_{31}\Omega_{11}^{-1} & 0 & 1 & \cdots & 0 \\ \vdots & \vdots & \vdots & \cdots & \vdots \\ -\Omega_{n1}\Omega_{11}^{-1} & 0 & 0 & \cdots & 1 \end{bmatrix}. \qquad [4.4.3]$$

This matrix always exists, provided that $\Omega_{11} \neq 0$. This is ensured in the present case, because Ω_{11} is equal to $\mathbf{e}_1'\mathbf{\Omega}\mathbf{e}_1$, where $\mathbf{e}_1' = [1 \quad 0 \quad 0 \cdots 0]$. Since $\mathbf{\Omega}$ is positive definite, $\mathbf{e}_1'\mathbf{\Omega}\mathbf{e}_1$ must be greater than zero.

When $\mathbf{\Omega}$ is premultiplied by $\mathbf{E}_1$ and postmultiplied by $\mathbf{E}_1'$ the result is

$$\mathbf{E}_1\mathbf{\Omega}\mathbf{E}_1' = \mathbf{H}, \qquad [4.4.4]$$

where

$$\mathbf{H} = \begin{bmatrix} h_{11} & 0 & 0 & \cdots & 0 \\ 0 & h_{22} & h_{23} & \cdots & h_{2n} \\ 0 & h_{32} & h_{33} & \cdots & h_{3n} \\ \vdots & \vdots & \vdots & \cdots & \vdots \\ 0 & h_{n2} & h_{n3} & \cdots & h_{nn} \end{bmatrix} \qquad [4.4.5]$$

$$= \begin{bmatrix} \Omega_{11} & 0 & 0 & \cdots & 0 \\ 0 & \Omega_{22} - \Omega_{21}\Omega_{11}^{-1}\Omega_{12} & \Omega_{23} - \Omega_{21}\Omega_{11}^{-1}\Omega_{13} & \cdots & \Omega_{2n} - \Omega_{21}\Omega_{11}^{-1}\Omega_{1n} \\ 0 & \Omega_{32} - \Omega_{31}\Omega_{11}^{-1}\Omega_{12} & \Omega_{33} - \Omega_{31}\Omega_{11}^{-1}\Omega_{13} & \cdots & \Omega_{3n} - \Omega_{31}\Omega_{11}^{-1}\Omega_{1n} \\ \vdots & \vdots & \vdots & \cdots & \vdots \\ 0 & \Omega_{n2} - \Omega_{n1}\Omega_{11}^{-1}\Omega_{12} & \Omega_{n3} - \Omega_{n1}\Omega_{11}^{-1}\Omega_{13} & \cdots & \Omega_{nn} - \Omega_{n1}\Omega_{11}^{-1}\Omega_{1n} \end{bmatrix}.$$

We next proceed in exactly the same way with the second column of $\mathbf{H}$. The approach now will be to multiply the second row of $\mathbf{H}$ by $h_{32}h_{22}^{-1}$ and subtract the result from the third row. Similarly, we multiply the second row of $\mathbf{H}$ by $h_{42}h_{22}^{-1}$ and subtract the result from the fourth row, and so on down through the second

column of $\mathbf{H}$. These operations can be represented as premultiplying $\mathbf{H}$ by the following matrix:

$$\mathbf{E}_2 = \begin{bmatrix} 1 & 0 & 0 & \cdots & 0 \\ 0 & 1 & 0 & \cdots & 0 \\ 0 & -h_{32}h_{22}^{-1} & 1 & \cdots & 0 \\ \vdots & \vdots & \vdots & \cdots & \vdots \\ 0 & -h_{n2}h_{22}^{-1} & 0 & \cdots & 1 \end{bmatrix}. \qquad [4.4.6]$$

This matrix always exists provided that $h_{22} \neq 0$. But h_{22} can be calculated as $h_{22} = \mathbf{e}_2'\mathbf{H}\mathbf{e}_2$, where $\mathbf{e}_2' = [0 \;\; 1 \;\; 0 \cdots 0]$. Moreover, $\mathbf{H} = \mathbf{E}_1\mathbf{\Omega}\mathbf{E}_1'$, where $\mathbf{\Omega}$ is positive definite and $\mathbf{E}_1$ is given by [4.4.3]. Since $\mathbf{E}_1$ is lower triangular, its determinant is the product of terms along the principal diagonal, which are all unity. Thus $\mathbf{E}_1$ is nonsingular, meaning that $\mathbf{H} = \mathbf{E}_1\mathbf{\Omega}\mathbf{E}_1'$ is positive definite and so $h_{22} = \mathbf{e}_2'\mathbf{H}\mathbf{e}_2$ must be strictly positive. Thus the matrix in [4.4.6] can always be calculated.

If $\mathbf{H}$ is premultiplied by the matrix in [4.4.6] and postmultiplied by the transpose, the result is

$$\mathbf{E}_2\mathbf{H}\mathbf{E}_2' = \mathbf{K},$$

where

$$\mathbf{K} = \begin{bmatrix} h_{11} & 0 & 0 & \cdots & 0 \\ 0 & h_{22} & 0 & \cdots & 0 \\ 0 & 0 & h_{33} - h_{32}h_{22}^{-1}h_{23} & \cdots & h_{3n} - h_{32}h_{22}^{-1}h_{2n} \\ \vdots & \vdots & \vdots & \cdots & \vdots \\ 0 & 0 & h_{n3} - h_{n2}h_{22}^{-1}h_{23} & \cdots & h_{nn} - h_{n2}h_{22}^{-1}h_{2n} \end{bmatrix}.$$

Again, since $\mathbf{H}$ is positive definite and since $\mathbf{E}_2$ is nonsingular, $\mathbf{K}$ is positive definite and in particular k_{33} is positive. Proceeding through each of the columns with the same approach, we see that for any positive definite symmetric matrix $\mathbf{\Omega}$ there exist matrices $\mathbf{E}_1, \mathbf{E}_2, \ldots, \mathbf{E}_{n-1}$ such that

$$\mathbf{E}_{n-1} \cdots \mathbf{E}_2\mathbf{E}_1\mathbf{\Omega}\mathbf{E}_1'\mathbf{E}_2' \cdots \mathbf{E}_{n-1}' = \mathbf{D}, \qquad [4.4.7]$$

where

$$\mathbf{D} = \begin{bmatrix} \Omega_{11} & 0 & 0 & \cdots & 0 \\ 0 & \Omega_{22} - \Omega_{21}\Omega_{11}^{-1}\Omega_{12} & 0 & \cdots & 0 \\ 0 & 0 & h_{33} - h_{32}h_{22}^{-1}h_{23} & \cdots & 0 \\ \vdots & \vdots & \vdots & \cdots & \vdots \\ 0 & 0 & 0 & \cdots & c_{nn} - c_{n,n-1}c_{n-1,n-1}^{-1}c_{n-1,n} \end{bmatrix},$$

with all the diagonal entries of $\mathbf{D}$ strictly positive. The matrices $\mathbf{E}_1$ and $\mathbf{E}_2$ in [4.4.7] are given by [4.4.3] and [4.4.6]. In general, $\mathbf{E}_j$ is a matrix with nonzero values in the jth column below the principal diagonal, 1s along the principal diagonal, and zeros everywhere else.

Thus each $\mathbf{E}_j$ is lower triangular with unit determinant. Hence $\mathbf{E}_j^{-1}$ exists, and the following matrix exists:

$$\mathbf{A} = (\mathbf{E}_{n-1} \cdots \mathbf{E}_2\mathbf{E}_1)^{-1} = \mathbf{E}_1^{-1}\mathbf{E}_2^{-1} \cdots \mathbf{E}_{n-1}^{-1}. \qquad [4.4.8]$$

If [4.4.7] is premultiplied by $\mathbf{A}$ and postmultiplied by $\mathbf{A}'$, the result is

$$\mathbf{\Omega} = \mathbf{ADA}'. \qquad [4.4.9]$$

Recall that $\mathbf{E}_1$ represents the operation of multiplying the first row of $\mathbf{\Omega}$ by certain numbers and subtracting the results from each of the subsequent rows. Its inverse $\mathbf{E}_1^{-1}$ undoes this operation, which would be achieved by multiplying the first row by these same numbers and *adding* the results to the subsequent rows. Thus

$$\mathbf{E}_1^{-1} = \begin{bmatrix} 1 & 0 & 0 & \cdots & 0 \\ \Omega_{21}\Omega_{11}^{-1} & 1 & 0 & \cdots & 0 \\ \Omega_{31}\Omega_{11}^{-1} & 0 & 1 & \cdots & 0 \\ \vdots & \vdots & \vdots & \cdots & \vdots \\ \Omega_{n1}\Omega_{11}^{-1} & 0 & 0 & \cdots & 1 \end{bmatrix}, \qquad [4.4.10]$$

as may be verified directly by multiplying [4.4.3] by [4.4.10] to obtain the identity matrix. Similarly,

$$\mathbf{E}_2^{-1} = \begin{bmatrix} 1 & 0 & 0 & \cdots & 0 \\ 0 & 1 & 0 & \cdots & 0 \\ 0 & h_{32}h_{22}^{-1} & 1 & \cdots & 0 \\ \vdots & \vdots & \vdots & \cdots & \vdots \\ 0 & h_{n2}h_{22}^{-1} & 0 & \cdots & 1 \end{bmatrix},$$

and so on. Because of this special structure, the series of multiplications in [4.4.8] turns out to be trivial to carry out:

$$\mathbf{A} = \begin{bmatrix} 1 & 0 & 0 & \cdots & 0 \\ \Omega_{21}\Omega_{11}^{-1} & 1 & 0 & \cdots & 0 \\ \Omega_{31}\Omega_{11}^{-1} & h_{32}h_{22}^{-1} & 1 & \cdots & 0 \\ \vdots & \vdots & \vdots & \cdots & \vdots \\ \Omega_{n1}\Omega_{11}^{-1} & h_{n2}h_{22}^{-1} & k_{n3}k_{33}^{-1} & \cdots & 1 \end{bmatrix}. \qquad [4.4.11]$$

That is, the jth column of $\mathbf{A}$ is just the jth column of $\mathbf{E}_j^{-1}$.

We should emphasize that the simplicity of carrying out these matrix multiplications is due not just to the special structure of the $\mathbf{E}_j^{-1}$ matrices but also to the order in which they are multiplied. For example, $\mathbf{A}^{-1} = \mathbf{E}_{n-1}\mathbf{E}_{n-2} \cdots \mathbf{E}_1$ cannot be calculated simply by using the jth column of $\mathbf{E}_j$ for the jth column of $\mathbf{A}^{-1}$.

Since the matrix $\mathbf{A}$ in [4.4.11] is lower triangular with 1s along the principal diagonal, expression [4.4.9] is the triangular factorization of $\mathbf{\Omega}$.

For illustration, the triangular factorization $\mathbf{\Omega} = \mathbf{ADA}'$ of a (2×2) matrix is

$$\begin{bmatrix} \Omega_{11} & \Omega_{12} \\ \Omega_{21} & \Omega_{22} \end{bmatrix} = \begin{bmatrix} 1 & 0 \\ \Omega_{21}\Omega_{11}^{-1} & 1 \end{bmatrix} \times \begin{bmatrix} \Omega_{11} & 0 \\ 0 & \Omega_{22} - \Omega_{21}\Omega_{11}^{-1}\Omega_{12} \end{bmatrix} \begin{bmatrix} 1 & \Omega_{11}^{-1}\Omega_{12} \\ 0 & 1 \end{bmatrix}, \qquad [4.4.12]$$

while that of a (3×3) matrix is

$$\begin{bmatrix} \Omega_{11} & \Omega_{12} & \Omega_{13} \\ \Omega_{21} & \Omega_{22} & \Omega_{23} \\ \Omega_{31} & \Omega_{32} & \Omega_{33} \end{bmatrix} = \begin{bmatrix} 1 & 0 & 0 \\ \Omega_{21}\Omega_{11}^{-1} & 1 & 0 \\ \Omega_{31}\Omega_{11}^{-1} & h_{32}h_{22}^{-1} & 1 \end{bmatrix}$$
$$\times \begin{bmatrix} \Omega_{11} & 0 & 0 \\ 0 & h_{22} & 0 \\ 0 & 0 & h_{33} - h_{32}h_{22}^{-1}h_{23} \end{bmatrix} \begin{bmatrix} 1 & \Omega_{11}^{-1}\Omega_{12} & \Omega_{11}^{-1}\Omega_{13} \\ 0 & 1 & h_{22}^{-1}h_{23} \\ 0 & 0 & 1 \end{bmatrix}, \qquad [4.4.13]$$

where $h_{22} = (\Omega_{22} - \Omega_{21}\Omega_{11}^{-1}\Omega_{12})$, $h_{33} = (\Omega_{33} - \Omega_{31}\Omega_{11}^{-1}\Omega_{13})$, and $h_{23} = h_{32} = (\Omega_{23} - \Omega_{21}\Omega_{11}^{-1}\Omega_{13})$.

Uniqueness of the Triangular Factorization

We next establish that the triangular factorization is unique. Suppose that

$$\mathbf{\Omega} = \mathbf{A}_1\mathbf{D}_1\mathbf{A}_1' = \mathbf{A}_2\mathbf{D}_2\mathbf{A}_2', \qquad [4.4.14]$$

where $\mathbf{A}_1$ and $\mathbf{A}_2$ are both lower triangular with 1s along the principal diagonal and $\mathbf{D}_1$ and $\mathbf{D}_2$ are both diagonal with positive entries along the principal diagonal. Then all the matrices have inverses. Premultiplying [4.4.14] by $\mathbf{D}_1^{-1}\mathbf{A}_1^{-1}$ and postmultiplying by $[\mathbf{A}_2']^{-1}$ yields

$$\mathbf{A}_1'[\mathbf{A}_2']^{-1} = \mathbf{D}_1^{-1}\mathbf{A}_1^{-1}\mathbf{A}_2\mathbf{D}_2. \qquad [4.4.15]$$

Since $\mathbf{A}_2'$ is upper triangular with 1s along the principal diagonal, $[\mathbf{A}_2']^{-1}$ must likewise be upper triangular with 1s along the principal diagonal. Since $\mathbf{A}_1'$ is also of this form, the left side of [4.4.15] is upper triangular with 1s along the principal diagonal. By similar reasoning, the right side of [4.4.15] must be lower triangular. The only way an upper triangular matrix can equal a lower triangular matrix is if all the off-diagonal terms are zero. Moreover, since the diagonal entries on the left side of [4.4.15] are all unity, this matrix must be the identity matrix:

$$\mathbf{A}_1'[\mathbf{A}_2']^{-1} = \mathbf{I}_n.$$

Postmultiplication by $\mathbf{A}_2'$ establishes that $\mathbf{A}_1' = \mathbf{A}_2'$. Premultiplying [4.4.14] by $\mathbf{A}^{-1}$ and postmultiplying by $[\mathbf{A}']^{-1}$ then yields $\mathbf{D}_1 = \mathbf{D}_2$.

The Cholesky Factorization

A closely related factorization of a symmetric positive definite matrix $\mathbf{\Omega}$ is obtained as follows. Define $\mathbf{D}^{1/2}$ to be the $(n \times n)$ diagonal matrix whose diagonal entries are the square roots of the corresponding elements of the matrix $\mathbf{D}$ in the triangular factorization:

$$\mathbf{D}^{1/2} = \begin{bmatrix} \sqrt{d_{11}} & 0 & 0 & \cdots & 0 \\ 0 & \sqrt{d_{22}} & 0 & \cdots & 0 \\ 0 & 0 & \sqrt{d_{33}} & \cdots & 0 \\ \vdots & \vdots & \vdots & \cdots & \vdots \\ 0 & 0 & 0 & \cdots & \sqrt{d_{nn}} \end{bmatrix}.$$

Since the matrix $\mathbf{D}$ is unique and has strictly positive diagonal entries, the matrix $\mathbf{D}^{1/2}$ exists and is unique. Then the triangular factorization can be written

$$\mathbf{\Omega} = \mathbf{A}\mathbf{D}^{1/2}\mathbf{D}^{1/2}\mathbf{A}' = \mathbf{A}\mathbf{D}^{1/2}(\mathbf{A}\mathbf{D}^{1/2})'$$

or

$$\mathbf{\Omega} = \mathbf{PP}', \qquad [4.4.16]$$

where

$$\mathbf{P} \equiv \mathbf{AD}^{1/2}$$

$$= \begin{bmatrix} 1 & 0 & 0 & \cdots & 0 \\ a_{21} & 1 & 0 & \cdots & 0 \\ a_{31} & a_{32} & 1 & \cdots & 0 \\ \vdots & \vdots & \vdots & \cdots & \vdots \\ a_{n1} & a_{n2} & a_{n3} & \cdots & 1 \end{bmatrix} \begin{bmatrix} \sqrt{d_{11}} & 0 & 0 & \cdots & 0 \\ 0 & \sqrt{d_{22}} & 0 & \cdots & 0 \\ 0 & 0 & \sqrt{d_{33}} & \cdots & 0 \\ \vdots & \vdots & \vdots & \cdots & \vdots \\ 0 & 0 & 0 & \cdots & \sqrt{d_{nn}} \end{bmatrix}$$

$$= \begin{bmatrix} \sqrt{d_{11}} & 0 & 0 & \cdots & 0 \\ a_{21}\sqrt{d_{11}} & \sqrt{d_{22}} & 0 & \cdots & 0 \\ a_{31}\sqrt{d_{11}} & a_{32}\sqrt{d_{22}} & \sqrt{d_{33}} & \cdots & 0 \\ \vdots & \vdots & \vdots & \cdots & \vdots \\ a_{n1}\sqrt{d_{11}} & a_{n2}\sqrt{d_{22}} & a_{n3}\sqrt{d_{33}} & \cdots & \sqrt{d_{nn}} \end{bmatrix}.$$

Expression [4.4.16] is known as the *Cholesky factorization* of $\mathbf{\Omega}$. Note that $\mathbf{P}$, like $\mathbf{A}$, is lower triangular, though whereas $\mathbf{A}$ has 1s along the principal diagonal, the Cholesky factor has the square roots of the elements of $\mathbf{D}$ along the principal diagonal.

4.5. *Updating a Linear Projection*

Triangular Factorization of a Second-Moment Matrix and Linear Projection

Let $\mathbf{Y} = (Y_1, Y_2, \ldots, Y_n)'$ be an $(n \times 1)$ vector of random variables whose second-moment matrix is given by

$$\mathbf{\Omega} = E(\mathbf{YY}'). \qquad [4.5.1]$$

Let $\mathbf{\Omega} = \mathbf{ADA}'$ be the triangular factorization of $\mathbf{\Omega}$, and define

$$\tilde{\mathbf{Y}} \equiv \mathbf{A}^{-1}\mathbf{Y}. \qquad [4.5.2]$$

The second-moment matrix of these transformed variables is given by

$$E(\tilde{\mathbf{Y}}\tilde{\mathbf{Y}}') = E(\mathbf{A}^{-1}\mathbf{YY}'[\mathbf{A}']^{-1}) = \mathbf{A}^{-1}E(\mathbf{YY}')[\mathbf{A}']^{-1}. \qquad [4.5.3]$$

Substituting [4.5.1] into [4.5.3], the second-moment matrix of $\tilde{\mathbf{Y}}$ is seen to be diagonal:

$$E(\tilde{\mathbf{Y}}\tilde{\mathbf{Y}}') = \mathbf{A}^{-1}\mathbf{\Omega}[\mathbf{A}']^{-1} = \mathbf{A}^{-1}\mathbf{ADA}'[\mathbf{A}']^{-1} = \mathbf{D}. \qquad [4.5.4]$$

That is,

$$E(\tilde{Y}_i\tilde{Y}_j) = \begin{cases} d_{ii} & \text{for } i = j \\ 0 & \text{for } i \neq j. \end{cases} \qquad [4.5.5]$$

Thus the $\tilde{Y}$'s form a series of random variables that are uncorrelated with one another.[4] To see the implication of this, premultiply [4.5.2] by $\mathbf{A}$:

$$\mathbf{A}\tilde{\mathbf{Y}} = \mathbf{Y}. \qquad [4.5.6]$$

[4]We will use "Y_i and Y_j are uncorrelated" to mean "$E(Y_iY_j) = 0$." The terminology will be correct if Y_i and Y_j have zero means or if a constant term is included in the linear projection.

Expression [4.4.11] can be used to write out [4.5.6] explicitly as

$$\begin{bmatrix} 1 & 0 & 0 & \cdots & 0 \\ \Omega_{21}\Omega_{11}^{-1} & 1 & 0 & \cdots & 0 \\ \Omega_{31}\Omega_{11}^{-1} & h_{32}h_{22}^{-1} & 1 & \cdots & 0 \\ \vdots & \vdots & \vdots & \cdots & \vdots \\ \Omega_{n1}\Omega_{11}^{-1} & h_{n2}h_{22}^{-1} & k_{n3}k_{33}^{-1} & \cdots & 1 \end{bmatrix} \begin{bmatrix} \tilde{Y}_1 \\ \tilde{Y}_2 \\ \tilde{Y}_3 \\ \vdots \\ \tilde{Y}_n \end{bmatrix} = \begin{bmatrix} Y_1 \\ Y_2 \\ Y_3 \\ \vdots \\ Y_n \end{bmatrix}. \qquad [4.5.7]$$

The first equation in [4.5.7] states that

$$\tilde{Y}_1 = Y_1, \qquad [4.5.8]$$

so the first elements of the vectors $\mathbf{Y}$ and $\tilde{\mathbf{Y}}$ represent the same random variable.

The second equation in [4.5.7] asserts that

$$\Omega_{21}\Omega_{11}^{-1}\tilde{Y}_1 + \tilde{Y}_2 = Y_2,$$

or, using [4.5.8],

$$\tilde{Y}_2 = Y_2 - \Omega_{21}\Omega_{11}^{-1}Y_1 \equiv Y_2 - \alpha Y_1, \qquad [4.5.9]$$

where we have defined $\alpha \equiv \Omega_{21}\Omega_{11}^{-1}$. The fact that $\tilde{Y}_2$ is uncorrelated with $\tilde{Y}_1$ implies

$$E(\tilde{Y}_2\tilde{Y}_1) = E[(Y_2 - \alpha Y_1)Y_1] = 0. \qquad [4.5.10]$$

But, recalling [4.1.10], the value of α that satisfies [4.5.10] is defined as the coefficient of the linear projection of Y_2 on Y_1. Thus the triangular factorization of $\mathbf{\Omega}$ can be used to infer that the coefficient of a linear projection of Y_2 on Y_1 is given by $\alpha = \Omega_{21}\Omega_{11}^{-1}$, confirming the earlier result [4.1.13]. In general, the row i, column 1 entry of $\mathbf{A}$ is $\Omega_{i1}\Omega_{11}^{-1}$, which is the coefficient from a linear projection of Y_i on Y_1.

Since $\tilde{Y}_2$ has the interpretation as the residual from a projection of Y_2 on Y_1, from [4.5.5] d_{22} gives the *MSE* of this projection:

$$E(\tilde{Y}_2^2) = d_{22} = \Omega_{22} - \Omega_{21}\Omega_{11}^{-1}\Omega_{12}.$$

This confirms the formula for the *MSE* of a linear projection derived earlier (equation [4.1.15]).

The third equation in [4.5.7] states that

$$\Omega_{31}\Omega_{11}^{-1}\tilde{Y}_1 + h_{32}h_{22}^{-1}\tilde{Y}_2 + \tilde{Y}_3 = Y_3.$$

Substituting in from [4.5.8] and [4.5.9] and rearranging,

$$\tilde{Y}_3 = Y_3 - \Omega_{31}\Omega_{11}^{-1}Y_1 - h_{32}h_{22}^{-1}(Y_2 - \Omega_{21}\Omega_{11}^{-1}Y_1). \qquad [4.5.11]$$

Thus $\tilde{Y}_3$ is the residual from subtracting a particular linear combination of Y_1 and Y_2 from Y_3. From [4.5.5], this residual is uncorrelated with either $\tilde{Y}_1$ or $\tilde{Y}_2$:

$$E[Y_3 - \Omega_{31}\Omega_{11}^{-1}Y_1 - h_{32}h_{22}^{-1}(Y_2 - \Omega_{21}\Omega_{11}^{-1}Y_1)]\tilde{Y}_j = 0 \qquad \text{for } j = 1 \text{ or } 2.$$

Thus this residual is uncorrelated with either Y_1 or Y_2, meaning that $\tilde{Y}_3$ has the interpretation as the residual from a linear projection of Y_3 on Y_1 and Y_2. According to [4.5.11], the linear projection is given by

$$\hat{P}(Y_3|Y_2,Y_1) = \Omega_{31}\Omega_{11}^{-1}Y_1 + h_{32}h_{22}^{-1}(Y_2 - \Omega_{21}\Omega_{11}^{-1}Y_1). \qquad [4.5.12]$$

The *MSE* of the linear projection is the variance of $\tilde{Y}_3$, which from [4.5.5] is given by d_{33}:

$$E[Y_3 - \hat{P}(Y_3|Y_2,Y_1)]^2 = h_{33} - h_{32}h_{22}^{-1}h_{23}. \qquad [4.5.13]$$

Expression [4.5.12] gives a convenient formula for updating a linear projection. Suppose we are interested in forecasting the value of Y_3. Let Y_1 be some initial information on which this forecast might be based. A forecast of Y_3 on the basis of Y_1 alone takes the form

$$\hat{P}(Y_3|Y_1) = \Omega_{31}\Omega_{11}^{-1}Y_1.$$

Let Y_2 represent some new information with which we could update this forecast. If we were asked to guess the magnitude of this second variable on the basis of Y_1 alone, the answer would be

$$\hat{P}(Y_2|Y_1) = \Omega_{21}\Omega_{11}^{-1}Y_1.$$

Equation [4.5.12] states that

$$\hat{P}(Y_3|Y_2,Y_1) = \hat{P}(Y_3|Y_1) + h_{32}h_{22}^{-1}[Y_2 - \hat{P}(Y_2|Y_1)]. \qquad [4.5.14]$$

We can thus optimally update the initial forecast $\hat{P}(Y_3|Y_1)$ by adding to it a multiple $(h_{32}h_{22}^{-1})$ of the unanticipated component of the new information $[Y_2 - \hat{P}(Y_2|Y_1)]$. This multiple $(h_{32}h_{22}^{-1})$ can also be interpreted as the coefficient on Y_2 in a linear projection of Y_3 on Y_2 and Y_1.

To understand the nature of the multiplier $(h_{32}h_{22}^{-1})$, define the $(n \times 1)$ vector $\tilde{\mathbf{Y}}(1)$ by

$$\tilde{\mathbf{Y}}(1) \equiv \mathbf{E}_1\mathbf{Y}, \qquad [4.5.15]$$

where $\mathbf{E}_1$ is the matrix given in [4.4.3]. Notice that the second-moment matrix of $\tilde{\mathbf{Y}}(1)$ is given by

$$E\{\tilde{\mathbf{Y}}(1)[\tilde{\mathbf{Y}}(1)]'\} = E\{\mathbf{E}_1\mathbf{Y}\mathbf{Y}'\mathbf{E}_1'\} = \mathbf{E}_1\mathbf{\Omega}\mathbf{E}_1'.$$

But from [4.4.4] this is just the matrix $\mathbf{H}$. Thus $\mathbf{H}$ has the interpretation as the second-moment matrix of $\tilde{\mathbf{Y}}(1)$. Substituting [4.4.3] into [4.5.15],

$$\tilde{\mathbf{Y}}(1) = \begin{bmatrix} Y_1 \\ Y_2 - \Omega_{21}\Omega_{11}^{-1}Y_1 \\ Y_3 - \Omega_{31}\Omega_{11}^{-1}Y_1 \\ \vdots \\ Y_n - \Omega_{n1}\Omega_{11}^{-1}Y_1 \end{bmatrix}.$$

The first element of $\tilde{\mathbf{Y}}(1)$ is thus just Y_1 itself, while the ith element of $\tilde{\mathbf{Y}}(1)$ for $i = 2, 3, \ldots, n$ is the residual from a projection of Y_i on Y_1. The matrix $\mathbf{H}$ is thus the second-moment matrix of the residuals from projections of each of the variables on Y_1. In particular, h_{22} is the *MSE* from a projection of Y_2 on Y_1:

$$h_{22} = E[Y_2 - \hat{P}(Y_2|Y_1)]^2,$$

while h_{32} is the expected product of this error with the error from a projection of Y_3 on Y_1:

$$h_{32} = E\{[Y_3 - \hat{P}(Y_3|Y_1)][Y_2 - \hat{P}(Y_2|Y_1)]\}.$$

Thus equation [4.5.14] states that a linear projection can be updated using the following formula:

$$\begin{aligned}\hat{P}(Y_3|Y_2,Y_1) = {} & \hat{P}(Y_3|Y_1) \\ & + \{E[Y_3 - \hat{P}(Y_3|Y_1)][Y_2 - \hat{P}(Y_2|Y_1)]\} \\ & \times \{E[Y_2 - \hat{P}(Y_2|Y_1)]^2\}^{-1} \times [Y_2 - \hat{P}(Y_2|Y_1)].\end{aligned} \qquad [4.5.16]$$

For example, suppose that Y_1 is a constant term, so that $\hat{P}(Y_2|Y_1)$ is just μ_2, the mean of Y_2, while $\hat{P}(Y_3|Y_1) = \mu_3$. Equation [4.5.16] then states that

$$\hat{P}(Y_3|Y_2,1) = \mu_3 + \text{Cov}(Y_3, Y_2)\cdot[\text{Var}(Y_2)]^{-1}\cdot(Y_2 - \mu_2).$$

The *MSE* associated with this updated linear projection can also be calculated from the triangular factorization. From [4.5.5], the *MSE* from a linear projection of Y_3 on Y_2 and Y_1 can be calculated from

$$\begin{aligned} E[Y_3 - \hat{P}(Y_3|Y_2,Y_1)]^2 &= E(\tilde{Y}_3^2) \\ &= d_{33} \\ &= h_{33} - h_{32}h_{22}^{-1}h_{23}. \end{aligned}$$

In general, for $i > 2$, the coefficient on Y_2 in a linear projection of Y_i on Y_2 and Y_1 is given by the ith element of the second column of the matrix **A**. For any $i > j$, the coefficients on Y_j in a linear projection of Y_i on $Y_j, Y_{j-1}, \ldots, Y_1$ is given by the row i, column j element of **A**. The magnitude d_{ii} gives the *MSE* for a linear projection of Y_i on $Y_{i-1}, Y_{i-2}, \ldots, Y_1$.

Application: Exact Finite-Sample Forecasts for an MA*(1) Process*

As an example of applying these results, suppose that Y_t follows an $MA(1)$ process:

$$Y_t = \mu + \varepsilon_t + \theta\varepsilon_{t-1},$$

where ε_t is a white noise process with variance σ^2 and θ is unrestricted. Suppose we want to forecast the value of Y_n on the basis of the previous $n - 1$ values $(Y_1, Y_2, \ldots, Y_{n-1})$. Let

$$\mathbf{Y}' \equiv [(Y_1 - \mu) \quad (Y_2 - \mu) \quad \cdots \quad (Y_{n-1} - \mu) \quad (Y_n - \mu)],$$

and let $\mathbf{\Omega}$ denote the $(n \times n)$ variance-covariance matrix of **Y**:

$$\mathbf{\Omega} = E(\mathbf{Y}\mathbf{Y}') = \sigma^2\begin{bmatrix} 1+\theta^2 & \theta & 0 & \cdots & 0 \\ \theta & 1+\theta^2 & \theta & \cdots & 0 \\ 0 & \theta & 1+\theta^2 & \cdots & 0 \\ \vdots & \vdots & \vdots & \cdots & \vdots \\ 0 & 0 & 0 & \cdots & 1+\theta^2 \end{bmatrix}. \qquad [4.5.17]$$

Appendix 4.B to this chapter shows that the triangular factorization of $\mathbf{\Omega}$ is

$$\mathbf{A} = \qquad [4.5.18]$$

$$\begin{bmatrix} 1 & 0 & 0 & \cdots & 0 & 0 \\ \dfrac{\theta}{1+\theta^2} & 1 & 0 & \cdots & 0 & 0 \\ 0 & \dfrac{\theta(1+\theta^2)}{1+\theta^2+\theta^4} & 1 & \cdots & 0 & 0 \\ \vdots & \vdots & \vdots & \cdots & \vdots & \vdots \\ 0 & 0 & 0 & \cdots & \dfrac{\theta[1+\theta^2+\theta^4+\cdots+\theta^{2(n-2)}]}{1+\theta^2+\theta^4+\cdots+\theta^{2(n-1)}} & 1 \end{bmatrix}$$

$$\mathbf{D} = \sigma^2 \begin{bmatrix} 1+\theta^2 & 0 & 0 & \cdots & 0 \\ 0 & \dfrac{1+\theta^2+\theta^4}{1+\theta^2} & 0 & \cdots & 0 \\ 0 & 0 & \dfrac{1+\theta^2+\theta^4+\theta^6}{1+\theta^2+\theta^4} & \cdots & 0 \\ \vdots & \vdots & \vdots & \cdots & \vdots \\ 0 & 0 & 0 & \cdots & \dfrac{1+\theta^2+\theta^4+\cdots+\theta^{2n}}{1+\theta^2+\theta^4+\cdots+\theta^{2(n-1)}} \end{bmatrix}. \qquad [4.5.19]$$

To use the triangular factorization to calculate exact finite-sample forecasts, recall that $\tilde{Y}_i$, the ith element of $\tilde{\mathbf{Y}} = \mathbf{A}^{-1}\mathbf{Y}$, has the interpretation as the residual from a linear projection of Y_i on a constant and its previous values:

$$\tilde{Y}_i = Y_i - \hat{E}(Y_i|Y_{i-1}, Y_{i-2}, \ldots, Y_1).$$

The system of equations $\mathbf{A}\tilde{\mathbf{Y}} = \mathbf{Y}$ can be written out explicitly as

$$\begin{aligned} \tilde{Y}_1 &= Y_1 - \mu \\ \frac{\theta}{1+\theta^2}\tilde{Y}_1 + \tilde{Y}_2 &= Y_2 - \mu \\ \frac{\theta(1+\theta^2)}{1+\theta^2+\theta^4}\tilde{Y}_2 + \tilde{Y}_3 &= Y_3 - \mu \\ &\vdots \\ \frac{\theta[1+\theta^2+\theta^4+\cdots+\theta^{2(n-2)}]}{1+\theta^2+\theta^4+\cdots+\theta^{2(n-1)}}\tilde{Y}_{n-1} + \tilde{Y}_n &= Y_n - \mu. \end{aligned}$$

Solving the last equation for $\tilde{Y}_n$,

$$\begin{aligned} Y_n - \hat{E}(Y_n|Y_{n-1}, Y_{n-2}, \ldots, Y_1) &= Y_n - \mu \\ &- \frac{\theta[1+\theta^2+\theta^4+\cdots+\theta^{2(n-2)}]}{1+\theta^2+\theta^4+\cdots+\theta^{2(n-1)}}[Y_{n-1} - \hat{E}(Y_{n-1}|Y_{n-2}, Y_{n-3}, \ldots, Y_1)], \end{aligned}$$

implying

$$\begin{aligned} \hat{E}(Y_n|Y_{n-1}, Y_{n-2}, \ldots, Y_1) &= \mu \\ &+ \frac{\theta[1+\theta^2+\theta^4+\cdots+\theta^{2(n-2)}]}{1+\theta^2+\theta^4+\cdots+\theta^{2(n-1)}}[Y_{n-1} - \hat{E}(Y_{n-1}|Y_{n-2}, Y_{n-3}, \ldots, Y_1)]. \end{aligned} \qquad [4.5.20]$$

The *MSE* of this forecast is given by d_{nn}:

$$MSE[\hat{E}(Y_n|Y_{n-1}, Y_{n-2}, \ldots, Y_1)] = \sigma^2\frac{1+\theta^2+\theta^4+\cdots+\theta^{2n}}{1+\theta^2+\theta^4+\cdots+\theta^{2(n-1)}}. \qquad [4.5.21]$$

It is interesting to note the behavior of this optimal forecast as the number of observations (n) becomes large. First, suppose that the moving average representation is invertible ($|\theta| < 1$). In this case, as $n \to \infty$, the coefficient in [4.5.20] tends to θ:

$$\frac{\theta[1+\theta^2+\theta^4+\cdots+\theta^{2(n-2)}]}{1+\theta^2+\theta^4+\cdots+\theta^{2(n-1)}} \to \theta,$$

while the *MSE* [4.5.21] tends to σ^2, the variance of the fundamental innovation. Thus the optimal forecast for a finite number of observations [4.5.20] eventually tends toward the forecast rule used for an infinite number of observations [4.2.32].

Alternatively, the calculations that produced [4.5.20] are equally valid for a noninvertible representation with $|\theta| > 1$. In this case the coefficient in [4.5.20] tends toward θ^{-1}:

$$\begin{aligned}\frac{\theta[1 + \theta^2 + \theta^4 + \cdots + \theta^{2(n-2)}]}{1 + \theta^2 + \theta^4 + \cdots + \theta^{2(n-1)}} &= \frac{\theta[1 - \theta^{2(n-1)}]/(1 - \theta^2)}{(1 - \theta^{2n})/(1 - \theta^2)} \\ &= \frac{\theta(\theta^{-2n} - \theta^{-2})}{\theta^{-2n} - 1} \\ &\to \frac{\theta(-\theta^{-2})}{-1} \\ &= \theta^{-1}.\end{aligned}$$

Thus, the coefficient in [4.5.20] tends to θ^{-1} in this case, which is the moving average coefficient associated with the invertible representation. The *MSE* [4.5.21] tends to $\sigma^2\theta^2$:

$$\sigma^2 \frac{[1 - \theta^{2(n+1)}]/(1 - \theta^2)}{(1 - \theta^{2n})/(1 - \theta^2)} \to \sigma^2\theta^2,$$

which will be recognized from [3.7.7] as the variance of the innovation associated with the fundamental representation.

This observation explains the use of the expression "fundamental" in this context. The fundamental innovation ε_t has the property that

$$Y_t - \hat{E}(Y_t|Y_{t-1}, Y_{t-2}, \ldots, Y_{t-m}) \xrightarrow{m.s.} \varepsilon_t \qquad [4.5.22]$$

as $m \to \infty$ where $\xrightarrow{m.s.}$ denotes mean square convergence. Thus when $|\theta| > 1$, the coefficient θ in the approximation in [4.3.3] should be replaced by θ^{-1}. When this is done, expression [4.3.3] will approach the correct forecast as $m \to \infty$.

It is also instructive to consider the borderline case $\theta = 1$. The optimal finite-sample forecast for an $MA(1)$ process with $\theta = 1$ is seen from [4.5.20] to be given by

$$\hat{E}(Y_n|Y_{n-1}, Y_{n-2}, \ldots, Y_1) = \mu + \frac{n-1}{n}[Y_{n-1} - \hat{E}(Y_{n-1}|Y_{n-2}, Y_{n-3}, \ldots, Y_1)],$$

which, after recursive substitution, becomes

$$\begin{aligned}&\hat{E}(Y_n|Y_{n-1}, Y_{n-2}, \ldots, Y_1) \\ &= \mu + \frac{n-1}{n}(Y_{n-1} - \mu) - \frac{n-2}{n}(Y_{n-2} - \mu) \qquad [4.5.23] \\ &\quad + \frac{n-3}{n}(Y_{n-3} - \mu) - \cdots + (-1)^n \frac{1}{n}(Y_1 - \mu).\end{aligned}$$

The *MSE* of this forecast is given by [4.5.21]:

$$\sigma^2(n + 1)/n \to \sigma^2.$$

Thus the variance of the forecast error again tends toward that of ε_t. Hence the innovation ε_t is again fundamental for this case in the sense of [4.5.22]. Note the contrast between the optimal forecast [4.5.23] and a forecast based on a naive application of [4.3.3],

$$\begin{aligned}&\mu + (Y_{n-1} - \mu) - (Y_{n-2} - \mu) + (Y_{n-3} - \mu) \\ &\qquad - \cdots + (-1)^n(Y_1 - \mu).\end{aligned} \qquad [4.5.24]$$

The approximation [4.3.3] was derived under the assumption that the moving average representation was invertible, and the borderline case $\theta = 1$ is not invertible. For this

reason [4.5.24] does not converge to the optimal forecast [4.5.23] as n grows large. When $\theta = 1$, $Y_t = \mu + \varepsilon_t + \varepsilon_{t-1}$ and [4.5.24] can be written as

$$\mu + (\varepsilon_{n-1} + \varepsilon_{n-2}) - (\varepsilon_{n-2} + \varepsilon_{n-3}) + (\varepsilon_{n-3} + \varepsilon_{n-4}) - \cdots + (-1)^n(\varepsilon_1 + \varepsilon_0) = \mu + \varepsilon_{n-1} + (-1)^n\varepsilon_0.$$

The difference between this and Y_n, the value being forecast, is $\varepsilon_n - (-1)^n\varepsilon_0$, which has MSE $2\sigma^2$ for all n. Thus, whereas [4.5.23] converges to the optimal forecast as $n \to \infty$, [4.5.24] does not.

Block Triangular Factorization

Suppose we have observations on two sets of variables. The first set of variables is collected in an $(n_1 \times 1)$ vector $\mathbf{Y}_1$ and the second set in an $(n_2 \times 1)$ vector $\mathbf{Y}_2$. Their second-moment matrix can be written in partitioned form as

$$\mathbf{\Omega} \equiv \begin{bmatrix} E(\mathbf{Y}_1\mathbf{Y}_1') & E(\mathbf{Y}_1\mathbf{Y}_2') \\ E(\mathbf{Y}_2\mathbf{Y}_1') & E(\mathbf{Y}_2\mathbf{Y}_2') \end{bmatrix} = \begin{bmatrix} \mathbf{\Omega}_{11} & \mathbf{\Omega}_{12} \\ \mathbf{\Omega}_{21} & \mathbf{\Omega}_{22} \end{bmatrix},$$

where $\mathbf{\Omega}_{11}$ is an $(n_1 \times n_1)$ matrix, $\mathbf{\Omega}_{22}$ is an $(n_2 \times n_2)$ matrix, and the $(n_1 \times n_2)$ matrix $\mathbf{\Omega}_{12}$ is the transpose of the $(n_2 \times n_1)$ matrix $\mathbf{\Omega}_{21}$.

We can put zeros in the lower left $(n_2 \times n_1)$ block of $\mathbf{\Omega}$ by premultiplying $\mathbf{\Omega}$ by the following matrix:

$$\overline{\mathbf{E}}_1 = \begin{bmatrix} \mathbf{I}_{n_1} & \mathbf{0} \\ -\mathbf{\Omega}_{21}\mathbf{\Omega}_{11}^{-1} & \mathbf{I}_{n_2} \end{bmatrix}.$$

If $\mathbf{\Omega}$ is premultiplied by $\overline{\mathbf{E}}_1$ and postmultiplied by $\overline{\mathbf{E}}_1'$, the result is

$$\begin{bmatrix} \mathbf{I}_{n_1} & \mathbf{0} \\ -\mathbf{\Omega}_{21}\mathbf{\Omega}_{11}^{-1} & \mathbf{I}_{n_2} \end{bmatrix} \begin{bmatrix} \mathbf{\Omega}_{11} & \mathbf{\Omega}_{12} \\ \mathbf{\Omega}_{21} & \mathbf{\Omega}_{22} \end{bmatrix} \begin{bmatrix} \mathbf{I}_{n_1} & -\mathbf{\Omega}_{11}^{-1}\mathbf{\Omega}_{12} \\ \mathbf{0} & \mathbf{I}_{n_2} \end{bmatrix} = \begin{bmatrix} \mathbf{\Omega}_{11} & \mathbf{0} \\ \mathbf{0} & \mathbf{\Omega}_{22} - \mathbf{\Omega}_{21}\mathbf{\Omega}_{11}^{-1}\mathbf{\Omega}_{12} \end{bmatrix}. \qquad [4.5.25]$$

Define

$$\overline{\mathbf{A}} \equiv \overline{\mathbf{E}}_1^{-1} = \begin{bmatrix} \mathbf{I}_{n_1} & \mathbf{0} \\ \mathbf{\Omega}_{21}\mathbf{\Omega}_{11}^{-1} & \mathbf{I}_{n_2} \end{bmatrix}.$$

If [4.5.25] is premultiplied by $\overline{\mathbf{A}}$ and postmultiplied by $\overline{\mathbf{A}}'$, the result is

$$\begin{aligned} \begin{bmatrix} \mathbf{\Omega}_{11} & \mathbf{\Omega}_{12} \\ \mathbf{\Omega}_{21} & \mathbf{\Omega}_{22} \end{bmatrix} &= \begin{bmatrix} \mathbf{I}_{n_1} & \mathbf{0} \\ \mathbf{\Omega}_{21}\mathbf{\Omega}_{11}^{-1} & \mathbf{I}_{n_2} \end{bmatrix} \\ &\quad \times \begin{bmatrix} \mathbf{\Omega}_{11} & \mathbf{0} \\ \mathbf{0} & \mathbf{\Omega}_{22} - \mathbf{\Omega}_{21}\mathbf{\Omega}_{11}^{-1}\mathbf{\Omega}_{12} \end{bmatrix} \begin{bmatrix} \mathbf{I}_{n_1} & \mathbf{\Omega}_{11}^{-1}\mathbf{\Omega}_{12} \\ \mathbf{0} & \mathbf{I}_{n_2} \end{bmatrix} \qquad [4.5.26] \\ &= \overline{\mathbf{A}}\,\overline{\mathbf{D}}\,\overline{\mathbf{A}}'. \end{aligned}$$

This is similar to the triangular factorization $\mathbf{\Omega} = \mathbf{ADA}'$, except that $\overline{\mathbf{D}}$ is a block-diagonal matrix rather than a truly diagonal matrix:

$$\overline{\mathbf{D}} = \begin{bmatrix} \mathbf{\Omega}_{11} & \mathbf{0} \\ \mathbf{0} & \mathbf{\Omega}_{22} - \mathbf{\Omega}_{21}\mathbf{\Omega}_{11}^{-1}\mathbf{\Omega}_{12} \end{bmatrix}.$$

As in the earlier case, $\bar{\mathbf{D}}$ can be interpreted as the second-moment matrix of the vector $\tilde{\mathbf{Y}} = \bar{\mathbf{A}}^{-1}\mathbf{Y}$,

$$\begin{bmatrix} \tilde{\mathbf{Y}}_1 \\ \tilde{\mathbf{Y}}_2 \end{bmatrix} = \begin{bmatrix} \mathbf{I}_{n_1} & \mathbf{0} \\ -\mathbf{\Omega}_{21}\mathbf{\Omega}_{11}^{-1} & \mathbf{I}_{n_2} \end{bmatrix} \begin{bmatrix} \mathbf{Y}_1 \\ \mathbf{Y}_2 \end{bmatrix};$$

that is, $\tilde{\mathbf{Y}}_1 = \mathbf{Y}_1$ and $\tilde{\mathbf{Y}}_2 = \mathbf{Y}_2 - \mathbf{\Omega}_{21}\mathbf{\Omega}_{11}^{-1}\mathbf{Y}_1$. The ith element of $\tilde{\mathbf{Y}}_2$ is given by Y_{2i} minus a linear combination of the elements of $\mathbf{Y}_1$. The block-diagonality of $\bar{\mathbf{D}}$ implies that the product of any element of $\tilde{\mathbf{Y}}_2$ with any element of $\mathbf{Y}_1$ has expectation zero. Thus $\mathbf{\Omega}_{21}\mathbf{\Omega}_{11}^{-1}$ gives the matrix of coefficients associated with the linear projection of the vector $\mathbf{Y}_2$ on the vector $\mathbf{Y}_1$,

$$\hat{P}(\mathbf{Y}_2|\mathbf{Y}_1) = \mathbf{\Omega}_{21}\mathbf{\Omega}_{11}^{-1}\mathbf{Y}_1, \qquad [4.5.27]$$

as claimed in [4.1.23]. The *MSE* matrix associated with this linear projection is

$$\begin{aligned} E\{[\mathbf{Y}_2 - \hat{P}(\mathbf{Y}_2|\mathbf{Y}_1)][\mathbf{Y}_2 - \hat{P}(\mathbf{Y}_2|\mathbf{Y}_1)]'\} &= E(\tilde{\mathbf{Y}}_2\tilde{\mathbf{Y}}_2') \\ &= \bar{\mathbf{D}}_{22} \qquad [4.5.28] \\ &= \mathbf{\Omega}_{22} - \mathbf{\Omega}_{21}\mathbf{\Omega}_{11}^{-1}\mathbf{\Omega}_{12}, \end{aligned}$$

as claimed in [4.1.24].

The calculations for a (3×3) matrix similarly extend to a (3×3) block matrix without complications. Let $\mathbf{Y}_1$, $\mathbf{Y}_2$, and $\mathbf{Y}_3$ be $(n_1 \times 1)$, $(n_2 \times 1)$, and $(n_3 \times 1)$ vectors. A block-triangular factorization of their second-moment matrix is obtained from a simple generalization of equation [4.4.13]:

$$\begin{bmatrix} \mathbf{\Omega}_{11} & \mathbf{\Omega}_{12} & \mathbf{\Omega}_{13} \\ \mathbf{\Omega}_{21} & \mathbf{\Omega}_{22} & \mathbf{\Omega}_{23} \\ \mathbf{\Omega}_{31} & \mathbf{\Omega}_{32} & \mathbf{\Omega}_{33} \end{bmatrix} = \begin{bmatrix} \mathbf{I}_{n_1} & \mathbf{0} & \mathbf{0} \\ \mathbf{\Omega}_{21}\mathbf{\Omega}_{11}^{-1} & \mathbf{I}_{n_2} & \mathbf{0} \\ \mathbf{\Omega}_{31}\mathbf{\Omega}_{11}^{-1} & \mathbf{H}_{32}\mathbf{H}_{22}^{-1} & \mathbf{I}_{n_3} \end{bmatrix}$$

$$\times \begin{bmatrix} \mathbf{\Omega}_{11} & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{H}_{22} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \mathbf{H}_{33} - \mathbf{H}_{32}\mathbf{H}_{22}^{-1}\mathbf{H}_{23} \end{bmatrix} \begin{bmatrix} \mathbf{I}_{n_1} & \mathbf{\Omega}_{11}^{-1}\mathbf{\Omega}_{12} & \mathbf{\Omega}_{11}^{-1}\mathbf{\Omega}_{13} \\ \mathbf{0} & \mathbf{I}_{n_2} & \mathbf{H}_{22}^{-1}\mathbf{H}_{23} \\ \mathbf{0} & \mathbf{0} & \mathbf{I}_{n_3} \end{bmatrix} \qquad [4.5.29]$$

where $\mathbf{H}_{22} = (\mathbf{\Omega}_{22} - \mathbf{\Omega}_{21}\mathbf{\Omega}_{11}^{-1}\mathbf{\Omega}_{12})$, $\mathbf{H}_{33} = (\mathbf{\Omega}_{33} - \mathbf{\Omega}_{31}\mathbf{\Omega}_{11}^{-1}\mathbf{\Omega}_{13})$, and $\mathbf{H}_{23} = \mathbf{H}_{32}' = (\mathbf{\Omega}_{23} - \mathbf{\Omega}_{21}\mathbf{\Omega}_{11}^{-1}\mathbf{\Omega}_{13})$.

This allows us to generalize the earlier result [4.5.12] on updating a linear projection. The optimal forecast of $\mathbf{Y}_3$ conditional on $\mathbf{Y}_2$ and $\mathbf{Y}_1$ can be read off the last block row of $\bar{\mathbf{A}}$:

$$\begin{aligned} \hat{P}(\mathbf{Y}_3|\mathbf{Y}_2,\mathbf{Y}_1) &= \mathbf{\Omega}_{31}\mathbf{\Omega}_{11}^{-1}\mathbf{Y}_1 + \mathbf{H}_{32}\mathbf{H}_{22}^{-1}(\mathbf{Y}_2 - \mathbf{\Omega}_{21}\mathbf{\Omega}_{11}^{-1}\mathbf{Y}_1) \\ &= \hat{P}(\mathbf{Y}_3|\mathbf{Y}_1) + \mathbf{H}_{32}\mathbf{H}_{22}^{-1}[\mathbf{Y}_2 - \hat{P}(\mathbf{Y}_2|\mathbf{Y}_1)], \end{aligned} \qquad [4.5.30]$$

where

$$\mathbf{H}_{22} = E\{[\mathbf{Y}_2 - \hat{P}(\mathbf{Y}_2|\mathbf{Y}_1)][\mathbf{Y}_2 - \hat{P}(\mathbf{Y}_2|\mathbf{Y}_1)]'\}$$

$$\mathbf{H}_{32} = E\{[\mathbf{Y}_3 - \hat{P}(\mathbf{Y}_3|\mathbf{Y}_1)][\mathbf{Y}_2 - \hat{P}(\mathbf{Y}_2|\mathbf{Y}_1)]'\}.$$

The *MSE* of this forecast is the matrix generalization of [4.5.13],

$$E\{[\mathbf{Y}_3 - \hat{P}(\mathbf{Y}_3|\mathbf{Y}_2,\mathbf{Y}_1)][\mathbf{Y}_3 - \hat{P}(\mathbf{Y}_3|\mathbf{Y}_2,\mathbf{Y}_1)]'\} = \mathbf{H}_{33} - \mathbf{H}_{32}\mathbf{H}_{22}^{-1}\mathbf{H}_{23}, \qquad [4.5.31]$$

where

$$\mathbf{H}_{33} = E\{[\mathbf{Y}_3 - \hat{P}(\mathbf{Y}_3|\mathbf{Y}_1)][\mathbf{Y}_3 - \hat{P}(\mathbf{Y}_3|\mathbf{Y}_1)]'\}.$$

Law of Iterated Projections

Another useful result, the law of iterated projections, can be inferred immediately from [4.5.30]. What happens if the projection $\hat{P}(\mathbf{Y}_3|\mathbf{Y}_2,\mathbf{Y}_1)$ is itself projected on $\mathbf{Y}_1$? The *law of iterated projections* says that this projection is equal to the simple projection of $\mathbf{Y}_3$ on $\mathbf{Y}_1$:

$$\hat{P}[\hat{P}(\mathbf{Y}_3|\mathbf{Y}_2,\mathbf{Y}_1)|\mathbf{Y}_1] = \hat{P}(\mathbf{Y}_3|\mathbf{Y}_1). \qquad [4.5.32]$$

To verify this claim, we need to show that the difference between $\hat{P}(\mathbf{Y}_3|\mathbf{Y}_2,\mathbf{Y}_1)$ and $\hat{P}(\mathbf{Y}_3|\mathbf{Y}_1)$ is uncorrelated with $\mathbf{Y}_1$. But from [4.5.30], this difference is given by

$$\hat{P}(\mathbf{Y}_3|\mathbf{Y}_2,\mathbf{Y}_1) - \hat{P}(\mathbf{Y}_3|\mathbf{Y}_1) = \mathbf{H}_{32}\mathbf{H}_{22}^{-1}[\mathbf{Y}_2 - \hat{P}(\mathbf{Y}_2|\mathbf{Y}_1)],$$

which indeed is uncorrelated with $\mathbf{Y}_1$ by the definition of the linear projection $\hat{P}(\mathbf{Y}_2|\mathbf{Y}_1)$.

4.6. *Optimal Forecasts for Gaussian Processes*

The forecasting rules developed in this chapter are optimal within the class of linear functions of the variables on which the forecast is based. For Gaussian processes, we can make the stronger claim that as long as a constant term is included among the variables on which the forecast is based, the optimal unrestricted forecast turns out to have a linear form and thus is given by the linear projection.

To verify this, let $\mathbf{Y}_1$ be an $(n_1 \times 1)$ vector with mean $\boldsymbol{\mu}_1$, and $\mathbf{Y}_2$ an $(n_2 \times 1)$ vector with mean $\boldsymbol{\mu}_2$, where the variance-covariance matrix is given by

$$\begin{bmatrix} E(\mathbf{Y}_1 - \boldsymbol{\mu}_1)(\mathbf{Y}_1 - \boldsymbol{\mu}_1)' & E(\mathbf{Y}_1 - \boldsymbol{\mu}_1)(\mathbf{Y}_2 - \boldsymbol{\mu}_2)' \\ E(\mathbf{Y}_2 - \boldsymbol{\mu}_2)(\mathbf{Y}_1 - \boldsymbol{\mu}_1)' & E(\mathbf{Y}_2 - \boldsymbol{\mu}_2)(\mathbf{Y}_2 - \boldsymbol{\mu}_2)' \end{bmatrix} = \begin{bmatrix} \boldsymbol{\Omega}_{11} & \boldsymbol{\Omega}_{12} \\ \boldsymbol{\Omega}_{21} & \boldsymbol{\Omega}_{22} \end{bmatrix}.$$

If $\mathbf{Y}_1$ and $\mathbf{Y}_2$ are Gaussian, then the joint probability density is

$$\begin{aligned} f_{\mathbf{Y}_1,\mathbf{Y}_2}(\mathbf{y}_1, \mathbf{y}_2) &= \frac{1}{(2\pi)^{(n_1+n_2)/2}} \begin{vmatrix} \boldsymbol{\Omega}_{11} & \boldsymbol{\Omega}_{12} \\ \boldsymbol{\Omega}_{21} & \boldsymbol{\Omega}_{22} \end{vmatrix}^{-1/2} \\ &\quad \times \exp\left\{-\frac{1}{2}[(\mathbf{y}_1 - \boldsymbol{\mu}_1)' \ (\mathbf{y}_2 - \boldsymbol{\mu}_2)'] \begin{bmatrix} \boldsymbol{\Omega}_{11} & \boldsymbol{\Omega}_{12} \\ \boldsymbol{\Omega}_{21} & \boldsymbol{\Omega}_{22} \end{bmatrix}^{-1} \begin{bmatrix} \mathbf{y}_1 - \boldsymbol{\mu}_1 \\ \mathbf{y}_2 - \boldsymbol{\mu}_2 \end{bmatrix}\right\}. \end{aligned} \qquad [4.6.1]$$

The inverse of $\boldsymbol{\Omega}$ is readily found by inverting [4.5.26]:

$$\begin{aligned} \boldsymbol{\Omega}^{-1} &= [\bar{\mathbf{A}}\bar{\mathbf{D}}\bar{\mathbf{A}}']^{-1} \\ &= [\bar{\mathbf{A}}']^{-1}\bar{\mathbf{D}}^{-1}\bar{\mathbf{A}}^{-1} \\ &= \begin{bmatrix} \mathbf{I}_{n_1} & -\boldsymbol{\Omega}_{11}^{-1}\boldsymbol{\Omega}_{12} \\ \mathbf{0} & \mathbf{I}_{n_2} \end{bmatrix} \begin{bmatrix} \boldsymbol{\Omega}_{11}^{-1} & \mathbf{0} \\ \mathbf{0} & (\boldsymbol{\Omega}_{22} - \boldsymbol{\Omega}_{21}\boldsymbol{\Omega}_{11}^{-1}\boldsymbol{\Omega}_{12})^{-1} \end{bmatrix} \\ &\quad \times \begin{bmatrix} \mathbf{I}_{n_1} & \mathbf{0} \\ -\boldsymbol{\Omega}_{21}\boldsymbol{\Omega}_{11}^{-1} & \mathbf{I}_{n_2} \end{bmatrix}. \end{aligned} \qquad [4.6.2]$$

Likewise, the determinant of $\boldsymbol{\Omega}$ can be found by taking the determinant of [4.5.26]:

$$|\boldsymbol{\Omega}| = |\bar{\mathbf{A}}| \cdot |\bar{\mathbf{D}}| \cdot |\bar{\mathbf{A}}'|.$$

But $\bar{\mathbf{A}}$ is a lower triangular matrix. Its determinant is therefore given by the product of terms along the principal diagonal, all of which are unity. Hence $|\bar{\mathbf{A}}| = 1$ and $|\mathbf{\Omega}| = |\bar{\mathbf{D}}|$:[5]

$$\begin{vmatrix} \mathbf{\Omega}_{11} & \mathbf{\Omega}_{12} \\ \mathbf{\Omega}_{21} & \mathbf{\Omega}_{22} \end{vmatrix} = \begin{vmatrix} \mathbf{\Omega}_{11} & \mathbf{0} \\ \mathbf{0} & \mathbf{\Omega}_{22} - \mathbf{\Omega}_{21}\mathbf{\Omega}_{11}^{-1}\mathbf{\Omega}_{12} \end{vmatrix} \qquad [4.6.3]$$
$$= |\mathbf{\Omega}_{11}| \cdot |\mathbf{\Omega}_{22} - \mathbf{\Omega}_{21}\mathbf{\Omega}_{11}^{-1}\mathbf{\Omega}_{12}|.$$

Substituting [4.6.2] and [4.6.3] into [4.6.1], the joint density can be written

$$\begin{aligned}
&f_{\mathbf{Y}_1,\mathbf{Y}_2}(\mathbf{y}_1, \mathbf{y}_2) \\
&= \frac{1}{(2\pi)^{(n_1+n_2)/2}} |\mathbf{\Omega}_{11}|^{-1/2} \cdot |\mathbf{\Omega}_{22} - \mathbf{\Omega}_{21}\mathbf{\Omega}_{11}^{-1}\mathbf{\Omega}_{12}|^{-1/2} \\
&\quad \times \exp\left\{ -\frac{1}{2} [(\mathbf{y}_1 - \boldsymbol{\mu}_1)' \ (\mathbf{y}_2 - \boldsymbol{\mu}_2)'] \begin{bmatrix} \mathbf{I}_{n_1} & -\mathbf{\Omega}_{11}^{-1}\mathbf{\Omega}_{12} \\ \mathbf{0} & \mathbf{I}_{n_2} \end{bmatrix} \right. \\
&\quad \times \left. \begin{bmatrix} \mathbf{\Omega}_{11}^{-1} & \mathbf{0} \\ \mathbf{0} & (\mathbf{\Omega}_{22} - \mathbf{\Omega}_{21}\mathbf{\Omega}_{11}^{-1}\mathbf{\Omega}_{12})^{-1} \end{bmatrix} \begin{bmatrix} \mathbf{I}_{n_1} & \mathbf{0} \\ -\mathbf{\Omega}_{21}\mathbf{\Omega}_{11}^{-1} & \mathbf{I}_{n_2} \end{bmatrix} \begin{bmatrix} \mathbf{y}_1 - \boldsymbol{\mu}_1 \\ \mathbf{y}_2 - \boldsymbol{\mu}_2 \end{bmatrix} \right\} \\
&= \frac{1}{(2\pi)^{(n_1+n_2)/2}} |\mathbf{\Omega}_{11}|^{-1/2} \cdot |\mathbf{\Omega}_{22} - \mathbf{\Omega}_{21}\mathbf{\Omega}_{11}^{-1}\mathbf{\Omega}_{12}|^{-1/2} \\
&\quad \times \exp\left\{ -\frac{1}{2} [(\mathbf{y}_1 - \boldsymbol{\mu}_1)' \ (\mathbf{y}_2 - \mathbf{m})'] \right. \qquad [4.6.4] \\
&\quad \times \left. \begin{bmatrix} \mathbf{\Omega}_{11}^{-1} & \mathbf{0} \\ \mathbf{0} & (\mathbf{\Omega}_{22} - \mathbf{\Omega}_{21}\mathbf{\Omega}_{11}^{-1}\mathbf{\Omega}_{12})^{-1} \end{bmatrix} \begin{bmatrix} \mathbf{y}_1 - \boldsymbol{\mu}_1 \\ \mathbf{y}_2 - \mathbf{m} \end{bmatrix} \right\} \\
&= \frac{1}{(2\pi)^{(n_1+n_2)/2}} |\mathbf{\Omega}_{11}|^{-1/2} \cdot |\mathbf{\Omega}_{22} - \mathbf{\Omega}_{21}\mathbf{\Omega}_{11}^{-1}\mathbf{\Omega}_{12}|^{-1/2} \\
&\quad \times \exp\left\{ -\frac{1}{2} (\mathbf{y}_1 - \boldsymbol{\mu}_1)'\mathbf{\Omega}_{11}^{-1}(\mathbf{y}_1 - \boldsymbol{\mu}_1) \right. \\
&\quad \left. - \frac{1}{2} (\mathbf{y}_2 - \mathbf{m})'(\mathbf{\Omega}_{22} - \mathbf{\Omega}_{21}\mathbf{\Omega}_{11}^{-1}\mathbf{\Omega}_{12})^{-1}(\mathbf{y}_2 - \mathbf{m}) \right\},
\end{aligned}$$

where

$$\mathbf{m} \equiv \boldsymbol{\mu}_2 + \mathbf{\Omega}_{21}\mathbf{\Omega}_{11}^{-1}(\mathbf{y}_1 - \boldsymbol{\mu}_1). \qquad [4.6.5]$$

The conditional density of $\mathbf{Y}_2$ given $\mathbf{Y}_1$ is found by dividing the joint density [4.6.4] by the marginal density:

$$f_{\mathbf{Y}_1}(\mathbf{y}_1) = \frac{1}{(2\pi)^{n_1/2}} |\mathbf{\Omega}_{11}|^{-1/2} \exp\left[-\frac{1}{2} (\mathbf{y}_1 - \boldsymbol{\mu}_1)'\mathbf{\Omega}_{11}^{-1}(\mathbf{y}_1 - \boldsymbol{\mu}_1) \right].$$

[5]Write $\mathbf{\Omega}_{11}$ in Jordan form as $\mathbf{M}_1\mathbf{J}_1\mathbf{M}_1^{-1}$, where $\mathbf{J}_1$ is upper triangular with eigenvalues of $\mathbf{\Omega}_{11}$ along the principal diagonal. Write $\mathbf{\Omega}_{22} - \mathbf{\Omega}_{21}\mathbf{\Omega}_{11}^{-1}\mathbf{\Omega}_{12}$ as $\mathbf{M}_2\mathbf{J}_2\mathbf{M}_2^{-1}$. Then $\mathbf{\Omega} = \mathbf{M}\mathbf{J}\mathbf{M}^{-1}$, where

$$\mathbf{M} = \begin{bmatrix} \mathbf{M}_1 & \mathbf{0} \\ \mathbf{0} & \mathbf{M}_2 \end{bmatrix} \qquad \mathbf{J} = \begin{bmatrix} \mathbf{J}_1 & \mathbf{0} \\ \mathbf{0} & \mathbf{J}_2 \end{bmatrix}.$$

Thus $\mathbf{\Omega}$ has the same determinant as $\mathbf{J}$. Because $\mathbf{J}$ is upper triangular, its determinant is the product of terms along the principal diagonal, or $|\mathbf{J}| = |\mathbf{J}_1| \cdot |\mathbf{J}_2|$. Hence $|\mathbf{\Omega}| = |\mathbf{\Omega}_{11}| \cdot |\mathbf{\Omega}_{22} - \mathbf{\Omega}_{21}\mathbf{\Omega}_{11}^{-1}\mathbf{\Omega}_{12}|$.

The result of this division is

$$f_{\mathbf{Y}_2|\mathbf{Y}_1}(\mathbf{y}_2|\mathbf{y}_1) = \frac{f_{\mathbf{Y}_1,\mathbf{Y}_2}(\mathbf{y}_1, \mathbf{y}_2)}{f_{\mathbf{Y}_1}(\mathbf{y}_1)}$$

$$= \frac{1}{(2\pi)^{n_2/2}} |\mathbf{H}|^{-1/2} \exp\left[-\frac{1}{2} (\mathbf{y}_2 - \mathbf{m})'\mathbf{H}^{-1}(\mathbf{y}_2 - \mathbf{m}) \right],$$

where

$$\mathbf{H} \equiv \mathbf{\Omega}_{22} - \mathbf{\Omega}_{21}\mathbf{\Omega}_{11}^{-1}\mathbf{\Omega}_{12}. \qquad [4.6.6]$$

In other words,

$$\begin{aligned} \mathbf{Y}_2|\mathbf{Y}_1 &\sim N(\mathbf{m}, \mathbf{H}) \\ &\sim N\Big([\boldsymbol{\mu}_2 + \mathbf{\Omega}_{21}\mathbf{\Omega}_{11}^{-1}(\mathbf{y}_1 - \boldsymbol{\mu}_1)], [\mathbf{\Omega}_{22} - \mathbf{\Omega}_{21}\mathbf{\Omega}_{11}^{-1}\mathbf{\Omega}_{12}] \Big). \end{aligned} \qquad [4.6.7]$$

We saw in Section 4.1 that the optimal unrestricted forecast is given by the conditional expectation. For a Gaussian process, the optimal forecast is thus

$$E(\mathbf{Y}_2|\mathbf{Y}_1) = \boldsymbol{\mu}_2 + \mathbf{\Omega}_{21}\mathbf{\Omega}_{11}^{-1}(\mathbf{y}_1 - \boldsymbol{\mu}_1).$$

On the other hand, for any distribution, the linear projection of the vector $\mathbf{Y}_2$ on a vector $\mathbf{Y}_1$ and a constant term is given by

$$\hat{E}(\mathbf{Y}_2|\mathbf{Y}_1) = \boldsymbol{\mu}_2 + \mathbf{\Omega}_{21}\mathbf{\Omega}_{11}^{-1}(\mathbf{y}_1 - \boldsymbol{\mu}_1).$$

Hence, for a Gaussian process, the linear projection gives the unrestricted optimal forecast.

4.7. *Sums of* ARMA *Processes*

This section explores the nature of series that result from adding two different *ARMA* processes together, beginning with an instructive example.

Sum of an MA*(1) Process Plus White Noise*

Suppose that a series X_t follows a zero-mean $MA(1)$ process:

$$X_t = u_t + \delta u_{t-1}, \qquad [4.7.1]$$

where u_t is white noise:

$$E(u_t u_{t-j}) = \begin{cases} \sigma_u^2 & \text{for } j = 0 \\ 0 & \text{otherwise.} \end{cases}$$

The autocovariances of X_t are thus

$$E(X_t X_{t-j}) = \begin{cases} (1 + \delta^2)\sigma_u^2 & \text{for } j = 0 \\ \delta\sigma_u^2 & \text{for } j = \pm 1 \\ 0 & \text{otherwise.} \end{cases} \qquad [4.7.2]$$

Let v_t indicate a separate white noise series:

$$E(v_t v_{t-j}) = \begin{cases} \sigma_v^2 & \text{for } j = 0 \\ 0 & \text{otherwise.} \end{cases} \qquad [4.7.3]$$

Suppose, furthermore, that v and u are uncorrelated at all leads and lags:

$$E(u_t v_{t-j}) = 0 \qquad \text{for all } j,$$

implying

$$E(X_t v_{t-j}) = 0 \qquad \text{for all } j. \tag{4.7.4}$$

Let an observed series Y_t represent the sum of the $MA(1)$ and the white noise process:

$$\begin{aligned} Y_t &= X_t + v_t \\ &= u_t + \delta u_{t-1} + v_t. \end{aligned} \tag{4.7.5}$$

The question now posed is, What are the time series properties of Y?

Clearly, Y_t has mean zero, and its autocovariances can be deduced from [4.7.2] through [4.7.4]:

$$\begin{aligned} E(Y_t Y_{t-j}) &= E(X_t + v_t)(X_{t-j} + v_{t-j}) \\ &= E(X_t X_{t-j}) + E(v_t v_{t-j}) \\ &= \begin{cases} (1 + \delta^2)\sigma_u^2 + \sigma_v^2 & \text{for } j = 0 \\ \delta\sigma_u^2 & \text{for } j = \pm 1 \\ 0 & \text{otherwise.} \end{cases} \end{aligned} \tag{4.7.6}$$

Thus, the sum $X_t + v_t$ is covariance-stationary, and its autocovariances are zero beyond one lag, as are those for an $MA(1)$. We might naturally then ask whether there exists a zero-mean $MA(1)$ representation for Y,

$$Y_t = \varepsilon_t + \theta\varepsilon_{t-1}, \tag{4.7.7}$$

with

$$E(\varepsilon_t \varepsilon_{t-j}) = \begin{cases} \sigma^2 & \text{for } j = 0 \\ 0 & \text{otherwise,} \end{cases}$$

whose autocovariances match those implied by [4.7.6]. The autocovariances of [4.7.7] would be given by

$$E(Y_t Y_{t-j}) = \begin{cases} (1 + \theta^2)\sigma^2 & \text{for } j = 0 \\ \theta\sigma^2 & \text{for } j = \pm 1 \\ 0 & \text{otherwise.} \end{cases}$$

In order to be consistent with [4.7.6], it would have to be the case that

$$(1 + \theta^2)\sigma^2 = (1 + \delta^2)\sigma_u^2 + \sigma_v^2 \tag{4.7.8}$$

and

$$\theta\sigma^2 = \delta\sigma_u^2. \tag{4.7.9}$$

Equation [4.7.9] can be solved for σ^2,

$$\sigma^2 = \delta\sigma_u^2/\theta, \tag{4.7.10}$$

and then substituted into [4.7.8] to deduce

$$\begin{aligned} (1 + \theta^2)(\delta\sigma_u^2/\theta) &= (1 + \delta^2)\sigma_u^2 + \sigma_v^2 \\ (1 + \theta^2)\delta &= [(1 + \delta^2) + (\sigma_v^2/\sigma_u^2)]\theta \\ \delta\theta^2 - [(1 + \delta^2) + (\sigma_v^2/\sigma_u^2)]\theta + \delta &= 0. \end{aligned} \tag{4.7.11}$$

For given values of δ, σ_u^2, and σ_v^2, two values of θ that satisfy [4.7.11] can be found from the quadratic formula:

$$\theta = \frac{[(1 + \delta^2) + (\sigma_v^2/\sigma_u^2)] \pm \sqrt{[(1 + \delta^2) + (\sigma_v^2/\sigma_u^2)]^2 - 4\delta^2}}{2\delta}. \qquad [4.7.12]$$

If σ_v^2 were equal to zero, the quadratic equation in [4.7.11] would just be

$$\delta\theta^2 - (1 + \delta^2)\theta + \delta = \delta(\theta - \delta)(\theta - \delta^{-1}) = 0, \qquad [4.7.13]$$

whose solutions are $\theta = \delta$ and $\tilde{\theta} = \delta^{-1}$, the moving average parameter for X_t from the invertible and noninvertible representations, respectively. Figure 4.1 graphs equations [4.7.11] and [4.7.13] as functions of θ assuming positive autocorrelation for X_t ($\delta > 0$). For $\theta > 0$ and $\sigma_v^2 > 0$, equation [4.7.11] is everywhere lower than [4.7.13] by the amount $(\sigma_v^2/\sigma_u^2)\theta$, implying that [4.7.11] has two real solutions for θ, an invertible solution θ^* satisfying

$$0 < |\theta^*| < |\delta|, \qquad [4.7.14]$$

and a noninvertible solution $\tilde{\theta}^*$ characterized by

$$1 < |\delta^{-1}| < |\tilde{\theta}^*|.$$

Taking the values associated with the invertible representation (θ^*, σ^{*2}), let us consider whether [4.7.7] could indeed characterize the data $\{Y_t\}$ generated by [4.7.5]. This would require

$$(1 + \theta^* L)\varepsilon_t = (1 + \delta L)u_t + v_t, \qquad [4.7.15]$$

or

$$\begin{aligned} \varepsilon_t &= (1 + \theta^* L)^{-1}[(1 + \delta L)u_t + v_t] \\ &= (u_t - \theta^* u_{t-1} + \theta^{*2} u_{t-2} - \theta^{*3} u_{t-3} + \cdots) \\ &\quad + \delta(u_{t-1} - \theta^* u_{t-2} + \theta^{*2} u_{t-3} - \theta^{*3} u_{t-4} + \cdots) \\ &\quad + (v_t - \theta^* v_{t-1} + \theta^{*2} v_{t-2} - \theta^{*3} v_{t-3} + \cdots). \end{aligned} \qquad [4.7.16]$$

The series ε_t defined in [4.7.16] is a distributed lag on past values of u and v, so it might seem to possess a rich autocorrelation structure. In fact, it turns out to be

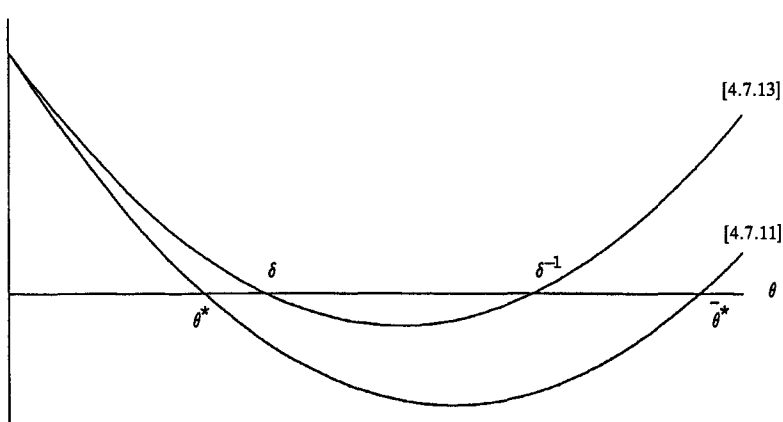


FIGURE 4.1 Graphs of equations [4.7.13] and [4.7.11].

white noise! To see this, note from [4.7.6] that the autocovariance-generating function of Y can be written

$$g_Y(z) = (1 + \delta z)\sigma_u^2(1 + \delta z^{-1}) + \sigma_v^2, \qquad [4.7.17]$$

so that the autocovariance-generating function of $\varepsilon_t = (1 + \theta^* L)^{-1} Y_t$ is

$$g_\varepsilon(z) = \frac{(1 + \delta z)\sigma_u^2(1 + \delta z^{-1}) + \sigma_v^2}{(1 + \theta^* z)(1 + \theta^* z^{-1})}. \qquad [4.7.18]$$

But θ^* and σ^{*2} were chosen so as to make the autocovariance-generating function of $(1 + \theta^* L)\varepsilon_t$, namely,

$$(1 + \theta^* z)\sigma^{*2}(1 + \theta^* z^{-1}),$$

identical to the right side of [4.7.17]. Thus, [4.7.18] is simply equal to

$$g_\varepsilon(z) = \sigma^{*2},$$

a white noise series.

To summarize, adding an $MA(1)$ process to a white noise series with which it is uncorrelated at all leads and lags produces a new $MA(1)$ process characterized by [4.7.7].

Note that the series ε_t in [4.7.16] could not be forecast as a linear function of lagged ε or of lagged Y. Clearly, ε could be forecast, however, on the basis of lagged u or lagged v. The histories $\{u_t\}$ and $\{v_t\}$ contain more information than $\{\varepsilon_t\}$ or $\{Y_t\}$. The optimal forecast of Y_{t+1} on the basis of $\{Y_t, Y_{t-1}, \ldots\}$ would be

$$\hat{E}(Y_{t+1}|Y_t, Y_{t-1}, \ldots) = \theta^* \varepsilon_t$$

with associated mean squared error σ^{*2}. By contrast, the optimal linear forecast of Y_{t+1} on the basis of $\{u_t, u_{t-1}, \ldots, v_t, v_{t-1}, \ldots\}$ would be

$$\hat{E}(Y_{t+1}|u_t, u_{t-1}, \ldots, v_t, v_{t-1}, \ldots) = \delta u_t$$

with associated mean squared error $\sigma_u^2 + \sigma_v^2$. Recalling from [4.7.14] that $|\theta^*| < |\delta|$, it appears from [4.7.9] that $(\theta^{*2})\sigma^{*2} < \delta^2\sigma_u^2$, meaning from [4.7.8] that $\sigma^2 > \sigma_u^2 + \sigma_v^2$. In other words, past values of Y contain less information than past values of u and v.

This example can be useful for thinking about the consequences of differing information sets. One can always make a sensible forecast on the basis of what one knows, $\{Y_t, Y_{t-1}, \ldots\}$, though usually there is other information that could have helped more. An important feature of such settings is that even though ε_t, u_t, and v_t are all white noise, there are complicated correlations between these white noise series.

Another point worth noting is that all that can be estimated on the basis of $\{Y_t\}$ are the two parameters θ^* and σ^{*2}, whereas the true "structural" model [4.7.5] has three parameters (δ, σ_u^2, and σ_v^2). Thus the parameters of the structural model are *unidentified* in the sense in which econometricians use this term—there exists a family of alternative configurations of δ, σ_u^2, and σ_v^2 with $|\delta| < 1$ that would produce the identical value for the likelihood function of the observed data $\{Y_t\}$.

The processes that were added together for this example both had mean zero. Adding constant terms to the processes will not change the results in any interesting way—if X_t is an $MA(1)$ process with mean μ_X and if v_t is white noise plus a constant μ_v, then $X_t + v_t$ will be an $MA(1)$ process with mean given by $\mu_X + \mu_v$. Thus, nothing is lost by restricting the subsequent discussion to sums of zero-mean processes.

Adding Two Moving Average Processes

Suppose next that X_t is a zero-mean $MA(q_1)$ process:

$$X_t = (1 + \delta_1 L + \delta_2 L^2 + \cdots + \delta_{q_1} L^{q_1})u_t \equiv \delta(L)u_t,$$

with

$$E(u_t u_{t-j}) = \begin{cases} \sigma_u^2 & \text{for } j = 0 \\ 0 & \text{otherwise.} \end{cases}$$

Let W_t be a zero-mean $MA(q_2)$ process:

$$W_t = (1 + \kappa_1 L + \kappa_2 L^2 + \cdots + \kappa_{q_2} L^{q_2})v_t \equiv \kappa(L)v_t,$$

with

$$E(v_t v_{t-j}) = \begin{cases} \sigma_v^2 & \text{for } j = 0 \\ 0 & \text{otherwise.} \end{cases}$$

Thus, X has autocovariances $\gamma_0^X, \gamma_1^X, \ldots, \gamma_{q_1}^X$ of the form of [3.3.12] while W has autocovariances $\gamma_0^W, \gamma_1^W, \ldots, \gamma_{q_2}^W$ of the same basic structure. Assume that X and W are uncorrelated with each other at all leads and lags:

$$E(X_t W_{t-j}) = 0 \qquad \text{for all } j;$$

and suppose we observe

$$Y_t = X_t + W_t.$$

Define q to be the larger of q_1 or q_2:

$$q \equiv \max\{q_1, q_2\}.$$

Then the jth autocovariance of Y is given by

$$\begin{aligned} E(Y_t Y_{t-j}) &= E(X_t + W_t)(X_{t-j} + W_{t-j}) \\ &= E(X_t X_{t-j}) + E(W_t W_{t-j}) \\ &= \begin{cases} \gamma_j^X + \gamma_j^W & \text{for } j = 0, \pm 1, \pm 2, \ldots, \pm q \\ 0 & \text{otherwise.} \end{cases} \end{aligned}$$

Thus the autocovariances are zero beyond q lags, suggesting that Y_t might be represented as an $MA(q)$ process.

What more would we need to show to be fully convinced that Y_t is indeed an $MA(q)$ process? This question can be posed in terms of autocovariance-generating functions. Since

$$\gamma_j^Y = \gamma_j^X + \gamma_j^W,$$

it follows that

$$\sum_{j=-\infty}^{\infty} \gamma_j^Y z^j = \sum_{j=-\infty}^{\infty} \gamma_j^X z^j + \sum_{j=-\infty}^{\infty} \gamma_j^W z^j.$$

But these are just the definitions of the respective autocovariance-generating functions,

$$g_Y(z) = g_X(z) + g_W(z). \qquad [4.7.19]$$

Equation [4.7.19] is a quite general result—if one adds together two covariance-stationary processes that are uncorrelated with each other at all leads and lags, the

autocovariance-generating function of the sum is the sum of the autocovariance-generating functions of the individual series.

If Y_t is to be expressed as an $MA(q)$ process,

$$Y_t = (1 + \theta_1 L + \theta_2 L^2 + \cdots + \theta_q L^q)\varepsilon_t \equiv \theta(L)\varepsilon_t$$

with

$$E(\varepsilon_t \varepsilon_{t-j}) = \begin{cases} \sigma^2 & \text{for } j = 0 \\ 0 & \text{otherwise,} \end{cases}$$

then its autocovariance-generating function would be

$$g_Y(z) = \theta(z)\theta(z^{-1})\sigma^2.$$

The question is thus whether there always exist values of $(\theta_1, \theta_2, \ldots, \theta_q, \sigma^2)$ such that [4.7.19] is satisfied:

$$\theta(z)\theta(z^{-1})\sigma^2 = \delta(z)\delta(z^{-1})\sigma_u^2 + \kappa(z)\kappa(z^{-1})\sigma_v^2. \qquad [4.7.20]$$

It turns out that there do. Thus, the conjecture turns out to be correct that if two moving average processes that are uncorrelated with each other at all leads and lags are added together, the result is a new moving average process whose order is the larger of the order of the original two series:

$$MA(q_1) + MA(q_2) = MA(\max\{q_1, q_2\}). \qquad [4.7.21]$$

A proof of this assertion, along with a constructive algorithm for achieving the factorization in [4.7.20], will be provided in Chapter 13.

Adding Two Autoregressive Processes

Suppose now that X_t and W_t are two $AR(1)$ processes:

$$(1 - \pi L)X_t = u_t \qquad [4.7.22]$$

$$(1 - \rho L)W_t = v_t, \qquad [4.7.23]$$

where u_t and v_t are each white noise with u_t uncorrelated with v_τ for all t and τ. Again suppose that we observe

$$Y_t = X_t + W_t$$

and want to forecast Y_{t+1} on the basis of its own lagged values.

If, by chance, X and W share the same autoregressive parameter, or

$$\pi = \rho,$$

then [4.7.22] could simply be added directly to [4.7.23] to deduce

$$(1 - \pi L)X_t + (1 - \pi L)W_t = u_t + v_t$$

or

$$(1 - \pi L)(X_t + W_t) = u_t + v_t.$$

But the sum $u_t + v_t$ is white noise (as a special case of result [4.7.21]), meaning that Y_t has an $AR(1)$ representation

$$(1 - \pi L)Y_t = \varepsilon_t.$$

In the more likely case that the autoregressive parameters π and ρ are different, then [4.7.22] can be multiplied by $(1 - \rho L)$:

$$(1 - \rho L)(1 - \pi L)X_t = (1 - \rho L)u_t; \qquad [4.7.24]$$

and similarly, [4.7.23] could be multiplied by $(1 - \pi L)$:

$$(1 - \pi L)(1 - \rho L)W_t = (1 - \pi L)v_t. \qquad [4.7.25]$$

Adding [4.7.24] to [4.7.25] produces

$$(1 - \rho L)(1 - \pi L)(X_t + W_t) = (1 - \rho L)u_t + (1 - \pi L)v_t. \qquad [4.7.26]$$

From [4.7.21], the right side of [4.7.26] has an $MA(1)$ representation. Thus, we could write

$$(1 - \phi_1 L - \phi_2 L^2)Y_t = (1 + \theta L)\varepsilon_t,$$

where

$$(1 - \phi_1 L - \phi_2 L^2) = (1 - \rho L)(1 - \pi L)$$

and

$$(1 + \theta L)\varepsilon_t = (1 - \rho L)u_t + (1 - \pi L)v_t.$$

In other words,

$$AR(1) + AR(1) = ARMA(2, 1). \qquad [4.7.27]$$

In general, adding an $AR(p_1)$ process

$$\pi(L)X_t = u_t,$$

to an $AR(p_2)$ process with which it is uncorrelated at all leads and lags,

$$\rho(L)W_t = v_t,$$

produces an $ARMA(p_1 + p_2, \max\{p_1, p_2\})$ process,

$$\phi(L)Y_t = \theta(L)\varepsilon_t,$$

where

$$\phi(L) = \pi(L)\rho(L)$$

and

$$\theta(L)\varepsilon_t = \rho(L)u_t + \pi(L)v_t.$$

4.8. *Wold's Decomposition and the Box-Jenkins Modeling Philosophy*

Wold's Decomposition

All of the covariance-stationary processes considered in Chapter 3 can be written in the form

$$Y_t = \mu + \sum_{j=0}^{\infty} \psi_j \varepsilon_{t-j}, \qquad [4.8.1]$$

where ε_t is the white noise error one would make in forecasting Y_t as a linear function of lagged Y and where $\Sigma_{j=0}^{\infty}\psi_j^2 < \infty$ with $\psi_0 = 1$.

One might think that we were able to write all these processes in the form of [4.8.1] because the discussion was restricted to a convenient class of models. However, the following result establishes that the representation [4.8.1] is in fact fundamental for any covariance-stationary time series.

Proposition 4.1: (Wold's decomposition). Any zero-mean covariance-stationary process Y_t can be represented in the form

$$Y_t = \sum_{j=0}^{\infty} \psi_j \varepsilon_{t-j} + \kappa_t, \qquad [4.8.2]$$

where $\psi_0 = 1$ and $\Sigma_{j=0}^{\infty}\psi_j^2 < \infty$. The term ε_t is white noise and represents the error made in forecasting Y_t on the basis of a linear function of lagged Y:

$$\varepsilon_t \equiv Y_t - \hat{E}(Y_t|Y_{t-1}, Y_{t-2}, \ldots). \qquad [4.8.3]$$

The value of κ_t is uncorrelated with ε_{t-j} for any j, though κ_t can be predicted arbitrarily well from a linear function of past values of Y:

$$\kappa_t = \hat{E}(\kappa_t|Y_{t-1}, Y_{t-2}, \ldots).$$

The term κ_t is called the *linearly deterministic* component of Y_t, while $\Sigma_{j=0}^{\infty}\psi_j\varepsilon_{t-j}$ is called the *linearly indeterministic* component. If $\kappa_t \equiv 0$, then the process is called *purely linearly indeterministic*.

This proposition was first proved by Wold (1938).[6] The proposition relies on stable second moments of Y but makes no use of higher moments. It thus describes only optimal linear forecasts of Y.

Finding the Wold representation in principle requires fitting an infinite number of parameters $(\psi_1, \psi_2, \ldots)$ to the data. With a finite number of observations on $(Y_1, Y_2, \ldots, Y_T)$, this will never be possible. As a practical matter, we therefore need to make some additional assumptions about the nature of $(\psi_1, \psi_2, \ldots)$. A typical assumption in Chapter 3 was that $\psi(L)$ can be expressed as the ratio of two finite-order polynomials:

$$\sum_{j=0}^{\infty} \psi_j L^j = \frac{\theta(L)}{\phi(L)} \equiv \frac{1 + \theta_1 L + \theta_2 L^2 + \cdots + \theta_q L^q}{1 - \phi_1 L - \phi_2 L^2 - \cdots - \phi_p L^p}. \qquad [4.8.4]$$

Another approach, based on the presumed "smoothness" of the population spectrum, will be explored in Chapter 6.

The Box-Jenkins Modeling Philosophy

Many forecasters are persuaded of the benefits of parsimony, or using as few parameters as possible. Box and Jenkins (1976) have been influential advocates of this view. They noted that in practice, analysts end up replacing the true operators $\theta(L)$ and $\phi(L)$ with estimates $\hat{\theta}(L)$ and $\hat{\phi}(L)$ based on the data. The more parameters to estimate, the more room there is to go wrong.

Although complicated models can track the data very well over the historical period for which parameters are estimated, they often perform poorly when used for out-of-sample forecasting. For example, the 1960s saw the development of a number of large macroeconometric models purporting to describe the economy using hundreds of macroeconomic variables and equations. Part of the disillusionment with such efforts was the discovery that univariate *ARMA* models with small values of p or q often produced better forecasts than the big models (see for example Nelson, 1972).[7] As we shall see in later chapters, large size alone was hardly the only liability of these large-scale macroeconometric models. Even so, the claim that simpler models provide more robust forecasts has a great many believers across disciplines.

[6]See Sargent (1987, pp. 286–90) for a nice sketch of the intuition behind this result.

[7]For more recent pessimistic evidence about current large-scale models, see Ashley (1988).

The approach to forecasting advocated by Box and Jenkins can be broken down into four steps:

(1) Transform the data, if necessary, so that the assumption of covariance-stationarity is a reasonable one.
(2) Make an initial guess of small values for p and q for an $ARMA(p, q)$ model that might describe the transformed series.
(3) Estimate the parameters in $\phi(L)$ and $\theta(L)$.
(4) Perform diagnostic analysis to confirm that the model is indeed consistent with the observed features of the data.

The first step, selecting a suitable transformation of the data, is discussed in Chapter 15. For now we merely remark that for economic series that grow over time, many researchers use the change in the natural logarithm of the raw data. For example, if X_t is the level of real GNP in year t, then

$$Y_t = \log X_t - \log X_{t-1} \qquad [4.8.5]$$

might be the variable that an $ARMA$ model purports to describe.

The third and fourth steps, estimation and diagnostic testing, will be discussed in Chapters 5 and 14. Analysis of seasonal dynamics can also be an important part of step 2 of the procedure; this is briefly discussed in Section 6.4. The remainder of this section is devoted to an exposition of the second step in the Box-Jenkins procedure on nonseasonal data, namely, selecting candidate values for p and q.[8]

Sample Autocorrelations

An important part of this selection procedure is to form an estimate $\hat{\rho}_j$ of the population autocorrelation ρ_j. Recall that ρ_j was defined as

$$\rho_j \equiv \gamma_j/\gamma_0$$

where

$$\gamma_j = E(Y_t - \mu)(Y_{t-j} - \mu).$$

A natural estimate of the population autocorrelation ρ_j is provided by the corresponding sample moments:

$$\hat{\rho}_j = \hat{\gamma}_j/\hat{\gamma}_0,$$

where

$$\hat{\gamma}_j = \frac{1}{T}\sum_{t=j+1}^{T}(y_t - \bar{y})(y_{t-j} - \bar{y}) \qquad \text{for } j = 0, 1, 2, \ldots, T-1 \qquad [4.8.6]$$

$$\bar{y} = \frac{1}{T}\sum_{t=1}^{T} y_t. \qquad [4.8.7]$$

Note that even though only $T - j$ observations are used to construct $\hat{\gamma}_j$, the denominator in [4.8.6] is T rather than $T - j$. Thus, for large j, expression [4.8.6] shrinks the estimates toward zero, as indeed the population autocovariances go to zero as $j \to \infty$, assuming covariance-stationarity. Also, the full sample of observations is used to construct $\bar{y}$.

[8]Box and Jenkins refer to this step as "identification" of the appropriate model. We avoid Box and Jenkins's terminology, because "identification" has a quite different meaning for econometricians.

Recall that if the data really follow an $MA(q)$ process, then ρ_j will be zero for $j > q$. By contrast, if the data follow an $AR(p)$ process, then ρ_j will gradually decay toward zero as a mixture of exponentials or damped sinusoids. One guide for distinguishing between MA and AR representations, then, would be the decay properties of ρ_j. Often, we are interested in a quick assessment of whether $\rho_j = 0$ for $j = q + 1, q + 2, \ldots$. If the data were really generated by a Gaussian $MA(q)$ process, then the variance of the estimate $\hat{\rho}_j$ could be approximated by[9]

$$\text{Var}(\hat{\rho}_j) \cong \frac{1}{T}\left\{1 + 2\sum_{i=1}^{q} \rho_i^2\right\} \qquad \text{for } j = q + 1, q + 2, \ldots. \qquad [4.8.8]$$

Thus, in particular, if we suspect that the data were generated by Gaussian white noise, then $\hat{\rho}_j$ for any $j \neq 0$ should lie between $\pm 2/\sqrt{T}$ about 95% of the time.

In general, if there is autocorrelation in the process that generated the original data $\{Y_t\}$, then the estimate $\hat{\rho}_j$ will be correlated with $\hat{\rho}_i$ for $i \neq j$.[10] Thus patterns in the estimated $\hat{\rho}_j$ may represent sampling error rather than patterns in the true ρ_j.

Partial Autocorrelation

Another useful measure is the *partial autocorrelation*. The mth population partial autocorrelation (denoted $\alpha_m^{(m)}$) is defined as the last coefficient in a linear projection of Y on its m most recent values (equation [4.3.7]):

$$\hat{Y}_{t+1|t} - \mu = \alpha_1^{(m)}(Y_t - \mu) + \alpha_2^{(m)}(Y_{t-1} - \mu) + \cdots + \alpha_m^{(m)}(Y_{t-m+1} - \mu).$$

We saw in equation [4.3.8] that the vector $\boldsymbol{\alpha}^{(m)}$ can be calculated from

$$\begin{bmatrix} \alpha_1^{(m)} \\ \alpha_2^{(m)} \\ \vdots \\ \alpha_m^{(m)} \end{bmatrix} = \begin{bmatrix} \gamma_0 & \gamma_1 & \cdots & \gamma_{m-1} \\ \gamma_1 & \gamma_0 & \cdots & \gamma_{m-2} \\ \vdots & \vdots & \cdots & \vdots \\ \gamma_{m-1} & \gamma_{m-2} & \cdots & \gamma_0 \end{bmatrix}^{-1} \begin{bmatrix} \gamma_1 \\ \gamma_2 \\ \vdots \\ \gamma_m \end{bmatrix}.$$

Recall that if the data were really generated by an $AR(p)$ process, only the p most recent values of Y would be useful for forecasting. In this case, the projection coefficients on Y's more than p periods in the past are equal to zero:

$$\alpha_m^{(m)} = 0 \qquad \text{for } m = p + 1, p + 2, \ldots.$$

By contrast, if the data really were generated by an $MA(q)$ process with $q \geq 1$, then the partial autocorrelation $\alpha_m^{(m)}$ asymptotically approaches zero instead of cutting off abruptly.

A natural estimate of the mth partial autocorrelation is the last coefficient in an OLS regression of y on a constant and its m most recent values:

$$y_{t+1} = \hat{c} + \hat{\alpha}_1^{(m)} y_t + \hat{\alpha}_2^{(m)} y_{t-1} + \cdots + \hat{\alpha}_m^{(m)} y_{t-m+1} + \hat{e}_t,$$

where $\hat{e}_t$ denotes the OLS regression residual. If the data were really generated by an $AR(p)$ process, then the sample estimate $(\hat{\alpha}_m^{(m)})$ would have a variance around the true value (0) that could be approximated by[11]

$$\text{Var}(\hat{\alpha}_m^{(m)}) \cong 1/T \qquad \text{for } m = p + 1, p + 2, \ldots.$$

[9]See Box and Jenkins (1976, p. 35).

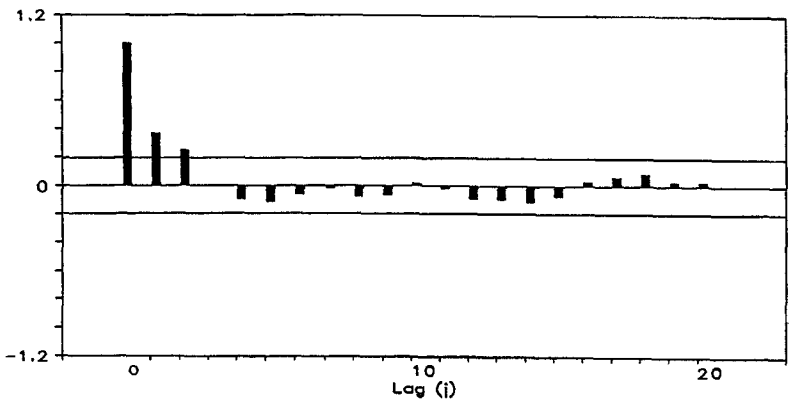
[10]Again, see Box and Jenkins (1976, p. 35).

[11]Box and Jenkins (1976, p. 65).

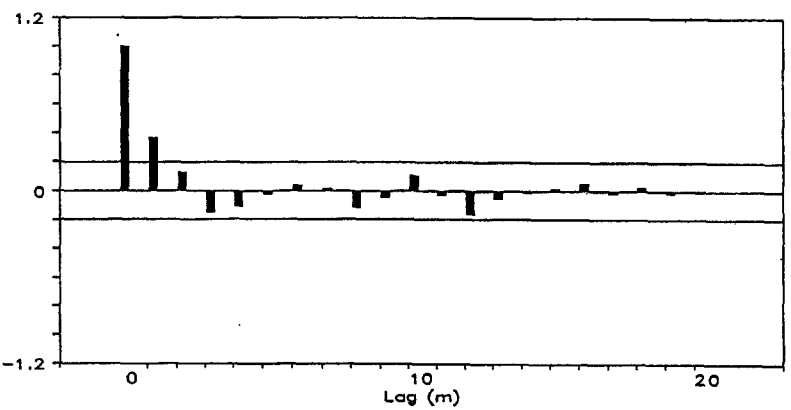
Moreover, if the data were really generated by an $AR(p)$ process, then $\hat{\alpha}_i^{(i)}$ and $\hat{\alpha}_j^{(j)}$ would be asymptotically independent for $i, j > p$.

Example 4.1

We illustrate the Box-Jenkins approach with seasonally adjusted quarterly data on U.S. real GNP from 1947 through 1988. The raw data (x_t) were converted to log changes (y_t) as in [4.8.5]. Panel (a) of Figure 4.2 plots the sample autocorrelations of y ($\hat{\rho}_j$ for $j = 0, 1, \ldots, 20$), while panel (b) displays the sample partial autocorrelations ($\hat{\alpha}_m^{(m)}$ for $m = 0, 1, \ldots, 20$). Ninety-five percent confidence bands ($\pm 2/\sqrt{T}$) are plotted on both panels; for panel (a), these are appropriate under the null hypothesis that the data are really white noise, whereas for panel (b) these are appropriate if the data are really generated by an $AR(p)$ process for p less than m.



(a) Sample autocorrelations



(b) Sample partial autocorrelations

FIGURE 4.2 Sample autocorrelations and partial autocorrelations for U.S. quarterly real GNP growth, 1947:II to 1988:IV. Ninety-five percent confidence intervals are plotted as $\pm 2/\sqrt{T}$.

The first two autocorrelations appear nonzero, suggesting that $q = 2$ would be needed to describe these data as coming from a moving average process. On the other hand, the pattern of autocorrelations appears consistent with the simple geometric decay of an $AR(1)$ process,

$$\rho_j = \phi^j$$

with $\phi \cong 0.4$. The partial autocorrelation could also be viewed as dying out after one lag, also consistent with the $AR(1)$ hypothesis. Thus, one's initial guess for a parsimonious model might be that GNP growth follows an $AR(1)$ process, with $MA(2)$ as another possibility to be considered.

APPENDIX 4.A. *Parallel Between* OLS *Regression and Linear Projection*

This appendix discusses the parallel between ordinary least squares regression and linear projection. This parallel is developed by introducing an artificial random variable specifically constructed so as to have population moments identical to the sample moments of a particular sample. Say that in some particular sample on which we intend to perform *OLS* we have observed T particular values for the explanatory vector, denoted $\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_T$. Consider an artificial discrete-valued random variable ξ that can take on only one of these particular T values, each with probability $(1/T)$:

$$\begin{aligned} P\{\xi = \mathbf{x}_1\} &= 1/T \\ P\{\xi = \mathbf{x}_2\} &= 1/T \\ &\vdots \\ P\{\xi = \mathbf{x}_T\} &= 1/T. \end{aligned}$$

Thus ξ is an artificially constructed random variable whose population probability distribution is given by the empirical distribution function of $\mathbf{x}_t$. The population mean of the random variable ξ is

$$E(\xi) = \sum_{t=1}^{T} \mathbf{x}_t \cdot P\{\xi = \mathbf{x}_t\} = \frac{1}{T}\sum_{t=1}^{T} \mathbf{x}_t.$$

Thus, the population mean of ξ equals the observed sample mean of the true random variable $\mathbf{X}_t$. The population second moment of ξ is

$$E(\xi\xi') = \frac{1}{T}\sum_{t=1}^{T} \mathbf{x}_t\mathbf{x}_t', \qquad [4.A.1]$$

which is the sample second moment of $(\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_T)$.

We can similarly construct a second artificial variable ω that can take on one of the discrete values $(y_2, y_3, \ldots, y_{T+1})$. Suppose that the joint distribution of ω and ξ is given by

$$P\{\xi = \mathbf{x}_t, \omega = y_{t+1}\} = 1/T \qquad \text{for } t = 1, 2, \ldots, T.$$

Then

$$E(\xi\omega) = \frac{1}{T}\sum_{t=1}^{T} \mathbf{x}_t y_{t+1}. \qquad [4.A.2]$$

The coefficient for a linear projection of ω on ξ is the value of $\boldsymbol{\alpha}$ that minimizes

$$E(\omega - \boldsymbol{\alpha}'\xi)^2 = \frac{1}{T}\sum_{t=1}^{T} (y_{t+1} - \boldsymbol{\alpha}'\mathbf{x}_t)^2. \qquad [4.A.3]$$

This is algebraically the same problem as choosing $\boldsymbol{\beta}$ so as to minimize [4.1.17]. Thus, ordinary least squares regression (choosing $\boldsymbol{\beta}$ so as to minimize [4.1.17]) can be viewed as a special case of linear projection (choosing $\boldsymbol{\alpha}$ so as to minimize [4.A.3]). The value of $\boldsymbol{\alpha}$

that minimizes [4.A.3] can be found from substituting the expressions for the population moments of the artificial random variables (equations [4.A.1] and [4.A.2]) into the formula for a linear projection (equation [4.1.13]):

$$\boldsymbol{\alpha} = [E(\boldsymbol{\xi}\boldsymbol{\xi}')]^{-1}E(\boldsymbol{\xi}\omega) = \left[\frac{1}{T}\sum_{t=1}^{T}\mathbf{x}_t\mathbf{x}_t'\right]^{-1}\left[\frac{1}{T}\sum_{t=1}^{T}\mathbf{x}_t y_{t+1}\right].$$

Thus the formula for the *OLS* estimate **b** in [4.1.18] can be obtained as a special case of the formula for the linear projection coefficient $\boldsymbol{\alpha}$ in [4.1.13].

Because linear projections and *OLS* regressions share the same mathematical structure, statements about one have a parallel in the other. This can be a useful device for remembering results or confirming algebra. For example, the statement about population moments,

$$E(Y^2) = \text{Var}(Y) + [E(Y)]^2, \qquad [4.A.4]$$

has the sample analog

$$\frac{1}{T}\sum_{t=1}^{T} y_t^2 = \frac{1}{T}\sum_{t=1}^{T}(y_t - \bar{y})^2 + (\bar{y})^2 \qquad [4.A.5]$$

with $\bar{y} = (1/T)\Sigma_{t=1}^{T} y_t$.

As a second example, suppose that we estimate a series of n *OLS* regressions, with y_{it} the dependent variable for the ith regression and $\mathbf{x}_t$ a $(k \times 1)$ vector of explanatory variables common to each regression. Let $\mathbf{y}_t \equiv (y_{1t}, y_{2t}, \ldots, y_{nt})'$ and write the regression model as

$$\mathbf{y}_t = \boldsymbol{\Pi}'\mathbf{x}_t + \mathbf{u}_t$$

for $\boldsymbol{\Pi}'$ an $(n \times k)$ matrix of regression coefficients. Then the sample variance-covariance matrix of the *OLS* residuals can be inferred from [4.1.24]:

$$\frac{1}{T}\sum_{t=1}^{T}\hat{\mathbf{u}}_t\hat{\mathbf{u}}_t' = \left[\frac{1}{T}\sum_{t=1}^{T}\mathbf{y}_t\mathbf{y}_t'\right] - \left[\frac{1}{T}\sum_{t=1}^{T}\mathbf{y}_t\mathbf{x}_t'\right]\left[\frac{1}{T}\sum_{t=1}^{T}\mathbf{x}_t\mathbf{x}_t'\right]^{-1}\left[\frac{1}{T}\sum_{t=1}^{T}\mathbf{x}_t\mathbf{y}_t'\right], \qquad [4.A.6]$$

where $\hat{\mathbf{u}}_t = \mathbf{y}_t - \hat{\boldsymbol{\Pi}}'\mathbf{x}_t$ and the ith row of $\hat{\boldsymbol{\Pi}}'$ is given by

$$\hat{\boldsymbol{\pi}}_i' = \left\{\left[\frac{1}{T}\sum_{t=1}^{T}\mathbf{x}_t\mathbf{x}_t'\right]^{-1}\left[\frac{1}{T}\sum_{t=1}^{T}\mathbf{x}_t y_{it}\right]\right\}'.$$

APPENDIX 4.B. *Triangular Factorization of the Covariance Matrix for an* MA(*1*) *Process*

This appendix establishes that the triangular factorization of $\boldsymbol{\Omega}$ in [4.5.17] is given by [4.5.18] and [4.5.19].

The magnitude σ^2 is simply a constant term that will end up multiplying every term in the **D** matrix. Recognizing this, we can initially solve the factorization assuming that $\sigma^2 = 1$, and then multiply the resulting **D** matrix by σ^2 to obtain the result for the general case. The (1, 1) element of **D** (ignoring the factor σ^2) is given by the (1, 1) element of $\boldsymbol{\Omega}$: $d_{11} = (1 + \theta^2)$. To put a zero in the (2, 1) position of $\boldsymbol{\Omega}$, we multiply the first row of $\boldsymbol{\Omega}$ by $\theta/(1 + \theta^2)$ and subtract the result from the second; hence, $a_{21} = \theta/(1 + \theta^2)$. This operation changes the (2, 2) element of $\boldsymbol{\Omega}$ to

$$d_{22} = (1 + \theta^2) - \frac{\theta^2}{1 + \theta^2} = \frac{(1 + \theta^2)^2 - \theta^2}{1 + \theta^2} = \frac{1 + \theta^2 + \theta^4}{1 + \theta^2}.$$

To put a zero in the (3, 2) element of $\boldsymbol{\Omega}$, the second row of the new matrix must be multiplied by θ/d_{22} and then subtracted from the third row; hence,

$$a_{32} = \theta/d_{22} = \frac{\theta(1 + \theta^2)}{1 + \theta^2 + \theta^4}.$$

This changes the (3, 3) element to

$$\begin{aligned} d_{33} &= (1 + \theta^2) - \frac{\theta^2(1 + \theta^2)}{1 + \theta^2 + \theta^4} \\ &= \frac{(1 + \theta^2)(1 + \theta^2 + \theta^4) - \theta^2(1 + \theta^2)}{1 + \theta^2 + \theta^4} \\ &= \frac{(1 + \theta^2 + \theta^4) + \theta^2(1 + \theta^2 + \theta^4) - \theta^2(1 + \theta^2)}{1 + \theta^2 + \theta^4} \\ &= \frac{1 + \theta^2 + \theta^4 + \theta^6}{1 + \theta^2 + \theta^4}. \end{aligned}$$

In general, for the ith row,

$$d_{ii} = \frac{1 + \theta^2 + \theta^4 + \cdots + \theta^{2i}}{1 + \theta^2 + \theta^4 + \cdots + \theta^{2(i-1)}}.$$

To put a zero in the $(i + 1, i)$ position, multiply by

$$a_{i+1,i} = \theta/d_{ii} = \frac{\theta[1 + \theta^2 + \theta^4 + \cdots + \theta^{2(i-1)}]}{1 + \theta^2 + \theta^4 + \cdots + \theta^{2i}}$$

and subtract from the $(i + 1)$th row, producing

$$\begin{aligned} d_{i+1,i+1} &= (1 + \theta^2) - \frac{\theta^2[1 + \theta^2 + \theta^4 + \cdots + \theta^{2(i-1)}]}{1 + \theta^2 + \theta^4 + \cdots + \theta^{2i}} \\ &= \frac{(1 + \theta^2 + \theta^4 + \cdots + \theta^{2i}) + \theta^2(1 + \theta^2 + \theta^4 + \cdots + \theta^{2i})}{1 + \theta^2 + \theta^4 + \cdots + \theta^{2i}} \\ &\quad - \frac{\theta^2[1 + \theta^2 + \theta^4 + \cdots + \theta^{2(i-1)}]}{1 + \theta^2 + \theta^4 + \cdots + \theta^{2i}} \\ &= \frac{1 + \theta^2 + \theta^4 + \cdots + \theta^{2(i+1)}}{1 + \theta^2 + \theta^4 + \cdots + \theta^{2i}}. \end{aligned}$$

Chapter 4 Exercises

4.1. Use formula [4.3.6] to show that for a covariance-stationary process, the projection of Y_{t+1} on a constant and Y_t is given by

$$\hat{E}(Y_{t+1}|Y_t) = (1 - \rho_1)\mu + \rho_1 Y_t$$

where $\mu = E(Y_t)$ and $\rho_1 = \gamma_1/\gamma_0$.

(a) Show that for the $AR(1)$ process, this reproduces equation [4.2.19] for $s = 1$.
(b) Show that for the $MA(1)$ process, this reproduces equation [4.5.20] for $n = 2$.
(c) Show that for an $AR(2)$ process, the implied forecast is

$$\mu + [\phi_1/(1 - \phi_2)](Y_t - \mu).$$

Is the error associated with this forecast correlated with Y_t? Is it correlated with Y_{t-1}?

4.2. Verify equation [4.3.3].

4.3. Find the triangular factorization of the following matrix:

$$\begin{bmatrix} 1 & -2 & 3 \\ -2 & 6 & -4 \\ 3 & -4 & 12 \end{bmatrix}.$$

4.4. Can the coefficient on Y_2 from a linear projection of Y_4 on Y_3, Y_2, and Y_1 be found from the (4, 2) element of the matrix $\mathbf{A}$ from the triangular factorization of $\mathbf{\Omega} = E(\mathbf{YY}')$?

4.5. Suppose that X_t follows an $AR(p)$ process and v_t is a white noise process that is uncorrelated with X_{t-j} for all j. Show that the sum

$$Y_t = X_t + v_t$$

follows an $ARMA(p, p)$ process.

4.6. Generalize Exercise 4.5 to deduce that if one adds together an $AR(p)$ process with an $MA(q)$ process and if these two processes are uncorrelated with each other at all leads and lags, then the result is an $ARMA(p, p + q)$ process.

Chapter 4 References

Ashley, Richard. 1988. "On the Relative Worth of Recent Macroeconomic Forecasts." *International Journal of Forecasting* 4:363–76.

Box, George E. P., and Gwilym M. Jenkins. 1976. *Time Series Analysis: Forecasting and Control*, rev. ed. San Francisco: Holden-Day.

Nelson, Charles R. 1972. "The Prediction Performance of the F.R.B.–M.I.T.–PENN Model of the U.S. Economy." *American Economic Review* 62:902–17.

Sargent, Thomas J. 1987. *Macroeconomic Theory*, 2d ed. Boston: Academic Press.

Wold, Herman. 1938 (2d ed. 1954). *A Study in the Analysis of Stationary Time Series*. Uppsala, Sweden: Almqvist and Wiksell.

5 Maximum Likelihood Estimation

5.1. *Introduction*

Consider an *ARMA* model of the form

$$Y_t = c + \phi_1 Y_{t-1} + \phi_2 Y_{t-2} + \cdots + \phi_p Y_{t-p} + \varepsilon_t + \theta_1 \varepsilon_{t-1} + \theta_2 \varepsilon_{t-2} + \cdots + \theta_q \varepsilon_{t-q}, \qquad [5.1.1]$$

with ε_t white noise:

$$E(\varepsilon_t) = 0 \qquad [5.1.2]$$

$$E(\varepsilon_t \varepsilon_\tau) = \begin{cases} \sigma^2 & \text{for } t = \tau \\ 0 & \text{otherwise.} \end{cases} \qquad [5.1.3]$$

The previous chapters assumed that the population parameters $(c, \phi_1, \ldots, \phi_p, \theta_1, \ldots, \theta_q, \sigma^2)$ were known and showed how population moments such as $E(Y_t Y_{t-j})$ and linear forecasts $\hat{E}(Y_{t+s} | Y_t, Y_{t-1}, \ldots)$ could be calculated as functions of these population parameters. This chapter explores how to estimate the values of $(c, \phi_1, \ldots, \phi_p, \theta_1, \ldots, \theta_q, \sigma^2)$ on the basis of observations on Y.

The primary principle on which estimation will be based is *maximum likelihood*. Let $\boldsymbol{\theta} \equiv (c, \phi_1, \ldots, \phi_p, \theta_1, \ldots, \theta_q, \sigma^2)'$ denote the vector of population parameters. Suppose we have observed a sample of size T $(y_1, y_2, \ldots, y_T)$. The approach will be to calculate the probability density

$$f_{Y_T, Y_{T-1}, \ldots, Y_1}(y_T, y_{T-1}, \ldots, y_1; \boldsymbol{\theta}), \qquad [5.1.4]$$

which might loosely be viewed as the probability of having observed this particular sample. The maximum likelihood estimate (*MLE*) of $\boldsymbol{\theta}$ is the value for which this sample is most likely to have been observed; that is, it is the value of $\boldsymbol{\theta}$ that maximizes [5.1.4].

This approach requires specifying a particular distribution for the white noise process ε_t. Typically we will assume that ε_t is Gaussian white noise:

$$\varepsilon_t \sim \text{i.i.d. } N(0, \sigma^2). \qquad [5.1.5]$$

Although this assumption is strong, the estimates of $\boldsymbol{\theta}$ that result from it will often turn out to be sensible for non-Gaussian processes as well.

Finding maximum likelihood estimates conceptually involves two steps. First, the likelihood function [5.1.4] must be calculated. Second, values of $\boldsymbol{\theta}$ must be found that maximize this function. This chapter is organized around these two steps. Sections 5.2 through 5.6 show how to calculate the likelihood function for different Gaussian *ARMA* specifications, while subsequent sections review general techniques for numerical optimization.

5.2. The Likelihood Function for a Gaussian AR(1) Process

Evaluating the Likelihood Function

A Gaussian $AR(1)$ process takes the form

$$Y_t = c + \phi Y_{t-1} + \varepsilon_t, \qquad [5.2.1]$$

with $\varepsilon_t \sim$ i.i.d. $N(0, \sigma^2)$. For this case, the vector of population parameters to be estimated consists of $\boldsymbol{\theta} \equiv (c, \phi, \sigma^2)'$.

Consider the probability distribution of Y_1, the first observation in the sample. From equations [3.4.3] and [3.4.4] this is a random variable with mean

$$E(Y_1) = \mu = c/(1 - \phi)$$

and variance

$$E(Y_1 - \mu)^2 = \sigma^2/(1 - \phi^2).$$

Since $\{\varepsilon_t\}_{t=-\infty}^{\infty}$ is Gaussian, Y_1 is also Gaussian. Hence, the density of the first observation takes the form

$$\begin{aligned} f_{Y_1}(y_1; \boldsymbol{\theta}) &= f_{Y_1}(y_1; c, \phi, \sigma^2) \\ &= \frac{1}{\sqrt{2\pi}\sqrt{\sigma^2/(1-\phi^2)}} \exp\left[\frac{-\{y_1 - [c/(1-\phi)]\}^2}{2\sigma^2/(1-\phi^2)}\right]. \end{aligned} \qquad [5.2.2]$$

Next consider the distribution of the second observation Y_2 conditional on observing $Y_1 = y_1$. From [5.2.1],

$$Y_2 = c + \phi Y_1 + \varepsilon_2. \qquad [5.2.3]$$

Conditioning on $Y_1 = y_1$ means treating the random variable Y_1 as if it were the deterministic constant y_1. For this case, [5.2.3] gives Y_2 as the constant $(c + \phi y_1)$ plus the $N(0, \sigma^2)$ variable ε_2. Hence,

$$(Y_2 | Y_1 = y_1) \sim N((c + \phi y_1), \sigma^2),$$

meaning

$$f_{Y_2|Y_1}(y_2|y_1; \boldsymbol{\theta}) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left[\frac{-(y_2 - c - \phi y_1)^2}{2\sigma^2}\right]. \qquad [5.2.4]$$

The joint density of observations 1 and 2 is then just the product of [5.2.4] and [5.2.2]:

$$f_{Y_2,Y_1}(y_2, y_1; \boldsymbol{\theta}) = f_{Y_2|Y_1}(y_2|y_1; \boldsymbol{\theta}) \cdot f_{Y_1}(y_1; \boldsymbol{\theta}).$$

Similarly, the distribution of the third observation conditional on the first two is

$$f_{Y_3|Y_2,Y_1}(y_3|y_2, y_1; \boldsymbol{\theta}) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left[\frac{-(y_3 - c - \phi y_2)^2}{2\sigma^2}\right],$$

from which

$$f_{Y_3,Y_2,Y_1}(y_3, y_2, y_1; \boldsymbol{\theta}) = f_{Y_3|Y_2,Y_1}(y_3|y_2, y_1; \boldsymbol{\theta}) \cdot f_{Y_2,Y_1}(y_2, y_1; \boldsymbol{\theta}).$$

In general, the values of $Y_1, Y_2, \ldots, Y_{t-1}$ matter for Y_t only through the value of Y_{t-1}, and the density of observation t conditional on the preceding $t - 1$ observations is given by

$$\begin{aligned} f_{Y_t|Y_{t-1},Y_{t-2},\ldots,Y_1}(y_t|y_{t-1}, y_{t-2}, \ldots, y_1; \boldsymbol{\theta}) \\ &= f_{Y_t|Y_{t-1}}(y_t|y_{t-1}; \boldsymbol{\theta}) \\ &= \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left[\frac{-(y_t - c - \phi y_{t-1})^2}{2\sigma^2}\right]. \end{aligned} \qquad [5.2.5]$$

The joint density of the first t observations is then

$$f_{Y_t,Y_{t-1},\ldots,Y_1}(y_t, y_{t-1}, \ldots, y_1; \boldsymbol{\theta}) = f_{Y_t|Y_{t-1}}(y_t|y_{t-1}; \boldsymbol{\theta}) \cdot f_{Y_{t-1},Y_{t-2},\ldots,Y_1}(y_{t-1}, y_{t-2}, \ldots, y_1; \boldsymbol{\theta}). \quad [5.2.6]$$

The likelihood of the complete sample can thus be calculated as

$$f_{Y_T,Y_{T-1},\ldots,Y_1}(y_T, y_{T-1}, \ldots, y_1; \boldsymbol{\theta}) = f_{Y_1}(y_1; \boldsymbol{\theta}) \cdot \prod_{t=2}^{T} f_{Y_t|Y_{t-1}}(y_t|y_{t-1}; \boldsymbol{\theta}). \quad [5.2.7]$$

The log likelihood function (denoted $\mathcal{L}(\boldsymbol{\theta})$) can be found by taking logs of [5.2.7]:

$$\mathcal{L}(\boldsymbol{\theta}) = \log f_{Y_1}(y_1; \boldsymbol{\theta}) + \sum_{t=2}^{T} \log f_{Y_t|Y_{t-1}}(y_t|y_{t-1}; \boldsymbol{\theta}). \quad [5.2.8]$$

Clearly, the value of $\boldsymbol{\theta}$ that maximizes [5.2.8] is identical to the value that maximizes [5.2.7]. However, Section 5.8 presents a number of useful results that can be calculated as a by-product of the maximization if one always poses the problem as maximization of the log likelihood function [5.2.8] rather than the likelihood function [5.2.7].

Substituting [5.2.2] and [5.2.5] into [5.2.8], the log likelihood for a sample of size T from a Gaussian $AR(1)$ process is seen to be

$$\begin{aligned}\mathcal{L}(\boldsymbol{\theta}) = & -\tfrac{1}{2}\log(2\pi) - \tfrac{1}{2}\log[\sigma^2/(1-\phi^2)] \\ & - \frac{\{y_1 - [c/(1-\phi)]\}^2}{2\sigma^2/(1-\phi^2)} - [(T-1)/2]\log(2\pi) \\ & - [(T-1)/2]\log(\sigma^2) - \sum_{t=2}^{T}\left[\frac{(y_t - c - \phi y_{t-1})^2}{2\sigma^2}\right].\end{aligned} \quad [5.2.9]$$

An Alternative Expression for the Likelihood Function

A different description of the likelihood function for a sample of size T from a Gaussian $AR(1)$ process is sometimes useful. Collect the full set of observations in a $(T \times 1)$ vector,

$$\underset{(T\times 1)}{\mathbf{y}} \equiv (y_1, y_2, \ldots, y_T)'.$$

This vector could be viewed as a single realization from a T-dimensional Gaussian distribution. The mean of this $(T \times 1)$ vector is

$$\begin{bmatrix} E(Y_1) \\ E(Y_2) \\ \vdots \\ E(Y_T) \end{bmatrix} = \begin{bmatrix} \mu \\ \mu \\ \vdots \\ \mu \end{bmatrix}, \quad [5.2.10]$$

where, as before, $\mu = c/(1-\phi)$. In vector form, [5.2.10] could be written

$$E(\mathbf{Y}) = \boldsymbol{\mu},$$

where $\boldsymbol{\mu}$ denotes the $(T \times 1)$ vector on the right side of [5.2.10]. The variance-covariance matrix of $\mathbf{Y}$ is given by

$$E[(\mathbf{Y} - \boldsymbol{\mu})(\mathbf{Y} - \boldsymbol{\mu})'] = \boldsymbol{\Omega}, \quad [5.2.11]$$

where

$$\mathbf{\Omega} = \begin{bmatrix} E(Y_1 - \mu)^2 & E(Y_1 - \mu)(Y_2 - \mu) & \cdots & E(Y_1 - \mu)(Y_T - \mu) \\ E(Y_2 - \mu)(Y_1 - \mu) & E(Y_2 - \mu)^2 & \cdots & E(Y_2 - \mu)(Y_T - \mu) \\ \vdots & \vdots & \cdots & \vdots \\ E(Y_T - \mu)(Y_1 - \mu) & E(Y_T - \mu)(Y_2 - \mu) & \cdots & E(Y_T - \mu)^2 \end{bmatrix}. \qquad [5.2.12]$$

The elements of this matrix correspond to autocovariances of Y. Recall that the jth autocovariance for an $AR(1)$ process is given by

$$E(Y_t - \mu)(Y_{t-j} - \mu) = \sigma^2\phi^j/(1 - \phi^2). \qquad [5.2.13]$$

Hence, [5.2.12] can be written as

$$\mathbf{\Omega} = \sigma^2\mathbf{V}, \qquad [5.2.14]$$

where

$$\mathbf{V} = \frac{1}{1 - \phi^2}\begin{bmatrix} 1 & \phi & \phi^2 & \cdots & \phi^{T-1} \\ \phi & 1 & \phi & \cdots & \phi^{T-2} \\ \phi^2 & \phi & 1 & \cdots & \phi^{T-3} \\ \vdots & \vdots & \vdots & \cdots & \vdots \\ \phi^{T-1} & \phi^{T-2} & \phi^{T-3} & \cdots & 1 \end{bmatrix}. \qquad [5.2.15]$$

Viewing the observed sample $\mathbf{y}$ as a single draw from a $N(\boldsymbol{\mu}, \mathbf{\Omega})$ distribution, the sample likelihood could be written down immediately from the formula for the multivariate Gaussian density:

$$f_{\mathbf{Y}}(\mathbf{y}; \boldsymbol{\theta}) = (2\pi)^{-T/2}\,|\mathbf{\Omega}^{-1}|^{1/2}\exp[-\tfrac{1}{2}(\mathbf{y} - \boldsymbol{\mu})'\mathbf{\Omega}^{-1}(\mathbf{y} - \boldsymbol{\mu})], \qquad [5.2.16]$$

with log likelihood

$$\mathcal{L}(\boldsymbol{\theta}) = (-T/2)\log(2\pi) + \tfrac{1}{2}\log|\mathbf{\Omega}^{-1}| - \tfrac{1}{2}(\mathbf{y} - \boldsymbol{\mu})'\mathbf{\Omega}^{-1}(\mathbf{y} - \boldsymbol{\mu}). \qquad [5.2.17]$$

Evidently, [5.2.17] and [5.2.9] must represent the identical function of $(y_1, y_2, \ldots, y_T)$. To verify that this is indeed the case, define

$$\underset{(T\times T)}{\mathbf{L}} \equiv \begin{bmatrix} \sqrt{1 - \phi^2} & 0 & 0 & \cdots & 0 & 0 \\ -\phi & 1 & 0 & \cdots & 0 & 0 \\ 0 & -\phi & 1 & \cdots & 0 & 0 \\ \vdots & \vdots & \vdots & \cdots & \vdots & \vdots \\ 0 & 0 & 0 & \cdots & -\phi & 1 \end{bmatrix}. \qquad [5.2.18]$$

It is straightforward to show that[1]

$$\mathbf{L}'\mathbf{L} = \mathbf{V}^{-1}, \qquad [5.2.19]$$

[1]By direct multiplication, one calculates

$$\mathbf{LV} = \frac{1}{1 - \phi^2}\begin{bmatrix} \sqrt{1 - \phi^2} & \phi\sqrt{1 - \phi^2} & \phi^2\sqrt{1 - \phi^2} & \cdots & \phi^{T-1}\sqrt{1 - \phi^2} \\ 0 & (1 - \phi^2) & \phi(1 - \phi^2) & \cdots & \phi^{T-2}(1 - \phi^2) \\ 0 & 0 & (1 - \phi^2) & \cdots & \phi^{T-3}(1 - \phi^2) \\ \vdots & \vdots & \vdots & \cdots & \vdots \\ 0 & 0 & 0 & \cdots & (1 - \phi^2) \end{bmatrix},$$

and premultiplying this by $\mathbf{L}'$ produces the $(T \times T)$ identity matrix. Thus, $\mathbf{L}'\mathbf{LV} = \mathbf{I}_T$, confirming [5.2.19].

implying from [5.2.14] that

$$\mathbf{\Omega}^{-1} = \sigma^{-2}\mathbf{L}'\mathbf{L}. \qquad [5.2.20]$$

Substituting [5.2.20] into [5.2.17] results in

$$\mathcal{L}(\boldsymbol{\theta}) = (-T/2)\log(2\pi) + \tfrac{1}{2}\log|\sigma^{-2}\mathbf{L}'\mathbf{L}| - \tfrac{1}{2}(\mathbf{y} - \boldsymbol{\mu})'\sigma^{-2}\mathbf{L}'\mathbf{L}(\mathbf{y} - \boldsymbol{\mu}). \qquad [5.2.21]$$

Define the $(T \times 1)$ vector $\tilde{\mathbf{y}}$ to be

$$\begin{aligned}
\tilde{\mathbf{y}} &\equiv \mathbf{L}(\mathbf{y} - \boldsymbol{\mu}) \\
&= \begin{bmatrix} \sqrt{1-\phi^2} & 0 & 0 & \cdots & 0 & 0 \\ -\phi & 1 & 0 & \cdots & 0 & 0 \\ 0 & -\phi & 1 & \cdots & 0 & 0 \\ \vdots & \vdots & \vdots & \cdots & \vdots & \vdots \\ 0 & 0 & 0 & \cdots & -\phi & 1 \end{bmatrix} \begin{bmatrix} y_1 - \mu \\ y_2 - \mu \\ y_3 - \mu \\ \vdots \\ y_T - \mu \end{bmatrix} \qquad [5.2.22] \\
&= \begin{bmatrix} \sqrt{1-\phi^2}\,(y_1 - \mu) \\ (y_2 - \mu) - \phi(y_1 - \mu) \\ (y_3 - \mu) - \phi(y_2 - \mu) \\ \vdots \\ (y_T - \mu) - \phi(y_{T-1} - \mu) \end{bmatrix}.
\end{aligned}$$

Substituting $\mu = c/(1 - \phi)$, this becomes

$$\tilde{\mathbf{y}} = \begin{bmatrix} \sqrt{1-\phi^2}\,[y_1 - c/(1-\phi)] \\ y_2 - c - \phi y_1 \\ y_3 - c - \phi y_2 \\ \vdots \\ y_T - c - \phi y_{T-1} \end{bmatrix}.$$

The last term in [5.2.21] can thus be written

$$\begin{aligned}
\tfrac{1}{2}(\mathbf{y} - \boldsymbol{\mu})'\sigma^{-2}\mathbf{L}'\mathbf{L}(\mathbf{y} - \boldsymbol{\mu}) &= [1/(2\sigma^2)]\tilde{\mathbf{y}}'\tilde{\mathbf{y}} \\
&= [1/(2\sigma^2)](1 - \phi^2)[y_1 - c/(1-\phi)]^2 \qquad [5.2.23] \\
&\quad + [1/(2\sigma^2)] \sum_{t=2}^{T} (y_t - c - \phi y_{t-1})^2.
\end{aligned}$$

The middle term in [5.2.21] is similarly

$$\begin{aligned}
\tfrac{1}{2}\log|\sigma^{-2}\mathbf{L}'\mathbf{L}| &= \tfrac{1}{2}\log\{\sigma^{-2T} \cdot |\mathbf{L}'\mathbf{L}|\} \\
&= -\tfrac{1}{2}\log\sigma^{2T} + \tfrac{1}{2}\log|\mathbf{L}'\mathbf{L}| \qquad [5.2.24] \\
&= (-T/2)\log\sigma^2 + \log|\mathbf{L}|,
\end{aligned}$$

where use has been made of equations [A.4.8], [A.4.9], and [A.4.11] in the Mathematical Review (Appendix A) at the end of the book. Moreover, since $\mathbf{L}$ is lower triangular, its determinant is given by the product of the terms along the principal diagonal: $|\mathbf{L}| = \sqrt{1-\phi^2}$. Thus, [5.2.24] states that

$$\tfrac{1}{2}\log|\sigma^{-2}\mathbf{L}'\mathbf{L}| = (-T/2)\log\sigma^2 + \tfrac{1}{2}\log(1 - \phi^2). \qquad [5.2.25]$$

Substituting [5.2.23] and [5.2.25] into [5,2.21] reproduces [5.2.9]. Thus, equations [5.2.17] and [5.2.9] are just two different expressions for the same magnitude, as claimed. Either expression accurately describes the log likelihood function.

Expression [5.2.17] requires inverting a $(T \times T)$ matrix, whereas [5.2.9] does not. Thus, expression [5.2.9] is clearly to be preferred for computations. It avoids inverting a $(T \times T)$ matrix by writing Y_t as the sum of a forecast $(c + \phi Y_{t-1})$ and a forecast error (ε_t). The forecast error is independent from previous observations by construction, so the log of its density is simply added to the log likelihood of the preceding observations. This approach is known as a *prediction-error decomposition* of the likelihood function.

Exact Maximum Likelihood Estimates for the Gaussian AR*(1) Process*

The *MLE* $\hat{\boldsymbol{\theta}}$ is the value for which [5.2.9] is maximized. In principle, this requires differentiating [5.2.9] and setting the result equal to zero. In practice, when an attempt is made to carry this out, the result is a system of nonlinear equations in $\boldsymbol{\theta}$ and $(y_1, y_2, \ldots, y_T)$ for which there is no simple solution for $\boldsymbol{\theta}$ in terms of $(y_1, y_2, \ldots, y_T)$. Maximization of [5.2.9] thus requires iterative or numerical procedures described in Section 5.7.

Conditional Maximum Likelihood Estimates

An alternative to numerical maximization of the exact likelihood function is to regard the value of y_1 as deterministic and maximize the likelihood conditioned on the first observation,

$$f_{Y_T, Y_{T-1}, \ldots, Y_2 | Y_1}(y_T, y_{T-1}, \ldots, y_2 | y_1; \boldsymbol{\theta}) = \prod_{t=2}^{T} f_{Y_t | Y_{t-1}}(y_t | y_{t-1}; \boldsymbol{\theta}), \qquad [5.2.26]$$

the objective then being to maximize

$$\begin{aligned} &\log f_{Y_T, Y_{T-1}, \ldots, Y_2 | Y_1}(y_T, y_{T-1}, \ldots, y_2 | y_1; \boldsymbol{\theta}) \\ &= -[(T-1)/2] \log(2\pi) - [(T-1)/2] \log(\sigma^2) \\ &\quad - \sum_{t=2}^{T} \left[\frac{(y_t - c - \phi y_{t-1})^2}{2\sigma^2} \right]. \end{aligned} \qquad [5.2.27]$$

Maximization of [5.2.27] with respect to c and ϕ is equivalent to minimization of

$$\sum_{t=2}^{T} (y_t - c - \phi y_{t-1})^2, \qquad [5.2.28]$$

which is achieved by an ordinary least squares (*OLS*) regression of y_t on a constant and its own lagged value. The conditional maximum likelihood estimates of c and ϕ are therefore given by

$$\begin{bmatrix} \hat{c} \\ \hat{\phi} \end{bmatrix} = \begin{bmatrix} T-1 & \Sigma y_{t-1} \\ \Sigma y_{t-1} & \Sigma y_{t-1}^2 \end{bmatrix}^{-1} \begin{bmatrix} \Sigma y_t \\ \Sigma y_{t-1} y_t \end{bmatrix},$$

where Σ denotes summation over $t = 2, 3, \ldots, T$.

The conditional maximum likelihood estimate of the innovation variance is found by differentiating [5.2.27] with respect to σ^2 and setting the result equal to zero:

$$\frac{-(T-1)}{2\sigma^2} + \sum_{t=2}^{T} \left[\frac{(y_t - c - \phi y_{t-1})^2}{2\sigma^4} \right] = 0,$$

or

$$\hat{\sigma}^2 = \sum_{t=2}^{T} \left[\frac{(y_t - \hat{c} - \hat{\phi} y_{t-1})^2}{T - 1} \right].$$

In other words, the conditional *MLE* is the average squared residual from the *OLS* regression [5.2.28].

In contrast to exact maximum likelihood estimates, the conditional maximum likelihood estimates are thus trivial to compute. Moreover, if the sample size T is sufficiently large, the first observation makes a negligible contribution to the total likelihood. The exact *MLE* and conditional *MLE* turn out to have the same large-sample distribution, provided that $|\phi| < 1$. And when $|\phi| > 1$, the conditional *MLE* continues to provide consistent estimates, whereas maximization of [5.2.9] does not. This is because [5.2.9] is derived from [5.2.2], which does not accurately describe the density of Y_1 when $|\phi| > 1$. For these reasons, in most applications the parameters of an autoregression are estimated by *OLS* (conditional maximum likelihood) rather than exact maximum likelihood.

5.3. *The Likelihood Function for a Gaussian* AR(p) *Process*

This section discusses a Gaussian $AR(p)$ process,

$$Y_t = c + \phi_1 Y_{t-1} + \phi_2 Y_{t-2} + \cdots + \phi_p Y_{t-p} + \varepsilon_t, \qquad [5.3.1]$$

with $\varepsilon_t \sim$ i.i.d. $N(0, \sigma^2)$. In this case, the vector of population parameters to be estimated is $\boldsymbol{\theta} = (c, \phi_1, \phi_2, \ldots, \phi_p, \sigma^2)'$.

Evaluating the Likelihood Function

A combination of the two methods described for the $AR(1)$ case is used to calculate the likelihood function for a sample of size T for an $AR(p)$ process. The first p observations in the sample $(y_1, y_2, \ldots, y_p)$ are collected in a $(p \times 1)$ vector $\mathbf{y}_p$, which is viewed as the realization of a p-dimensional Gaussian variable. The mean of this vector is $\boldsymbol{\mu}_p$, which denotes a $(p \times 1)$ vector each of whose elements is given by

$$\mu = c/(1 - \phi_1 - \phi_2 - \cdots - \phi_p). \qquad [5.3.2]$$

Let $\sigma^2 \mathbf{V}_p$ denote the $(p \times p)$ variance-covariance matrix of $(Y_1, Y_2, \ldots, Y_p)$:

$$\sigma^2 \mathbf{V}_p = \begin{bmatrix} E(Y_1 - \mu)^2 & E(Y_1 - \mu)(Y_2 - \mu) & \cdots & E(Y_1 - \mu)(Y_p - \mu) \\ E(Y_2 - \mu)(Y_1 - \mu) & E(Y_2 - \mu)^2 & \cdots & E(Y_2 - \mu)(Y_p - \mu) \\ \vdots & \vdots & \cdots & \vdots \\ E(Y_p - \mu)(Y_1 - \mu) & E(Y_p - \mu)(Y_2 - \mu) & \cdots & E(Y_p - \mu)^2 \end{bmatrix}. \qquad [5.3.3]$$

For example, for a first-order autoregression $(p = 1)$, $\mathbf{V}_p$ is the scalar $1/(1 - \phi^2)$. For a general pth-order autoregression,

$$\sigma^2 \mathbf{V}_p = \begin{bmatrix} \gamma_0 & \gamma_1 & \gamma_2 & \cdots & \gamma_{p-1} \\ \gamma_1 & \gamma_0 & \gamma_1 & \cdots & \gamma_{p-2} \\ \gamma_2 & \gamma_1 & \gamma_0 & \cdots & \gamma_{p-3} \\ \vdots & \vdots & \vdots & \cdots & \vdots \\ \gamma_{p-1} & \gamma_{p-2} & \gamma_{p-3} & \cdots & \gamma_0 \end{bmatrix},$$

where γ_j, the jth autocovariance for an $AR(p)$ process, can be calculated using the methods in Chapter 3. The density of the first p observations is then that of a $N(\boldsymbol{\mu}_p, \sigma^2\mathbf{V}_p)$ variable:

$$\begin{aligned} f_{Y_p,Y_{p-1},\ldots,Y_1}(y_p, y_{p-1}, \ldots, y_1; \boldsymbol{\theta}) &= (2\pi)^{-p/2}|\sigma^{-2}\mathbf{V}_p^{-1}|^{1/2} \exp\left[-\frac{1}{2\sigma^2}(\mathbf{y}_p - \boldsymbol{\mu}_p)'\mathbf{V}_p^{-1}(\mathbf{y}_p - \boldsymbol{\mu}_p)\right] \\ &= (2\pi)^{-p/2}(\sigma^{-2})^{p/2}|\mathbf{V}_p^{-1}|^{1/2} \exp\left[-\frac{1}{2\sigma^2}(\mathbf{y}_p - \boldsymbol{\mu}_p)'\mathbf{V}_p^{-1}(\mathbf{y}_p - \boldsymbol{\mu}_p)\right], \end{aligned} \tag{5.3.4}$$

where use has been made of result [A.4.8].

For the remaining observations in the sample, $(y_{p+1}, y_{p+2}, \ldots, y_T)$, the prediction-error decomposition can be used. Conditional on the first $t-1$ observations, the tth observation is Gaussian with mean

$$c + \phi_1 y_{t-1} + \phi_2 y_{t-2} + \cdots + \phi_p y_{t-p}$$

and variance σ^2. Only the p most recent observations matter for this distribution. Hence, for $t > p$,

$$\begin{aligned} f_{Y_t|Y_{t-1},Y_{t-2},\ldots,Y_1}(y_t|y_{t-1}, y_{t-2}, \ldots, y_1; \boldsymbol{\theta}) &= f_{Y_t|Y_{t-1},Y_{t-2},\ldots,Y_{t-p}}(y_t|y_{t-1}, y_{t-2}, \ldots, y_{t-p}; \boldsymbol{\theta}) \\ &= \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left[\frac{-(y_t - c - \phi_1 y_{t-1} - \phi_2 y_{t-2} - \cdots - \phi_p y_{t-p})^2}{2\sigma^2}\right]. \end{aligned}$$

The likelihood function for the complete sample is then

$$\begin{aligned} f_{Y_T,Y_{T-1},\ldots,Y_1}(y_T, y_{T-1}, \ldots, y_1; \boldsymbol{\theta}) &= f_{Y_p,Y_{p-1},\ldots,Y_1}(y_p, y_{p-1}, \ldots, y_1; \boldsymbol{\theta}) \\ &\quad \times \prod_{t=p+1}^{T} f_{Y_t|Y_{t-1},Y_{t-2},\ldots,Y_{t-p}}(y_t|y_{t-1}, y_{t-2}, \ldots, y_{t-p}; \boldsymbol{\theta}), \end{aligned} \tag{5.3.5}$$

and the log likelihood is therefore

$$\begin{aligned} \mathcal{L}(\boldsymbol{\theta}) &= \log f_{Y_T,Y_{T-1},\ldots,Y_1}(y_T, y_{T-1}, \ldots, y_1; \boldsymbol{\theta}) \\ &= -\frac{p}{2}\log(2\pi) - \frac{p}{2}\log(\sigma^2) + \frac{1}{2}\log|\mathbf{V}_p^{-1}| \\ &\quad - \frac{1}{2\sigma^2}(\mathbf{y}_p - \boldsymbol{\mu}_p)'\mathbf{V}_p^{-1}(\mathbf{y}_p - \boldsymbol{\mu}_p) \\ &\quad - \frac{T-p}{2}\log(2\pi) - \frac{T-p}{2}\log(\sigma^2) \\ &\quad - \sum_{t=p+1}^{T} \frac{(y_t - c - \phi_1 y_{t-1} - \phi_2 y_{t-2} - \cdots - \phi_p y_{t-p})^2}{2\sigma^2} \\ &= -\frac{T}{2}\log(2\pi) - \frac{T}{2}\log(\sigma^2) + \frac{1}{2}\log|\mathbf{V}_p^{-1}| \\ &\quad - \frac{1}{2\sigma^2}(\mathbf{y}_p - \boldsymbol{\mu}_p)'\mathbf{V}_p^{-1}(\mathbf{y}_p - \boldsymbol{\mu}_p) \\ &\quad - \sum_{t=p+1}^{T} \frac{(y_t - c - \phi_1 y_{t-1} - \phi_2 y_{t-2} - \cdots - \phi_p y_{t-p})^2}{2\sigma^2}. \end{aligned} \tag{5.3.6}$$

Evaluation of [5.3.6] requires inverting the $(p \times p)$ matrix $\mathbf{V}_p$. Denote the row i, column j element of $\mathbf{V}_p^{-1}$ by $v^{ij}(p)$. Galbraith and Galbraith (1974, equation

16, p. 70) showed that

$$v^{ij}(p) = \left[\sum_{k=0}^{i-1} \phi_k\phi_{k+j-i} - \sum_{k=p+1-j}^{p+i-j} \phi_k\phi_{k+j-i}\right] \qquad \text{for } 1 \le i \le j \le p, \qquad [5.3.7]$$

where $\phi_0 \equiv -1$. Values of $v^{ij}(p)$ for $i > j$ can be inferred from the fact that $\mathbf{V}_p^{-1}$ is symmetric ($v^{ij}(p) = v^{ji}(p)$). For example, for an $AR(1)$ process, $\mathbf{V}_p^{-1}$ is a scalar whose value is found by taking $i = j = p = 1$:

$$\mathbf{V}_1^{-1} = \left[\sum_{k=0}^{0} \phi_k\phi_k - \sum_{k=1}^{1} \phi_k\phi_k\right] = (\phi_0^2 - \phi_1^2) = (1 - \phi^2).$$

Thus $\sigma^2\mathbf{V}_1 = \sigma^2/(1 - \phi^2)$, which indeed reproduces the formula for the variance of an $AR(1)$ process. For $p = 2$, equation [5.3.7] implies

$$\mathbf{V}_2^{-1} = \begin{bmatrix} (1 - \phi_2^2) & -(\phi_1 + \phi_1\phi_2) \\ -(\phi_1 + \phi_1\phi_2) & (1 - \phi_2^2) \end{bmatrix},$$

from which one readily calculates

$$|\mathbf{V}_2^{-1}| = \left|(1 + \phi_2)\begin{bmatrix} (1 - \phi_2) & -\phi_1 \\ -\phi_1 & (1 - \phi_2) \end{bmatrix}\right| = (1 + \phi_2)^2[(1 - \phi_2)^2 - \phi_1^2]$$

and

$$\begin{aligned}
&(\mathbf{y}_2 - \boldsymbol{\mu}_2)'\mathbf{V}_2^{-1}(\mathbf{y}_2 - \boldsymbol{\mu}_2) \\
&\quad = [(y_1 - \mu) \quad (y_2 - \mu)](1 + \phi_2)\begin{bmatrix} (1 - \phi_2) & -\phi_1 \\ -\phi_1 & (1 - \phi_2) \end{bmatrix}\begin{bmatrix} (y_1 - \mu) \\ (y_2 - \mu) \end{bmatrix} \\
&\quad = (1 + \phi_2) \times \{(1 - \phi_2)(y_1 - \mu)^2 \\
&\qquad - 2\phi_1(y_1 - \mu)(y_2 - \mu) + (1 - \phi_2)(y_2 - \mu)^2\}.
\end{aligned}$$

The exact log likelihood for a Gaussian $AR(2)$ process is thus given by

$$\begin{aligned}
\mathcal{L}(\boldsymbol{\theta}) = &-\frac{T}{2}\log(2\pi) - \frac{T}{2}\log(\sigma^2) + \frac{1}{2}\log\{(1 + \phi_2)^2[(1 - \phi_2)^2 - \phi_1^2]\} \\
&- \left\{\frac{1 + \phi_2}{2\sigma^2}\right\} \times \{(1 - \phi_2)(y_1 - \mu)^2 \\
&- 2\phi_1(y_1 - \mu)(y_2 - \mu) + (1 - \phi_2)(y_2 - \mu)^2\} \\
&- \sum_{t=3}^{T} \frac{(y_t - c - \phi_1 y_{t-1} - \phi_2 y_{t-2})^2}{2\sigma^2},
\end{aligned} \qquad [5.3.8]$$

where $\mu = c/(1 - \phi_1 - \phi_2)$.

Conditional Maximum Likelihood Estimates

Maximization of the exact log likelihood for an $AR(p)$ process [5.3.6] must be accomplished numerically. In contrast, the log of the likelihood conditional on the first p observations assumes the simple form

$$\begin{aligned}
&\log f_{Y_T, Y_{T-1}, \ldots, Y_{p+1}|Y_p, \ldots, Y_1}(y_T, y_{T-1}, \ldots, y_{p+1}|y_p, \ldots, y_1; \boldsymbol{\theta}) \\
&\quad = -\frac{T - p}{2}\log(2\pi) - \frac{T - p}{2}\log(\sigma^2) \\
&\qquad - \sum_{t=p+1}^{T} \frac{(y_t - c - \phi_1 y_{t-1} - \phi_2 y_{t-2} - \cdots - \phi_p y_{t-p})^2}{2\sigma^2}.
\end{aligned} \qquad [5.3.9]$$

The values of $c, \phi_1, \phi_2, \ldots, \phi_p$ that maximize [5.3.9] are the same as those that minimize

$$\sum_{t=p+1}^{T} (y_t - c - \phi_1 y_{t-1} - \phi_2 y_{t-2} - \cdots - \phi_p y_{t-p})^2. \qquad [5.3.10]$$

Thus, the conditional maximum likelihood estimates of these parameters can be obtained from an *OLS* regression of y_t on a constant and p of its own lagged values. The conditional maximum likelihood estimate of σ^2 turns out to be the average squared residual from this regression:

$$\hat{\sigma}^2 = \frac{1}{T-p} \sum_{t=p+1}^{T} (y_t - \hat{c} - \hat{\phi}_1 y_{t-1} - \hat{\phi}_2 y_{t-2} - \cdots - \hat{\phi}_p y_{t-p})^2.$$

The exact maximum likelihood estimates and the conditional maximum likelihood estimates again have the same large-sample distribution.

Maximum Likelihood Estimation for Non-Gaussian Time Series

We noted in Chapter 4 that an *OLS* regression of a variable on a constant and p of its lags would yield a consistent estimate of the coefficients of the linear projection,

$$\hat{E}(Y_t | Y_{t-1}, Y_{t-2}, \ldots, Y_{t-p}),$$

provided that the process is ergodic for second moments. This *OLS* regression also maximizes the Gaussian conditional log likelihood [5.3.9]. Thus, even if the process is non-Gaussian, if we mistakenly form a Gaussian log likelihood function and maximize it, the resulting estimates $(\hat{c}, \hat{\phi}_1, \hat{\phi}_2, \ldots, \hat{\phi}_p)$ will provide consistent estimates of the population parameters in [5.3.1].

An estimate that maximizes a misspecified likelihood function (for example, an *MLE* calculated under the assumption of a Gaussian process when the true data are non-Gaussian) is known as a *quasi-maximum likelihood estimate*. Sometimes, as turns out to be the case here, quasi-maximum likelihood estimation provides consistent estimates of the population parameters of interest. However, standard errors for the estimated coefficients that are calculated under the Gaussianity assumption need not be correct if the true data are non-Gaussian.[2]

Alternatively, if the raw data are non-Gaussian, sometimes a simple transformation such as taking logs will produce a Gaussian time series. For a positive random variable Y_t, Box and Cox (1964) proposed the general class of transformations

$$Y_t^{(\lambda)} = \begin{cases} \dfrac{Y_t^{\lambda} - 1}{\lambda} & \text{for } \lambda \neq 0 \\ \log Y_t & \text{for } \lambda = 0. \end{cases}$$

One approach is to pick a particular value of λ and maximize the likelihood function for $Y_t^{(\lambda)}$ under the assumption that $Y_t^{(\lambda)}$ is a Gaussian *ARMA* process. The value of λ that is associated with the highest value of the maximized likelihood is taken as the best transformation. However, Nelson and Granger (1979) reported discouraging results from this method in practice.

[2]These points were first raised by White (1982) and are discussed further in Sections 5.8 and 14.4.

Li and McLeod (1988) and Janacek and Swift (1990) described approaches to maximum likelihood estimation for some non-Gaussian *ARMA* models. Martin (1981) discussed robust time series estimation for contaminated data.

5.4. *The Likelihood Function for a Gaussian* MA*(1) Process*

Conditional Likelihood Function

Calculation of the likelihood function for an autoregression turned out to be much simpler if we conditioned on initial values for the Y's. Similarly, calculation of the likelihood function for a moving average process is simpler if we condition on initial values for the ε's.

Consider the Gaussian $MA(1)$ process

$$Y_t = \mu + \varepsilon_t + \theta\varepsilon_{t-1} \qquad [5.4.1]$$

with $\varepsilon_t \sim$ i.i.d. $N(0, \sigma^2)$. Let $\boldsymbol{\theta} = (\mu, \theta, \sigma^2)'$ denote the population parameters to be estimated. If the value of ε_{t-1} were known with certainty, then

$$Y_t|\varepsilon_{t-1} \sim N((\mu + \theta\varepsilon_{t-1}), \sigma^2)$$

or

$$f_{Y_t|\varepsilon_{t-1}}(y_t|\varepsilon_{t-1}; \boldsymbol{\theta}) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left[\frac{-(y_t - \mu - \theta\varepsilon_{t-1})^2}{2\sigma^2}\right]. \qquad [5.4.2]$$

Suppose that we knew for certain that $\varepsilon_0 = 0$. Then

$$(Y_1|\varepsilon_0 = 0) \sim N(\mu, \sigma^2).$$

Moreover, given observation of y_1, the value of ε_1 is then known with certainty as well:

$$\varepsilon_1 = y_1 - \mu,$$

allowing application of [5.4.2] again:

$$f_{Y_2|Y_1,\varepsilon_0=0}(y_2|y_1, \varepsilon_0 = 0; \boldsymbol{\theta}) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left[\frac{-(y_2 - \mu - \theta\varepsilon_1)^2}{2\sigma^2}\right].$$

Since ε_1 is known with certainty, ε_2 can be calculated from

$$\varepsilon_2 = y_2 - \mu - \theta\varepsilon_1.$$

Proceeding in this fashion, it is clear that given knowledge that $\varepsilon_0 = 0$, the full sequence $\{\varepsilon_1, \varepsilon_2, \ldots, \varepsilon_T\}$ can be calculated from $\{y_1, y_2, \ldots, y_T\}$ by iterating on

$$\varepsilon_t = y_t - \mu - \theta\varepsilon_{t-1} \qquad [5.4.3]$$

for $t = 1, 2, \ldots, T$, starting from $\varepsilon_0 = 0$. The conditional density of the tth observation can then be calculated from [5.4.2] as

$$\begin{aligned} &f_{Y_t|Y_{t-1},Y_{t-2},\ldots,Y_1,\varepsilon_0=0}(y_t|y_{t-1}, y_{t-2}, \ldots, y_1, \varepsilon_0 = 0; \boldsymbol{\theta}) \\ &= f_{Y_t|\varepsilon_{t-1}}(y_t|\varepsilon_{t-1}; \boldsymbol{\theta}) \qquad [5.4.4] \\ &= \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left[\frac{-\varepsilon_t^2}{2\sigma^2}\right]. \end{aligned}$$

The sample likelihood would then be the product of these individual densities:

$$f_{Y_T,Y_{T-1},\ldots,Y_1|\varepsilon_0=0}(y_T, y_{T-1}, \ldots, y_1|\varepsilon_0 = 0; \boldsymbol{\theta})$$
$$= f_{Y_1|\varepsilon_0=0}(y_1|\varepsilon_0 = 0; \boldsymbol{\theta}) \prod_{t=2}^{T} f_{Y_t|Y_{t-1},Y_{t-2},\ldots,Y_1,\varepsilon_0=0}(y_t|y_{t-1}, y_{t-2}, \ldots, y_1, \varepsilon_0 = 0; \boldsymbol{\theta}).$$

The conditional log likelihood is

$$\begin{aligned}\mathcal{L}(\boldsymbol{\theta}) &= \log f_{Y_T,Y_{T-1},\ldots,Y_1|\varepsilon_0=0}(y_T, y_{T-1}, \ldots, y_1|\varepsilon_0 = 0; \boldsymbol{\theta}) \\ &= -\frac{T}{2}\log(2\pi) - \frac{T}{2}\log(\sigma^2) - \sum_{t=1}^{T}\frac{\varepsilon_t^2}{2\sigma^2}.\end{aligned} \qquad [5.4.5]$$

For a particular numerical value of $\boldsymbol{\theta}$, we thus calculate the sequence of ε's implied by the data from [5.4.3]. The conditional log likelihood [5.4.5] is then a function of the sum of squares of these ε's. Although it is simple to program this iteration by computer, the log likelihood is a fairly complicated nonlinear function of μ and θ, so that an analytical expression for the maximum likelihood estimates of μ and θ is not readily calculated. Hence, even the conditional maximum likelihood estimates for an $MA(1)$ process must be found by numerical optimization.

Iteration on [5.4.3] from an arbitrary starting value of ε_0 will result in

$$\begin{aligned}\varepsilon_t = (y_t - \mu) - \theta(y_{t-1} - \mu) + \theta^2(y_{t-2} - \mu) - \cdots \\ + (-1)^{t-1}\theta^{t-1}(y_1 - \mu) + (-1)^t\theta^t\varepsilon_0.\end{aligned}$$

If $|\theta|$ is substantially less than unity, the effect of imposing $\varepsilon_0 = 0$ will quickly die out and the conditional likelihood [5.4.4] will give a good approximation to the unconditional likelihood for a reasonably large sample size. By contrast, if $|\theta| > 1$, the consequences of imposing $\varepsilon_0 = 0$ accumulate over time. The conditional approach is not reasonable in such a case. If numerical optimization of [5.4.5] results in a value of θ that exceeds 1 in absolute value, the results must be discarded. The numerical optimization should be attempted again with the reciprocal of $\hat{\theta}$ used as a starting value for the numerical search procedure.

Exact Likelihood Function

Two convenient algorithms are available for calculating the exact likelihood function for a Gaussian $MA(1)$ process. One approach is to use the Kalman filter discussed in Chapter 13. A second approach uses the triangular factorization of the variance-covariance matrix. The second approach is described here.

As in Section 5.2, the observations on y can be collected in a $(T \times 1)$ vector $\mathbf{y} \equiv (y_1, y_2, \ldots, y_T)'$ with mean $\boldsymbol{\mu} \equiv (\mu, \mu, \ldots, \mu)'$ and $(T \times T)$ variance-covariance matrix

$$\boldsymbol{\Omega} = E(\mathbf{Y} - \boldsymbol{\mu})(\mathbf{Y} - \boldsymbol{\mu})'.$$

The variance-covariance matrix for T consecutive draws from an $MA(1)$ process is

$$\boldsymbol{\Omega} = \sigma^2 \begin{bmatrix} (1+\theta^2) & \theta & 0 & \cdots & 0 \\ \theta & (1+\theta^2) & \theta & \cdots & 0 \\ 0 & \theta & (1+\theta^2) & \cdots & 0 \\ \vdots & \vdots & \vdots & \cdots & \vdots \\ 0 & 0 & 0 & \cdots & (1+\theta^2) \end{bmatrix}.$$

The likelihood function is then

$$f_{\mathbf{Y}}(\mathbf{y}; \boldsymbol{\theta}) = (2\pi)^{-T/2}|\boldsymbol{\Omega}|^{-1/2}\exp[-\tfrac{1}{2}(\mathbf{y} - \boldsymbol{\mu})'\boldsymbol{\Omega}^{-1}(\mathbf{y} - \boldsymbol{\mu})]. \qquad [5.4.6]$$

A prediction-error decomposition of the likelihood is provided from the triangular factorization of $\mathbf{\Omega}$,

$$\mathbf{\Omega} = \mathbf{ADA}', \qquad [5.4.7]$$

where $\mathbf{A}$ is the lower triangular matrix given in [4.5.18] and $\mathbf{D}$ is the diagonal matrix in [4.5.19]. Substituting [5.4.7] into [5.4.6] gives

$$\begin{aligned} f_{\mathbf{Y}}(\mathbf{y}; \boldsymbol{\theta}) &= (2\pi)^{-T/2}|\mathbf{ADA}'|^{-1/2} \\ &\quad \times \exp[-\tfrac{1}{2}(\mathbf{y} - \boldsymbol{\mu})'[\mathbf{A}']^{-1}\mathbf{D}^{-1}\mathbf{A}^{-1}(\mathbf{y} - \boldsymbol{\mu})]. \end{aligned} \qquad [5.4.8]$$

But $\mathbf{A}$ is a lower triangular matrix with 1s along the principal diagonal. Hence, $|\mathbf{A}| = 1$ and

$$|\mathbf{ADA}'| = |\mathbf{A}|\cdot|\mathbf{D}|\cdot|\mathbf{A}'| = |\mathbf{D}|.$$

Further defining

$$\tilde{\mathbf{y}} \equiv \mathbf{A}^{-1}(\mathbf{y} - \boldsymbol{\mu}), \qquad [5.4.9]$$

the likelihood [5.4.8] can be written

$$f_{\mathbf{Y}}(\mathbf{y}; \boldsymbol{\theta}) = (2\pi)^{-T/2}\,|\mathbf{D}|^{-1/2} \exp[-\tfrac{1}{2}\tilde{\mathbf{y}}'\mathbf{D}^{-1}\tilde{\mathbf{y}}]. \qquad [5.4.10]$$

Notice that [5.4.9] implies

$$\mathbf{A}\tilde{\mathbf{y}} = \mathbf{y} - \boldsymbol{\mu}.$$

The first row of this system states that $\tilde{y}_1 = y_1 - \mu$, while the tth row implies that

$$\tilde{y}_t = y_t - \mu - \frac{\theta[1 + \theta^2 + \theta^4 + \cdots + \theta^{2(t-2)}]}{1 + \theta^2 + \theta^4 + \cdots + \theta^{2(t-1)}}\,\tilde{y}_{t-1}. \qquad [5.4.11]$$

The vector $\tilde{\mathbf{y}}$ can thus be calculated by iterating on [5.4.11] for $t = 2, 3, \ldots, T$ starting from $\tilde{y}_1 = y_1 - \mu$. The variable $\tilde{y}_t$ has the interpretation as the residual from a linear projection of y_t on a constant and $y_{t-1}, y_{t-2}, \ldots, y_1$, while the tth diagonal element of $\mathbf{D}$ gives the MSE of this linear projection:

$$d_{tt} = E(\tilde{Y}_t^2) = \sigma^2\,\frac{1 + \theta^2 + \theta^4 + \cdots + \theta^{2t}}{1 + \theta^2 + \theta^4 + \cdots + \theta^{2(t-1)}}. \qquad [5.4.12]$$

Since $\mathbf{D}$ is diagonal, its determinant is the product of the terms along the principal diagonal,

$$|\mathbf{D}| = \prod_{t=1}^{T} d_{tt}, \qquad [5.4.13]$$

while the inverse of $\mathbf{D}$ is obtained by taking reciprocals of the terms along the principal diagonal. Hence,

$$\tilde{\mathbf{y}}'\mathbf{D}^{-1}\tilde{\mathbf{y}} = \sum_{t=1}^{T} \frac{\tilde{y}_t^2}{d_{tt}}. \qquad [5.4.14]$$

Substituting [5.4.13] and [5.4.14] into [5.4.10], the likelihood function is

$$f_{\mathbf{Y}}(\mathbf{y}; \boldsymbol{\theta}) = (2\pi)^{-T/2}\left[\prod_{t=1}^{T} d_{tt}\right]^{-1/2} \exp\left[-\frac{1}{2}\sum_{t=1}^{T}\frac{\tilde{y}_t^2}{d_{tt}}\right]. \qquad [5.4.15]$$

The exact log likelihood for a Gaussian $MA(1)$ process is therefore

$$\mathcal{L}(\boldsymbol{\theta}) = \log f_{\mathbf{Y}}(\mathbf{y}; \boldsymbol{\theta}) = -\frac{T}{2}\log(2\pi) - \frac{1}{2}\sum_{t=1}^{T}\log(d_{tt}) - \frac{1}{2}\sum_{t=1}^{T}\frac{\tilde{y}_t^2}{d_{tt}}. \qquad [5.4.16]$$

Given numerical values for μ, θ, and σ^2, the sequence $\tilde{y}_t$ is calculated by iterating on [5.4.11] starting with $\tilde{y}_1 = y_1 - \mu$, while d_{tt} is given by [5.4.12].

In contrast to the conditional log likelihood function [5.4.5], expression [5.4.16] will be valid regardless of whether θ is associated with an invertible $MA(1)$ representation. The value of [5.4.16] at $\theta = \bar{\theta}$, $\sigma^2 = \bar{\sigma}^2$ will be identical to its value at $\theta = \bar{\theta}^{-1}$, $\sigma^2 = \bar{\theta}^2\bar{\sigma}^2$; see Exercise 5.1.

5.5. *The Likelihood Function for a Gaussian* MA(q) *Process*

Conditional Likelihood Function

For the $MA(q)$ process,

$$Y_t = \mu + \varepsilon_t + \theta_1\varepsilon_{t-1} + \theta_2\varepsilon_{t-2} + \cdots + \theta_q\varepsilon_{t-q}, \qquad [5.5.1]$$

a simple approach is to condition on the assumption that the first q values for ε were all zero:

$$\varepsilon_0 = \varepsilon_{-1} = \cdots = \varepsilon_{-q+1} = 0. \qquad [5.5.2]$$

From these starting values we can iterate on

$$\varepsilon_t = y_t - \mu - \theta_1\varepsilon_{t-1} - \theta_2\varepsilon_{t-2} - \cdots - \theta_q\varepsilon_{t-q} \qquad [5.5.3]$$

for $t = 1, 2, \ldots, T$. Let $\boldsymbol{\varepsilon}_0$ denote the $(q \times 1)$ vector $(\varepsilon_0, \varepsilon_{-1}, \ldots, \varepsilon_{-q+1})'$. The conditional log likelihood is then

$$\begin{aligned}\mathcal{L}(\boldsymbol{\theta}) &= \log f_{Y_T,Y_{T-1},\ldots,Y_1|\boldsymbol{\varepsilon}_0=\mathbf{0}}(y_T, y_{T-1}, \ldots, y_1|\boldsymbol{\varepsilon}_0 = \mathbf{0}; \boldsymbol{\theta}) \\ &= -\frac{T}{2}\log(2\pi) - \frac{T}{2}\log(\sigma^2) - \sum_{t=1}^{T}\frac{\varepsilon_t^2}{2\sigma^2},\end{aligned} \qquad [5.5.4]$$

where $\boldsymbol{\theta} = (\mu, \theta_1, \theta_2, \ldots, \theta_q, \sigma^2)'$. Again, expression [5.5.4] is useful only if all values of z for which

$$1 + \theta_1 z + \theta_2 z^2 + \cdots + \theta_q z^q = 0$$

lie outside the unit circle.

Exact Likelihood Function

The exact likelihood function is given by

$$f_{\mathbf{Y}}(\mathbf{y}; \boldsymbol{\theta}) = (2\pi)^{-T/2}|\boldsymbol{\Omega}|^{-1/2}\exp[-\tfrac{1}{2}(\mathbf{y} - \boldsymbol{\mu})'\boldsymbol{\Omega}^{-1}(\mathbf{y} - \boldsymbol{\mu})], \qquad [5.5.5]$$

where as before $\mathbf{y} \equiv (y_1, y_2, \ldots, y_T)'$ and $\boldsymbol{\mu} \equiv (\mu, \mu, \ldots, \mu)'$. Here $\boldsymbol{\Omega}$ represents the variance-covariance matrix of T consecutive draws from an $MA(q)$ process:

$$\boldsymbol{\Omega} = \begin{bmatrix} \gamma_0 & \gamma_1 & \cdots & \gamma_q & 0 & \cdots & 0 \\ \gamma_1 & \gamma_0 & \gamma_1 & \cdots & \gamma_q & 0 & \vdots \\ \vdots & \gamma_1 & \ddots & \ddots & \ddots & \ddots & \gamma_q \\ \gamma_q & \cdots & \ddots & \ddots & \ddots & \ddots & 0 \\ 0 & \gamma_q & \ddots & \ddots & \ddots & \ddots & \vdots \\ \vdots & \ddots & \ddots & \ddots & \ddots & \ddots & \gamma_1 \\ 0 & \cdots & \gamma_q & \cdots & \gamma_1 & \gamma_0 & \end{bmatrix}. \qquad [5.5.6]$$

The row i, column j element of $\mathbf{\Omega}$ is given by $\gamma_{|i-j|}$, where γ_k is the kth autocovariance of an $MA(q)$ process:

$$\gamma_k = \begin{cases} \sigma^2(\theta_k + \theta_{k+1}\theta_1 + \theta_{k+2}\theta_2 + \cdots + \theta_q\theta_{q-k}) & \text{for } k = 0, 1, \ldots, q \\ 0 & \text{for } k > q, \end{cases} \qquad [5.5.7]$$

where $\theta_0 \equiv 1$. Again, the exact likelihood function [5.5.5] can be evaluated using either the Kalman filter of Chapter 13 or the triangular factorization of $\mathbf{\Omega}$,

$$\mathbf{\Omega} = \mathbf{ADA}', \qquad [5.5.8]$$

where $\mathbf{A}$ is the lower triangular matrix given by [4.4.11] and $\mathbf{D}$ is the diagonal matrix given by [4.4.7]. Note that the band structure of $\mathbf{\Omega}$ in [5.5.6] makes $\mathbf{A}$ and $\mathbf{D}$ simple to calculate. After the first $(q + 1)$ rows, all the subsequent entries in the first column of $\mathbf{\Omega}$ are already zero, so no multiple of the first row need be added to make these zero. Hence, $a_{i1} = 0$ for $i > q + 1$. Similarly, beyond the first $(q + 2)$ rows of the second column, no multiple of the second row need be added to make these entries zero, meaning that $a_{i2} = 0$ for $i > q + 2$. Thus $\mathbf{A}$ is a lower triangular band matrix with $a_{ij} = 0$ for $i > q + j$:

$$\mathbf{A} = \begin{bmatrix} 1 & 0 & 0 & \cdots & 0 & 0 \\ a_{21} & 1 & 0 & \cdots & 0 & 0 \\ a_{31} & a_{32} & 1 & \cdots & 0 & 0 \\ \vdots & \vdots & \vdots & \cdots & \vdots & \vdots \\ a_{q+1,1} & a_{q+1,2} & a_{q+1,3} & \cdots & 0 & 0 \\ 0 & a_{q+2,2} & a_{q+2,3} & \cdots & 0 & 0 \\ \vdots & \vdots & \vdots & \cdots & \vdots & \vdots \\ 0 & 0 & 0 & \cdots & a_{T,T-1} & 1 \end{bmatrix}.$$

A computer can be programmed to calculate these matrices quickly for a given numerical value for $\boldsymbol{\theta}$.

Substituting [5.5.8] into [5.5.5], the exact likelihood function for a Gaussian $MA(q)$ process can be written as in [5.4.10]:

$$f_{\mathbf{Y}}(\mathbf{y}; \boldsymbol{\theta}) = (2\pi)^{-T/2}|\mathbf{D}|^{-1/2} \exp[-\tfrac{1}{2}\tilde{\mathbf{y}}'\mathbf{D}^{-1}\tilde{\mathbf{y}}]$$

where

$$\mathbf{A}\tilde{\mathbf{y}} = \mathbf{y} - \boldsymbol{\mu}. \qquad [5.5.9]$$

The elements of $\tilde{\mathbf{y}}$ can be calculated recursively by working down the rows of [5.5.9]:

$$\begin{aligned} \tilde{y}_1 &= y_1 - \mu \\ \tilde{y}_2 &= (y_2 - \mu) - a_{21}\tilde{y}_1 \\ \tilde{y}_3 &= (y_3 - \mu) - a_{32}\tilde{y}_2 - a_{31}\tilde{y}_1 \\ &\vdots \\ \tilde{y}_t &= (y_t - \mu) - a_{t,t-1}\tilde{y}_{t-1} - a_{t,t-2}\tilde{y}_{t-2} - \cdots - a_{t,t-q}\tilde{y}_{t-q}. \end{aligned}$$

The exact log likelihood function can then be calculated as in [5.4.16]:

$$\mathcal{L}(\boldsymbol{\theta}) = \log f_{\mathbf{Y}}(\mathbf{y}; \boldsymbol{\theta}) = -\frac{T}{2}\log(2\pi) - \frac{1}{2}\sum_{t=1}^{T}\log(d_{tt}) - \frac{1}{2}\sum_{t=1}^{T}\frac{\tilde{y}_t^2}{d_{tt}}. \qquad [5.5.10]$$

5.6. *The Likelihood Function for a Gaussian* ARMA(p, q) *Process*

Conditional Likelihood Function

A Gaussian $ARMA(p, q)$ process takes the form

$$\begin{aligned} Y_t = c + \phi_1 Y_{t-1} + \phi_2 Y_{t-2} + \cdots + \phi_p Y_{t-p} + \varepsilon_t \\ + \theta_1 \varepsilon_{t-1} + \theta_2 \varepsilon_{t-2} + \cdots + \theta_q \varepsilon_{t-q}, \end{aligned} \tag{5.6.1}$$

where $\varepsilon_t \sim$ i.i.d. $N(0, \sigma^2)$. The goal is to estimate the vector of population parameters $\boldsymbol{\theta} = (c, \phi_1, \phi_2, \ldots, \phi_p, \theta_1, \theta_2, \ldots, \theta_q, \sigma^2)'$.

The approximation to the likelihood function for an autoregression conditioned on initial values of the y's. The approximation to the likelihood function for a moving average process conditioned on initial values of the ε's. A common approximation to the likelihood function for an $ARMA(p, q)$ process conditions on both y's and ε's.

Taking initial values for $\mathbf{y}_0 \equiv (y_0, y_{-1}, \ldots, y_{-p+1})'$ and $\boldsymbol{\varepsilon}_0 \equiv (\varepsilon_0, \varepsilon_{-1}, \ldots, \varepsilon_{-q+1})'$ as given, the sequence $\{\varepsilon_1, \varepsilon_2, \ldots, \varepsilon_T\}$ can be calculated from $\{y_1, y_2, \ldots, y_T\}$ by iterating on

$$\begin{aligned} \varepsilon_t = y_t - c - \phi_1 y_{t-1} - \phi_2 y_{t-2} - \cdots - \phi_p y_{t-p} \\ - \theta_1 \varepsilon_{t-1} - \theta_2 \varepsilon_{t-2} - \cdots - \theta_q \varepsilon_{t-q} \end{aligned} \tag{5.6.2}$$

for $t = 1, 2, \ldots, T$. The conditional log likelihood is then

$$\begin{aligned} \mathcal{L}(\boldsymbol{\theta}) &= \log f_{Y_T, Y_{T-1}, \ldots, Y_1 | \mathbf{Y}_0, \boldsymbol{\varepsilon}_0}(y_T, y_{T-1}, \ldots, y_1 | \mathbf{y}_0, \boldsymbol{\varepsilon}_0; \boldsymbol{\theta}) \\ &= -\frac{T}{2}\log(2\pi) - \frac{T}{2}\log(\sigma^2) - \sum_{t=1}^{T} \frac{\varepsilon_t^2}{2\sigma^2}. \end{aligned} \tag{5.6.3}$$

One option is to set initial y's and ε's equal to their expected values. That is, set $y_s = c/(1 - \phi_1 - \phi_2 - \cdots - \phi_p)$ for $s = 0, -1, \ldots, -p + 1$ and set $\varepsilon_s = 0$ for $s = 0, -1, \ldots, -q + 1$, and then proceed with the iteration in [5.6.2] for $t = 1, 2, \ldots, T$. Alternatively, Box and Jenkins (1976, p. 211) recommended setting ε's to zero but y's equal to their actual values. Thus, iteration on [5.6.2] is started at date $t = p + 1$ with $y_1, y_2, \ldots, y_p$ set to the observed values and

$$\varepsilon_p = \varepsilon_{p-1} = \cdots = \varepsilon_{p-q+1} = 0.$$

Then the conditional likelihood calculated is

$$\begin{aligned} &\log f(y_T, \ldots, y_{p+1} | y_p, \ldots, y_1, \varepsilon_p = 0, \ldots, \varepsilon_{p-q+1} = 0) \\ &\qquad = -\frac{T-p}{2}\log(2\pi) - \frac{T-p}{2}\log(\sigma^2) - \sum_{t=p+1}^{T} \frac{\varepsilon_t^2}{2\sigma^2}. \end{aligned}$$

As in the case for the moving average processes, these approximations should be used only if all values of z satisfying

$$1 + \theta_1 z + \theta_2 z^2 + \cdots + \theta_q z^q = 0$$

lie outside the unit circle.

Alternative Algorithms

The simplest approach to calculating the exact likelihood function for a Gaussian $ARMA$ process is to use the Kalman filter described in Chapter 13. For more

details on exact and approximate maximum likelihood estimation of *ARMA* models, see Galbraith and Galbraith (1974), Box and Jenkins (1976, Chapter 6), Hannan and Rissanen (1982), and Koreisha and Pukkila (1989).

5.7. *Numerical Optimization*

Previous sections of this chapter have shown how to calculate the log likelihood function

$$\mathcal{L}(\boldsymbol{\theta}) = \log f_{Y_T, Y_{T-1}, \ldots, Y_1}(y_T, y_{T-1}, \ldots, y_1; \boldsymbol{\theta}) \qquad [5.7.1]$$

for various specifications of the process thought to have generated the observed data $y_1, y_2, \ldots, y_T$. Given the observed data, the formulas given could be used to calculate the value of $\mathcal{L}(\boldsymbol{\theta})$ for any given numerical value of $\boldsymbol{\theta}$.

This section discusses how to find the value of $\hat{\boldsymbol{\theta}}$ that maximizes $\mathcal{L}(\boldsymbol{\theta})$ given no more knowledge than this ability to calculate the value of $\mathcal{L}(\boldsymbol{\theta})$ for any particular value of $\boldsymbol{\theta}$. The general approach is to write a procedure that enables a computer to calculate the numerical value of $\mathcal{L}(\boldsymbol{\theta})$ for any particular numerical values for $\boldsymbol{\theta}$ and the observed data $y_1, y_2, \ldots, y_T$. We can think of this procedure as a "black box" that enables us to guess some value of $\boldsymbol{\theta}$ and see what the resulting value of $\mathcal{L}(\boldsymbol{\theta})$ would be:

Input		*Procedure*		*Output*
values of $y_1, y_2, \ldots, y_T$ and $\boldsymbol{\theta}$	$\rightarrow$	calculates $\mathcal{L}(\boldsymbol{\theta})$	$\rightarrow$	value of $\mathcal{L}(\boldsymbol{\theta})$

The idea will be to make a series of different guesses for $\boldsymbol{\theta}$, compare the value of $\mathcal{L}(\boldsymbol{\theta})$ for each guess, and try to infer from these values for $\mathcal{L}(\boldsymbol{\theta})$ the value $\hat{\boldsymbol{\theta}}$ for which $\mathcal{L}(\boldsymbol{\theta})$ is largest. Such methods are described as *numerical maximization*.

Grid Search

The simplest approach to numerical maximization is known as the *grid search* method. To illustrate this approach, suppose we have data generated by an $AR(1)$ process, for which the log likelihood was seen to be given by [5.2.9]. To keep the example very simple, it is assumed to be known that the mean of the process is zero ($c = 0$) and that the innovations have unit variance ($\sigma^2 = 1$). Thus the only unknown parameter is the autoregressive coefficient ϕ, and [5.2.9] simplifies to

$$\begin{aligned}\mathcal{L}(\phi) = &-\frac{T}{2}\log(2\pi) + \frac{1}{2}\log(1 - \phi^2) \\ &- \frac{1}{2}(1 - \phi^2)y_1^2 - \frac{1}{2}\sum_{t=2}^{T}(y_t - \phi y_{t-1})^2.\end{aligned} \qquad [5.7.2]$$

Suppose that the observed sample consists of the following $T = 5$ observations:

$$y_1 = 0.8 \qquad y_2 = 0.2 \qquad y_3 = -1.2 \qquad y_4 = -0.4 \qquad y_5 = 0.0.$$

If we make an arbitrary guess as to the value of ϕ, say, $\phi = 0.0$, and plug this guess into expression [5.7.2], we calculate that $\mathcal{L}(\phi) = -5.73$ at $\phi = .0.0$. Trying another guess ($\phi = 0.1$), we calculate $\mathcal{L}(\phi) = -5.71$ at $\phi = 0.1$—the log likelihood is higher at $\phi = 0.1$ than at $\phi = 0.0$. Continuing in this fashion, we could calculate the value of $\mathcal{L}(\phi)$ for every value of ϕ between -0.9 and $+0.9$ in increments of

0.1. The results are reported in Figure 5.1. It appears from these calculations that the log likelihood function $\mathcal{L}(\phi)$ is nicely behaved with a unique maximum at some value of ϕ between 0.1 and 0.3. We could then focus on this subregion of the parameter space and evaluate $\mathcal{L}(\phi)$ at a finer grid, calculating the value of $\mathcal{L}(\phi)$ for all values of ϕ between 0.1 and 0.3 in increments of 0.02. Proceeding in this fashion, it should be possible to get arbitrarily close to the value of ϕ that maximizes $\mathcal{L}(\phi)$ by making the grid finer and finer.

Note that this procedure does not find the *exact MLE* $\hat{\phi}$, but instead approximates it with any accuracy desired. In general, this will be the case with any numerical maximization algorithm. To use these algorithms we therefore have to specify a *convergence criterion*, or some way of deciding when we are close enough to the true maximum. For example, suppose we want an estimate $\hat{\phi}$ that differs from the true *MLE* by no more than ± 0.0001. Then we would continue refining the grid until the increments are in steps of 0.0001, and the best estimate among the elements of that grid would be the numerical *MLE* of ϕ.

For the simple $AR(1)$ example in Figure 5.1, the log likelihood function is *unimodal*—there is a unique value $\boldsymbol{\theta}$ for which $\partial\mathcal{L}(\boldsymbol{\theta})/\partial\boldsymbol{\theta} = \mathbf{0}$. For a general numerical maximization problem, this need not be the case. For example, suppose that we are interested in estimating a scalar parameter θ for which the log likelihood function is as displayed in Figure 5.2. The value $\theta = -0.6$ is a *local maximum*, meaning that the likelihood function is higher there than for any other θ in a neighborhood around $\theta = -0.6$. However, the *global maximum* occurs around $\theta = 0.2$. The grid search method should work well for a unimodal likelihood as long as $\mathcal{L}(\boldsymbol{\theta})$ is continuous. When there are multiple local maxima, the grid must be sufficiently fine to reveal all of the local "hills" on the likelihood surface.

Steepest Ascent

Grid search can be a very good method when there is a single unknown parameter to estimate. However, it quickly becomes intractable when the number of elements of $\boldsymbol{\theta}$ becomes large. An alternative numerical method that often suc-

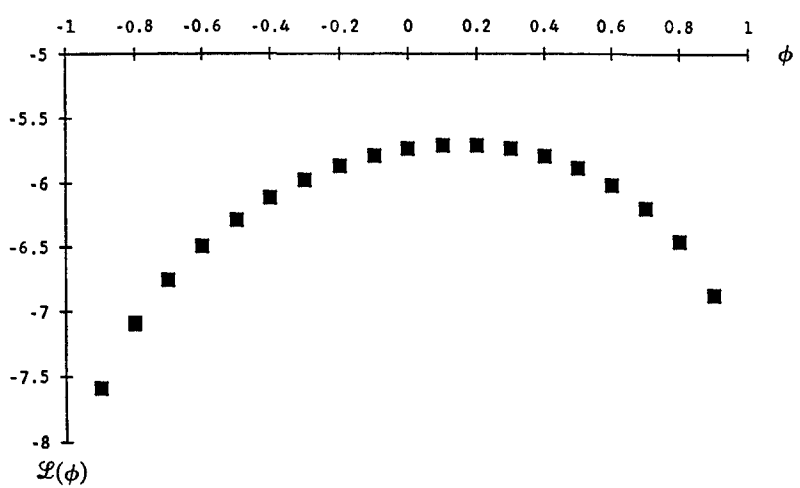


FIGURE 5.1 Log likelihood for an $AR(1)$ process for various guesses of ϕ.

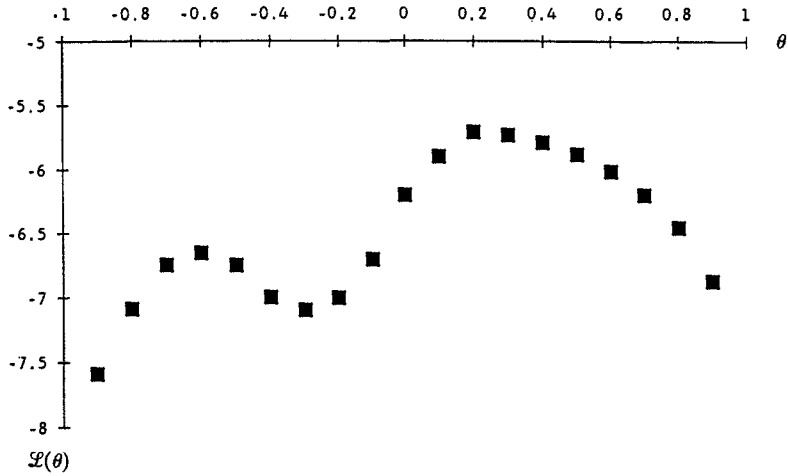


FIGURE 5.2 Bimodal log likelihood function.

ceeds in maximizing a continuously differentiable function of a large number of parameters is known as *steepest ascent*.

To understand this approach, let us temporarily disregard the "black box" nature of the investigation and instead examine how we would proceed analytically with a particular maximization problem. Suppose we have an initial estimate of the parameter vector, denoted $\boldsymbol{\theta}^{(0)}$, and wish to come up with a better estimate $\boldsymbol{\theta}^{(1)}$. Imagine that we are constrained to choose $\boldsymbol{\theta}^{(1)}$ so that the squared distance between $\boldsymbol{\theta}^{(0)}$ and $\boldsymbol{\theta}^{(1)}$ is some fixed number k:

$$\{\boldsymbol{\theta}^{(1)} - \boldsymbol{\theta}^{(0)}\}'\{\boldsymbol{\theta}^{(1)} - \boldsymbol{\theta}^{(0)}\} = k.$$

The optimal value to choose for $\boldsymbol{\theta}^{(1)}$ would then be the solution to the following constrained maximization problem:

$$\max_{\boldsymbol{\theta}^{(1)}} \mathcal{L}(\boldsymbol{\theta}^{(1)}) \qquad \text{subject to} \quad \{\boldsymbol{\theta}^{(1)} - \boldsymbol{\theta}^{(0)}\}'\{\boldsymbol{\theta}^{(1)} - \boldsymbol{\theta}^{(0)}\} = k.$$

To characterize the solution to this problem,[3] form the Lagrangean,

$$J(\boldsymbol{\theta}^{(1)}) = \mathcal{L}(\boldsymbol{\theta}^{(1)}) + \lambda[k - \{\boldsymbol{\theta}^{(1)} - \boldsymbol{\theta}^{(0)}\}'\{\boldsymbol{\theta}^{(1)} - \boldsymbol{\theta}^{(0)}\}], \qquad [5.7.3]$$

where λ denotes a Lagrange multiplier. Differentiating [5.7.3] with respect to $\boldsymbol{\theta}^{(1)}$ and setting the result equal to zero yields

$$\left.\frac{\partial \mathcal{L}(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}}\right|_{\boldsymbol{\theta}=\boldsymbol{\theta}^{(1)}} - (2\lambda)\{\boldsymbol{\theta}^{(1)} - \boldsymbol{\theta}^{(0)}\} = \mathbf{0}. \qquad [5.7.4]$$

Let $\mathbf{g}(\boldsymbol{\theta})$ denote the gradient vector of the log likelihood function:

$$\mathbf{g}(\boldsymbol{\theta}) \equiv \frac{\partial \mathcal{L}(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}}.$$

If there are a elements of $\boldsymbol{\theta}$, then $\mathbf{g}(\boldsymbol{\theta})$ is an $(a \times 1)$ vector whose ith element represents the derivative of the log likelihood with respect to the ith element of $\boldsymbol{\theta}$.

[3]See Chiang (1974) for an introduction to the use of Lagrange multipliers for solving a constrained optimization problem.

Using this notation, expression [5.7.4] can be written as

$$\boldsymbol{\theta}^{(1)} - \boldsymbol{\theta}^{(0)} = [1/(2\lambda)] \cdot \mathbf{g}(\boldsymbol{\theta}^{(1)}). \qquad [5.7.5]$$

Expression [5.7.5] asserts that if we are allowed to change $\boldsymbol{\theta}$ by only a fixed amount, the biggest increase in the log likelihood function will be achieved if the change in $\boldsymbol{\theta}$ (the magnitude $\boldsymbol{\theta}^{(1)} - \boldsymbol{\theta}^{(0)}$) is chosen to be a constant $1/(2\lambda)$ times the gradient vector $\mathbf{g}(\boldsymbol{\theta}^{(1)})$. If we are contemplating a very small step (so that k is near zero), the value $\mathbf{g}(\boldsymbol{\theta}^{(1)})$ will approach $\mathbf{g}(\boldsymbol{\theta}^{(0)})$. In other words, the gradient vector $\mathbf{g}(\boldsymbol{\theta}^{(0)})$ gives the direction in which the log likelihood function increases most steeply from $\boldsymbol{\theta}^{(0)}$.

For illustration, suppose that $a = 2$ and let the log likelihood be

$$\mathcal{L}(\boldsymbol{\theta}) = -1.5\theta_1^2 - 2\theta_2^2. \qquad [5.7.6]$$

We can easily see analytically for this example that the *MLE* is given by $\hat{\boldsymbol{\theta}} = (0, 0)'$. Let us nevertheless use this example to illustrate how the method of steepest ascent works. The elements of the gradient vector are

$$\frac{\partial \mathcal{L}(\boldsymbol{\theta})}{\partial \theta_1} = -3\theta_1 \qquad \frac{\partial \mathcal{L}(\boldsymbol{\theta})}{\partial \theta_2} = -4\theta_2. \qquad [5.7.7]$$

Suppose that the initial guess is $\boldsymbol{\theta}^{(0)} = (-1, 1)'$. Then

$$\left.\frac{\partial \mathcal{L}(\boldsymbol{\theta})}{\partial \theta_1}\right|_{\boldsymbol{\theta}=\boldsymbol{\theta}^{(0)}} = 3 \qquad \left.\frac{\partial \mathcal{L}(\boldsymbol{\theta})}{\partial \theta_2}\right|_{\boldsymbol{\theta}=\boldsymbol{\theta}^{(0)}} = -4.$$

An increase in θ_1 would increase the likelihood, while an increase in θ_2 would decrease the likelihood. The gradient vector evaluated at $\boldsymbol{\theta}^{(0)}$ is

$$\mathbf{g}(\boldsymbol{\theta}^{(0)}) = \begin{bmatrix} 3 \\ -4 \end{bmatrix},$$

so that the optimal step $\boldsymbol{\theta}^{(1)} - \boldsymbol{\theta}^{(0)}$ should be proportional to $(3, -4)'$. For example, with $k = 1$ we would choose

$$\theta_1^{(1)} - \theta_1^{(0)} = \tfrac{3}{5}$$
$$\theta_2^{(1)} - \theta_2^{(0)} = -\tfrac{4}{5};$$

that is, the new guesses would be $\theta_1^{(1)} = -0.4$ and $\theta_2^{(1)} = 0.2$. To increase the likelihood by the greatest amount, we want to increase θ_1 and decrease θ_2 relative to their values at the initial guess $\boldsymbol{\theta}^{(0)}$. Since a one-unit change in θ_2 has a bigger effect on $\mathcal{L}(\boldsymbol{\theta})$ than would a one-unit change in θ_1, the change in θ_2 is larger in absolute value than the change in θ_1.

Let us now return to the black box perspective, where the only capability we have is to calculate the value of $\mathcal{L}(\boldsymbol{\theta})$ for a specified numerical value of $\boldsymbol{\theta}$. We might start with an arbitrary initial guess for the value of $\boldsymbol{\theta}$, denoted $\boldsymbol{\theta}^{(0)}$. Suppose we then calculate the value of the gradient vector at $\boldsymbol{\theta}^{(0)}$:

$$\mathbf{g}(\boldsymbol{\theta}^{(0)}) = \left.\frac{\partial \mathcal{L}(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}}\right|_{\boldsymbol{\theta}=\boldsymbol{\theta}^{(0)}}. \qquad [5.7.8]$$

This gradient could in principle be calculated analytically, by differentiating the general expression for $\mathcal{L}(\boldsymbol{\theta})$ with respect to $\boldsymbol{\theta}$ and writing a computer procedure to calculate each element of $\mathbf{g}(\boldsymbol{\theta})$ given the data and a numerical value for $\boldsymbol{\theta}$. For example, expression [5.7.7] could be used to calculate $\mathbf{g}(\boldsymbol{\theta})$ for any particular value of $\boldsymbol{\theta}$. Alternatively, if it is too hard to differentiate $\mathcal{L}(\boldsymbol{\theta})$ analytically, we can always

get a numerical approximation to the gradient by seeing how $\mathcal{L}(\boldsymbol{\theta})$ changes for a small change in each element of $\boldsymbol{\theta}$. In particular, the ith element of $\mathbf{g}(\boldsymbol{\theta}^{(0)})$ might be approximated by

$$g_i(\boldsymbol{\theta}^{(0)}) \cong \frac{1}{\Delta}\{\mathcal{L}(\theta_1^{(0)}, \theta_2^{(0)}, \ldots, \theta_{i-1}^{(0)}, \theta_i^{(0)} + \Delta, \theta_{i+1}^{(0)}, \theta_{i+2}^{(0)}, \ldots, \theta_a^{(0)}) - \mathcal{L}(\theta_1^{(0)}, \theta_2^{(0)}, \ldots, \theta_{i-1}^{(0)}, \theta_i^{(0)}, \theta_{i+1}^{(0)}, \theta_{i+2}^{(0)}, \ldots, \theta_a^{(0)})\}, \qquad [5.7.9]$$

where Δ represents some arbitrarily chosen small scalar such as $\Delta = 10^{-6}$. By numerically calculating the value of $\mathcal{L}(\boldsymbol{\theta})$ at $\boldsymbol{\theta}^{(0)}$ and at a different values of $\boldsymbol{\theta}$ corresponding to small changes in each of the individual elements of $\boldsymbol{\theta}^{(0)}$, an estimate of the full vector $\mathbf{g}(\boldsymbol{\theta}^{(0)})$ can be uncovered.

Result [5.7.5] suggests that we should change the value of $\boldsymbol{\theta}$ in the direction of the gradient, choosing

$$\boldsymbol{\theta}^{(1)} - \boldsymbol{\theta}^{(0)} = s \cdot \mathbf{g}(\boldsymbol{\theta}^{(0)})$$

for some positive scalar s. A suitable choice for s could be found by an adaptation of the grid search method. For example, we might calculate the value of $\mathcal{L}\{\boldsymbol{\theta}^{(0)} + s \cdot \mathbf{g}(\boldsymbol{\theta}^{(0)})\}$ for $s = \frac{1}{16}, \frac{1}{8}, \frac{1}{4}, \frac{1}{2}, 1, 2, 4, 8$, and 16 and choose as the new estimate $\boldsymbol{\theta}^{(1)}$ the value of $\boldsymbol{\theta}^{(0)} + s \cdot \mathbf{g}(\boldsymbol{\theta}^{(0)})$ for which $\mathcal{L}(\boldsymbol{\theta})$ is largest. Smaller or larger values of s could also be explored if the maximum appears to be at one of the extremes. If none of the values of s improves the likelihood, then a very small value for s such as the value $\Delta = 10^{-6}$ used to approximate the derivative should be tried.

We can then repeat the process, taking $\boldsymbol{\theta}^{(1)} = \boldsymbol{\theta}^{(0)} + s \cdot \mathbf{g}(\boldsymbol{\theta}^{(0)})$ as the starting point, evaluating the gradient at the new location $\mathbf{g}(\boldsymbol{\theta}^{(1)})$, and generating a new estimate $\boldsymbol{\theta}^{(2)}$ according to

$$\boldsymbol{\theta}^{(2)} = \boldsymbol{\theta}^{(1)} + s \cdot \mathbf{g}(\boldsymbol{\theta}^{(1)})$$

for the best choice of s. The process is iterated, calculating

$$\boldsymbol{\theta}^{(m+1)} = \boldsymbol{\theta}^{(m)} + s \cdot \mathbf{g}(\boldsymbol{\theta}^{(m)})$$

for $m = 0, 1, 2, \ldots$ until some convergence criterion is satisfied, such as that the gradient vector $\mathbf{g}(\boldsymbol{\theta}^{(m)})$ is within some specified tolerance of zero, the distance between $\boldsymbol{\theta}^{(m+1)}$ and $\boldsymbol{\theta}^{(m)}$ is less than some specified threshold, or the change between $\mathcal{L}(\boldsymbol{\theta}^{(m+1)})$ and $\mathcal{L}(\boldsymbol{\theta}^{(m)})$ is smaller than some desired amount.

Figure 5.3 illustrates the method of steepest ascent when $\boldsymbol{\theta}$ contains $a = 2$ elements. The figure displays contour lines for the log likelihood $\mathcal{L}(\boldsymbol{\theta})$; along a given contour, the log likelihood $\mathcal{L}(\boldsymbol{\theta})$ is constant. If the iteration is started at the initial guess $\boldsymbol{\theta}^{(0)}$, the gradient $\mathbf{g}(\boldsymbol{\theta}^{(0)})$ describes the direction of steepest ascent. Finding the optimal step in that direction produces the new estimate $\boldsymbol{\theta}^{(1)}$. The gradient at that point $\mathbf{g}(\boldsymbol{\theta}^{(1)})$ then determines a new search direction on which a new estimate $\boldsymbol{\theta}^{(2)}$ is based, until the top of the hill is reached.

Figure 5.3 also illustrates a multivariate generalization of the problem with multiple local maxima seen earlier in Figure 5.2. The procedure should converge to a local maximum, which in this case is different from the global maximum $\boldsymbol{\theta}^*$. In Figure 5.3, it appears that if $\boldsymbol{\theta}^{(0)*}$ were used to begin the iteration in place of $\boldsymbol{\theta}^{(0)}$, the procedure would converge to the true global maximum $\boldsymbol{\theta}^*$. In practice, the only way to ensure that a global maximum is found is to begin the iteration from a number of different starting values for $\boldsymbol{\theta}^{(0)}$ and to continue the sequence from each starting value until the top of the hill associated with that starting value is discovered.

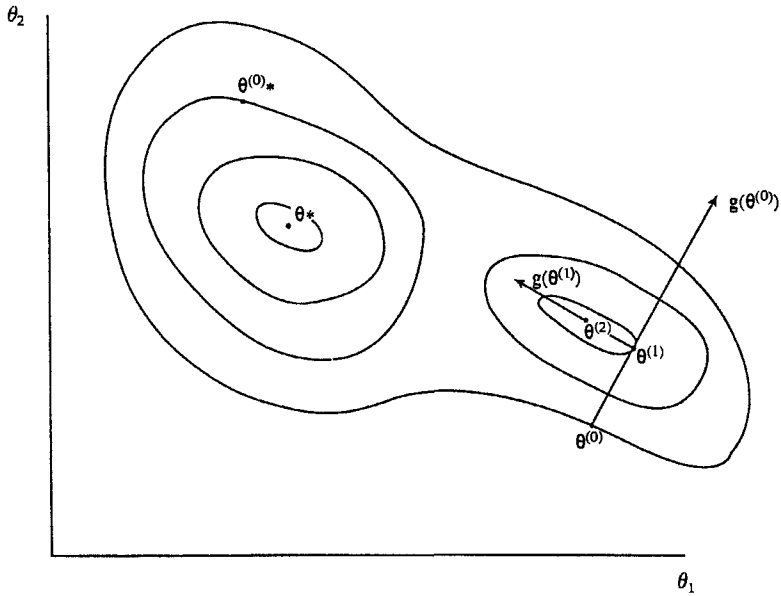


FIGURE 5.3 Likelihood contours and maximization by steepest ascent.

Newton-Raphson

One drawback to the steepest-ascent method is that it may require a very large number of iterations to close in on the local maximum. An alternative method known as *Newton-Raphson* often converges more quickly provided that (1) second derivatives of the log likelihood function $\mathcal{L}(\boldsymbol{\theta})$ exist and (2) the function $\mathcal{L}(\boldsymbol{\theta})$ is concave, meaning that -1 times the matrix of second derivatives is everywhere positive definite.

Suppose that $\boldsymbol{\theta}$ is an $(a \times 1)$ vector of parameters to be estimated. Let $\mathbf{g}(\boldsymbol{\theta}^{(0)})$ denote the gradient vector of the log likelihood function at $\boldsymbol{\theta}^{(0)}$:

$$\underset{(a\times 1)}{\mathbf{g}(\boldsymbol{\theta}^{(0)})} = \left.\frac{\partial \mathcal{L}(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}}\right|_{\boldsymbol{\theta}=\boldsymbol{\theta}^{(0)}};$$

and let $\mathbf{H}(\boldsymbol{\theta}^{(0)})$ denote -1 times the matrix of second derivatives of the log likelihood function:

$$\underset{(a\times a)}{\mathbf{H}(\boldsymbol{\theta}^{(0)})} = -\left.\frac{\partial^2 \mathcal{L}(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}\, \partial \boldsymbol{\theta}'}\right|_{\boldsymbol{\theta}=\boldsymbol{\theta}^{(0)}}.$$

Consider approximating $\mathcal{L}(\boldsymbol{\theta})$ with a second-order Taylor series around $\boldsymbol{\theta}^{(0)}$:

$$\mathcal{L}(\boldsymbol{\theta}) \cong \mathcal{L}(\boldsymbol{\theta}^{(0)}) + [\mathbf{g}(\boldsymbol{\theta}^{(0)})]'[\boldsymbol{\theta} - \boldsymbol{\theta}^{(0)}] - \tfrac{1}{2}[\boldsymbol{\theta} - \boldsymbol{\theta}^{(0)}]'\mathbf{H}(\boldsymbol{\theta}^{(0)})[\boldsymbol{\theta} - \boldsymbol{\theta}^{(0)}]. \qquad [5.7.10]$$

The idea behind the Newton-Raphson method is to choose $\boldsymbol{\theta}$ so as to maximize [5.7.10]. Setting the derivative of [5.7.10] with respect to $\boldsymbol{\theta}$ equal to zero results in

$$\mathbf{g}(\boldsymbol{\theta}^{(0)}) - \mathbf{H}(\boldsymbol{\theta}^{(0)})[\boldsymbol{\theta} - \boldsymbol{\theta}^{(0)}] = \mathbf{0}. \qquad [5.7.11]$$

Let $\boldsymbol{\theta}^{(0)}$ denote an initial guess as to the value of $\boldsymbol{\theta}$. One can calculate the derivative of the log likelihood at that initial guess ($\mathbf{g}(\boldsymbol{\theta}^{(0)})$) either analytically, as in [5.7.7], or numerically, as in [5.7.9]. One can also use analytical or numerical methods to calculate the negative of the matrix of second derivatives at the initial guess ($\mathbf{H}(\boldsymbol{\theta}^{(0)})$). Expression [5.7.11] suggests that an improved estimate of $\boldsymbol{\theta}$ (denoted $\boldsymbol{\theta}^{(1)}$) would satisfy

$$\mathbf{g}(\boldsymbol{\theta}^{(0)}) = \mathbf{H}(\boldsymbol{\theta}^{(0)})[\boldsymbol{\theta}^{(1)} - \boldsymbol{\theta}^{(0)}]$$

or

$$\boldsymbol{\theta}^{(1)} - \boldsymbol{\theta}^{(0)} = [\mathbf{H}(\boldsymbol{\theta}^{(0)})]^{-1}\mathbf{g}(\boldsymbol{\theta}^{(0)}). \qquad [5.7.12]$$

One could next calculate the gradient and Hessian at $\boldsymbol{\theta}^{(1)}$ and use these to find a new estimate $\boldsymbol{\theta}^{(2)}$ and continue iterating in this fashion. The mth step in the iteration updates the estimate of $\boldsymbol{\theta}$ by using the formula

$$\boldsymbol{\theta}^{(m+1)} = \boldsymbol{\theta}^{(m)} + [\mathbf{H}(\boldsymbol{\theta}^{(m)})]^{-1}\mathbf{g}(\boldsymbol{\theta}^{(m)}). \qquad [5.7.13]$$

If the log likelihood function happens to be a perfect quadratic function, then [5.7.10] holds exactly and [5.7.12] will generate the exact *MLE* in a single step:

$$\boldsymbol{\theta}^{(1)} = \hat{\boldsymbol{\theta}}_{MLE}.$$

If the quadratic approximation is reasonably good, Newton-Raphson should converge to the local maximum more quickly than the steepest-ascent method. However, if the likelihood function is not concave, Newton-Raphson behaves quite poorly. Thus, steepest ascent is often slower to converge but sometimes proves to be more robust compared with Newton-Raphson.

Since [5.7.10] is usually only an approximation to the true log likelihood function, the iteration on [5.7.13] is often modified as follows. Expression [5.7.13] is taken to suggest the search direction. The value of the log likelihood function at several points in that direction is then calculated, and the best value determines the length of the step. This strategy calls for replacing [5.7.13] by

$$\boldsymbol{\theta}^{(m+1)} = \boldsymbol{\theta}^{(m)} + s[\mathbf{H}(\boldsymbol{\theta}^{(m)})]^{-1}\mathbf{g}(\boldsymbol{\theta}^{(m)}), \qquad [5.7.14]$$

where s is a scalar controlling the step length. One calculates $\boldsymbol{\theta}^{(m+1)}$ and the associated value for the log likelihood $\mathcal{L}(\boldsymbol{\theta}^{(m+1)})$ for various values of s in [5.7.14] and chooses as the estimate $\boldsymbol{\theta}^{(m+1)}$ the value that produces the biggest value for the log likelihood.

Davidon-Fletcher-Powell

If $\boldsymbol{\theta}$ contains a unknown parameters, then the symmetric matrix $\mathbf{H}(\boldsymbol{\theta})$ has $a(a + 1)/2$ separate elements. Calculating all these elements can be extremely time-consuming if a is large. An alternative approach reasons as follows. The matrix of second derivatives ($-\mathbf{H}(\boldsymbol{\theta})$) corresponds to the first derivatives of the gradient vector ($\mathbf{g}(\boldsymbol{\theta})$), which tell us how $\mathbf{g}(\boldsymbol{\theta})$ changes as $\boldsymbol{\theta}$ changes. We get some independent information about this by comparing $\mathbf{g}(\boldsymbol{\theta}^{(1)}) - \mathbf{g}(\boldsymbol{\theta}^{(0)})$ with $\boldsymbol{\theta}^{(1)} - \boldsymbol{\theta}^{(0)}$. This is not enough information by itself to estimate $\mathbf{H}(\boldsymbol{\theta})$, but it is information that could be used to update an initial guess about the value of $\mathbf{H}(\boldsymbol{\theta})$. Thus, rather than evaluate $\mathbf{H}(\boldsymbol{\theta})$ directly at each iteration, the idea will be to start with an initial guess about $\mathbf{H}(\boldsymbol{\theta})$ and update the guess solely on the basis of how much $\mathbf{g}(\boldsymbol{\theta})$ changes between iterations, given the magnitude of the change in $\boldsymbol{\theta}$. Such methods are sometimes described as *modified Newton-Raphson*.

One of the most popular modified Newton-Raphson methods was proposed by Davidon (1959) and Fletcher and Powell (1963). Since it is $\mathbf{H}^{-1}$ rather than $\mathbf{H}$

itself that appears in the updating formula [5.7.14], the Davidon-Fletcher-Powell algorithm updates an estimate of $\mathbf{H}^{-1}$ at each step on the basis of the size of the change in $\mathbf{g}(\boldsymbol{\theta})$ relative to the change in $\boldsymbol{\theta}$. Specifically, let $\boldsymbol{\theta}^{(m)}$ denote an estimate of $\boldsymbol{\theta}$ that has been calculated at the mth iteration, and let $\mathbf{A}^{(m)}$ denote an estimate of $[\mathbf{H}(\boldsymbol{\theta}^{(m)})]^{-1}$. The new estimate $\boldsymbol{\theta}^{(m+1)}$ is given by

$$\boldsymbol{\theta}^{(m+1)} = \boldsymbol{\theta}^{(m)} + s\mathbf{A}^{(m)}\mathbf{g}(\boldsymbol{\theta}^{(m)}) \qquad [5.7.15]$$

for s the positive scalar that maximizes $\mathcal{L}\{\boldsymbol{\theta}^{(m)} + s\mathbf{A}^{(m)}\mathbf{g}(\boldsymbol{\theta}^{(m)})\}$. Once $\boldsymbol{\theta}^{(m+1)}$ and the gradient at $\boldsymbol{\theta}^{(m+1)}$ have been calculated, a new estimate $\mathbf{A}^{(m+1)}$ is found from

$$\begin{aligned}\mathbf{A}^{(m+1)} = \mathbf{A}^{(m)} &- \frac{\mathbf{A}^{(m)}(\Delta\mathbf{g}^{(m+1)})(\Delta\mathbf{g}^{(m+1)})'\mathbf{A}^{(m)}}{(\Delta\mathbf{g}^{(m+1)})'\mathbf{A}^{(m)}(\Delta\mathbf{g}^{(m+1)})} \\ &- \frac{(\Delta\boldsymbol{\theta}^{(m+1)})(\Delta\boldsymbol{\theta}^{(m+1)})'}{(\Delta\mathbf{g}^{(m+1)})'(\Delta\boldsymbol{\theta}^{(m+1)})}\end{aligned} \qquad [5.7.16]$$

where

$$\Delta\boldsymbol{\theta}^{(m+1)} \equiv \boldsymbol{\theta}^{(m+1)} - \boldsymbol{\theta}^{(m)}$$
$$\Delta\mathbf{g}^{(m+1)} \equiv \mathbf{g}(\boldsymbol{\theta}^{(m+1)}) - \mathbf{g}(\boldsymbol{\theta}^{(m)}).$$

In what sense should $\mathbf{A}^{(m+1)}$ as calculated from [5.7.16] be regarded as an estimate of the inverse of $\mathbf{H}(\boldsymbol{\theta}^{(m+1)})$? Consider first the case when $\boldsymbol{\theta}$ is a scalar ($a = 1$). Then [5.7.16] simplifies to

$$\begin{aligned}A^{(m+1)} &= A^{(m)} - \frac{(A^{(m)})^2(\Delta g^{(m+1)})^2}{(\Delta g^{(m+1)})^2(A^{(m)})} - \frac{(\Delta\theta^{(m+1)})^2}{(\Delta g^{(m+1)})(\Delta\theta^{(m+1)})} \\ &= A^{(m)} - A^{(m)} - \frac{\Delta\theta^{(m+1)}}{\Delta g^{(m+1)}} \\ &= -\frac{\Delta\theta^{(m+1)}}{\Delta g^{(m+1)}}.\end{aligned}$$

In this case,

$$[A^{(m+1)}]^{-1} = -\frac{\Delta g^{(m+1)}}{\Delta\theta^{(m+1)}},$$

which is the natural discrete approximation to

$$H(\theta^{(m+1)}) = -\left.\frac{\partial^2\mathcal{L}}{\partial\theta^2}\right|_{\theta=\theta^{(m+1)}} = -\left.\frac{\partial g}{\partial\theta}\right|_{\theta=\theta^{(m+1)}}.$$

More generally (for $a > 1$), an estimate of the derivative of $\mathbf{g}(\cdot)$ should be related to the observed change in $\mathbf{g}(\cdot)$ according to

$$\mathbf{g}(\boldsymbol{\theta}^{(m+1)}) \cong \mathbf{g}(\boldsymbol{\theta}^{(m)}) + \left.\frac{\partial\mathbf{g}}{\partial\boldsymbol{\theta}'}\right|_{\boldsymbol{\theta}=\boldsymbol{\theta}^{(m+1)}}[\boldsymbol{\theta}^{(m+1)} - \boldsymbol{\theta}^{(m)}].$$

That is,

$$\mathbf{g}(\boldsymbol{\theta}^{(m+1)}) \cong \mathbf{g}(\boldsymbol{\theta}^{(m)}) - \mathbf{H}(\boldsymbol{\theta}^{(m+1)})[\boldsymbol{\theta}^{(m+1)} - \boldsymbol{\theta}^{(m)}]$$

or

$$\Delta\boldsymbol{\theta}^{(m+1)} \cong -[\mathbf{H}(\boldsymbol{\theta}^{(m+1)})]^{-1}\,\Delta\mathbf{g}^{(m+1)}.$$

Hence an estimate $\mathbf{A}^{(m+1)}$ of $[\mathbf{H}(\boldsymbol{\theta}^{(m+1)})]^{-1}$ should satisfy

$$\mathbf{A}^{(m+1)}\,\Delta\mathbf{g}^{(m+1)} = -\Delta\boldsymbol{\theta}^{(m+1)}. \qquad [5.7.17]$$

Postmultiplication of [5.7.16] by $\Delta\mathbf{g}^{(m+1)}$ confirms that [5.7.17] is indeed satisfied by the Davidon-Fletcher-Powell estimate $\mathbf{A}^{(m+1)}$:

$$\begin{aligned}\mathbf{A}^{(m+1)}\,\Delta\mathbf{g}^{(m+1)} &= \mathbf{A}^{(m)}\,\Delta\mathbf{g}^{(m+1)} \\ &\quad - \frac{\mathbf{A}^{(m)}(\Delta\mathbf{g}^{(m+1)})(\Delta\mathbf{g}^{(m+1)})'\mathbf{A}^{(m)}(\Delta\mathbf{g}^{(m+1)})}{(\Delta\mathbf{g}^{(m+1)})'\mathbf{A}^{(m)}(\Delta\mathbf{g}^{(m+1)})} \\ &\quad - \frac{(\Delta\boldsymbol{\theta}^{(m+1)})(\Delta\boldsymbol{\theta}^{(m+1)})'(\Delta\mathbf{g}^{(m+1)})}{(\Delta\mathbf{g}^{(m+1)})'(\Delta\boldsymbol{\theta}^{(m+1)})} \\ &= \mathbf{A}^{(m)}\,\Delta\mathbf{g}^{(m+1)} - \mathbf{A}^{(m)}\,\Delta\mathbf{g}^{(m+1)} - \Delta\boldsymbol{\theta}^{(m+1)} \\ &= -\Delta\boldsymbol{\theta}^{(m+1)}.\end{aligned}$$

Thus, calculation of [5.7.16] produces an estimate of $[\mathbf{H}(\boldsymbol{\theta}^{(m+1)})]^{-1}$ that is consistent with the magnitude of the observed change between $\mathbf{g}(\boldsymbol{\theta}^{(m+1)})$ and $\mathbf{g}(\boldsymbol{\theta}^{(m)})$ given the size of the change between $\boldsymbol{\theta}^{(m+1)}$ and $\boldsymbol{\theta}^{(m)}$.

The following proposition (proved in Appendix 5.A at the end of the chapter) establishes some further useful properties of the updating formula [5.7.16].

Proposition 5.1: (Fletcher and Powell (1963)). Consider $\mathcal{L}(\boldsymbol{\theta})$, where $\mathcal{L}$: $\mathbb{R}^a \to \mathbb{R}^1$ has continuous first derivatives denoted

$$\underset{(a\times 1)}{\mathbf{g}(\boldsymbol{\theta}^{(m)})} = \left.\frac{\partial\mathcal{L}(\boldsymbol{\theta})}{\partial\boldsymbol{\theta}}\right|_{\boldsymbol{\theta}=\boldsymbol{\theta}^{(m)}}.$$

Suppose that some element of $\mathbf{g}(\boldsymbol{\theta}^{(m)})$ is nonzero, and let $\mathbf{A}^{(m)}$ be a positive definite symmetric $(a \times a)$ matrix. Then the following hold.

(a) *There exists a scalar $s > 0$ such that $\mathcal{L}(\boldsymbol{\theta}^{(m+1)}) > \mathcal{L}(\boldsymbol{\theta}^{(m)})$ for*

$$\boldsymbol{\theta}^{(m+1)} = \boldsymbol{\theta}^{(m)} + s\mathbf{A}^{(m)}\mathbf{g}(\boldsymbol{\theta}^{(m)}). \qquad [5.7.18]$$

(b) *If s in [5.7.18] is chosen so as to maximize $\mathcal{L}(\boldsymbol{\theta}^{(m+1)})$, then the first-order conditions for an interior maximum imply that*

$$[\mathbf{g}(\boldsymbol{\theta}^{(m+1)})]'[\boldsymbol{\theta}^{(m+1)} - \boldsymbol{\theta}^{(m)}] = 0. \qquad [5.7.19]$$

(c) *Provided that [5.7.19] holds and that some element of $\mathbf{g}(\boldsymbol{\theta}^{(m+1)}) - \mathbf{g}(\boldsymbol{\theta}^{(m)})$ is nonzero, then $\mathbf{A}^{(m+1)}$ described by [5.7.16] is a positive definite symmetric matrix.*

Result (a) establishes that as long as we are not already at an optimum ($\mathbf{g}(\boldsymbol{\theta}^{(m)}) \neq \mathbf{0}$), there exists a step in the direction suggested by the algorithm that will increase the likelihood further, provided that $\mathbf{A}^{(m)}$ is a positive definite matrix. Result (c) establishes that provided that the iteration is begun with $\mathbf{A}^{(0)}$ a positive definite matrix, then the sequence of matrices $\{\mathbf{A}^{(m)}\}_{m=1}^{N}$ should all be positive definite, meaning that each step of the iteration should increase the likelihood function. A standard procedure is to start the iteration with $\mathbf{A}^{(0)} = \mathbf{I}_a$, the $(a \times a)$ identity matrix.

If the function $\mathcal{L}(\boldsymbol{\theta})$ is exactly quadratic, so that

$$\mathcal{L}(\boldsymbol{\theta}) = \mathcal{L}(\boldsymbol{\theta}^{(0)}) + \mathbf{g}'[\boldsymbol{\theta} - \boldsymbol{\theta}^{(0)}] - \tfrac{1}{2}[\boldsymbol{\theta} - \boldsymbol{\theta}^{(0)}]'\mathbf{H}[\boldsymbol{\theta} - \boldsymbol{\theta}^{(0)}],$$

with $\mathbf{H}$ positive definite, then Fletcher and Powell (1963) showed that iteration on [5.7.15] and [5.7.16] will converge to the true global maximum in a steps:

$$\boldsymbol{\theta}^{(a)} = \hat{\boldsymbol{\theta}}_{MLE} = \boldsymbol{\theta}^{(0)} + \mathbf{H}^{-1}\mathbf{g};$$

and the weighting matrix will converge to the inverse of -1 times the matrix of second derivatives:

$$\mathbf{A}^{(a)} = \mathbf{H}^{-1}.$$

More generally, if $\mathcal{L}(\boldsymbol{\theta})$ is well approximated by a quadratic function, then the Davidon-Fletcher-Powell search procedure should approach the global maximum more quickly than the steepest-ascent method,

$$\boldsymbol{\theta}^{(N)} \cong \hat{\boldsymbol{\theta}}_{MLE}$$

for large N, while $\mathbf{A}^{(m)}$ should converge to the negative of the matrix of second derivatives of the log likelihood function:

$$\mathbf{A}^{(N)} \cong -\left[\left.\frac{\partial^2 \mathcal{L}(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}\, \partial \boldsymbol{\theta}'}\right|_{\boldsymbol{\theta}=\hat{\boldsymbol{\theta}}_{MLE}}\right]^{-1}. \qquad [5.7.20]$$

In practice, however, the approximation in [5.7.20] can be somewhat poor, and it is better to evaluate the matrix of second derivatives numerically for purposes of calculating standard errors, as discussed in Section 5.8.

If the function $\mathcal{L}(\boldsymbol{\theta})$ is not globally concave or if the starting value $\boldsymbol{\theta}^{(0)}$ is far from the true maximum, the Davidon-Fletcher-Powell procedure can do very badly. If problems are encountered, it often helps to try a different starting value $\boldsymbol{\theta}^{(0)}$, to rescale the data or parameters so that the elements of $\boldsymbol{\theta}$ are in comparable units, or to rescale the initial matrix $\mathbf{A}^{(0)}$—for example, by setting

$$\mathbf{A}^{(0)} = (1 \times 10^{-4})\mathbf{I}_a.$$

Other Numerical Optimization Methods

A variety of other modified Newton-Raphson methods are available, which use alternative techniques for updating an estimate of $\mathbf{H}(\boldsymbol{\theta}^{(m)})$ or its inverse. Two of the more popular methods are those of Broyden (1965, 1967) and Berndt, Hall, Hall, and Hausman (1974). Surveys of these and a variety of other approaches are provided by Judge, Griffiths, Hill, and Lee (1980, pp. 719–72) and Quandt (1983).

Obviously, these same methods can be used to minimize a function $Q(\boldsymbol{\theta})$ with respect to $\boldsymbol{\theta}$. We simply multiply the objective function by -1 and then maximize the function $-Q(\boldsymbol{\theta})$.

5.8. *Statistical Inference with Maximum Likelihood Estimation*

The previous section discussed ways to find the maximum likelihood estimate $\hat{\boldsymbol{\theta}}$ given only the numerical ability to evaluate the log likelihood function $\mathcal{L}(\boldsymbol{\theta})$. This section summarizes general approaches that can be used to test a hypothesis about $\boldsymbol{\theta}$. The section merely summarizes a number of useful results without providing any proofs. We will return to these issues in more depth in Chapter 14, where the statistical foundation behind many of these claims will be developed.

Before detailing these results, however, it is worth calling attention to two of the key assumptions behind the formulas presented in this section. First, it is assumed that the observed data are strictly stationary. Second, it is assumed that neither the estimate $\hat{\boldsymbol{\theta}}$ nor the true value $\boldsymbol{\theta}_0$ falls on a boundary of the allowable parameter space. For example, suppose that the first element of $\boldsymbol{\theta}$ is a parameter corresponding to the probability of a particular event, which must be between 0 and 1. If the event did not occur in the sample, the maximum likelihood estimate of the probability might be zero. This is an example where the estimate $\hat{\boldsymbol{\theta}}$ falls on the boundary of the allowable parameter space, in which case the formulas presented in this section will not be valid.

Asymptotic Standard Errors for Maximum Likelihood Estimates

If the sample size T is sufficiently large, it often turns out that the distribution of the maximum likelihood estimate $\hat{\boldsymbol{\theta}}$ can be well approximated by the following distribution:

$$\hat{\boldsymbol{\theta}} \approx N(\boldsymbol{\theta}_0, T^{-1}\mathcal{I}^{-1}), \qquad [5.8.1]$$

where $\boldsymbol{\theta}_0$ denotes the true parameter vector. The matrix $\mathcal{I}$ is known as the *information matrix* and can be estimated in either of two ways.

The *second-derivative estimate* of the information matrix is

$$\hat{\mathcal{I}}_{2D} = -T^{-1}\left.\frac{\partial^2\mathcal{L}(\boldsymbol{\theta})}{\partial\boldsymbol{\theta}\,\partial\boldsymbol{\theta}'}\right|_{\boldsymbol{\theta}=\hat{\boldsymbol{\theta}}}. \qquad [5.8.2]$$

Here $\mathcal{L}(\boldsymbol{\theta})$ denotes the log likelihood:

$$\mathcal{L}(\boldsymbol{\theta}) = \sum_{t=1}^{T} \log f_{Y_t|\mathcal{Y}_{t-1}}(y_t|\mathcal{Y}_{t-1};\, \boldsymbol{\theta});$$

and $\mathcal{Y}_t$ denotes the history of observations on y obtained through date t. The matrix of second derivatives of the log likelihood is often calculated numerically. Substituting [5.8.2] into [5.8.1], the terms involving the sample size T cancel out so that the variance-covariance matrix of $\hat{\boldsymbol{\theta}}$ can be approximated by

$$E(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0)(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0)' \cong \left[-\left.\frac{\partial^2\mathcal{L}(\boldsymbol{\theta})}{\partial\boldsymbol{\theta}\,\partial\boldsymbol{\theta}'}\right|_{\boldsymbol{\theta}=\hat{\boldsymbol{\theta}}}\right]^{-1}. \qquad [5.8.3]$$

A second estimate of the information matrix $\mathcal{I}$ in [5.8.1] is called the *outer-product estimate*:

$$\hat{\mathcal{I}}_{OP} = T^{-1}\sum_{t=1}^{T} [\mathbf{h}(\hat{\boldsymbol{\theta}}, \mathcal{Y}_t)] \cdot [\mathbf{h}(\hat{\boldsymbol{\theta}}, \mathcal{Y}_t)]'. \qquad [5.8.4]$$

Here $\mathbf{h}(\hat{\boldsymbol{\theta}}, \mathcal{Y}_t)$ denotes the $(a \times 1)$ vector of derivatives of the log of the conditional density of the tth observation with respect to the a elements of the parameter vector $\boldsymbol{\theta}$, with this derivative evaluated at the maximum likelihood estimate $\hat{\boldsymbol{\theta}}$:

$$\mathbf{h}(\hat{\boldsymbol{\theta}}, \mathcal{Y}_t) = \left.\frac{\partial \log f(y_t|y_{t-1}, y_{t-2}, \ldots;\, \boldsymbol{\theta})}{\partial\boldsymbol{\theta}}\right|_{\boldsymbol{\theta}=\hat{\boldsymbol{\theta}}}.$$

In this case, the variance-covariance matrix of $\hat{\boldsymbol{\theta}}$ is approximated by

$$E(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0)(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0)' \cong \left[\sum_{t=1}^{T} [\mathbf{h}(\hat{\boldsymbol{\theta}}, \mathcal{Y}_t)] \cdot [\mathbf{h}(\hat{\boldsymbol{\theta}}, \mathcal{Y}_t)]'\right]^{-1}.$$

As an illustration of how such approximations can be used, suppose that the log likelihood is given by expression [5.7.6]. For this case, one can see analytically that

$$\frac{\partial^2\mathcal{L}(\boldsymbol{\theta})}{\partial\boldsymbol{\theta}\,\partial\boldsymbol{\theta}'} = \begin{bmatrix} -3 & 0 \\ 0 & -4 \end{bmatrix},$$

and so result [5.8.3] suggests that the variance of the maximum likelihood estimate $\hat{\theta}_2$ can be approximated by $\frac{1}{4}$. The *MLE* for this example was $\hat{\theta}_2 = 0$. Thus an

approximate 95% confidence interval for θ_2 is given by

$$0 \pm 2\sqrt{\tfrac{1}{4}} = \pm 1.$$

Note that unless the off-diagonal elements of $\hat{\mathcal{I}}$ are zero, in general one needs to calculate all the elements of the matrix $\hat{\mathcal{I}}$ and invert this full matrix in order to obtain a standard error for any given parameter.

Which estimate of the information matrix, $\hat{\mathcal{I}}_{2D}$ or $\hat{\mathcal{I}}_{OP}$, is it better to use in practice? Expression [5.8.1] is only an approximation to the true distribution of $\hat{\boldsymbol{\theta}}$, and $\hat{\mathcal{I}}_{2D}$ and $\hat{\mathcal{I}}_{OP}$ are in turn only approximations to the true value of $\mathcal{I}$. The theory that justifies these approximations does not give any clear guidance to which is better to use, and typically, researchers rely on whichever estimate of the information matrix is easiest to calculate. If the two estimates differ a great deal, this may mean that the model is misspecified. White (1982) developed a general test of model specification based on this idea. One option for constructing standard errors when the two estimates differ significantly is to use the "quasi-maximum likelihood" standard errors discussed at the end of this section.

Likelihood Ratio Test

Another popular approach to testing hypotheses about parameters that are estimated by maximum likelihood is the *likelihood ratio test*. Suppose a null hypothesis implies a set of m different restrictions on the value of the $(a \times 1)$ parameter vector $\boldsymbol{\theta}$. First, we maximize the likelihood function ignoring these restrictions to obtain the unrestricted maximum likelihood estimate $\hat{\boldsymbol{\theta}}$. Next, we find an estimate $\tilde{\boldsymbol{\theta}}$ that makes the likelihood as large as possible while still satisfying all the restrictions. In practice, this is usually achieved by defining a new $[(a - m) \times 1]$ vector $\boldsymbol{\lambda}$ in terms of which all of the elements of $\boldsymbol{\theta}$ can be expressed when the restrictions are satisfied. For example, if the restriction is that the last m elements of $\boldsymbol{\theta}$ are zero, then $\boldsymbol{\lambda}$ consists of the first $a - m$ elements of $\boldsymbol{\theta}$. Let $\mathcal{L}(\hat{\boldsymbol{\theta}})$ denote the value of the log likelihood function at the unrestricted estimate, and let $\mathcal{L}(\tilde{\boldsymbol{\theta}})$ denote the value of the log likelihood function at the restricted estimate. Clearly $\mathcal{L}(\hat{\boldsymbol{\theta}}) > \mathcal{L}(\tilde{\boldsymbol{\theta}})$, and it often proves to be the case that

$$2[\mathcal{L}(\hat{\boldsymbol{\theta}}) - \mathcal{L}(\tilde{\boldsymbol{\theta}})] \approx \chi^2(m). \qquad [5.8.5]$$

For example, suppose that $a = 2$ and we are interested in testing the hypothesis that $\theta_2 = \theta_1 + 1$. Under this null hypothesis the vector $(\theta_1, \theta_2)'$ can be written as $(\lambda, \lambda + 1)'$, where $\lambda = \theta_1$. Suppose that the log likelihood is given by expression [5.7.6]. One can find the restricted *MLE* by replacing θ_2 by $\theta_1 + 1$ and maximizing the resulting expression with respect to θ_1:

$$\mathcal{L}(\theta_1) = -1.5\theta_1^2 - 2(\theta_1 + 1)^2.$$

The first-order condition for maximization of $\mathcal{L}(\theta_1)$ is

$$-3\theta_1 - 4(\theta_1 + 1) = 0,$$

or $\theta_1 = -\frac{4}{7}$. The restricted *MLE* is thus $\tilde{\boldsymbol{\theta}} = (-\frac{4}{7}, \frac{3}{7})'$, and the maximum value attained for the log likelihood while satisfying the restriction is

$$\begin{aligned}\mathcal{L}(\tilde{\boldsymbol{\theta}}) &= (-\tfrac{3}{2})(-\tfrac{4}{7})^2 - (\tfrac{4}{2})(\tfrac{3}{7})^2 \\ &= -\{(3 \cdot 4)/(2 \cdot 7 \cdot 7)\}\{4 + 3\} \\ &= -\tfrac{6}{7}.\end{aligned}$$

The unrestricted *MLE* is $\hat{\boldsymbol{\theta}} = \mathbf{0}$, at which $\mathcal{L}(\hat{\boldsymbol{\theta}}) = 0$. Hence, [5.8.5] would be

$$2[\mathcal{L}(\hat{\boldsymbol{\theta}}) - \mathcal{L}(\tilde{\boldsymbol{\theta}})] = \tfrac{12}{7} = 1.71.$$

The test here involves a single restriction, so $m = 1$. From Table B.2 in Appendix B, the probability that a $\chi^2(1)$ variable exceeds 3.84 is 0.05. Since $1.71 < 3.84$, we accept the null hypothesis that $\theta_2 = \theta_1 + 1$ at the 5% significance level.

Lagrange Multiplier Test

In order to use the standard errors from [5.8.2] or [5.8.4] to test a hypothesis about $\boldsymbol{\theta}$, we need only to find the unrestricted *MLE* $\hat{\boldsymbol{\theta}}$. In order to use the likelihood ratio test [5.8.5], it is necessary to find both the unrestricted *MLE* $\hat{\boldsymbol{\theta}}$ and the restricted *MLE* $\tilde{\boldsymbol{\theta}}$. The *Lagrange multiplier test* provides a third principle with which to test a null hypothesis that requires only the restricted *MLE* $\tilde{\boldsymbol{\theta}}$. This test is useful when it is easier to calculate the restricted estimate $\tilde{\boldsymbol{\theta}}$ than the unrestricted estimate $\hat{\boldsymbol{\theta}}$.

Let $\boldsymbol{\theta}$ be an $(a \times 1)$ vector of parameters, and let $\tilde{\boldsymbol{\theta}}$ be an estimate of $\boldsymbol{\theta}$ that maximizes the log likelihood subject to a set of m restrictions on $\boldsymbol{\theta}$. Let $f(y_t | y_{t-1}, y_{t-2}, \ldots ; \boldsymbol{\theta})$ be the conditional density of the tth observation, and let $\mathbf{h}(\tilde{\boldsymbol{\theta}}, \mathcal{Y}_t)$ denote the $(a \times 1)$ vector of derivatives of the log of this conditional density evaluated at the restricted estimate $\tilde{\boldsymbol{\theta}}$:

$$\mathbf{h}(\tilde{\boldsymbol{\theta}}, \mathcal{Y}_t) = \left. \frac{\partial \log f(y_t | y_{t-1}, y_{t-2}, \ldots ; \boldsymbol{\theta})}{\partial \boldsymbol{\theta}} \right|_{\boldsymbol{\theta} = \tilde{\boldsymbol{\theta}}}.$$

The Lagrange multiplier test of the null hypothesis that the restrictions are true is given by the following statistic:

$$T^{-1} \left[\sum_{t=1}^{T} \mathbf{h}(\tilde{\boldsymbol{\theta}}, \mathcal{Y}_t) \right]' \mathcal{I}^{-1} \left[\sum_{t=1}^{T} \mathbf{h}(\tilde{\boldsymbol{\theta}}, \mathcal{Y}_t) \right]. \qquad [5.8.6]$$

If the null hypothesis is true, then for large T this should approximately have a $\chi^2(m)$ distribution. The information matrix $\mathcal{I}$ can again be estimated as in [5.8.2] or [5.8.4] with $\hat{\boldsymbol{\theta}}$ replaced by $\tilde{\boldsymbol{\theta}}$.

Quasi-Maximum Likelihood Standard Errors

It was mentioned earlier in this section that if the data were really generated from the assumed density and the sample size is sufficiently large, the second-derivative estimate $\hat{\mathcal{I}}_{2D}$ and the outer-product estimate $\hat{\mathcal{I}}_{OP}$ of the information matrix should be reasonably close to each other. However, maximum likelihood estimation may still be a reasonable way to estimate parameters even if the data were not generated by the assumed density. For example, we noted in Section 5.2 that the conditional *MLE* for a Gaussian $AR(1)$ process is obtained from an *OLS* regression of y_t on y_{t-1}. This *OLS* regression is often a very sensible way to estimate parameters of an $AR(1)$ process even if the true innovations ε_t are not i.i.d. Gaussian. Although maximum likelihood may be yielding a reasonable estimate of $\boldsymbol{\theta}$, when the innovations are not i.i.d. Gaussian, the standard errors proposed in [5.8.2] or [5.8.4] may no longer be valid. An approximate variance-covariance matrix for $\hat{\boldsymbol{\theta}}$ that is sometimes valid even if the probability density is misspecified is given by

$$E(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0)(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0)' \cong T^{-1}\{\mathcal{I}_{2D}\mathcal{I}_{OP}^{-1}\mathcal{I}_{2D}\}^{-1}. \qquad [5.8.7]$$

This variance-covariance matrix was proposed by White (1982), who described this approach as *quasi-maximum likelihood estimation*.

5.9. *Inequality Constraints*

A Common Pitfall with Numerical Maximization

Suppose we were to apply one of the methods discussed in Section 5.7 such as steepest ascent to the $AR(1)$ likelihood [5.7.2]. We start with an arbitrary initial guess, say, $\phi = 0.1$. We calculate the gradient at this point, and find that it is positive. The computer is then programmed to try to improve this estimate by evaluating the log likelihood at points described by $\phi^{(1)} = \phi^{(0)} + s \cdot g(\phi^{(0)})$ for various values of s, seeing what works best. But if the computer were to try a value for s such that $\phi^{(1)} = \phi^{(0)} + s \cdot g(\phi^{(0)}) = 1.1$, calculation of [5.7.2] would involve finding the log of $(1 - 1.1^2) = -0.21$. Attempting to calculate the log of a negative number would typically be a fatal execution error, causing the search procedure to crash.

Often such problems can be avoided by using modified Newton-Raphson procedures, provided that the initial estimate $\boldsymbol{\theta}^{(0)}$ is chosen wisely and provided that the initial search area is kept fairly small. The latter might be accomplished by setting the initial weighting matrix $\mathbf{A}^{(0)}$ in [5.7.15] and [5.7.16] equal to a small multiple of the identity matrix, such as $\mathbf{A}^{(0)} = (1 \times 10^{-4}) \cdot \mathbf{I}_a$. In later iterations, the algorithm should use the shape of the likelihood function in the vicinity of the maximum to keep the search conservative. However, if the true MLE is close to one of the boundaries (for example, if $\hat{\phi}_{MLE} = 0.998$ in the $AR(1)$ example), it will be virtually impossible to keep a numerical algorithm from exploring what happens when ϕ is greater than unity, which would induce a fatal crash.

Solving the Problem by Reparameterizing the Likelihood Function

One simple way to ensure that a numerical search always stays within certain specified boundaries is to reparameterize the likelihood function in terms of an $(a \times 1)$ vector $\boldsymbol{\lambda}$ for which $\boldsymbol{\theta} = \mathbf{g}(\boldsymbol{\lambda})$, where the function $\mathbf{g}$: $\mathbb{R}^a \rightarrow \mathbb{R}^a$ incorporates the desired restrictions. The scheme is then as follows:

Input		*Procedure*		*Output*
values of $y_1, y_2, \ldots, y_T$ and $\boldsymbol{\lambda}$	$\rightarrow$	set $\boldsymbol{\theta} = \mathbf{g}(\boldsymbol{\lambda})$; calculate $\mathcal{L}(\boldsymbol{\theta})$	$\rightarrow$	value of $\mathcal{L}(\mathbf{g}(\boldsymbol{\lambda}))$

For example, to ensure that ϕ is always between ± 1, we could take

$$\phi = g(\lambda) = \frac{\lambda}{1 + |\lambda|}. \qquad [5.9.1]$$

The goal is to find the value of λ that produces the biggest value for the log likelihood. We start with an initial guess such as $\lambda = 3$. The procedure to evaluate the log likelihood function first calculates

$$\phi = 3/(1 + 3) = 0.75$$

and then finds the value for the log likelihood associated with this value of ϕ from [5.7.2]. No matter what value for λ the computer guesses, the value of ϕ in [5.9.1] will always be less than 1 in absolute value and the likelihood function will be well

defined. Once we have found the value of $\hat{\lambda}$ that maximizes the likelihood function, the maximum likelihood estimate of ϕ is then given by

$$\hat{\phi} = \frac{\hat{\lambda}}{1 + |\hat{\lambda}|}.$$

This technique of reparameterizing the likelihood function so that estimates always satisfy any necessary constraints is often very easy to implement. However, one note of caution should be mentioned. If a standard error is calculated from the matrix of second derivatives of the log likelihood as in [5.8.3], this represents the standard error of $\hat{\lambda}$, not the standard error of $\hat{\phi}$. To obtain a standard error for $\hat{\phi}$, the best approach is first to parameterize the likelihood function in terms of λ to find the *MLE*, and then to reparameterize in terms of ϕ to calculate the matrix of second derivatives evaluated at $\hat{\phi}$ to get the final standard error for $\hat{\phi}$. Alternatively, one can calculate an approximation to the standard error for $\hat{\phi}$ from the standard error for $\hat{\lambda}$, based on the formula for a Wald test of a nonlinear hypothesis described in Chapter 14.

Parameterizations for a Variance-Covariance Matrix

Another common restriction one needs to impose is that a variance parameter σ^2 be positive. An obvious way to achieve this is to parameterize the likelihood in terms of λ which represents ± 1 times the standard deviation. The procedure to evaluate the log likelihood then begins by squaring this parameter λ:

$$\sigma^2 = \lambda^2;$$

and if the standard deviation σ is itself called, it is calculated as

$$\sigma = \sqrt{\lambda^2}.$$

More generally, let $\mathbf{\Omega}$ denote an $(n \times n)$ variance-covariance matrix:

$$\mathbf{\Omega} = \begin{bmatrix} \sigma_{11} & \sigma_{12} & \cdots & \sigma_{1n} \\ \sigma_{21} & \sigma_{22} & \cdots & \sigma_{2n} \\ \vdots & \vdots & \cdots & \vdots \\ \sigma_{n1} & \sigma_{n2} & \cdots & \sigma_{nn} \end{bmatrix}.$$

Here one needs to impose the condition that $\mathbf{\Omega}$ is positive definite and symmetric. The best approach is to parameterize $\mathbf{\Omega}$ in terms of the $n(n + 1)/2$ distinct elements of the Cholesky decomposition of $\mathbf{\Omega}$:

$$\mathbf{\Omega} = \mathbf{PP}', \qquad [5.9.2]$$

where

$$\mathbf{P} = \begin{bmatrix} \lambda_{11} & 0 & 0 & \cdots & 0 \\ \lambda_{21} & \lambda_{22} & 0 & \cdots & 0 \\ \vdots & \vdots & \vdots & \cdots & \vdots \\ \lambda_{n1} & \lambda_{n2} & \lambda_{n3} & \cdots & \lambda_{nn} \end{bmatrix}.$$

No matter what values the computer guesses for $\lambda_{11}, \lambda_{21}, \ldots, \lambda_{nn}$, the matrix $\mathbf{\Omega}$ calculated from [5.9.2] will be symmetric and positive semidefinite.

Parameterizations for Probabilities

Sometimes some of the unknown parameters are probabilities $p_1, p_2, \ldots, p_K$ which must satisfy the restrictions

$$0 \le p_i \le 1 \qquad \text{for } i = 1, 2, \ldots, K$$
$$p_1 + p_2 + \cdots + p_K = 1.$$

In this case, one approach is to parameterize the probabilities in terms of $\lambda_1, \lambda_2, \ldots, \lambda_{K-1}$, where

$$p_i = \lambda_i^2/(1 + \lambda_1^2 + \lambda_2^2 + \cdots + \lambda_{K-1}^2) \qquad \text{for } i = 1, 2, \ldots, K-1$$
$$p_K = 1/(1 + \lambda_1^2 + \lambda_2^2 + \cdots + \lambda_{K-1}^2).$$

More General Inequality Constraints

For more complicated inequality constraints that do not admit a simple reparameterization, an approach that sometimes works is to put a branching statement in the procedure to evaluate the log likelihood function. The procedure first checks whether the constraint is satisfied. If it is, then the likelihood function is evaluated in the usual way. If it is not, then the procedure returns a large negative number in place of the value of the log likelihood function. Sometimes such an approach will allow an *MLE* satisfying the specified conditions to be found with simple numerical search procedures.

If these measures prove inadequate, more complicated algorithms are available. Judge, Griffiths, Hill, and Lee (1980, pp. 747–49) described some of the possible approaches.

APPENDIX 5.A. *Proofs of Chapter 5 Propositions*

■ **Proof of Proposition 5.1.**

(a) By Taylor's theorem,

$$\mathcal{L}(\boldsymbol{\theta}^{(m+1)}) = \mathcal{L}(\boldsymbol{\theta}^{(m)}) + [\mathbf{g}(\boldsymbol{\theta}^{(m)})]'[\boldsymbol{\theta}^{(m+1)} - \boldsymbol{\theta}^{(m)}] + R_1(\boldsymbol{\theta}^{(m)}, \boldsymbol{\theta}^{(m+1)}). \qquad [5.A.1]$$

Substituting [5.7.18] into [5.A.1],

$$\mathcal{L}(\boldsymbol{\theta}^{(m+1)}) - \mathcal{L}(\boldsymbol{\theta}^{(m)}) = [\mathbf{g}(\boldsymbol{\theta}^{(m)})]'s\mathbf{A}^{(m)}\mathbf{g}(\boldsymbol{\theta}^{(m)}) + R_1(\boldsymbol{\theta}^{(m)}, \boldsymbol{\theta}^{(m+1)}). \qquad [5.A.2]$$

Since $\mathbf{A}^{(m)}$ is positive definite and since $\mathbf{g}(\boldsymbol{\theta}^{(m)}) \neq \mathbf{0}$, expression [5.A.2] establishes that

$$\mathcal{L}(\boldsymbol{\theta}^{(m+1)}) - \mathcal{L}(\boldsymbol{\theta}^{(m)}) = s\kappa(\boldsymbol{\theta}^{(m)}) + R_1(\boldsymbol{\theta}^{(m)}, \boldsymbol{\theta}^{(m+1)}),$$

where $\kappa(\boldsymbol{\theta}^{(m)}) > 0$. Moreover, $s^{-1} \cdot R_1(\boldsymbol{\theta}^{(m)}, \boldsymbol{\theta}^{(m+1)}) \to 0$ as $s \to 0$. Hence, there exists an s such that $\mathcal{L}(\boldsymbol{\theta}^{(m+1)}) - \mathcal{L}(\boldsymbol{\theta}^{(m)}) > 0$, as claimed.

(b) Direct differentiation reveals

$$\begin{aligned}\frac{\partial \mathcal{L}(\boldsymbol{\theta}^{(m+1)})}{\partial s} &= \frac{\partial \mathcal{L}}{\partial \theta_1}\frac{\partial \theta_1}{\partial s} + \frac{\partial \mathcal{L}}{\partial \theta_2}\frac{\partial \theta_2}{\partial s} + \cdots + \frac{\partial \mathcal{L}}{\partial \theta_a}\frac{\partial \theta_a}{\partial s} \\ &= [\mathbf{g}(\boldsymbol{\theta}^{(m+1)})]' \frac{\partial \boldsymbol{\theta}^{(m+1)}}{\partial s} \qquad [5.A.3] \\ &= [\mathbf{g}(\boldsymbol{\theta}^{(m+1)})]'\mathbf{A}^{(m)}\mathbf{g}(\boldsymbol{\theta}^{(m)}),\end{aligned}$$

with the last line following from [5.7.18]. The first-order conditions set [5.A.3] equal to zero, which implies

$$0 = [\mathbf{g}(\boldsymbol{\theta}^{(m+1)})]'s\mathbf{A}^{(m)}\mathbf{g}(\boldsymbol{\theta}^{(m)}) = [\mathbf{g}(\boldsymbol{\theta}^{(m+1)})]'[\boldsymbol{\theta}^{(m+1)} - \boldsymbol{\theta}^{(m)}],$$

with the last line again following from [5.7.18]. This establishes the claim in [5.7.19].

(c) Let $\mathbf{y}$ be any $(a \times 1)$ nonzero vector. The task is to show that $\mathbf{y}'\mathbf{A}^{(m+1)}\mathbf{y} > 0$. Observe from [5.7.16] that

$$\mathbf{y}'\mathbf{A}^{(m+1)}\mathbf{y} = \mathbf{y}'\mathbf{A}^{(m)}\mathbf{y} - \frac{\mathbf{y}'\mathbf{A}^{(m)}(\Delta\mathbf{g}^{(m+1)})(\Delta\mathbf{g}^{(m+1)})'\mathbf{A}^{(m)}\mathbf{y}}{(\Delta\mathbf{g}^{(m+1)})'\mathbf{A}^{(m)}(\Delta\mathbf{g}^{(m+1)})} - \frac{\mathbf{y}'(\Delta\boldsymbol{\theta}^{(m+1)})(\Delta\boldsymbol{\theta}^{(m+1)})'\mathbf{y}}{(\Delta\mathbf{g}^{(m+1)})'(\Delta\boldsymbol{\theta}^{(m+1)})}. \qquad [5.A.4]$$

Since $\mathbf{A}^{(m)}$ is positive definite, there exists a nonsingular matrix $\mathbf{P}$ such that

$$\mathbf{A}^{(m)} = \mathbf{P}\mathbf{P}'.$$

Define

$$\mathbf{y}^* \equiv \mathbf{P}'\mathbf{y}$$
$$\mathbf{x}^* \equiv \mathbf{P}'\Delta\mathbf{g}^{(m+1)}.$$

Then [5.A.4] can be written as

$$\begin{aligned}\mathbf{y}'\mathbf{A}^{(m+1)}\mathbf{y} &= \mathbf{y}'\mathbf{P}\mathbf{P}'\mathbf{y} - \frac{\mathbf{y}'\mathbf{P}\mathbf{P}'(\Delta\mathbf{g}^{(m+1)})(\Delta\mathbf{g}^{(m+1)})'\mathbf{P}\mathbf{P}'\mathbf{y}}{(\Delta\mathbf{g}^{(m+1)})'\mathbf{P}\mathbf{P}'(\Delta\mathbf{g}^{(m+1)})} \\ &\quad - \frac{\mathbf{y}'(\Delta\boldsymbol{\theta}^{(m+1)})(\Delta\boldsymbol{\theta}^{(m+1)})'\mathbf{y}}{(\Delta\mathbf{g}^{(m+1)})'(\Delta\boldsymbol{\theta}^{(m+1)})} \qquad [5.A.5] \\ &= \mathbf{y}^{*\prime}\mathbf{y}^* - \frac{(\mathbf{y}^{*\prime}\mathbf{x}^*)(\mathbf{x}^{*\prime}\mathbf{y}^*)}{\mathbf{x}^{*\prime}\mathbf{x}^*} - \frac{\mathbf{y}'(\Delta\boldsymbol{\theta}^{(m+1)})(\Delta\boldsymbol{\theta}^{(m+1)})'\mathbf{y}}{(\Delta\mathbf{g}^{(m+1)})'(\Delta\boldsymbol{\theta}^{(m+1)})}.\end{aligned}$$

Recalling equation [4.A.6], the first two terms in the last line of [5.A.5] represent the sum of squared residuals from an *OLS* regression of $\mathbf{y}^*$ on $\mathbf{x}^*$. This cannot be negative,

$$\mathbf{y}^{*\prime}\mathbf{y}^* - \frac{(\mathbf{y}^{*\prime}\mathbf{x}^*)(\mathbf{x}^{*\prime}\mathbf{y}^*)}{\mathbf{x}^{*\prime}\mathbf{x}^*} \geq 0; \qquad [5.A.6]$$

it would equal zero only if the *OLS* regression has a perfect fit, or if $\mathbf{y}^* = \beta\mathbf{x}^*$ or $\mathbf{P}'\mathbf{y} = \beta\mathbf{P}'\Delta\mathbf{g}^{(m+1)}$ for some β. Since $\mathbf{P}$ is nonsingular, expression [5.A.6] would only be zero if $\mathbf{y} = \beta\Delta\mathbf{g}^{(m+1)}$ for some β. Consider two cases.

Case 1. There is no β such that $\mathbf{y} = \beta\Delta\mathbf{g}^{(m+1)}$. In this case, the inequality [5.A.6] is strict and [5.A.5] implies

$$\mathbf{y}'\mathbf{A}^{(m+1)}\mathbf{y} > -\frac{[\mathbf{y}'\Delta\boldsymbol{\theta}^{(m+1)}]^2}{(\Delta\mathbf{g}^{(m+1)})'(\Delta\boldsymbol{\theta}^{(m+1)})}.$$

Since $[\mathbf{y}'\Delta\boldsymbol{\theta}^{(m+1)}]^2 \geq 0$, it follows that $\mathbf{y}'\mathbf{A}^{(m+1)}\mathbf{y} > 0$, provided that

$$(\Delta\mathbf{g}^{(m+1)})'(\Delta\boldsymbol{\theta}^{(m+1)}) < 0. \qquad [5.A.7]$$

But, from [5.7.19],

$$\begin{aligned}(\Delta\mathbf{g}^{(m+1)})'(\Delta\boldsymbol{\theta}^{(m+1)}) &= [\mathbf{g}(\boldsymbol{\theta}^{(m+1)}) - \mathbf{g}(\boldsymbol{\theta}^{(m)})]'(\Delta\boldsymbol{\theta}^{(m+1)}) \\ &= -\mathbf{g}(\boldsymbol{\theta}^{(m)})'(\Delta\boldsymbol{\theta}^{(m+1)}) \qquad [5.A.8] \\ &= -\mathbf{g}(\boldsymbol{\theta}^{(m)})'s\mathbf{A}^{(m)}\mathbf{g}(\boldsymbol{\theta}^{(m)}),\end{aligned}$$

with the last line following from [5.7.18]. But the final term in [5.A.8] must be negative, by virtue of the facts that $\mathbf{A}^{(m)}$ is positive definite, $s > 0$, and $\mathbf{g}(\boldsymbol{\theta}^{(m)}) \neq \mathbf{0}$. Hence, [5.A.7] holds, meaning that $\mathbf{A}^{(m+1)}$ is positive definite for this case.

Case 2. There exists a β such that $\mathbf{y} = \beta\Delta\mathbf{g}^{(m+1)}$. In this case, [5.A.6] is zero, so that [5.A.5] becomes

$$\begin{aligned}\mathbf{y}'\mathbf{A}^{(m+1)}\mathbf{y} &= -\frac{\mathbf{y}'(\Delta\boldsymbol{\theta}^{(m+1)})(\Delta\boldsymbol{\theta}^{(m+1)})'\mathbf{y}}{(\Delta\mathbf{g}^{(m+1)})'(\Delta\boldsymbol{\theta}^{(m+1)})} \\ &= -\frac{\beta(\Delta\mathbf{g}^{(m+1)})'(\Delta\boldsymbol{\theta}^{(m+1)})(\Delta\boldsymbol{\theta}^{(m+1)})'\beta(\Delta\mathbf{g}^{(m+1)})}{(\Delta\mathbf{g}^{(m+1)})'(\Delta\boldsymbol{\theta}^{(m+1)})} \\ &= -\beta^2(\Delta\mathbf{g}^{(m+1)})'(\Delta\boldsymbol{\theta}^{(m+1)}) = \beta^2\mathbf{g}(\boldsymbol{\theta}^{(m)})'s\mathbf{A}^{(m)}\mathbf{g}(\boldsymbol{\theta}^{(m)}) > 0,\end{aligned}$$

as in [5.A.8]. ■

Chapter 5 Exercises

5.1. Show that the value of [5.4.16] at $\theta = \tilde{\theta}$, $\sigma^2 = \tilde{\sigma}^2$ is identical to its value at $\theta = \tilde{\theta}^{-1}$, $\sigma^2 = \tilde{\theta}^2\tilde{\sigma}^2$.

5.2. Verify that expression [5.7.12] calculates the maximum of [5.7.6] in a single step from the initial estimate $\theta^{(0)} = (-1, 1)'$.

5.3. Let $(y_1, y_2, \ldots, y_T)$ be a sample of size T drawn from an i.i.d. $N(\mu, \sigma^2)$ distribution.

(a) Show that the maximum likelihood estimates are given by

$$\hat{\mu} = T^{-1}\sum_{t=1}^{T} y_t$$

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

(b) Show that the matrix $\hat{\mathcal{J}}_{2D}$ in [5.8.2] is

$$\hat{\mathcal{J}}_{2D} = \begin{bmatrix} 1/\hat{\sigma}^2 & 0 \\ 0 & 1/(2\hat{\sigma}^4) \end{bmatrix}.$$

(c) Show that for this example result [5.8.1] suggests

$$\begin{bmatrix} \hat{\mu} \\ \hat{\sigma}^2 \end{bmatrix} \approx N\left(\begin{bmatrix} \mu \\ \sigma^2 \end{bmatrix}, \begin{bmatrix} \hat{\sigma}^2/T & 0 \\ 0 & 2\hat{\sigma}^4/T \end{bmatrix}\right).$$

Chapter 5 References

Anderson, Brian D. O., and John B. Moore. 1979. *Optimal Filtering*. Englewood Cliffs, N.J.: Prentice-Hall.

Berndt, E. K., B. H. Hall, R. E. Hall, and J. A. Hausman. 1974. "Estimation and Inference in Nonlinear Structural Models." *Annals of Economic and Social Measurement* 3:653–65.

Box, George E. P., and D. R. Cox. 1964. "An Analysis of Transformations." *Journal of the Royal Statistical Society* Series B, 26:211–52.

——— and Gwilym M. Jenkins. 1976. *Time Series Analysis: Forecasting and Control*, rev. ed. San Francisco: Holden-Day.

Broyden, C. G. 1965. "A Class of Methods for Solving Nonlinear Simultaneous Equations." *Mathematics of Computation* 19:577–93.

———. 1967. "Quasi-Newton Methods and Their Application to Function Minimization." *Mathematics of Computation* 21:368–81.

Chiang, Alpha C. 1974. *Fundamental Methods of Mathematical Economics*, 2d ed. New York: McGraw-Hill.

Davidon, W. C. 1959. "Variable Metric Method of Minimization." A.E.C. Research and Development Report ANL-5990 (rev.).

Fletcher, R., and M. J. D. Powell. 1963. "A Rapidly Convergent Descent Method for Minimization." *Computer Journal* 6:163–68.

Galbraith, R. F., and J. I. Galbraith. 1974. "On the Inverses of Some Patterned Matrices Arising in the Theory of Stationary Time Series." *Journal of Applied Probability* 11:63–71.

Hannan, E., and J. Rissanen. 1982. "Recursive Estimation of Mixed Autoregressive–Moving Average Order." *Biometrika* 69:81–94.

Janacek, G. J., and A. L. Swift. 1990. "A Class of Models for Non-Normal Time Series." *Journal of Time Series Analysis* 11:19–31.

Judge, George G., William E. Griffiths, R. Carter Hill, and Tsoung-Chao Lee. 1980. *The Theory and Practice of Econometrics*. New York: Wiley.

Koreisha, Sergio, and Tarmo Pukkila. 1989. "Fast Linear Estimation Methods for Vector Autoregressive Moving-Average Models." *Journal of Time Series Analysis* 10:325–39.

Li, W. K., and A. I. McLeod. 1988. "ARMA Modelling with Non-Gaussian Innovations." *Journal of Time Series Analysis* 9:155–68.

Martin, R. D. 1981. "Robust Methods for Time Series," in D. F. Findley, ed., *Applied Time Series*, Vol. II. New York: Academic Press.

Nelson, Harold L., and C. W. J. Granger. 1979. "Experience with Using the Box-Cox Transformation When Forecasting Economic Time Series." *Journal of Econometrics* 10:57–69.

Quandt, Richard E. 1983. "Computational Problems and Methods," in Zvi Griliches and Michael D. Intriligator, eds., *Handbook of Econometrics*, Vol. 1. Amsterdam: North-Holland.

White, Halbert. 1982. "Maximum Likelihood Estimation of Misspecified Models." *Econometrica* 50:1–25.

6 Spectral Analysis

Up to this point in the book, the value of a variable Y_t at date t has typically been described in terms of a sequence of innovations $\{\varepsilon_t\}_{t=-\infty}^{\infty}$ in models of the form

$$Y_t = \mu + \sum_{j=0}^{\infty} \psi_j \varepsilon_{t-j}.$$

The focus has been on the implications of such a representation for the covariance between Y_t and Y_τ at distinct dates t and τ. This is known as analyzing the properties of $\{Y_t\}_{t=-\infty}^{\infty}$ in the *time domain*.

This chapter instead describes the value of Y_t as a weighted sum of periodic functions of the form $\cos(\omega t)$ and $\sin(\omega t)$, where ω denotes a particular frequency:

$$Y_t = \mu + \int_0^{\pi} \alpha(\omega)\cdot\cos(\omega t)\, d\omega + \int_0^{\pi} \delta(\omega)\cdot\sin(\omega t)\, d\omega.$$

The goal will be to determine how important cycles of different frequencies are in accounting for the behavior of Y. This is known as *frequency-domain* or *spectral* analysis. As we will see, the two kinds of analysis are not mutually exclusive. Any covariance-stationary process has both a time-domain representation and a frequency-domain representation, and any feature of the data that can be described by one representation can equally well be described by the other representation. For some features, the time-domain description may be simpler, while for other features the frequency-domain description may be simpler.

Section 6.1 describes the properties of the population spectrum and introduces the spectral representation theorem, which can be viewed as a frequency-domain version of Wold's theorem. Section 6.2 introduces the sample analog of the population spectrum and uses an *OLS* regression framework to motivate the spectral representation theorem and to explain the sense in which the spectrum identifies the contributions to the variance of the observed data of periodic components with different cycles. Section 6.3 discusses strategies for estimating the population spectrum. Section 6.4 provides an example of applying spectral techniques and discusses some of the ways they can be used in practice. More detailed discussions of spectral analysis are provided by Anderson (1971), Bloomfield (1976), and Fuller (1976).

6.1. *The Population Spectrum*

The Population Spectrum and Its Properties

Let $\{Y_t\}_{t=-\infty}^{\infty}$ be a covariance-stationary process with mean $E(Y_t) = \mu$ and jth autocovariance

$$E(Y_t - \mu)(Y_{t-j} - \mu) = \gamma_j.$$

Assuming that these autocovariances are absolutely summable, the autocovariance-generating function is given by

$$g_Y(z) \equiv \sum_{j=-\infty}^{\infty} \gamma_j z^j, \qquad [6.1.1]$$

where z denotes a complex scalar. If [6.1.1] is divided by 2π and evaluated at some z represented by $z = e^{-i\omega}$ for $i = \sqrt{-1}$ and ω a real scalar, the result is called the population spectrum of Y:

$$s_Y(\omega) = \frac{1}{2\pi} g_Y(e^{-i\omega}) = \frac{1}{2\pi} \sum_{j=-\infty}^{\infty} \gamma_j e^{-i\omega j}. \qquad [6.1.2]$$

Note that the spectrum is a function of ω: given any particular value of ω and a sequence of autocovariances $\{\gamma_j\}_{j=-\infty}^{\infty}$, we could in principle calculate the value of $s_Y(\omega)$.

De Moivre's theorem allows us to write $e^{-i\omega j}$ as

$$e^{-i\omega j} = \cos(\omega j) - i\cdot\sin(\omega j). \qquad [6.1.3]$$

Substituting [6.1.3] into [6.1.2], it appears that the spectrum can equivalently be written

$$s_Y(\omega) = \frac{1}{2\pi} \sum_{j=-\infty}^{\infty} \gamma_j[\cos(\omega j) - i\cdot\sin(\omega j)]. \qquad [6.1.4]$$

Note that for a covariance-stationary process, $\gamma_j = \gamma_{-j}$. Hence, [6.1.4] implies

$$\begin{aligned} s_Y(\omega) = {} & \frac{1}{2\pi}\gamma_0[\cos(0) - i\cdot\sin(0)] \\ & + \frac{1}{2\pi}\left\{\sum_{j=1}^{\infty} \gamma_j[\cos(\omega j) + \cos(-\omega j) - i\cdot\sin(\omega j) - i\cdot\sin(-\omega j)]\right\}. \end{aligned} \qquad [6.1.5]$$

Next, we make use of the following results from trigonometry:[1]

$$\begin{aligned} \cos(0) &= 1 \\ \sin(0) &= 0 \\ \sin(-\theta) &= -\sin(\theta) \\ \cos(-\theta) &= \cos(\theta). \end{aligned}$$

Using these relations, [6.1.5] simplifies to

$$s_Y(\omega) = \frac{1}{2\pi}\left\{\gamma_0 + 2\sum_{j=1}^{\infty} \gamma_j \cos(\omega j)\right\}. \qquad [6.1.6]$$

Assuming that the sequence of autocovariances $\{\gamma_j\}_{j=-\infty}^{\infty}$ is absolutely summable, expression [6.1.6] implies that the population spectrum exists and that $s_Y(\omega)$ is a continuous, real-valued function of ω. It is possible to go a bit further and show that if the γ_j's represent autocovariances of a covariance-stationary process, then $s_Y(\omega)$ will be nonnegative for all ω.[2] Since $\cos(\omega j) = \cos(-\omega j)$ for any ω, the spectrum is symmetric around $\omega = 0$. Finally, since $\cos[(\omega + 2\pi k)\cdot j] = \cos(\omega j)$ for any integers k and j, it follows from [6.1.6] that $s_Y(\omega + 2\pi k) = s_Y(\omega)$ for any integer k. Hence, the spectrum is a periodic function of ω. If we know the value of $s_Y(\omega)$ for all ω between 0 and π, we can infer the value of $s_Y(\omega)$ for any ω.

[1]These are reviewed in Section A.1 of the Mathematical Review (Appendix A) at the end of the book.

[2]See, for example, Fuller (1976, p. 110).

Calculating the Population Spectrum for Various Processes

Let Y_t follow an $MA(\infty)$ process:

$$Y_t = \mu + \psi(L)\varepsilon_t, \qquad [6.1.7]$$

where

$$\psi(L) = \sum_{j=0}^{\infty} \psi_j L^j$$

$$\sum_{j=0}^{\infty} |\psi_j| < \infty$$

$$E(\varepsilon_t \varepsilon_\tau) = \begin{cases} \sigma^2 & \text{for } t = \tau \\ 0 & \text{otherwise.} \end{cases}$$

Recall from expression [3.6.8] that the autocovariance-generating function for Y is given by

$$g_Y(z) = \sigma^2 \psi(z)\psi(z^{-1}).$$

Hence, from [6.1.2], the population spectrum for an $MA(\infty)$ process is given by

$$s_Y(\omega) = (2\pi)^{-1} \cdot \sigma^2 \psi(e^{-i\omega})\psi(e^{i\omega}). \qquad [6.1.8]$$

For example, for a white noise process, $\psi(z) = 1$ and the population spectrum is a constant for all ω:

$$s_Y(\omega) = \sigma^2/2\pi. \qquad [6.1.9]$$

Next, consider an $MA(1)$ process:

$$Y_t = \varepsilon_t + \theta\varepsilon_{t-1}.$$

Here, $\psi(z) = 1 + \theta z$ and the population spectrum is

$$\begin{aligned} s_Y(\omega) &= (2\pi)^{-1} \cdot \sigma^2 (1 + \theta e^{-i\omega})(1 + \theta e^{i\omega}) \\ &= (2\pi)^{-1} \cdot \sigma^2 (1 + \theta e^{-i\omega} + \theta e^{i\omega} + \theta^2). \end{aligned} \qquad [6.1.10]$$

But notice that

$$e^{-i\omega} + e^{i\omega} = \cos(\omega) - i \cdot \sin(\omega) + \cos(\omega) + i \cdot \sin(\omega) = 2 \cdot \cos(\omega), \qquad [6.1.11]$$

so that [6.1.10] becomes

$$s_Y(\omega) = (2\pi)^{-1} \cdot \sigma^2 [1 + \theta^2 + 2\theta \cdot \cos(\omega)]. \qquad [6.1.12]$$

Recall that $\cos(\omega)$ goes from 1 to -1 as ω goes from 0 to π. Hence, when $\theta > 0$, the spectrum $s_Y(\omega)$ is a monotonically decreasing function of ω for ω in $[0, \pi]$, whereas when $\theta < 0$, the spectrum is monotonically increasing.

For an $AR(1)$ process

$$Y_t = c + \phi Y_{t-1} + \varepsilon_t,$$

we have $\psi(z) = 1/(1 - \phi z)$ as long as $|\phi| < 1$. Thus, the spectrum is

$$\begin{aligned} s_Y(\omega) &= \frac{1}{2\pi} \frac{\sigma^2}{(1 - \phi e^{-i\omega})(1 - \phi e^{i\omega})} \\ &= \frac{1}{2\pi} \frac{\sigma^2}{(1 - \phi e^{-i\omega} - \phi e^{i\omega} + \phi^2)} \\ &= \frac{1}{2\pi} \frac{\sigma^2}{[1 + \phi^2 - 2\phi \cdot \cos(\omega)]}. \end{aligned} \qquad [6.1.13]$$

When $\phi > 0$, the denominator is monotonically increasing in ω over $[0, \pi]$, meaning that $s_Y(\omega)$ is monotonically decreasing. When $\phi < 0$, the spectrum $s_Y(\omega)$ is a monotonically increasing function of ω.

In general, for an $ARMA(p, q)$ process

$$Y_t = c + \phi_1 Y_{t-1} + \phi_2 Y_{t-2} + \cdots + \phi_p Y_{t-p} + \varepsilon_t + \theta_1 \varepsilon_{t-1} + \theta_2 \varepsilon_{t-2} + \cdots + \theta_q \varepsilon_{t-q},$$

the population spectrum is given by

$$s_Y(\omega) = \frac{\sigma^2}{2\pi} \frac{(1 + \theta_1 e^{-i\omega} + \theta_2 e^{-i2\omega} + \cdots + \theta_q e^{-iq\omega})}{(1 - \phi_1 e^{-i\omega} - \phi_2 e^{-i2\omega} - \cdots - \phi_p e^{-ip\omega})} \times \frac{(1 + \theta_1 e^{i\omega} + \theta_2 e^{i2\omega} + \cdots + \theta_q e^{iq\omega})}{(1 - \phi_1 e^{i\omega} - \phi_2 e^{i2\omega} - \cdots - \phi_p e^{ip\omega})}. \qquad [6.1.14]$$

If the moving average and autoregressive polynomials are factored as follows:

$$1 + \theta_1 z + \theta_2 z^2 + \cdots + \theta_q z^q = (1 - \eta_1 z)(1 - \eta_2 z) \cdots (1 - \eta_q z)$$
$$1 - \phi_1 z - \phi_2 z^2 - \cdots - \phi_p z^p = (1 - \lambda_1 z)(1 - \lambda_2 z) \cdots (1 - \lambda_p z),$$

then the spectral density in [6.1.14] can be written

$$s_Y(\omega) = \frac{\sigma^2 \prod_{j=1}^{q} [1 + \eta_j^2 - 2\eta_j \cdot \cos(\omega)]}{2\pi \prod_{j=1}^{p} [1 + \lambda_j^2 - 2\lambda_j \cdot \cos(\omega)]}.$$

Calculating the Autocovariances from the Population Spectrum

If we know the sequence of autocovariances $\{\gamma_j\}_{j=-\infty}^{\infty}$, in principle we can calculate the value of $s_Y(\omega)$ for any ω from [6.1.2] or [6.1.6]. The converse is also true: if we know the value of $s_Y(\omega)$ for all ω in $[0, \pi]$, we can calculate the value of the kth autocovariance γ_k for any given k. This means that the population spectrum $s_Y(\omega)$ and the sequence of autocovariances contain exactly the same information—neither one can tell us anything about the process that is not possible to infer from the other.

The following proposition (proved in Appendix 6.A at the end of this chapter) provides a formula for calculating any autocovariance from the population spectrum.

Proposition 6.1: *Let $\{\gamma_j\}_{j=-\infty}^{\infty}$ be an absolutely summable sequence of autocovariances, and define $s_Y(\omega)$ as in [6.1.2]. Then*

$$\int_{-\pi}^{\pi} s_Y(\omega) e^{i\omega k}\, d\omega = \gamma_k. \qquad [6.1.15]$$

Result [6.1.15] can equivalently be written as

$$\int_{-\pi}^{\pi} s_Y(\omega) \cos(\omega k)\, d\omega = \gamma_k. \qquad [6.1.16]$$

Interpreting the Population Spectrum

The following result obtains as a special case of Proposition 6.1 by setting $k = 0$:

$$\int_{-\pi}^{\pi} s_Y(\omega)\, d\omega = \gamma_0. \qquad [6.1.17]$$

In other words, the area under the population spectrum between $\pm\pi$ gives γ_0, the variance of Y_t.

More generally—since $s_Y(\omega)$ is nonnegative—if we were to calculate

$$\int_{-\omega_1}^{\omega_1} s_Y(\omega)\, d\omega$$

for any ω_1 between 0 and π, the result would be a positive number that we could interpret as the portion of the variance of Y_t that is associated with frequencies ω that are less than ω_1 in absolute value. Recalling that $s_Y(\omega)$ is symmetric, the claim is that

$$2 \cdot \int_{0}^{\omega_1} s_Y(\omega)\, d\omega \qquad [6.1.18]$$

represents the portion of the variance of Y that could be attributed to periodic random components with frequency less than or equal to ω_1.

What does it mean to attribute a certain portion of the variance of Y to cycles with frequency less than or equal to ω_1? To explore this question, let us consider the following rather special stochastic process. Suppose that the value of Y at date t is determined by

$$Y_t = \sum_{j=1}^{M} [\alpha_j \cdot \cos(\omega_j t) + \delta_j \cdot \sin(\omega_j t)]. \qquad [6.1.19]$$

Here, α_j and δ_j are zero-mean random variables, meaning that $E(Y_t) = 0$ for all t. The sequences $\{\alpha_j\}_{j=1}^{M}$ and $\{\delta_j\}_{j=1}^{M}$ are serially uncorrelated and mutually uncorrelated:

$$E(\alpha_j \alpha_k) = \begin{cases} \sigma_j^2 & \text{for } j = k \\ 0 & \text{for } j \neq k \end{cases}$$

$$E(\delta_j \delta_k) = \begin{cases} \sigma_j^2 & \text{for } j = k \\ 0 & \text{for } j \neq k \end{cases}$$

$$E(\alpha_j \delta_k) = 0 \qquad \text{for all } j \text{ and } k.$$

The variance of Y_t is then

$$\begin{aligned} E(Y_t^2) &= \sum_{j=1}^{M} \left[E(\alpha_j^2) \cdot \cos^2(\omega_j t) + E(\delta_j^2) \cdot \sin^2(\omega_j t) \right] \\ &= \sum_{j=1}^{M} \sigma_j^2 \left[\cos^2(\omega_j t) + \sin^2(\omega_j t) \right] \qquad [6.1.20] \\ &= \sum_{j=1}^{M} \sigma_j^2, \end{aligned}$$

with the last line following from equation [A.1.12]. Thus, for this process, the portion of the variance of Y that is due to cycles of frequency ω_j is given by σ_j^2.

If the frequencies are ordered $0 < \omega_1 < \omega_2 < \cdots < \omega_M < \pi$, the portion of the variance of Y that is due to cycles of frequency less than or equal to ω_j is given by $\sigma_1^2 + \sigma_2^2 + \cdots + \sigma_j^2$.

The kth autocovariance of Y is

$$\begin{aligned} E(Y_t Y_{t-k}) &= \sum_{j=1}^{M} \{E(\alpha_j^2)\cdot\cos(\omega_j t)\cdot\cos[\omega_j(t-k)] \\ &\quad + E(\delta_j^2)\cdot\sin(\omega_j t)\cdot\sin[\omega_j(t-k)]\} \\ &= \sum_{j=1}^{M} \sigma_j^2\{\cos(\omega_j t)\cdot\cos[\omega_j(t-k)] \\ &\quad + \sin(\omega_j t)\cdot\sin[\omega_j(t-k)]\}. \end{aligned} \qquad [6.1.21]$$

Recall the trigonometric identity[3]

$$\cos(A - B) = \cos(A)\cdot\cos(B) + \sin(A)\cdot\sin(B). \qquad [6.1.22]$$

For $A = \omega_j t$ and $B = \omega_j(t - k)$, we have $A - B = \omega_j k$, so that [6.1.21] becomes

$$E(Y_t Y_{t-k}) = \sum_{j=1}^{M} \sigma_j^2\cdot\cos(\omega_j k). \qquad [6.1.23]$$

Since the mean and the autocovariances of Y are not functions of time, the process described by [6.1.19] is covariance-stationary, although [6.1.23] implies that the sequence of autocovariances $\{\gamma_k\}_{k=0}^{\infty}$ is not absolutely summable.

We were able to attribute a certain portion of the variance of Y_t to cycles of less than a given frequency for the process in [6.1.19] because that is a rather special covariance-stationary process. However, there is a general result known as the *spectral representation theorem* which says that any covariance-stationary process Y_t can be expressed in terms of a generalization of [6.1.19]. For any fixed frequency ω in $[0, \pi]$, we define random variables $\alpha(\omega)$ and $\delta(\omega)$ and propose to write a stationary process with absolutely summable autocovariances in the form

$$Y_t = \mu + \int_0^{\pi} [\alpha(\omega)\cdot\cos(\omega t) + \delta(\omega)\cdot\sin(\omega t)]\, d\omega.$$

The random processes represented by $\alpha(\cdot)$ and $\delta(\cdot)$ have zero mean and the further properties that for any frequencies $0 < \omega_1 < \omega_2 < \omega_3 < \omega_4 < \pi$, the variable $\int_{\omega_1}^{\omega_2} \alpha(\omega)\, d\omega$ is uncorrelated with $\int_{\omega_3}^{\omega_4} \alpha(\omega)\, d\omega$ and the variable $\int_{\omega_1}^{\omega_2} \delta(\omega)\, d\omega$ is uncorrelated with $\int_{\omega_3}^{\omega_4} \delta(\omega)\, d\omega$, while for any $0 < \omega_1 < \omega_2 < \pi$ and $0 < \omega_3 < \omega_4 < \pi$, the variable $\int_{\omega_1}^{\omega_2} \alpha(\omega)\, d\omega$ is uncorrelated with $\int_{\omega_3}^{\omega_4} \delta(\omega)\, d\omega$. For such a process, one can calculate the portion of the variance of Y_t that is due to cycles with frequency less than or equal to some specified value ω_1 through a generalization of the procedure used to analyze [6.1.19]. Moreover, this magnitude turns out to be given by the expression in [6.1.18].

We shall not attempt a proof of the spectral representation theorem here; for details the reader is referred to Cramér and Leadbetter (1967, pp. 128–38). Instead, the next section provides a formal derivation of a finite-sample version of these results, showing the sense in which the sample analog of [6.1.18] gives the portion of the sample variance of an observed series that can be attributed to cycles with frequencies less than or equal to ω_1.

[3]See, for example, Thomas (1972, p. 176).

6.2. *The Sample Periodogram*

For a covariance-stationary process Y_t with absolutely summable autocovariances, we have defined the value of the population spectrum at frequency ω to be

$$s_Y(\omega) = \frac{1}{2\pi}\sum_{j=-\infty}^{\infty}\gamma_j e^{-i\omega j}, \qquad [6.2.1]$$

where

$$\gamma_j \equiv E(Y_t - \mu)(Y_{t-j} - \mu)$$

and $\mu = E(Y_t)$. Note that the population spectrum is expressed in terms of $\{\gamma_j\}_{j=0}^{\infty}$, which represents population second moments.

Given an observed sample of T observations denoted $y_1, y_2, \ldots, y_T$, we can calculate up to $T - 1$ sample autocovariances from the formulas

$$\hat{\gamma}_j = \begin{cases} T^{-1}\sum_{t=j+1}^{T}(y_t - \bar{y})(y_{t-j} - \bar{y}) & \text{for } j = 0, 1, 2, \ldots, T-1 \\ \hat{\gamma}_{-j} & \text{for } j = -1, -2, \ldots, -T+1, \end{cases} \qquad [6.2.2]$$

where $\bar{y}$ is the sample mean:

$$\bar{y} = T^{-1}\sum_{t=1}^{T} y_t. \qquad [6.2.3]$$

For any given ω we can then construct the sample analog of [6.2.1], which is known as the *sample periodogram*:

$$\hat{s}_y(\omega) = \frac{1}{2\pi}\sum_{j=-T+1}^{T-1}\hat{\gamma}_j e^{-i\omega j}. \qquad [6.2.4]$$

As in [6.1.6], the sample periodogram can equivalently be expressed as

$$\hat{s}_y(\omega) = \frac{1}{2\pi}\left[\hat{\gamma}_0 + 2\sum_{j=1}^{T-1}\hat{\gamma}_j\cos(\omega j)\right]. \qquad [6.2.5]$$

The same calculations that led to [6.1.17] can be used to show that the area under the periodogram is the sample variance of y:

$$\int_{-\pi}^{\pi}\hat{s}_y(\omega)\,d\omega = \hat{\gamma}_0.$$

Like the population spectrum, the sample periodogram is symmetric around $\omega = 0$, so that we could equivalently write

$$\hat{\gamma}_0 = 2\int_0^{\pi}\hat{s}_y(\omega)\,d\omega.$$

There also turns out to be a sample analog to the spectral representation theorem, which we now develop. In particular, we will see that given any T observations on a process $(y_1, y_2, \ldots, y_T)$, there exist frequencies $\omega_1, \omega_2, \ldots, \omega_M$ and coefficients $\hat{\mu}, \hat{\alpha}_1, \hat{\alpha}_2, \ldots, \hat{\alpha}_M, \hat{\delta}_1, \hat{\delta}_2, \ldots, \hat{\delta}_M$ such that the value for y at date t can be expressed as

$$y_t = \hat{\mu} + \sum_{j=1}^{M}\{\hat{\alpha}_j\cdot\cos[\omega_j(t-1)] + \hat{\delta}_j\cdot\sin[\omega_j(t-1)]\}, \qquad [6.2.6]$$

where the variable $\hat{\alpha}_j \cdot \cos[\omega_j(t-1)]$ is orthogonal in the sample to $\hat{\alpha}_k \cdot \cos[\omega_k(t-1)]$ for $j \neq k$, the variable $\hat{\delta}_j \cdot \sin[\omega_j(t-1)]$ is orthogonal to $\hat{\delta}_k \cdot \sin[\omega_k(t-1)]$ for $j \neq k$, and the variable $\hat{\alpha}_j \cdot \cos[\omega_j(t-1)]$ is orthogonal to $\hat{\delta}_k \cdot \sin[\omega_k(t-1)]$ for all j and k. The sample variance of y is $T^{-1}\Sigma_{t=1}^{T}(y_t - \bar{y})^2$, and the portion of this variance that can be attributed to cycles with frequency ω_j can be inferred from the sample periodogram $\hat{s}_y(\omega_j)$.

We will develop this claim for the case when the sample size T is an odd number. In this case y_t will be expressed in terms of periodic functions with $M \equiv (T-1)/2$ different frequencies in [6.2.6]. The frequencies $\omega_1, \omega_2, \ldots, \omega_M$ are specified as follows:

$$\begin{aligned} \omega_1 &= 2\pi/T \\ \omega_2 &= 4\pi/T \\ &\vdots \\ \omega_M &= 2M\pi/T. \end{aligned} \qquad [6.2.7]$$

Thus, the highest frequency considered is

$$\omega_M = \frac{2(T-1)\pi}{2T} < \pi.$$

Consider an *OLS* regression of the value of y_t on a constant and on the various cosine and sine terms,

$$y_t = \mu + \sum_{j=1}^{M} \{\alpha_j \cdot \cos[\omega_j(t-1)] + \delta_j \cdot \sin[\omega_j(t-1)]\} + u_t.$$

This can be viewed as a standard regression model of the form

$$y_t = \boldsymbol{\beta}'\mathbf{x}_t + u_t, \qquad [6.2.8]$$

where

$$\mathbf{x}_t = \Big[1 \quad \cos[\omega_1(t-1)] \quad \sin[\omega_1(t-1)] \quad \cos[\omega_2(t-1)] \quad \sin[\omega_2(t-1)] \quad \cdots \cos[\omega_M(t-1)] \quad \sin[\omega_M(t-1)]\Big]', \qquad [6.2.9]$$

$$\boldsymbol{\beta}' = [\mu \quad \alpha_1 \quad \delta_1 \quad \alpha_2 \quad \delta_2 \quad \cdots \quad \alpha_M \quad \delta_M]. \qquad [6.2.10]$$

Note that $\mathbf{x}_t$ has $(2M+1) = T$ elements, so that there are as many explanatory variables as observations. We will show that the elements of $\mathbf{x}_t$ are linearly independent, meaning that an *OLS* regression of y_t on $\mathbf{x}_t$ yields a perfect fit. Thus, the fitted values for this regression are of the form of [6.2.6] with no error term u_t. Moreover, the coefficients of this regression have the property that $\frac{1}{2}(\hat{\alpha}_j^2 + \hat{\delta}_j^2)$ represents the portion of the sample variance of y that can be attributed to cycles with frequency ω_j. This magnitude $\frac{1}{2}(\hat{\alpha}_j^2 + \hat{\delta}_j^2)$ further turns out to be proportional to the sample periodogram evaluated at ω_j. In other words, any observed series $y_1, y_2, \ldots, y_T$ can be expressed in terms of periodic functions as in [6.2.6], and the portion of the sample variance that is due to cycles with frequency ω_j can be found from the sample periodogram. These points are established formally in the following proposition, which is proved in Appendix 6.A at the end of this chapter.

Proposition 6.2: Let T denote an odd integer and let $M \equiv (T - 1)/2$. Let $\omega_j = 2\pi j/T$ for $j = 1, 2, \ldots, M$, and let $\mathbf{x}_t$ be the $(T \times 1)$ vector in [6.2.9]. Then

$$\sum_{t=1}^{T} \mathbf{x}_t\mathbf{x}_t' = \begin{bmatrix} T & \mathbf{0}' \\ \mathbf{0} & (T/2)\cdot\mathbf{I}_{T-1} \end{bmatrix}. \qquad [6.2.11]$$

Furthermore, let $\{y_1, y_2, \ldots, y_T\}$ be any T numbers. Then the following are true:

(*a*) *The value of y_t can be expressed as*

$$y_t = \hat{\mu} + \sum_{j=1}^{M} \{\hat{\alpha}_j\cdot\cos[\omega_j(t-1)] + \hat{\delta}_j\cdot\sin[\omega_j(t-1)]\},$$

with $\hat{\mu} = \bar{y}$ (the sample mean from [6.2.3]) and

$$\hat{\alpha}_j = (2/T)\sum_{t=1}^{T} y_t\cdot\cos[\omega_j(t-1)] \qquad \text{for } j = 1, 2, \ldots, M \qquad [6.2.12]$$

$$\hat{\delta}_j = (2/T)\sum_{t=1}^{T} y_t\cdot\sin[\omega_j(t-1)] \qquad \text{for } j = 1, 2, \ldots, M. \qquad [6.2.13]$$

(*b*) *The sample variance of y_t can be expressed as*

$$(1/T)\sum_{t=1}^{T}(y_t - \bar{y})^2 = (1/2)\sum_{j=1}^{M}(\hat{\alpha}_j^2 + \hat{\delta}_j^2), \qquad [6.2.14]$$

and the portion of the sample variance of y that can be attributed to cycles of frequency ω_j is given by $\frac{1}{2}(\hat{\alpha}_j^2 + \hat{\delta}_j^2)$.

(*c*) *The portion of the sample variance of y that can be attributed to cycles of frequency ω_j can equivalently be expressed as*

$$(1/2)(\hat{\alpha}_j^2 + \hat{\delta}_j^2) = (4\pi/T)\cdot\hat{s}_y(\omega_j), \qquad [6.2.15]$$

where $\hat{s}_y(\omega_j)$ is the sample periodogram at frequency ω_j.

Result [6.2.11] establishes that $\Sigma_{t=1}^{T}\mathbf{x}_t\mathbf{x}_t'$ is a diagonal matrix, meaning that the explanatory variables contained in $\mathbf{x}_t$ are mutually orthogonal. The proposition asserts that any observed time series $(y_1, y_2, \ldots, y_T)$ with T odd can be written as a constant plus a weighted sum of $(T - 1)$ periodic functions with $(T - 1)/2$ different frequencies; a related result can also be developed when T is an even integer. Hence, the proposition gives a finite-sample analog of the spectral representation theorem. The proposition further shows that the sample periodogram captures the portion of the sample variance of y that can be attributed to cycles of different frequencies.

Note that the frequencies ω_j in terms of which the variance of y is explained all lie in $[0, \pi]$. Why aren't negative frequencies $\omega < 0$ employed as well? Suppose that the data were actually generated by a special case of the process in [6.1.19],

$$Y_t = \alpha\cdot\cos(-\omega t) + \delta\cdot\sin(-\omega t), \qquad [6.2.16]$$

where $-\omega < 0$ represents some particular negative frequency and where α and δ are zero-mean random variables. Since $\cos(-\omega t) = \cos(\omega t)$ and $\sin(-\omega t) = -\sin(\omega t)$, the process [6.2.16] can equivalently be written

$$Y_t = \alpha\cdot\cos(\omega t) - \delta\cdot\sin(\omega t). \qquad [6.2.17]$$

Thus there is no way of using observed data on y to decide whether the data are generated by a cycle with frequency $-\omega$ as in [6.2.16] or by a cycle with frequency

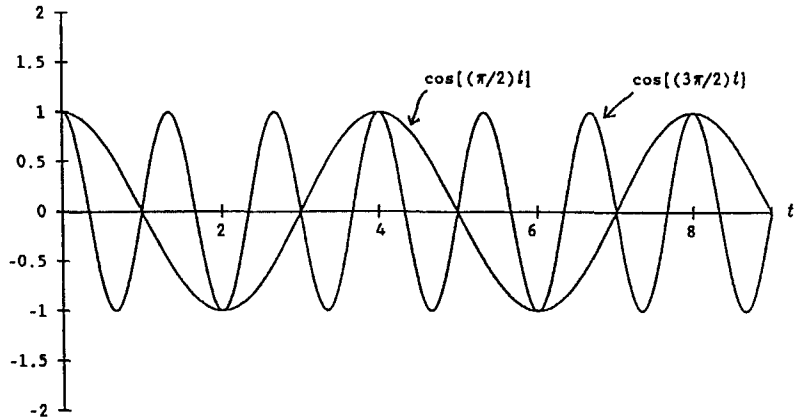


FIGURE 6.1 Aliasing: plots of $\cos[(\pi/2)t]$ and $\cos[(3\pi/2)t]$ as functions of t.

$+\omega$ as in [6.2.17]. It is simply a matter of convention that we choose to focus only on positive frequencies.

Why is $\omega = \pi$ the largest frequency considered? Suppose the data were generated from a periodic function with frequency $\omega > \pi$, say, $\omega = 3\pi/2$ for illustration:

$$Y_t = \alpha\cdot\cos[(3\pi/2)t] + \delta\cdot\sin[(3\pi/2)t]. \qquad [6.2.18]$$

Again, the properties of the sine and cosine function imply that [6.2.18] is equivalent to

$$Y_t = \alpha\cdot\cos[(-\pi/2)t] + \delta\cdot\sin[(-\pi/2)t]. \qquad [6.2.19]$$

Thus, by the previous argument, a representation with cycles of frequency $(3\pi/2)$ is observationally indistinguishable from one with cycles of frequency $(\pi/2)$.

To summarize, if the data-generating process actually includes cycles with negative frequencies or with frequencies greater than π, these will be imputed to cycles with frequencies between 0 and π. This is known as *aliasing*.

Another way to think about aliasing is as follows. Recall that the value of the function $\cos(\omega t)$ repeats itself every $2\pi/\omega$ periods, so that a frequency of ω is associated with a period of $2\pi/\omega$.[4] We have argued that the highest-frequency cycle that one can observe is $\omega = \pi$. Another way to express this conclusion is that the shortest-period cycle that one can observe is one that repeats itself every $2\pi/\pi = 2$ periods. If $\omega = 3\pi/2$, the cycle repeats itself every $\frac{4}{3}$ periods. But if the data are observed only at integer dates, the sampled data will exhibit cycles that are repeated every four periods, corresponding to the frequency $\omega = \pi/2$. This is illustrated in Figure 6.1, which plots $\cos[(\pi/2)t]$ and $\cos[(3\pi/2)t]$ as functions of t. When sampled at integer values of t, these two functions appear identical. Even though the function $\cos[(3\pi/2)t]$ repeats itself every time that t increases by $\frac{4}{3}$, one would have to observe y_t at four distinct dates $(y_t, y_{t+1}, y_{t+2}, y_{t+3})$ before one would see the value of $\cos[(3\pi/2)t]$ repeat itself for an integer value of t.

[4]See Section A.1 of the Mathematical Review (Appendix A) at the end of the book for a further discussion of this point.

Note that in a particular finite sample, the lowest frequency used to account for variation in y is $\omega_1 = 2\pi/T$, which corresponds to a period of T. If a cycle takes longer than T periods to repeat itself, there is not much that one could infer about it if one has only T observations available.

Result (c) of Proposition 6.2 indicates that the portion of the sample variance of y that can be attributed to cycles of frequency ω_j is proportional to the sample periodogram evaluated at ω_j, with $4\pi/T$ the constant of proportionality. Thus, the proposition develops the formal basis for the claim that the sample periodogram reflects the portion of the sample variance of y that can be attributed to cycles of various frequencies.

Why is the constant of proportionality in [6.2.15] equal to $4\pi/T$? The population spectrum $s_Y(\omega)$ could be evaluated at any ω in the continuous set of points between 0 and π. In this respect it is much like a probability density $f_X(x)$, where X is a continuous random variable. Although we might loosely think of the value of $f_X(x)$ as the "probability" that $X = x$, it is more accurate to say that the integral $\int_{x_1}^{x_2} f_X(x)\, dx$ represents the probability that X takes on a value between x_1 and x_2. As $x_2 - x_1$ becomes smaller, the probability that X will be observed to lie between x_1 and x_2 becomes smaller, and the probability that X would take on precisely the value x is effectively equal to zero. In just the same way, although we can loosely think of the value of $s_Y(\omega)$ as the contribution that cycles with frequency ω make to the variance of Y, it is more accurate to say that the integral

$$\int_{-\omega_1}^{\omega_1} s_Y(\omega)\, d\omega = \int_0^{\omega_1} 2s_Y(\omega)\, d\omega$$

represents the contribution that cycles of frequency less than or equal to ω_1 make to the variance of Y, and that $\int_{\omega_1}^{\omega_2} 2s_Y(\omega)\, d\omega$ represents the contribution that cycles with frequencies between ω_1 and ω_2 make to the variance of Y. Assuming that $s_Y(\omega)$ is continuous, the contribution that a cycle of any particular frequency ω makes is technically zero.

Although the population spectrum $s_Y(\omega)$ is defined at any ω in $[0, \pi]$, the representation in [6.2.6] attributes all of the sample variance of y to the particular frequencies $\omega_1, \omega_2, \ldots, \omega_M$. Any variation in Y that is in reality due to cycles with frequencies other than these M particular values is attributed by [6.2.6] to one of these M frequencies. If we are thinking of the regression in [6.2.6] as telling us something about the population spectrum, we should interpret $\frac{1}{2}(\hat{\alpha}_j^2 + \hat{\delta}_j^2)$ not as the portion of the variance of Y that is due to cycles with frequency exactly equal to ω_j, but rather as the portion of the variance of Y that is due to cycles with frequency near ω_j. Thus [6.2.15] is not an estimate of the height of the population spectrum, but an estimate of the area under the population spectrum.

This is illustrated in Figure 6.2. Suppose we thought of $\frac{1}{2}(\hat{\alpha}_j^2 + \hat{\delta}_j^2)$ as an estimate of the portion of the variance of Y that is due to cycles with frequency between ω_{j-1} and ω_j, that is, an estimate of 2 times the area under $s_Y(\omega)$ between ω_{j-1} and ω_j. Since $\omega_j = 2\pi j/T$, the difference $\omega_j - \omega_{j-1}$ is equal to $2\pi/T$. If $\hat{s}_y(\omega_j)$ is an estimate of $s_Y(\omega_j)$, then the area under $s_Y(\omega)$ between ω_{j-1} and ω_j could be approximately estimated by the area of a rectangle with width $2\pi/T$ and height $\hat{s}_y(\omega_j)$. The area of such a rectangle is $(2\pi/T)\cdot\hat{s}_j(\omega_j)$. Since $\frac{1}{2}(\hat{\alpha}_j^2 + \hat{\delta}_j^2)$ is an estimate of 2 times the area under $s_Y(\omega)$ between ω_{j-1} and ω_j, we have $\frac{1}{2}(\hat{\alpha}_j^2 + \hat{\delta}_j^2) = (4\pi/T)\cdot\hat{s}_y(\omega_j)$, as claimed in equation [6.2.15].

Proposition 6.2 also provides a convenient formula for calculating the value of the sample periodogram at frequency $\omega_j = 2\pi j/T$ for $j = 1, 2, \ldots, (T - 1)/2$,

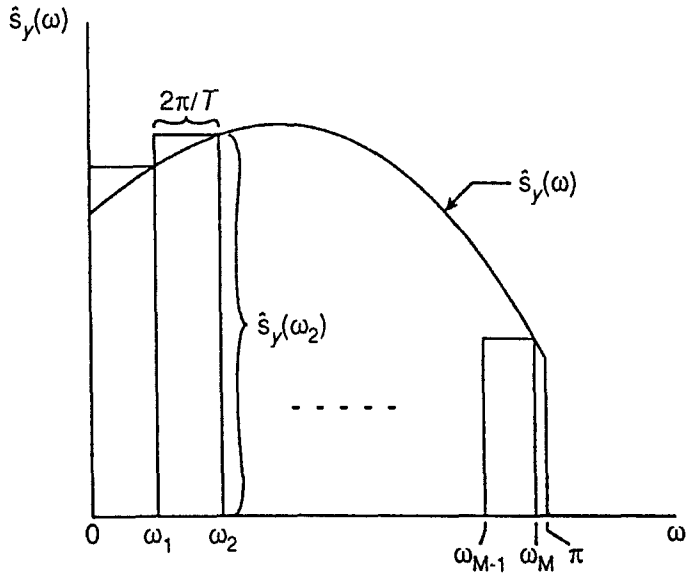


FIGURE 6.2 The area under the sample periodogram and the portion of the variance of y attributable to cycles of different frequencies.

namely,

$$\hat{s}_y(\omega_j) = [T/(8\pi)](\hat{\alpha}_j^2 + \hat{\delta}_j^2),$$

where

$$\hat{\alpha}_j = (2/T) \sum_{t=1}^{T} y_t \cdot \cos[\omega_j(t - 1)]$$

$$\hat{\delta}_j = (2/T) \sum_{t=1}^{T} y_t \cdot \sin[\omega_j(t - 1)].$$

That is,

$$\hat{s}_y(\omega_j) = \frac{1}{2\pi T}\left\{\left[\sum_{t=1}^{T} y_t \cdot \cos[\omega_j(t - 1)]\right]^2 + \left[\sum_{t=1}^{T} y_t \cdot \sin[\omega_j(t - 1)]\right]^2\right\}.$$

6.3. *Estimating the Population Spectrum*

Section 6.1 introduced the population spectrum $s_Y(\omega)$, which indicates the portion of the population variance of Y that can be attributed to cycles of frequency ω. This section addresses the following question: Given an observed sample $\{y_1, y_2, \ldots, y_T\}$, how might $s_Y(\omega)$ be estimated?

Large-Sample Properties of the Sample Periodogram

One obvious approach would be to estimate the population spectrum $s_Y(\omega)$ by the sample periodogram $\hat{s}_y(\omega)$. However, this approach turns out to have some

serious limitations. Suppose that

$$Y_t = \sum_{j=0}^{\infty} \psi_j \varepsilon_{t-j},$$

where $\{\psi_j\}_{j=0}^{\infty}$ is absolutely summable and where $\{\varepsilon_t\}_{t=-\infty}^{\infty}$ is an i.i.d. sequence with $E(\varepsilon_t) = 0$ and $E(\varepsilon_t^2) = \sigma^2$. Let $s_Y(\omega)$ be the population spectrum defined in [6.1.2], and suppose that $s_Y(\omega) > 0$ for all ω. Let $\hat{s}_y(\omega)$ be the sample periodogram defined in [6.2.4]. Fuller (1976, p. 280) showed that for $\omega \neq 0$ and a sufficiently large sample size T, twice the ratio of the sample periodogram to the population spectrum has approximately the following distribution:

$$\frac{2 \cdot \hat{s}_y(\omega)}{s_Y(\omega)} \approx \chi^2(2). \qquad [6.3.1]$$

Moreover, if $\lambda \neq \omega$, the quantity

$$\frac{2 \cdot \hat{s}_y(\lambda)}{s_Y(\lambda)} \qquad [6.3.2]$$

also has an approximate $\chi^2(2)$ distribution, with the variable in [6.3.1] approximately independent of that in [6.3.2].

Since a $\chi^2(2)$ variable has a mean of 2, result [6.3.1] suggests that

$$E\left[\frac{2 \cdot \hat{s}_y(\omega)}{s_Y(\omega)}\right] \cong 2,$$

or since $s_Y(\omega)$ is a population magnitude rather than a random variable,

$$E[\hat{s}_y(\omega)] \cong s_Y(\omega).$$

Thus, if the sample size is sufficiently large, the sample periodogram affords an approximately unbiased estimate of the population spectrum.

Note from Table B.2 that 95% of the time, a $\chi^2(2)$ variable will fall between 0.05 and 7.4. Thus, from [6.3.1], $\hat{s}_y(\omega)$ is unlikely to be as small as 0.025 times the true value of $s_Y(\omega)$, and $\hat{s}_y(\omega)$ is unlikely to be any larger than 3.7 times as big as $s_Y(\omega)$. Given such a large confidence interval, we would have to say that $\hat{s}_y(\omega)$ is not an altogether satisfactory estimate of $s_Y(\omega)$.

Another feature of result [6.3.1] is that the estimate $\hat{s}_y(\omega)$ is not getting any more accurate as the sample size T increases. Typically, one expects an econometric estimate to get better and better as the sample size grows. For example, the variance for the sample autocorrelation coefficient $\hat{\rho}_j$ given in [4.8.8] goes to zero as $T \to \infty$, so that given a sufficiently large sample, we would be able to infer the true value of ρ_j with virtual certainty. The estimate $\hat{s}_y(\omega)$ defined in [6.2.4] does not have this property, because we have tried to estimate as many parameters ($\gamma_0, \gamma_1, \ldots, \gamma_{T-1}$) as we had observations ($y_1, y_2, \ldots, y_T$).

Parametric Estimates of the Population Spectrum

Suppose we believe that the data could be represented with an $ARMA(p, q)$ model,

$$Y_t = \mu + \phi_1 Y_{t-1} + \phi_2 Y_{t-2} + \cdots + \phi_p Y_{t-p} + \varepsilon_t + \theta_1 \varepsilon_{t-1} + \theta_2 \varepsilon_{t-2} + \cdots + \theta_q \varepsilon_{t-q}, \qquad [6.3.3]$$

where ε_t is white noise with variance σ^2. Then an excellent approach to estimating the population spectrum is first to estimate the parameters $\mu, \phi_1, \ldots, \phi_p, \theta_1,$

$\ldots, \theta_q$ and σ^2 by maximum likelihood as described in the previous chapter. The maximum likelihood estimates $(\hat{\phi}_1, \ldots, \hat{\phi}_p, \hat{\theta}_1, \ldots, \hat{\theta}_q, \hat{\sigma}^2)$ could then be plugged into a formula such as [6.1.14] to estimate the population spectrum $s_Y(\omega)$ at any frequency ω. If the model is correctly specified, the maximum likelihood estimates $(\hat{\phi}_1, \ldots, \hat{\phi}_p, \hat{\theta}_1, \ldots, \hat{\theta}_q, \hat{\sigma}^2)$ will get closer and closer to the true values as the sample size grows; hence, the resulting estimate of the population spectrum should have this same property.

Even if the model is incorrectly specified, if the autocovariances of the true process are reasonably close to those for an $ARMA(p, q)$ specification, this procedure should provide a useful estimate of the population spectrum.

Nonparametric Estimates of the Population Spectrum

The assumption in [6.3.3] is that Y_t can be reasonably approximated by an $ARMA(p, q)$ process with p and q small. An alternative assumption is that $s_Y(\omega)$ will be close to $s_Y(\lambda)$ when ω is close to λ. This assumption forms the basis for another class of estimates of the population spectrum known as *nonparametric* or *kernel estimates*.

If $s_Y(\omega)$ is close to $s_Y(\lambda)$ when ω is close to λ, this suggests that $s_Y(\omega)$ might be estimated with a weighted average of the values of $\hat{s}_y(\lambda)$ for values of λ in a neighborhood around ω, where the weights depend on the distance between ω and λ. Let $\hat{s}_Y(\omega)$ denote such an estimate of $s_Y(\omega)$ and let $\omega_j = 2\pi j/T$. The suggestion is to take

$$\hat{s}_Y(\omega_j) = \sum_{m=-h}^{h} \kappa(\omega_{j+m}, \omega_j) \cdot \hat{s}_y(\omega_{j+m}). \qquad [6.3.4]$$

Here, h is a *bandwidth* parameter indicating how many different frequencies $\{\omega_{j\pm 1}, \omega_{j\pm 2}, \ldots, \omega_{j\pm h}\}$ are viewed as useful for estimating $s_Y(\omega_j)$. The *kernel* $\kappa(\omega_{j+m}, \omega_j)$ indicates how much weight each frequency is to be given. The kernel weights sum to unity:

$$\sum_{m=-h}^{h} \kappa(\omega_{j+m}, \omega_j) = 1.$$

One approach is to take $\kappa(\omega_{j+m}, \omega_j)$ to be proportional to $h + 1 - |m|$. One can show that[5]

$$\sum_{m=-h}^{h} [h + 1 - |m|] = (h + 1)^2.$$

Hence, in order to satisfy the property that the weights sum to unity, the proposed kernel is

$$\kappa(\omega_{j+m}, \omega_j) = \frac{h + 1 - |m|}{(h + 1)^2} \qquad [6.3.5]$$

[5]Notice that

$$\begin{aligned}\sum_{m=-h}^{h} [h + 1 - |m|] &= \sum_{m=-h}^{h} (h + 1) - \sum_{m=-h}^{h} |m| \\ &= (h + 1) \sum_{m=-h}^{h} 1 - 2\sum_{s=0}^{h} s \\ &= (2h + 1)(h + 1) - 2h(h + 1)/2 \\ &= (h + 1)^2.\end{aligned}$$

and the estimator [6.3.4] becomes

$$\hat{s}_Y(\omega_j) = \sum_{m=-h}^{h} \left[\frac{h + 1 - |m|}{(h + 1)^2}\right] \hat{s}_y(\omega_{j+m}). \qquad [6.3.6]$$

For example, for $h = 2$, this is

$$\hat{s}_Y(\omega_j) = \tfrac{1}{5}\hat{s}_y(\omega_{j-2}) + \tfrac{2}{5}\hat{s}_y(\omega_{j-1}) + \tfrac{3}{5}\hat{s}_y(\omega_j) + \tfrac{2}{5}\hat{s}_y(\omega_{j+1}) + \tfrac{1}{5}\hat{s}_y(\omega_{j+2}).$$

Recall from [6.3.1] and [6.3.2] that the estimates $\hat{s}_y(\omega)$ and $\hat{s}_y(\lambda)$ are approximately independent in large samples for $\omega \neq \lambda$. Because the kernel estimate averages over a number of different frequencies, it should give a much better estimate than does the periodogram.

Averaging $\hat{s}_y(\omega)$ over different frequencies can equivalently be represented as multiplying the jth sample autocovariance $\hat{\gamma}_j$ for $j > 0$ in the formula for the sample periodogram [6.2.5] by a weight κ_j^*. For example, consider an estimate of the spectrum at frequency ω that is obtained by taking a simple average of the value of $\hat{s}_y(\lambda)$ for λ between $\omega - \nu$ and $\omega + \nu$:

$$\hat{s}_Y(\omega) = (2\nu)^{-1} \int_{\omega-\nu}^{\omega+\nu} \hat{s}_y(\lambda)\, d\lambda. \qquad [6.3.7]$$

Substituting [6.2.5] into [6.3.7], such an estimate could equivalently be expressed as

$$\begin{aligned}
\hat{s}_Y(\omega) &= (4\nu\pi)^{-1} \int_{\omega-\nu}^{\omega+\nu} \left[\hat{\gamma}_0 + 2\sum_{j=1}^{T-1} \hat{\gamma}_j \cos(\lambda j)\right] d\lambda \\
&= (4\nu\pi)^{-1}(2\nu)\hat{\gamma}_0 + (2\nu\pi)^{-1} \sum_{j=1}^{T-1} \hat{\gamma}_j (1/j)\cdot\left[\sin(\lambda j)\right]_{\lambda=\omega-\nu}^{\omega+\nu} \qquad [6.3.8] \\
&= (2\pi)^{-1}\hat{\gamma}_0 + (2\nu\pi)^{-1} \sum_{j=1}^{T-1} \hat{\gamma}_j (1/j)\cdot\{\sin[(\omega + \nu)j] - \sin[(\omega - \nu)j]\}.
\end{aligned}$$

Using the trigonometric identity[6]

$$\sin(A + B) - \sin(A - B) = 2\cdot\cos(A)\cdot\sin(B), \qquad [6.3.9]$$

expression [6.3.8] can be written

$$\begin{aligned}
\hat{s}_Y(\omega) &= (2\pi)^{-1}\hat{\gamma}_0 + (2\nu\pi)^{-1} \sum_{j=1}^{T-1} \hat{\gamma}_j (1/j)\cdot[2\cdot\cos(\omega j)\cdot\sin(\nu j)] \\
&= (2\pi)^{-1}\left\{\hat{\gamma}_0 + 2\sum_{j=1}^{T-1}\left[\frac{\sin(\nu j)}{\nu j}\right]\hat{\gamma}_j \cos(\omega j)\right\}.
\end{aligned} \qquad [6.3.10]$$

Notice that expression [6.3.10] is of the following form:

$$\hat{s}_Y(\omega) = (2\pi)^{-1}\left\{\hat{\gamma}_0 + 2\sum_{j=1}^{T-1} \kappa_j^* \hat{\gamma}_j \cos(\omega j)\right\}, \qquad [6.3.11]$$

where

$$\kappa_j^* = \left[\frac{\sin(\nu j)}{\nu j}\right]. \qquad [6.3.12]$$

The sample periodogram can be regarded as a special case of [6.3.11] when $\kappa_j^* = 1$. Expression [6.3.12] cannot exceed 1 in absolute value, and so the estimate [6.3.11] essentially downweights $\hat{\gamma}_j$ relative to the sample periodogram.

[6]See, for example, Thomas (1972, pp. 174–75).

Recall that $\sin(\pi j) = 0$ for any integer j. Hence, if $\nu = \pi$, then $\kappa_j^* = 0$ for all j and [6.3.11] becomes

$$\hat{s}_Y(\omega) = (2\pi)^{-1}\hat{\gamma}_0. \qquad [6.3.13]$$

In this case, all autocovariances other than $\hat{\gamma}_0$ would be shrunk to zero. When $\nu = \pi$, the estimate [6.3.7] is an unweighted average of $\hat{s}_y(\lambda)$ over all possible values of λ, and the resulting estimate would be the flat spectrum for a white noise process.

Specification of a kernel function $\kappa(\omega_{j+m}, \omega_j)$ in [6.3.4] can equivalently be described in terms of a weighting sequence $\{\kappa_j^*\}_{j=1}^{T-1}$ in [6.3.11]. Because they are just two different representations for the same idea, the weight κ_j^* is also sometimes called a kernel. Smaller values of κ_j^* impose more smoothness on the spectrum. Smoothing schemes may be chosen either because they provide a convenient specification for $\kappa(\omega_{j+m}, \omega_j)$ or because they provide a convenient specification for κ_j^*.

One popular estimate of the spectrum employs the modified *Bartlett kernel*, which is given by

$$\kappa_j^* = \begin{cases} 1 - \dfrac{j}{q+1} & \text{for } j = 1, 2, \ldots, q \\ 0 & \text{for } j > q. \end{cases} \qquad [6.3.14]$$

The Bartlett estimate of the spectrum is thus

$$\hat{s}_Y(\omega) = (2\pi)^{-1}\left\{\hat{\gamma}_0 + 2\sum_{j=1}^{q}[1 - j/(q+1)]\hat{\gamma}_j\cos(\omega j)\right\}. \qquad [6.3.15]$$

Autocovariances γ_j for $j > q$ are treated as if they were zero, or as if Y_t followed an $MA(q)$ process. For $j \leq q$, the estimated autocovariances $\hat{\gamma}_j$ are shrunk toward zero, with the shrinkage greater the larger the value of j.

How is one to choose the bandwidth parameter h in [6.3.6] or q in [6.3.15]? The periodogram itself is asymptotically unbiased but has a large variance. If one constructs an estimate based on averaging the periodogram at different frequencies, this reduces the variance but introduces some bias. The severity of the bias depends on the steepness of the population spectrum and the size of the bandwidth. One practical guide is to plot an estimate of the spectrum using several different bandwidths and rely on subjective judgment to choose the bandwidth that produces the most plausible estimate.

6.4. *Uses of Spectral Analysis*

We illustrate some of the uses of spectral analysis with data on manufacturing production in the United States. The data are plotted in Figure 6.3. The series is the Federal Reserve Board's seasonally unadjusted monthly index from January 1947 to November 1989. Economic recessions in 1949, 1954, 1958, 1960, 1970, 1974, 1980, and 1982 appear as roughly year-long episodes of falling production. There are also strong seasonal patterns in this series; for example, production almost always declines in July and recovers in August.

The sample periodogram for the raw data is plotted in Figure 6.4, which displays $\hat{s}_y(\omega_j)$ as a function of j where $\omega_j = 2\pi j/T$. The contribution to the sample variance of the lowest-frequency components (j near zero) is several orders of magnitude larger than the contributions of economic recessions or the seasonal factors. This is due to the clear upward trend of the series in Figure 6.3. Let y_t

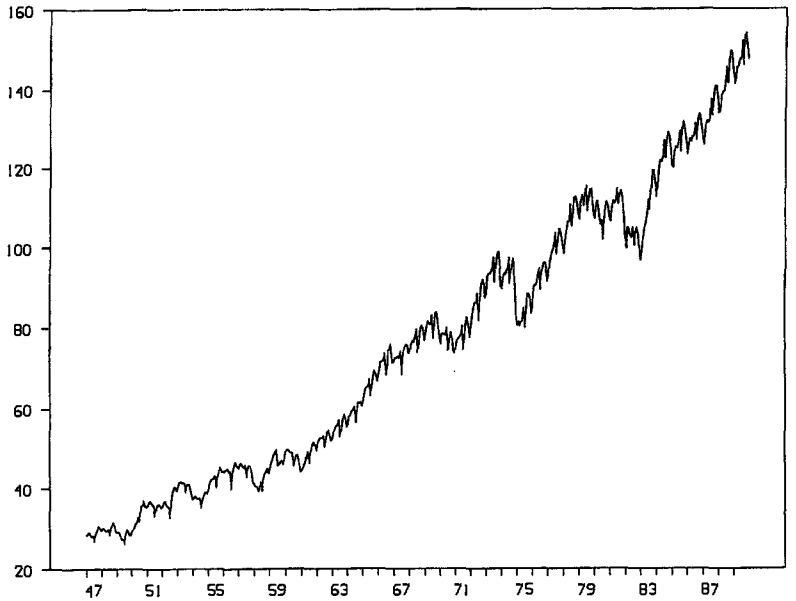


FIGURE 6.3 Federal Reserve Board's seasonally unadjusted index of industrial production for U.S. manufacturing, monthly 1947:1 to 1989:11.

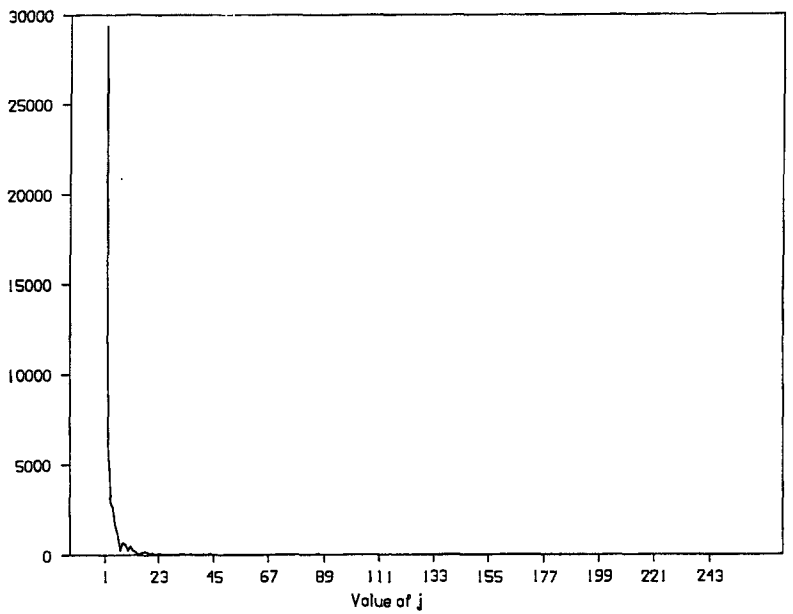


FIGURE 6.4 Sample periodogram for the data plotted in Figure 6.3. The figure plots $\hat{s}_y(\omega_j)$ as a function of j, where $\omega_j = 2\pi j/T$.

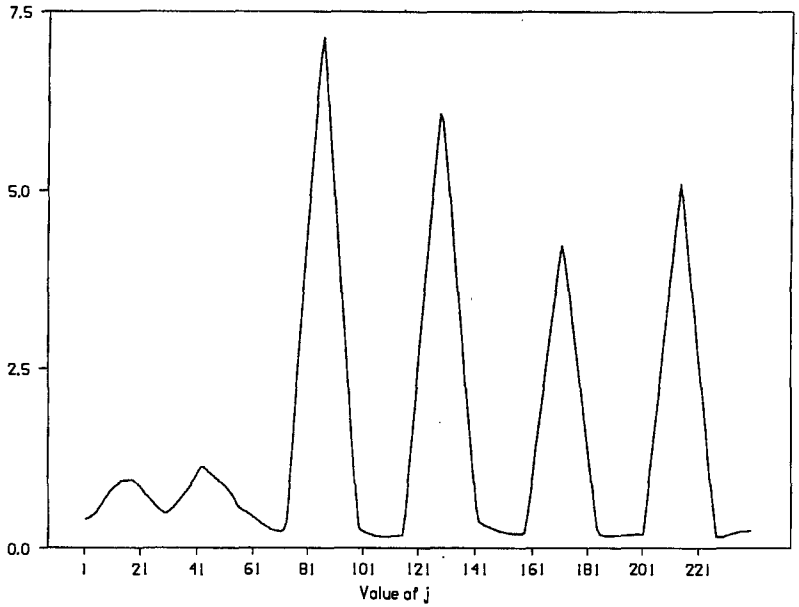


FIGURE 6.5 Estimate of the spectrum for monthly growth rate of industrial production, or spectrum of 100 times the first difference of the log of the series in Figure 6.3.

represent the series plotted in Figure 6.3. If one were trying to describe this with a sine function

$$y_t = \delta \cdot \sin(\omega t),$$

the presumption would have to be that ω is so small that even at date $t = T$ the magnitude ωT would still be less than $\pi/2$. Figure 6.4 thus indicates that the trend or low-frequency components are by far the most important determinants of the sample variance of y.

The definition of the population spectrum in equation [6.1.2] assumed that the process is covariance-stationary, which is not a good assumption for the data in Figure 6.3. We might instead try to analyze the monthly growth rate defined by

$$x_t = 100 \cdot [\log(y_t) - \log(y_{t-1})]. \qquad [6.4.1]$$

Figure 6.5 plots the estimate of the population spectrum of X as described in equation [6.3.6] with $h = 12$.

In interpreting a plot such as Figure 6.5 it is often more convenient to think in terms of the period of a cyclic function rather than its frequency. Recall that if the frequency of a cycle is ω, the period of the cycle is $2\pi/\omega$. Thus, a frequency of $\omega_j = 2\pi j/T$ corresponds to a period of $2\pi/\omega_j = T/j$. The sample size is $T = 513$ observations, and the first peak in Figure 6.5 occurs around $j = 18$. This corresponds to a cycle with a period of $513/18 = 28.5$ months, or about $2\frac{1}{2}$ years. Given the dates of the economic recessions noted previously, this is sometimes described as a "business cycle frequency," and the area under this hill might be viewed as telling us how much of the variability in monthly growth rates is due to economic recessions.

The second peak in Figure 6.5 occurs at $j = 44$ and corresponds to a period of $513/44 = 11.7$ months. This is natural to view as a 12-month cycle associated with seasonal effects. The four subsequent peaks correspond to cycles with periods of 6, 4, 3, and 2.4 months, respectively, and again seem likely to be picking up seasonal and calendar effects.

Since manufacturing typically falls temporarily in July, the growth rate is negative in July and positive in August. This induces negative first-order serial correlation to the series in [6.4.1] and a variety of calendar patterns for x_t that may account for the high-frequency peaks in Figure 6.5. An alternative strategy for detrending would use year-to-year growth rates, or the percentage change between y_t and its value for the corresponding month in the previous year:

$$w_t = 100 \cdot [\log(y_t) - \log(y_{t-12})]. \qquad [6.4.2]$$

The estimate of the sample spectrum for this series is plotted in Figure 6.6. When the data are detrended in this way, virtually all the variance that remains is attributed to components associated with the business cycle frequencies.

Filters

Apart from the scale parameter, the monthly growth rate x_t in [6.4.1] is obtained from $\log(y_t)$ by applying the filter

$$x_t = (1 - L)\log(y_t), \qquad [6.4.3]$$

where L is the lag operator. To discuss such transformations in general terms, let Y_t be any covariance-stationary series with absolutely summable autocovariances.

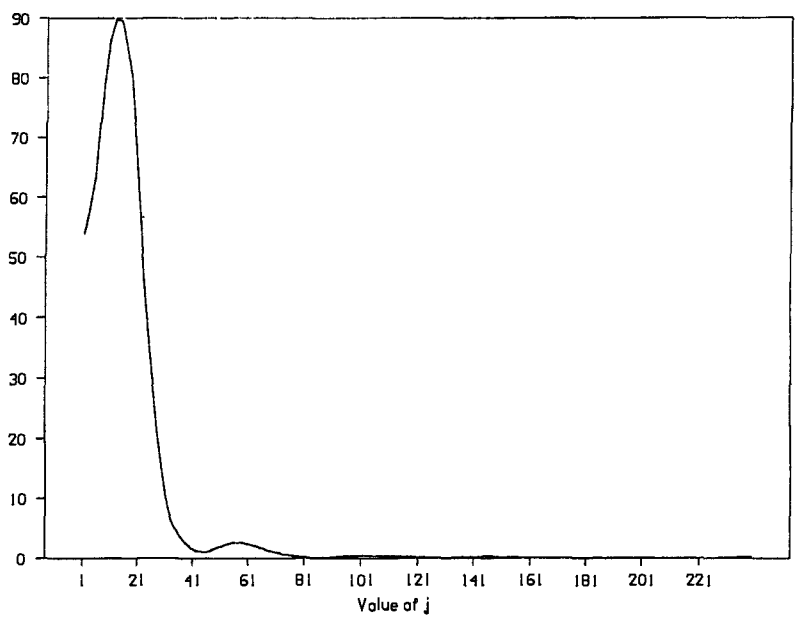


FIGURE 6.6 Estimate of the spectrum for year-to-year growth rate of monthly industrial production, or spectrum of 100 times the seasonal difference of the log of the series in Figure 6.3.

Denote the autocovariance-generating function of Y by $g_Y(z)$, and denote the population spectrum of Y by $s_Y(\omega)$. Recall that

$$s_Y(\omega) = (2\pi)^{-1} g_Y(e^{-i\omega}). \qquad [6.4.4]$$

Suppose we transform Y according to

$$X_t = h(L)Y_t,$$

where

$$h(L) = \sum_{j=-\infty}^{\infty} h_j L^j$$

and

$$\sum_{j=-\infty}^{\infty} |h_j| < \infty.$$

Recall from equation [3.6.17] that the autocovariance-generating function of X can be calculated from the autocovariance-generating function of Y using the formula

$$g_X(z) = h(z)h(z^{-1})g_Y(z). \qquad [6.4.5]$$

The population spectrum of X is thus

$$s_X(\omega) = (2\pi)^{-1} g_X(e^{-i\omega}) = (2\pi)^{-1} h(e^{-i\omega})h(e^{i\omega})g_Y(e^{-i\omega}). \qquad [6.4.6]$$

Substituting [6.4.4] into [6.4.6] reveals that the population spectrum of X is related to the population spectrum of Y according to

$$s_X(\omega) = h(e^{-i\omega})h(e^{i\omega})s_Y(\omega). \qquad [6.4.7]$$

Operating on a series Y_t with the filter $h(L)$ has the effect of multiplying the spectrum by the function $h(e^{-i\omega})h(e^{i\omega})$.

For the difference operator in [6.4.3], the filter is $h(L) = 1 - L$ and the function $h(e^{-i\omega})h(e^{i\omega})$ would be

$$\begin{aligned} h(e^{-i\omega})h(e^{i\omega}) &= (1 - e^{-i\omega})(1 - e^{i\omega}) \\ &= 1 - e^{-i\omega} - e^{i\omega} + 1 \\ &= 2 - 2\cdot\cos(\omega), \end{aligned} \qquad [6.4.8]$$

where the last line follows from [6.1.11]. If $X_t = (1 - L)Y_t$, then, to find the value of the population spectrum of X at any frequency ω, we first find the value of the population spectrum of Y at ω and then multiply by $2 - 2\cdot\cos(\omega)$. For example, the spectrum at frequency $\omega = 0$ is multiplied by zero, the spectrum at frequency $\omega = \pi/2$ is multiplied by 2, and the spectrum at frequency $\omega = \pi$ is multiplied by 4. Differencing the data removes the low-frequency components and accentuates the high-frequency components.

Of course, this calculation assumes that the original process Y_t is covariance-stationary, so that $s_Y(\omega)$ exists. If the original process is nonstationary, as appears to be the case in Figure 6.3, the differenced data $(1 - L)Y_t$ in general would not have a population spectrum that is zero at frequency zero.

The seasonal difference filter used in [6.4.2] is $h(L) = 1 - L^{12}$, for which

$$\begin{aligned} h(e^{-i\omega})h(e^{i\omega}) &= (1 - e^{-12i\omega})(1 - e^{12i\omega}) \\ &= 1 - e^{-12i\omega} - e^{12i\omega} + 1 \\ &= 2 - 2\cdot\cos(12\omega). \end{aligned}$$

This function is equal to zero when $12\omega = 0, 2\pi, 4\pi, 6\pi, 8\pi, 10\pi$, or 12π; that is, it is zero at frequencies $\omega = 0, 2\pi/12, 4\pi/12, 6\pi/12, 8\pi/12, 10\pi/12$, and π. Thus, seasonally differencing not only eliminates the low-frequency ($\omega = 0$) components of a stationary process, but further eliminates any contribution from cycles with periods of 12, 6, 4, 3, 2.4, or 2 months.

Composite Stochastic Processes

Let X_t be covariance-stationary with absolutely summable autocovariances, autocovariance-generating function $g_X(z)$, and population spectrum $s_X(\omega)$. Let W_t be a different covariance-stationary series with absolutely summable autocovariances, autocovariance-generating function $g_W(z)$, and population spectrum $s_W(\omega)$, where X_t is uncorrelated with W_τ for all t and τ. Suppose we observe the sum of these two processes,

$$Y_t = X_t + W_t.$$

Recall from [4.7.19] that the autocovariance-generating function of the sum is the sum of the autocovariance-generating functions:

$$g_Y(z) = g_X(z) + g_W(z).$$

It follows from [6.1.2] that the spectrum of the sum is the sum of the spectra:

$$s_Y(\omega) = s_X(\omega) + s_W(\omega). \qquad [6.4.9]$$

For example, if a white noise series W_t with variance σ^2 is added to a series X_t and if X_t is uncorrelated with W_τ for all t and τ, the effect is to shift the population spectrum everywhere up by the constant $\sigma^2/(2\pi)$. More generally, if X has a peak in its spectrum at frequency ω_1 and if W has a peak in its spectrum at ω_2, then typically the sum $X + W$ will have peaks at both ω_1 and ω_2.

As another example, suppose that

$$Y_t = c + \sum_{j=-\infty}^{\infty} h_j X_{t-j} + \varepsilon_t,$$

where X_t is covariance-stationary with absolutely summable autocovariances and spectrum $s_X(\omega)$. Suppose that the sequence $\{h_j\}_{j=-\infty}^{\infty}$ is absolutely summable and that ε_t is a white noise process with variance σ^2 where ε is uncorrelated with X at all leads and lags. It follows from [6.4.7] that the random variable $\Sigma_{j=-\infty}^{\infty} h_j X_{t-j}$ has spectrum $h(e^{-i\omega})h(e^{i\omega})s_X(\omega)$, and so, from [6.4.9], the spectrum of Y is

$$s_Y(\omega) = h(e^{-i\omega})h(e^{i\omega})s_X(\omega) + \sigma^2/(2\pi).$$

APPENDIX 6.A. *Proofs of Chapter 6 Propositions*

■ **Proof of Proposition 6.1.** Notice that

$$\begin{aligned}\int_{-\pi}^{\pi} s_Y(\omega)e^{i\omega k}\,d\omega &= \frac{1}{2\pi}\int_{-\pi}^{\pi}\sum_{j=-\infty}^{\infty}\gamma_j e^{-i\omega j}e^{i\omega k}\,d\omega \\ &= \frac{1}{2\pi}\sum_{j=-\infty}^{\infty}\gamma_j\int_{-\pi}^{\pi}e^{i\omega(k-j)}\,d\omega \\ &= \frac{1}{2\pi}\sum_{j=-\infty}^{\infty}\gamma_j\int_{-\pi}^{\pi}\{\cos[\omega(k-j)] + i\cdot\sin[\omega(k-j)]\}\,d\omega.\end{aligned} \qquad [6.A.1]$$

Consider the integral in [6.A.1]. For $k = j$, this would be

$$\int_{-\pi}^{\pi} \{\cos[\omega(k-j)] + i\cdot\sin[\omega(k-j)]\}\, d\omega = \int_{-\pi}^{\pi} \{\cos(0) + i\cdot\sin(0)\}\, d\omega$$
$$= \int_{-\pi}^{\pi} d\omega \qquad [6.A.2]$$
$$= 2\pi.$$

For $k \neq j$, the integral in [6.A.1] would be

$$\int_{-\pi}^{\pi} \{\cos[\omega(k-j)] + i\cdot\sin[\omega(k-j)]\}\, d\omega$$
$$= \left.\frac{\sin[\omega(k-j)]}{k-j}\right|_{\omega=-\pi}^{\pi} - i\cdot\left.\frac{\cos[\omega(k-j)]}{k-j}\right|_{\omega=-\pi}^{\pi} \qquad [6.A.3]$$
$$= (k-j)^{-1}\{\sin[\pi(k-j)] - \sin[-\pi(k-j)]$$
$$- i\cdot\cos[\pi(k-j)] + i\cdot\cos[-\pi(k-j)]\}.$$

But the difference between the frequencies $\pi(k-j)$ and $-\pi(k-j)$ is $2\pi(k-j)$, which is an integer multiple of 2π. Since the sine and cosine functions are periodic, the magnitude in [6.A.3] is zero. Hence, only the term for $j = k$ in the sum in [6.A.1] is nonzero, and using [6.A.2], this sum is seen to be

$$\int_{-\pi}^{\pi} s_Y(\omega)e^{i\omega k}\, d\omega = \frac{1}{2\pi}\gamma_k \int_{-\pi}^{\pi} [\cos(0) + i\cdot\sin(0)]\, d\omega = \gamma_k,$$

as claimed in [6.1.15].

To derive [6.1.16], notice that since $s_Y(\omega)$ is symmetric around $\omega = 0$,

$$\int_{-\pi}^{\pi} s_Y(\omega)e^{i\omega k}\, d\omega = \int_{-\pi}^{0} s_Y(\omega)e^{i\omega k}\, d\omega + \int_{0}^{\pi} s_Y(\omega)e^{i\omega k}\, d\omega$$
$$= \int_{0}^{\pi} s_Y(-\omega)e^{-i\omega k}\, d\omega + \int_{0}^{\pi} s_Y(\omega)e^{i\omega k}\, d\omega$$
$$= \int_{0}^{\pi} s_Y(\omega)(e^{-i\omega k} + e^{i\omega k})\, d\omega$$
$$= \int_{0}^{\pi} s_Y(\omega)\cdot 2\cdot\cos(\omega k)\, d\omega,$$

where the last line follows from [6.1.11]. Again appealing to the symmetry of $s_Y(\omega)$,

$$\int_{0}^{\pi} s_Y(\omega)\cdot 2\cdot\cos(\omega k)\, d\omega = \int_{-\pi}^{\pi} s_Y(\omega)\cos(\omega k)\, d\omega,$$

so that

$$\int_{-\pi}^{\pi} s_Y(\omega)e^{i\omega k}\, d\omega = \int_{-\pi}^{\pi} s_Y(\omega)\cos(\omega k)\, d\omega,$$

as claimed. ■

■ **Derivation of Equation [6.2.11] in Proposition 6.2.** We begin by establishing the following result:

$$\sum_{t=1}^{T} \exp[i(2\pi s/T)(t-1)] = \begin{cases} T & \text{for } s = 0 \\ 0 & \text{for } s = \pm 1, \pm 2, \ldots, \pm(T-1). \end{cases} \qquad [6.A.4]$$

That [6.A.4] holds for $s = 0$ is an immediate consequence of the fact that $\exp(0) = 1$. To see that it holds for the other cases in [6.A.4], define

$$z \equiv \exp[i(2\pi s/T)]. \qquad [6.A.5]$$

Then the expression to be evaluated in [6.A.4] can be written

$$\sum_{t=1}^{T} \exp[i(2\pi s/T)(t-1)] = \sum_{t=1}^{T} z^{(t-1)}. \qquad [6.A.6]$$

We now show that for any N,

$$\sum_{t=1}^{N} z^{(t-1)} = \frac{1-z^N}{1-z}, \qquad [6.A.7]$$

provided that $z \neq 1$, which is the case whenever $0 < |s| < T$. Expression [6.A.7] can be verified by induction. Clearly, it holds for $N = 1$, for then

$$\sum_{t=1}^{N} z^{(t-1)} = z^{(0)} = 1.$$

Given that [6.A.7] holds for N, we see that

$$\begin{aligned}\sum_{t=1}^{N+1} z^{(t-1)} &= \sum_{t=1}^{N} z^{(t-1)} + z^N \\ &= \frac{1-z^N}{1-z} + z^N \\ &= \frac{1-z^N + z^N(1-z)}{1-z} \\ &= \frac{1-z^{N+1}}{1-z},\end{aligned}$$

as claimed in [6.A.7].

Setting $N = T$ in [6.A.7] and substituting the result into [6.A.6], we see that

$$\sum_{t=1}^{T} \exp[i(2\pi s/T)(t-1)] = \frac{1-z^T}{1-z} \qquad [6.A.8]$$

for $0 < |s| < T$. But it follows from the definition of z in [6.A.5] that

$$\begin{aligned} z^T &= \exp[i(2\pi s/T)\cdot T] \\ &= \exp[i(2\pi s)] \\ &= \cos(2\pi s) + i\cdot\sin(2\pi s) \\ &= 1 \qquad \text{for } s = \pm 1, \pm 2, \ldots, \pm(T-1). \end{aligned} \qquad [6.A.9]$$

Substituting [6.A.9] into [6.A.8] produces

$$\sum_{t=1}^{T} \exp[i(2\pi s/T)(t-1)] = 0 \qquad \text{for } s = \pm 1, \pm 2, \ldots, \pm(T-1),$$

as claimed in [6.A.4].

To see how [6.A.4] can be used to deduce expression [6.2.11], notice that the first column of $\Sigma_{t=1}^{T} \mathbf{x}_t\mathbf{x}_t'$ is given by

$$\begin{bmatrix} T \\ \sum \cos[\omega_1(t-1)] \\ \sum \sin[\omega_1(t-1)] \\ \vdots \\ \sum \cos[\omega_M(t-1)] \\ \sum \sin[\omega_M(t-1)] \end{bmatrix}, \qquad [6.A.10]$$

where Σ indicates summation over t from 1 to T. The first row of $\Sigma_{t=1}^{T} \mathbf{x}_t\mathbf{x}_t'$ is the transpose of [6.A.10]. To show that all the terms in [6.A.10] other than the first element are zero,

we must show that

$$\sum_{t=1}^{T} \cos[\omega_j(t-1)] = 0 \qquad \text{for } j = 1, 2, \ldots, M \tag{6.A.11}$$

and

$$\sum_{t=1}^{T} \sin[\omega_j(t-1)] = 0 \qquad \text{for } j = 1, 2, \ldots, M \tag{6.A.12}$$

for ω_j the frequencies specified in [6.2.7]. But [6.A.4] establishes that

$$\begin{aligned} 0 &= \sum_{t=1}^{T} \exp[i(2\pi j/T)(t-1)] \\ &= \sum_{t=1}^{T} \cos[(2\pi j/T)(t-1)] + i\cdot\sum_{t=1}^{T} \sin[(2\pi j/T)(t-1)] \end{aligned} \tag{6.A.13}$$

for $j = 1, 2, \ldots, M$. For [6.A.13] to equal zero, both the real and the imaginary component must equal zero. Since $\omega_j = 2\pi j/T$, results [6.A.11] and [6.A.12] follow immediately from [6.A.13].

Result [6.A.4] can also be used to calculate the other elements of $\Sigma_{t=1}^{T}\mathbf{x}_t\mathbf{x}_t'$. To see how, note that

$$\begin{aligned} \frac{1}{2}[e^{i\theta} + e^{-i\theta}] &= \frac{1}{2}[\cos(\theta) + i\cdot\sin(\theta) + \cos(\theta) - i\cdot\sin(\theta)] \\ &= \cos(\theta) \end{aligned} \tag{6.A.14}$$

and similarly

$$\begin{aligned} \frac{1}{2i}[e^{i\theta} - e^{-i\theta}] &= \frac{1}{2i}\{\cos(\theta) + i\cdot\sin(\theta) - [\cos(\theta) - i\cdot\sin(\theta)]\} \\ &= \sin(\theta). \end{aligned} \tag{6.A.15}$$

Thus, for example, the elements of $\Sigma_{t=1}^{T}\mathbf{x}_t\mathbf{x}_t'$ corresponding to products of the cosine terms can be calculated as

$$\begin{aligned} &\sum_{t=1}^{T} \cos[\omega_j(t-1)]\cdot\cos[\omega_k(t-1)] \\ &\quad = \frac{1}{4}\sum_{t=1}^{T} \{\exp[i\omega_j(t-1)] + \exp[-i\omega_j(t-1)]\} \\ &\qquad \times \{\exp[i\omega_k(t-1)] + \exp[-i\omega_k(t-1)]\} \\ &\quad = \frac{1}{4}\sum_{t=1}^{T} \{\exp[i(\omega_j + \omega_k)(t-1)] + \exp[i(-\omega_j + \omega_k)(t-1)] \\ &\qquad + \exp[i(\omega_j - \omega_k)(t-1)] + \exp[i(-\omega_j - \omega_k)(t-1)]\} \\ &\quad = \frac{1}{4}\sum_{t=1}^{T} \{\exp[i(2\pi/T)(j+k)(t-1)] + \exp[i(2\pi/T)(k-j)(t-1)] \\ &\qquad + \exp[i(2\pi/T)(j-k)(t-1)] + \exp[i(2\pi/T)(-j-k)(t-1)]\}. \end{aligned} \tag{6.A.16}$$

For any $j = 1, 2, \ldots, M$ and any $k = 1, 2, \ldots, M$ where $k \neq j$, expression [6.A.16] is zero by virtue of [6.A.4]. For $k = j$, the first and last sums in the last line of [6.A.16] are zero, so that the total is equal to

$$(1/4)\sum_{t=1}^{T} (1 + 1) = T/2.$$

Similarly, elements of $\Sigma_{t=1}^{T}\mathbf{x}_t\mathbf{x}_t'$ corresponding to cross products of the sine terms can be found from

$$\begin{aligned}
\sum_{t=1}^{T} \sin[\omega_j(t-1)]\cdot\sin[\omega_k(t-1)] &= -\frac{1}{4}\sum_{t=1}^{T}\{\exp[i\omega_j(t-1)] - \exp[-i\omega_j(t-1)]\} \\
&\quad \times \{\exp[i\omega_k(t-1)] - \exp[-i\omega_k(t-1)]\} \\
&= -\frac{1}{4}\sum_{t=1}^{T}\{\exp[i(2\pi/T)(j+k)(t-1)] - \exp[i(2\pi/T)(k-j)(t-1)] \\
&\quad - \exp[i(2\pi/T)(j-k)(t-1)] + \exp[i(2\pi/T)(-j-k)(t-1)]\} \\
&= \begin{cases} T/2 & \text{for } j = k \\ 0 & \text{otherwise.} \end{cases}
\end{aligned}$$

Finally, elements of $\Sigma_{t=1}^{T}\mathbf{x}_t\mathbf{x}_t'$ corresponding to cross products of the sine and cosine terms are given by

$$\begin{aligned}
\sum_{t=1}^{T} \cos[\omega_j(t-1)]\cdot\sin[\omega_k(t-1)] &= \frac{1}{4i}\sum_{t=1}^{T}\{\exp[i\omega_j(t-1)] + \exp[-i\omega_j(t-1)]\} \\
&\quad \times \{\exp[i\omega_k(t-1)] - \exp[-i\omega_k(t-1)]\} \\
&= \frac{1}{4i}\sum_{t=1}^{T}\{\exp[i(2\pi/T)(j+k)(t-1)] + \exp[i(2\pi/T)(k-j)(t-1)] \\
&\quad - \exp[i(2\pi/T)(j-k)(t-1)] - \exp[i(2\pi/T)(-j-k)(t-1)]\},
\end{aligned}$$

which equals zero for all j and k. This completes the derivation of [6.2.11]. ■

■ **Proof of Proposition 6.2(a).** Let $\mathbf{b}$ denote the estimate of $\boldsymbol{\beta}$ based on *OLS* estimation of the regression in [6.2.8]:

$$\begin{aligned}
\mathbf{b} &= \left\{\sum_{t=1}^{T}\mathbf{x}_t\mathbf{x}_t'\right\}^{-1}\left\{\sum_{t=1}^{T}\mathbf{x}_t y_t\right\} \\
&= \begin{bmatrix} T & \mathbf{0}' \\ \mathbf{0} & (T/2)\cdot\mathbf{I}_{T-1} \end{bmatrix}^{-1}\left\{\sum_{t=1}^{T}\mathbf{x}_t y_t\right\} \qquad [6.A.17] \\
&= \begin{bmatrix} T^{-1} & \mathbf{0}' \\ \mathbf{0} & (2/T)\cdot\mathbf{I}_{T-1} \end{bmatrix}\left\{\sum_{t=1}^{T}\mathbf{x}_t y_t\right\}.
\end{aligned}$$

But the definition of $\mathbf{x}_t$ in [6.2.9] implies that

$$\begin{aligned}
\sum_{t=1}^{T}\mathbf{x}_t y_t = \Big[&\sum y_t \quad \sum y_t\cdot\cos[\omega_1(t-1)] \quad \sum y_t\cdot\sin[\omega_1(t-1)] \\
&\sum y_t\cdot\cos[\omega_2(t-1)] \quad \sum y_t\cdot\sin[\omega_2(t-1)] \quad \cdots \qquad [6.A.18] \\
&\sum y_t\cdot\cos[\omega_M(t-1)] \quad \sum y_t\cdot\sin[\omega_M(t-1)]\Big]',
\end{aligned}$$

where Σ again denotes summation over t from 1 to T. Substituting [6.A.18] into [6.A.17] produces result (a) of Proposition 6.2. ■

■ **Proof of Proposition 6.2(b).** Recall from expression [4.A.6] that the residual sum of squares associated with *OLS* estimation of [6.2.8] is

$$\sum_{t=1}^{T}\hat{u}_t^2 = \sum_{t=1}^{T}y_t^2 - \left[\sum_{t=1}^{T}y_t\mathbf{x}_t'\right]\left[\sum_{t=1}^{T}\mathbf{x}_t\mathbf{x}_t'\right]^{-1}\left[\sum_{t=1}^{T}\mathbf{x}_t y_t\right]. \qquad [6.A.19]$$

Since there are as many explanatory variables as observations and since the explanatory variables are linearly independent, the OLS residuals $\hat{u}_t$ are all zero. Hence, [6.A.19] implies that

$$\sum_{t=1}^{T} y_t^2 = \left[\sum_{t=1}^{T} y_t \mathbf{x}_t'\right]\left[\sum_{t=1}^{T} \mathbf{x}_t \mathbf{x}_t'\right]^{-1}\left[\sum_{t=1}^{T} \mathbf{x}_t y_t\right]. \qquad [6.A.20]$$

But [6.A.17] allows us to write

$$\sum_{t=1}^{T} \mathbf{x}_t y_t = \begin{bmatrix} T & \mathbf{0}' \\ \mathbf{0} & (T/2)\cdot\mathbf{I}_{T-1} \end{bmatrix}\mathbf{b}. \qquad [6.A.21]$$

Substituting [6.A.21] and [6.2.11] into [6.A.20] establishes that

$$\begin{aligned}\sum_{t=1}^{T} y_t^2 &= \mathbf{b}'\begin{bmatrix} T & \mathbf{0}' \\ \mathbf{0} & (T/2)\cdot\mathbf{I}_{T-1} \end{bmatrix}\begin{bmatrix} T & \mathbf{0}' \\ \mathbf{0} & (T/2)\cdot\mathbf{I}_{T-1} \end{bmatrix}^{-1}\begin{bmatrix} T & \mathbf{0}' \\ \mathbf{0} & (T/2)\cdot\mathbf{I}_{T-1} \end{bmatrix}\mathbf{b} \\ &= \mathbf{b}'\begin{bmatrix} T & \mathbf{0}' \\ \mathbf{0} & (T/2)\cdot\mathbf{I}_{T-1} \end{bmatrix}\mathbf{b} \\ &= T\cdot\hat{\mu}^2 + (T/2)\sum_{j=1}^{M}(\hat{\alpha}_j^2 + \hat{\delta}_j^2)\end{aligned}$$

so that

$$(1/T)\sum_{t=1}^{T} y_t^2 = \hat{\mu}^2 + (1/2)\sum_{j=1}^{M}(\hat{\alpha}_j^2 + \hat{\delta}_j^2). \qquad [6.A.22]$$

Finally, observe from [4.A.5] and the fact that $\hat{\mu} = \bar{y}$ that

$$(1/T)\sum_{t=1}^{T} y_t^2 - \hat{\mu}^2 = (1/T)\sum_{t=1}^{T}(y_t - \bar{y})^2,$$

allowing [6.A.22] to be written as

$$(1/T)\sum_{t=1}^{T}(y_t - \bar{y})^2 = (1/2)\sum_{j=1}^{M}(\hat{\alpha}_j^2 + \hat{\delta}_j^2),$$

as claimed in [6.2.14]. Since the regressors are all orthogonal, the term $\frac{1}{2}(\hat{\alpha}_j^2 + \hat{\delta}_j^2)$ can be interpreted as the portion of the sample variance that can be attributed to the regressors $\cos[\omega_j(t-1)]$ and $\sin[\omega_j(t-1)]$. ■

■ **Proof of Proposition 6.2(c).** Notice that

$$(\hat{\alpha}_j^2 + \hat{\delta}_j^2) = (\hat{\alpha}_j + i\cdot\hat{\delta}_j)(\hat{\alpha}_j - i\cdot\hat{\delta}_j). \qquad [6.A.23]$$

But from result (a) of Proposition 6.2,

$$\hat{\alpha}_j = (2/T)\sum_{t=1}^{T} y_t\cdot\cos[\omega_j(t-1)] = (2/T)\sum_{t=1}^{T}(y_t - \bar{y})\cdot\cos[\omega_j(t-1)], \qquad [6.A.24]$$

where the second equality follows from [6.A.11]. Similarly,

$$\hat{\delta}_j = (2/T)\sum_{t=1}^{T}(y_t - \bar{y})\cdot\sin[\omega_j(t-1)]. \qquad [6.A.25]$$

It follows from [6.A.24] and [6.A.25] that

$$\begin{aligned}\hat{\alpha}_j + i\cdot\hat{\delta}_j &= (2/T)\left\{\sum_{t=1}^{T}(y_t - \bar{y})\cdot\cos[\omega_j(t-1)]\right. \\ &\qquad \left. + i\cdot\sum_{t=1}^{T}(y_t - \bar{y})\cdot\sin[\omega_j(t-1)]\right\} \qquad [6.A.26] \\ &= (2/T)\sum_{t=1}^{T}(y_t - \bar{y})\cdot\exp[i\omega_j(t-1)].\end{aligned}$$

Similarly,

$$\hat{\alpha}_j - i\cdot\hat{\delta}_j = (2/T)\sum_{\tau=1}^{T}(y_\tau - \bar{y})\cdot\exp[-i\omega_j(\tau - 1)]. \qquad [6.A.27]$$

Substituting [6.A.26] and [6.A.27] into [6.A.23] produces

$$\begin{aligned}
\hat{\alpha}_j^2 + \hat{\delta}_j^2 &= (4/T^2)\left\{\sum_{t=1}^{T}(y_t - \bar{y})\cdot\exp[i\omega_j(t-1)]\right\} \\
&\quad \times \left\{\sum_{\tau=1}^{T}(y_\tau - \bar{y})\cdot\exp[-i\omega_j(\tau-1)]\right\} \\
&= (4/T^2)\sum_{t=1}^{T}\sum_{\tau=1}^{T}(y_t - \bar{y})(y_\tau - \bar{y})\cdot\exp[i\omega_j(t-\tau)] \\
&= (4/T^2)\left\{\sum_{t=1}^{T}(y_t - \bar{y})^2 + \sum_{t=1}^{T-1}(y_t - \bar{y})(y_{t+1} - \bar{y})\cdot\exp[-i\omega_j]\right. \\
&\quad + \sum_{t=2}^{T}(y_t - \bar{y})(y_{t-1} - \bar{y})\cdot\exp[i\omega_j] \\
&\quad + \sum_{t=1}^{T-2}(y_t - \bar{y})(y_{t+2} - \bar{y})\cdot\exp[-2i\omega_j] \\
&\quad + \sum_{t=3}^{T}(y_t - \bar{y})(y_{t-2} - \bar{y})\cdot\exp[2i\omega_j] + \cdots \qquad [6.A.28] \\
&\quad + (y_1 - \bar{y})(y_T - \bar{y})\cdot\exp[-(T-1)i\omega_j] \\
&\quad \left. + (y_T - \bar{y})(y_1 - \bar{y})\cdot\exp[(T-1)i\omega_j]\right\} \\
&= (4/T)\left\{\hat{\gamma}_0 + \hat{\gamma}_1\cdot\exp[-i\omega_j] + \hat{\gamma}_{-1}\cdot\exp[i\omega_j]\right. \\
&\quad + \hat{\gamma}_2\cdot\exp[-2i\omega_j] + \hat{\gamma}_{-2}\cdot\exp[2i\omega_j] + \cdots \\
&\quad \left. + \hat{\gamma}_{T-1}\cdot\exp[-(T-1)i\omega_j] + \hat{\gamma}_{-T+1}\cdot\exp[(T-1)i\omega_j]\right\} \\
&= (4/T)(2\pi)\hat{s}_y(\omega_j),
\end{aligned}$$

from which equation [6.2.15] follows. ■

Chapter 6 Exercises

6.1. Derive [6.1.12] directly from expression [6.1.6] and the formulas for the autocovariances of an $MA(1)$ process.

6.2. Integrate [6.1.9] and [6.1.12] to confirm independently that [6.1.17] holds for white noise and an $MA(1)$ process.

Chapter 6 References

Anderson, T. W. 1971. *The Statistical Analysis of Time Series*. New York: Wiley.

Bloomfield, Peter. 1976. *Fourier Analysis of Time Series: An Introduction*. New York: Wiley.

Cramér, Harald, and M. R. Leadbetter. 1967. *Stationary and Related Stochastic Processes*. New York: Wiley.

Fuller, Wayne A. 1976. *Introduction to Statistical Time Series*. New York: Wiley.

Thomas, George B., Jr. 1972. *Calculus and Analytic Geometry*, alternate ed. Reading, Mass.: Addison-Wesley.

7 Asymptotic Distribution Theory

Suppose a sample of T observations $(y_1, y_2, \ldots, y_T)$ has been used to construct $\hat{\boldsymbol{\theta}}$, an estimate of the vector of population parameters. For example, the parameter vector $\boldsymbol{\theta} = (c, \phi_1, \phi_2, \ldots, \phi_p, \sigma^2)'$ for an $AR(p)$ process might have been estimated from an OLS regression of y_t on lagged y's. We would like to know how far this estimate $\hat{\boldsymbol{\theta}}$ is likely to be from the true value $\boldsymbol{\theta}$ and how to test a hypothesis about the true value based on the observed sample of y's.

Much of the distribution theory used to answer these questions is *asymptotic*: that is, it describes the properties of estimators as the sample size (T) goes to infinity. This chapter develops the basic asymptotic results that will be used in subsequent chapters. The first section summarizes the key tools of asymptotic analysis and presents limit theorems for the sample mean of a sequence of i.i.d. random variables. Section 7.2 develops limit theorems for serially dependent variables with time-varying marginal distributions.

7.1. *Review of Asymptotic Distribution Theory*

Limits of Deterministic Sequences

Let $\{c_T\}_{T=1}^{\infty}$ denote a sequence of deterministic numbers. The sequence is said to *converge* to c if for any $\varepsilon > 0$, there exists an N such that $|c_T - c| < \varepsilon$ whenever $T \geq N$; in other words, c_T will be as close as desired to c so long as T is sufficiently large. This is indicated as

$$\lim_{T\to\infty} c_T = c, \qquad [7.1.1]$$

or, equivalently,

$$c_T \to c.$$

For example, $c_T = 1/T$ denotes the sequence $\{1, \frac{1}{2}, \frac{1}{3}, \ldots\}$, for which

$$\lim_{T\to\infty} c_T = 0.$$

A sequence of deterministic $(m \times n)$ matrices $\{\mathbf{C}_T\}_{T=1}^{\infty}$ converges to $\mathbf{C}$ if each element of $\mathbf{C}_T$ converges to the corresponding element of $\mathbf{C}$.

Convergence in Probability

Consider a sequence of scalar random variables, $\{X_T\}_{T=1}^{\infty}$. The sequence is said to *converge in probability* to c if for every $\varepsilon > 0$ and every $\delta > 0$ there exists a value N such that, for all $T \geq N$,

$$P\{|X_T - c| > \delta\} < \varepsilon. \qquad [7.1.2]$$

In words, if we go far enough along in the sequence, the probability that X_T differs from c by more than δ can be made arbitrarily small for any δ.

When [7.1.2] is satisfied, the number c is called the *probability limit*, or *plim*, of the sequence $\{X_T\}$. This is indicated as

$$\text{plim } X_T = c,$$

or, equivalently,

$$X_T \xrightarrow{p} c.$$

Recall that if $\{c_T\}_{T=1}^{\infty}$ is a deterministic sequence converging to c, then there exists an N such that $|c_T - c| < \delta$ for all $T \geq N$. Then $P\{|c_T - c| > \delta\} = 0$ for all $T \geq N$. Thus, if a deterministic sequence converges to c, then we could also say that $c_T \xrightarrow{p} c$.

A sequence of $(m \times n)$ matrices of random variables $\{\mathbf{X}_T\}$ converges in probability to the $(m \times n)$ matrix $\mathbf{C}$ if each element of $\mathbf{X}_T$ converges in probability to the corresponding element of $\mathbf{C}$.

More generally, if $\{\mathbf{X}_T\}$ and $\{\mathbf{Y}_T\}$ are sequences of $(m \times n)$ matrices, we will use the notation

$$\mathbf{X}_T \xrightarrow{p} \mathbf{Y}_T$$

to indicate that the difference between the two sequences converges in probability to zero:

$$\mathbf{X}_T - \mathbf{Y}_T \xrightarrow{p} \mathbf{0}.$$

An example of a sequence of random variables of interest is the following. Suppose we have a sample of T observations on a random variable $\{Y_1, Y_2, \ldots, Y_T\}$. Consider the sample mean,

$$\overline{Y}_T \equiv (1/T) \sum_{t=1}^{T} Y_t, \qquad [7.1.3]$$

as an estimator of the population mean,

$$\hat{\mu}_T = \overline{Y}_T.$$

We append the subscript T to this estimator to emphasize that it describes the mean of a sample of size T. The primary focus will be on the behavior of this estimator as T grows large. Thus, we will be interested in the properties of the sequence $\{\hat{\mu}_T\}_{T=1}^{\infty}$.

When the plim of a sequence of estimators (such as $\{\hat{\mu}_T\}_{T=1}^{\infty}$) is equal to the true population parameter (in this case, μ), the estimator is said to be *consistent*. If an estimator is consistent, then there exists a sufficiently large sample such that we can be assured with very high probability that the estimate will be within any desired tolerance band around the true value.

The following result is quite helpful in finding plims; a proof of this and some of the other propositions of this chapter are provided in Appendix 7.A at the end of the chapter.

Proposition 7.1: *Let $\{\mathbf{X}_T\}$ denote a sequence of $(n \times 1)$ random vectors with plim $\mathbf{c}$, and let $\mathbf{g}(\mathbf{c})$ be a vector-valued function, $\mathbf{g}: \mathbb{R}^n \to \mathbb{R}^m$, where $\mathbf{g}(\cdot)$ is continuous at $\mathbf{c}$ and does not depend on T. Then $\mathbf{g}(\mathbf{X}_T) \xrightarrow{p} \mathbf{g}(\mathbf{c})$.*

The basic idea behind this proposition is that, since $\mathbf{g}(\cdot)$ is continuous, $\mathbf{g}(\mathbf{X}_T)$ will be close to $\mathbf{g}(\mathbf{c})$ provided that $\mathbf{X}_T$ is close to $\mathbf{c}$. By choosing a sufficiently large value of T, the probability that $\mathbf{X}_T$ is close to $\mathbf{c}$ (and thus that $\mathbf{g}(\mathbf{X}_T)$ is close to $\mathbf{g}(\mathbf{c})$) can be brought as near to unity as desired.

Note that $\mathbf{g}(\mathbf{X}_T)$ depends on the *value* of $\mathbf{X}_T$ but cannot depend on the index T itself. Thus, $g(X_T, T) = T \cdot X_T^2$ is not a function covered by Proposition 7.1.

Example 7.1

If $X_{1T} \xrightarrow{p} c_1$ and $X_{2T} \xrightarrow{p} c_2$, then $(X_{1T} + X_{2T}) \xrightarrow{p} (c_1 + c_2)$. This follows immediately, since $g(X_{1T}, X_{2T}) \equiv (X_{1T} + X_{2T})$ is a continuous function of (X_{1T}, X_{2T}).

Example 7.2

Let $\{\mathbf{X}_{1T}\}$ denote a sequence of $(n \times n)$ random matrices with $\mathbf{X}_{1T} \xrightarrow{p} \mathbf{C}_1$, a nonsingular matrix. Let $\mathbf{X}_{2T}$ denote a sequence of $(n \times 1)$ random vectors with $\mathbf{X}_{2T} \xrightarrow{p} \mathbf{c}_2$. Then $[\mathbf{X}_{1T}]^{-1}\mathbf{X}_{2T} \xrightarrow{p} [\mathbf{C}_1]^{-1}\mathbf{c}_2$. To see this, note that the elements of the matrix $[\mathbf{X}_{1T}]^{-1}$ are continuous functions of the elements of $\mathbf{X}_{1T}$ at $\mathbf{X}_{1T} = \mathbf{C}_1$, since $[\mathbf{C}_1]^{-1}$ exists. Thus, $[\mathbf{X}_{1T}]^{-1} \xrightarrow{p} [\mathbf{C}_1]^{-1}$. Similarly, the elements of $[\mathbf{X}_{1T}]^{-1}\mathbf{X}_{2T}$ are sums of products of elements of $[\mathbf{X}_{1T}]^{-1}$ with those of $\mathbf{X}_{2T}$. Since each sum is again a continuous function of $\mathbf{X}_{1T}$ and $\mathbf{X}_{2T}$,

$$\text{plim } [\mathbf{X}_{1T}]^{-1}\mathbf{X}_{2T} = [\text{plim } \mathbf{X}_{1T}]^{-1} \text{ plim } \mathbf{X}_{2T} = [\mathbf{C}_1]^{-1}\mathbf{c}_2.$$

Proposition 7.1 also holds if some of the elements of $\mathbf{X}_T$ are deterministic with conventional limits as in expression [7.1.1]. Specifically, let $\mathbf{X}_T' = (\mathbf{X}_{1T}', \mathbf{c}_{2T}')$, where $\mathbf{X}_{1T}$ is a stochastic $(n_1 \times 1)$ vector and $\mathbf{c}_{2T}$ is a deterministic $(n_2 \times 1)$ vector. If plim $\mathbf{X}_{1T} \doteq \mathbf{c}_1$ and $\lim_{T\to\infty} \mathbf{c}_{2T} = \mathbf{c}_2$, then $\mathbf{g}(\mathbf{X}_{1T}, \mathbf{c}_{2T}) \xrightarrow{p} \mathbf{g}(\mathbf{c}_1, \mathbf{c}_2)$. (See Exercise 7.1.)

Example 7.3

Consider an alternative estimator of the mean given by $\overline{Y}_T^* \equiv [1/(T-1)] \times \Sigma_{t=1}^T Y_t$. This can be written as $c_{1T}\overline{Y}_T$, where $c_{1T} \equiv [T/(T-1)]$ and $\overline{Y}_T \equiv (1/T)\Sigma_{t=1}^T Y_t$. Under general conditions detailed in Section 7.2, the sample mean is a consistent estimator of the population mean, implying that $\overline{Y}_T \xrightarrow{p} \mu$. It is also easy to verify that $c_{1T} \to 1$. Since $c_{1T}\overline{Y}_T$ is a continuous function of c_{1T} and $\overline{Y}_T$, it follows that $c_{1T}\overline{Y}_T \xrightarrow{p} 1 \cdot \mu = \mu$. Thus, $\overline{Y}_T^*$, like $\overline{Y}_T$, is a consistent estimator of μ.

Convergence in Mean Square and Chebyshev's Inequality

A stronger condition than convergence in probability is *mean square convergence*. The random sequence $\{X_T\}$ is said to converge in mean square to c, indicated as

$$X_T \xrightarrow{m.s.} c,$$

if for every $\varepsilon > 0$ there exists a value N such that, for all $T \geq N$,

$$E(X_T - c)^2 < \varepsilon. \qquad [7.1.4]$$

Another useful result is the following.

Proposition 7.2: *(Generalized Chebyshev's inequality). Let X be a random variable with $E(|X|^r)$ finite for some $r > 0$. Then, for any $\delta > 0$ and any value of c,*

$$P\{|X - c| > \delta\} \leq \frac{E|X - c|^r}{\delta^r}. \qquad [7.1.5]$$

An implication of Chebyshev's inequality is that if $X_T \xrightarrow{m.s.} c$, then $X_T \xrightarrow{p} c$. To see this, note that if $X_T \xrightarrow{m.s.} c$, then for any $\varepsilon > 0$ and $\delta > 0$ there exists an N such that $E(X_T - c)^2 < \delta^2\varepsilon$ for all $T \geq N$. This would ensure that

$$\frac{E(X_T - c)^2}{\delta^2} < \varepsilon$$

for all $T \geq N$. From Chebyshev's inequality, this also implies

$$P\{|X_T - c| > \delta\} < \varepsilon$$

for all $T \geq N$, or that $X_T \xrightarrow{p} c$.

Law of Large Numbers for Independent and Identically Distributed Variables

Let us now consider the behavior of the sample mean $\overline{Y}_T = (1/T)\Sigma_{t=1}^{T} Y_t$ where $\{Y_t\}$ is i.i.d. with mean μ and variance σ^2. For this case, $\overline{Y}_T$ has expectation μ and variance

$$E(\overline{Y}_T - \mu)^2 = (1/T^2)\,\mathrm{Var}\left(\sum_{t=1}^{T} Y_t\right) = (1/T^2)\sum_{t=1}^{T}\mathrm{Var}(Y_t) = \sigma^2/T.$$

Since $\sigma^2/T \to 0$ as $T \to \infty$, this means that $\overline{Y}_T \xrightarrow{m.s.} \mu$, implying also that $\overline{Y}_T \xrightarrow{p} \mu$.

Figure 7.1 graphs an example of the density of the sample mean $f_{\overline{Y}_T}(\bar{y}_T)$ for three different values of T. As T becomes large, the density becomes increasingly concentrated in a spike centered at μ.

The result that the sample mean is a consistent estimate of the population mean is known as the *law of large numbers.*[1] It was proved here for the special case of i.i.d. variables with finite variance. In fact, it turns out also to be true of any sequence of i.i.d. variables with finite mean μ.[2] Section 7.2 explores some of the circumstances under which it also holds for serially dependent variables with time-varying marginal distributions.

Convergence in Distribution

Let $\{X_T\}_{T=1}^{\infty}$ be a sequence of random variables, and let $F_{X_T}(x)$ denote the cumulative distribution function of X_T. Suppose that there exists a cumulative distribution function $F_X(x)$ such that

$$\lim_{T\to\infty} F_{X_T}(x) = F_X(x)$$

[1]This is often described as the weak law of large numbers. An analogous result known as the strong law of large numbers refers to almost sure convergence rather than convergence in probability of the sample mean.

[2]This is known as *Khinchine's theorem.* See, for example, Rao (1973, p. 112).

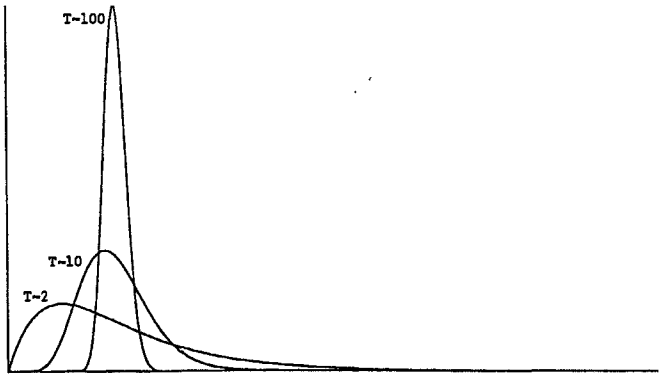


FIGURE 7.1 Density of the sample mean for a sample of size T.

at any value x at which $F_X(\cdot)$ is continuous. Then X_T is said to *converge in distribution* (or in law) to X, denoted

$$X_T \xrightarrow{L} X.$$

When $F_X(x)$ is of a common form, such as the cumulative distribution function for a $N(\mu, \sigma^2)$ variable, we will equivalently write

$$X_T \xrightarrow{L} N(\mu, \sigma^2).$$

The definitions are unchanged if the scalar X_T is replaced with an $(n \times 1)$ vector $\mathbf{X}_T$. A simple way to verify convergence in distribution of a vector is the following.[3] If the scalar $(\lambda_1 X_{1T} + \lambda_2 X_{2T} + \cdots + \lambda_n X_{nT})$ converges in distribution to $(\lambda_1 X_1 + \lambda_2 X_2 + \cdots + \lambda_n X_n)$ for any real values of $(\lambda_1, \lambda_2, \ldots, \lambda_n)$, then the vector $\mathbf{X}_T \equiv (X_{1T}, X_{2T}, \ldots, X_{nT})'$ converges in distribution to the vector $\mathbf{X} \equiv (X_1, X_2, \ldots, X_n)'$.

The following results are useful in determining limiting distributions.[4]

Proposition 7.3:

(*a*) *Let* $\{\mathbf{Y}_T\}$ *be a sequence of* $(n \times 1)$ *random vectors with* $\mathbf{Y}_T \xrightarrow{L} \mathbf{Y}$. *Suppose that* $\{\mathbf{X}_T\}$ *is a sequence of* $(n \times 1)$ *random vectors such that* $(\mathbf{X}_T - \mathbf{Y}_T) \xrightarrow{p} \mathbf{0}$. *Then* $\mathbf{X}_T \xrightarrow{L} \mathbf{Y}$; *that is,* $\mathbf{X}_T$ *and* $\mathbf{Y}_T$ *have the same limiting distribution.*

(*b*) *Let* $\{\mathbf{X}_T\}$ *be a sequence of random* $(n \times 1)$ *vectors with* $\mathbf{X}_T \xrightarrow{p} \mathbf{c}$, *and let* $\{\mathbf{Y}_T\}$ *be a sequence of random* $(n \times 1)$ *vectors with* $\mathbf{Y}_T \xrightarrow{L} \mathbf{Y}$. *Then the sequence constructed from the sum* $\{\mathbf{X}_T + \mathbf{Y}_T\}$ *converges in distribution to* $\mathbf{c} + \mathbf{Y}$ *and the sequence constructed from the product* $\{\mathbf{X}_T'\mathbf{Y}_T\}$ *converges in distribution to* $\mathbf{c}'\mathbf{Y}$.

(*c*) *Let* $\{\mathbf{X}_T\}$ *be a sequence of random* $(n \times 1)$ *vectors with* $\mathbf{X}_T \xrightarrow{L} \mathbf{X}$, *and let* $\mathbf{g}(\mathbf{X})$, $\mathbf{g}: \mathbb{R}^n \to \mathbb{R}^m$ *be a continuous function (not dependent on* T*). Then the sequence of random variables* $\{\mathbf{g}(\mathbf{X}_T)\}$ *converges in distribution to* $\mathbf{g}(\mathbf{X})$.

[3]This is known as the Cramér-Wold theorem. See Rao (1973, p. 123).

[4]See Rao (1973, pp. 122–24).

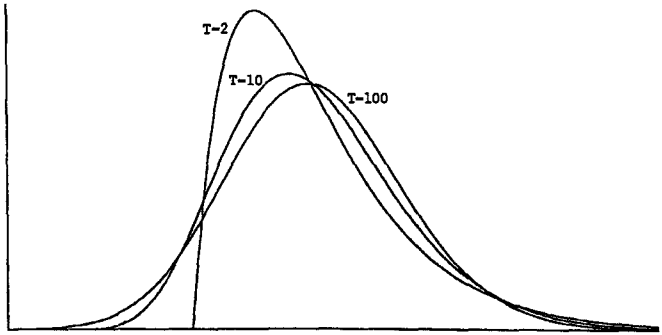


FIGURE 7.2 Density of $\sqrt{T}(\overline{Y}_T - \mu)$.

Example 7.4
Suppose that $X_T \xrightarrow{p} c$ and $Y_T \xrightarrow{L} Y$, where $Y \sim N(\mu, \sigma^2)$. Then, by Proposition 7.3(b), the sequence $X_T Y_T$ has the same limiting probability law as that of c times a $N(\mu, \sigma^2)$ variable. In other words, $X_T Y_T \xrightarrow{L} N(c\mu, c^2\sigma^2)$.

Example 7.5
Generalizing the previous result, let $\{\mathbf{X}_T\}$ be a sequence of random $(m \times n)$ matrices and $\{\mathbf{Y}_T\}$ a sequence of random $(n \times 1)$ vectors with $\mathbf{X}_T \xrightarrow{p} \mathbf{C}$ and $\mathbf{Y}_T \xrightarrow{L} \mathbf{Y}$, with $\mathbf{Y} \sim N(\boldsymbol{\mu}, \boldsymbol{\Omega})$. Then the limiting distribution of $\mathbf{X}_T\mathbf{Y}_T$ is the same as that of $\mathbf{CY}$; that is, $\mathbf{X}_T\mathbf{Y}_T \xrightarrow{L} N(\mathbf{C}\boldsymbol{\mu}, \mathbf{C}\boldsymbol{\Omega}\mathbf{C}')$.

Example 7.6
Suppose that $X_T \xrightarrow{L} N(0, 1)$. Then Proposition 7.3(c) implies that the square of X_T asymptotically behaves as the square of a $N(0, 1)$ variable: $X_T^2 \xrightarrow{p} \chi^2(1)$.

Central Limit Theorem

We have seen that the sample mean $\overline{Y}_T$ for an i.i.d. sequence has a degenerate probability density as $T \to \infty$, collapsing toward a point mass at μ as the sample size grows. For statistical inference we would like to describe the distribution of $\overline{Y}_T$ in more detail. For this purpose, note that the random variable $\sqrt{T}(\overline{Y}_T - \mu)$ has mean zero and variance given by $(\sqrt{T})^2 \operatorname{Var}(\overline{Y}_T) = \sigma^2$ for all T, and thus, in contrast to $\overline{Y}_T$, the random variable $\sqrt{T}(\overline{Y}_T - \mu)$ might be expected to converge to a nondegenerate random variable as T goes to infinity.

The *central limit theorem* is the result that, as T increases, the sequence $\sqrt{T}(\overline{Y}_T - \mu)$ converges in distribution to a Gaussian random variable. The most familiar, albeit restrictive, version of the central limit theorem establishes that if Y_t is i.i.d. with mean μ and variance σ^2, then[5]

$$\sqrt{T}(\overline{Y}_T - \mu) \xrightarrow{L} N(0, \sigma^2). \qquad [7.1.6]$$

Result [7.1.6] also holds under much more general conditions, some of which are explored in the next section.

Figure 7.2 graphs an example of the density of $\sqrt{T}(\overline{Y}_T - \mu)$ for three different

[5]See, for example, White (1984, pp. 108–9).

values of T. Each of these densities has mean zero and variance σ^2. As T becomes large, the density converges to that of a $N(0, \sigma^2)$ variable.

A final useful result is the following.

Proposition 7.4: *Let $\{\mathbf{X}_t\}$ be a sequence of random $(n \times 1)$ vectors such that $\sqrt{T}(\mathbf{X}_T - \mathbf{c}) \xrightarrow{L} \mathbf{X}$, and let $\mathbf{g}$: $\mathbb{R}^n \to \mathbb{R}^m$ have continuous first derivatives with $\mathbf{G}$ denoting the $(m \times n)$ matrix of derivatives evaluated at $\mathbf{c}$:*

$$\mathbf{G} \equiv \left.\frac{\partial \mathbf{g}}{\partial \mathbf{x}'}\right|_{\mathbf{x}=\mathbf{c}}.$$

Then $\sqrt{T}[\mathbf{g}(\mathbf{X}_T) - \mathbf{g}(\mathbf{c})] \xrightarrow{L} \mathbf{GX}$.

Example 7.7

Let $\{Y_1, Y_2, \ldots, Y_T\}$ be an i.i.d. sample of size T drawn from a distribution with mean $\mu \neq 0$ and variance σ^2. Consider the distribution of the reciprocal of the sample mean, $S_T = 1/\overline{Y}_T$, where $\overline{Y}_T \equiv (1/T)\Sigma_{t=1}^{T} Y_t$. We know from the central limit theorem that $\sqrt{T}(\overline{Y}_T - \mu) \xrightarrow{L} Y$, where $Y \sim N(0, \sigma^2)$. Also, $g(y) = 1/y$ is continuous at $y = \mu$. Let $G \equiv (\partial g/\partial y)|_{y=\mu} = (-1/\mu^2)$. Then $\sqrt{T}[S_T - (1/\mu)] \xrightarrow{L} G \cdot Y$; in other words, $\sqrt{T}[S_T - (1/\mu)] \xrightarrow{L} N(0, \sigma^2/\mu^4)$.

7.2. *Limit Theorems for Serially Dependent Observations*

The previous section stated the law of large numbers and central limit theorem for independent and identically distributed random variables with finite second moments. This section develops analogous results for heterogeneously distributed variables with various forms of serial dependence. We first develop a law of large numbers for a general covariance-stationary process.

Law of Large Numbers for a Covariance-Stationary Process

Let $(Y_1, Y_2, \ldots, Y_T)$ represent a sample of size T from a covariance-stationary process with

$$E(Y_t) = \mu \quad \text{for all } t \qquad [7.2.1]$$

$$E(Y_t - \mu)(Y_{t-j} - \mu) = \gamma_j \quad \text{for all } t \qquad [7.2.2]$$

$$\sum_{j=0}^{\infty} |\gamma_j| < \infty. \qquad [7.2.3]$$

Consider the properties of the sample mean,

$$\overline{Y}_T = (1/T) \sum_{t=1}^{T} Y_t. \qquad [7.2.4]$$

Taking expectations of [7.2.4] reveals that the sample mean provides an unbiased estimate of the population mean,

$$E(\overline{Y}_T) = \mu,$$

while the variance of the sample mean is

$$
\begin{aligned}
E(\overline{Y}_T - \mu)^2 &= E\left[(1/T)\sum_{t=1}^{T}(Y_t - \mu)\right]^2 \\
&= (1/T^2)E\{[(Y_1 - \mu) + (Y_2 - \mu) + \cdots + (Y_T - \mu)] \\
&\quad \times [(Y_1 - \mu) + (Y_2 - \mu) + \cdots + (Y_T - \mu)]\} \\
&= (1/T^2)E\{(Y_1 - \mu)[(Y_1 - \mu) + (Y_2 - \mu) + \cdots + (Y_T - \mu)] \\
&\quad + (Y_2 - \mu)[(Y_1 - \mu) + (Y_2 - \mu) + \cdots + (Y_T - \mu)] \\
&\quad + (Y_3 - \mu)[(Y_1 - \mu) + (Y_2 - \mu) + \cdots + (Y_T - \mu)] \\
&\quad + \cdots + (Y_T - \mu)[(Y_1 - \mu) + (Y_2 - \mu) + \cdots + (Y_T - \mu)]\} \\
&= (1/T^2)\{[\gamma_0 + \gamma_1 + \gamma_2 + \gamma_3 + \cdots + \gamma_{T-1}] \\
&\quad + [\gamma_1 + \gamma_0 + \gamma_1 + \gamma_2 + \cdots + \gamma_{T-2}] \\
&\quad + [\gamma_2 + \gamma_1 + \gamma_0 + \gamma_1 + \cdots + \gamma_{T-3}] \\
&\quad + \cdots + [\gamma_{T-1} + \gamma_{T-2} + \gamma_{T-3} + \cdots + \gamma_0]\}.
\end{aligned}
$$

Thus,

$$
\begin{aligned}
E(\overline{Y}_T - \mu)^2 = (1/T^2)\{&T\gamma_0 + 2(T-1)\gamma_1 \\
&+ 2(T-2)\gamma_2 + 2(T-3)\gamma_3 + \cdots + 2\gamma_{T-1}\}
\end{aligned}
$$

or

$$
\begin{aligned}
E(\overline{Y}_T - \mu)^2 = (1/T)\{&\gamma_0 + [(T-1)/T](2\gamma_1) + [(T-2)/T](2\gamma_2) \\
&+ [(T-3)/T](2\gamma_3) + \cdots + [1/T](2\gamma_{T-1})\}.
\end{aligned}
\tag{7.2.5}
$$

It is easy to see that this expression goes to zero as the sample size grows—that is, that $\overline{Y}_T \xrightarrow{m.s.} \mu$:

$$
\begin{aligned}
T\cdot E(\overline{Y}_T - \mu)^2 &= |\gamma_0 + [(T-1)/T](2\gamma_1) + [(T-2)/T](2\gamma_2) \\
&\quad + [(T-3)/T](2\gamma_3) + \cdots + [1/T](2\gamma_{T-1})| \\
&\le \{|\gamma_0| + [(T-1)/T]\cdot 2|\gamma_1| + [(T-2)/T]\cdot 2|\gamma_2| \\
&\quad + [(T-3)/T]\cdot 2|\gamma_3| + \cdots + [1/T]\cdot 2|\gamma_{T-1}|\} \\
&\le \{|\gamma_0| + 2|\gamma_1| + 2|\gamma_2| + 2|\gamma_3| + \cdots\}.
\end{aligned}
\tag{7.2.6}
$$

Hence, $T\cdot E(\overline{Y}_T - \mu)^2 < \infty$, by [7.2.3], and so $E(\overline{Y}_T - \mu)^2 \to 0$, as claimed.

It is also of interest to calculate the limiting value of $T\cdot E(\overline{Y}_T - \mu)^2$. Result [7.2.5] expresses this variance for finite T as a weighted average of the first $T-1$ autocovariances γ_j. For large j, these autocovariances approach zero and will not affect the sum. For small j, the autocovariances are given a weight that approaches unity as the sample size grows. Thus, we might guess that

$$
\lim_{T\to\infty} T\cdot E(\overline{Y}_T - \mu)^2 = \sum_{j=-\infty}^{\infty}\gamma_j = \gamma_0 + 2\gamma_1 + 2\gamma_2 + 2\gamma_3 + \cdots. \tag{7.2.7}
$$

This conjecture is indeed correct. To verify this, note that the assumption [7.2.3] means that for any $\varepsilon > 0$ there exists a q such that

$$
2|\gamma_{q+1}| + 2|\gamma_{q+2}| + 2|\gamma_{q+3}| + \cdots < \varepsilon/2.
$$

Now

$$\left|\sum_{j=-\infty}^{\infty}\gamma_j - T\cdot E(\overline{Y}_T - \mu)^2\right|$$

$$\begin{aligned}
&= |\{\gamma_0 + 2\gamma_1 + 2\gamma_2 + 2\gamma_3 + \cdots\} \\
&\quad - \{\gamma_0 + [(T-1)/T]\cdot 2\gamma_1 + [(T-2)/T]\cdot 2\gamma_2 \\
&\quad + [(T-3)/T]\cdot 2\gamma_3 + \cdots + [1/T]\cdot 2\gamma_{T-1}\}| \\
&\le (1/T)\cdot 2|\gamma_1| + (2/T)\cdot 2|\gamma_2| + (3/T)\cdot 2|\gamma_3| + \cdots \\
&\quad + (q/T)\cdot 2|\gamma_q| + 2|\gamma_{q+1}| + 2|\gamma_{q+2}| + 2|\gamma_{q+3}| + \cdots \\
&\le (1/T)\cdot 2|\gamma_1| + (2/T)\cdot 2|\gamma_2| + (3/T)\cdot 2|\gamma_3| + \cdots \\
&\quad + (q/T)\cdot 2|\gamma_q| + \varepsilon/2.
\end{aligned}$$

Moreover, for this given q, we can find an N such that

$$(1/T)\cdot 2|\gamma_1| + (2/T)\cdot 2|\gamma_2| + (3/T)\cdot 2|\gamma_3| + \cdots + (q/T)\cdot 2|\gamma_q| < \varepsilon/2$$

for all $T \ge N$, ensuring that

$$\left|\sum_{j=-\infty}^{\infty}\gamma_j - T\cdot E(\overline{Y}_T - \mu)^2\right| < \varepsilon,$$

as was to be shown.

These results can be summarized as follows.

Proposition 7.5: *Let Y_t be a covariance-stationary process with moments given by [7.2.1] and [7.2.2] and with absolutely summable autocovariances as in [7.2.3]. Then the sample mean [7.2.4] satisfies*

(*a*) $\overline{Y}_T \xrightarrow{m.s.} \mu$

(*b*) $\lim_{T\to\infty}\{T\cdot E(\overline{Y}_T - \mu)^2\} = \sum_{j=-\infty}^{\infty}\gamma_j.$

Recall from Chapter 3 that condition [7.2.3] is satisfied for any covariance-stationary $ARMA(p, q)$ process,

$$(1 - \phi_1 L - \phi_2 L^2 - \cdots - \phi_p L^p)Y_t = \mu + (1 + \theta_1 L + \theta_2 L^2 + \cdots + \theta_q L^q)\varepsilon_t,$$

with roots of $(1 - \phi_1 z - \phi_2 z^2 - \cdots - \phi_p z^p) = 0$ outside the unit circle.

Alternative expressions for the variance in result (b) of Proposition 7.5 are sometimes used. Recall that the autocovariance-generating function for Y_t is defined as

$$g_Y(z) = \sum_{j=-\infty}^{\infty}\gamma_j z^j,$$

while the spectrum is given by

$$s_Y(\omega) = \frac{1}{2\pi}g_Y(e^{-i\omega}).$$

Thus, result (b) could equivalently be described as the autocovariance-generating function evaluated at $z = 1$,

$$\sum_{j=-\infty}^{\infty} \gamma_j = g_Y(1),$$

or as 2π times the spectrum at frequency $\omega = 0$,

$$\sum_{j=-\infty}^{\infty} \gamma_j = 2\pi s_Y(0),$$

the last result coming from the fact that $e^0 = 1$. For example, consider the $MA(\infty)$ process

$$Y_t = \mu + \sum_{j=0}^{\infty} \psi_j \varepsilon_{t-j} \equiv \mu + \psi(L)\varepsilon_t$$

with $E(\varepsilon_t \varepsilon_\tau) = \sigma^2$ if $t = \tau$ and 0 otherwise and with $\Sigma_{j=0}^{\infty}|\psi_j| < \infty$. Recall that its autocovariance-generating fucntion is given by

$$g_Y(z) = \psi(z)\sigma^2\psi(z^{-1}).$$

Evaluating this at $z = 1$,

$$\sum_{j=-\infty}^{\infty} \gamma_j = \psi(1)\sigma^2\psi(1) = \sigma^2[1 + \psi_1 + \psi_2 + \psi_3 + \cdots]^2. \qquad [7.2.8]$$

Martingale Difference Sequence

Some very useful limit theorems pertain to *martingale difference sequences*. Let $\{Y_t\}_{t=1}^{\infty}$ denote a sequence of random scalars with $E(Y_t) = 0$ for all t.[6] Let Ω_t denote information available at date t, where this information includes current and lagged values of Y.[7] For example, we might have

$$\Omega_t = \{Y_t, Y_{t-1}, \ldots, Y_1, X_t, X_{t-1}, \ldots, X_1\},$$

where X_t is a second random variable. If

$$E(Y_t|\Omega_{t-1}) = 0 \qquad \text{for } t = 2, 3, \ldots, \qquad [7.2.9]$$

then $\{Y_t\}$ is said to be a *martingale difference sequence* with respect to $\{\Omega_t\}$.

Where no information set is specifies, Ω_t is presumed to consist solely of current and lagged values of Y:

$$\Omega_t = \{Y_t, Y_{t-1}, \ldots, Y_1\}.$$

Thus, if a sequence of scalars $\{Y_t\}_{t=1}^{\infty}$ satisfied $E(Y_t) = 0$ for all t and

$$E(Y_t|Y_{t-1}, Y_{t-2}, \ldots, Y_1) = 0, \qquad [7.2.10]$$

for $t = 2, 3, \ldots$, then we will say simply that $\{Y_t\}$ is a martingale difference sequence. Note that [7.2.10] is implied by [7.2.9] by the law of iterated expectations.

A sequence of $(n \times 1)$ vectors $\{\mathbf{Y}_t\}_{t=1}^{\infty}$ satisfying $E(\mathbf{Y}_t) = \mathbf{0}$ and $E(\mathbf{Y}_t|\mathbf{Y}_{t-1}, \mathbf{Y}_{t-2}, \ldots, \mathbf{Y}_1) = \mathbf{0}$ is said to form a *vector martingale difference sequence*.

[6]Wherever an expectation is indicated, it is taken as implicit that the integral exists, that is, that $E|Y_t|$ is finite.

[7]More formally, $\{\Omega_t\}_{t=1}^{\infty}$ denotes an increasing sequence of σ-fields ($\Omega_{t-1} \subset \Omega_t$) with Y_t measurable with respect to Ω_t. See, for example, White (1984, p. 56).

Note that condition [7.2.10] is stronger than the condition that Y_t is serially uncorrelated. A serially uncorrelated sequence cannot be forecast on the basis of a linear function of its past values. No function of past values, linear or nonlinear, can forecast a martingale difference sequence. While stronger than absence of serial correlation, the martingale difference condition is weaker than independence, since it does not rule out the possibility that higher moments such as $E(Y_t^2|Y_{t-1}, Y_{t-2}, \ldots, Y_1)$ might depend on past Y's.

Example 7.8
If $\varepsilon_t \sim$ i.i.d. $N(0, \sigma^2)$, then $Y_t = \varepsilon_t\varepsilon_{t-1}$ is a martingale difference sequence but not serially independent.

L^1-*Mixingales*

A more general class of processes known as L^1-*mixingales* was introduced by Andrews (1988). Consider a sequence of random variables $\{Y_t\}_{t=1}^{\infty}$ with $E(Y_t) = 0$ for $t = 1, 2, \ldots$. Let Ω_t denote information available at time t, as before, where Ω_t includes current and lagged values of Y. Suppose that we can find sequences of nonnegative deterministic constants $\{c_t\}_{t=1}^{\infty}$ and $\{\xi_m\}_{m=0}^{\infty}$ such that $\lim_{m\to\infty} \xi_m = 0$ and

$$E\left|E(Y_t|\Omega_{t-m})\right| \leq c_t\xi_m \qquad [7.2.11]$$

for all $t \geq 1$ and all $m \geq 0$. Then $\{Y_t\}$ is said to follow an L^1-mixingale with respect to $\{\Omega_t\}$.

Thus, a zero-mean process for which the m-period-ahead forecast $E(Y_t|\Omega_{t-m})$ converges (in absolute expected value) to the unconditional mean of zero is described as an L^1-mixingale.

Example 7.9
Let $\{Y_t\}$ be a martingale difference sequence. Let $c_t = E|Y_t|$, and choose $\xi_0 = 1$ and $\xi_m = 0$ for $m = 1, 2, \ldots$. Then [7.2.11] is satisfied for $\Omega_t = \{Y_t, Y_{t-1}, \ldots, Y_1\}$, so that $\{Y_t\}$ could be described as an L^1-mixingale sequence.

Example 7.10
Let $Y_t = \Sigma_{j=0}^{\infty}\psi_j\varepsilon_{t-j}$, where $\Sigma_{j=0}^{\infty}|\psi_j| < \infty$ and $\{\varepsilon_t\}$ is a martingale difference sequence with $E|\varepsilon_t| < M$ for all t for some $M < \infty$. Then $\{Y_t\}$ is an L^1-mixingale with respect to $\Omega_t = \{\varepsilon_t, \varepsilon_{t-1}, \ldots\}$. To see this, notice that

$$E\left|E(Y_t|\varepsilon_{t-m}, \varepsilon_{t-m-1}, \ldots)\right| = E\left|\sum_{j=m}^{\infty}\psi_j\varepsilon_{t-j}\right| \leq E\left\{\sum_{j=m}^{\infty}|\psi_j\varepsilon_{t-j}|\right\}.$$

Since $\{\psi_j\}_{j=0}^{\infty}$ is absolutely summable and $E|\varepsilon_{t-j}| < M$, we can interchange the order of expectation and summation:

$$E\left\{\sum_{j=m}^{\infty}|\psi_j\varepsilon_{t-j}|\right\} = \sum_{j=m}^{\infty}|\psi_j|\cdot E|\varepsilon_{t-j}| \leq \sum_{j=m}^{\infty}|\psi_j|\cdot M.$$

Then [7.2.11] is satisfied with $c_t = M$ and $\xi_m = \Sigma_{j=m}^{\infty}|\psi_j|$. Moreover, $\lim_{m\to\infty}\xi_m = 0$, because of absolute summability of $\{\psi_j\}_{j=0}^{\infty}$. Hence, $\{Y_t\}$ is an L^1-mixingale.

Law of Large Numbers For L^1-Mixingales

Andrews (1988) derived the following law of large numbers for L^1-mixingales.[8]

Proposition 7.6: *Let $\{Y_t\}$ be an L^1-mixingale. If (a) $\{Y_t\}$ is uniformly integrable and (b) there exists a choice for $\{c_t\}$ such that*

$$\lim_{T\to\infty} (1/T)\sum_{t=1}^{T} c_t < \infty,$$

then $(1/T)\Sigma_{t=1}^{T} Y_t \xrightarrow{p} 0$.

To apply this result, we need to verify that a sequence is uniformly integrable. A sequence $\{Y_t\}$ is said to be *uniformly integrable* if for every $\varepsilon > 0$ there exists a number $c > 0$ such that

$$E(|Y_t|\cdot\delta_{[|Y_t|\geq c]}) < \varepsilon \qquad [7.2.12]$$

for all t, where $\delta_{[|Y_t|\geq c]} = 1$ if $|Y_t| \geq c$ and 0 otherwise. The following proposition gives sufficient conditions for uniform integrability.

Proposition 7.7: *(a) Suppose there exist an $r > 1$ and an $M' < \infty$ such that $E(|Y_t|^r) < M'$ for all t. Then $\{Y_t\}$ is uniformly integrable. (b) Suppose there exist an $r > 1$ and an $M' < \infty$ such that $E(|X_t|^r) < M'$ for all t. If $Y_t = \Sigma_{j=-\infty}^{\infty} h_j X_{t-j}$ with $\Sigma_{j=-\infty}^{\infty} |h_j| < \infty$, then $\{Y_t\}$ is uniformly integrable.*

Condition (a) requires us to find a moment higher than the first that exists. Typically, we would use $r = 2$. However, even if a variable has infinite variance, it can still be uniformly integrable as long as $E|Y_t|^r$ exists for some r between 1 and 2.

Example 7.11

Let $\overline{Y}_T$ be the sample mean from a martingale difference sequence, $\overline{Y}_T = (1/T)\Sigma_{t=1}^{T} Y_t$ with $E|Y_t|^r < M'$ for some $r > 1$ and $M' < \infty$. Note that this also implies that there exists an $M < \infty$ such that $E|Y_t| < M$. From Proposition 7.7(a), $\{Y_t\}$ is uniformly integrable. Moreover, from Example 7.9, $\{Y_t\}$ can be viewed as an L^1-mixingale with $c_t = M$. Thus, $\lim_{T\to\infty} (1/T)\Sigma_{t=1}^{T} c_t = M < \infty$, and so, from Proposition 7.6, $\overline{Y}_T \xrightarrow{p} 0$.

Example 7.12

Let $Y_t = \Sigma_{j=0}^{\infty}\psi_j\varepsilon_{t-j}$, where $\Sigma_{j=0}^{\infty}|\psi_j| < \infty$ and $\{\varepsilon_t\}$ is a martingale difference sequence with $E|\varepsilon_t|^r < M' < \infty$ for some $r > 1$ and some $M' < \infty$. Then, from Proposition 7.7(b), $\{Y_t\}$ is uniformly integrable. Moreover, from Example 7.10, $\{Y_t\}$ is an L^1-mixingale with $c_t = M$, where M represents the largest value of $E|\varepsilon_t|$ for any t. Then $\lim_{T\to\infty} (1/T)\Sigma_{t=1}^{T} c_t = M < \infty$, establishing again that $\overline{Y}_T \xrightarrow{p} 0$.

Proposition 7.6 can also be applied to a double-indexed array $\{Y_{t,T}\}$; that is, each sample size T can be associated with a different sequence $\{Y_{1,T}, Y_{2,T}, \ldots, Y_{T,T}\}$. The array is said to be an L^1-mixingale with respect to an information set $\Omega_{t,T}$ that includes $\{Y_{1,T}, Y_{2,T}, \ldots, Y_{T,T}\}$ if there exist nonnegative constants ξ_m and $c_{t,T}$ such that $\lim_{m\to\infty} \xi_m = 0$ and

$$E|E(Y_{t,T}|\Omega_{t-m,T})| \leq c_{t,T}\xi_m$$

[8]Andrews replaced part (b) of the proposition with the weaker condition $\overline{\lim}_{T\to\infty} (1/T)\,\Sigma_{t=1}^{T} c_t < \infty$. See Royden (1968, p. 36) on the relation between "lim" and "$\overline{\lim}$."

for all $m \geq 0$, $T \geq 1$, and $t = 1, 2, \ldots, T$. If the array is uniformly integrable with $\lim_{T\to\infty} (1/T)\Sigma_{t=1}^{T} c_{t,T} < \infty$, then $(1/T)\Sigma_{t=1}^{T} Y_{t,T} \xrightarrow{p} 0$.

Example 7.13

Let $\{\varepsilon_t\}_{t=1}^{\infty}$ be a martingale difference sequence with $E|\varepsilon_t|^r < M'$ for some $r > 1$ and $M' < \infty$, and define $Y_{t,T} \equiv (t/T)\varepsilon_t$. Then the array $\{Y_{t,T}\}$ is a uniformly integrable L^1-mixingale with $c_{t,T} = M$, where M denotes the maximal value for $E|\varepsilon_t|$, $\xi_0 = 1$, and $\xi_m = 0$ for $m > 0$. Hence, $(1/T)\Sigma_{t=1}^{T}(t/T)\varepsilon_t \xrightarrow{p} 0$.

Consistent Estimation of Second Moments

Next consider the conditions under which

$$(1/T)\sum_{t=1}^{T} Y_t Y_{t-k} \xrightarrow{p} E(Y_t Y_{t-k})$$

(for notational simplicity, we assume here that the sample consists of $T + k$ observations on Y). Suppose that $Y_t = \Sigma_{j=0}^{\infty}\psi_j\varepsilon_{t-j}$, where $\Sigma_{j=0}^{\infty}|\psi_j| < \infty$ and $\{\varepsilon_t\}$ is an i.i.d. sequence with $E|\varepsilon_t|^r < \infty$ for some $r > 2$. Note that the population second moment can be written[9]

$$\begin{aligned} E(Y_t Y_{t-k}) &= E\left(\sum_{u=0}^{\infty}\psi_u\varepsilon_{t-u}\right)\left(\sum_{v=0}^{\infty}\psi_v\varepsilon_{t-k-v}\right) \\ &= E\left(\sum_{u=0}^{\infty}\sum_{v=0}^{\infty}\psi_u\psi_v\varepsilon_{t-u}\varepsilon_{t-k-v}\right) \qquad [7.2.13] \\ &= \sum_{u=0}^{\infty}\sum_{v=0}^{\infty}\psi_u\psi_v E(\varepsilon_{t-u}\varepsilon_{t-k-v}). \end{aligned}$$

Define $X_{t,k}$ to be the following random variable:

$$\begin{aligned} X_{t,k} &\equiv Y_t Y_{t-k} - E(Y_t Y_{t-k}) \\ &= \left(\sum_{u=0}^{\infty}\sum_{v=0}^{\infty}\psi_u\psi_v\varepsilon_{t-u}\varepsilon_{t-k-v}\right) - \left(\sum_{u=0}^{\infty}\sum_{v=0}^{\infty}\psi_u\psi_v E(\varepsilon_{t-u}\varepsilon_{t-k-v})\right) \\ &= \sum_{u=0}^{\infty}\sum_{v=0}^{\infty}\psi_u\psi_v[\varepsilon_{t-u}\varepsilon_{t-k-v} - E(\varepsilon_{t-u}\varepsilon_{t-k-v})]. \end{aligned}$$

Consider a forecast of $X_{t,k}$ on the basis of $\Omega_{t-m} \equiv \{\varepsilon_{t-m}, \varepsilon_{t-m-1}, \ldots\}$ for $m > k$:

$$E(X_{t,k}|\Omega_{t-m}) = \sum_{u=m}^{\infty}\sum_{v=m-k}^{\infty}\psi_u\psi_v[\varepsilon_{t-u}\varepsilon_{t-k-v} - E(\varepsilon_{t-u}\varepsilon_{t-k-v})].$$

[9]Notice that

$$\sum_{u=0}^{\infty}\sum_{v=0}^{\infty}|\psi_u\psi_v| = \sum_{u=0}^{\infty}|\psi_u|\sum_{v=0}^{\infty}|\psi_v| < \infty$$

and $E|\varepsilon_{t-u}\varepsilon_{t-k-v}| < \infty$, permitting us to move the expectation operator inside the summation signs in the last line of [7.2.13].

The expected absolute value of this forecast is bounded by

$$E\left|E(X_{t,k}|\Omega_{t-m})\right| = E\left|\sum_{u=m}^{\infty}\sum_{v=m-k}^{\infty}\psi_u\psi_v[\varepsilon_{t-u}\varepsilon_{t-k-v} - E(\varepsilon_{t-u}\varepsilon_{t-k-v})]\right|$$

$$\leq E\left(\sum_{u=m}^{\infty}\sum_{v=m-k}^{\infty}|\psi_u\psi_v|\cdot|\varepsilon_{t-u}\varepsilon_{t-k-v} - E(\varepsilon_{t-u}\varepsilon_{t-k-v})|\right)$$

$$\leq \sum_{u=m}^{\infty}\sum_{v=m-k}^{\infty}|\psi_u\psi_v|\cdot M$$

for some $M < \infty$. Define

$$\xi_m \equiv \sum_{u=m}^{\infty}\sum_{v=m-k}^{\infty}|\psi_u\psi_v| = \sum_{u=m}^{\infty}|\psi_u|\sum_{v=m-k}^{\infty}|\psi_v|.$$

Since $\{\psi_j\}_{j=0}^{\infty}$ is absolutely summable, $\lim_{m\to\infty}\Sigma_{u=m}^{\infty}|\psi_u| = 0$ and $\lim_{m\to\infty}\xi_m = 0$. It follows that $X_{t,k}$ is an L^1-mixingale with respect to Ω_t with coefficient $c_t = M$. Moreover, $X_{t,k}$ is uniformly integrable, from a simple adaptation of the argument in Proposition 7.7(b) (see Exercise 7.5). Hence,

$$(1/T)\sum_{t=1}^{T}X_{t,k} = (1/T)\sum_{t=1}^{T}[Y_tY_{t-k} - E(Y_tY_{t-k})] \xrightarrow{p} 0,$$

from which

$$(1/T)\sum_{t=1}^{T}Y_tY_{t-k} \xrightarrow{p} E(Y_tY_{t-k}). \qquad [7.2.14]$$

It is straightforward to deduce from [7.2.14] that the jth sample autocovariance for a sample of size T gives a consistent estimate of the population autocovariance,

$$(1/T)\sum_{t=k+1}^{T}(Y_t - \overline{Y}_T)(Y_{t-k} - \overline{Y}_T) \xrightarrow{p} E(Y_t - \mu)(Y_{t-k} - \mu), \qquad [7.2.15]$$

where $\overline{Y}_T = (1/T)\Sigma_{t=1}^{T}Y_t$; see Exercise 7.6.

Central Limit Theorem for a Martingale Difference Sequence

Next we consider the asymptotic distribution of $\sqrt{T}$ times the sample mean. The following version of the central limit theorem can often be applied.

Proposition 7.8: *(White, 1984, Corollary 5.25, p. 130). Let $\{Y_t\}_{t=1}^{\infty}$ be a scalar martingale difference sequence with $\overline{Y}_T = (1/T)\Sigma_{t=1}^{T}Y_t$. Suppose that (a) $E(Y_t^2) = \sigma_t^2 > 0$ with $(1/T)\Sigma_{t=1}^{T}\sigma_t^2 \to \sigma^2 > 0$, (b) $E|Y_t|^r < \infty$ for some $r > 2$ and all t, and (c) $(1/T)\Sigma_{t=1}^{T}Y_t^2 \xrightarrow{p} \sigma^2$. Then $\sqrt{T}\,\overline{Y}_T \xrightarrow{L} N(0, \sigma^2)$.*

Again, Proposition 7.8 can be extended to arrays $\{Y_{t,T}\}$ as follows. Let $\{Y_{t,T}\}_{t=1}^{T}$ be a martingale difference sequence with $E(Y_{t,T}^2) = \sigma_{t,T}^2 > 0$. Let $\{Y_{t,T+1}\}_{t=1}^{T+1}$ be a potentially different martingale difference sequence with $E(Y_{t,T+1}^2) = \sigma_{t,T+1}^2 > 0$. If (a) $(1/T)\Sigma_{t=1}^{T}\sigma_{t,T}^2 \to \sigma^2$, (b) $E|Y_{t,T}|^r < \infty$ for some $r > 2$ and all t and T, and (c) $(1/T)\Sigma_{t=1}^{T}Y_{t,T}^2 \xrightarrow{p} \sigma^2$, then $\sqrt{T}\,\overline{Y}_T \xrightarrow{L} N(0, \sigma^2)$.

Proposition 7.8 also readily generalizes to vector martingale difference sequences.

Proposition 7.9: Let $\{\mathbf{Y}_t\}_{t=1}^{\infty}$ be an n-dimensional vector martingale difference sequence with $\bar{\mathbf{Y}}_T = (1/T)\Sigma_{t=1}^{T}\mathbf{Y}_t$. Suppose that (a) $E(\mathbf{Y}_t\mathbf{Y}_t') = \mathbf{\Omega}_t$, a positive definite matrix with $(1/T)\Sigma_{t=1}^{T}\mathbf{\Omega}_t \to \mathbf{\Omega}$, a positive definite matrix; (b) $E(Y_{it}Y_{jt}Y_{lt}Y_{mt}) < \infty$ for all t and all i, j, l, and m (including $i = j = l = m$), where Y_{it} is the ith element of the vector $\mathbf{Y}_t$; and (c) $(1/T)\Sigma_{t=1}^{T}\mathbf{Y}_t\mathbf{Y}_t' \xrightarrow{p} \mathbf{\Omega}$. Then $\sqrt{T}\,\bar{\mathbf{Y}}_T \xrightarrow{L} N(\mathbf{0}, \mathbf{\Omega})$.

Again, Proposition 7.9 holds for arrays $\{\mathbf{Y}_{t,T}\}_{t=1}^{T}$ satisfying the stated conditions.

To apply Proposition 7.9, we will often need to assume that a certain process has finite fourth moments. The following result can be useful for this purpose.

Proposition 7.10: Let X_t be a strictly stationary stochastic process with $E(X_t^4) = \mu_4 < \infty$. Let $Y_t = \Sigma_{j=0}^{\infty}h_jX_{t-j}$, where $\Sigma_{j=0}^{\infty}|h_j| < \infty$. Then Y_t is a strictly stationary stochastic process with $E|Y_tY_sY_uY_v| < \infty$ for all t, s, u, and v.

Example 7.14

Let $Y_t = \phi_1Y_{t-1} + \phi_2Y_{t-2} + \cdots + \phi_pY_{t-p} + \varepsilon_t$, where $\{\varepsilon_t\}$ is an i.i.d. sequence and where roots of $(1 - \phi_1z - \phi_2z^2 - \cdots - \phi_pz^p) = 0$ lie outside the unit circle. We saw in Chapter 3 that Y_t can be written as $\Sigma_{j=0}^{\infty}\psi_j\varepsilon_{t-j}$ with $\Sigma_{j=0}^{\infty}|\psi_j| < \infty$. Proposition 7.10 states that if ε_t has finite fourth moments, then so does Y_t.

Example 7.15

Let $Y_t = \Sigma_{j=0}^{\infty}\psi_j\varepsilon_{t-j}$ with $\Sigma_{j=0}^{\infty}|\psi_j| < \infty$ and ε_t i.i.d. with $E(\varepsilon_t) = 0$, $E(\varepsilon_t^2) = \sigma^2$, and $E(\varepsilon_t^4) < \infty$. Consider the random variable X_t defined by $X_t \equiv \varepsilon_tY_{t-k}$ for $k > 0$. Then X_t is a martingale difference sequence with variance $E(X_t^2) = \sigma^2\cdot E(Y_t^2)$ and with fourth moment $E(\varepsilon_t^4)\cdot E(Y_t^4) < \infty$, by Example 7.14. Hence, if we can show that

$$(1/T)\sum_{t=1}^{T}X_t^2 \xrightarrow{p} E(X_t^2), \qquad [7.2.16]$$

then Proposition 7.8 can be applied to deduce that

$$(1/\sqrt{T})\sum_{t=1}^{T}X_t \xrightarrow{L} N\Big(0, E(X_t^2)\Big)$$

or

$$(1/\sqrt{T})\sum_{t=1}^{T}\varepsilon_tY_{t-k} \xrightarrow{L} N\Big(0, \sigma^2\cdot E(Y_t^2)\Big). \qquad [7.2.17]$$

To verify [7.2.16], notice that

$$\begin{aligned}(1/T)\sum_{t=1}^{T}X_t^2 &= (1/T)\sum_{t=1}^{T}\varepsilon_t^2Y_{t-k}^2 \\ &= (1/T)\sum_{t=1}^{T}(\varepsilon_t^2 - \sigma^2)Y_{t-k}^2 + (1/T)\sum_{t=1}^{T}\sigma^2Y_{t-k}^2.\end{aligned} \qquad [7.2.18]$$

But $(\varepsilon_t^2 - \sigma^2)Y_{t-k}^2$ is a martingale difference sequence with finite second moment, so, from Example 7.11,

$$(1/T)\sum_{t=1}^{T}(\varepsilon_t^2 - \sigma^2)Y_{t-k}^2 \xrightarrow{p} 0.$$

It further follows from result [7.2.14] that

$$(1/T)\sum_{t=1}^{T}\sigma^2 Y_{t-k}^2 \xrightarrow{p} \sigma^2 \cdot E(Y_t^2).$$

Thus, [7.2.18] implies

$$(1/T)\sum_{t=1}^{T} X_t^2 \xrightarrow{p} \sigma^2 \cdot E(Y_t^2),$$

as claimed in [7.2.16].

Central Limit Theorem for Stationary Stochastic Processes

We now present a central limit theorem for a serially correlated sequence. Recall from Proposition 7.5 that the sample mean has asymptotic variance given by $(1/T)\Sigma_{j=-\infty}^{\infty}\gamma_j$. Thus, we would expect the central limit theorem to take the form $\sqrt{T}(\bar{Y}_T - \mu) \xrightarrow{L} N(0, \Sigma_{j=-\infty}^{\infty}\gamma_j)$. The next proposition gives a result of this type.

Proposition 7.11: *(Anderson, 1971, p. 429). Let*

$$Y_t = \mu + \sum_{j=0}^{\infty}\psi_j\varepsilon_{t-j},$$

where $\{\varepsilon_t\}$ is a sequence of i.i.d. random variables with $E(\varepsilon_t^2) < \infty$ and $\Sigma_{j=0}^{\infty}|\psi_j| < \infty$. Then

$$\sqrt{T}(\bar{Y}_T - \mu) \xrightarrow{L} N(0, \sum_{j=-\infty}^{\infty}\gamma_j). \qquad [7.2.19]$$

A version of [7.2.19] can also be developed for $\{\varepsilon_t\}$ a martingale difference sequence satisfying certain restrictions; see Phillips and Solo (1992).

APPENDIX 7.A. *Proofs of Chapter 7 Propositions*

■ **Proof of Proposition 7.1.** Let $g_j(\mathbf{c})$ denote the jth element of $\mathbf{g}(\mathbf{c})$, g_j: $\mathbb{R}^n \to \mathbb{R}^1$. We need to show that for any $\delta > 0$ and $\varepsilon > 0$ there exists an N such that for all $T \geq N$,

$$P\{|g_j(\mathbf{X}_T) - g_j(\mathbf{c})| > \delta\} < \varepsilon. \qquad [7.A.1]$$

Continuity of $g_j(\cdot)$ implies that there exists an η such that $|g_j(\mathbf{X}_T) - g_j(\mathbf{c})| > \delta$ only if

$$[(X_{1T} - c_1)^2 + (X_{2T} - c_2)^2 + \cdots + (X_{nT} - c_n)^2] > \eta^2. \qquad [7.A.2]$$

This would be the case only if $(X_{iT} - c_i)^2 > \eta^2/n$ for some i. But from the fact that plim $X_{iT} = c_i$, for any i and specified values of ε and η we can find a value of N such that

$$P\{|X_{iT} - c_i| > \eta/\sqrt{n}\} < \varepsilon/n$$

for all $T \geq N$.

Recall the elementary addition rule for the probability of any events A and B,

$$P\{A \text{ or } B\} \leq P\{A\} + P\{B\},$$

from which it follows that

$$P\{(|X_{1T} - c_1| > \eta/\sqrt{n}) \text{ or } (|X_{2T} - c_2| > \eta/\sqrt{n}) \text{ or } \cdots \text{ or } (|X_{nT} - c_n| > \eta/\sqrt{n})\}$$
$$< (\varepsilon/n) + (\varepsilon/n) + \cdots + (\varepsilon/n).$$

Hence,

$$P\{[(X_{1T} - c_1)^2 + (X_{2T} - c_2)^2 + \cdots + (X_{nT} - c_n)^2] > \eta^2\} < \varepsilon$$

for all $T \geq N$. Since [7.A.2] was a necessary condition for $|g_j(\mathbf{X}_T) - g_j(\mathbf{c})|$ to be greater than δ, it follows that the probability that $|g_j(\mathbf{X}_T) - g_j(\mathbf{c})|$ is greater than δ is less than ε, which was to be shown. ■

■ **Proof of Proposition 7.2.** Let S denote the set of all x such that $|x - c| > \delta$, and let $\bar{S}$ denote its complement (all x such that $|x - c| \leq \delta$). Then, for $f_X(x)$ the density of x,

$$\begin{aligned} E|X - c|^r &= \int |x - c|^r f_X(x)\, dx \\ &= \int_S |x - c|^r f_X(x)\, dx + \int_{\bar{S}} |x - c|^r f_X(x)\, dx \\ &\geq \int_S |x - c|^r f_X(x)\, dx \\ &\geq \int_S \delta^r f_X(x)\, dx \\ &= \delta^r P\{|X - c| > \delta\}, \end{aligned}$$

so that

$$E|X - c|^r \geq \delta^r P\{|X - c| > \delta\},$$

as claimed. ■

■ **Proof of Proposition 7.4.** Consider any real $(m \times 1)$ vector $\boldsymbol{\lambda}$, and form the function $h\colon \mathbb{R}^n \to \mathbb{R}^1$ defined by $h(\mathbf{x}) \equiv \boldsymbol{\lambda}'\mathbf{g}(\mathbf{x})$, noting that $h(\cdot)$ is differentiable. The *mean-value theorem* states that for a differentiable function $h(\cdot)$, there exists an $(n \times 1)$ vector $\mathbf{c}_T$ between $\mathbf{X}_T$ and $\mathbf{c}$ such that[10]

$$h(\mathbf{X}_T) - h(\mathbf{c}) = \left.\frac{\partial h(\mathbf{x})}{\partial \mathbf{x}'}\right|_{\mathbf{x}=\mathbf{c}_T} \times (\mathbf{X}_T - \mathbf{c})$$

and therefore

$$\sqrt{T}\,[h(\mathbf{X}_T) - h(\mathbf{c})] = \left.\frac{\partial h(\mathbf{x})}{\partial \mathbf{x}'}\right|_{\mathbf{x}=\mathbf{c}_T} \times \sqrt{T}(\mathbf{X}_T - \mathbf{c}). \qquad [7.A.3]$$

Since $\mathbf{c}_T$ is between $\mathbf{X}_T$ and $\mathbf{c}$ and since $\mathbf{X}_T \xrightarrow{p} \mathbf{c}$, we know that $\mathbf{c}_T \xrightarrow{p} \mathbf{c}$. Moreover, the derivative $\partial h(\mathbf{x})/\partial \mathbf{x}'$ is itself a continuous function of $\mathbf{x}$. Thus, from Proposition 7.1,

$$\left.\frac{\partial h(\mathbf{x})}{\partial \mathbf{x}'}\right|_{\mathbf{x}=\mathbf{c}_T} \xrightarrow{p} \left.\frac{\partial h(\mathbf{x})}{\partial \mathbf{x}'}\right|_{\mathbf{x}=\mathbf{c}}.$$

Given that $\sqrt{T}(\mathbf{X}_T - \mathbf{c}) \xrightarrow{L} \mathbf{X}$, Proposition 7.3(b) applied to expression [7.A.3] gives

$$\sqrt{T}\,[h(\mathbf{X}_T) - h(\mathbf{c})] \xrightarrow{L} \left.\frac{\partial h(\mathbf{x})}{\partial \mathbf{x}'}\right|_{\mathbf{x}=\mathbf{c}} \mathbf{X},$$

or, in terms of the original function $\mathbf{g}(\cdot)$,

$$\boldsymbol{\lambda}'\{\sqrt{T}\,[\mathbf{g}(\mathbf{X}_T) - \mathbf{g}(\mathbf{c})]\} \xrightarrow{L} \boldsymbol{\lambda}' \left.\frac{\partial \mathbf{g}(\mathbf{x})}{\partial \mathbf{x}'}\right|_{\mathbf{x}=\mathbf{c}} \mathbf{X}.$$

Since this is true for any $\boldsymbol{\lambda}$, we conclude that

$$\sqrt{T}\,[\mathbf{g}(\mathbf{X}_T) - \mathbf{g}(\mathbf{c})] \xrightarrow{L} \left.\frac{\partial \mathbf{g}(\mathbf{x})}{\partial \mathbf{x}'}\right|_{\mathbf{x}=\mathbf{c}} \mathbf{X},$$

as claimed. ■

[10]That is, for any given $\mathbf{X}_T$ there exists a scalar μ_T with $0 \leq \mu_T \leq 1$ such that $\mathbf{c}_T = \mu_T\mathbf{X}_T + (1 - \mu_T)\mathbf{c}$. See, for example, Marsden (1974, pp. 174–75).

■ **Proof of Proposition 7.7.** Part (a) is established as in Andrews (1988, p. 463) using *Hölder's inequality* (see, for example, White, 1984, p. 30), which states that for $r > 1$, if $E[|Y|^r] < \infty$ and $E[|W|^{r/(r-1)}] < \infty$, then

$$E|YW| \leq \{E[|Y|^r]\}^{1/r} \times \{E[|W|^{r/(r-1)}]\}^{(r-1)/r}.$$

This implies that

$$E(|Y_t|\cdot\delta_{[|Y_t|\geq c]}) \leq \{E[|Y_t|^r]\}^{1/r} \times \{E[(\delta_{[|Y_t|\geq c]})^{r/(r-1)}]\}^{(r-1)/r}. \qquad [7.A.4]$$

Since $\delta_{[|Y_t|\geq c]}$ is either 0 or 1, it follows that

$$(\delta_{[|Y_t|\geq c]})^{r/(r-1)} = \delta_{[|Y_t|\geq c]}$$

and so

$$E[(\delta_{[|Y_t|\geq c]})^{r/(r-1)}] = E[\delta_{[|Y_t|\geq c]}] = \int_{|Y_t|\geq c} 1\cdot f_{Y_t}(y_t)\, dy_t = P\{|Y_t| \geq c\} \leq \frac{E|Y_t|}{c}, \qquad [7.A.5]$$

where the last result follows from Chebyshev's inequality. Substituting [7.A.5] into [7.A.4],

$$E(|Y_t|\cdot\delta_{[|Y_t|\geq c]}) \leq \{E[|Y_t|^r]\}^{1/r} \times \left\{\frac{E|Y_t|}{c}\right\}^{(r-1)/r}. \qquad [7.A.6]$$

Recall that $E[|Y_t|^r] < M'$ for all t, implying that there also exists an $M < \infty$ such that $E|Y_t| < M$ for all t. Hence,

$$E(|Y_t|\cdot\delta_{[|Y_t|\geq c]}) \leq (M')^{1/r} \times (M/c)^{(r-1)/r}.$$

This expression can be made as small as desired by choosing c sufficiently large. Thus condition [7.2.12] holds, ensuring that $\{Y_t\}$ is uniformly integrable.

To establish (b), notice that

$$E(|Y_t|\cdot\delta_{[|Y_t|\geq c]}) = E\left|\sum_{j=-\infty}^{\infty} h_j X_{t-j}\cdot\delta_{[|Y_t|\geq c]}\right| \leq E\left\{\sum_{j=-\infty}^{\infty} |h_j|\cdot|X_{t-j}|\cdot\delta_{[|Y_t|\geq c]}\right\}. \qquad [7.A.7]$$

Since $E[|X_{t-j}|^r] < M'$ and since $\delta_{[|Y_t|\geq c]} \leq 1$, it follows that $E\{|X_{t-j}|\cdot\delta_{[|Y_t|\geq c]}\}$ is bounded. Since $\{h_j\}_{j=-\infty}^{\infty}$ is absolutely summable, we can bring the expectation operator inside the summation in the last expression of [7.A.7] to deduce that

$$\begin{aligned} E\left\{\sum_{j=-\infty}^{\infty} |h_j|\cdot|X_{t-j}|\cdot\delta_{[|Y_t|\geq c]}\right\} &= \sum_{j=-\infty}^{\infty} |h_j|\cdot E\{|X_{t-j}|\cdot\delta_{[|Y_t|\geq c]}\} \\ &\leq \sum_{j=-\infty}^{\infty} |h_j|\cdot\{E[|X_{t-j}|^r]\}^{1/r} \times \left\{\frac{E|Y_t|}{c}\right\}^{(r-1)/r}, \end{aligned}$$

where the last inequality follows from the same arguments as in [7.A.6]. Hence, [7.A.7] becomes

$$E(|Y_t|\cdot\delta_{[|Y_t|\geq c]}) \leq \sum_{j=-\infty}^{\infty} |h_j| \times (M')^{1/r} \times \left\{\frac{E|Y_t|}{c}\right\}^{(r-1)/r}. \qquad [7.A.8]$$

But certainly, $E|Y_t|$ is bounded:

$$E|Y_t| = E\left|\sum_{j=-\infty}^{\infty} h_j X_{t-j}\right| \leq \sum_{j=-\infty}^{\infty} |h_j|\cdot E|X_{t-j}| = K < \infty.$$

Thus, from [7.A.8],

$$E(|Y_t|\cdot\delta_{[|Y_t|\geq c]}) \leq (M')^{1/r}(K/c)^{(r-1)/r} \sum_{j=-\infty}^{\infty} |h_j|. \qquad [7.A.9]$$

Since $\Sigma_{j=-\infty}^{\infty}|h_j|$ is finite, [7.A.9] can again be made as small as desired by choosing c sufficiently large. ■

■ **Proof of Proposition 7.9.** Consider $Y_t = \boldsymbol{\lambda}'\mathbf{Y}_t$ for $\boldsymbol{\lambda}$ any real $(n \times 1)$ vector. Then Y_t is a martingale difference sequence. We next verify that each of the conditions of Proposition

7.8 is satisfied. (a) $E(Y_t^2) = \boldsymbol{\lambda}'\boldsymbol{\Omega}_t\boldsymbol{\lambda} \equiv \sigma_t^2 > 0$, by positive definiteness of $\boldsymbol{\Omega}_t$. Likewise,

$$(1/T)\sum_{t=1}^{T}\sigma_t^2 = \boldsymbol{\lambda}'(1/T)\sum_{t=1}^{T}\boldsymbol{\Omega}_t\boldsymbol{\lambda} \to \boldsymbol{\lambda}'\boldsymbol{\Omega}\boldsymbol{\lambda} \equiv \sigma^2,$$

with $\sigma^2 > 0$, by positive definiteness of $\boldsymbol{\Omega}$. (b) $E(Y_t^4)$ is a finite sum of terms of the form $\lambda_i\lambda_j\lambda_l\lambda_m E(Y_{it}Y_{jt}Y_{lt}Y_{mt})$ and so is bounded for all t by condition (b) of Proposition 7.9; hence, Y_t satisfies condition (b) of Proposition 7.8 for $r = 4$. (c) Define $S_T \equiv (1/T) \times \Sigma_{t=1}^{T} Y_t^2$ and $\mathbf{S}_T \equiv (1/T)\Sigma_{t=1}^{T}\mathbf{Y}_t\mathbf{Y}_t'$, noticing that $S_T = \boldsymbol{\lambda}'\mathbf{S}_T\boldsymbol{\lambda}$. Since S_T is a continuous function of $\mathbf{S}_T$, we know that $\operatorname{plim} S_T = \boldsymbol{\lambda}'\boldsymbol{\Omega}\boldsymbol{\lambda} \equiv \sigma^2$, where $\boldsymbol{\Omega}$ is given as the plim of $\mathbf{S}_T$. Thus Y_t satisfies conditions (a) through (c) of Proposition 7.8, and so $\sqrt{T}\,\bar{Y}_T \xrightarrow{L} N(0, \sigma^2)$, or $\sqrt{T}\,\bar{Y}_T \xrightarrow{L} \boldsymbol{\lambda}'\mathbf{Y}$, where $\mathbf{Y} \sim (\mathbf{0}, \boldsymbol{\Omega})$. Since this is true for any $\boldsymbol{\lambda}$, this confirms the claim that $\sqrt{T}\,\bar{\mathbf{Y}}_T \xrightarrow{L} N(\mathbf{0}, \boldsymbol{\Omega})$. ■

■ **Proof of Proposition 7.10.** Let $Y \equiv X_tX_s$ and $W \equiv X_uX_v$. Then Hölder's inequality implies that for $r > 1$,

$$E|X_tX_sX_uX_v| \le \{E|X_tX_s|^r\}^{1/r} \times \{E|X_uX_v|^{r/(r-1)}\}^{(r-1)/r}.$$

For $r = 2$, this means

$$E|X_tX_sX_uX_v| \le \{E(X_tX_s)^2\}^{1/2} \times \{E(X_uX_v)^2\}^{1/2} \le \max\{E(X_tX_s)^2, E(X_uX_v)^2\}.$$

A second application of Hölder's inequality with $Y \equiv X_t^2$ and $W \equiv X_t^2$ reveals that

$$E(X_tX_s)^2 = E(X_t^2X_s^2) \le \{E(X_t^2)^r\}^{1/r} \times \{E(X_s^2)^{r/(r-1)}\}^{(r-1)/r}.$$

Again for $r = 2$, this implies from the strict stationarity of $\{X_t\}$ that

$$E(X_tX_s)^2 \le E(X_t^4).$$

Hence, if $\{X_t\}$ is strictly stationary with finite fourth moment, then

$$E|X_tX_sX_uX_v| \le E(X_t^4) = \mu_4$$

for all t, s, u, and v.

Observe further that

$$\begin{aligned}
E|Y_tY_sY_uY_v| &= E\left|\sum_{i=0}^{\infty}h_iX_{t-i}\sum_{j=0}^{\infty}h_jX_{s-j}\sum_{l=0}^{\infty}h_lX_{u-l}\sum_{m=0}^{\infty}h_mX_{v-m}\right| \\
&= E\left|\sum_{i=0}^{\infty}\sum_{j=0}^{\infty}\sum_{l=0}^{\infty}\sum_{m=0}^{\infty}h_ih_jh_lh_mX_{t-i}X_{s-j}X_{u-l}X_{v-m}\right| \\
&\le E\left\{\sum_{i=0}^{\infty}\sum_{j=0}^{\infty}\sum_{l=0}^{\infty}\sum_{m=0}^{\infty}|h_ih_jh_lh_m|\cdot|X_{t-i}X_{s-j}X_{u-l}X_{v-m}|\right\}.
\end{aligned}$$

But

$$\begin{aligned}
\sum_{i=0}^{\infty}\sum_{j=0}^{\infty}\sum_{l=0}^{\infty}\sum_{m=0}^{\infty}|h_ih_jh_lh_m| &= \sum_{i=0}^{\infty}|h_i|\sum_{j=0}^{\infty}|h_j|\sum_{l=0}^{\infty}|h_l|\sum_{m=0}^{\infty}|h_m| \\
&< \infty
\end{aligned}$$

and

$$E|X_{t-i}X_{s-j}X_{u-l}X_{v-m}| < \mu_4$$

for any value of any of the indices. Hence,

$$\begin{aligned}
E|Y_tY_sY_uY_v| &< \sum_{i=0}^{\infty}\sum_{j=0}^{\infty}\sum_{l=0}^{\infty}\sum_{m=0}^{\infty}|h_ih_jh_lh_m|\cdot\mu_4 \\
&< \infty. \ ■
\end{aligned}$$

Chapter 7 Exercises

7.1. Let $\{X_T\}$ denote a sequence of random scalars with $\operatorname{plim} X_T = \xi$. Let $\{c_T\}$ denote a sequence of deterministic scalars with $\lim_{T\to\infty} c_T = c$. Let $g: \mathbb{R}^2 \to \mathbb{R}^1$ be continuous at (ξ, c). Show that $g(X_T, c_T) \xrightarrow{p} g(\xi, c)$.

7.2. Let $Y_t = 0.8Y_{t-1} + \varepsilon_t$ with $E(\varepsilon_t\varepsilon_\tau) = 1$ for $t = \tau$ and zero otherwise.

(a) Calculate $\lim_{T\to\infty} T\cdot\mathrm{Var}(\overline{Y}_T)$.

(b) How large a sample would we need in order to have 95% confidence that $\overline{Y}_T$ differed from the true value zero by no more than 0.1?

7.3. Does a martingale difference sequence have to be covariance-stationary?

7.4. Let $Y_t = \Sigma_{j=0}^{\infty}\psi_j\varepsilon_{t-j}$, where $\Sigma_{j=0}^{\infty}|\psi_j| < \infty$ and $\{\varepsilon_t\}$ is a martingale difference sequence with $E(\varepsilon_t^2) = \sigma^2$. Is Y_t covariance-stationary?

7.5. Define $X_{t,k} \equiv \Sigma_{u=0}^{\infty}\Sigma_{v=0}^{\infty}\psi_u\psi_v[\varepsilon_{t-u}\varepsilon_{t-k-v} - E(\varepsilon_{t-u}\varepsilon_{t-k-v})]$, where ε_t is an i.i.d. sequence with $E|\varepsilon_t|^r < M''$ for some $r > 2$ and $M'' < \infty$ with $\Sigma_{j=0}^{\infty}|\psi_j| < \infty$. Show that $X_{t,k}$ is uniformly integrable.

7.6. Derive result [7.2.15].

7.7. Let Y_t follow an $ARMA(p, q)$ process,

$$(1 - \phi_1 L - \phi_2 L^2 - \cdots - \phi_p L^p)(Y_t - \mu) = (1 + \theta_1 L + \theta_2 L^2 + \cdots + \theta_q L^q)\varepsilon_t,$$

with roots of $(1 - \phi_1 z - \phi_2 z^2 - \cdots - \phi_p z^p) = 0$ and $(1 + \theta_1 z + \theta_2 z^2 + \cdots + \theta_q z^q) = 0$ outside the unit circle. Suppose ε_t has mean zero and is independent of ε_τ for $t \neq \tau$ with $E(\varepsilon_t^2) = \sigma^2$ and $E(\varepsilon_t^4) < \infty$ for all t. Prove the following:

(a) $(1/T)\sum_{t=1}^{T} Y_t \xrightarrow{p} \mu$

(b) $[1/(T-k)]\sum_{t=k+1}^{T} Y_t Y_{t-k} \xrightarrow{p} E(Y_t Y_{t-k})$.

Chapter 7 References

Anderson, T. W. 1971. *The Statistical Analysis of Time Series*. New York: Wiley.

Andrews, Donald W. K. 1988. "Laws of Large Numbers for Dependent Non-Identically Distributed Random Variables." *Econometric Theory* 4:458–67.

Hoel, Paul G., Sidney C. Port, and Charles J. Stone. 1971. *Introduction to Probability Theory*. Boston: Houghton Mifflin.

Marsden, Jerrold E. 1974. *Elementary Classical Analysis*. San Francisco: Freeman.

Phillips, Peter C. B., and Victor Solo. 1992. "Asymptotics for Linear Processes." *Annals of Statistics* 20:971–1001.

Rao, C. Radhakrishna. 1973. *Linear Statistical Inference and Its Applications*, 2d ed. New York: Wiley.

Royden, H. L. 1968. *Real Analysis*, 2d ed. New York: Macmillan.

Theil, Henri. 1971. *Principles of Econometrics*. New York: Wiley.

White, Halbert. 1984. *Asymptotic Theory for Econometricians*. Orlando, Fla.: Academic Press.

8 Linear Regression Models

We have seen that one convenient way to estimate the parameters of an autoregression is with ordinary least squares regression, an estimation technique that is also useful for a number of other models. This chapter reviews the properties of linear regression. Section 8.1 analyzes the simplest case, in which the explanatory variables are nonrandom and the disturbances are i.i.d. Gaussian. Section 8.2 develops analogous results for ordinary least squares estimation of more general models such as autoregressions and regressions in which the disturbances are non-Gaussian, heteroskedastic, or autocorrelated. Linear regression models can also be estimated by generalized least squares, which is described in Section 8.3.

8.1. *Review of Ordinary Least Squares with Deterministic Regressors and i.i.d. Gaussian Disturbances*

Suppose that a scalar y_t is related to a $(k \times 1)$ vector $\mathbf{x}_t$ and a disturbance term u_t according to the regression model

$$y_t = \mathbf{x}_t'\boldsymbol{\beta} + u_t. \qquad [8.1.1]$$

This relation could be used to describe either the random variables or their realization. In discussing regression models, it proves cumbersome to distinguish notationally between random variables and their realization, and standard practice is to use small letters for either.

This section reviews estimation and hypothesis tests about $\boldsymbol{\beta}$ under the assumptions that $\mathbf{x}_t$ is deterministic and u_t is i.i.d. Gaussian. The next sections discuss regression under more general assumptions. First, however, we summarize the mechanics of linear regression and present some formulas that hold regardless of statistical assumptions.

The Algebra of Linear Regression

Given an observed sample $(y_1, y_2, \ldots, y_T)$, the *ordinary least squares* (*OLS*) estimate of $\boldsymbol{\beta}$ (denoted $\mathbf{b}$) is the value of $\boldsymbol{\beta}$ that minimizes the residual sum of squares (*RSS*):

$$RSS \equiv \sum_{t=1}^{T} (y_t - \mathbf{x}_t'\boldsymbol{\beta})^2. \qquad [8.1.2]$$

We saw in Appendix 4.A to Chapter 4 that the *OLS* estimate is given by

$$\mathbf{b} = \left[\sum_{t=1}^{T} (\mathbf{x}_t\mathbf{x}_t')\right]^{-1}\left[\sum_{t=1}^{T} (\mathbf{x}_t y_t)\right], \qquad [8.1.3]$$

assuming that the $(k \times k)$ matrix $\Sigma_{t=1}^{T}(\mathbf{x}_t\mathbf{x}_t')$ is nonsingular. The *OLS* sample residual for observation t is

$$\hat{u}_t \equiv y_t - \mathbf{x}_t'\mathbf{b}. \qquad [8.1.4]$$

Often the model in [8.1.1] is written in matrix notation as

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{u}, \qquad [8.1.5]$$

where

$$\underset{(T\times 1)}{\mathbf{y}} \equiv \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_T \end{bmatrix} \quad \underset{(T\times k)}{\mathbf{X}} \equiv \begin{bmatrix} \mathbf{x}_1' \\ \mathbf{x}_2' \\ \vdots \\ \mathbf{x}_T' \end{bmatrix} \quad \underset{(T\times 1)}{\mathbf{u}} \equiv \begin{bmatrix} u_1 \\ u_2 \\ \vdots \\ u_T \end{bmatrix}.$$

Then the *OLS* estimate in [8.1.3] can be written as

$$\begin{aligned} \mathbf{b} &= \left\{[\mathbf{x}_1 \quad \mathbf{x}_2 \cdots \mathbf{x}_T]\begin{bmatrix} \mathbf{x}_1' \\ \mathbf{x}_2' \\ \vdots \\ \mathbf{x}_T' \end{bmatrix}\right\}^{-1}\left\{[\mathbf{x}_1 \quad \mathbf{x}_2 \cdots \mathbf{x}_T]\begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_T \end{bmatrix}\right\} \qquad [8.1.6] \\ &= (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}. \end{aligned}$$

Similarly, the vector of *OLS* sample residuals [8.1.4] can be written as

$$\hat{\mathbf{u}} = \mathbf{y} - \mathbf{X}\mathbf{b} = \mathbf{y} - \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y} = [\mathbf{I}_T - \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}']\mathbf{y} = \mathbf{M}_\mathbf{X}\mathbf{y}, \qquad [8.1.7]$$

where $\mathbf{M}_\mathbf{X}$ is defined as the following $(T \times T)$ matrix:

$$\mathbf{M}_\mathbf{X} \equiv \mathbf{I}_T - \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'. \qquad [8.1.8]$$

One can readily verify that $\mathbf{M}_\mathbf{X}$ is symmetric:

$$\mathbf{M}_\mathbf{X} = \mathbf{M}_\mathbf{X}';$$

idempotent:

$$\mathbf{M}_\mathbf{X}\mathbf{M}_\mathbf{X} = \mathbf{M}_\mathbf{X};$$

and orthogonal to the columns of $\mathbf{X}$:

$$\mathbf{M}_\mathbf{X}\mathbf{X} = \mathbf{0}. \qquad [8.1.9]$$

Thus, from [8.1.7], the *OLS* sample residuals are orthogonal to the explanatory variables in $\mathbf{X}$:

$$\hat{\mathbf{u}}'\mathbf{X} = \mathbf{y}'\mathbf{M}_\mathbf{X}'\mathbf{X} = \mathbf{0}'. \qquad [8.1.10]$$

The *OLS* sample residual $(\hat{u}_t)$ should be distinguished from the population residual (u_t). The sample residual is constructed from the sample estimate $\mathbf{b}$ $(\hat{u}_t = y_t - \mathbf{x}_t'\mathbf{b})$, whereas the population residual is a hypothetical construct based on the true population value $\boldsymbol{\beta}$ $(u_t = y_t - \mathbf{x}_t'\boldsymbol{\beta})$. The relation between the sample

and population residuals can be found by substituting [8.1.5] into [8.1.7]:

$$\hat{\mathbf{u}} = \mathbf{M_X}(\mathbf{X}\boldsymbol{\beta} + \mathbf{u}) = \mathbf{M_X}\mathbf{u}. \qquad [8.1.11]$$

The difference between the *OLS* estimate $\mathbf{b}$ and the true population parameter $\boldsymbol{\beta}$ is found by substituting [8.1.5] into [8.1.6]:

$$\mathbf{b} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'[\mathbf{X}\boldsymbol{\beta} + \mathbf{u}] = \boldsymbol{\beta} + (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{u}. \qquad [8.1.12]$$

The fit of an *OLS* regression is sometimes described in terms of the sample multiple correlation coefficient, or R^2. The *uncentered* R^2 (denoted R_u^2) is defined as the sum of squares of the fitted values ($\mathbf{x}_t'\mathbf{b}$) of the regression as a fraction of the sum of squares of y:

$$R_u^2 \equiv \frac{\sum_{t=1}^{T} (\mathbf{b}'\mathbf{x}_t\mathbf{x}_t'\mathbf{b})}{\sum_{t=1}^{T} y_t^2} = \frac{\mathbf{b}'\mathbf{X}'\mathbf{X}\mathbf{b}}{\mathbf{y}'\mathbf{y}} = \frac{\mathbf{y}'\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}}{\mathbf{y}'\mathbf{y}}. \qquad [8.1.13]$$

If the only explanatory variable in the regression were a constant term ($\mathbf{x}_t = 1$), then the fitted value for each observation would just be the sample mean $\bar{y}$ and the sum of squares of the fitted values would be $T\bar{y}^2$. This sum of squares is often compared with the sum of squares when a vector of variables $\mathbf{x}_t$ is included in the regression. The *centered* R^2 (denoted R_c^2) is defined as

$$R_c^2 \equiv \frac{\mathbf{y}'\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y} - T\bar{y}^2}{\mathbf{y}'\mathbf{y} - T\bar{y}^2}. \qquad [8.1.14]$$

Most regression software packages report the centered R^2 rather than the uncentered R^2. If the regression includes a constant term, then R_c^2 must be between zero and unity. However, if the regression does not include a constant term, then R_c^2 can be negative.

The Classical Regression Assumptions

Statistical inference requires assumptions about the properties of the explanatory variables $\mathbf{x}_t$ and the population residuals u_t. The simplest case to analyze is the following.

Assumption 8.1: *(a) $\mathbf{x}_t$ is a vector of deterministic variables (for example, $\mathbf{x}_t$ might include a constant term and deterministic functions of t); (b) u_t is i.i.d. with mean 0 and variance σ^2; (c) u_t is Gaussian.*

To highlight the role of each of these assumptions, we first note the implications of Assumption 8.1(a) and (b) alone and then comment on the added implications that follow from (c).

Properties of the Estimated OLS *Coefficient Vector Under Assumption 8.1(a) and (b)*

In vector form, Assumption 8.1(b) could be written $E(\mathbf{u}) = \mathbf{0}$ and $E(\mathbf{u}\mathbf{u}') = \sigma^2\mathbf{I}_T$.

Taking expectations of [8.1.12] and using these conditions establishes that $\mathbf{b}$ is unbiased,

$$E(\mathbf{b}) = \boldsymbol{\beta} + (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'[E(\mathbf{u})] = \boldsymbol{\beta}, \qquad [8.1.15]$$

with variance-covariance matrix given by

$$\begin{aligned} E[(\mathbf{b} - \boldsymbol{\beta})(\mathbf{b} - \boldsymbol{\beta})'] &= E[(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{u}\mathbf{u}'\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}] \\ &= (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'[E(\mathbf{u}\mathbf{u}')]\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1} \\ &= \sigma^2(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1} \\ &= \sigma^2(\mathbf{X}'\mathbf{X})^{-1}. \end{aligned} \qquad [8.1.16]$$

The *OLS* coefficient estimate $\mathbf{b}$ is unbiased and is a linear function of $\mathbf{y}$. The *Gauss-Markov theorem* states that the variance-covariance matrix of any alternative estimator of $\boldsymbol{\beta}$, if that estimator is also unbiased and a linear function of $\mathbf{y}$, differs from the variance-covariance matrix of $\mathbf{b}$ by a positive semidefinite matrix.[1] This means that an inference based on $\mathbf{b}$ about any linear combination of the elements of $\boldsymbol{\beta}$ will have a smaller variance than the corresponding inference based on any alternative linear unbiased estimator. The Gauss-Markov theorem thus establishes the optimality of the *OLS* estimate within a certain limited class.

Properties of the Estimated Coefficient Vector Under Assumption 8.1(a) Through (c)

When $\mathbf{u}$ is Gaussian, [8.1.12] implies that $\mathbf{b}$ is Gaussian. Hence, the preceding results imply

$$\mathbf{b} \sim N(\boldsymbol{\beta}, \sigma^2(\mathbf{X}'\mathbf{X})^{-1}). \qquad [8.1.17]$$

It can further be shown that under Assumption 8.1(a) through (c), no unbiased estimator of $\boldsymbol{\beta}$ is more efficient than the *OLS* estimator $\mathbf{b}$.[2] Thus, with Gaussian residuals, the *OLS* estimator is optimal.

Properties of Estimated Residual Variance Under Assumption 8.1(a) and (b)

The *OLS* estimate of the variance of the disturbances σ^2 is

$$s^2 = RSS/(T - k) = \hat{\mathbf{u}}'\hat{\mathbf{u}}/(T - k) = \mathbf{u}'\mathbf{M}_{\mathbf{X}}'\mathbf{M}_{\mathbf{X}}\mathbf{u}/(T - k) \qquad [8.1.18]$$

for $\mathbf{M}_{\mathbf{X}}$ the matrix in [8.1.8]. Recalling that $\mathbf{M}_{\mathbf{X}}$ is symmetric and idempotent, [8.1.18] becomes

$$s^2 = \mathbf{u}'\mathbf{M}_{\mathbf{X}}\mathbf{u}/(T - k). \qquad [8.1.19]$$

Also, since $\mathbf{M}_{\mathbf{X}}$ is symmetric, there exists a $(T \times T)$ matrix $\mathbf{P}$ such that[3]

$$\mathbf{M}_{\mathbf{X}} = \mathbf{P}\boldsymbol{\Lambda}\mathbf{P}' \qquad [8.1.20]$$

and

$$\mathbf{P}'\mathbf{P} = \mathbf{I}_T, \qquad [8.1.21]$$

where $\boldsymbol{\Lambda}$ is a $(T \times T)$ matrix with the eigenvalues of $\mathbf{M}_{\mathbf{X}}$ along the principal diagonal and zeros elsewhere. Note from [8.1.9] that $\mathbf{M}_{\mathbf{X}}\mathbf{v} = \mathbf{0}$ if $\mathbf{v}$ should be given by one of the k columns of $\mathbf{X}$. Assuming that the columns of $\mathbf{X}$ are linearly independent, the k columns of $\mathbf{X}$ thus represent k different eigenvectors of $\mathbf{M}_{\mathbf{X}}$ each associated

[1]See, for example, Theil (1971, pp. 119–20).

[2]See, for example, Theil (1971, pp. 390–91).

[3]See, for example, O'Nan (1976, p. 296).

with an eigenvalue equal to zero. Also from [8.1.8], $\mathbf{M_X v} = \mathbf{v}$ for any vector $\mathbf{v}$ that is orthogonal to the columns of $\mathbf{X}$ (that is, any vector $\mathbf{v}$ such that $\mathbf{X'v} = \mathbf{0}$); $(T - k)$ such vectors that are linearly independent can be found, associated with $(T - k)$ eigenvalues equal to unity. Thus, $\mathbf{\Lambda}$ contains k zeros and $(T - k)$ 1s along its principal diagonal. Notice from [8.1.20] that

$$\begin{aligned} \mathbf{u'M_X u} &= \mathbf{u'P\Lambda P'u} \\ &= (\mathbf{P'u})'\mathbf{\Lambda}(\mathbf{P'u}) \\ &= \mathbf{w'\Lambda w} \\ &= w_1^2\lambda_1 + w_2^2\lambda_2 + \cdots + w_T^2\lambda_T, \end{aligned} \qquad [8.1.22]$$

where

$$\mathbf{w} \equiv \mathbf{P'u}.$$

Furthermore,

$$E(\mathbf{ww'}) = E(\mathbf{P'uu'P}) = \mathbf{P'}E(\mathbf{uu'})\mathbf{P} = \sigma^2\mathbf{P'P} = \sigma^2\mathbf{I}_T.$$

Thus, the elements of $\mathbf{w}$ are uncorrelated, with mean zero and variance σ^2. Since k of the λ's are zero and the remaining $T - k$ are unity, [8.1.22] becomes

$$\mathbf{u'M_X u} = w_1^2 + w_2^2 + \cdots + w_{T-k}^2. \qquad [8.1.23]$$

Furthermore, each w_i^2 has expectation σ^2, so that

$$E(\mathbf{u'M_X u}) = (T - k)\sigma^2,$$

and from [8.1.19], s^2 gives an unbiased estimate of σ^2:

$$E(s^2) = \sigma^2.$$

Properties of Estimated Residual Variance Under Assumption 8.1(a) Through (c)

When u_t is Gaussian, w_t is also Gaussian and expression [8.1.23] is the sum of squares of $(T - k)$ independent $N(0, \sigma^2)$ variables. Thus,

$$RSS/\sigma^2 = \mathbf{u'M_X u}/\sigma^2 \sim \chi^2(T - k). \qquad [8.1.24]$$

Again, it is possible to show that under Assumption 8.1(a) through (c), no other unbiased estimator of σ^2 has a smaller variance than does s^2.[4]

Notice also from [8.1.11] and [8.1.12] that $\mathbf{b}$ and $\hat{\mathbf{u}}$ are uncorrelated:

$$E[\hat{\mathbf{u}}(\mathbf{b} - \boldsymbol{\beta})'] = E[\mathbf{M_X uu'X}(\mathbf{X'X})^{-1}] = \sigma^2\mathbf{M_X X}(\mathbf{X'X})^{-1} = \mathbf{0}. \qquad [8.1.25]$$

Under Assumption 8.1(a) through (c), both $\mathbf{b}$ and $\hat{\mathbf{u}}$ are Gaussian, so that absence of correlation implies that $\mathbf{b}$ and $\hat{\mathbf{u}}$ are independent. This means that $\mathbf{b}$ and s^2 are independent.

t *Tests About* $\boldsymbol{\beta}$ *Under Assumption 8.1(a) Through (c)*

Suppose that we wish to test the null hypothesis that β_i, the ith element of $\boldsymbol{\beta}$, is equal to some particular value β_i^0. The *OLS* t statistic for testing this null hypothesis is given by

$$t = \frac{(b_i - \beta_i^0)}{\hat{\sigma}_{b_i}} = \frac{(b_i - \beta_i^0)}{s(\xi^{ii})^{1/2}}, \qquad [8.1.26]$$

[4]See Rao (1973, p. 319).

where ξ^{ii} denotes the row i, column i element of $(\mathbf{X}'\mathbf{X})^{-1}$ and $\hat{\sigma}_{b_i} \equiv \sqrt{s^2\xi^{ii}}$ is the standard error of the *OLS* estimate of the ith coefficient. The magnitude in [8.1.26] has an exact t distribution with $T - k$ degrees of freedom so long as $\mathbf{x}_t$ is deterministic and u_t is i.i.d. Gaussian. To verify this claim, note from [8.1.17] that under the null hypothesis, $b_i \sim N(\beta_i^0, \sigma^2\xi^{ii})$, meaning that $(b_i - \beta_i^0)/\sqrt{\sigma^2\xi^{ii}} \sim N(0, 1)$. Thus, if [8.1.26] is written as

$$t = \frac{(b_i - \beta_i^0)/\sqrt{\sigma^2\xi^{ii}}}{\sqrt{s^2/\sigma^2}},$$

the numerator is $N(0, 1)$ while from [8.1.24] the denominator is the square root of a $\chi^2(T - k)$ variable divided by its degrees of freedom. Recalling [8.1.25], the numerator and denominator are independent, confirming the exact t distribution claimed for [8.1.26].

F *Tests About* $\boldsymbol{\beta}$ *Under Assumption 8.1(a) Through (c)*

More generally, suppose we want a joint test of m different linear restrictions about $\boldsymbol{\beta}$, as represented by

$$H_0\colon \mathbf{R}\boldsymbol{\beta} = \mathbf{r}. \qquad [8.1.27]$$

Here $\mathbf{R}$ is a known $(m \times k)$ matrix representing the particular linear combinations of $\boldsymbol{\beta}$ about which we entertain hypotheses and $\mathbf{r}$ is a known $(m \times 1)$ vector of the values that we believe these linear combinations take on. For example, to represent the simple hypothesis $\beta_i = \beta_i^0$ used previously, we would have $m = 1$, $\mathbf{R}$ a $(1 \times k)$ vector with unity in the ith position and zeros elsewhere, and $\mathbf{r}$ the scalar β_i^0. As a second example, consider a regression with $k = 4$ explanatory variables and the joint hypothesis that $\beta_1 + \beta_2 = 1$ and $\beta_3 = \beta_4$. In this case, $m = 2$ and

$$\mathbf{R} = \begin{bmatrix} 1 & 1 & 0 & 0 \\ 0 & 0 & 1 & -1 \end{bmatrix} \qquad \mathbf{r} = \begin{bmatrix} 1 \\ 0 \end{bmatrix}. \qquad [8.1.28]$$

Notice from [8.1.17] that under H_0,

$$\mathbf{Rb} \sim N(\mathbf{r}, \sigma^2\mathbf{R}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{R}'). \qquad [8.1.29]$$

A *Wald test* of H_0 is based on the following result.

Proposition 8.1: *Consider an* $(n \times 1)$ *vector* $\mathbf{z} \sim N(\mathbf{0}, \boldsymbol{\Omega})$ *with* $\boldsymbol{\Omega}$ *nonsingular. Then* $\mathbf{z}'\boldsymbol{\Omega}^{-1}\mathbf{z} \sim \chi^2(n)$.

For the scalar case ($n = 1$), observe that if $z \sim N(0, \sigma^2)$, then $(z/\sigma) \sim N(0, 1)$ and $z^2/\sigma^2 \sim \chi^2(1)$, as asserted by the proposition.

To verify Proposition 8.1 for the vector case, since $\boldsymbol{\Omega}$ is symmetric, there exists a matrix $\mathbf{P}$, as in [8.1.20] and [8.1.21], such that $\boldsymbol{\Omega} = \mathbf{P}\boldsymbol{\Lambda}\mathbf{P}'$ and $\mathbf{P}'\mathbf{P} = \mathbf{I}_n$ with $\boldsymbol{\Lambda}$ containing the eigenvalues of $\boldsymbol{\Omega}$. Since $\boldsymbol{\Omega}$ is positive definite, the diagonal elements of $\boldsymbol{\Lambda}$ are positive. Then

$$\begin{aligned} \mathbf{z}'\boldsymbol{\Omega}^{-1}\mathbf{z} &= \mathbf{z}'(\mathbf{P}\boldsymbol{\Lambda}\mathbf{P}')^{-1}\mathbf{z} \\ &= \mathbf{z}'[\mathbf{P}']^{-1}\boldsymbol{\Lambda}^{-1}\mathbf{P}^{-1}\mathbf{z} \\ &= [\mathbf{P}^{-1}\mathbf{z}]'\boldsymbol{\Lambda}^{-1}\mathbf{P}^{-1}\mathbf{z} \\ &= \mathbf{w}'\boldsymbol{\Lambda}^{-1}\mathbf{w} \qquad [8.1.30] \\ &= \sum_{i=1}^{n} w_i^2/\lambda_i, \end{aligned}$$

where $\mathbf{w} \equiv \mathbf{P}^{-1}\mathbf{z}$. Notice that $\mathbf{w}$ is Gaussian with mean zero and variance

$$E(\mathbf{ww}') = E(\mathbf{P}^{-1}\mathbf{zz}'[\mathbf{P}']^{-1}) = \mathbf{P}^{-1}\mathbf{\Omega}[\mathbf{P}']^{-1} = \mathbf{P}^{-1}\mathbf{P}\mathbf{\Lambda}\mathbf{P}'[\mathbf{P}']^{-1} = \mathbf{\Lambda}.$$

Thus [8.1.30] is the sum of squares of n independent Normal variables, each divided by its variance λ_i. It accordingly has a $\chi^2(n)$ distribution, as claimed.

Applying Proposition 8.1 directly to [8.1.29], under H_0;

$$(\mathbf{Rb} - \mathbf{r})'[\sigma^2\mathbf{R}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{R}']^{-1}(\mathbf{Rb} - \mathbf{r}) \sim \chi^2(m). \qquad [8.1.31]$$

Replacing σ^2 with the estimate s^2 and dividing by the number of restrictions gives the Wald form of the *OLS* F test of a linear hypothesis:

$$F = (\mathbf{Rb} - \mathbf{r})'[s^2\mathbf{R}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{R}']^{-1}(\mathbf{Rb} - \mathbf{r})/m. \qquad [8.1.32]$$

Note that [8.1.32] can be written

$$F = \frac{(\mathbf{Rb} - \mathbf{r})'[\sigma^2\mathbf{R}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{R}']^{-1}(\mathbf{Rb} - \mathbf{r})/m}{[RSS/(T - k)]/\sigma^2}.$$

The numerator is a $\chi^2(m)$ variable divided by its degrees of freedom, while the denominator is a $\chi^2(T - k)$ variable divided by its degrees of freedom. Again, since $\mathbf{b}$ and $\hat{\mathbf{u}}$ are independent, the numerator and denominator are independent of each other. Hence, [8.1.32] has an exact $F(m, T - k)$ distribution under H_0 when $\mathbf{x}_t$ is nonstochastic and u_t is i.i.d. Gaussian.

Notice that the t test of the simple hypothesis $\beta_i = \beta_i^0$ is a special case of the general formula [8.1.32], for which

$$F = (b_i - \beta_i^0)[s^2\xi^{ii}]^{-1}(b_i - \beta_i^0). \qquad [8.1.33]$$

This is the square of the t statistic in [8.1.26]. Since an $F(1, T - k)$ variable is just the square of a $t(T - k)$ variable, the identical answer results from (1) calculating [8.1.26] and using t tables to find the probability of so large an absolute value for a $t(T - k)$ variable, or (2) calculating [8.1.33] and using F tables to find the probability of so large a value for an $F(1, T - k)$ variable.

A Convenient Alternative Expression for the F *Test*

It is often straightforward to estimate the model in [8.1.1] subject to the restrictions in [8.1.27]. For example, to impose a constraint $\beta_1 = \beta_1^0$ on the first element of $\boldsymbol{\beta}$, we could just do an ordinary least squares regression of $y_t - \beta_1^0 x_{1t}$ on $x_{2t}, x_{3t}, \ldots, x_{kt}$. The resulting estimates $b_2^*, b_3^*, \ldots, b_k^*$ minimize $\Sigma_{t=1}^T [(y_t - \beta_1^0 x_{1t}) - b_2^* x_{2t} - b_3^* x_{3t} - \cdots - b_k^* x_{kt}]^2$ with respect to $b_2^*, b_3^*, \ldots, b_k^*$ and thus minimize the residual sum of squares [8.1.2] subject to the constraint that $\beta_1 = \beta_1^0$. Alternatively, to impose the constraint in [8.1.28], we could regress $y_t - x_{2t}$ on $(x_{1t} - x_{2t})$ and $(x_{3t} + x_{4t})$:

$$y_t - x_{2t} = \beta_1(x_{1t} - x_{2t}) + \beta_3(x_{3t} + x_{4t}) + u_t.$$

The *OLS* estimates b_1^* and b_3^* minimize

$$\begin{aligned}\sum_{t=1}^{T} [(y_t - x_{2t}) - b_1^*(x_{1t} - x_{2t}) - b_3^*(x_{3t} + x_{4t})]^2 \\ = \sum_{t=1}^{T} [y_t - b_1^* x_{1t} - (1 - b_1^*)x_{2t} - b_3^* x_{3t} - b_3^* x_{4t}]^2\end{aligned} \qquad [8.1.34]$$

and thus minimize [8.1.2] subject to [8.1.28].

Whenever the constraints in [8.1.27] can be imposed through a simple *OLS* regression on transformed variables, there is an easy way to calculate the F statistic

[8.1.32] just by comparing the residual sum of squares for the constrained and unconstrained regressions. The following result is established in Appendix 8.A at the end of this chapter.

Proposition 8.2: *Let* **b** *denote the unconstrained OLS estimate [8.1.6] and let* RSS_1 *be the residual sum of squares resulting from using this estimate:*

$$RSS_1 = \sum_{t=1}^{T} (y_t - \mathbf{x}_t'\mathbf{b})^2. \qquad [8.1.35]$$

Let $\mathbf{b}^*$ *denote the constrained OLS estimate and* RSS_0 *the residual sum of squares from the constrained OLS estimation:*

$$RSS_0 = \sum_{t=1}^{T} (y_t - \mathbf{x}_t'\mathbf{b}^*)^2. \qquad [8.1.36]$$

Then the Wald form of the OLS F test of a linear hypothesis [8.1.32] can equivalently be calculated as

$$F = \frac{(RSS_0 - RSS_1)/m}{RSS_1/(T - k)}. \qquad [8.1.37]$$

Expressions [8.1.37] and [8.1.32] will generate exactly the same number, regardless of whether the null hypothesis and the model are valid or not.

For example, suppose the sample size is $T = 50$ observations and the null hypothesis is $\beta_3 = \beta_4 = 0$ in an *OLS* regression with $k = 4$ explanatory variables. First regress y_t on x_{1t}, x_{2t}, x_{3t}, x_{4t} and call the residual sum of squares from this regression RSS_1. Next, regress y_t on just x_{1t} and x_{2t} and call the residual sum of squares from this restricted regression RSS_0. If

$$\frac{(RSS_0 - RSS_1)/2}{RSS_1/(50 - 4)}$$

is greater than 3.20 (the 5% critical value for an $F(2, 46)$ random variable), then the null hypothesis should be rejected.

8.2. *Ordinary Least Squares Under More General Conditions*

The previous section analyzed the regression model

$$y_t = \mathbf{x}_t'\boldsymbol{\beta} + u_t$$

under the maintained Assumption 8.1 ($\mathbf{x}_t$ is deterministic and u_t is i.i.d. Gaussian). We will hereafter refer to this assumption as "case 1." This section generalizes this assumption to describe specifications likely to arise in time series analysis. Some of the key results are summarized in Table 8.1.

Case 2. Error Term i.i.d. Gaussian and Independent of Explanatory Variables

Consider the case in which **X** is stochastic but completely independent of **u**.

Assumption 8.2:[5] *(a)* $\mathbf{x}_t$ *stochastic and independent of* u_s *for all t, s; (b)* $u_t \sim$ *i.i.d.* $N(0, \sigma^2)$.

[5]This could be replaced with the assumption $\mathbf{u}|\mathbf{X} \sim N(\mathbf{0}, \sigma^2\mathbf{I}_T)$ with all the results to follow unchanged.

Many of the results for deterministic regressors continue to apply for this case. For example, taking expectations of [8.1.12] and exploiting the independence assumption,

$$E(\mathbf{b}) = \boldsymbol{\beta} + \{E[(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}']\}\{E(\mathbf{u})\} = \boldsymbol{\beta}, \qquad [8.2.1]$$

so that the *OLS* coefficient remains unbiased.

The distribution of test statistics for this case can be found by a two-step procedure. The first step evaluates the distribution conditional on $\mathbf{X}$; that is, it treats $\mathbf{X}$ as deterministic just as in the earlier analysis. The second step multiplies by the density of $\mathbf{X}$ and integrates over $\mathbf{X}$ to find the true unconditional distribution. For example, [8.1.17] implies that

$$\mathbf{b}|\mathbf{X} \sim N(\boldsymbol{\beta}, \sigma^2(\mathbf{X}'\mathbf{X})^{-1}). \qquad [8.2.2]$$

If this density is multiplied by the density of $\mathbf{X}$ and integrated over $\mathbf{X}$, the result is no longer a Gaussian distribution; thus, $\mathbf{b}$ is non-Gaussian under Assumption 8.2. On the other hand, [8.1.24] implies that

$$RSS|\mathbf{X} \sim \sigma^2 \cdot \chi^2(T - k).$$

But this density is the same for all $\mathbf{X}$. Thus, when we multiply the density of $RSS|\mathbf{X}$ by the density of $\mathbf{X}$ and integrate, we will get exactly the same density. Hence, [8.1.24] continues to give the correct unconditional distribution for Assumption 8.2.

The same is true for the t and F statistics in [8.1.26] and [8.1.32]. Conditional on $\mathbf{X}$, $(b_i - \beta_i^0)/[\sigma(\xi^{ii})^{1/2}] \sim N(0, 1)$ and s/σ is the square root of an independent $[1/(T - k)]\cdot\chi^2(T - k)$ variable. Hence, conditional on $\mathbf{X}$, the statistic in [8.1.26] has a $t(T - k)$ distribution. Since this is true for any $\mathbf{X}$, when we multiply by the density of $\mathbf{X}$ and integrate over $\mathbf{X}$ we obtain the same distribution.

Case 3. Error Term i.i.d. Non-Gaussian and Independent of Explanatory Variables

Next consider the following specification.

Assumption 8.3: *(a) $\mathbf{x}_t$ stochastic and independent of u_s for all t, s; (b) u_t non-Gaussian but i.i.d. with mean zero, variance σ^2, and $E(u_t^4) = \mu_4 < \infty$; (c) $E(\mathbf{x}_t\mathbf{x}_t') = \mathbf{Q}_t$, a positive definite matrix with $(1/T)\Sigma_{t=1}^T \mathbf{Q}_t \to \mathbf{Q}$, a positive definite matrix; (d) $E(x_{it}x_{jt}x_{lt}x_{mt}) < \infty$ for all i, j, l, m, and t; (e) $(1/T)\Sigma_{t=1}^T(\mathbf{x}_t\mathbf{x}_t') \xrightarrow{p} \mathbf{Q}$.*

Since result [8.2.1] required only the independence assumption, $\mathbf{b}$ continues to be unbiased in this case. However, for hypothesis tests, the small-sample distributions of s^2 and the t and F statistics are no longer the same as when the population residuals are Gaussian. To justify the usual *OLS* inference rules, we have to appeal to asymptotic results, for which purpose Assumption 8.3 includes conditions (c) through (e). To understand these conditions, note that if $\mathbf{x}_t$ is covariance-stationary, then $E(\mathbf{x}_t\mathbf{x}_t')$ does not depend on t. Then $\mathbf{Q}_t = \mathbf{Q}$ for all t and condition (e) simply requires that $\mathbf{x}_t$ be ergodic for second moments. Assumption 8.3 also allows more general processes in that $E(\mathbf{x}_t\mathbf{x}_t')$ might be different for different t, so long as the limit of $(1/T)\Sigma_{t=1}^T E(\mathbf{x}_t\mathbf{x}_t')$ can be consistently estimated by $(1/T)\Sigma_{t=1}^T(\mathbf{x}_t\mathbf{x}_t')$.

TABLE 8.1
Properties of *OLS* Estimates and Test Statistics Under Various Assumptions

	Coefficient **b**	*Variance* s^2	*t statistic*	*F statistic*
Case 1	unbiased $\mathbf{b} \sim N(\boldsymbol{\beta}, \sigma^2(\mathbf{X}'\mathbf{X})^{-1})$	unbiased $(T-k)s^2/\sigma^2 \sim \chi^2(T-k)$	exact $t(T-k)$	exact $F(m, T-k)$
Case 2	unbiased non-Gaussian	unbiased $(T-k)s^2/\sigma^2 \sim \chi^2(T-k)$	exact $t(T-k)$	exact $F(m, T-k)$
Case 3	unbiased $\sqrt{T}(\mathbf{b}_T - \boldsymbol{\beta}) \xrightarrow{L} N(\mathbf{0}, \sigma^2\mathbf{Q}^{-1})$	unbiased $\sqrt{T}(s_T^2 - \sigma^2) \xrightarrow{L} N(0, \mu_4 - \sigma^4)$	$t_T \xrightarrow{L} N(0, 1)$	$mF_T \xrightarrow{L} \chi^2(m)$
Case 4	biased $\sqrt{T}(\mathbf{b}_T - \boldsymbol{\beta}) \xrightarrow{L} N(\mathbf{0}, \sigma^2\mathbf{Q}^{-1})$	biased $\sqrt{T}(s_T^2 - \sigma^2) \xrightarrow{L} N(0, \mu_4 - \sigma^4)$	$t_T \xrightarrow{L} N(0, 1)$	$mF_T \xrightarrow{L} \chi^2(m)$

Regression model is $\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{u}$, **b** is given by [8.1.6], s^2 by [8.1.18], t statistic by [8.1.26], and F statistic by [8.1.32]; μ_4 denotes $E(u_t^4)$.
Case 1: **X** nonstochastic, $\mathbf{u} \sim N(\mathbf{0}, \sigma^2\mathbf{I}_T)$.
Case 2: **X** stochastic, $\mathbf{u} \sim N(0, \sigma^2\mathbf{I}_T)$, **X** independent of **u**.
Case 3: **X** stochastic, $\mathbf{u} \sim$ non-Gaussian $(\mathbf{0}, \sigma^2\mathbf{I}_T)$, **X** independent of **u**, $T^{-1}\Sigma\mathbf{x}_t\mathbf{x}_t' \xrightarrow{P} \mathbf{Q}$.
Case 4: Stationary autoregression with independent errors, **Q** given by [8.2.27].

To describe the asymptotic results, we denote the *OLS* estimator [8.1.3] by $\mathbf{b}_T$ to emphasize that it is based on a sample of size T. Our interest is in the behavior of $\mathbf{b}_T$ as T becomes large. We first establish that the *OLS* coefficient estimator is consistent under Assumption 8.3, that is, that $\mathbf{b}_T \xrightarrow{p} \boldsymbol{\beta}$.

Note that [8.1.12] implies

$$\begin{aligned}\mathbf{b}_T - \boldsymbol{\beta} &= \left[\sum_{t=1}^{T} \mathbf{x}_t\mathbf{x}_t'\right]^{-1}\left[\sum_{t=1}^{T} \mathbf{x}_t u_t\right] \\ &= \left[(1/T)\sum_{t=1}^{T} \mathbf{x}_t\mathbf{x}_t'\right]^{-1}\left[(1/T)\sum_{t=1}^{T} \mathbf{x}_t u_t\right].\end{aligned} \tag{8.2.3}$$

Consider the first term in [8.2.3]. Assumption 8.3(e) and Proposition 7.1 imply that

$$\left[(1/T)\sum_{t=1}^{T} \mathbf{x}_t\mathbf{x}_t'\right]^{-1} \xrightarrow{p} \mathbf{Q}^{-1}. \tag{8.2.4}$$

Considering next the second term in [8.2.3], notice that $\mathbf{x}_t u_t$ is a martingale difference sequence with variance-covariance matrix given by

$$E(\mathbf{x}_t u_t \mathbf{x}_t' u_t) = \{E(\mathbf{x}_t\mathbf{x}_t')\}\cdot\sigma^2,$$

which is finite. Thus, from Example 7.11,

$$\left[(1/T)\sum_{t=1}^{T} \mathbf{x}_t u_t\right] \xrightarrow{p} \mathbf{0}. \tag{8.2.5}$$

Applying Example 7.2 to [8.2.3] through [8.2.5],

$$\mathbf{b}_T - \boldsymbol{\beta} \xrightarrow{p} \mathbf{Q}^{-1}\cdot\mathbf{0} = \mathbf{0},$$

verifying that the *OLS* estimator is consistent.

Next turn to the asymptotic distribution of $\mathbf{b}$. Notice from [8.2.3] that

$$\sqrt{T}(\mathbf{b}_T - \boldsymbol{\beta}) = \left[(1/T)\sum_{t=1}^{T} \mathbf{x}_t\mathbf{x}_t'\right]^{-1}\left[(1/\sqrt{T})\sum_{t=1}^{T} \mathbf{x}_t u_t\right]. \tag{8.2.6}$$

We saw in [8.2.4] that the first term converges in probability to $\mathbf{Q}^{-1}$. The second term is $\sqrt{T}$ times the sample mean of $\mathbf{x}_t u_t$, where $\mathbf{x}_t u_t$ is a martingale difference sequence with variance $\sigma^2\cdot E(\mathbf{x}_t\mathbf{x}_t') = \sigma^2\mathbf{Q}_t$ and $(1/T)\Sigma_{t=1}^{\infty}\sigma^2\mathbf{Q}_t \to \sigma^2\mathbf{Q}$. Notice that under Assumption 8.3 we can apply Proposition 7.9:

$$\left[(1/\sqrt{T})\sum_{t=1}^{T} \mathbf{x}_t u_t\right] \xrightarrow{L} N(\mathbf{0}, \sigma^2\mathbf{Q}). \tag{8.2.7}$$

Combining [8.2.6], [8.2.4], and [8.2.7], we see as in Example 7.5 that

$$\sqrt{T}(\mathbf{b}_T - \boldsymbol{\beta}) \xrightarrow{L} N(\mathbf{0}, [\mathbf{Q}^{-1}\cdot(\sigma^2\mathbf{Q})\cdot\mathbf{Q}^{-1}]) = N(\mathbf{0}, \sigma^2\mathbf{Q}^{-1}). \tag{8.2.8}$$

In other words, we can act as if

$$\mathbf{b}_T \approx N(\boldsymbol{\beta}, \sigma^2\mathbf{Q}^{-1}/T), \tag{8.2.9}$$

where the symbol $\approx$ means "is approximately distributed." Recalling Assumption 8.3(e), in large samples $\mathbf{Q}$ should be close to $(1/T)\Sigma_{t=1}^{T}\mathbf{x}_t\mathbf{x}_t'$. Thus $\mathbf{Q}^{-1}/T$ should be close to $[\Sigma_{t=1}^{T}\mathbf{x}_t\mathbf{x}_t']^{-1} = (\mathbf{X}_T'\mathbf{X}_T)^{-1}$ for $\mathbf{X}_T$ the same $(T \times k)$ matrix that was represented in [8.1.5] simply by $\mathbf{X}$ (again, the subscript T is added at this point to

emphasize that the dimensions of this matrix depend on T). Thus, [8.2.9] can be approximated by

$$\mathbf{b}_T \approx N(\boldsymbol{\beta}, \sigma^2(\mathbf{X}_T'\mathbf{X}_T)^{-1}).$$

This, of course, is the same result obtained in [8.1.17], which assumed Gaussian disturbances. With non-Gaussian disturbances the distribution is not exact, but provides an increasingly good approximation as the sample size grows.

Next, consider consistency of the variance estimate s_T^2. Notice that the population residual sum of squares can be written

$$\begin{aligned}(\mathbf{y}_T - \mathbf{X}_T\boldsymbol{\beta})'(\mathbf{y}_T - \mathbf{X}_T\boldsymbol{\beta}) \\ &= (\mathbf{y}_T - \mathbf{X}_T\mathbf{b}_T + \mathbf{X}_T\mathbf{b}_T - \mathbf{X}_T\boldsymbol{\beta})'(\mathbf{y}_T - \mathbf{X}_T\mathbf{b}_T + \mathbf{X}_T\mathbf{b}_T - \mathbf{X}_T\boldsymbol{\beta}) \qquad [8.2.10] \\ &= (\mathbf{y}_T - \mathbf{X}_T\mathbf{b}_T)'(\mathbf{y}_T - \mathbf{X}_T\mathbf{b}_T) + (\mathbf{X}_T\mathbf{b}_T - \mathbf{X}_T\boldsymbol{\beta})'(\mathbf{X}_T\mathbf{b}_T - \mathbf{X}_T\boldsymbol{\beta}),\end{aligned}$$

where cross-product terms have vanished, since

$$(\mathbf{y}_T - \mathbf{X}_T\mathbf{b}_T)'\mathbf{X}_T(\mathbf{b}_T - \boldsymbol{\beta}) = 0,$$

by the *OLS* orthogonality condition [8.1.10]. Dividing [8.2.10] by T,

$$\begin{aligned}(1/T)(\mathbf{y}_T - \mathbf{X}_T\boldsymbol{\beta})'(\mathbf{y}_T - \mathbf{X}_T\boldsymbol{\beta}) \\ &= (1/T)(\mathbf{y}_T - \mathbf{X}_T\mathbf{b}_T)'(\mathbf{y}_T - \mathbf{X}_T\mathbf{b}_T) + (1/T)(\mathbf{b}_T - \boldsymbol{\beta})'\mathbf{X}_T'\mathbf{X}_T(\mathbf{b}_T - \boldsymbol{\beta}),\end{aligned}$$

or

$$\begin{aligned}(1/T)(\mathbf{y}_T - \mathbf{X}_T\mathbf{b}_T)'(\mathbf{y}_T - \mathbf{X}_T\mathbf{b}_T) \\ &= (1/T)(\mathbf{u}_T'\mathbf{u}_T) - (\mathbf{b}_T - \boldsymbol{\beta})'(\mathbf{X}_T'\mathbf{X}_T/T)(\mathbf{b}_T - \boldsymbol{\beta}). \qquad [8.2.11]\end{aligned}$$

Now, $(1/T)(\mathbf{u}_T'\mathbf{u}_T) = (1/T)\Sigma_{t=1}^T u_t^2$, where $\{u_t^2\}$ is an i.i.d. sequence with mean σ^2. Thus, by the law of large numbers,

$$(1/T)(\mathbf{u}_T'\mathbf{u}_T) \xrightarrow{p} \sigma^2.$$

For the second term in [8.2.11], we have $(\mathbf{X}_T'\mathbf{X}_T/T) \xrightarrow{p} \mathbf{Q}$ and $(\mathbf{b}_T - \boldsymbol{\beta}) \xrightarrow{p} \mathbf{0}$, and so, from Proposition 7.1,

$$(\mathbf{b}_T - \boldsymbol{\beta})'(\mathbf{X}_T'\mathbf{X}_T/T)(\mathbf{b}_T - \boldsymbol{\beta}) \xrightarrow{p} \mathbf{0}'\mathbf{Q}\mathbf{0} = 0.$$

Substituting these results into [8.2.11],

$$(1/T)(\mathbf{y}_T - \mathbf{X}_T\mathbf{b}_T)'(\mathbf{y}_T - \mathbf{X}_T\mathbf{b}_T) \xrightarrow{p} \sigma^2. \qquad [8.2.12]$$

Now, [8.2.12] describes an estimate of the variance, which we denote $\hat{\sigma}_T^2$:

$$\hat{\sigma}_T^2 \equiv (1/T)(\mathbf{y}_T - \mathbf{X}_T\mathbf{b}_T)'(\mathbf{y}_T - \mathbf{X}_T\mathbf{b}_T). \qquad [8.2.13]$$

The *OLS* estimator given in [8.1.18],

$$s_T^2 = [1/(T - k)](\mathbf{y}_T - \mathbf{X}_T\mathbf{b}_T)'(\mathbf{y}_T - \mathbf{X}_T\mathbf{b}_T), \qquad [8.2.14]$$

differs from $\hat{\sigma}_T^2$ by a term that vanishes as $T \to \infty$,

$$s_T^2 = a_T \cdot \hat{\sigma}_T^2,$$

where $a_T \equiv [T/(T - k)]$ with $\lim_{T\to\infty} a_T = 1$. Hence, from Proposition 7.1,

$$\operatorname{plim} s_T^2 = 1 \cdot \sigma^2,$$

establishing consistency of s_T^2.

To find the asymptotic distribution of s_T^2, consider first $\sqrt{T}(\hat{\sigma}_T^2 - \sigma^2)$. From [8.2.11], this equals

$$\begin{aligned}\sqrt{T}(\hat{\sigma}_T^2 - \sigma^2) &= (1/\sqrt{T})(\mathbf{u}_T'\mathbf{u}_T) - \sqrt{T}\sigma^2 \\ &\quad - \sqrt{T}(\mathbf{b}_T - \boldsymbol{\beta})'(\mathbf{X}_T'\mathbf{X}_T/T)(\mathbf{b}_T - \boldsymbol{\beta}).\end{aligned} \tag{8.2.15}$$

But

$$(1/\sqrt{T})(\mathbf{u}_T'\mathbf{u}_T) - \sqrt{T}\sigma^2 = (1/\sqrt{T}) \sum_{t=1}^{T} (u_t^2 - \sigma^2),$$

where $\{u_t^2 - \sigma^2\}$ is a sequence of i.i.d. variables with mean zero and variance $E(u_t^2 - \sigma^2)^2 = E(u_t^4) - 2\sigma^2 E(u_t^2) + \sigma^4 = \mu_4 - \sigma^4$. Hence, by the central limit theorem,

$$(1/\sqrt{T})(\mathbf{u}_T'\mathbf{u}_T) - \sqrt{T}\sigma^2 \xrightarrow{L} N(0, (\mu_4 - \sigma^4)). \tag{8.2.16}$$

For the last term in [8.2.15], we have $\sqrt{T}(\mathbf{b}_T - \boldsymbol{\beta}) \xrightarrow{L} N(\mathbf{0}, \sigma^2\mathbf{Q}^{-1})$, $(\mathbf{X}_T'\mathbf{X}_T/T) \xrightarrow{p} \mathbf{Q}$, and $(\mathbf{b}_T - \boldsymbol{\beta}) \xrightarrow{p} \mathbf{0}$. Hence,

$$\sqrt{T}(\mathbf{b}_T - \boldsymbol{\beta})'(\mathbf{X}_T'\mathbf{X}_T/T)(\mathbf{b}_T - \boldsymbol{\beta}) \xrightarrow{p} 0. \tag{8.2.17}$$

Putting [8.2.16] and [8.2.17] into [8.2.15], we conclude

$$\sqrt{T}(\hat{\sigma}_T^2 - \sigma^2) \xrightarrow{L} N(0, (\mu_4 - \sigma^4)). \tag{8.2.18}$$

To see that s_T^2 has this same limiting distribution, notice that

$$\begin{aligned}\sqrt{T}(s_T^2 - \sigma^2) - \sqrt{T}(\hat{\sigma}_T^2 - \sigma^2) &= \sqrt{T}\{[T/(T-k)]\hat{\sigma}_T^2 - \hat{\sigma}_T^2\} \\ &= [(k\sqrt{T})/(T-k)]\hat{\sigma}_T^2.\end{aligned}$$

But $\lim_{T\to\infty} [(k\sqrt{T})/(T-k)] = 0$, establishing that

$$\sqrt{T}(s_T^2 - \sigma^2) - \sqrt{T}(\hat{\sigma}_T^2 - \sigma^2) \xrightarrow{p} 0 \cdot \sigma^2 = 0$$

and hence, from Proposition 7.3(a),

$$\sqrt{T}(s_T^2 - \sigma^2) \xrightarrow{L} N(0, (\mu_4 - \sigma^4)). \tag{8.2.19}$$

Notice that if we are relying on asymptotic justifications for test statistics, theory offers us no guidance for choosing between s^2 and $\hat{\sigma}^2$ as estimates of σ^2, since they have the same limiting distribution.

Next consider the asymptotic distribution of the *OLS* t test of the null hypothesis $\beta_i = \beta_i^0$,

$$t_T = \frac{(b_{iT} - \beta_i^0)}{s_T\sqrt{\xi_T^{ii}}} = \frac{\sqrt{T}(b_{iT} - \beta_i^0)}{s_T\sqrt{T\xi_T^{ii}}}, \tag{8.2.20}$$

where ξ_T^{ii} denotes the row i, column i element of $(\mathbf{X}_T'\mathbf{X}_T)^{-1}$. We have seen that $\sqrt{T}(b_{i,T} - \beta^0) \xrightarrow{L} N(0, \sigma^2 q^{ii})$, where q^{ii} denotes the row i, column i element of $\mathbf{Q}^{-1}$. Similarly, $T\xi_T^{ii}$ is the row i, column i element of $(\mathbf{X}_T'\mathbf{X}_T/T)^{-1}$ and converges in probability to q^{ii}. Also, $s_T \xrightarrow{p} \sigma$. Hence, the t statistic [8.2.20] has a limiting distribution that is the same as a $N(0, \sigma^2 q^{ii})$ variable divided by $\sqrt{\sigma^2 q^{ii}}$; that is,

$$t_T \xrightarrow{L} N(0, 1). \tag{8.2.21}$$

Now, under the more restrictive conditions of Assumption 8.2, we saw that t_T would have a t distribution with $(T - k)$ degrees of freedom. Recall that a t variable with N degrees of freedom has the distribution of the ratio of a $N(0, 1)$ variable to the square root of $(1/N)$ times an independent $\chi^2(N)$ variable. But a $\chi^2(N)$ variable in turn is the sum of N squares of independent $N(0, 1)$ variables.

Thus, letting Z denote a $N(0, 1)$ variable, a t variable with N degrees of freedom has the same distribution as

$$t_N = \frac{Z}{\{(Z_1^2 + Z_2^2 + \cdots + Z_N^2)/N\}^{1/2}}.$$

By the law of large numbers,

$$(Z_1^2 + Z_2^2 + \cdots + Z_N^2)/N \xrightarrow{p} E(Z_i^2) = 1,$$

and so $t_N \xrightarrow{L} N(0, 1)$. Hence, the critical value for a t variable with N degrees of freedom will be arbitrarily close to that for a $N(0, 1)$ variable as N becomes large. Even though the statistic calculated in [8.2.20] does not have an exact $t(T - k)$ distribution under Assumption 8.3, if we treat it as if it did, then we will not be far wrong if our sample is sufficiently large.

The same is true of [8.1.32], the F test of m different restrictions:

$$\begin{aligned} F_T &= (\mathbf{R}\mathbf{b}_T - \mathbf{r})'[s_T^2\mathbf{R}(\mathbf{X}_T'\mathbf{X}_T)^{-1}\mathbf{R}']^{-1}(\mathbf{R}\mathbf{b}_T - \mathbf{r})/m \\ &= \sqrt{T}(\mathbf{R}\mathbf{b}_T - \mathbf{r})'[s_T^2\mathbf{R}(\mathbf{X}_T'\mathbf{X}_T/T)^{-1}\mathbf{R}']^{-1}\sqrt{T}(\mathbf{R}\mathbf{b}_T - \mathbf{r})/m. \end{aligned} \qquad [8.2.22]$$

Here $s_T^2 \xrightarrow{p} \sigma^2$, $\mathbf{X}_T'\mathbf{X}_T/T \xrightarrow{p} \mathbf{Q}$, and, under the null hypothesis,

$$\begin{aligned} \sqrt{T}(\mathbf{R}\mathbf{b}_T - \mathbf{r}) &= [\mathbf{R}\sqrt{T}(\mathbf{b}_T - \boldsymbol{\beta})] \\ &\xrightarrow{L} N(\mathbf{0}, \sigma^2\mathbf{R}\mathbf{Q}^{-1}\mathbf{R}'). \end{aligned}$$

Hence, under the null hypothesis,

$$m \cdot F_T \xrightarrow{p} [\mathbf{R}\sqrt{T}(\mathbf{b}_T - \boldsymbol{\beta})]'[\sigma^2\mathbf{R}\mathbf{Q}^{-1}\mathbf{R}']^{-1}[\mathbf{R}\sqrt{T}(\mathbf{b}_T - \boldsymbol{\beta})].$$

This is a quadratic function of a Normal vector of the type described by Proposition 8.1, from which

$$m \cdot F_T \xrightarrow{L} \chi^2(m).$$

Thus an asymptotic inference can be based on the approximation

$$(\mathbf{R}\mathbf{b}_T - \mathbf{r})'[s_T^2\mathbf{R}(\mathbf{X}_T'\mathbf{X}_T)^{-1}\mathbf{R}']^{-1}(\mathbf{R}\mathbf{b}_T - \mathbf{r}) \approx \chi^2(m). \qquad [8.2.23]$$

This is known as the *Wald form* of the *OLS* χ^2 test.

As in the case of the t and limiting Normal distributions, viewing [8.2.23] as $\chi^2(m)$ and viewing [8.2.22] as $F(m, T - k)$ asymptotically amount to the same test. Recall that an $F(m, N)$ variable is a ratio of a $\chi^2(m)$ variable to an independent $\chi^2(N)$ variable, each divided by its degrees of freedom. Thus, if Z_i denotes a $N(0, 1)$ variable and X a $\chi^2(m)$ variable,

$$F_{m,N} = \frac{X/m}{(Z_1^2 + Z_2^2 + \cdots + Z_N^2)/N}.$$

For the denominator,

$$(Z_1^2 + Z_2^2 + \cdots + Z_N^2)/N \xrightarrow{p} E(Z_i^2) = 1,$$

implying

$$F_{m,N} \underset{N\to\infty}{\xrightarrow{L}} X/m.$$

Hence, comparing [8.2.23] with a $\chi^2(m)$ critical value or comparing [8.2.22] with an $F(m, T - k)$ critical value will result in the identical test for sufficiently large T (see Exercise 8.2).

For a given sample of size T, the small-sample distribution (the t or F distribution) implies wider confidence intervals than the large-sample distribution (the

Normal or χ^2 distribution). Even when the justification for using the t or F distribution is only asymptotic, many researchers prefer to use the t or F tables rather than the Normal or χ^2 tables on the grounds that the former are more conservative and may represent a better approximation to the true small-sample distribution.

If we are relying only on the asymptotic distribution, the Wald test statistic [8.2.23] can be generalized to allow a test of a nonlinear set of restrictions on $\boldsymbol{\beta}$. Consider a null hypothesis consisting of m separate nonlinear restrictions of the form $\mathbf{g}(\boldsymbol{\beta}) = \mathbf{0}$ where $\mathbf{g}: \mathbb{R}^k \to \mathbb{R}^m$ and $\mathbf{g}(\cdot)$ has continuous first derivatives. Result [8.2.8] and Proposition 7.4 imply that

$$\sqrt{T}[\mathbf{g}(\mathbf{b}_T) - \mathbf{g}(\boldsymbol{\beta}_0)] \xrightarrow{L} \left[\left.\frac{\partial \mathbf{g}}{\partial \boldsymbol{\beta}'}\right|_{\boldsymbol{\beta}=\boldsymbol{\beta}_0}\right]\mathbf{z},$$

where $\mathbf{z} \sim N(\mathbf{0}, \sigma^2\mathbf{Q}^{-1})$ and

$$\left.\frac{\partial \mathbf{g}}{\partial \boldsymbol{\beta}'}\right|_{\boldsymbol{\beta}=\boldsymbol{\beta}_0}$$

denotes the $(m \times k)$ matrix of derivatives of $\mathbf{g}(\cdot)$ with respect to $\boldsymbol{\beta}$, evaluated at the true value $\boldsymbol{\beta}_0$. Under the null hypothesis that $\mathbf{g}(\boldsymbol{\beta}_0) = \mathbf{0}$, it follows from Proposition 8.1 that

$$\{\sqrt{T}\cdot\mathbf{g}(\mathbf{b}_T)\}'\left\{\left[\left.\frac{\partial \mathbf{g}}{\partial \boldsymbol{\beta}'}\right|_{\boldsymbol{\beta}=\boldsymbol{\beta}_0}\right]\sigma^2\mathbf{Q}^{-1}\left[\left.\frac{\partial \mathbf{g}}{\partial \boldsymbol{\beta}'}\right|_{\boldsymbol{\beta}=\boldsymbol{\beta}_0}\right]'\right\}^{-1}\{\sqrt{T}\cdot\mathbf{g}(\mathbf{b}_T)\} \xrightarrow{L} \chi^2(m).$$

Recall that $\mathbf{Q}$ is the plim of $(1/T)(\mathbf{X}_T'\mathbf{X}_T)$. Since $\partial\mathbf{g}/\partial\boldsymbol{\beta}'$ is continuous and since $\mathbf{b}_T \xrightarrow{p} \boldsymbol{\beta}_0$, it follows from Proposition 7.1 that

$$\left[\left.\frac{\partial \mathbf{g}}{\partial \boldsymbol{\beta}'}\right|_{\boldsymbol{\beta}=\mathbf{b}_T}\right] \xrightarrow{p} \left[\left.\frac{\partial \mathbf{g}}{\partial \boldsymbol{\beta}'}\right|_{\boldsymbol{\beta}=\boldsymbol{\beta}_0}\right].$$

Hence a set of m nonlinear restrictions about $\boldsymbol{\beta}$ of the form $\mathbf{g}(\boldsymbol{\beta}) = \mathbf{0}$ can be tested with the statistic

$$\{\mathbf{g}(\mathbf{b}_T)\}'\left\{\left[\left.\frac{\partial \mathbf{g}}{\partial \boldsymbol{\beta}'}\right|_{\boldsymbol{\beta}=\mathbf{b}_T}\right]s_T^2(\mathbf{X}_T'\mathbf{X}_T)^{-1}\left[\left.\frac{\partial \mathbf{g}}{\partial \boldsymbol{\beta}'}\right|_{\boldsymbol{\beta}=\mathbf{b}_T}\right]'\right\}^{-1}\{\mathbf{g}(\mathbf{b}_T)\} \xrightarrow{L} \chi^2(m).$$

Note that the Wald test for linear restrictions [8.2.23] can be obtained as a special case of this more general formula by setting $\mathbf{g}(\boldsymbol{\beta}) = \mathbf{R}\boldsymbol{\beta} - \mathbf{r}$.

One disadvantage of the Wald test for nonlinear restrictions is that the answer one obtains can be different depending on how the restrictions $\mathbf{g}(\boldsymbol{\beta}) = \mathbf{0}$ are parameterized. For example, the hypotheses $\beta_1 = \beta_2$ and $\beta_1/\beta_2 = 1$ are equivalent, and asymptotically a Wald test based on either parameterization should give the same answer. However, in a particular finite sample the answers could be quite different. In effect, the nonlinear Wald test approximates the restriction $\mathbf{g}(\mathbf{b}_T) = \mathbf{0}$ by the linear restriction

$$\mathbf{g}(\boldsymbol{\beta}_0) + \left[\left.\frac{\partial \mathbf{g}}{\partial \boldsymbol{\beta}'}\right|_{\boldsymbol{\beta}=\boldsymbol{\beta}_0}\right](\mathbf{b}_T - \boldsymbol{\beta}_0) = \mathbf{0}.$$

Some care must be taken to ensure that this linearization is reasonable over the range of plausible values for $\boldsymbol{\beta}$. See Gregory and Veall (1985), Lafontaine and White (1986), and Phillips and Park (1988) for further discussion.

Case 4. Estimating Parameters for an Autoregression

Consider now estimation of the parameters of a pth-order autoregression by *OLS*.

Assumption 8.4: *The regression model is*

$$y_t = c + \phi_1 y_{t-1} + \phi_2 y_{t-2} + \cdots + \phi_p y_{t-p} + \varepsilon_t, \qquad [8.2.24]$$

with roots of $(1 - \phi_1 z - \phi_2 z^2 - \cdots - \phi_p z^p) = 0$ *outside the unit circle and with* $\{\varepsilon_t\}$ *an i.i.d. sequence with mean zero, variance* σ^2, *and finite fourth moment* μ_4.

An autoregression has the form of the standard regression model $y_t = \mathbf{x}_t'\boldsymbol{\beta} + u_t$ with $\mathbf{x}_t' = (1, y_{t-1}, y_{t-2}, \ldots, y_{t-p})$ and $u_t = \varepsilon_t$. Note, however, that an autoregression cannot satisfy condition (a) of Assumption 8.2 or 8.3. Even though u_t is independent of $\mathbf{x}_t$ under Assumption 8.4, it will not be the case that u_t is independent of $\mathbf{x}_{t+1}$. Without this independence, none of the small-sample results for case 1 applies. Specifically, even if ε_t is Gaussian, the *OLS* coefficient $\mathbf{b}$ gives a biased estimate of $\boldsymbol{\beta}$ for an autoregression, and the standard t and F statistics can only be justified asymptotically.

However, the asymptotic results for case 4 are the same as for case 3 and are derived in essentially the same way. To adapt the earlier notation, suppose that the sample consists of $T + p$ observations on y_t, numbered $(y_{-p+1}, y_{-p+2}, \ldots, y_0, y_1, \ldots, y_T)$; *OLS* estimation will thus use observations 1 through T. Then, as in [8.2.6],

$$\sqrt{T}(\mathbf{b}_T - \boldsymbol{\beta}) = \left[(1/T)\sum_{t=1}^{T} \mathbf{x}_t\mathbf{x}_t'\right]^{-1}\left[(1/\sqrt{T})\sum_{t=1}^{T} \mathbf{x}_t u_t\right]. \qquad [8.2.25]$$

The first term in [8.2.25] is

$$\left[(1/T)\sum_{t=1}^{T} \mathbf{x}_t\mathbf{x}_t'\right]^{-1}$$

$$= \begin{bmatrix} 1 & T^{-1}\Sigma y_{t-1} & T^{-1}\Sigma y_{t-2} & \cdots & T^{-1}\Sigma y_{t-p} \\ T^{-1}\Sigma y_{t-1} & T^{-1}\Sigma y_{t-1}^2 & T^{-1}\Sigma y_{t-1}y_{t-2} & \cdots & T^{-1}\Sigma y_{t-1}y_{t-p} \\ T^{-1}\Sigma y_{t-2} & T^{-1}\Sigma y_{t-2}y_{t-1} & T^{-1}\Sigma y_{t-2}^2 & \cdots & T^{-1}\Sigma y_{t-2}y_{t-p} \\ \vdots & \vdots & \vdots & \cdots & \vdots \\ T^{-1}\Sigma y_{t-p} & T^{-1}\Sigma y_{t-p}y_{t-1} & T^{-1}\Sigma y_{t-p}y_{t-2} & \cdots & T^{-1}\Sigma y_{t-p}^2 \end{bmatrix}^{-1}.$$

where Σ denotes summation over $t = 1$ to T. The elements in the first row or column are of the form $T^{-1}\Sigma y_{t-j}$ and converge in probability to $\mu = E(y_t)$, by Proposition 7.5. Other elements are of the form $T^{-1}\Sigma y_{t-i}y_{t-j}$, which, from [7.2.14], converges in probability to

$$E(y_{t-i}y_{t-j}) = \gamma_{|i-j|} + \mu^2.$$

Hence

$$\left[(1/T)\sum_{t=1}^{T} \mathbf{x}_t\mathbf{x}_t'\right]^{-1} \xrightarrow{p} \mathbf{Q}^{-1} \qquad [8.2.26]$$

where

$$\mathbf{Q} \equiv \begin{bmatrix} 1 & \mu & \mu & \cdots & \mu \\ \mu & \gamma_0 + \mu^2 & \gamma_1 + \mu^2 & \cdots & \gamma_{p-1} + \mu^2 \\ \mu & \gamma_1 + \mu^2 & \gamma_0 + \mu^2 & \cdots & \gamma_{p-2} + \mu^2 \\ \vdots & \vdots & \vdots & \cdots & \vdots \\ \mu & \gamma_{p-1} + \mu^2 & \gamma_{p-2} + \mu^2 & \cdots & \gamma_0 + \mu^2 \end{bmatrix}. \qquad [8.2.27]$$

For the second term in [8.2.25], observe that $\mathbf{x}_t u_t$ is a martingale difference sequence with positive definite variance-covariance matrix given by

$$E(\mathbf{x}_t u_t u_t \mathbf{x}_t') = E(u_t^2) \cdot E(\mathbf{x}_t \mathbf{x}_t') = \sigma^2 \mathbf{Q}.$$

Using an argument similar to that in Example 7.15, it can be shown that

$$\left[(1/\sqrt{T}) \sum_{t=1}^{T} \mathbf{x}_t u_t \right] \xrightarrow{L} N(\mathbf{0}, \sigma^2 \mathbf{Q}) \qquad [8.2.28]$$

(see Exercise 8.3). Substituting [8.2.26] and [8.2.28] into [8.2.25],

$$\sqrt{T}(\mathbf{b}_T - \boldsymbol{\beta}) \xrightarrow{L} N(\mathbf{0}, \sigma^2 \mathbf{Q}^{-1}). \qquad [8.2.29]$$

It is straightforward to verify further that $\mathbf{b}_T$ and s_T^2 are consistent for this case. From [8.2.26], the asymptotic variance-covariance matrix of $\sqrt{T}(\mathbf{b}_T - \boldsymbol{\beta})$ can be estimated consistently by $s_T^2(\mathbf{X}_T'\mathbf{X}_T/T)^{-1}$, meaning that standard t and F statistics that treat $\mathbf{b}_T$ as if it were $N(\boldsymbol{\beta}, s_T^2(\mathbf{X}_T'\mathbf{X}_T)^{-1})$ will yield asymptotically valid tests of hypotheses about the coefficients of an autoregression.

As a special case of [8.2.29], consider *OLS* estimation of a first-order autoregression,

$$y_t = \phi y_{t-1} + \varepsilon_t,$$

with $|\phi| < 1$. Then $\mathbf{Q}$ is the scalar $E(y_{t-1}^2) = \gamma_0$, the variance of an $AR(1)$ process. We saw in Chapter 3 that this is given by $\sigma^2/(1 - \phi^2)$. Hence, for $\hat{\phi}$ the *OLS* coefficient,

$$\hat{\phi}_T = \frac{\sum_{t=1}^{T} y_{t-1} y_t}{\sum_{t=1}^{T} y_{t-1}^2},$$

result [8.2.29] implies that

$$\sqrt{T}(\hat{\phi}_T - \phi) \xrightarrow{L} N(0, \sigma^2 \cdot [\sigma^2/(1 - \phi^2)]^{-1}) = N(0, 1 - \phi^2). \qquad [8.2.30]$$

If more precise results than the asymptotic approximation in equation [8.2.29] are desired, the exact small-sample distribution of $\hat{\phi}_T$ can be calculated in either of two ways. If the errors in the autoregression [8.2.24] are $N(0, \sigma^2)$, then for any specified numerical value for $\phi_1, \phi_2, \ldots, \phi_p$ and c the exact small-sample distribution can be calculated using numerical routines developed by Imhof (1961); for illustrations of this method, see Evans and Savin (1981) and Flavin (1983). An alternative is to approximate the small-sample distribution by *Monte Carlo* methods. Here the idea is to use a computer to generate pseudo-random variables $\varepsilon_1, \ldots, \varepsilon_T$, each distributed $N(0, \sigma^2)$ from numerical algorithms such as that described in Kinderman and Ramage (1976). For fixed starting values $y_{-p+1}, \ldots, y_1$, the

values for $y_1, y_2, \ldots, y_T$ can then be calculated by iterating on [8.2.24].[6] One then estimates the parameters of [8.2.24] with an *OLS* regression on this artificial sample. A new sample is generated for which a new *OLS* regression is estimated. By performing, say, 10,000 such regressions, an estimate of the exact small-sample distribution of the *OLS* estimates can be obtained.

For the case of a first-order autoregression, it is known from such calculations that $\hat{\phi}_T$ is downward-biased in small samples, with the bias becoming more severe as ϕ approaches unity. For example, for a sample of size $T = 25$ generated by [8.2.24] with $p = 1$, $c = 0$, and $\phi = 1$, the estimate $\hat{\phi}_T$ based on *OLS* estimation of [8.2.24] (with a constant term included) will be less than the true value of 1 in 95% of the samples, and will even fall below 0.6 in 10% of the samples.[7]

Case 5. Errors Gaussian with Known Variance-Covariance Matrix

Next consider the following case.

Assumption 8.5: *(a)* $\mathbf{x}_t$ *stochastic; (b) conditional on the full matrix* $\mathbf{X}$, *the vector* $\mathbf{u}$ *is* $N(\mathbf{0}, \sigma^2\mathbf{V})$; *(c)* $\mathbf{V}$ *is a known positive definite matrix.*

When the errors for different dates have different variances but are uncorrelated with each other (that is, $\mathbf{V}$ is diagonal), then the errors are said to exhibit *heteroskedasticity*. For $\mathbf{V}$ nondiagonal, the errors are said to be *autocorrelated*. Writing the variance-covariance matrix as the product of some scalar σ^2 and a matrix $\mathbf{V}$ is a convention that will help simplify the algebra and interpretation for some examples of heteroskedasticity and autocorrelation. Note again that Assumption 8.5(b) could not hold for an autoregression, since conditional on $\mathbf{x}_{t+1} = (1, y_t, y_{t-1}, \ldots, y_{t-p+1})'$ and $\mathbf{x}_t$, the value of u_t is known with certainty.

Recall from [8.1.12] that

$$(\mathbf{b} - \boldsymbol{\beta}) = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{u}.$$

Taking expectations conditional on $\mathbf{X}$,

$$E[(\mathbf{b} - \boldsymbol{\beta})|\mathbf{X}] = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\cdot E(\mathbf{u}) = \mathbf{0},$$

and by the law of iterated expectations,

$$E(\mathbf{b} - \boldsymbol{\beta}) = E_{\mathbf{X}}\{E[(\mathbf{b} - \boldsymbol{\beta})|\mathbf{X}]\} = \mathbf{0}.$$

Hence, the *OLS* coefficient estimate is unbiased.

The variance of $\mathbf{b}$ conditional on $\mathbf{X}$ is

$$\begin{aligned} E\{[(\mathbf{b} - \boldsymbol{\beta})(\mathbf{b} - \boldsymbol{\beta})']|\mathbf{X}\} &= E\{[(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{u}\mathbf{u}'\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}]|\mathbf{X}\} \\ &= \sigma^2(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{V}\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}. \end{aligned} \qquad [8.2.31]$$

Thus, conditional on $\mathbf{X}$,

$$\mathbf{b}|\mathbf{X} \sim N\Big(\boldsymbol{\beta}, \sigma^2(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{V}\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\Big).$$

[6]Alternatively, one can generate the initial values for y with a draw from the appropriate unconditional distribution. Specifically, generate a $(p \times 1)$ vector $\mathbf{v} \sim N(\mathbf{0}, \mathbf{I}_p)$ and set $(y_{-p+1}, \ldots, y_0)' = \mu\cdot\mathbf{1} + \mathbf{P}\cdot\mathbf{v}$, where $\mu = c/(1 - \phi_1 - \phi_2 - \cdots - \phi_p)$, $\mathbf{1}$ denotes a $(p \times 1)$ vector of 1s, and $\mathbf{P}$ is the Cholesky factor such that $\mathbf{P}\cdot\mathbf{P}' = \boldsymbol{\Gamma}$ for $\boldsymbol{\Gamma}$ the $(p \times p)$ matrix whose columns stacked in a $(p^2 \times 1)$ vector comprise the first column of the matrix $\sigma^2[\mathbf{I}_{p^2} - (\mathbf{F} \otimes \mathbf{F})]^{-1}$, where $\mathbf{F}$ is the $(p \times p)$ matrix defined in equation [1.2.3] in Chapter 1.

[7]These values can be inferred from Table B.5.

Unless $\mathbf{V} = \mathbf{I}_T$, this is not the same variance matrix as in [8.1.17], so that the *OLS* t statistic [8.1.26] does not have the interpretation as a Gaussian variable divided by an estimate of its standard deviation. Thus [8.1.26] will not have a $t(T - k)$ distribution in small samples, nor will it even asymptotically be $N(0, 1)$. A valid test of the hypothesis that $\beta_i = \beta_i^0$ for case 5 would be based not on [8.1.26] but rather on

$$t^* = \frac{(b_i - \beta_i^0)}{s\sqrt{d_{ii}}}, \qquad [8.2.32]$$

where d_{ii} denotes the row i, column i element of $(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{V}\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}$. This statistic will be asymptotically $N(0, 1)$.

Although one could form an inference based on [8.2.32], in this case in which $\mathbf{V}$ is known, a superior estimator and test procedure are described in Section 8.3. First, however, we consider a more general case in which $\mathbf{V}$ is of unknown form.

Case 6. Errors Serially Uncorrelated but with General Heteroskedasticity

It may be possible to design asymptotically valid tests even in the presence of heteroskedasticity of a completely unknown form. This point was first observed by Eicker (1967) and White (1980) and extended to time series regressions by Hansen (1982) and Nicholls and Pagan (1983).

Assumption 8.6: *(a) $\mathbf{x}_t$ stochastic, including perhaps lagged values of y; (b) $\mathbf{x}_t u_t$ is a martingale difference sequence; (c) $E(u_t^2\mathbf{x}_t\mathbf{x}_t') = \mathbf{\Omega}_t$, a positive definite matrix, with $(1/T)\Sigma_{t=1}^T\mathbf{\Omega}_t$ converging to the positive definite matrix $\mathbf{\Omega}$ and $(1/T)\Sigma_{t=1}^T u_t^2\mathbf{x}_t\mathbf{x}_t' \xrightarrow{p} \mathbf{\Omega}$; (d) $E(u_t^4 x_{it}x_{jt}x_{lt}x_{mt}) < \infty$ for all i, j, l, m, and t; (e) plims of $(1/T)\Sigma_{t=1}^T u_t x_{it}\mathbf{x}_t\mathbf{x}_t'$ and $(1/T)\Sigma_{t=1}^T x_{it}x_{jt}\mathbf{x}_t\mathbf{x}_t'$ exist and are finite for all i and j and $(1/T)\Sigma_{t=1}^T\mathbf{x}_t\mathbf{x}_t' \xrightarrow{p} \mathbf{Q}$, a nonsingular matrix.*

Assumption 8.6(b) requires u_t to be uncorrelated with its own lagged values and with current and lagged values of $\mathbf{x}$. Although the errors are presumed to be serially uncorrelated, Assumption 8.6(c) allows a broad class of conditional heteroskedasticity for the errors. As an example of such heteroskedasticity, consider a regression with a single i.i.d. explanatory variable x_t with $E(x_t^2) = \mu_2$ and $E(x_t^4) = \mu_4$. Suppose that the variance of the residual for date t is given by $E(u_t^2|x_t) = a + bx_t^2$. Then $E(u_t^2x_t^2) = E_x[E(u_t^2|x_t)\cdot x_t^2] = E_x[(a + bx_t^2)\cdot x_t^2] = a\mu_2 + b\mu_4$. Thus, $\Omega_t = a\mu_2 + b\mu_4 = \Omega$ for all t. By the law of large numbers, $(1/T)\Sigma_{t=1}^T u_t^2x_t^2$ will converge to the population moment Ω. Assumption 8.6(c) allows more general conditional heteroskedasticity in that $E(u_t^2x_t^2)$ might be a function of t, provided that the time average of $(u_t^2x_t^2)$ converges. Assumption 8.6(d) and (e) impose bounds on higher moments of $\mathbf{x}$ and u.

Consistency of $\mathbf{b}$ is established using the same arguments as in case 3. The asymptotic variance is found from writing

$$\sqrt{T}(\mathbf{b}_T - \boldsymbol{\beta}) = \left[(1/T)\sum_{t=1}^T \mathbf{x}_t\mathbf{x}_t'\right]^{-1}\left[(1/\sqrt{T})\sum_{t=1}^T \mathbf{x}_t u_t\right].$$

Assumption 8.6(e) ensures that

$$\left[(1/T)\sum_{t=1}^T \mathbf{x}_t\mathbf{x}_t'\right]^{-1} \xrightarrow{p} \mathbf{Q}^{-1}$$

for some nonsingular matrix $\mathbf{Q}$. Similarly, $\mathbf{x}_t u_t$ satisfies the conditions of Proposition 7.9, from which

$$\left[(1/\sqrt{T})\sum_{t=1}^{T}\mathbf{x}_t u_t\right] \xrightarrow{L} N(\mathbf{0}, \mathbf{\Omega}).$$

The asymptotic distribution of the *OLS* estimate is thus given by

$$\sqrt{T}(\mathbf{b}_T - \boldsymbol{\beta}) \xrightarrow{L} N(\mathbf{0}, \mathbf{Q}^{-1}\mathbf{\Omega}\mathbf{Q}^{-1}). \qquad [8.2.33]$$

White's proposal was to estimate the asymptotic variance matrix consistently by substituting $\hat{\mathbf{Q}}_T = (1/T)\Sigma_{t=1}^{T}\mathbf{x}_t\mathbf{x}_t'$ and $\hat{\mathbf{\Omega}}_T = (1/T)\Sigma_{t=1}^{T}\hat{u}_t^2\mathbf{x}_t\mathbf{x}_t'$ into [8.2.33], where $\hat{u}_t$ denotes the *OLS* residual [8.1.4]. The following result is established in Appendix 8.A to this chapter.

Proposition 8.3: *With heteroskedasticity of unknown form satisfying Assumption 8.6, the asymptotic variance-covariance matrix of the OLS coefficient vector can be consistently estimated by*

$$\hat{\mathbf{Q}}_T^{-1}\hat{\mathbf{\Omega}}_T\hat{\mathbf{Q}}_T^{-1} \xrightarrow{p} \mathbf{Q}^{-1}\mathbf{\Omega}\mathbf{Q}^{-1}. \qquad [8.2.34]$$

Recalling [8.2.33], the *OLS* estimate $\mathbf{b}_T$ can be treated as if

$$\mathbf{b}_T \approx N(\boldsymbol{\beta}, \hat{\mathbf{V}}_T/T)$$

where

$$\begin{aligned}\hat{\mathbf{V}}_T &= \hat{\mathbf{Q}}_T^{-1}\hat{\mathbf{\Omega}}_T\hat{\mathbf{Q}}_T^{-1}\\ &= (\mathbf{X}_T'\mathbf{X}_T/T)^{-1}\left[(1/T)\sum_{t=1}^{T}\hat{u}_t^2\mathbf{x}_t\mathbf{x}_t'\right](\mathbf{X}_T'\mathbf{X}_T/T)^{-1} \qquad [8.2.35]\\ &= T\cdot(\mathbf{X}_T'\mathbf{X}_T)^{-1}\left[\sum_{t=1}^{T}\hat{u}_t^2\mathbf{x}_t\mathbf{x}_t'\right](\mathbf{X}_T'\mathbf{X}_t)^{-1}.\end{aligned}$$

The square root of the row i, column i element of $\hat{\mathbf{V}}_T/T$ is known as a *heteroskedasticity-consistent standard error* for the *OLS* estimate b_i. We can, of course, also use $(\hat{\mathbf{V}}_T/T)$ to test a joint hypothesis of the form $\mathbf{R}\boldsymbol{\beta} = \mathbf{r}$, where $\mathbf{R}$ is an $(m \times k)$ matrix summarizing m separate hypotheses about $\boldsymbol{\beta}$. Specifically,

$$(\mathbf{R}\mathbf{b}_T - \mathbf{r})'[\mathbf{R}(\hat{\mathbf{V}}_T/T)\mathbf{R}']^{-1}(\mathbf{R}\mathbf{b}_T - \mathbf{r}) \qquad [8.2.36]$$

has the same asymptotic distribution as

$$[\sqrt{T}(\mathbf{R}\mathbf{b}_T - \mathbf{r})]'(\mathbf{R}\mathbf{Q}^{-1}\mathbf{\Omega}\mathbf{Q}^{-1}\mathbf{R}')^{-1}[\sqrt{T}(\mathbf{R}\mathbf{b}_T - \mathbf{r})],$$

which, from [8.2.33], is a quadratic form of an asymptotically Normal $(m \times 1)$ vector $\sqrt{T}(\mathbf{R}\mathbf{b}_T - \mathbf{r})$ with weighting matrix the inverse of its variance-covariance matrix, $(\mathbf{R}\mathbf{Q}^{-1}\mathbf{\Omega}\mathbf{Q}^{-1}\mathbf{R}')$. Hence, [8.2.36] has an asymptotic χ^2 distribution with m degrees of freedom.

It is also possible to develop an estimate of the asymptotic variance-covariance matrix of $\mathbf{b}_T$ that is robust with respect to both heteroskedasticity and autocorrelation:

$$\begin{aligned}&(\hat{\mathbf{V}}_T/T)\\ &= (\mathbf{X}_T'\mathbf{X}_T)^{-1}\left[\sum_{t=1}^{T}\hat{u}_t^2\mathbf{x}_t\mathbf{x}_t'\right.\\ &\quad\left.+\sum_{\nu=1}^{q}\left[1-\frac{\nu}{q+1}\right]\sum_{t=\nu+1}^{T}(\mathbf{x}_t\hat{u}_t\hat{u}_{t-\nu}\mathbf{x}_{t-\nu}' + \mathbf{x}_{t-\nu}\hat{u}_{t-\nu}\hat{u}_t\mathbf{x}_t')\right](\mathbf{X}_T'\mathbf{X}_T)^{-1}.\end{aligned}$$

Here q is a parameter representing the number of autocorrelations used to approximate the dynamics for u_t. The square root of the row i, column i element of $(\hat{\mathbf{V}}_T/T)$ is known as the Newey-West (1987) heteroskedasticity- and autocorrelation-consistent standard error for the *OLS* estimator. The basis for this expression and alternative ways to calculate heteroskedasticity- and autocorrelation-consistent standard errors will be discussed in Chapter 10.

8.3. *Generalized Least Squares*

The previous section evaluated *OLS* estimation under a variety of assumptions, including $E(\mathbf{uu}') \neq \sigma^2\mathbf{I}_T$. Although *OLS* can be used in this last case, generalized least squares (*GLS*) is usually preferred.

GLS *with Known Covariance Matrix*

Let us reconsider data generated according to Assumption 8.5, under which $\mathbf{u}|\mathbf{X} \sim N(\mathbf{0}, \sigma^2\mathbf{V})$ with $\mathbf{V}$ a known $(T \times T)$ matrix. Since $\mathbf{V}$ is symmetric and positive definite, there exists a nonsingular $(T \times T)$ matrix $\mathbf{L}$ such that[8]

$$\mathbf{V}^{-1} = \mathbf{L}'\mathbf{L}. \qquad [8.3.1]$$

Imagine transforming the population residuals $\mathbf{u}$ by $\mathbf{L}$:

$$\underset{(T\times 1)}{\tilde{\mathbf{u}}} \equiv \mathbf{Lu}.$$

This would generate a new set of residuals $\tilde{\mathbf{u}}$ with mean $\mathbf{0}$ and variance conditional on $\mathbf{X}$ given by

$$E(\tilde{\mathbf{u}}\tilde{\mathbf{u}}'|\mathbf{X}) = \mathbf{L}\cdot E(\mathbf{uu}'|\mathbf{X})\mathbf{L}' = \mathbf{L}\sigma^2\mathbf{VL}'.$$

But $\mathbf{V} = [\mathbf{V}^{-1}]^{-1} = [\mathbf{L}'\mathbf{L}]^{-1}$, meaning

$$E(\tilde{\mathbf{u}}\tilde{\mathbf{u}}'|\mathbf{X}) = \sigma^2\mathbf{L}[\mathbf{L}'\mathbf{L}]^{-1}\mathbf{L}' = \sigma^2\mathbf{I}_T. \qquad [8.3.2]$$

We can thus take the matrix equation that characterizes the basic regression model,

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{u},$$

and premultiply both sides by $\mathbf{L}$:

$$\mathbf{Ly} = \mathbf{LX}\boldsymbol{\beta} + \mathbf{Lu},$$

to produce a new regression model

$$\tilde{\mathbf{y}} = \tilde{\mathbf{X}}\boldsymbol{\beta} + \tilde{\mathbf{u}}, \qquad [8.3.3]$$

where

$$\tilde{\mathbf{y}} \equiv \mathbf{Ly} \qquad \tilde{\mathbf{X}} \equiv \mathbf{LX} \qquad \tilde{\mathbf{u}} \equiv \mathbf{Lu} \qquad [8.3.4]$$

with $\tilde{\mathbf{u}}|\mathbf{X} \sim N(\mathbf{0}, \sigma^2\mathbf{I}_T)$. Hence, the transformed model [8.3.3] satisfies Assumption 8.2, meaning that all the results for that case apply to [8.3.3]. Specifically, the estimator

$$\tilde{\mathbf{b}} = (\tilde{\mathbf{X}}'\tilde{\mathbf{X}})^{-1}\tilde{\mathbf{X}}'\tilde{\mathbf{y}} = (\mathbf{X}'\mathbf{L}'\mathbf{LX})^{-1}\mathbf{X}'\mathbf{L}'\mathbf{Ly} = (\mathbf{X}'\mathbf{V}^{-1}\mathbf{X})^{-1}\mathbf{X}'\mathbf{V}^{-1}\mathbf{y} \qquad [8.3.5]$$

[8]We know that there exists a nonsingular matrix $\mathbf{P}$ such that $\mathbf{V} = \mathbf{PP}'$ and so $\mathbf{V}^{-1} = [\mathbf{P}']^{-1}\mathbf{P}^{-1}$. Take $\mathbf{L} = \mathbf{P}^{-1}$ to deduce [8.3.1].

is Gaussian with mean $\boldsymbol{\beta}$ and variance $\sigma^2(\tilde{\mathbf{X}}'\tilde{\mathbf{X}})^{-1} = \sigma^2(\mathbf{X}'\mathbf{V}^{-1}\mathbf{X})^{-1}$ conditional on $\mathbf{X}$ and is the minimum-variance unbiased estimator conditional on $\mathbf{X}$. The estimator [8.3.5] is known as the *generalized least squares* (*GLS*) estimator. Similarly,

$$\tilde{s}^2 = [1/(T - k)] \sum_{t=1}^{T} (\tilde{y}_t - \tilde{\mathbf{x}}_t'\tilde{\mathbf{b}})^2 \qquad [8.3.6]$$

has an exact $[\sigma^2/(T - k)]\cdot\chi^2(T - K)$ distribution under Assumption 8.5, while

$$(\mathbf{R}\tilde{\mathbf{b}} - \mathbf{r})'[\tilde{s}^2\mathbf{R}(\mathbf{X}'\mathbf{V}^{-1}\mathbf{X})^{-1}\mathbf{R}']^{-1}(\mathbf{R}\tilde{\mathbf{b}} - \mathbf{r})/m$$

has an exact $F(m, T - k)$ distribution under the null hypothesis $\mathbf{R}\boldsymbol{\beta} = \mathbf{r}$.

We now discuss several examples to make these ideas concrete.

Heteroskedasticity

A simple case to analyze is one for which the variance of u_t is presumed to be proportional to the square of one of the explanatory variables for that equation, say, x_{1t}^2:

$$E(\mathbf{u}\mathbf{u}'|\mathbf{X}) = \sigma^2 \begin{bmatrix} x_{11}^2 & 0 & \cdots & 0 \\ 0 & x_{12}^2 & \cdots & 0 \\ \vdots & \vdots & \cdots & \vdots \\ 0 & 0 & \cdots & x_{1T}^2 \end{bmatrix} = \sigma^2\mathbf{V}.$$

Then it is easy to see that

$$\mathbf{L} = \begin{bmatrix} 1/|x_{11}| & 0 & \cdots & 0 \\ 0 & 1/|x_{12}| & \cdots & 0 \\ \vdots & \vdots & \cdots & \vdots \\ 0 & 0 & \cdots & 1/|x_{1T}| \end{bmatrix}$$

satisfies conditions [8.3.1] and [8.3.2]. Hence, if we regress $y_t/|x_{1t}|$ on $\mathbf{x}_t/|x_{1t}|$, all the standard *OLS* output from the regression will be valid.

Autocorrelation

As a second example, consider

$$u_t = \rho u_{t-1} + \varepsilon_t, \qquad [8.3.7]$$

where $|\rho| < 1$ and ε_t is Gaussian white noise with variance σ^2. Then

$$E(\mathbf{u}\mathbf{u}'|\mathbf{X}) = \frac{\sigma^2}{1 - \rho^2} \begin{bmatrix} 1 & \rho & \rho^2 & \cdots & \rho^{T-1} \\ \rho & 1 & \rho & \cdots & \rho^{T-2} \\ \vdots & \vdots & \vdots & \cdots & \vdots \\ \rho^{T-1} & \rho^{T-2} & \rho^{T-3} & \cdots & 1 \end{bmatrix} = \sigma^2\mathbf{V}. \qquad [8.3.8]$$

Notice from expression [5.2.18] that the matrix

$$\mathbf{L} = \begin{bmatrix} \sqrt{1-\rho^2} & 0 & 0 & \cdots & 0 & 0 \\ -\rho & 1 & 0 & \cdots & 0 & 0 \\ 0 & -\rho & 1 & \cdots & 0 & 0 \\ \vdots & \vdots & \vdots & \cdots & \vdots & \vdots \\ 0 & 0 & 0 & \cdots & -\rho & 1 \end{bmatrix} \qquad [8.3.9]$$

satisfies [8.3.1]. The *GLS* estimates are found from an *OLS* regression of $\tilde{\mathbf{y}} = \mathbf{L}\mathbf{y}$ on $\tilde{\mathbf{X}} = \mathbf{L}\mathbf{X}$; that is, regress $y_1\sqrt{1-\rho^2}$ on $x_1\sqrt{1-\rho^2}$ and $y_t - \rho y_{t-1}$ on $\mathbf{x}_t - \rho\mathbf{x}_{t-1}$ for $t = 2, 3, \ldots, T$.

GLS *and Maximum Likelihood Estimation*

Assumption 8.5 asserts that $\mathbf{y}|\mathbf{X} \sim N(\mathbf{X}\boldsymbol{\beta}, \sigma^2\mathbf{V})$. Hence, the log of the likelihood of $\mathbf{y}$ conditioned on $\mathbf{X}$ is given by

$$(-T/2)\log(2\pi) - (1/2)\log|\sigma^2\mathbf{V}| - (1/2)(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})'(\sigma^2\mathbf{V})^{-1}(\mathbf{y} - \mathbf{X}\boldsymbol{\beta}). \qquad [8.3.10]$$

Notice that [8.3.1] can be used to write the last term in [8.3.10] as

$$\begin{aligned} -(1/2)(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})'(\sigma^2\mathbf{V})^{-1}(\mathbf{y} - \mathbf{X}\boldsymbol{\beta}) \\ &= -[1/(2\sigma^2)](\mathbf{y} - \mathbf{X}\boldsymbol{\beta})'(\mathbf{L}'\mathbf{L})(\mathbf{y} - \mathbf{X}\boldsymbol{\beta}) \\ &= -[1/(2\sigma^2)](\mathbf{L}\mathbf{y} - \mathbf{L}\mathbf{X}\boldsymbol{\beta})'(\mathbf{L}\mathbf{y} - \mathbf{L}\mathbf{X}\boldsymbol{\beta}) \\ &= -[1/(2\sigma^2)](\tilde{\mathbf{y}} - \tilde{\mathbf{X}}\boldsymbol{\beta})'(\tilde{\mathbf{y}} - \tilde{\mathbf{X}}\boldsymbol{\beta}). \end{aligned} \qquad [8.3.11]$$

Similarly, the middle term in [8.3.10] can be written as in [5.2.24]:

$$-(1/2)\log|\sigma^2\mathbf{V}| = -(T/2)\log(\sigma^2) + \log|\det(\mathbf{L})|, \qquad [8.3.12]$$

where $|\det(\mathbf{L})|$ denotes the absolute value of the determinant of $\mathbf{L}$. Substituting [8.3.11] and [8.3.12] into [8.3.10], the conditional log likelihood can be written as

$$\begin{aligned} -(T/2)\log(2\pi) - (T/2)\log(\sigma^2) + \log|\det(\mathbf{L})| \\ - [1/(2\sigma^2)](\tilde{\mathbf{y}} - \tilde{\mathbf{X}}\boldsymbol{\beta})'(\tilde{\mathbf{y}} - \tilde{\mathbf{X}}\boldsymbol{\beta}). \end{aligned} \qquad [8.3.13]$$

Thus, the log likelihood is maximized with respect to $\boldsymbol{\beta}$ by an *OLS* regression of $\tilde{\mathbf{y}}$ on $\tilde{\mathbf{X}}$,[9] meaning that the *GLS* estimate [8.3.5] is also the maximum likelihood estimate under Assumption 8.5.

The *GLS* estimate $\tilde{\mathbf{b}}$ is still likely to be reasonable even if the residuals $\mathbf{u}$ are non-Gaussian. Specifically, the residuals of the transformed regression [8.3.3] have mean $\mathbf{0}$ and variance $\sigma^2\mathbf{I}_T$, and so this regression satisfies the conditions of the Gauss-Markov theorem—even if the residuals are non-Gaussian, $\tilde{\mathbf{b}}$ will have minimum variance (conditional on $\mathbf{X}$) among the class of all unbiased estimators that are linear functions of $\mathbf{y}$. Hence, maximization of [8.3.13], or quasi-maximum likelihood estimation, may offer a useful estimating principle even for non-Gaussian $\mathbf{u}$.

GLS *When the Variance Matrix of Residuals Must Be Estimated from the Data*

Up to this point we have been assuming that the elements of $\mathbf{V}$ are known a priori. More commonly, $\mathbf{V}$ is posited to be of a particular form $\mathbf{V}(\boldsymbol{\theta})$, where $\boldsymbol{\theta}$ is a

[9]This assumes that the parameters of $\mathbf{L}$ do not involve $\boldsymbol{\beta}$, as is implied by Assumption 8.5.

vector of parameters that must be estimated from the data. For example, with first-order serial correlation of residuals as in [8.3.7], $\mathbf{V}$ is the matrix in [8.3.8] and $\boldsymbol{\theta}$ is the scalar ρ. As a second example, we might postulate that the variance of observation t depends on the explanatory variables according to

$$E(u_t^2|\mathbf{x}_t) = \sigma^2(1 + \alpha_1 x_{1t}^2 + \alpha_2 x_{2t}^2),$$

in which case $\boldsymbol{\theta} = (\alpha_1, \alpha_2)'$.

Our task is then to estimate $\boldsymbol{\theta}$ and $\boldsymbol{\beta}$ jointly from the data. One approach is to use as estimates the values of $\boldsymbol{\theta}$ and $\boldsymbol{\beta}$ that maximize [8.3.13]. Since one can always form [8.3.13] and maximize it numerically, this approach has the appeal of offering a single rule to follow whenever $E(\mathbf{uu}'|\mathbf{X})$ is not of the simple form $\sigma^2\mathbf{I}_T$. However, other, simpler estimators can also have desirable properties.

It often turns out to be the case that

$$\sqrt{T}(\mathbf{X}_T'[\mathbf{V}_T(\hat{\boldsymbol{\theta}}_T)]^{-1}\mathbf{X}_T)^{-1}(\mathbf{X}_T'[\mathbf{V}_T(\hat{\boldsymbol{\theta}}_T)]^{-1}\mathbf{y}_T)$$
$$\xrightarrow{p} \sqrt{T}(\mathbf{X}_T'[\mathbf{V}_T(\boldsymbol{\theta}_0)]^{-1}\mathbf{X}_T)^{-1}(\mathbf{X}_T'[\mathbf{V}_T(\boldsymbol{\theta}_0)]^{-1}\mathbf{y}_T),$$

where $\mathbf{V}_T(\boldsymbol{\theta}_0)$ denotes the true variance of errors and $\hat{\boldsymbol{\theta}}_T$ is any consistent estimate of $\boldsymbol{\theta}$. Moreover, a consistent estimate of $\boldsymbol{\theta}$ can often be obtained from a simple analysis of *OLS* residuals. Thus, an estimate coming from a few simple *OLS* and *GLS* regressions can have the same asymptotic distribution as the maximum likelihood estimator. Since regressions are much easier to implement than numerical maximization, the simpler estimates are often used.

Estimation with First-Order Autocorrelation of Regression Residuals and No Lagged Endogenous Variables

We illustrate these issues by considering a regression whose residuals follow the $AR(1)$ process [8.3.7]. For now we maintain the assumption that $\mathbf{u}|\mathbf{X}$ has mean zero and variance $\sigma^2\mathbf{V}(\rho)$, noting that this rules out lagged endogenous variables; that is, we assume that $\mathbf{x}_t$ is uncorrelated with u_{t-s}. The following subsection comments on the importance of this assumption. Recalling that the determinant of a lower triangular matrix is just the product of the terms on the principal diagonal, we see from [8.3.9] that $\det(\mathbf{L}) = \sqrt{1 - \rho^2}$. Thus, the log likelihood [8.3.13] for this case is

$$\begin{aligned} -(T/2)\log(2\pi) &- (T/2)\log(\sigma^2) + (1/2)\log(1 - \rho^2) \\ &- [(1 - \rho^2)/(2\sigma^2)](y_1 - \mathbf{x}_1'\boldsymbol{\beta})^2 \\ &- [1/(2\sigma^2)]\sum_{t=2}^{T}[(y_t - \mathbf{x}_t'\boldsymbol{\beta}) - \rho(y_{t-1} - \mathbf{x}_{t-1}'\boldsymbol{\beta})]^2. \end{aligned} \qquad [8.3.14]$$

One approach, then, is to maximize [8.3.14] numerically with respect to $\boldsymbol{\beta}$, ρ, and σ^2. The reader may recognize [8.3.14] as the exact log likelihood function for an $AR(1)$ process (equation [5.2.9]) with $(y_t - \mu)$ replaced by $(y_t - \mathbf{x}_t'\boldsymbol{\beta})$.

Just as in the $AR(1)$ case, simpler estimates (with the same asymptotic distribution) are obtained if we condition on the first observation, seeking to maximize

$$\begin{aligned} -[(T - 1)/2]\log(2\pi) &- [(T - 1)/2]\log(\sigma^2) \\ &- [1/(2\sigma^2)]\sum_{t=2}^{T}[(y_t - \mathbf{x}_t'\boldsymbol{\beta}) - \rho(y_{t-1} - \mathbf{x}_{t-1}'\boldsymbol{\beta})]^2. \end{aligned} \qquad [8.3.15]$$

If we knew the value of ρ, then the value of $\boldsymbol{\beta}$ that maximizes [8.3.15] could be found by an *OLS* regression of $(y_t - \rho y_{t-1})$ on $(\mathbf{x}_t - \rho\mathbf{x}_{t-1})$ for $t = 2, 3, \ldots,$

T (call this regression A). Conversely, if we knew the value of $\boldsymbol{\beta}$, then the value of ρ that maximizes [8.3.15] would be found by an *OLS* regression of $(y_t - \mathbf{x}_t'\boldsymbol{\beta})$ on $(y_{t-1} - \mathbf{x}_{t-1}'\boldsymbol{\beta})$ for $t = 2, 3, \ldots, T$ (call this regression B). We can thus start with an initial guess for ρ (often $\rho = 0$), and perform regression A to get an initial estimate of $\boldsymbol{\beta}$. For $\rho = 0$, this initial estimate of $\boldsymbol{\beta}$ would just be the *OLS* estimate $\mathbf{b}$. This estimate of $\boldsymbol{\beta}$ can be used in regression B to get an updated estimate of ρ, for example, by regressing the *OLS* residual $\hat{u}_t = y_t - \mathbf{x}_t'\mathbf{b}$ on its own lagged value. This new estimate of ρ can be used to repeat the two regressions. Zigzagging back and forth between A and B is known as the *iterated Cochrane-Orcutt* method and will converge to a local maximum of [8.3.15].

Alternatively, consider the estimate of ρ that results from the first iteration alone,

$$\hat{\rho} = \frac{(1/T)\sum_{t=1}^{T} \hat{u}_{t-1}\hat{u}_t}{(1/T)\sum_{t=1}^{T} \hat{u}_{t-1}^2}, \qquad [8.3.16]$$

where $\hat{u}_t = y_t - \mathbf{x}_t'\mathbf{b}$ and $\mathbf{b}$ is the *OLS* estimate of $\boldsymbol{\beta}$. To simplify expressions, we have renormalized the number of observations in the original sample to $T + 1$, denoted $y_0, y_1, \ldots, y_T$, so that T observations are used in the conditional maximum likelihood estimation. Notice that

$$\hat{u}_t = (y_t - \boldsymbol{\beta}'\mathbf{x}_t + \boldsymbol{\beta}'\mathbf{x}_t - \mathbf{b}'\mathbf{x}_t) = u_t + (\boldsymbol{\beta} - \mathbf{b})'\mathbf{x}_t,$$

allowing the numerator of [8.3.16] to be written

$$\begin{aligned}(1/T)\sum_{t=1}^{T} \hat{u}_t\hat{u}_{t-1} &= (1/T)\sum_{t=1}^{T} [u_t + (\boldsymbol{\beta} - \mathbf{b})'\mathbf{x}_t][u_{t-1} + (\boldsymbol{\beta} - \mathbf{b})'\mathbf{x}_{t-1}] \\ &= (1/T)\sum_{t=1}^{T} (u_t u_{t-1}) + (\boldsymbol{\beta} - \mathbf{b})'(1/T)\sum_{t=1}^{T} (u_t\mathbf{x}_{t-1} + u_{t-1}\mathbf{x}_t) \\ &\quad + (\boldsymbol{\beta} - \mathbf{b})'\left[(1/T)\sum_{t=1}^{T} \mathbf{x}_t\mathbf{x}_{t-1}'\right](\boldsymbol{\beta} - \mathbf{b}).\end{aligned} \qquad [8.3.17]$$

As long as $\mathbf{b}$ is a consistent estimate of $\boldsymbol{\beta}$ and boundedness conditions ensure that plims of $(1/T)\Sigma_{t=1}^{T} u_t\mathbf{x}_{t-1}$, $(1/T)\Sigma_{t=1}^{T} u_{t-1}\mathbf{x}_t$, and $(1/T)\Sigma_{t=1}^{T} \mathbf{x}_t\mathbf{x}_{t-1}'$ exist, then

$$\begin{aligned}(1/T)\sum_{t=1}^{T} \hat{u}_t\hat{u}_{t-1} &\xrightarrow{p} (1/T)\sum_{t=1}^{T} u_t u_{t-1} \\ &= (1/T)\sum_{t=1}^{T} (\varepsilon_t + \rho u_{t-1})u_{t-1} \\ &\xrightarrow{p} \rho\cdot\mathrm{Var}(u).\end{aligned} \qquad [8.3.18]$$

Similar analysis establishes that the denominator of [8.3.16] converges in probability to $\mathrm{Var}(u)$, so that $\hat{\rho} \xrightarrow{p} \rho$.

If u_t is uncorrelated with $\mathbf{x}_s$ for $s = t - 1$, t, and $t + 1$, one can make the stronger claim that an estimate of ρ based on an autoregression of the *OLS* residuals $\hat{u}_t$ (expression [8.3.16]) has the same asymptotic distribution as an estimate of ρ based on the true population residuals u_t. Specifically, if $\mathrm{plim}[(1/T)\Sigma_{t=1}^{T} u_t\mathbf{x}_{t-1}] =$

$\text{plim}[(1/T)\Sigma_{t=1}^{T} u_{t-1}\mathbf{x}_t] = \mathbf{0}$, then multiplying [8.3.17] by $\sqrt{T}$, we find

$$
\begin{aligned}
(1/\sqrt{T}) \sum_{t=1}^{T} \hat{u}_t\hat{u}_{t-1} \\
&= (1/\sqrt{T}) \sum_{t=1}^{T} (u_t u_{t-1}) + \sqrt{T}(\boldsymbol{\beta} - \mathbf{b})'(1/T) \sum_{t=1}^{T} (u_t \mathbf{x}_{t-1} + u_{t-1}\mathbf{x}_t) \\
&\quad + \sqrt{T}(\boldsymbol{\beta} - \mathbf{b})'\left[(1/T) \sum_{t=1}^{T} \mathbf{x}_t\mathbf{x}_{t-1}'\right](\boldsymbol{\beta} - \mathbf{b}) \\
&\xrightarrow{p} (1/\sqrt{T}) \sum_{t=1}^{T} (u_t u_{t-1}) + \sqrt{T}(\boldsymbol{\beta} - \mathbf{b})'\mathbf{0} \\
&\quad + \sqrt{T}(\boldsymbol{\beta} - \mathbf{b})'\,\text{plim}\left[(1/T) \sum_{t=1}^{T} \mathbf{x}_t\mathbf{x}_{t-1}'\right]\mathbf{0} \\
&= (1/\sqrt{T}) \sum_{t=1}^{T} (u_t u_{t-1}).
\end{aligned}
\tag{8.3.19}
$$

Hence,

$$
\sqrt{T}\left[\frac{(1/T) \sum_{t=1}^{T} \hat{u}_{t-1}\hat{u}_t}{(1/T) \sum_{t=1}^{T} \hat{u}_{t-1}^2}\right] \xrightarrow{p} \sqrt{T}\left[\frac{(1/T) \sum_{t=1}^{T} u_{t-1}u_t}{(1/T) \sum_{t=1}^{T} u_{t-1}^2}\right]. \tag{8.3.20}
$$

The *OLS* estimate of ρ based on the population residuals would have an asymptotic distribution given by [8.2.30]:

$$
\sqrt{T}\left[\frac{(1/T) \sum_{t=1}^{T} \hat{u}_{t-1}\hat{u}_t}{(1/T) \sum_{t=1}^{T} \hat{u}_{t-1}^2} - \rho\right] \xrightarrow{L} N(0, (1 - \rho^2)). \tag{8.3.21}
$$

Result [8.3.20] implies that an estimate of ρ has the same asymptotic distribution when based on any consistent estimate of $\boldsymbol{\beta}$. If the Cochrane-Orcutt iterations are stopped after just one evaluation of $\hat{\rho}$, the resulting estimate of ρ has the same asymptotic distribution as the estimate of ρ emerging from any subsequent step of the iteration.

The same also turns out to be true of the *GLS* estimate $\tilde{\mathbf{b}}$.

Proposition 8.4: *Suppose that Assumption 8.5(a) and (b) holds with* $\mathbf{V}$ *given by [8.3.8] and* $|\rho| < 1$. *Suppose in addition that* $(1/T)\Sigma_{t=1}^{T}\mathbf{x}_t u_s \xrightarrow{p} \mathbf{0}$ *for all s and that* $(1/T)\Sigma_{t=1}^{T}\mathbf{x}_t\mathbf{x}_t'$ *and* $(1/T)\Sigma_{t=1}^{T}\mathbf{x}_t\mathbf{x}_{t-1}'$ *have finite plims. Then the GLS estimate* $\tilde{\mathbf{b}}$ *constructed from* $\mathbf{V}(\hat{\rho})$ *for* $\hat{\rho}$ *given by [8.3.16] has the same asymptotic distribution as* $\tilde{\mathbf{b}}$ *constructed from* $\mathbf{V}(\rho)$ *for the true value of* ρ.

Serial Correlation with Lagged Endogenous Variables

An *endogenous variable* is a variable that is correlated with the regression error term u_t. Many of the preceding results about serially correlated errors no

longer hold if the regression contains lagged endogenous variables. For example, consider estimation of

$$y_t = \beta y_{t-1} + \gamma x_t + u_t, \qquad [8.3.22]$$

where u_t follows an $AR(1)$ process as in [8.3.7]. Since (1) u_t is correlated with u_{t-1} and (2) u_{t-1} is correlated with y_{t-1}, it follows that u_t is correlated with the explanatory variable y_{t-1}. Accordingly, it is not the case that $\text{plim}[(1/T)\Sigma_{t=1}^{T}\mathbf{x}_t u_t] = \mathbf{0}$, the key condition required for consistency of the *OLS* estimator $\mathbf{b}$. Hence, $\hat{\rho}$ in [8.3.16] is not a consistent estimate of ρ.

If one nevertheless iterates on the Cochrane-Orcutt procedure, then the algorithm will converge to a *local* maximum of [8.3.15]. However, the resulting *GLS* estimate $\tilde{\mathbf{b}}$ need not be a consistent estimate of $\boldsymbol{\beta}$. Notwithstanding, the *global* maximum of [8.3.15] should provide a consistent estimate of $\boldsymbol{\beta}$. By experimenting with start-up values for iterated Cochrane-Orcutt other than $\rho = 0$, one should find this global maximum.[10]

A simple estimate of ρ that is consistent in the presence of lagged endogenous variables was suggested by Durbin (1960). Multiplying [8.3.22] by $(1 - \rho L)$ gives

$$y_t = (\rho + \beta)y_{t-1} - \rho\beta y_{t-2} + \gamma x_t - \rho\gamma x_{t-1} + \varepsilon_t. \qquad [8.3.23]$$

This is a restricted version of the regression model

$$y_t = \alpha_1 y_{t-1} + \alpha_2 y_{t-2} + \alpha_3 x_t + \alpha_4 x_{t-1} + \varepsilon_t, \qquad [8.3.24]$$

where the four regression coefficients $(\alpha_1, \alpha_2, \alpha_3, \alpha_4)$ are restricted to be nonlinear functions of three underlying parameters (ρ, β, γ). Minimization of the sum of squared ε's in [8.3.23] is equivalent to maximum likelihood estimation conditioning on the first two observations. Moreover, the error term in equation [8.3.24] is uncorrelated with the explanatory variables, and so the α's can be estimated consistently by *OLS* estimation of [8.3.24]. Then $-\hat{\alpha}_4/\hat{\alpha}_3$ provides a consistent estimate of ρ despite the presence of lagged endogenous variables in [8.3.24].

Even if consistent estimates of ρ and $\boldsymbol{\beta}$ are obtained, Durbin (1970) emphasized that with lagged endogenous variables it will still not be the case that an estimate of ρ based on $(y_t - \mathbf{x}_t'\hat{\boldsymbol{\beta}})$ has the same asymptotic distribution as an estimate based on $(y_t - \mathbf{x}_t'\boldsymbol{\beta})$. To see this, note that if $\mathbf{x}_t$ contains lagged endogenous variables, then [8.3.19] would no longer be valid. If $\mathbf{x}_t$ includes y_{t-1}, for example, then $\mathbf{x}_t$ and u_{t-1} will be correlated and $\text{plim}[(1/T)\Sigma_{t=1}^{T}u_{t-1}\mathbf{x}_t] \neq \mathbf{0}$, as was assumed in arriving at [8.3.19]. Hence, [8.3.20] will not hold when $\mathbf{x}_t$ includes lagged endogenous variables. Again, an all-purpose procedure that will work is to maximize the log likelihood function [8.3.15] numerically.

Higher-Order Serial Correlation[11]

Consider next the case when the distribution of $\mathbf{u}|\mathbf{X}$ can be described by a pth-order autoregression,

$$u_t = \rho_1 u_{t-1} + \rho_2 u_{t-2} + \cdots + \rho_p u_{t-p} + \varepsilon_t.$$

[10]See Betancourt and Kelejian (1981).

[11]This discussion is based on Harvey (1981, pp. 204–6).

The log likelihood conditional on $\mathbf{X}$ for this case becomes

$$\begin{aligned}
&-(T/2)\log(2\pi) - (T/2)\log(\sigma^2) - (1/2)\log|\mathbf{V}_p| \\
&\quad - [1/(2\sigma^2)](\mathbf{y}_p - \mathbf{X}_p\boldsymbol{\beta})'\mathbf{V}_p^{-1}(\mathbf{y}_p - \mathbf{X}_p\boldsymbol{\beta}) \\
&\quad - [1/(2\sigma^2)]\sum_{t=p+1}^{T}\Big[(y_t - \mathbf{x}_t'\boldsymbol{\beta}) - \rho_1(y_{t-1} - \mathbf{x}_{t-1}'\boldsymbol{\beta}) \qquad [8.3.25] \\
&\quad - \rho_2(y_{t-2} - \mathbf{x}_{t-2}'\boldsymbol{\beta}) - \cdots - \rho_p(y_{t-p} - \mathbf{x}_{t-p}'\boldsymbol{\beta})\Big]^2,
\end{aligned}$$

where the $(p \times 1)$ vector $\mathbf{y}_p$ denotes the first p observations on y, $\mathbf{X}_p$ is the $(p \times k)$ matrix of explanatory variables associated with these first p observations, and $\sigma^2\mathbf{V}_p$ is the $(p \times p)$ variance-covariance matrix of $(\mathbf{y}_p|\mathbf{X}_p)$. The row i, column j element of $\sigma^2\mathbf{V}_p$ is given by $\gamma_{|i-j|}$ for γ_k the kth autocovariance of an $AR(p)$ process with autoregressive parameters $\rho_1, \rho_2, \ldots, \rho_p$ and innovation variance σ^2. Letting $\mathbf{L}_p$ denote a $(p \times p)$ matrix such that $\mathbf{L}_p'\mathbf{L}_p = \mathbf{V}_p^{-1}$, *GLS* can be obtained by regressing $\tilde{\mathbf{y}}_p = \mathbf{L}_p\mathbf{y}_p$ on $\tilde{\mathbf{X}}_p = \mathbf{L}_p\mathbf{X}_p$ and $\tilde{y}_t = y_t - \rho_1 y_{t-1} - \rho_2 y_{t-2} - \cdots - \rho_p y_{t-p}$ on $\tilde{\mathbf{x}}_t = \mathbf{x}_t - \rho_1\mathbf{x}_{t-1} - \rho_2\mathbf{x}_{t-2} - \cdots - \rho_p\mathbf{x}_{t-p}$ for $t = p+1, p+2, \ldots, T$. Equation [8.3.14] is a special case of [8.3.25] with $p = 1$, $\mathbf{V}_p = 1/(1-\rho^2)$, and $\mathbf{L}_p = \sqrt{1-\rho^2}$.

If we are willing to condition on the first p observations, the task is to choose $\boldsymbol{\beta}$ and $\rho_1, \rho_2, \ldots, \rho_p$ so as to minimize

$$\sum_{t=p+1}^{T}\Big[(y_t - \mathbf{x}_t'\boldsymbol{\beta}) - \rho_1(y_{t-1} - \mathbf{x}_{t-1}'\boldsymbol{\beta}) - \rho_2(y_{t-2} - \mathbf{x}_{t-2}'\boldsymbol{\beta}) - \cdots - \rho_p(y_{t-p} - \mathbf{x}_{t-p}'\boldsymbol{\beta})\Big]^2.$$

Again, in the absence of lagged endogenous variables we can iterate as in Cochrane-Orcutt, first taking the ρ_i's as given and regressing $\tilde{y}_t$ on $\tilde{\mathbf{x}}_t$, and then taking $\boldsymbol{\beta}$ as given and regressing $\hat{u}_t$ on $\hat{u}_{t-1}, \hat{u}_{t-2}, \ldots, \hat{u}_{t-p}$.

Any covariance-stationary process for the errors can always be approximated by a finite autoregression, provided that the order of the approximating autoregression (p) is sufficiently large. Amemiya (1973) demonstrated that by letting p go to infinity at a slower rate than the sample size T, this iterated *GLS* estimate will have the same asymptotic distribution as would the *GLS* estimate for the case when $\mathbf{V}$ is known. Alternatively, if theory implies an $ARMA(p, q)$ structure for the errors with p and q known, one can find exact or approximate maximum likelihood estimates by adapting the methods in Chapter 5, replacing μ in the expressions in Chapter 5 with $\mathbf{x}_t'\boldsymbol{\beta}$.

Further Remarks on Heteroskedasticity

Heteroskedasticity can arise from a variety of sources, and the solution depends on the nature of the problem identified. Using logs rather than levels of variables, allowing the explanatory variables to enter nonlinearly in the regression equation, or adding previously omitted explanatory variables to the regression may all be helpful. Judge, Griffiths, Hill, and Lee (1980) discussed a variety of solutions when the heteroskedasticity is thought to be related to the explanatory variables. In time series regressions, the explanatory variables themselves exhibit dynamic behavior, and such specifications then imply a dynamic structure for the conditional variance. An example of such a model is the autoregressive conditional heteroskedasticity specification of Engle (1982). Dynamic models of heteroskedasticity will be discussed in Chapter 21.

APPENDIX 8.A. *Proofs of Chapter 8 Propositions*

■ **Proof of Proposition 8.2.** The restricted estimate $\mathbf{b}^*$ that minimizes [8.1.2] subject to [8.1.27] can be calculated using the Lagrangean:

$$J = (1/2)\sum_{t=1}^{T}(y_t - \mathbf{x}_t'\boldsymbol{\beta})^2 + \boldsymbol{\lambda}'(\mathbf{R}\boldsymbol{\beta} - \mathbf{r}). \qquad [8.A.1]$$

Here $\boldsymbol{\lambda}$ denotes an $(m \times 1)$ vector of Lagrange multipliers; λ_i is associated with the constraint represented by the ith row of $\mathbf{R}\boldsymbol{\beta} = \mathbf{r}$. The term $\frac{1}{2}$ is a normalizing constant to simplify the expressions that follow. The constrained minimum is found by setting the derivative of [8.A.1] with respect to $\boldsymbol{\beta}$ equal to zero:[12]

$$\begin{aligned}\frac{\partial J}{\partial \boldsymbol{\beta}'} &= (1/2)\sum_{t=1}^{T} 2(y_t - \mathbf{x}_t'\boldsymbol{\beta})\frac{\partial(y_t - \mathbf{x}_t'\boldsymbol{\beta})}{\partial \boldsymbol{\beta}'} + \boldsymbol{\lambda}'\mathbf{R}\\ &= -\sum_{t=1}^{T}(y_t - \boldsymbol{\beta}'\mathbf{x}_t)\mathbf{x}_t' + \boldsymbol{\lambda}'\mathbf{R} = \mathbf{0}',\end{aligned}$$

or

$$\mathbf{b}^{*\prime}\sum_{t=1}^{T}\mathbf{x}_t\mathbf{x}_t' = \sum_{t=1}^{T}y_t\mathbf{x}_t' - \boldsymbol{\lambda}'\mathbf{R}.$$

Taking transposes,

$$\begin{aligned}\left[\sum_{t=1}^{T}\mathbf{x}_t\mathbf{x}_t'\right]\mathbf{b}^* &= \sum_{t=1}^{T}\mathbf{x}_t y_t - \mathbf{R}'\boldsymbol{\lambda}\\ \mathbf{b}^* &= \left[\sum_{t=1}^{T}\mathbf{x}_t\mathbf{x}_t'\right]^{-1}\left[\sum_{t=1}^{T}\mathbf{x}_t y_t\right] - \left[\sum_{t=1}^{T}\mathbf{x}_t\mathbf{x}_t'\right]^{-1}\mathbf{R}'\boldsymbol{\lambda} \qquad [8.A.2]\\ &= \mathbf{b} - (\mathbf{X}'\mathbf{X})^{-1}\mathbf{R}'\boldsymbol{\lambda},\end{aligned}$$

where $\mathbf{b}$ denotes the unrestricted *OLS* estimate. Premultiplying [8.A.2] by $\mathbf{R}$ (and recalling that $\mathbf{b}^*$ satisfies $\mathbf{R}\mathbf{b}^* = \mathbf{r}$),

$$\mathbf{R}\mathbf{b} - \mathbf{r} = \mathbf{R}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{R}'\boldsymbol{\lambda}$$

or

$$\boldsymbol{\lambda} = [\mathbf{R}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{R}']^{-1}(\mathbf{R}\mathbf{b} - \mathbf{r}). \qquad [8.A.3]$$

Substituting [8.A.3] into [8.A.2],

$$\mathbf{b} - \mathbf{b}^* = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{R}'[\mathbf{R}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{R}']^{-1}(\mathbf{R}\mathbf{b} - \mathbf{r}). \qquad [8.A.4]$$

Notice from [8.A.4] that

$$\begin{aligned}(\mathbf{b} - \mathbf{b}^*)'(\mathbf{X}'\mathbf{X})(\mathbf{b} - \mathbf{b}^*) &= \{(\mathbf{R}\mathbf{b} - \mathbf{r})'[\mathbf{R}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{R}']^{-1}\mathbf{R}(\mathbf{X}'\mathbf{X})^{-1}\}(\mathbf{X}'\mathbf{X})\\ &\quad\times\{(\mathbf{X}'\mathbf{X})^{-1}\mathbf{R}'[\mathbf{R}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{R}']^{-1}(\mathbf{R}\mathbf{b} - \mathbf{r})\}\\ &= (\mathbf{R}\mathbf{b} - \mathbf{r})'[\mathbf{R}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{R}']^{-1}[\mathbf{R}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{R}'] \qquad [8.A.5]\\ &\quad\times[\mathbf{R}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{R}']^{-1}(\mathbf{R}\mathbf{b} - \mathbf{r})\\ &= (\mathbf{R}\mathbf{b} - \mathbf{r})'[\mathbf{R}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{R}']^{-1}(\mathbf{R}\mathbf{b} - \mathbf{r}).\end{aligned}$$

Thus, the magnitude in [8.1.32] is numerically identical to

$$F = \frac{(\mathbf{b} - \mathbf{b}^*)'\mathbf{X}'\mathbf{X}(\mathbf{b} - \mathbf{b}^*)/m}{s^2} = \frac{(\mathbf{b} - \mathbf{b}^*)'\mathbf{X}'\mathbf{X}(\mathbf{b} - \mathbf{b}^*)/m}{RSS_1/(T - k)}.$$

Comparing this with [8.1.37], we will have completed the demonstration of the equivalence of [8.1.32] with [8.1.37] if it is the case that

$$RSS_0 - RSS_1 = (\mathbf{b} - \mathbf{b}^*)'(\mathbf{X}'\mathbf{X})(\mathbf{b} - \mathbf{b}^*). \qquad [8.A.6]$$

[12]We have used the fact that $\partial\mathbf{x}_t'\boldsymbol{\beta}/\partial\boldsymbol{\beta}' = \mathbf{x}_t'$. See the Mathematical Review (Appendix A) at the end of the book on the use of derivatives with respect to vectors.

Now, notice that

$$\begin{aligned} RSS_0 &= (\mathbf{y} - \mathbf{X}\mathbf{b}^*)'(\mathbf{y} - \mathbf{X}\mathbf{b}^*) \\ &= (\mathbf{y} - \mathbf{X}\mathbf{b} + \mathbf{X}\mathbf{b} - \mathbf{X}\mathbf{b}^*)'(\mathbf{y} - \mathbf{X}\mathbf{b} + \mathbf{X}\mathbf{b} - \mathbf{X}\mathbf{b}^*) \\ &= (\mathbf{y} - \mathbf{X}\mathbf{b})'(\mathbf{y} - \mathbf{X}\mathbf{b}) + (\mathbf{b} - \mathbf{b}^*)'\mathbf{X}'\mathbf{X}(\mathbf{b} - \mathbf{b}^*), \end{aligned} \qquad [8.A.7]$$

where the cross-product term has vanished, since $(\mathbf{y} - \mathbf{X}\mathbf{b})'\mathbf{X} = \mathbf{0}$ by the least squares property [8.1.10]. Equation [8.A.7] states that

$$RSS_0 = RSS_1 + (\mathbf{b} - \mathbf{b}^*)'\mathbf{X}'\mathbf{X}(\mathbf{b} - \mathbf{b}^*), \qquad [8.A.8]$$

confirming [8.A.6]. ■

■ **Proof of Proposition 8.3.** Assumption 8.6(e) guarantees that $\hat{\mathbf{Q}}_T \xrightarrow{p} \mathbf{Q}$, so the issue is whether $\hat{\mathbf{\Omega}}_T$ gives a consistent estimate of $\mathbf{\Omega}$. Define $\mathbf{\Omega}_T^* \equiv (1/T)\Sigma_{t=1}^T u_t^2\mathbf{x}_t\mathbf{x}_t'$, noting that $\mathbf{\Omega}_T^*$ converges in probability to $\mathbf{\Omega}$ by Assumption 8.6(c). Thus, if we can show that $\hat{\mathbf{\Omega}}_T - \mathbf{\Omega}_T^* \xrightarrow{p} \mathbf{0}$, then $\hat{\mathbf{\Omega}}_T \xrightarrow{p} \mathbf{\Omega}$. Now,

$$\hat{\mathbf{\Omega}}_T - \mathbf{\Omega}_T^* = (1/T)\sum_{t=1}^T (\hat{u}_t^2 - u_t^2)\mathbf{x}_t\mathbf{x}_t'. \qquad [8.A.9]$$

But

$$\begin{aligned} (\hat{u}_t^2 - u_t^2) &= (\hat{u}_t + u_t)(\hat{u}_t - u_t) \\ &= [(y_t - \mathbf{b}_T'\mathbf{x}_t) + (y_t - \boldsymbol{\beta}'\mathbf{x}_t)][(y_t - \mathbf{b}_T'\mathbf{x}_t) - (y_t - \boldsymbol{\beta}'\mathbf{x}_t)] \\ &= [2(y_t - \boldsymbol{\beta}'\mathbf{x}_t) - (\mathbf{b}_T - \boldsymbol{\beta})'\mathbf{x}_t][-(\mathbf{b}_T - \boldsymbol{\beta})'\mathbf{x}_t] \\ &= -2u_t(\mathbf{b}_T - \boldsymbol{\beta})'\mathbf{x}_t + [(\mathbf{b}_T - \boldsymbol{\beta})'\mathbf{x}_t]^2, \end{aligned}$$

allowing [8.A.9] to be written as

$$\hat{\mathbf{\Omega}}_T - \mathbf{\Omega}_T^* = (-2/T)\sum_{t=1}^T u_t(\mathbf{b}_T - \boldsymbol{\beta})'\mathbf{x}_t(\mathbf{x}_t\mathbf{x}_t') + (1/T)\sum_{t=1}^T [(\mathbf{b}_T - \boldsymbol{\beta})'\mathbf{x}_t]^2(\mathbf{x}_t\mathbf{x}_t'). \qquad [8.A.10]$$

The first term in [8.A.10] can be written

$$(-2/T)\sum_{t=1}^T u_t(\mathbf{b}_T - \boldsymbol{\beta})'\mathbf{x}_t(\mathbf{x}_t\mathbf{x}_t') = -2\sum_{i=1}^k (b_{iT} - \beta_i)\left[(1/T)\sum_{t=1}^T u_t x_{it}(\mathbf{x}_t\mathbf{x}_t')\right]. \qquad [8.A.11]$$

The second term in [8.A.11] has a finite plim by Assumption 8.6(e), and $(b_{iT} - \beta_i) \xrightarrow{p} 0$ for each i. Hence, the probability limit of [8.A.11] is zero.

Turning next to the second term in [8.A.10],

$$(1/T)\sum_{t=1}^T [(\mathbf{b}_T - \boldsymbol{\beta})'\mathbf{x}_t]^2(\mathbf{x}_t\mathbf{x}_t') = \sum_{i=1}^k\sum_{j=1}^k (b_{iT} - \beta_i)(b_{jT} - \beta_j)\left[(1/T)\sum_{t=1}^T x_{it}x_{jt}(\mathbf{x}_t\mathbf{x}_t')\right],$$

which again has plim zero. Hence, from [8.A.10],

$$\hat{\mathbf{\Omega}}_T - \mathbf{\Omega}_T^* \xrightarrow{p} \mathbf{0}. \quad ■$$

■ **Proof of Proposition 8.4.** Recall from [8.2.6] that

$$\begin{aligned} \sqrt{T}(\tilde{\mathbf{b}}_T - \boldsymbol{\beta}) &= \left[(1/T)\sum_{t=1}^T \tilde{\mathbf{x}}_t\tilde{\mathbf{x}}_t'\right]^{-1}\left[(1/\sqrt{T})\sum_{t=1}^T \tilde{\mathbf{x}}_t\tilde{u}_t\right] \\ &= \left[(1/T)\sum_{t=1}^T (\mathbf{x}_t - \hat{\rho}\mathbf{x}_{t-1})(\mathbf{x}_t - \hat{\rho}\mathbf{x}_{t-1})'\right]^{-1} \\ &\quad \times \left[(1/\sqrt{T})\sum_{t=1}^T (\mathbf{x}_t - \hat{\rho}\mathbf{x}_{t-1})(u_t - \hat{\rho}u_{t-1})\right]. \end{aligned} \qquad [8.A.12]$$

We will now show that $[(1/T)\Sigma_{t=1}^T(\mathbf{x}_t - \hat{\rho}\mathbf{x}_{t-1})(\mathbf{x}_t - \hat{\rho}\mathbf{x}_{t-1})']$ has the same plim as $[(1/T)\Sigma_{t=1}^T(\mathbf{x}_t - \rho\mathbf{x}_{t-1})(\mathbf{x}_t - \rho\mathbf{x}_{t-1})']$ and that $[(1/\sqrt{T})\Sigma_{t=1}^T(\mathbf{x}_t - \hat{\rho}\mathbf{x}_{t-1})(u_t - \hat{\rho}u_{t-1})]$ has the same asymptotic distribution as $[(1/\sqrt{T})\Sigma_{t=1}^T(\mathbf{x}_t - \rho\mathbf{x}_{t-1})(u_t - \rho u_{t-1})]$.

Consider the first term in [8.A.12]:

$$
\begin{aligned}
(1/T)\sum_{t=1}^{T}(\mathbf{x}_t - \hat{\rho}\mathbf{x}_{t-1})(\mathbf{x}_t - \hat{\rho}\mathbf{x}_{t-1})' &= (1/T)\sum_{t=1}^{T}[\mathbf{x}_t - \rho\mathbf{x}_{t-1} + (\rho - \hat{\rho})\mathbf{x}_{t-1}][\mathbf{x}_t - \rho\mathbf{x}_{t-1} + (\rho - \hat{\rho})\mathbf{x}_{t-1}]' \\
&= (1/T)\sum_{t=1}^{T}(\mathbf{x}_t - \rho\mathbf{x}_{t-1})(\mathbf{x}_t - \rho\mathbf{x}_{t-1})' \\
&\quad + (\rho - \hat{\rho})\cdot(1/T)\sum_{t=1}^{T}(\mathbf{x}_t - \rho\mathbf{x}_{t-1})\mathbf{x}_{t-1}' \qquad [8.A.13] \\
&\quad + (\rho - \hat{\rho})\cdot(1/T)\sum_{t=1}^{T}\mathbf{x}_{t-1}(\mathbf{x}_t - \rho\mathbf{x}_{t-1})' \\
&\quad + (\rho - \hat{\rho})^2\cdot(1/T)\sum_{t=1}^{T}\mathbf{x}_{t-1}\mathbf{x}_{t-1}'.
\end{aligned}
$$

But $(\rho - \hat{\rho}) \xrightarrow{p} 0$, and the plims of $(1/T)\Sigma_{t=1}^{T}\mathbf{x}_{t-1}\mathbf{x}_{t-1}'$ and $(1/T)\Sigma_{t=1}^{T}\mathbf{x}_t\mathbf{x}_{t-1}'$ are assumed to exist. Hence [8.A.13] has the same plim as $(1/T)\Sigma_{t=1}^{T}(\mathbf{x}_t - \rho\mathbf{x}_{t-1})(\mathbf{x}_t - \rho\mathbf{x}_{t-1})'$.

Consider next the second term in [8.A.12]:

$$
\begin{aligned}
(1/\sqrt{T})\sum_{t=1}^{T}(\mathbf{x}_t - \hat{\rho}\mathbf{x}_{t-1})(u_t - \hat{\rho}u_{t-1}) &= (1/\sqrt{T})\sum_{t=1}^{T}[\mathbf{x}_t - \rho\mathbf{x}_{t-1} + (\rho - \hat{\rho})\mathbf{x}_{t-1}][u_t - \rho u_{t-1} + (\rho - \hat{\rho})u_{t-1}] \qquad [8.A.14] \\
&= (1/\sqrt{T})\sum_{t=1}^{T}(\mathbf{x}_t - \rho\mathbf{x}_{t-1})(u_t - \rho u_{t-1}) \\
&\quad + \sqrt{T}(\rho - \hat{\rho})\cdot\left[(1/T)\sum_{t=1}^{T}\mathbf{x}_{t-1}(u_t - \rho u_{t-1})\right] \\
&\quad + \sqrt{T}(\rho - \hat{\rho})\cdot\left[(1/T)\sum_{t=1}^{T}(\mathbf{x}_t - \rho\mathbf{x}_{t-1})u_{t-1}\right] \\
&\quad + \sqrt{T}(\rho - \hat{\rho})^2\cdot\left[(1/T)\sum_{t=1}^{T}\mathbf{x}_{t-1}u_{t-1}\right].
\end{aligned}
$$

But [8.3.21] established that $\sqrt{T}(\rho - \hat{\rho})$ converges in distribution to a stable random variable. Since $\text{plim}[(1/T)\Sigma_{t=1}^{T}\mathbf{x}_t u_s] = \mathbf{0}$, the last three terms in [8.A.14] vanish asymptotically. Hence,

$$(1/\sqrt{T})\sum_{t=1}^{T}(\mathbf{x}_t - \hat{\rho}\mathbf{x}_{t-1})(u_t - \hat{\rho}u_{t-1}) \xrightarrow{p} (1/\sqrt{T})\sum_{t=1}^{T}(\mathbf{x}_t - \rho\mathbf{x}_{t-1})(u_t - \rho u_{t-1}),$$

which was to be shown.

Chapter 8 Exercises

8.1. Show that the uncentered R_u^2 [8.1.13] can equivalently be written as

$$R_u^2 = 1 - \left[\left(\sum_{t=1}^{T}\hat{u}_t^2\right) \div \left(\sum_{t=1}^{T}y_t^2\right)\right]$$

for $\hat{u}_t$ the *OLS* sample residual [8.1.4]. Show that the centered R_c^2 can be written as

$$R_c^2 = 1 - \left[\left(\sum_{t=1}^{T}\hat{u}_t^2\right) \div \left(\sum_{t=1}^{T}(y_t - \bar{y})^2\right)\right].$$

8.2. Consider a null hypothesis H_0 involving $m = 2$ linear restrictions on $\boldsymbol{\beta}$. How large a sample size T is needed before the 5% critical value based on the Wald form of the *OLS* F test of H_0 is within 1% of the critical value of the Wald form of the *OLS* χ^2 test of H_0?

8.3. Derive result [8.2.28].

8.4. Consider a covariance-stationary process given by

$$y_t = \mu + \sum_{j=0}^{\infty} \psi_j \varepsilon_{t-j},$$

where $\{\varepsilon_t\}$ is an i.i.d. sequence with mean zero, variance σ^2, and finite fourth moment and where $\sum_{j=0}^{\infty} |\psi_j| < \infty$. Consider estimating a pth-order autoregression by *OLS*:

$$y_t = c + \phi_1 y_{t-1} + \phi_2 y_{t-2} + \cdots + \phi_p y_{t-p} + u_t.$$

Show that the *OLS* coefficients give consistent estimates of the population parameters that characterize the linear projection of y_t on a constant and p of its lags—that is, the coefficients give consistent estimates of the parameters $c, \phi_1, \ldots, \phi_p$ defined by

$$\hat{E}(y_t | y_{t-1}, y_{t-2}, \ldots, y_{t-p}) = c + \phi_1 y_{t-1} + \phi_2 y_{t-2} + \cdots + \phi_p y_{t-p}$$

(HINT: Recall that $c, \phi_1, \ldots, \phi_p$ are characterized by equation [4.3.6]).

Chapter 8 References

Amemiya, Takeshi. 1973. "Generalized Least Squares with an Estimated Autocovariance Matrix." *Econometrica* 41:723–32.

Anderson, T. W. 1971. *The Statistical Analysis of Time Series*. New York: Wiley.

Betancourt, Roger, and Harry Kelejian. 1981. "Lagged Endogenous Variables and the Cochrane-Orcutt Procedure." *Econometrica* 49:1073–78.

Brillinger, David R. 1981. *Time Series: Data Anaysis and Theory*, expanded ed. San Francisco: Holden-Day.

Durbin, James. 1960. "Estimation of Parameters in Time-Series Regression Models." *Journal of the Royal Statistical Society* Series B, 22:139–53.

———. 1970. "Testing for Serial Correlation in Least-Squares Regression When Some of the Regressors Are Lagged Dependent Variables." *Econometrica* 38:410–21.

Eicker, F. 1967. "Limit Theorems for Regressions with Unequal and Dependent Errors." *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability*, Vol. 1, pp. 59–62. Berkeley: University of California Press.

Engle, Robert F. 1982. "Autoregressive Conditional Heteroscedasticity with Estimates of the Variance of United Kingdom Inflation." *Econometrica* 50:987–1007.

Evans, G. B. A., and N. E. Savin. 1981. "Testing for Unit Roots: 1. *Econometrica* 49:753–79.

Flavin, Marjorie A. 1983. "Excess Volatility in the Financial Markets: A Reassessment of the Empirical Evidence." *Journal of Political Economy* 91:929–56.

Gregory, Allan W., and Michael R. Veall. 1985. "Formulating Wald Tests of Nonlinear Restrictions." *Econometrica* 53:1465–68.

Hansen, Lars P. 1982. "Large Sample Properties of Generalized Method of Moments Estimators." *Econometrica* 50:1029–54.

Harvey, A. C. 1981. *The Econometric Analysis of Time Series*. New York: Wiley.

Hausman, Jerry A., and William E. Taylor. 1983. "Identification in Linear Simultaneous Equations Models with Covariance Restrictions: An Instrumental Variables Interpretation." *Econometrica* 51:1527–49.

Imhof, J. P. 1961. "Computing the Distribution of Quadratic Forms in Normal Variables." *Biometrika* 48:419–26.

Judge, George G., William E. Griffiths, R. Carter Hill, and Tsoung-Chao Lee. 1980. *The Theory and Practice of Econometrics*. New York: Wiley.

Kinderman, A. J., and J. G. Ramage. 1976. "Computer Generation of Normal Random Variables." *Journal of the American Statistical Association* 71:893–96.

Lafontaine, Francine, and Kenneth J. White. 1986. "Obtaining Any Wald Statistic You Want." *Economics Letters* 21:35–40.

Maddala, G. S. 1977. *Econometrics*. New York: McGraw-Hill.

Newey, Whitney K., and Kenneth D. West. 1987. "A Simple Positive Semi-Definite, Heteroskedasticity and Autocorrelation Consistent Covariance Matrix." *Econometrica* 55:703–8.

Nicholls, D. F., and A. R. Pagan. 1983. "Heteroscedasticity in Models with Lagged Dependent Variables." *Econometrica* 51:1233–42.

O'Nan, Michael. 1976. *Linear Algebra*, 2d ed. New York: Harcourt Brace Jovanovich.

Phillips, P. C. B., and Joon Y. Park. 1988. "On the Formulation of Wald Tests of Nonlinear Restrictions." *Econometrica* 56:1065–83.

Rao, C. Radhakrishna. 1973. *Linear Statistical Inference and Its Applications*, 2d ed. New York: Wiley.

Theil, Henri. 1971. *Principles of Econometrics*. New York: Wiley.

White, Halbert. 1980. "A Heteroskedasticity-Consistent Covariance Matrix Estimator and a Direct Test for Heteroskedasticity." *Econometrica* 48:817–38.

———. 1984. *Asymptotic Theory for Econometricians*. Orlando, Fla.: Academic Press.

9 Linear Systems of Simultaneous Equations

The previous chapter described a number of possible departures from the ideal regression model arising from errors that are non-Gaussian, heteroskedastic, or autocorrelated. We saw that while these factors can make a difference for the small-sample validity of t and F tests, under any of Assumptions 8.1 through 8.6, the *OLS* estimator $\mathbf{b}_T$ is either unbiased or consistent. This is because all these cases retained the crucial assumption that u_t, the error term for observation t, is uncorrelated with $\mathbf{x}_t$, the explanatory variables for that observation. Unfortunately, this critical assumption is unlikely to be satisfied in many important applications.

Section 9.1 discusses why this assumption often fails to hold, by examining a concrete example of *simultaneous equations bias*. Subsequent sections discuss a variety of techniques for dealing with this problem. These results will be used in the structural interpretation of vector autoregressions in Chapter 11 and for understanding generalized method of moments estimation in Chapter 14.

9.1. *Simultaneous Equations Bias*

To illustrate the difficulties with endogenous regressors, consider an investigation of the public's demand for oranges. Let p_t denote the log of the price of oranges in a particular year and q_t^d the log of the quantity the public is willing to buy. To keep the example very simple, suppose that price and quantity are covariance-stationary and that each is measured as deviations from its population mean. The demand curve is presumed to take the form

$$q_t^d = \beta p_t + \varepsilon_t^d, \qquad [9.1.1]$$

with $\beta < 0$; a higher price reduces the quantity that the public is willing to buy. Here ε_t^d represents factors that influence demand other than price. These are assumed to be independent and identically distributed with mean zero and variance σ_d^2.

The price also influences the supply of oranges brought to the market,

$$q_t^s = \gamma p_t + \varepsilon_t^s, \qquad [9.1.2]$$

where $\gamma > 0$ and ε_t^s represents factors that influence supply other than price. These omitted factors are again assumed to be i.i.d. with mean zero and variance σ_s^2, with the supply disturbance ε_t^s uncorrelated with the demand disturbance ε_t^d.

Equation [9.1.1] describes the behavior of buyers of oranges, and equation [9.1.2] describes the behavior of sellers. Market equilibrium requires $q_t^d = q_t^s$, or

$$\beta p_t + \varepsilon_t^d = \gamma p_t + \varepsilon_t^s.$$

Rearranging,

$$p_t = \frac{\varepsilon_t^d - \varepsilon_t^s}{\gamma - \beta}. \qquad [9.1.3]$$

Substituting this back into [9.1.2],

$$q_t = \gamma \frac{\varepsilon_t^d - \varepsilon_t^s}{\gamma - \beta} + \varepsilon_t^s = \frac{\gamma}{\gamma - \beta}\varepsilon_t^d - \frac{\beta}{\gamma - \beta}\varepsilon_t^s. \qquad [9.1.4]$$

Consider the consequences of trying to estimate [9.1.1] by *OLS*. A regression of quantity on price will produce the estimate

$$b_T = \frac{(1/T)\sum_{t=1}^{T} p_t q_t}{(1/T)\sum_{t=1}^{T} p_t^2}. \qquad [9.1.5]$$

Substituting [9.1.3] and [9.1.4] into the numerator in [9.1.5] results in

$$\begin{aligned}\frac{1}{T}\sum_{t=1}^{T} p_t q_t &= \frac{1}{T}\sum_{t=1}^{T}\left[\frac{1}{\gamma - \beta}\varepsilon_t^d - \frac{1}{\gamma - \beta}\varepsilon_t^s\right]\left[\frac{\gamma}{\gamma - \beta}\varepsilon_t^d - \frac{\beta}{\gamma - \beta}\varepsilon_t^s\right] \\ &= \frac{1}{T}\sum_{t=1}^{T}\left[\frac{\gamma}{(\gamma - \beta)^2}(\varepsilon_t^d)^2 + \frac{\beta}{(\gamma - \beta)^2}(\varepsilon_t^s)^2 - \frac{\gamma + \beta}{(\gamma - \beta)^2}\varepsilon_t^d\varepsilon_t^s\right] \\ &\xrightarrow{p} \frac{\gamma\sigma_d^2 + \beta\sigma_s^2}{(\gamma - \beta)^2}.\end{aligned}$$

Similarly, for the denominator,

$$\frac{1}{T}\sum_{t=1}^{T} p_t^2 = \frac{1}{T}\sum_{t=1}^{T}\left[\frac{1}{\gamma - \beta}\varepsilon_t^d - \frac{1}{\gamma - \beta}\varepsilon_t^s\right]^2 \xrightarrow{p} \frac{\sigma_d^2 + \sigma_s^2}{(\gamma - \beta)^2}.$$

Hence,

$$b_T \xrightarrow{p} \left[\frac{\sigma_d^2 + \sigma_s^2}{(\gamma - \beta)^2}\right]^{-1}\left[\frac{\gamma\sigma_d^2 + \beta\sigma_s^2}{(\gamma - \beta)^2}\right] = \frac{\gamma\sigma_d^2 + \beta\sigma_s^2}{\sigma_d^2 + \sigma_s^2}. \qquad [9.1.6]$$

OLS regression thus gives not the demand elasticity β but rather an average of β and the supply elasticity γ, with weights depending on the sizes of the variances σ_d^2 and σ_s^2. If the error in the demand curve is negligible ($\sigma_d^2 \to 0$) or if the error term in the supply curve has a big enough variance ($\sigma_s^2 \to \infty$), then [9.1.6] indicates that *OLS* would give a consistent estimate of the demand elasticity β. On the other hand, if $\sigma_d^2 \to \infty$ or $\sigma_s^2 \to 0$, then *OLS* gives a consistent estimate of the supply elasticity γ. In the cases in between, one economist might believe the regression was estimating the demand curve [9.1.1] and a second economist might perform the same regression calling it the supply curve [9.1.2]. The actual *OLS* estimates would represent a mixture of both. This phenomenon is known as *simultaneous equations bias*.

Figure 9.1 depicts the problem graphically.[1] At any date in the sample, there is some demand curve (determined by the value of ε_t^d) and a supply curve (determined by ε_t^s), with the observation on (p_t, q_t) given by the intersection of these two curves. For example, date 1 may have been associated with a small negative shock to demand, producing the curve D_1, and a large positive shock to supply, producing S_1. The date 1 observation will then be (p_1, q_1). Date 2 might have seen

[1]Economists usually display these figures with the axes reversed from those displayed in Figure 9.1.

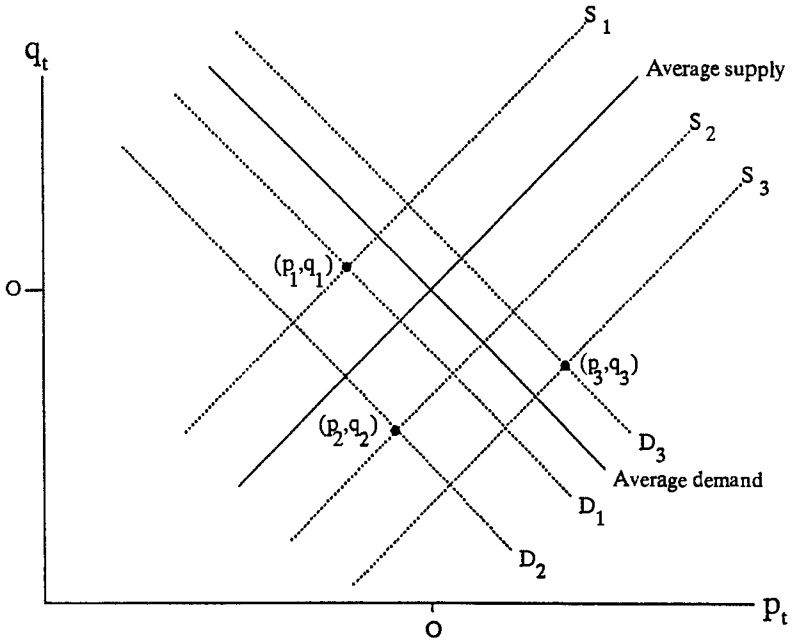


FIGURE 9.1 Observations on price and quantity implied by disturbances to both supply functions and demand functions.

a bigger negative shock to demand and a negative shock to supply, while date 3 as drawn reflects a modest positive shock to demand and a large negative shock to supply. *OLS* tries to fit a line through the scatter of points $\{p_t, q_t\}_{t=1}^{T}$.

If the shocks are known to be due to the supply curve and not the demand curve, then the scatter of points will trace out the demand curve, as in Figure 9.2. If the shocks are due to the demand curve rather than the supply curve, the scatter will trace out the supply curve, as in Figure 9.3.

The problem of simultaneous equations bias is extremely widespread in the social sciences. It is rare that the relation that we would like to estimate is the only possible reason why there might be a correlation among a group of variables.

Consistent Estimation of the Demand Elasticity

The above analysis suggests that consistent estimates of the demand elasticity might be obtained if we could find a variable that shifts the supply curve but not the demand curve. For example, let w_t represent the number of days of below-freezing temperatures in Florida during year t. Recalling that the supply disturbance ε_t^s was defined as factors influencing supply other than price, w_t seems likely to be an important component of ε_t^s. Define h to be the coefficient from a linear projection of ε_t^s on w_t, and write

$$\varepsilon_t^s = hw_t + u_t^s. \qquad [9.1.7]$$

Thus, u_t^s is uncorrelated with w_t, by the definition of h. Although Florida weather is likely to influence the supply of oranges, it is natural to assume that weather

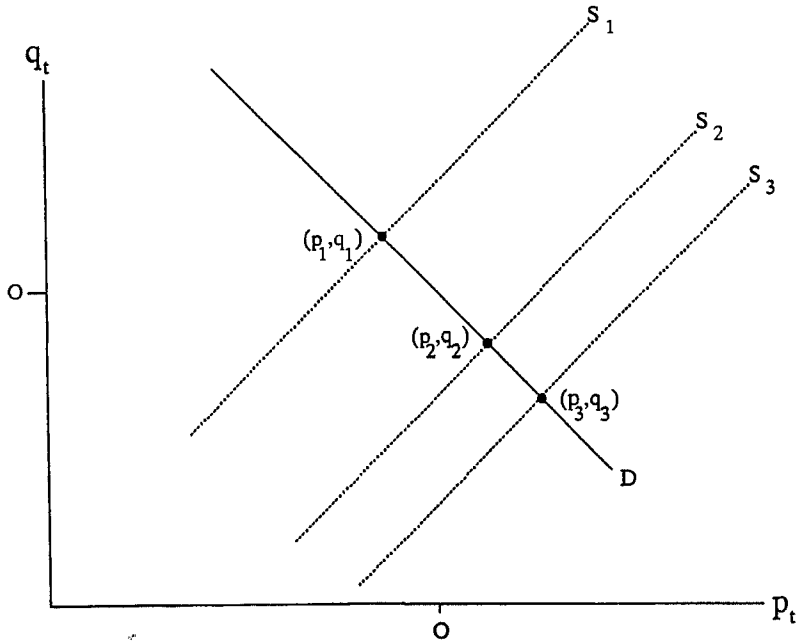


FIGURE 9.2 Observations on price and quantity implied by disturbances to supply function only.

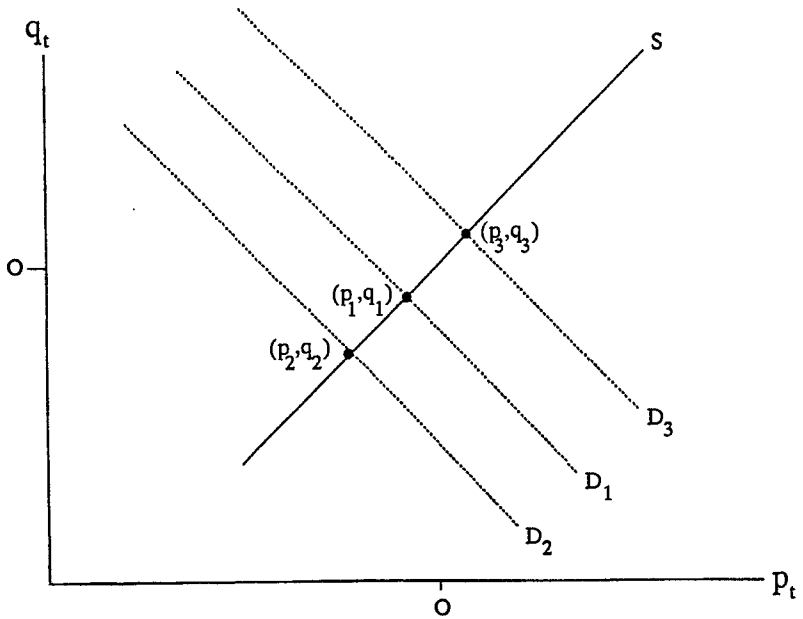


FIGURE 9.3 Observations on price and quantity implied by disturbances to demand function only.

matters for the public's demand for oranges only through its effect on the price. Under this assumption, both w_t and u_t^s are uncorrelated with ε_t^d. Changes in price that can be attributed to the weather represent supply shifts and not demand shifts.

Define p_t^* to be the linear projection of p_t on w_t. Substituting [9.1.7] into [9.1.3],

$$p_t = \frac{\varepsilon_t^d - hw_t - u_t^s}{\gamma - \beta}, \qquad [9.1.8]$$

and thus

$$p_t^* = \frac{-h}{\gamma - \beta} w_t, \qquad [9.1.9]$$

since ε_t^d and u_t^s are uncorrelated with w_t. Equation [9.1.8] can thus be written

$$p_t = p_t^* + \frac{\varepsilon_t^d - u_t^s}{\gamma - \beta},$$

and substituting this into [9.1.1],

$$q_t = \beta\left\{p_t^* + \frac{\varepsilon_t^d - u_t^s}{\gamma - \beta}\right\} + \varepsilon_t^d = \beta p_t^* + v_t, \qquad [9.1.10]$$

where

$$v_t \equiv \frac{-\beta u_t^s}{\gamma - \beta} + \frac{\gamma \varepsilon_t^d}{\gamma - \beta}.$$

Since u_t^s and ε_t^d are both uncorrelated with w_t, it follows that v_t is uncorrelated with p_t^*. Hence, if [9.1.10] were estimated by ordinary least squares, the result would be a consistent estimate of β:

$$\begin{aligned}
\hat{\beta}_T^* &= \frac{(1/T)\sum_{t=1}^{T} p_t^* q_t}{(1/T)\sum_{t=1}^{T} [p_t^*]^2} \\
&= \frac{(1/T)\sum_{t=1}^{T} p_t^*(\beta p_t^* + v_t)}{(1/T)\sum_{t=1}^{T} [p_t^*]^2} \qquad [9.1.11] \\
&= \beta + \frac{(1/T)\sum_{t=1}^{T} p_t^* v_t}{(1/T)\sum_{t=1}^{T} [p_t^*]^2} \\
&\xrightarrow{p} \beta.
\end{aligned}$$

The suggestion is thus to regress quantity on that component of price that is induced by the weather, that is, regress quantity on the linear projection of price on the weather.

In practice, we will not know the values of the population parameters h, γ, and β necessary to construct p_t^* in [9.1.9]. However, the linear projection p_t^* can be consistently estimated by the fitted value for observation t from an *OLS* regression of p on w,

$$\hat{p}_t = \hat{\delta}_T w_t, \qquad [9.1.12]$$

where

$$\hat{\delta}_T = \frac{(1/T)\sum_{t=1}^{T} w_t p_t}{(1/T)\sum_{t=1}^{T} w_t^2}.$$

The estimator [9.1.11] with p_t^* replaced by $\hat{p}_t$ is known as the *two-stage least squares* (*2SLS*) coefficient estimator:

$$\hat{\beta}_{2SLS} = \frac{(1/T)\sum_{t=1}^{T} \hat{p}_t q_t}{(1/T)\sum_{t=1}^{T} [\hat{p}_t]^2}. \qquad [9.1.13]$$

Like $\hat{\beta}_T^*$, the *2SLS* estimator is consistent, as will be shown in the following section.

9.2. *Instrumental Variables and Two-Stage Least Squares*

General Description of Two-Stage Least Squares

A generalization of the previous example is as follows. Suppose the objective is to estimate the vector $\boldsymbol{\beta}$ in the regression model

$$y_t = \boldsymbol{\beta}'\mathbf{z}_t + u_t, \qquad [9.2.1]$$

where $\mathbf{z}_t$ is a $(k \times 1)$ vector of explanatory variables. Some subset $n \leq k$ of the variables in $\mathbf{z}_t$ are thought to be endogenous, that is, correlated with u_t. The remaining $k - n$ variables in $\mathbf{z}_t$ are said to be *predetermined*, meaning that they are uncorrelated with u_t. Estimation of $\boldsymbol{\beta}$ requires variables known as *instruments*. To be a valid instrument, a variable must be correlated with an endogenous explanatory variable in $\mathbf{z}_t$ but uncorrelated with the regression disturbance u_t. In the supply-and-demand example, the weather variable w_t served as an instrument for price. At least one valid instrument must be found for each endogenous explanatory variable.

Collect the predetermined explanatory variables together with the instruments in an $(r \times 1)$ vector $\mathbf{x}_t$. For example, to estimate the demand curve, there were no predetermined explanatory variables in equation [9.1.1] and only a single instrument; hence, $r = 1$, and $\mathbf{x}_t$ would be the scalar w_t. As a second example, suppose that the equation to be estimated is

$$y_t = \beta_1 + \beta_2 z_{2t} + \beta_3 z_{3t} + \beta_4 z_{4t} + \beta_5 z_{5t} + u_t.$$

In this example, z_{4t} and z_{5t} are endogenous (meaning that they are correlated with u_t), z_{2t} and z_{3t} are predetermined (uncorrelated with u_t), and ξ_{1t}, ξ_{2t}, and ξ_{3t} are valid instruments (correlated with z_{4t} and z_{5t} but uncorrelated with u_t). Then $r = 6$ and $\mathbf{x}_t' = (1, z_{2t}, z_{3t}, \xi_{1t}, \xi_{2t}, \xi_{3t})$. The requirement that there be at least as many instruments as endogenous explanatory variables implies that $r \geq k$.

Consider an *OLS* regression of z_{it} (the ith explanatory variable in [9.2.1]) on $\mathbf{x}_t$:

$$z_{it} = \boldsymbol{\delta}_i'\mathbf{x}_t + e_{it}. \qquad [9.2.2]$$

The fitted values for the regression are given by

$$\hat{z}_{it} = \hat{\boldsymbol{\delta}}_i'\mathbf{x}_t, \qquad [9.2.3]$$

where

$$\hat{\boldsymbol{\delta}}_i = \left[\sum_{t=1}^{T} \mathbf{x}_t\mathbf{x}_t'\right]^{-1}\left[\sum_{t=1}^{T} \mathbf{x}_t z_{it}\right].$$

If z_{it} is one of the predetermined variables, then z_{it} is one of the elements of $\mathbf{x}_t$ and equation [9.2.3] simplifies to

$$\hat{z}_{it} = z_{it}.$$

This is because when the dependent variable (z_{it}) is included in the regressors ($\mathbf{x}_t$), a unit coefficient on z_{it} and zero coefficients on the other variables produce a perfect fit and thus minimize the residual sum of squares.

Collect the equations in [9.2.3] for $i = 1, 2, \ldots, k$ in a $(k \times 1)$ vector equation

$$\hat{\mathbf{z}}_t = \hat{\boldsymbol{\delta}}'\mathbf{x}_t, \qquad [9.2.4]$$

where the $(k \times r)$ matrix $\hat{\boldsymbol{\delta}}'$ is given by

$$\hat{\boldsymbol{\delta}}' = \begin{bmatrix} \hat{\boldsymbol{\delta}}_1' \\ \hat{\boldsymbol{\delta}}_2' \\ \vdots \\ \hat{\boldsymbol{\delta}}_k' \end{bmatrix} = \left[\sum_{t=1}^{T} \mathbf{z}_t\mathbf{x}_t'\right]\left[\sum_{t=1}^{T} \mathbf{x}_t\mathbf{x}_t'\right]^{-1}. \qquad [9.2.5]$$

The *two-stage least squares* (*2SLS*) estimate of $\boldsymbol{\beta}$ is found from an *OLS* regression of y_t on $\hat{\mathbf{z}}_t$:

$$\hat{\boldsymbol{\beta}}_{2SLS} = \left[\sum_{t=1}^{T} \hat{\mathbf{z}}_t\hat{\mathbf{z}}_t'\right]^{-1}\left[\sum_{t=1}^{T} \hat{\mathbf{z}}_t y_t\right]. \qquad [9.2.6]$$

An alternative way of writing [9.2.6] is sometimes useful. Let $\hat{e}_{it}$ denote the sample residual from *OLS* estimation of [9.2.2]; that is, let

$$z_{it} = \hat{\boldsymbol{\delta}}_i'\mathbf{x}_t + \hat{e}_{it} = \hat{z}_{it} + \hat{e}_{it}. \qquad [9.2.7]$$

OLS causes this residual to be orthogonal to $\mathbf{x}_t$:

$$\sum_{t=1}^{T} \mathbf{x}_t\hat{e}_{it} = \mathbf{0},$$

meaning that the residual is orthogonal to $\hat{z}_{jt}$:

$$\sum_{t=1}^{T} \hat{z}_{jt}\hat{e}_{it} = \hat{\boldsymbol{\delta}}_j' \sum_{t=1}^{T} \mathbf{x}_t\hat{e}_{it} = 0.$$

Hence, if [9.2.7] is multiplied by $\hat{z}_{jt}$ and summed over t, the result is

$$\sum_{t=1}^{T} \hat{z}_{jt}z_{it} = \sum_{t=1}^{T} \hat{z}_{jt}(\hat{z}_{it} + \hat{e}_{it}) = \sum_{t=1}^{T} \hat{z}_{jt}\hat{z}_{it}$$

for all i and j. This means that

$$\sum_{t=1}^{T} \hat{\mathbf{z}}_t\mathbf{z}_t' = \sum_{t=1}^{T} \hat{\mathbf{z}}_t\hat{\mathbf{z}}_t',$$

so that the *2SLS* estimator [9.2.6] can equivalently be written as

$$\hat{\boldsymbol{\beta}}_{2SLS} = \left[\sum_{t=1}^{T} \hat{\mathbf{z}}_t\mathbf{z}_t'\right]^{-1}\left[\sum_{t=1}^{T} \hat{\mathbf{z}}_t y_t\right]. \qquad [9.2.8]$$

Consistency of 2SLS Estimator

Substituting [9.2.1] into [9.2.8],

$$\begin{aligned}\hat{\boldsymbol{\beta}}_{2SLS,T} &= \left[\sum_{t=1}^{T} \hat{\mathbf{z}}_t \mathbf{z}_t'\right]^{-1}\left[\sum_{t=1}^{T} \hat{\mathbf{z}}_t(\mathbf{z}_t'\boldsymbol{\beta} + u_t)\right] \\ &= \boldsymbol{\beta} + \left[\sum_{t=1}^{T} \hat{\mathbf{z}}_t \mathbf{z}_t'\right]^{-1}\left[\sum_{t=1}^{T} \hat{\mathbf{z}}_t u_t\right],\end{aligned} \tag{9.2.9}$$

where the subscript T has been added to keep explicit track of the sample size T on which estimation is based. It follows from [9.2.9] that

$$\hat{\boldsymbol{\beta}}_{2SLS,T} - \boldsymbol{\beta} = \left[(1/T)\sum_{t=1}^{T} \hat{\mathbf{z}}_t \mathbf{z}_t'\right]^{-1}\left[(1/T)\sum_{t=1}^{T} \hat{\mathbf{z}}_t u_t\right]. \tag{9.2.10}$$

Consistency of the $2SLS$ estimator can then be shown as follows. First note from [9.2.4] and [9.2.5] that

$$\begin{aligned}(1/T)\sum_{t=1}^{T} \hat{\mathbf{z}}_t \mathbf{z}_t' &= \hat{\boldsymbol{\delta}}_T'(1/T)\sum_{t=1}^{T} \mathbf{x}_t \mathbf{z}_t' \\ &= \left[(1/T)\sum_{t=1}^{T} \mathbf{z}_t \mathbf{x}_t'\right]\left[(1/T)\sum_{t=1}^{T} \mathbf{x}_t \mathbf{x}_t'\right]^{-1}\left[(1/T)\sum_{t=1}^{T} \mathbf{x}_t \mathbf{z}_t'\right].\end{aligned} \tag{9.2.11}$$

Assuming that the process $(\mathbf{z}_t, \mathbf{x}_t)$ is covariance-stationary and ergodic for second moments,

$$(1/T)\sum_{t=1}^{T} \hat{\mathbf{z}}_t \mathbf{z}_t' \xrightarrow{p} \mathbf{Q}, \tag{9.2.12}$$

where

$$\mathbf{Q} = [E(\mathbf{z}_t \mathbf{x}_t')][E(\mathbf{x}_t \mathbf{x}_t')]^{-1}[E(\mathbf{x}_t \mathbf{z}_t')]. \tag{9.2.13}$$

Turning next to the second term in [9.2.10],

$$\left[(1/T)\sum_{t=1}^{T} \hat{\mathbf{z}}_t u_t\right] = \hat{\boldsymbol{\delta}}_T'(1/T)\sum_{t=1}^{T} \mathbf{x}_t u_t.$$

Again, ergodicity for second moments implies from [9.2.5] that

$$\hat{\boldsymbol{\delta}}_T' \xrightarrow{p} [E(\mathbf{z}_t \mathbf{x}_t')][E(\mathbf{x}_t \mathbf{x}_t')]^{-1}, \tag{9.2.14}$$

while the law of large numbers will typically ensure that

$$(1/T)\sum_{t=1}^{T} \mathbf{x}_t u_t \xrightarrow{p} E(\mathbf{x}_t u_t) = \mathbf{0},$$

under the assumed absence of correlation between $\mathbf{x}_t$ and u_t. Hence,

$$\left[(1/T)\sum_{t=1}^{T} \hat{\mathbf{z}}_t u_t\right] \xrightarrow{p} \mathbf{0}. \tag{9.2.15}$$

Substituting [9.2.12] and [9.2.15] into [9.2.10], it follows that

$$\hat{\boldsymbol{\beta}}_{2SLS,T} - \boldsymbol{\beta} \xrightarrow{p} \mathbf{Q}^{-1}\cdot\mathbf{0} = \mathbf{0}.$$

Hence, the $2SLS$ estimator is consistent as long as the matrix $\mathbf{Q}$ in [9.2.13] is nonsingular.

Notice that if none of the predetermined variables is correlated with z_{it}, then the ith row of $E(\mathbf{z}_t\mathbf{x}_t')$ contains all zeros and the corresponding row of $\mathbf{Q}$ in [9.2.13] contains all zeros, in which case $2SLS$ is not consistent. Alternatively, if z_{it} is correlated with $\mathbf{x}_t$ only through, say, the first element x_{1t} and z_{jt} is also correlated with $\mathbf{x}_t$ only through x_{1t}, then subtracting some multiple of the ith row of $\mathbf{Q}$ from the jth row produces a row of zeros, and $\mathbf{Q}$ again is not invertible. In general, consistency of the $2SLS$ estimator requires the rows of $E(\mathbf{z}_t\mathbf{x}_t')$ to be linearly independent. This essentially amounts to the requirement that there be a way of assigning instruments to endogenous variables such that each endogenous variable has an instrument associated with it, with no instrument counted twice for this purpose.

Asymptotic Distribution of 2SLS *Estimator*

Equation [9.2.10] implies that

$$\sqrt{T}(\hat{\boldsymbol{\beta}}_{2SLS,T} - \boldsymbol{\beta}) = \left[(1/T)\sum_{t=1}^{T}\hat{\mathbf{z}}_t\mathbf{z}_t'\right]^{-1}\left[(1/\sqrt{T})\sum_{t=1}^{T}\hat{\mathbf{z}}_t u_t\right], \qquad [9.2.16]$$

where

$$\left[(1/\sqrt{T})\sum_{t=1}^{T}\hat{\mathbf{z}}_t u_t\right] = \hat{\boldsymbol{\delta}}_T'(1/\sqrt{T})\sum_{t=1}^{T}\mathbf{x}_t u_t.$$

Hence, from [9.2.12] and [9.2.14],

$$\sqrt{T}(\hat{\boldsymbol{\beta}}_{2SLS,T} - \boldsymbol{\beta}) \xrightarrow{p} \mathbf{Q}^{-1}\cdot[E(\mathbf{z}_t\mathbf{x}_t')][E(\mathbf{x}_t\mathbf{x}_t')]^{-1}\left((1/\sqrt{T})\sum_{t=1}^{T}\mathbf{x}_t u_t\right). \qquad [9.2.17]$$

Suppose that $\mathbf{x}_t$ is covariance-stationary and that $\{u_t\}$ is an i.i.d. sequence with mean zero and variance σ^2 with u_t independent of $\mathbf{x}_s$ for all $s \le t$. Then $\{\mathbf{x}_t u_t\}$ is a martingale difference sequence with variance-covariance matrix given by $\sigma^2\cdot E(\mathbf{x}_t\mathbf{x}_t')$. If u_t and $\mathbf{x}_t$ have finite fourth moments, then we can expect from Proposition 7.9 that

$$\left((1/\sqrt{T})\sum_{t=1}^{T}\mathbf{x}_t u_t\right) \xrightarrow{L} N(\mathbf{0}, \sigma^2\cdot E(\mathbf{x}_t\mathbf{x}_t')). \qquad [9.2.18]$$

Thus, [9.2.17] implies that

$$\sqrt{T}(\hat{\boldsymbol{\beta}}_{2SLS,T} - \boldsymbol{\beta}) \xrightarrow{L} N(\mathbf{0}, \mathbf{V}), \qquad [9.2.19]$$

where

$$\begin{aligned}\mathbf{V} &= \mathbf{Q}^{-1}[E(\mathbf{z}_t\mathbf{x}_t')][E(\mathbf{x}_t\mathbf{x}_t')]^{-1}[\sigma^2\cdot E(\mathbf{x}_t\mathbf{x}_t')][E(\mathbf{x}_t\mathbf{x}_t')]^{-1}[E(\mathbf{x}_t\mathbf{z}_t')]\mathbf{Q}^{-1} \\ &= \sigma^2\mathbf{Q}^{-1}\cdot\mathbf{Q}\cdot\mathbf{Q}^{-1} \qquad [9.2.20] \\ &= \sigma^2\mathbf{Q}^{-1}\end{aligned}$$

for $\mathbf{Q}$ given in [9.2.13]. Hence,

$$\hat{\boldsymbol{\beta}}_{2SLS,T} \approx N(\boldsymbol{\beta}, (1/T)\sigma^2\mathbf{Q}^{-1}). \qquad [9.2.21]$$

Since $\hat{\boldsymbol{\beta}}_{2SLS}$ is a consistent estimate of $\boldsymbol{\beta}$, clearly a consistent estimate of the population residual for observation t is afforded by

$$\hat{u}_{t,T} \equiv y_t - \mathbf{z}_t'\hat{\boldsymbol{\beta}}_{2SLS,T} \xrightarrow{p} u_t. \qquad [9.2.22]$$

Similarly, it is straightforward to show that σ^2 can be consistently estimated by

$$\hat{\sigma}_T^2 = (1/T)\sum_{t=1}^{T}(y_t - \mathbf{z}_t'\hat{\boldsymbol{\beta}}_{2SLS,T})^2 \qquad [9.2.23]$$

(see Exercise 9.1). Note well that although $\hat{\boldsymbol{\beta}}_{2SLS}$ can be calculated from an *OLS* regression of y_t on $\hat{\mathbf{z}}_t$, the estimates $\hat{u}_t$ and $\hat{\sigma}^2$ in [9.2.22] and [9.2.23] are not based on the residuals from this regression:

$$\hat{u}_t \neq y_t - \hat{\mathbf{z}}_t'\hat{\boldsymbol{\beta}}_{2SLS}$$

$$\hat{\sigma}^2 \neq (1/T)\sum_{t=1}^{T}(y_t - \hat{\mathbf{z}}_t'\hat{\boldsymbol{\beta}}_{2SLS})^2.$$

The correct estimates [9.2.22] and [9.2.23] use the actual explanatory variables $\mathbf{z}_t$ rather than the fitted values $\hat{\mathbf{z}}_t$.

A consistent estimate of $\mathbf{Q}$ is provided by [9.2.11]:

$$\begin{aligned}\hat{\mathbf{Q}}_T &= (1/T)\sum_{t=1}^{T}\hat{\mathbf{z}}_t\hat{\mathbf{z}}_t' \\ &= \left[(1/T)\sum_{t=1}^{T}\mathbf{z}_t\mathbf{x}_t'\right]\left[(1/T)\sum_{t=1}^{T}\mathbf{x}_t\mathbf{x}_t'\right]^{-1}\left[(1/T)\sum_{t=1}^{T}\mathbf{x}_t\mathbf{z}_t'\right].\end{aligned} \qquad [9.2.24]$$

Substituting [9.2.23] and [9.2.24] into [9.2.21], the estimated variance-covariance matrix of the $2SLS$ estimator is

$$\begin{aligned}\hat{\mathbf{V}}_T/T &= \hat{\sigma}_T^2\cdot(1/T)\cdot\left[(1/T)\sum_{t=1}^{T}\hat{\mathbf{z}}_t\hat{\mathbf{z}}_t'\right]^{-1} \\ &= \hat{\sigma}_T^2\cdot\left\{\left[\sum_{t=1}^{T}\mathbf{z}_t\mathbf{x}_t'\right]\left[\sum_{t=1}^{T}\mathbf{x}_t\mathbf{x}_t'\right]^{-1}\left[\sum_{t=1}^{T}\mathbf{x}_t\mathbf{z}_t'\right]\right\}^{-1}.\end{aligned} \qquad [9.2.25]$$

A test of the null hypothesis $\mathbf{R}\boldsymbol{\beta} = \mathbf{r}$ can thus be based on

$$(\mathbf{R}\hat{\boldsymbol{\beta}}_{2SLS,T} - \mathbf{r})'[\mathbf{R}(\hat{\mathbf{V}}_T/T)\mathbf{R}']^{-1}(\mathbf{R}\hat{\boldsymbol{\beta}}_{2SLS,T} - \mathbf{r}), \qquad [9.2.26]$$

which, under the null hypothesis, has an asymptotic distribution that is χ^2 with degrees of freedom given by m, where m represents the number of restrictions or the number of rows of $\mathbf{R}$.

Heteroskedasticity- and autocorrelation-consistent standard errors for $2SLS$ estimation will be discussed in Chapter 14.

Instrumental Variable Estimation

Substituting [9.2.4] and [9.2.5] into [9.2.8], the $2SLS$ estimator can be written as

$$\begin{aligned}\hat{\boldsymbol{\beta}}_{2SLS} &= \left[\sum_{t=1}^{T}\hat{\boldsymbol{\delta}}'\mathbf{x}_t\mathbf{z}_t'\right]^{-1}\left[\sum_{t=1}^{T}\hat{\boldsymbol{\delta}}'\mathbf{x}_t y_t\right] \\ &= \left\{\left[\sum_{t=1}^{T}\mathbf{z}_t\mathbf{x}_t'\right]\left[\sum_{t=1}^{T}\mathbf{x}_t\mathbf{x}_t'\right]^{-1}\left[\sum_{t=1}^{T}\mathbf{x}_t\mathbf{z}_t'\right]\right\}^{-1}\left\{\left[\sum_{t=1}^{T}\mathbf{z}_t\mathbf{x}_t'\right]\left[\sum_{t=1}^{T}\mathbf{x}_t\mathbf{x}_t'\right]^{-1}\left[\sum_{t=1}^{T}\mathbf{x}_t y_t\right]\right\}.\end{aligned} \qquad [9.2.27]$$

Consider the special case in which the number of instruments is exactly equal to the number of endogenous explanatory variables, so that $r = k$, as was the case for estimation of the demand curve in Section 9.1. Then $\Sigma_{t=1}^{T} \mathbf{z}_t \mathbf{x}_t'$ is a $(k \times k)$ matrix and [9.2.27] becomes

$$\begin{aligned}\hat{\boldsymbol{\beta}}_{IV} &= \left\{\left[\sum_{t=1}^{T} \mathbf{x}_t \mathbf{z}_t'\right]^{-1}\left[\sum_{t=1}^{T} \mathbf{x}_t \mathbf{x}_t'\right]\left[\sum_{t=1}^{T} \mathbf{z}_t \mathbf{x}_t'\right]^{-1}\right\} \\ &\quad \times \left\{\left[\sum_{t=1}^{T} \mathbf{z}_t \mathbf{x}_t'\right]\left[\sum_{t=1}^{T} \mathbf{x}_t \mathbf{x}_t'\right]^{-1}\left[\sum_{t=1}^{T} \mathbf{x}_t y_t\right]\right\} \qquad [9.2.28] \\ &= \left[\sum_{t=1}^{T} \mathbf{x}_t \mathbf{z}_t'\right]^{-1}\left[\sum_{t=1}^{T} \mathbf{x}_t y_t\right].\end{aligned}$$

Expression [9.2.28] is known as the *instrumental variable* (*IV*) estimator.

A key property of the *IV* estimator can be seen by premultiplying both sides of [9.2.28] by $\Sigma_{t=1}^{T} \mathbf{x}_t \mathbf{z}_t'$:

$$\sum_{t=1}^{T} \mathbf{x}_t \mathbf{z}_t' \hat{\boldsymbol{\beta}}_{IV} = \sum_{t=1}^{T} \mathbf{x}_t y_t,$$

implying that

$$\sum_{t=1}^{T} \mathbf{x}_t (y_t - \mathbf{z}_t' \hat{\boldsymbol{\beta}}_{IV}) = \mathbf{0}. \qquad [9.2.29]$$

Thus, the *IV* sample residual $(y_t - \mathbf{z}_t' \hat{\boldsymbol{\beta}}_{IV})$ has the property that it is orthogonal to the instruments $\mathbf{x}_t$, in contrast to the *OLS* sample residual $(y_t - \mathbf{z}_t' \mathbf{b})$, which is orthogonal to the explanatory variables $\mathbf{z}_t$. The *IV* estimator is preferred to *OLS* because the population residual of the equation we are trying to estimate (u_t) is correlated with $\mathbf{z}_t$ but uncorrelated with $\mathbf{x}_t$.

Since the *IV* estimator is a special case of *2SLS*, it shares the consistency property of the *2SLS* estimator. Its estimated variance with i.i.d. residuals can be calculated from [9.2.25]:

$$\hat{\sigma}_T^2 \left[\sum_{t=1}^{T} \mathbf{x}_t \mathbf{z}_t'\right]^{-1}\left[\sum_{t=1}^{T} \mathbf{x}_t \mathbf{x}_t'\right]\left[\sum_{t=1}^{T} \mathbf{z}_t \mathbf{x}_t'\right]^{-1}. \qquad [9.2.30]$$

9.3. *Identification*

We noted in the supply-and-demand example in Section 9.1 that the demand elasticity β could not be estimated consistently by an *OLS* regression of quantity on price. Indeed, in the absence of a valid instrument such as w_t, the demand elasticity cannot be estimated by any method! To see this, recall that the system as written in [9.1.1] and [9.1.2] implied the expressions [9.1.4] and [9.1.3]:

$$\begin{aligned} q_t &= \frac{\gamma}{\gamma - \beta}\varepsilon_t^d - \frac{\beta}{\gamma - \beta}\varepsilon_t^s \\ p_t &= \frac{\varepsilon_t^d - \varepsilon_t^s}{\gamma - \beta}. \end{aligned}$$

If ε_t^d and ε_t^s are i.i.d. Gaussian, then these equations imply that the vector $(q_t, p_t)'$ is Gaussian with mean zero and variance-covariance matrix

$$\mathbf{\Omega} \equiv [1/(\gamma - \beta)^2] \cdot \begin{bmatrix} \gamma^2\sigma_d^2 + \beta^2\sigma_s^2 & \gamma\sigma_d^2 + \beta\sigma_s^2 \\ \gamma\sigma_d^2 + \beta\sigma_s^2 & \sigma_d^2 + \sigma_s^2 \end{bmatrix}.$$

This matrix is completely described by three magnitudes, these being the variances of q and p along with their covariance. Given a large enough sample, the values of these three magnitudes can be inferred with considerable confidence, but that is all that can be inferred, because these magnitudes can completely specify the process that generated the data under the maintained assumption of zero-mean i.i.d. Gaussian observations. There is no way to uncover the four parameters of the structural model $(\beta, \gamma, \sigma_d^2, \sigma_s^2)$ from these three magnitudes. For example, the values $(\beta, \gamma, \sigma_d^2, \sigma_s^2) = (1, 2, 3, 4)$ imply exactly the same observable properties for the data as would $(\beta, \gamma, \sigma_d^2, \sigma_s^2) = (2, 1, 4, 3)$.

If two different values for a parameter vector $\boldsymbol{\theta}$ imply the same probability distribution for the observed data, then the vector $\boldsymbol{\theta}$ is said to be *unidentified*.

When a third Gaussian white noise variable w_t is added to the set of observations, three additional magnitudes are available to characterize the process for observables, these being the variance of w, the covariance between w and p, and the covariance between w and q. If the new variable w enters both the demand and the supply equation, then three new parameters would be required to estimate the structural model—the parameter that summarizes the effect of w on demand, the parameter that summarizes its effect on supply, and the variance of w. With three more estimable magnitudes but three more parameters to estimate, we would be stuck with the same problem, having no basis for estimation of β.

Consistent estimation of the demand elasticity was achieved by using two-stage least squares because it was assumed that w appeared in the supply equation but was excluded from the demand equation. This is known as achieving identification through *exclusion restrictions*.

We showed in Section 9.2 that the parameters of an equation could be estimated (and thus must be identified) if (1) the number of instruments for that equation is at least as great as the number of endogenous explanatory variables for that equation and (2) the rows of $E(\mathbf{z}_t\mathbf{x}_t')$ are linearly independent. The first condition is known as the *order* condition for identification, and the second is known as the *rank* condition.

The rank condition for identification can be summarized more explicitly by specifying a complete system of equations for all of the endogenous variables. Let $\mathbf{y}_t$ denote an $(n \times 1)$ vector containing all of the endogenous variables in the system, and let $\mathbf{x}_t$ denote an $(m \times 1)$ vector containing all of the predetermined variables. Suppose that the system consists of n equations written as

$$\mathbf{B}\mathbf{y}_t + \mathbf{\Gamma}\mathbf{x}_t = \mathbf{u}_t, \qquad [9.3.1]$$

where $\mathbf{B}$ and $\mathbf{\Gamma}$ are $(n \times n)$ and $(n \times m)$ matrices of coefficients, respectively, and $\mathbf{u}_t$ is an $(n \times 1)$ vector of disturbances. The statement that $\mathbf{x}_t$ is predetermined is taken to mean that $E(\mathbf{x}_t\mathbf{u}_t') = \mathbf{0}$. For example, the demand and supply equations considered in Section 9.1 were

$$q_t = \beta p_t + u_t^d \qquad \text{(demand)} \qquad [9.3.2]$$

$$q_t = \gamma p_t + h w_t + u_t^s \qquad \text{(supply)}. \qquad [9.3.3]$$

For this system, there are $n = 2$ endogenous variables, with $\mathbf{y}_t = (q_t, p_t)'$; and $m = 1$ predetermined variable, so that $\mathbf{x}_t = w_t$. This system can be written in the form of [9.3.1] as

$$\begin{bmatrix} 1 & -\beta \\ 1 & -\gamma \end{bmatrix} \begin{bmatrix} q_t \\ p_t \end{bmatrix} + \begin{bmatrix} 0 \\ -h \end{bmatrix} w_t = \begin{bmatrix} u_t^d \\ u_t^s \end{bmatrix}. \qquad [9.3.4]$$

Suppose we are interested in the equation represented by the first row of the vector system of equations in [9.3.1]. Let y_{0t} be the dependent variable in the first equation, and let $\mathbf{y}_{1t}$ denote an $(n_1 \times 1)$ vector consisting of those endogenous variables that appear in the first equation as explanatory variables. Similarly, let $\mathbf{x}_{1t}$ denote an $(m_1 \times 1)$ vector consisting of those predetermined variables that appear in the first equation as explanatory variables. Then the first equation in the system is

$$y_{0t} + \mathbf{B}_{01}\mathbf{y}_{1t} + \mathbf{\Gamma}_{01}\mathbf{x}_{1t} = u_{0t},$$

where $\mathbf{B}_{01}$ is a $(1 \times n_1)$ vector and $\mathbf{\Gamma}_{01}$ is a $(1 \times m_1)$ vector. Let $\mathbf{y}_{2t}$ denote an $(n_2 \times 1)$ vector consisting of those endogenous variables that do not appear in the first equation; thus, $\mathbf{y}_t' = (y_{0t}, \mathbf{y}_{1t}', \mathbf{y}_{2t}')$ and $1 + n_1 + n_2 = n$. Similarly, let $\mathbf{x}_{2t}$ denote an $(m_2 \times 1)$ vector consisting of those predetermined variables that do not appear in the first equation, so that $\mathbf{x}_t' = (\mathbf{x}_{1t}', \mathbf{x}_{2t}')$ and $m_1 + m_2 = m$. Then the system in [9.3.1] can be written in partitioned form as

$$\begin{bmatrix} 1 & \mathbf{B}_{01} & \mathbf{0}' \\ \mathbf{B}_{10} & \mathbf{B}_{11} & \mathbf{B}_{12} \\ \mathbf{B}_{20} & \mathbf{B}_{21} & \mathbf{B}_{22} \end{bmatrix} \begin{bmatrix} y_{0t} \\ \mathbf{y}_{1t} \\ \mathbf{y}_{2t} \end{bmatrix} + \begin{bmatrix} \mathbf{\Gamma}_{01} & \mathbf{0}' \\ \mathbf{\Gamma}_{11} & \mathbf{\Gamma}_{12} \\ \mathbf{\Gamma}_{21} & \mathbf{\Gamma}_{22} \end{bmatrix} \begin{bmatrix} \mathbf{x}_{1t} \\ \mathbf{x}_{2t} \end{bmatrix} = \begin{bmatrix} u_{0t} \\ \mathbf{u}_{1t} \\ \mathbf{u}_{2t} \end{bmatrix}. \qquad [9.3.5]$$

Here, for example, $\mathbf{B}_{12}$ is an $(n_1 \times n_2)$ matrix consisting of rows 2 through $(n_1 + 1)$ and columns $(n_1 + 2)$ through n of the matrix $\mathbf{B}$.

An alternative useful representation of the system is obtained by moving $\mathbf{\Gamma}\mathbf{x}_t$ to the right side of [9.3.1] and premultiplying both sides by $\mathbf{B}^{-1}$:

$$\mathbf{y}_t = -\mathbf{B}^{-1}\mathbf{\Gamma}\mathbf{x}_t + \mathbf{B}^{-1}\mathbf{u}_t = \mathbf{\Pi}'\mathbf{x}_t + \mathbf{v}_t, \qquad [9.3.6]$$

where

$$\mathbf{\Pi}' = -\mathbf{B}^{-1}\mathbf{\Gamma} \qquad [9.3.7]$$

$$\mathbf{v}_t = \mathbf{B}^{-1}\mathbf{u}_t. \qquad [9.3.8]$$

Expression [9.3.6] is known as the *reduced-form* representation of the structural system [9.3.1]. In the reduced-form representation, each endogenous variable is expressed solely as a function of predetermined variables. For the example of [9.3.4], the reduced form is

$$\begin{aligned} \begin{bmatrix} q_t \\ p_t \end{bmatrix} &= -\begin{bmatrix} 1 & -\beta \\ 1 & -\gamma \end{bmatrix}^{-1} \begin{bmatrix} 0 \\ -h \end{bmatrix} w_t + \begin{bmatrix} 1 & -\beta \\ 1 & -\gamma \end{bmatrix}^{-1} \begin{bmatrix} u_t^d \\ u_t^s \end{bmatrix} \\ &= [1/(\beta - \gamma)] \begin{bmatrix} -\gamma & \beta \\ -1 & 1 \end{bmatrix} \begin{bmatrix} 0 \\ h \end{bmatrix} w_t \\ &\quad + [1/(\beta - \gamma)] \begin{bmatrix} -\gamma & \beta \\ -1 & 1 \end{bmatrix} \begin{bmatrix} u_t^d \\ u_t^s \end{bmatrix} \\ &= [1/(\beta - \gamma)] \begin{bmatrix} \beta h \\ h \end{bmatrix} w_t + [1/(\beta - \gamma)] \begin{bmatrix} -\gamma u_t^d + \beta u_t^s \\ -u_t^d + u_t^s \end{bmatrix}. \end{aligned} \qquad [9.3.9]$$

The reduced form for a general system can be written in partitioned form as

$$\begin{bmatrix} y_{0t} \\ \mathbf{y}_{1t} \\ \mathbf{y}_{2t} \end{bmatrix} = \begin{bmatrix} \mathbf{\Pi}_{01} & \mathbf{\Pi}_{02} \\ \mathbf{\Pi}_{11} & \mathbf{\Pi}_{12} \\ \mathbf{\Pi}_{21} & \mathbf{\Pi}_{22} \end{bmatrix} \begin{bmatrix} \mathbf{x}_{1t} \\ \mathbf{x}_{2t} \end{bmatrix} + \begin{bmatrix} v_{0t} \\ \mathbf{v}_{1t} \\ \mathbf{v}_{2t} \end{bmatrix}, \qquad [9.3.10]$$

where, for example, $\mathbf{\Pi}_{12}$ denotes an $(n_1 \times m_2)$ matrix consisting of rows 2 through $(n_1 + 1)$ and columns $(m_1 + 1)$ through m of the matrix $\mathbf{\Pi}'$.

To apply the rank condition for identification of the first equation stated earlier, we would form the matrix of cross products between the explanatory variables in the first equation ($\mathbf{x}_{1t}$ and $\mathbf{y}_{1t}$) and the predetermined variables for the whole system ($\mathbf{x}_{1t}$ and $\mathbf{x}_{2t}$):

$$\mathbf{M} = \begin{bmatrix} E(\mathbf{x}_{1t}\mathbf{x}_{1t}') & E(\mathbf{x}_{1t}\mathbf{x}_{2t}') \\ E(\mathbf{y}_{1t}\mathbf{x}_{1t}') & E(\mathbf{y}_{1t}\mathbf{x}_{2t}') \end{bmatrix}. \qquad [9.3.11]$$

In the earlier notation, the explanatory variables for the first equation consist of $\mathbf{z}_t = (\mathbf{x}_{1t}', \mathbf{y}_{1t}')'$, while the predetermined variables for the system as a whole consist of $\mathbf{x}_t = (\mathbf{x}_{1t}', \mathbf{x}_{2t}')'$. Thus, the rank condition, which required the rows of $E(\mathbf{z}_t\mathbf{x}_t')$ to be linearly independent, amounts to the requirement that the rows of the $[(m_1 + n_1) \times m]$ matrix $\mathbf{M}$ in [9.3.11] be linearly independent. The rank condition can equivalently be stated in terms of the structural parameter matrices $\mathbf{B}$ and $\mathbf{\Gamma}$ or the reduced-form parameter matrix $\mathbf{\Pi}$. The following proposition is adapted from Fisher (1966) and is proved in Appendix 9.A at the end of this chapter.

Proposition 9.1: *If the matrix* $\mathbf{B}$ *in [9.3.1] and the matrix of second moments of the predetermined variables* $E(\mathbf{x}_t\mathbf{x}_t')$ *are both nonsingular, then the following conditions are equivalent:*

(*a*) *The rows of the* $[(m_1 + n_1) \times m]$ *matrix* $\mathbf{M}$ *in [9.3.11] are linearly independent.*

(*b*) *The rows of the* $[(n_1 + n_2) \times (m_2 + n_2)]$ *matrix*

$$\begin{bmatrix} \mathbf{\Gamma}_{12} & \mathbf{B}_{12} \\ \mathbf{\Gamma}_{22} & \mathbf{B}_{22} \end{bmatrix} \qquad [9.3.12]$$

are linearly independent.

(*c*) *The rows of the* $(n_1 \times m_2)$ *matrix* $\mathbf{\Pi}_{12}$ *are linearly independent.*

For example, for the system in [9.3.4], no endogenous variables are excluded from the first equation, and so $y_{0t} = q_t$, $\mathbf{y}_{1t} = p_t$, and $\mathbf{y}_{2t}$ contains no elements. No predetermined variables appear in the first equation, and so $\mathbf{x}_{1t}$ contains no elements and $\mathbf{x}_{2t} = w_t$. The matrix in [9.3.12] is then just given by the parameter $\mathbf{\Gamma}_{12}$. This represents the coefficient on $\mathbf{x}_{2t}$ in the equation describing $\mathbf{y}_{1t}$ and is equal to the scalar parameter $-h$. Result (b) of Proposition 9.1 thus states that the first equation is identified provided that $h \neq 0$. The value of $\mathbf{\Pi}_{12}$ can be read directly off the coefficient on w_t in the second row of [9.3.9] and turns out to be given by $h/(\beta - \gamma)$. Since $\mathbf{B}$ is assumed to be nonsingular, $(\beta - \gamma)$ is nonzero, and so $\mathbf{\Gamma}_{12}$ is zero if and only if $\mathbf{\Pi}_{12}$ is zero.

Achieving Identification Through Covariance Restrictions

Another way in which parameters can be identified is through restrictions on the covariances of the errors of the structural equations. For example, consider

again the supply and demand model, [9.3.2] and [9.3.3]. We saw that the demand elasticity β was identified by the exclusion of w_t from the demand equation. Consider now estimation of the supply elasticity γ.

Suppose first that we somehow knew the value of the demand elasticity β with certainty. Then the error in the demand equation could be constructed from

$$u_t^d = q_t - \beta p_t.$$

Notice that u_t^d would then be a valid instrument for the supply equation [9.3.3], since u_t^d is correlated with the endogenous explanatory variable for that equation (p_t) but u_t^d is uncorrelated with the error for that equation (u_t^s). Since w_t is also uncorrelated with the error u_t^s, it follows that the parameters of the supply equation could be estimated consistently by instrumental variable estimation with $\mathbf{x}_t = (u_t^d, w_t)'$:

$$\begin{bmatrix} \hat{\gamma}_T^* \\ \hat{h}_T^* \end{bmatrix} = \begin{bmatrix} \Sigma u_t^d p_t & \Sigma u_t^d w_t \\ \Sigma w_t p_t & \Sigma w_t^2 \end{bmatrix}^{-1} \begin{bmatrix} \Sigma u_t^d q_t \\ \Sigma w_t q_t \end{bmatrix} \xrightarrow{p} \begin{bmatrix} \gamma \\ h \end{bmatrix}, \qquad [9.3.13]$$

where Σ indicates summation over $t = 1, 2, \ldots, T$.

Although in practice we do not know the true value of β, it can be estimated consistently by *IV* estimation of [9.3.2] with w_t as an instrument:

$$\hat{\beta} = (\Sigma w_t p_t)^{-1}(\Sigma w_t q_t).$$

Then the residual u_t^d can be estimated consistently with $\hat{u}_t^d = q_t - \hat{\beta} p_t$. Consider, therefore, the estimator [9.3.13] with the population residual u_t^d replaced by the *IV* sample residual:

$$\begin{bmatrix} \hat{\gamma}_T \\ \hat{h}_T \end{bmatrix} = \begin{bmatrix} \Sigma \hat{u}_t^d p_t & \Sigma \hat{u}_t^d w_t \\ \Sigma w_t p_t & \Sigma w_t^2 \end{bmatrix}^{-1} \begin{bmatrix} \Sigma \hat{u}_t^d q_t \\ \Sigma w_t q_t \end{bmatrix}. \qquad [9.3.14]$$

It is straightforward to use the fact that $\hat{\beta} \xrightarrow{p} \beta$ to deduce that the difference between the estimators in [9.3.14] and [9.3.13] converges in probability to zero. Hence, the estimator [9.3.14] is also consistent.

Two assumptions allowed the parameters of the supply equation (γ and h) to be estimated. First, an exclusion restriction allowed β to be estimated consistently. Second, a restriction on the covariance between u_t^d and u_t^s was necessary. If u_t^d were correlated with u_t^s, then u_t^d would not be a valid instrument for the supply equation and the estimator [9.3.13] would not be consistent.

Other Approaches to Identification

A good deal more can be said about identification. For example, parameters can also be identified through the imposition of certain restrictions on parameters such as $\beta_1 + \beta_2 = 1$. Useful references include Fisher (1966), Rothenberg (1971), and Hausman and Taylor (1983).

9.4. *Full-Information Maximum Likelihood Estimation*

Up to this point we have considered estimation of a single equation of the form $y_t = \boldsymbol{\beta}'\mathbf{z}_t + u_t$. A more general approach is to specify a similar equation for every endogenous variable in the system, calculate the joint density of the vector of all of the endogenous variables conditional on the predetermined variables, and maximize the joint likelihood function. This is known as *full-information maximum likelihood* estimation, or *FIML*.

For illustration, suppose in [9.3.1] that the $(n \times 1)$ vector of structural disturbances $\mathbf{u}_t$ for date t is distributed $N(\mathbf{0}, \mathbf{D})$. Assume, further, that $\mathbf{u}_t$ is independent of $\mathbf{u}_\tau$ for $t \neq \tau$ and that $\mathbf{u}_t$ is independent of $\mathbf{x}_\tau$ for all t and τ. Then the reduced-form disturbance $\mathbf{v}_t = \mathbf{B}^{-1}\mathbf{u}_t$ is distributed $N(\mathbf{0}, \mathbf{B}^{-1}\mathbf{D}(\mathbf{B}^{-1})')$, and the reduced-form representation [9.3.6] implies that

$$\mathbf{y}_t|\mathbf{x}_t \sim N\Big(\mathbf{\Pi}'\mathbf{x}_t, \mathbf{B}^{-1}\mathbf{D}(\mathbf{B}^{-1})'\Big) = N\Big(-\mathbf{B}^{-1}\mathbf{\Gamma}\mathbf{x}_t, \mathbf{B}^{-1}\mathbf{D}(\mathbf{B}^{-1})'\Big).$$

The conditional log likelihood can then be found from

$$\begin{aligned}
&\mathcal{L}(\mathbf{B}, \mathbf{\Gamma}, \mathbf{D}) \\
&\quad = \sum_{t=1}^{T} \log f(\mathbf{y}_t|\mathbf{x}_t; \mathbf{B}, \mathbf{\Gamma}, \mathbf{D}) \\
&\quad = -(Tn/2)\log(2\pi) - (T/2)\log|\mathbf{B}^{-1}\mathbf{D}(\mathbf{B}^{-1})'| \\
&\qquad - (1/2)\sum_{t=1}^{T} [\mathbf{y}_t + \mathbf{B}^{-1}\mathbf{\Gamma}\mathbf{x}_t]'[\mathbf{B}^{-1}\mathbf{D}(\mathbf{B}^{-1})']^{-1}[\mathbf{y}_t + \mathbf{B}^{-1}\mathbf{\Gamma}\mathbf{x}_t].
\end{aligned} \tag{9.4.1}$$

But

$$\begin{aligned}
&[\mathbf{y}_t + \mathbf{B}^{-1}\mathbf{\Gamma}\mathbf{x}_t]'[\mathbf{B}^{-1}\mathbf{D}(\mathbf{B}^{-1})']^{-1}[\mathbf{y}_t + \mathbf{B}^{-1}\mathbf{\Gamma}\mathbf{x}_t] \\
&\qquad = [\mathbf{y}_t + \mathbf{B}^{-1}\mathbf{\Gamma}\mathbf{x}_t]'[\mathbf{B}'\mathbf{D}^{-1}\mathbf{B}][\mathbf{y}_t + \mathbf{B}^{-1}\mathbf{\Gamma}\mathbf{x}_t] \\
&\qquad = [\mathbf{B}(\mathbf{y}_t + \mathbf{B}^{-1}\mathbf{\Gamma}\mathbf{x}_t)]'\mathbf{D}^{-1}[\mathbf{B}(\mathbf{y}_t + \mathbf{B}^{-1}\mathbf{\Gamma}\mathbf{x}_t)] \\
&\qquad = [\mathbf{B}\mathbf{y}_t + \mathbf{\Gamma}\mathbf{x}_t]'\mathbf{D}^{-1}[\mathbf{B}\mathbf{y}_t + \mathbf{\Gamma}\mathbf{x}_t].
\end{aligned} \tag{9.4.2}$$

Furthermore,

$$\begin{aligned}
|\mathbf{B}^{-1}\mathbf{D}(\mathbf{B}^{-1})'| &= |\mathbf{B}^{-1}|\cdot|\mathbf{D}|\cdot|\mathbf{B}^{-1}| \\
&= |\mathbf{D}|/|\mathbf{B}|^2.
\end{aligned} \tag{9.4.3}$$

Substituting [9.4.2] and [9.4.3] into [9.4.1],

$$\begin{aligned}
\mathcal{L}(\mathbf{B}, \mathbf{\Gamma}, \mathbf{D}) &= -(Tn/2)\log(2\pi) + (T/2)\log|\mathbf{B}|^2 \\
&\quad - (T/2)\log|\mathbf{D}| - (1/2)\sum_{t=1}^{T} [\mathbf{B}\mathbf{y}_t + \mathbf{\Gamma}\mathbf{x}_t]'\mathbf{D}^{-1}[\mathbf{B}\mathbf{y}_t + \mathbf{\Gamma}\mathbf{x}_t].
\end{aligned} \tag{9.4.4}$$

The *FIML* estimates are then the values of $\mathbf{B}$, $\mathbf{\Gamma}$, and $\mathbf{D}$ for which [9.4.4] is maximized.

For example, for the system of [9.3.4], the *FIML* estimates of β, γ, h, σ_d^2, and σ_s^2 are found by maximizing

$$\begin{aligned}
&\mathcal{L}(\beta, \gamma, h, \sigma_d^2, \sigma_s^2) \\
&\quad = -T\log(2\pi) + \frac{T}{2}\log\begin{vmatrix} 1 & -\beta \\ 1 & -\gamma \end{vmatrix}^2 - \frac{T}{2}\log\begin{vmatrix} \sigma_d^2 & 0 \\ 0 & \sigma_s^2 \end{vmatrix} \\
&\qquad - \frac{1}{2}\sum_{t=1}^{T}\left\{[q_t - \beta p_t \quad q_t - \gamma p_t - hw_t]\begin{bmatrix} \sigma_d^2 & 0 \\ 0 & \sigma_s^2 \end{bmatrix}^{-1}\begin{bmatrix} q_t - \beta p_t \\ q_t - \gamma p_t - hw_t \end{bmatrix}\right\} \\
&\quad = -T\log(2\pi) + T\log(\gamma - \beta) - (T/2)\log(\sigma_d^2) \\
&\qquad - (T/2)\log(\sigma_s^2) - (1/2)\sum_{t=1}^{T}(q_t - \beta p_t)^2/\sigma_d^2 \\
&\qquad - (1/2)\sum_{t=1}^{T}(q_t - \gamma p_t - hw_t)^2/\sigma_s^2.
\end{aligned} \tag{9.4.5}$$

The first-order conditions for maximization are

$$\frac{\partial \mathcal{L}}{\partial \beta} = -\frac{T}{\gamma - \beta} + \frac{\sum_{t=1}^{T} (q_t - \beta p_t)p_t}{\sigma_d^2} = 0 \qquad [9.4.6]$$

$$\frac{\partial \mathcal{L}}{\partial \gamma} = \frac{T}{\gamma - \beta} + \frac{\sum_{t=1}^{T} (q_t - \gamma p_t - hw_t)p_t}{\sigma_s^2} = 0 \qquad [9.4.7]$$

$$\frac{\partial \mathcal{L}}{\partial h} = \frac{\sum_{t=1}^{T} (q_t - \gamma p_t - hw_t)w_t}{\sigma_s^2} = 0 \qquad [9.4.8]$$

$$\frac{\partial \mathcal{L}}{\partial \sigma_d^2} = -\frac{T}{2\sigma_d^2} + \frac{\sum_{t=1}^{T} (q_t - \beta p_t)^2}{2\sigma_d^4} = 0 \qquad [9.4.9]$$

$$\frac{\partial \mathcal{L}}{\partial \sigma_s^2} = -\frac{T}{2\sigma_s^2} + \frac{\sum_{t=1}^{T} (q_t - \gamma p_t - hw_t)^2}{2\sigma_s^4} = 0. \qquad [9.4.10]$$

The last two equations characterize the maximum likelihood estimates of the variances as the average squared residuals:

$$\hat{\sigma}_d^2 = (1/T) \sum_{t=1}^{T} (q_t - \hat{\beta} p_t)^2 \qquad [9.4.11]$$

$$\hat{\sigma}_s^2 = (1/T) \sum_{t=1}^{T} (q_t - \hat{\gamma} p_t - \hat{h} w_t)^2. \qquad [9.4.12]$$

Multiplying equation [9.4.7] by $(\beta - \gamma)/T$ results in

$$\begin{aligned} 0 &= -1 + \sum_{t=1}^{T} (q_t - \gamma p_t - hw_t)(\beta p_t - \gamma p_t)/(T\sigma_s^2) \\ &= -1 + \sum_{t=1}^{T} (q_t - \gamma p_t - hw_t)(\beta p_t - q_t + q_t - \gamma p_t)/(T\sigma_s^2). \end{aligned} \qquad [9.4.13]$$

If [9.4.8] is multiplied by h/T and subtracted from [9.4.13], the result is

$$\begin{aligned} 0 &= -1 + \sum_{t=1}^{T} (q_t - \gamma p_t - hw_t)(\beta p_t - q_t + q_t - \gamma p_t - hw_t)/(T\sigma_s^2) \\ &= -1 + \sum_{t=1}^{T} (q_t - \gamma p_t - hw_t)(\beta p_t - q_t)/(T\sigma_s^2) \\ &\quad + \sum_{t=1}^{T} (q_t - \gamma p_t - hw_t)^2/(T\sigma_s^2) \\ &= -1 - \sum_{t=1}^{T} (q_t - \gamma p_t - hw_t)(q_t - \beta p_t)/(T\sigma_s^2) + 1, \end{aligned}$$

by virtue of [9.4.12]. Hence, the *MLE*s satisfy

$$\sum_{t=1}^{T} (q_t - \hat{\gamma} p_t - \hat{h} w_t)(q_t - \hat{\beta} p_t) = 0. \qquad [9.4.14]$$

Similarly, multiplying [9.4.6] by $(\gamma - \beta)/T$,

$$0 = -1 + \sum_{t=1}^{T} (q_t - \beta p_t)(\gamma p_t - q_t + q_t - \beta p_t)/(T\sigma_d^2)$$

$$= -1 - \sum_{t=1}^{T} (q_t - \beta p_t)(q_t - \gamma p_t)/(T\sigma_d^2) + \sum_{t=1}^{T} (q_t - \beta p_t)^2/(T\sigma_d^2).$$

Using [9.4.11],

$$\sum_{t=1}^{T} (q_t - \hat{\beta} p_t)(q_t - \hat{\gamma} p_t) = 0. \qquad [9.4.15]$$

Subtracting [9.4.14] from [9.4.15],

$$0 = \sum_{t=1}^{T} (q_t - \hat{\beta} p_t)[(q_t - \hat{\gamma} p_t) - (q_t - \hat{\gamma} p_t - \hat{h} w_t)] = \hat{h} \sum_{t=1}^{T} (q_t - \hat{\beta} p_t) w_t.$$

Assuming that $\hat{h} \neq 0$, the *FIML* estimate of β thus satisfies

$$\sum_{t=1}^{T} (q_t - \hat{\beta} p_t) w_t = 0;$$

that is, the demand elasticity is chosen so as to make the estimated residual for the demand equation orthogonal to w_t. Hence, the instrumental variable estimator $\hat{\beta}_{IV}$ turns out also to be the *FIML* estimator. Equations [9.4.8] and [9.4.14] assert that the parameters for the supply equation (γ and h) are chosen so as to make the residual for that equation orthogonal to w_t and to the demand residual $\hat{u}_t^d = q_t - \hat{\beta} p_t$. Hence, the *FIML* estimates for these parameters are the same as the instrumental-variable estimates suggested in [9.3.14].

For this example, two-stage least squares, instrumental variable estimation, and full-information maximum likelihood all produced the identical estimates. This is because the model is *just identified*. A model is said to be just identified if for any admissible value for the parameters of the reduced-form representation there exists a unique value for the structural parameters that implies those reduced-form parameters. A model is said to be *overidentified* if some admissible values for the reduced-form parameters are ruled out by the structural restrictions. In an overidentified model, *IV*, *2SLS*, and *FIML* estimation are not equivalent, and *FIML* typically produces the most efficient estimates.

For a general overidentified simultaneous equation system with no restrictions on the variance-covariance matrix, the *FIML* estimates can be calculated by iterating on a procedure known as *three-stage least squares*; see, for example, Maddala (1977, pp. 482–90). Rothenberg and Ruud (1990) discussed *FIML* estimation in the presence of covariance restrictions. *FIML* estimation of dynamic time series models will be discussed further in Chapter 11.

9.5 *Estimation Based on the Reduced Form*

If a system is just identified as in [9.3.2] and [9.3.3] with u_t^d uncorrelated with u_t^s, one approach is to maximize the likelihood function with respect to the reduced-form parameters. The values of the structural parameters associated with these values for the reduced-form parameters are the same as the *FIML* estimates in a just-identified model.

The log likelihood [9.4.1] can be expressed in terms of the reduced-form parameters $\mathbf{\Pi}$ and $\mathbf{\Omega}$ as

$$\begin{aligned}\mathcal{L}(\mathbf{\Pi}, \mathbf{\Omega}) &= \sum_{t=1}^{T} \log f(\mathbf{y}_t|\mathbf{x}_t; \mathbf{\Pi}, \mathbf{\Omega}) \\ &= -(Tn/2)\log(2\pi) - (T/2)\log|\mathbf{\Omega}| \\ &\quad - (1/2)\sum_{t=1}^{T}[\mathbf{y}_t - \mathbf{\Pi}'\mathbf{x}_t]'\mathbf{\Omega}^{-1}[\mathbf{y}_t - \mathbf{\Pi}'\mathbf{x}_t],\end{aligned} \tag{9.5.1}$$

where $\mathbf{\Omega} = E(\mathbf{v}_t\mathbf{v}_t') = \mathbf{B}^{-1}\mathbf{D}(\mathbf{B}^{-1})'$. The value of $\mathbf{\Pi}$ that maximizes [9.5.1] will be shown in Chapter 11 to be given by

$$\hat{\mathbf{\Pi}}' = \left[\sum_{t=1}^{T} \mathbf{y}_t\mathbf{x}_t'\right]\left[\sum_{t=1}^{T} \mathbf{x}_t\mathbf{x}_t'\right]^{-1};$$

in other words, the ith row of $\hat{\mathbf{\Pi}}'$ is obtained from an *OLS* regression of the ith endogenous variable on all of the predetermined variables:

$$\hat{\boldsymbol{\pi}}_i' = \left[\sum_{t=1}^{T} y_{it}\mathbf{x}_t'\right]\left[\sum_{t=1}^{T} \mathbf{x}_t\mathbf{x}_t'\right]^{-1}.$$

The *MLE* of $\mathbf{\Omega}$ turns out to be

$$\hat{\mathbf{\Omega}} = (1/T)\left[\sum_{t=1}^{T} (\mathbf{y}_t - \hat{\mathbf{\Pi}}'\mathbf{x}_t)(\mathbf{y}_t - \hat{\mathbf{\Pi}}'\mathbf{x}_t)'\right].$$

For a just-identified model, the *FIML* estimates are the values of $(\mathbf{B}, \mathbf{\Gamma}, \mathbf{D})$ for which $\hat{\mathbf{\Pi}}' = -\mathbf{B}^{-1}\mathbf{\Gamma}$ and $\hat{\mathbf{\Omega}} = \mathbf{B}^{-1}\mathbf{D}(\mathbf{B}^{-1})'$.

We now show that the estimates of $\mathbf{B}$, $\mathbf{\Gamma}$, and $\mathbf{D}$ inferred in this fashion from the reduced-form parameters for the just-identified supply-and-demand example are the same as the *FIML* estimates. The estimate $\hat{\pi}_1$ is found by *OLS* regression of q_t on w_t, while $\hat{\pi}_2$ is the coefficient from an *OLS* regression of p_t on w_t. These estimates satisfy

$$\sum_{t=1}^{T} (q_t - \hat{\pi}_1 w_t)w_t = 0 \tag{9.5.2}$$

$$\sum_{t=1}^{T} (p_t - \hat{\pi}_2 w_t)w_t = 0 \tag{9.5.3}$$

and

$$\begin{bmatrix}\hat{\Omega}_{11} & \hat{\Omega}_{12} \\ \hat{\Omega}_{21} & \hat{\Omega}_{22}\end{bmatrix} = (1/T)\begin{bmatrix}\Sigma(q_t - \hat{\pi}_1 w_t)^2 & \Sigma(q_t - \hat{\pi}_1 w_t)(p_t - \hat{\pi}_2 w_t) \\ \Sigma(p_t - \hat{\pi}_2 w_t)(q_t - \hat{\pi}_1 w_t) & \Sigma(p_t - \hat{\pi}_2 w_t)^2\end{bmatrix}. \tag{9.5.4}$$

The structural estimates satisfy $\mathbf{B}\hat{\mathbf{\Pi}}' = -\mathbf{\Gamma}$ or

$$\begin{bmatrix}1 & -\beta \\ 1 & -\gamma\end{bmatrix}\begin{bmatrix}\hat{\pi}_1 \\ \hat{\pi}_2\end{bmatrix} = \begin{bmatrix}0 \\ h\end{bmatrix}. \tag{9.5.5}$$

Multiplying [9.5.3] by β and subtracting the result from [9.5.2] produces

$$\begin{aligned} 0 &= \sum_{t=1}^{T} (q_t - \hat{\pi}_1 w_t - \beta p_t + \beta \hat{\pi}_2 w_t) w_t \\ &= \sum_{t=1}^{T} (q_t - \beta p_t) w_t - \sum_{t=1}^{T} (\hat{\pi}_1 - \beta \hat{\pi}_2) w_t^2 \\ &= \sum_{t=1}^{T} (q_t - \beta p_t) w_t, \end{aligned}$$

by virtue of the first row of [9.5.5]. Thus, the estimate of β inferred from the reduced-form parameters is the same as the *IV* or *FIML* estimate derived earlier. Similarly, multiplying [9.5.3] by γ and subtracting the result from [9.5.2] gives

$$\begin{aligned} 0 &= \sum_{t=1}^{T} (q_t - \hat{\pi}_1 w_t - \gamma p_t + \gamma \hat{\pi}_2 w_t) w_t \\ &= \sum_{t=1}^{T} [q_t - \gamma p_t - (\hat{\pi}_1 - \gamma \hat{\pi}_2) w_t] w_t \\ &= \sum_{t=1}^{T} [q_t - \gamma p_t - h w_t] w_t, \end{aligned}$$

by virtue of the second row of [9.5.5], reproducing the first-order condition [9.4.8] for *FIML*. Finally, we need to solve $\mathbf{D} = \mathbf{B}\hat{\mathbf{\Omega}}\mathbf{B}'$ for $\mathbf{D}$ and γ (the remaining element of $\mathbf{B}$). These equations are

$$\begin{aligned} &\begin{bmatrix} \sigma_d^2 & 0 \\ 0 & \sigma_s^2 \end{bmatrix} \\ &= \begin{bmatrix} 1 & -\beta \\ 1 & -\gamma \end{bmatrix} \begin{bmatrix} \hat{\Omega}_{11} & \hat{\Omega}_{12} \\ \hat{\Omega}_{21} & \hat{\Omega}_{22} \end{bmatrix} \begin{bmatrix} 1 & 1 \\ -\beta & -\gamma \end{bmatrix} \\ &= \frac{1}{T} \sum_{t=1}^{T} \left\{ \begin{bmatrix} 1 & -\beta \\ 1 & -\gamma \end{bmatrix} \begin{bmatrix} q_t - \hat{\pi}_1 w_t \\ p_t - \hat{\pi}_2 w_t \end{bmatrix} [q_t - \hat{\pi}_1 w_t \quad p_t - \hat{\pi}_2 w_t] \begin{bmatrix} 1 & 1 \\ -\beta & -\gamma \end{bmatrix} \right\} \\ &= \frac{1}{T} \sum_{t=1}^{T} \left\{ \begin{bmatrix} q_t - \beta p_t - (\hat{\pi}_1 - \beta \hat{\pi}_2) w_t \\ q_t - \gamma p_t - (\hat{\pi}_1 - \gamma \hat{\pi}_2) w_t \end{bmatrix} \begin{bmatrix} q_t - \beta p_t - (\hat{\pi}_1 - \beta \hat{\pi}_2) w_t \\ q_t - \gamma p_t - (\hat{\pi}_1 - \gamma \hat{\pi}_2) w_t \end{bmatrix}' \right\} \\ &= \frac{1}{T} \sum_{t=1}^{T} \left\{ \begin{bmatrix} q_t - \beta p_t \\ q_t - \gamma p_t - h w_t \end{bmatrix} [q_t - \beta p_t \quad q_t - \gamma p_t - h w_t] \right\}. \end{aligned}$$

The diagonal elements of this matrix system of equations reproduce the earlier formulas for the *FIML* estimates of the variance parameters, while the off-diagonal element reproduces the result [9.4.14].

9.6. *Overview of Simultaneous Equations Bias*

The problem of simultaneous equations bias is extremely widespread in the social sciences. It is rare that the relation that we are interested in estimating is the only possible reason why the dependent and explanatory variables might be correlated. For example, consider trying to estimate the effect of military service on an individual's subsequent income. This parameter cannot be estimated by a regression of income on a measure of military service and other observed variables. The error

term in such a regression represents other characteristics of the individual that influence income, and these omitted factors are also likely to have influenced the individual's military participation. As another example, consider trying to estimate the success of long prison sentences in deterring crime. This cannot be estimated by regressing the crime rate in a state on the average prison term in that state, because some states may have adopted stiffer prison sentences in response to higher crime. The error term in the regression, which represents other factors that influence crime, is thus likely also to be correlated with the explanatory variable. Regardless of whether the researcher is interested in the factors that determine military service or prison terms or has any theory about them, simultaneous equations bias must be recognized and dealt with.

Furthermore, it is not enough to find an instrument $\mathbf{x}_t$ that is uncorrelated with the residual u_t. In order to satisfy the rank condition, the instrument $\mathbf{x}_t$ must be correlated with the endogenous explanatory variables $\mathbf{z}_t$. The calculations by Nelson and Startz (1990) suggest that very poor estimates can result if $\mathbf{x}_t$ is only weakly correlated with $\mathbf{z}_t$.

Finding valid instruments is often extremely difficult and requires careful thought and a bit of good luck. For the question about military service, Angrist (1990) found an ingenious instrument for military service based on the institutional details of the draft in the United States during the Vietnam War. The likelihood that an individual was drafted into military service was determined by a lottery based on birthdays. Thus, an individual's birthday during the year would be correlated with military service but presumably uncorrelated with other factors influencing income. Unfortunately, it is unusual to be able to find such a compelling instrument for many questions that one would like to ask of the data.

APPENDIX 9.A. *Proofs of Chapter 9 Proposition*

■ **Proof of Proposition 9.1.** We first show that (a) implies (c). The middle block of [9.3.10] states that

$$\mathbf{y}_{1t} = \mathbf{\Pi}_{11}\mathbf{x}_{1t} + \mathbf{\Pi}_{12}\mathbf{x}_{2t} + \mathbf{v}_{1t}.$$

Hence,

$$\begin{aligned}
\mathbf{M} &= E\left\{\begin{bmatrix}\mathbf{x}_{1t}\\ \mathbf{y}_{1t}\end{bmatrix}[\mathbf{x}_{1t}' \quad \mathbf{x}_{2t}']\right\} \\
&= E\left\{\begin{bmatrix}\mathbf{I}_{m_1} & \mathbf{0}\\ \mathbf{\Pi}_{11} & \mathbf{\Pi}_{12}\end{bmatrix}\begin{bmatrix}\mathbf{x}_{1t}\\ \mathbf{x}_{2t}\end{bmatrix}[\mathbf{x}_{1t}' \quad \mathbf{x}_{2t}'] + \begin{bmatrix}\mathbf{0}\\ \mathbf{v}_{1t}\end{bmatrix}[\mathbf{x}_{1t}' \quad \mathbf{x}_{2t}']\right\} \qquad [9.A.1] \\
&= \begin{bmatrix}\mathbf{I}_{m_1} & \mathbf{0}\\ \mathbf{\Pi}_{11} & \mathbf{\Pi}_{12}\end{bmatrix}E(\mathbf{x}_t\mathbf{x}_t'),
\end{aligned}$$

since $\mathbf{x}_t$ is uncorrelated with $\mathbf{u}_t$ and thus uncorrelated with $\mathbf{v}_t$.

Suppose that the rows of $\mathbf{M}$ are linearly independent. This means that $[\boldsymbol{\lambda}' \quad \boldsymbol{\mu}']\mathbf{M} \neq \mathbf{0}'$ for any $(m_1 \times 1)$ vector $\boldsymbol{\lambda}$ and any $(n_1 \times 1)$ vector $\boldsymbol{\mu}$ that are not both zero. In particular, $[-\boldsymbol{\mu}'\mathbf{\Pi}_{11} \quad \boldsymbol{\mu}']\mathbf{M} \neq \mathbf{0}'$. But from the right side of [9.A.1], this implies that

$$[-\boldsymbol{\mu}'\mathbf{\Pi}_{11} \quad \boldsymbol{\mu}']\begin{bmatrix}\mathbf{I}_{m_1} & \mathbf{0}\\ \mathbf{\Pi}_{11} & \mathbf{\Pi}_{12}\end{bmatrix}E(\mathbf{x}_t\mathbf{x}_t') = [\mathbf{0}' \quad \boldsymbol{\mu}'\mathbf{\Pi}_{12}]E(\mathbf{x}_t\mathbf{x}_t') \neq \mathbf{0}'$$

for any nonzero $(n_1 \times 1)$ vector $\boldsymbol{\mu}$. But this could be true only if $\boldsymbol{\mu}'\mathbf{\Pi}_{12} \neq \mathbf{0}'$. Hence, if the rows of $\mathbf{M}$ are linearly independent, then the rows of $\mathbf{\Pi}_{12}$ are also linearly independent.

To prove that (c) implies (a), premultiply both sides of [9.A.1] by any nonzero vector $[\boldsymbol{\lambda}' \quad \boldsymbol{\mu}']$. The right side becomes

$$[\boldsymbol{\lambda}' \quad \boldsymbol{\mu}'] \begin{bmatrix} \mathbf{I}_{m_1} & \mathbf{0} \\ \boldsymbol{\Pi}_{11} & \boldsymbol{\Pi}_{12} \end{bmatrix} E(\mathbf{x}_t\mathbf{x}_t') = [(\boldsymbol{\lambda}' + \boldsymbol{\mu}'\boldsymbol{\Pi}_{11}) \quad \boldsymbol{\mu}'\boldsymbol{\Pi}_{12}]E(\mathbf{x}_t\mathbf{x}_t') \equiv \boldsymbol{\eta}'E(\mathbf{x}_t\mathbf{x}_t'),$$

where $\boldsymbol{\eta}' \equiv [(\boldsymbol{\lambda}' + \boldsymbol{\mu}'\boldsymbol{\Pi}_{11}) \quad \boldsymbol{\mu}'\boldsymbol{\Pi}_{12}]$. If the rows of $\boldsymbol{\Pi}_{12}$ are linearly independent, then $\boldsymbol{\eta}'$ cannot be the zero vector unless both $\boldsymbol{\mu}$ and $\boldsymbol{\lambda}$ are zero. To see this, note that if $\boldsymbol{\mu}$ is nonzero, then $\boldsymbol{\mu}'\boldsymbol{\Pi}_{12}$ cannot be the zero vector, while if $\boldsymbol{\mu} = \mathbf{0}$, then $\boldsymbol{\eta}$ will be zero only if $\boldsymbol{\lambda}$ is also the zero vector. Furthermore, since $E(\mathbf{x}_t\mathbf{x}_t')$ is nonsingular, a nonzero $\boldsymbol{\eta}$ means that $\boldsymbol{\eta}'E(\mathbf{x}_t\mathbf{x}_t')$ cannot be the zero vector. Thus, if the right side of [9.A.1] is premultiplied by any nonzero vector $(\boldsymbol{\lambda}', \boldsymbol{\mu}')$, the result is not zero. The same must be true of the left side: $[\boldsymbol{\lambda}' \quad \boldsymbol{\mu}']\mathbf{M} \neq \mathbf{0}'$ for any nonzero $(\boldsymbol{\lambda}', \boldsymbol{\mu}')$, establishing that linear independence of the rows of $\boldsymbol{\Pi}_{12}$ implies linear independence of the rows of $\mathbf{M}$.

To see that (b) implies (c), write [9.3.7] as

$$\begin{bmatrix} \boldsymbol{\Pi}_{01} & \boldsymbol{\Pi}_{02} \\ \boldsymbol{\Pi}_{11} & \boldsymbol{\Pi}_{12} \\ \boldsymbol{\Pi}_{21} & \boldsymbol{\Pi}_{22} \end{bmatrix} = -\mathbf{B}^{-1} \begin{bmatrix} \boldsymbol{\Gamma}_{01} & \mathbf{0}' \\ \boldsymbol{\Gamma}_{11} & \boldsymbol{\Gamma}_{12} \\ \boldsymbol{\Gamma}_{21} & \boldsymbol{\Gamma}_{22} \end{bmatrix}. \qquad [9.A.2]$$

We also have the identity

$$\begin{bmatrix} 1 & \mathbf{0}' & \mathbf{0}' \\ \mathbf{0} & \mathbf{I}_{n_1} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \mathbf{I}_{n_2} \end{bmatrix} = \mathbf{B}^{-1} \begin{bmatrix} 1 & \mathbf{B}_{01} & \mathbf{0}' \\ \mathbf{B}_{10} & \mathbf{B}_{11} & \mathbf{B}_{12} \\ \mathbf{B}_{20} & \mathbf{B}_{21} & \mathbf{B}_{22} \end{bmatrix}. \qquad [9.A.3]$$

The system of equations represented by the second block column of [9.A.2] and the third block column of [9.A.3] can be collected as

$$\begin{bmatrix} \boldsymbol{\Pi}_{02} & \mathbf{0}' \\ \boldsymbol{\Pi}_{12} & \mathbf{0} \\ \boldsymbol{\Pi}_{22} & \mathbf{I}_{n_2} \end{bmatrix} = \mathbf{B}^{-1} \begin{bmatrix} \mathbf{0}' & \mathbf{0}' \\ -\boldsymbol{\Gamma}_{12} & \mathbf{B}_{12} \\ -\boldsymbol{\Gamma}_{22} & \mathbf{B}_{22} \end{bmatrix}. \qquad [9.A.4]$$

If both sides of [9.A.4] are premultiplied by the row vector $[0 \quad \boldsymbol{\mu}_1' \quad \mathbf{0}']$ where $\boldsymbol{\mu}_1$ is any $(n_1 \times 1)$ vector, the result is

$$\begin{aligned} [\boldsymbol{\mu}_1'\boldsymbol{\Pi}_{12} \quad \mathbf{0}'] &= [0 \quad \boldsymbol{\mu}_1' \quad \mathbf{0}']\mathbf{B}^{-1} \begin{bmatrix} \mathbf{0}' & \mathbf{0}' \\ -\boldsymbol{\Gamma}_{12} & \mathbf{B}_{12} \\ -\boldsymbol{\Gamma}_{22} & \mathbf{B}_{22} \end{bmatrix} \\ &= [\lambda_0 \quad \boldsymbol{\lambda}_1' \quad \boldsymbol{\lambda}_2'] \begin{bmatrix} \mathbf{0}' & \mathbf{0}' \\ -\boldsymbol{\Gamma}_{12} & \mathbf{B}_{12} \\ -\boldsymbol{\Gamma}_{22} & \mathbf{B}_{22} \end{bmatrix} \qquad [9.A.5] \\ &= [\boldsymbol{\lambda}_1' \quad \boldsymbol{\lambda}_2'] \begin{bmatrix} -\boldsymbol{\Gamma}_{12} & \mathbf{B}_{12} \\ -\boldsymbol{\Gamma}_{22} & \mathbf{B}_{22} \end{bmatrix}, \end{aligned}$$

where

$$[\lambda_0 \quad \boldsymbol{\lambda}_1' \quad \boldsymbol{\lambda}_2'] \equiv [0 \quad \boldsymbol{\mu}_1' \quad \mathbf{0}']\mathbf{B}^{-1},$$

implying

$$[0 \quad \boldsymbol{\mu}_1' \quad \mathbf{0}'] = [\lambda_0 \quad \boldsymbol{\lambda}_1' \quad \boldsymbol{\lambda}_2']\mathbf{B}. \qquad [9.A.6]$$

Suppose that the rows of the matrix $\begin{bmatrix} \boldsymbol{\Gamma}_{12} & \mathbf{B}_{12} \\ \boldsymbol{\Gamma}_{22} & \mathbf{B}_{22} \end{bmatrix}$ are linearly independent. Then the only values for $\boldsymbol{\lambda}_1$ and $\boldsymbol{\lambda}_2$ for which the right side of [9.A.5] can be zero are $\boldsymbol{\lambda}_1 = \mathbf{0}$ and $\boldsymbol{\lambda}_2 = \mathbf{0}$. Substituting these values into [9.A.6], the only value of $\boldsymbol{\mu}_1$ for which the left side of [9.A.5] can be zero must satisfy

$$\begin{aligned} [0 \quad \boldsymbol{\mu}_1' \quad \mathbf{0}'] &= [\lambda_0 \quad \mathbf{0}' \quad \mathbf{0}']\mathbf{B} \\ &= [\lambda_0 \quad \lambda_0\mathbf{B}_{01} \quad \mathbf{0}']. \end{aligned}$$

Matching the first elements in these vectors implies $\lambda_0 = 0$, and thus matching the second elements requires $\boldsymbol{\mu}_1 = \mathbf{0}$. Thus, if condition (b) is satisfied, then the only value of $\boldsymbol{\mu}_1$ for

which the left side of [9.A.5] can be zero is $\boldsymbol{\mu}_1 = \mathbf{0}$, establishing that the rows of $\boldsymbol{\Pi}_{12}$ are linearly independent. Hence, condition (c) is satisfied whenever (b) holds.

Conversely, to see that (c) implies (b), let $\boldsymbol{\lambda}_1$ and $\boldsymbol{\lambda}_2$ denote any $(n_1 \times 1)$ and $(n_2 \times 1)$ vectors, and premultiply both sides of [9.A.4] by the row vector $[0 \quad \boldsymbol{\lambda}_1' \quad \boldsymbol{\lambda}_2']\mathbf{B}$:

$$[0 \quad \boldsymbol{\lambda}_1' \quad \boldsymbol{\lambda}_2']\mathbf{B}\begin{bmatrix}\boldsymbol{\Pi}_{02} & \mathbf{0}' \\ \boldsymbol{\Pi}_{12} & \mathbf{0} \\ \boldsymbol{\Pi}_{22} & \mathbf{I}_{n_2}\end{bmatrix} = [0 \quad \boldsymbol{\lambda}_1' \quad \boldsymbol{\lambda}_2']\begin{bmatrix}\mathbf{0}' & \mathbf{0}' \\ -\boldsymbol{\Gamma}_{12} & \mathbf{B}_{12} \\ -\boldsymbol{\Gamma}_{22} & \mathbf{B}_{22}\end{bmatrix}$$

or

$$[\mu_0 \quad \boldsymbol{\mu}_1' \quad \boldsymbol{\mu}_2']\begin{bmatrix}\boldsymbol{\Pi}_{02} & \mathbf{0}' \\ \boldsymbol{\Pi}_{12} & \mathbf{0} \\ \boldsymbol{\Pi}_{22} & \mathbf{I}_{n_2}\end{bmatrix} = [\boldsymbol{\lambda}_1' \quad \boldsymbol{\lambda}_2']\begin{bmatrix}-\boldsymbol{\Gamma}_{12} & \mathbf{B}_{12} \\ -\boldsymbol{\Gamma}_{22} & \mathbf{B}_{22}\end{bmatrix} \qquad [9.A.7]$$

where

$$[\mu_0 \quad \boldsymbol{\mu}_1' \quad \boldsymbol{\mu}_2'] \equiv [0 \quad \boldsymbol{\lambda}_1' \quad \boldsymbol{\lambda}_2']\mathbf{B}. \qquad [9.A.8]$$

Premultiplying both sides of equation [9.A.4] by $\mathbf{B}$ implies that

$$\begin{bmatrix}1 & \mathbf{B}_{01} & \mathbf{0}' \\ \mathbf{B}_{10} & \mathbf{B}_{11} & \mathbf{B}_{12} \\ \mathbf{B}_{20} & \mathbf{B}_{21} & \mathbf{B}_{22}\end{bmatrix}\begin{bmatrix}\boldsymbol{\Pi}_{02} & \mathbf{0}' \\ \boldsymbol{\Pi}_{12} & \mathbf{0} \\ \boldsymbol{\Pi}_{22} & \mathbf{I}_{n_2}\end{bmatrix} = \begin{bmatrix}\mathbf{0}' & \mathbf{0}' \\ -\boldsymbol{\Gamma}_{12} & \mathbf{B}_{12} \\ -\boldsymbol{\Gamma}_{22} & \mathbf{B}_{22}\end{bmatrix}.$$

The upper left element of this matrix system asserts that

$$\boldsymbol{\Pi}_{02} + \mathbf{B}_{01}\boldsymbol{\Pi}_{12} = \mathbf{0}'. \qquad [9.A.9]$$

Substituting [9.A.9] into [9.A.7],

$$[\mu_0 \quad \boldsymbol{\mu}_1' \quad \boldsymbol{\mu}_2']\begin{bmatrix}-\mathbf{B}_{01}\boldsymbol{\Pi}_{12} & \mathbf{0}' \\ \boldsymbol{\Pi}_{12} & \mathbf{0} \\ \boldsymbol{\Pi}_{22} & \mathbf{I}_{n_2}\end{bmatrix} = [\boldsymbol{\lambda}_1' \quad \boldsymbol{\lambda}_2']\begin{bmatrix}-\boldsymbol{\Gamma}_{12} & \mathbf{B}_{12} \\ -\boldsymbol{\Gamma}_{22} & \mathbf{B}_{22}\end{bmatrix}. \qquad [9.A.10]$$

In order for the left side of [9.A.10] to be zero, it must be the case that $\boldsymbol{\mu}_2 = \mathbf{0}$ and that

$$-\mu_0\mathbf{B}_{01}\boldsymbol{\Pi}_{12} + \boldsymbol{\mu}_1'\boldsymbol{\Pi}_{12} = (\boldsymbol{\mu}_1' - \mu_0\mathbf{B}_{01})\boldsymbol{\Pi}_{12} = \mathbf{0}'. \qquad [9.A.11]$$

But if the rows of $\boldsymbol{\Pi}_{12}$ are linearly independent, [9.A.11] can be zero only if

$$\boldsymbol{\mu}_1' = \mu_0\mathbf{B}_{01}. \qquad [9.A.12]$$

Substituting these results into [9.A.8], it follows that [9.A.10] can be zero only if

$$\begin{aligned}[0 \quad \boldsymbol{\lambda}_1' \quad \boldsymbol{\lambda}_2']\mathbf{B} &= [\mu_0 \quad \mu_0\mathbf{B}_{01} \quad \mathbf{0}'] \\ &= [\mu_0 \quad \mathbf{0}' \quad \mathbf{0}']\begin{bmatrix}1 & \mathbf{B}_{01} & \mathbf{0}' \\ \mathbf{B}_{10} & \mathbf{B}_{11} & \mathbf{B}_{12} \\ \mathbf{B}_{20} & \mathbf{B}_{21} & \mathbf{B}_{22}\end{bmatrix} \qquad [9.A.13] \\ &= [\mu_0 \quad \mathbf{0}' \quad \mathbf{0}']\mathbf{B}.\end{aligned}$$

Since $\mathbf{B}$ is nonsingular, both sides of [9.A.13] can be postmultiplied by $\mathbf{B}^{-1}$ to deduce that [9.A.10] can be zero only if

$$[0 \quad \boldsymbol{\lambda}_1' \quad \boldsymbol{\lambda}_2'] = [\mu_0 \quad \mathbf{0}' \quad \mathbf{0}'].$$

Thus, the right side of [9.A.10] can be zero only if $\boldsymbol{\lambda}_1$ and $\boldsymbol{\lambda}_2$ are both zero, establishing that the rows of the matrix in [9.3.12] must be linearly independent. ■

Chapter 9 Exercise

9.1. Verify that [9.2.23] gives a consistent estimate of σ^2.

Chapter 9 References

Angrist, Joshua D. 1990. "Lifetime Earnings and the Vietnam Era Draft Lottery: Evidence from Social Security Administration Records." *American Economic Review* 80:313–36. Errata, 1990, 80:1284–86.

Fisher, Franklin M. 1966. *The Identification Problem in Econometrics*. New York: McGraw-Hill.

Hausman, Jerry A., and William E. Taylor. 1983. "Identification in Linear Simultaneous Equations Models with Covariance Restrictions: An Instrumental Variables Interpretation." *Econometrica* 51:1527–49.

Maddala, G. S. 1977. *Econometrics*. New York: McGraw-Hill.

Nelson, Charles R., and Richard Startz. 1990. "Some Further Results on the Exact Small Sample Properties of the Instrumental Variable Estimator." *Econometrica* 58:967–76.

Rothenberg, Thomas J. 1971. "Identification in Parametric Models." *Econometrica* 39:577–91.

——— and Paul A. Ruud. 1990. "Simultaneous Equations with Covariance Restrictions." *Journal of Econometrics* 44:25–39.

10 Covariance-Stationary Vector Processes

This is the first of two chapters introducing vector time series. Chapter 10 is devoted to the theory of multivariate dynamic systems, while Chapter 11 focuses on empirical issues of estimating and interpreting vector autoregressions. Only the first section of Chapter 10 is necessary for understanding the material in Chapter 11.

Section 10.1 introduces some of the key ideas in vector time series analysis. Section 10.2 develops some convergence results that are useful for deriving the asymptotic properties of certain statistics and for characterizing the consequences of multivariate filters. Section 10.3 introduces the autocovariance-generating function for vector processes, which is used to analyze the multivariate spectrum in Section 10.4. Section 10.5 develops a multivariate generalization of Proposition 7.5, describing the asymptotic properties of the sample mean of a serially correlated vector process. These last results are useful for deriving autocorrelation- and heteroskedasticity-consistent estimators for *OLS*, for understanding the properties of generalized method of moments estimators discussed in Chapter 14, and for deriving some of the tests for unit roots discussed in Chapter 17.

10.1. *Introduction to Vector Autoregressions*

Chapter 3 proposed modeling a scalar time series y_t in terms of an autoregression:

$$y_t = c + \phi_1 y_{t-1} + \phi_2 y_{t-2} + \cdots + \phi_p y_{t-p} + \varepsilon_t, \qquad [10.1.1]$$

where

$$E(\varepsilon_t) = 0 \qquad [10.1.2]$$

$$E(\varepsilon_t \varepsilon_\tau) = \begin{cases} \sigma^2 & \text{for } t = \tau \\ 0 & \text{otherwise.} \end{cases} \qquad [10.1.3]$$

Note that we will continue to use the convention introduced in Chapter 8 of using lowercase letters to denote either a random variable or its realization. This chapter describes the dynamic interactions among a set of variables collected in an $(n \times 1)$ vector $\mathbf{y}_t$. For example, the first element of $\mathbf{y}_t$ (denoted y_{1t}) might represent the level of GNP in year t, the second element (y_{2t}) the interest rate paid on Treasury bills in year t, and so on. A *pth-order vector autoregression*, denoted $VAR(p)$, is a vector generalization of [10.1.1] through [10.1.3]:

$$\mathbf{y}_t = \mathbf{c} + \mathbf{\Phi}_1 \mathbf{y}_{t-1} + \mathbf{\Phi}_2 \mathbf{y}_{t-2} + \cdots + \mathbf{\Phi}_p \mathbf{y}_{t-p} + \boldsymbol{\varepsilon}_t. \qquad [10.1.4]$$

Here $\mathbf{c}$ denotes an $(n \times 1)$ vector of constants and $\mathbf{\Phi}_j$ an $(n \times n)$ matrix of autoregressive coefficients for $j = 1, 2, \ldots, p$. The $(n \times 1)$ vector $\boldsymbol{\varepsilon}_t$ is a vector

generalization of white noise:

$$E(\boldsymbol{\varepsilon}_t) = \mathbf{0} \qquad [10.1.5]$$

$$E(\boldsymbol{\varepsilon}_t\boldsymbol{\varepsilon}_\tau') = \begin{cases} \boldsymbol{\Omega} & \text{for } t = \tau \\ \mathbf{0} & \text{otherwise,} \end{cases} \qquad [10.1.6]$$

with $\boldsymbol{\Omega}$ an $(n \times n)$ symmetric positive definite matrix.

Let c_i denote the ith element of the vector $\mathbf{c}$ and let $\phi_{ij}^{(1)}$ denote the row i, column j element of the matrix $\boldsymbol{\Phi}_1$. Then the first row of the vector system in [10.1.4] specifies that

$$\begin{aligned} y_{1t} = c_1 &+ \phi_{11}^{(1)}y_{1,t-1} + \phi_{12}^{(1)}y_{2,t-1} + \cdots + \phi_{1n}^{(1)}y_{n,t-1} \\ &+ \phi_{11}^{(2)}y_{1,t-2} + \phi_{12}^{(2)}y_{2,t-2} + \cdots + \phi_{1n}^{(2)}y_{n,t-2} \\ &+ \cdots + \phi_{11}^{(p)}y_{1,t-p} + \phi_{12}^{(p)}y_{2,t-p} + \cdots + \phi_{1n}^{(p)}y_{n,t-p} + \varepsilon_{1t}. \end{aligned} \qquad [10.1.7]$$

Thus, a vector autoregression is a system in which each variable is regressed on a constant and p of its own lags as well as on p lags of each of the other variables in the *VAR*. Note that each regression has the same explanatory variables.

Using lag operator notation, [10.1.4] can be written in the form

$$[\mathbf{I}_n - \boldsymbol{\Phi}_1 L - \boldsymbol{\Phi}_2 L^2 - \cdots - \boldsymbol{\Phi}_p L^p]\mathbf{y}_t = \mathbf{c} + \boldsymbol{\varepsilon}_t$$

or

$$\boldsymbol{\Phi}(L)\mathbf{y}_t = \mathbf{c} + \boldsymbol{\varepsilon}_t.$$

Here $\boldsymbol{\Phi}(L)$ indicates an $(n \times n)$ matrix polynomial in the lag operator L. The row i, column j element of $\boldsymbol{\Phi}(L)$ is a scalar polynomial in L:

$$\boldsymbol{\Phi}(L) = [\delta_{ij} - \phi_{ij}^{(1)}L^1 - \phi_{ij}^{(2)}L^2 - \cdots - \phi_{ij}^{(p)}L^p],$$

where δ_{ij} is unity if $i = j$ and zero otherwise.

A vector process $\mathbf{y}_t$ is said to be covariance-stationary if its first and second moments ($E[\mathbf{y}_t]$ and $E[\mathbf{y}_t\mathbf{y}_{t-j}']$, respectively) are independent of the date t. If the process is covariance-stationary, we can take expectations of both sides of [10.1.4] to calculate the mean $\boldsymbol{\mu}$ of the process:

$$\boldsymbol{\mu} = \mathbf{c} + \boldsymbol{\Phi}_1\boldsymbol{\mu} + \boldsymbol{\Phi}_2\boldsymbol{\mu} + \cdots + \boldsymbol{\Phi}_p\boldsymbol{\mu},$$

or

$$\boldsymbol{\mu} = (\mathbf{I}_n - \boldsymbol{\Phi}_1 - \boldsymbol{\Phi}_2 - \cdots - \boldsymbol{\Phi}_p)^{-1}\mathbf{c}.$$

Equation [10.1.4] can then be written in terms of deviations from the mean as

$$\begin{aligned} (\mathbf{y}_t - \boldsymbol{\mu}) &= \boldsymbol{\Phi}_1(\mathbf{y}_{t-1} - \boldsymbol{\mu}) \\ &\quad + \boldsymbol{\Phi}_2(\mathbf{y}_{t-2} - \boldsymbol{\mu}) + \cdots + \boldsymbol{\Phi}_p(\mathbf{y}_{t-p} - \boldsymbol{\mu}) + \boldsymbol{\varepsilon}_t. \end{aligned} \qquad [10.1.8]$$

Rewriting a VAR(p) *as a* VAR(*1*)

As in the case of the univariate $AR(p)$ process, it is helpful to rewrite [10.1.8] in terms of a $VAR(1)$ process. Toward this end, define

$$\underset{(np \times 1)}{\boldsymbol{\xi}_t} \equiv \begin{bmatrix} \mathbf{y}_t - \boldsymbol{\mu} \\ \mathbf{y}_{t-1} - \boldsymbol{\mu} \\ \vdots \\ \mathbf{y}_{t-p+1} - \boldsymbol{\mu} \end{bmatrix} \qquad [10.1.9]$$

$$\underset{(np \times np)}{\mathbf{F}} \equiv \begin{bmatrix} \boldsymbol{\Phi}_1 & \boldsymbol{\Phi}_2 & \boldsymbol{\Phi}_3 & \cdots & \boldsymbol{\Phi}_{p-1} & \boldsymbol{\Phi}_p \\ \mathbf{I}_n & \mathbf{0} & \mathbf{0} & \cdots & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{I}_n & \mathbf{0} & \cdots & \mathbf{0} & \mathbf{0} \\ \vdots & \vdots & \vdots & \cdots & \vdots & \vdots \\ \mathbf{0} & \mathbf{0} & \mathbf{0} & \cdots & \mathbf{I}_n & \mathbf{0} \end{bmatrix} \qquad [10.1.10]$$

$$\underset{(np \times 1)}{\mathbf{v}_t} \equiv \begin{bmatrix} \boldsymbol{\varepsilon}_t \\ \mathbf{0} \\ \vdots \\ \mathbf{0} \end{bmatrix}.$$

The $VAR(p)$ in [10.1.8] can then be rewritten as the following $VAR(1)$:

$$\boldsymbol{\xi}_t = \mathbf{F}\boldsymbol{\xi}_{t-1} + \mathbf{v}_t, \qquad [10.1.11]$$

where

$$E(\mathbf{v}_t\mathbf{v}_\tau') = \begin{cases} \mathbf{Q} & \text{for } t = \tau \\ \mathbf{0} & \text{otherwise} \end{cases}$$

and

$$\underset{(np \times np)}{\mathbf{Q}} \equiv \begin{bmatrix} \boldsymbol{\Omega} & \mathbf{0} & \cdots & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \cdots & \mathbf{0} \\ \vdots & \vdots & \cdots & \vdots \\ \mathbf{0} & \mathbf{0} & \cdots & \mathbf{0} \end{bmatrix}$$

Conditions for Stationarity

Equation [10.1.11] implies that

$$\boldsymbol{\xi}_{t+s} = \mathbf{v}_{t+s} + \mathbf{F}\mathbf{v}_{t+s-1} + \mathbf{F}^2\mathbf{v}_{t+s-2} + \cdots + \mathbf{F}^{s-1}\mathbf{v}_{t+1} + \mathbf{F}^s\boldsymbol{\xi}_t. \qquad [10.1.12]$$

In order for the process to be covariance-stationary, the consequences of any given $\boldsymbol{\varepsilon}_t$ must eventually die out. If the eigenvalues of $\mathbf{F}$ all lie inside the unit circle, then the VAR turns out to be covariance-stationary.

The following result generalizes Proposition 1.1 from Chapter 1 (for a proof see Appendix 10.A at the end of this chapter).

Proposition 10.1: *The eigenvalues of the matrix* $\mathbf{F}$ *in [10.1.10] satisfy*

$$|\mathbf{I}_n\lambda^p - \boldsymbol{\Phi}_1\lambda^{p-1} - \boldsymbol{\Phi}_2\lambda^{p-2} - \cdots - \boldsymbol{\Phi}_p| = 0. \qquad [10.1.13]$$

Hence, a $VAR(p)$ is covariance-stationary as long as $|\lambda| < 1$ for all values of λ satisfying [10.1.13]. Equivalently, the VAR is covariance-stationary if all values of z satisfying

$$|\mathbf{I}_n - \boldsymbol{\Phi}_1 z - \boldsymbol{\Phi}_2 z^2 - \cdots - \boldsymbol{\Phi}_p z^p| = 0$$

lie outside the unit circle.

Vector MA(∞) *Representation*

The first n rows of the vector system represented in [10.1.12] constitute a vector generalization of equation [4.2.20]:

$$\begin{aligned}\mathbf{y}_{t+s} = \boldsymbol{\mu} + \boldsymbol{\varepsilon}_{t+s} + \boldsymbol{\Psi}_1\boldsymbol{\varepsilon}_{t+s-1} + \boldsymbol{\Psi}_2\boldsymbol{\varepsilon}_{t+s-2} + \cdots + \boldsymbol{\Psi}_{s-1}\boldsymbol{\varepsilon}_{t+1} \\ + \mathbf{F}_{11}^{(s)}(\mathbf{y}_t - \boldsymbol{\mu}) + \mathbf{F}_{12}^{(s)}(\mathbf{y}_{t-1} - \boldsymbol{\mu}) + \cdots + \mathbf{F}_{1p}^{(s)}(\mathbf{y}_{t-p+1} - \boldsymbol{\mu}).\end{aligned} \qquad [10.1.14]$$

Here $\boldsymbol{\Psi}_j = \mathbf{F}_{11}^{(j)}$ and $\mathbf{F}_{11}^{(j)}$ denotes the upper left block of $\mathbf{F}^j$, where $\mathbf{F}^j$ is the matrix $\mathbf{F}$ raised to the jth power—that is, the $(n \times n)$ matrix $\mathbf{F}_{11}^{(j)}$ indicates rows 1 through n and columns 1 through n of the $(np \times np)$ matrix $\mathbf{F}^j$. Similarly, $\mathbf{F}_{12}^{(j)}$ denotes the block of $\mathbf{F}^j$ consisting of rows 1 through n and columns $(n + 1)$ through $2n$, while $\mathbf{F}_{1p}^{(j)}$ denotes rows 1 through n and columns $[n(p - 1) + 1]$ through np of $\mathbf{F}^j$.

If the eigenvalues of $\mathbf{F}$ all lie inside the unit circle, then $\mathbf{F}^s \to \mathbf{0}$ as $s \to \infty$ and $\mathbf{y}_t$ can be expressed as a convergent sum of the history of $\boldsymbol{\varepsilon}$:

$$\mathbf{y}_t = \boldsymbol{\mu} + \boldsymbol{\varepsilon}_t + \boldsymbol{\Psi}_1\boldsymbol{\varepsilon}_{t-1} + \boldsymbol{\Psi}_2\boldsymbol{\varepsilon}_{t-2} + \boldsymbol{\Psi}_3\boldsymbol{\varepsilon}_{t-3} + \cdots \equiv \boldsymbol{\mu} + \boldsymbol{\Psi}(L)\boldsymbol{\varepsilon}_t, \qquad [10.1.15]$$

which is a vector $MA(\infty)$ representation.

Note that $\mathbf{y}_{t-j}$ is a linear function of $\boldsymbol{\varepsilon}_{t-j}, \boldsymbol{\varepsilon}_{t-j-1}, \ldots$, each of which is uncorrelated with $\boldsymbol{\varepsilon}_{t+1}$ for $j = 0, 1, \ldots$. It follows that $\boldsymbol{\varepsilon}_{t+1}$ is uncorrelated with $\mathbf{y}_{t-j}$ for any $j \geq 0$. Thus, the linear forecast of $\mathbf{y}_{t+1}$ on the basis of $\mathbf{y}_t, \mathbf{y}_{t-1}, \ldots$ is given by

$$\hat{\mathbf{y}}_{t+1|t} = \boldsymbol{\mu} + \boldsymbol{\Phi}_1(\mathbf{y}_t - \boldsymbol{\mu}) + \boldsymbol{\Phi}_2(\mathbf{y}_{t-1} - \boldsymbol{\mu}) + \cdots + \boldsymbol{\Phi}_p(\mathbf{y}_{t-p+1} - \boldsymbol{\mu}),$$

and $\boldsymbol{\varepsilon}_{t+1}$ can be interpreted as the fundamental innovation for $\mathbf{y}_{t+1}$, that is, the error in forecasting $\mathbf{y}_{t+1}$ on the basis of a linear function of a constant and $\mathbf{y}_t, \mathbf{y}_{t-1}, \ldots$. More generally, it follows from [10.1.14] that a forecast of $\mathbf{y}_{t+s}$ on the basis of $\mathbf{y}_t, \mathbf{y}_{t-1}, \ldots$ will take the form

$$\begin{aligned}\hat{\mathbf{y}}_{t+s|t} = \boldsymbol{\mu} + \mathbf{F}_{11}^{(s)}(\mathbf{y}_t - \boldsymbol{\mu}) + \mathbf{F}_{12}^{(s)}(\mathbf{y}_{t-1} - \boldsymbol{\mu}) \\ + \cdots + \mathbf{F}_{1p}^{(s)}(\mathbf{y}_{t-p+1} - \boldsymbol{\mu}).\end{aligned} \qquad [10.1.16]$$

The moving average matrices $\boldsymbol{\Psi}_j$ could equivalently be calculated as follows. The operators $\boldsymbol{\Phi}(L)$ and $\boldsymbol{\Psi}(L)$ are related by

$$\boldsymbol{\Psi}(L) = [\boldsymbol{\Phi}(L)]^{-1},$$

requiring

$$[\mathbf{I}_n - \boldsymbol{\Phi}_1 L - \boldsymbol{\Phi}_2 L^2 - \cdots - \boldsymbol{\Phi}_p L^p][\mathbf{I}_n + \boldsymbol{\Psi}_1 L + \boldsymbol{\Psi}_2 L^2 + \cdots] = \mathbf{I}_n.$$

Setting the coefficient on L^1 equal to the zero matrix, as in Exercise 3.3 of Chapter 3, produces

$$\boldsymbol{\Psi}_1 - \boldsymbol{\Phi}_1 = \mathbf{0}. \qquad [10.1.17]$$

Similarly, setting the coefficient on L^2 equal to zero gives

$$\boldsymbol{\Psi}_2 = \boldsymbol{\Phi}_1\boldsymbol{\Psi}_1 + \boldsymbol{\Phi}_2, \qquad [10.1.18]$$

and in general for L^s,

$$\boldsymbol{\Psi}_s = \boldsymbol{\Phi}_1\boldsymbol{\Psi}_{s-1} + \boldsymbol{\Phi}_2\boldsymbol{\Psi}_{s-2} + \cdots + \boldsymbol{\Phi}_p\boldsymbol{\Psi}_{s-p} \qquad \text{for } s = 1, 2, \ldots, \qquad [10.1.19]$$

with $\boldsymbol{\Psi}_0 = \mathbf{I}_n$ and $\boldsymbol{\Psi}_s = \mathbf{0}$ for $s < 0$.

Note that the innovation in the $MA(\infty)$ representation [10.1.15] is $\boldsymbol{\varepsilon}_t$, the fundamental innovation for $\mathbf{y}$. There are alternative moving average representations based on vector white noise processes other than $\boldsymbol{\varepsilon}_t$. Let $\mathbf{H}$ denote a nonsingular

$(n \times n)$ matrix, and define

$$\mathbf{u}_t \equiv \mathbf{H}\boldsymbol{\varepsilon}_t. \qquad [10.1.20]$$

Then certainly $\mathbf{u}_t$ is white noise. Moreover, from [10.1.15] we could write

$$\begin{aligned} \mathbf{y}_t &= \boldsymbol{\mu} + \mathbf{H}^{-1}\mathbf{H}\boldsymbol{\varepsilon}_t + \boldsymbol{\Psi}_1\mathbf{H}^{-1}\mathbf{H}\boldsymbol{\varepsilon}_{t-1} + \boldsymbol{\Psi}_2\mathbf{H}^{-1}\mathbf{H}\boldsymbol{\varepsilon}_{t-2} \\ &\quad + \boldsymbol{\Psi}_3\mathbf{H}^{-1}\mathbf{H}\boldsymbol{\varepsilon}_{t-3} + \cdots \qquad [10.1.21] \\ &= \boldsymbol{\mu} + \mathbf{J}_0\mathbf{u}_t + \mathbf{J}_1\mathbf{u}_{t-1} + \mathbf{J}_2\mathbf{u}_{t-2} + \mathbf{J}_3\mathbf{u}_{t-3} + \cdots, \end{aligned}$$

where

$$\mathbf{J}_s \equiv \boldsymbol{\Psi}_s\mathbf{H}^{-1}.$$

For example, $\mathbf{H}$ could be any matrix that diagonalizes $\boldsymbol{\Omega}$, the variance-covariance matrix of $\boldsymbol{\varepsilon}_t$:

$$\mathbf{H}\boldsymbol{\Omega}\mathbf{H}' = \mathbf{D},$$

with $\mathbf{D}$ a diagonal matrix. For such a choice of $\mathbf{H}$, the elements of $\mathbf{u}_t$ are uncorrelated with one another:

$$E(\mathbf{u}_t\mathbf{u}_t') = E(\mathbf{H}\boldsymbol{\varepsilon}_t\boldsymbol{\varepsilon}_t'\mathbf{H}') = \mathbf{D}.$$

Thus, it is always possible to write a stationary $VAR(p)$ process as a convergent infinite moving average of a white noise vector $\mathbf{u}_t$ whose elements are mutually uncorrelated.

There is one important difference between the $MA(\infty)$ representations [10.1.15] and [10.1.21], however. In [10.1.15], the leading MA parameter matrix ($\boldsymbol{\Psi}_0$) is the identity matrix, whereas in [10.1.21] the leading MA parameter matrix ($\mathbf{J}_0$) is not the identity matrix. To obtain the MA representation for the fundamental innovations, we must impose the normalization $\boldsymbol{\Psi}_0 = \mathbf{I}_n$.

Assumptions Implicit in a VAR

For a covariance-stationary process, the parameters $\mathbf{c}$ and $\boldsymbol{\Phi}_1, \ldots, \boldsymbol{\Phi}_p$ in equation [10.1.4] could be defined as the coefficients of the projection of $\mathbf{y}_t$ on a constant and $\mathbf{y}_{t-1}, \ldots, \mathbf{y}_{t-p}$. Thus, $\boldsymbol{\varepsilon}_t$ is uncorrelated with $\mathbf{y}_{t-1}, \ldots, \mathbf{y}_{t-p}$ by the definition of $\boldsymbol{\Phi}_1, \ldots, \boldsymbol{\Phi}_p$. The parameters of a vector autoregression can accordingly be estimated consistently with n *OLS* regressions of the form of [10.1.7]. The additional assumption implicit in a *VAR* is that the $\boldsymbol{\varepsilon}_t$ defined by this projection is further uncorrelated with $\mathbf{y}_{t-p-1}, \mathbf{y}_{t-p-2}, \ldots$. The assumption that $\mathbf{y}_t$ follows a vector autoregression is basically the assumption that p lags are sufficient to summarize all of the dynamic correlations between elements of $\mathbf{y}$.

10.2. *Autocovariances and Convergence Results for Vector Processes*

The j*th Autocovariance Matrix*

For a covariance-stationary n-dimensional vector process, the jth autocovariance is defined to be the following $(n \times n)$ matrix:

$$\boldsymbol{\Gamma}_j = E[(\mathbf{y}_t - \boldsymbol{\mu})(\mathbf{y}_{t-j} - \boldsymbol{\mu})']. \qquad [10.2.1]$$

Note that although $\gamma_j = \gamma_{-j}$ for a scalar process, the same is not true of a vector

process:

$$\mathbf{\Gamma}_j \neq \mathbf{\Gamma}_{-j}.$$

For example, the (1, 2) element of $\mathbf{\Gamma}_j$ gives the covariance between y_{1t} and $y_{2,t-j}$. The (1, 2) element of $\mathbf{\Gamma}_{-j}$ gives the covariance between y_{1t} and $y_{2,t+j}$. There is no reason that these should be related—the response of y_1 to previous movements in y_2 could be completely different from the response of y_2 to previous movements in y_1.

Instead, the correct relation is

$$\mathbf{\Gamma}_j' = \mathbf{\Gamma}_{-j}. \qquad [10.2.2]$$

To derive [10.2.2], notice that covariance-stationarity would mean that t in [10.2.1] could be replaced with $t + j$:

$$\mathbf{\Gamma}_j = E[(\mathbf{y}_{t+j} - \boldsymbol{\mu})(\mathbf{y}_{(t+j)-j} - \boldsymbol{\mu})'] = E[(\mathbf{y}_{t+j} - \boldsymbol{\mu})(\mathbf{y}_t - \boldsymbol{\mu})'].$$

Taking transposes,

$$\mathbf{\Gamma}_j' = E[(\mathbf{y}_t - \boldsymbol{\mu})(\mathbf{y}_{t+j} - \boldsymbol{\mu})'] = \mathbf{\Gamma}_{-j},$$

as claimed.

Vector MA(q) *Process*

A *vector moving average process* of order q takes the form

$$\mathbf{y}_t = \boldsymbol{\mu} + \boldsymbol{\varepsilon}_t + \mathbf{\Theta}_1\boldsymbol{\varepsilon}_{t-1} + \mathbf{\Theta}_2\boldsymbol{\varepsilon}_{t-2} + \cdots + \mathbf{\Theta}_q\boldsymbol{\varepsilon}_{t-q}, \qquad [10.2.3]$$

where $\boldsymbol{\varepsilon}_t$ is a vector white noise process satisfying [10.1.5] and [10.1.6] and $\mathbf{\Theta}_j$ denotes an $(n \times n)$ matrix of *MA* coefficients for $j = 1, 2, \ldots, q$. The mean of $\mathbf{y}_t$ is $\boldsymbol{\mu}$, and the variance is

$$\begin{aligned}\mathbf{\Gamma}_0 &= E[(\mathbf{y}_t - \boldsymbol{\mu})(\mathbf{y}_t - \boldsymbol{\mu})'] \\ &= E[\boldsymbol{\varepsilon}_t\boldsymbol{\varepsilon}_t'] + \mathbf{\Theta}_1 E[\boldsymbol{\varepsilon}_{t-1}\boldsymbol{\varepsilon}_{t-1}']\mathbf{\Theta}_1' + \mathbf{\Theta}_2 E[\boldsymbol{\varepsilon}_{t-2}\boldsymbol{\varepsilon}_{t-2}']\mathbf{\Theta}_2' \\ &\quad + \cdots + \mathbf{\Theta}_q E[\boldsymbol{\varepsilon}_{t-q}\boldsymbol{\varepsilon}_{t-q}']\mathbf{\Theta}_q' \\ &= \mathbf{\Omega} + \mathbf{\Theta}_1\mathbf{\Omega}\mathbf{\Theta}_1' + \mathbf{\Theta}_2\mathbf{\Omega}\mathbf{\Theta}_2' + \cdots + \mathbf{\Theta}_q\mathbf{\Omega}\mathbf{\Theta}_q',\end{aligned} \qquad [10.2.4]$$

with autocovariances

$$\mathbf{\Gamma}_j = \begin{cases} \mathbf{\Theta}_j\mathbf{\Omega} + \mathbf{\Theta}_{j+1}\mathbf{\Omega}\mathbf{\Theta}_1' + \mathbf{\Theta}_{j+2}\mathbf{\Omega}\mathbf{\Theta}_2' + \cdots + \mathbf{\Theta}_q\mathbf{\Omega}\mathbf{\Theta}_{q-j}' & \text{for } j = 1, 2, \ldots, q \\ \mathbf{\Omega}\mathbf{\Theta}_{-j}' + \mathbf{\Theta}_1\mathbf{\Omega}\mathbf{\Theta}_{-j+1}' + \mathbf{\Theta}_2\mathbf{\Omega}\mathbf{\Theta}_{-j+2}' + \cdots + \mathbf{\Theta}_{q+j}\mathbf{\Omega}\mathbf{\Theta}_q' & \text{for } j = -1, -2, \ldots, -q \\ \mathbf{0} & \text{for } |j| > q, \end{cases} \qquad [10.2.5]$$

where $\mathbf{\Theta}_0 = \mathbf{I}_n$. Thus, any vector $MA(q)$ process is covariance-stationary.

Vector MA(∞) *Process*

The vector $MA(\infty)$ process is written

$$\mathbf{y}_t = \boldsymbol{\mu} + \boldsymbol{\varepsilon}_t + \mathbf{\Psi}_1\boldsymbol{\varepsilon}_{t-1} + \mathbf{\Psi}_2\boldsymbol{\varepsilon}_{t-2} + \cdots \qquad [10.2.6]$$

for $\boldsymbol{\varepsilon}_t$ again satisfying [10.1.5] and [10.1.6].

A sequence of scalars $\{h_s\}_{s=-\infty}^{\infty}$ was said to be absolutely summable if $\Sigma_{s=-\infty}^{\infty}|h_s| < \infty$. For $\mathbf{H}_s$ an $(n \times m)$ matrix, the sequence of matrices $\{\mathbf{H}_s\}_{s=-\infty}^{\infty}$ is

absolutely summable if each of its elements forms an absolutely summable scalar sequence. For example, if $\psi_{ij}^{(s)}$ denotes the row i, column j element of the moving average parameter matrix $\mathbf{\Psi}_s$ associated with lag s, then the sequence $\{\mathbf{\Psi}_s\}_{s=0}^{\infty}$ is absolutely summable if

$$\sum_{s=0}^{\infty} |\psi_{ij}^{(s)}| < \infty \qquad \text{for } i = 1, 2, \ldots, n \text{ and } j = 1, 2, \ldots, n. \qquad [10.2.7]$$

Many of the results for scalar $MA(\infty)$ processes with absolutely summable coefficients go through for vector processes as well. This is summarized by the following theorem, proved in Appendix 10.A to this chapter.

Proposition 10.2: *Let* $\mathbf{y}_t$ *be an* $(n \times 1)$ *vector satisfying*

$$\mathbf{y}_t = \boldsymbol{\mu} + \sum_{k=0}^{\infty} \mathbf{\Psi}_k \boldsymbol{\varepsilon}_{t-k},$$

where $\boldsymbol{\varepsilon}_t$ *is vector white noise satisfying [10.1.5] and [10.1.6] and* $\{\mathbf{\Psi}_k\}_{k=0}^{\infty}$ *is absolutely summable. Let* y_{it} *denote the ith element of* $\mathbf{y}_t$, *and let* μ_i *denote the ith element of* $\boldsymbol{\mu}$. *Then*

(*a*) *the autocovariance between the ith variable at time t and the jth variable s periods earlier,* $E(y_{it} - \mu_i)(y_{j,t-s} - \mu_j)$, *exists and is given by the row i, column j element of*

$$\mathbf{\Gamma}_s = \sum_{v=0}^{\infty} \mathbf{\Psi}_{s+v} \mathbf{\Omega} \mathbf{\Psi}_v' \qquad \text{for } s = 0, 1, 2, \ldots;$$

(*b*) *the sequence of matrices* $\{\mathbf{\Gamma}_s\}_{s=0}^{\infty}$ *is absolutely summable.*

If, furthermore, $\{\boldsymbol{\varepsilon}_t\}_{t=-\infty}^{\infty}$ *is an i.i.d. sequence with* $E|\varepsilon_{i_1,t}\varepsilon_{i_2,t}\varepsilon_{i_3,t}\varepsilon_{i_4,t}| < \infty$ *for* $i_1, i_2, i_3, i_4 = 1, 2, \ldots, n$, *then also,*

(*c*) $E|y_{i_1,t_1}y_{i_2,t_2}y_{i_3,t_3}y_{i_4,t_4}| < \infty$ *for* $i_1, i_2, i_3, i_4 = 1, 2, \ldots, n$ *and for all* t_1, t_2, t_3, t_4;

(*d*) $(1/T)\sum_{t=1}^{T} y_{it}y_{j,t-s} \xrightarrow{p} E(y_{it}y_{j,t-s})$ *for* $i, j = 1, 2, \ldots, n$ *and for all s.*

Result (a) implies that the second moments of an $MA(\infty)$ vector process with absolutely summable coefficients can be found by taking the limit of [10.2.5] as $q \to \infty$. Result (b) is a convergence condition on these moments that will turn out to ensure that the vector process is ergodic for the mean (see Proposition 10.5 later in this chapter). Result (c) says that $\mathbf{y}_t$ has bounded fourth moments, while result (d) establishes that $\mathbf{y}_t$ is ergodic for second moments.

Note that the vector $MA(\infty)$ representation of a stationary vector autoregression calculated from [10.1.4] satisfies the absolute summability condition. To see this, recall from [10.1.14] that $\mathbf{\Psi}_s$ is a block of the matrix $\mathbf{F}^s$. If $\mathbf{F}$ has np distinct eigenvalues $(\lambda_1, \lambda_2, \ldots, \lambda_{np})$, then any element of $\mathbf{\Psi}_s$ can be written as a weighted average of these eigenvalues as in equation [1.2.20];

$$\psi_{ij}^{(s)} = c_1(i, j)\cdot\lambda_1^s + c_2(i, j)\cdot\lambda_2^s + \cdots + c_{np}(i, j)\cdot\lambda_{np}^s,$$

where $c_v(i, j)$ denotes a constant that depends on v, i, and j but not s. Absolute summability [10.2.7] then follows from the same arguments as in Exercise 3.5.

Multivariate Filters

Suppose that the $(n \times 1)$ vector $\mathbf{y}_t$ follows an $MA(\infty)$ process:

$$\mathbf{y}_t = \boldsymbol{\mu}_Y + \boldsymbol{\Psi}(L)\boldsymbol{\varepsilon}_t, \qquad [10.2.8]$$

with $\{\boldsymbol{\Psi}_k\}_{k=0}^{\infty}$ absolutely summable. Let $\{\mathbf{H}_k\}_{k=-\infty}^{\infty}$ be an absolutely summable sequence of $(r \times n)$ matrices and suppose that an $(r \times 1)$ vector $\mathbf{x}_t$ is related to $\mathbf{y}_t$ according to

$$\mathbf{x}_t = \mathbf{H}(L)\mathbf{y}_t = \sum_{k=-\infty}^{\infty} \mathbf{H}_k \mathbf{y}_{t-k}. \qquad [10.2.9]$$

That is,

$$\begin{aligned} \mathbf{x}_t &= \mathbf{H}(L)[\boldsymbol{\mu}_Y + \boldsymbol{\Psi}(L)\boldsymbol{\varepsilon}_t] \\ &= \mathbf{H}(1)\boldsymbol{\mu}_Y + \mathbf{H}(L)\boldsymbol{\Psi}(L)\boldsymbol{\varepsilon}_t \\ &= \boldsymbol{\mu}_X + \mathbf{B}(L)\boldsymbol{\varepsilon}_t, \end{aligned} \qquad [10.2.10]$$

where $\boldsymbol{\mu}_X \equiv \mathbf{H}(1)\boldsymbol{\mu}_Y$ and $\mathbf{B}(L)$ is the compound operator given by

$$\mathbf{B}(L) = \sum_{k=-\infty}^{\infty} \mathbf{B}_k L^k = \mathbf{H}(L)\boldsymbol{\Psi}(L). \qquad [10.2.11]$$

The following proposition establishes that $\mathbf{x}_t$ follows an absolutely summable two-sided $MA(\infty)$ process.

Proposition 10.3: Let $\{\boldsymbol{\Psi}_k\}_{k=0}^{\infty}$ be an absolutely summable sequence of $(n \times n)$ matrices and let $\{\mathbf{H}_k\}_{k=-\infty}^{\infty}$ be an absolutely summable sequence of $(r \times n)$ matrices. Then the sequence of matrices $\{\mathbf{B}_k\}_{k=-\infty}^{\infty}$ associated with the operator $\mathbf{B}(L) = \mathbf{H}(L)\boldsymbol{\Psi}(L)$ is absolutely summable.

If $\{\boldsymbol{\varepsilon}_t\}$ in [10.2.8] is i.i.d. with finite fourth moments, then $\{\mathbf{x}_t\}$ in [10.2.9] has finite fourth moments and is ergodic for second moments.

Vector Autoregression

Next we derive expressions for the second moments for $\mathbf{y}_t$ following a $VAR(p)$. Let $\boldsymbol{\xi}_t$ be as defined in equation [10.1.9]. Assuming that $\boldsymbol{\xi}$ and $\mathbf{y}$ are covariance-stationary, let $\boldsymbol{\Sigma}$ denote the variance of $\boldsymbol{\xi}$,

$$\begin{aligned} \boldsymbol{\Sigma} &= E(\boldsymbol{\xi}_t \boldsymbol{\xi}_t') \\ &= E\left\{ \begin{bmatrix} \mathbf{y}_t - \boldsymbol{\mu} \\ \mathbf{y}_{t-1} - \boldsymbol{\mu} \\ \vdots \\ \mathbf{y}_{t-p+1} - \boldsymbol{\mu} \end{bmatrix} \right. \\ &\qquad \left. \times [(\mathbf{y}_t - \boldsymbol{\mu})' \quad (\mathbf{y}_{t-1} - \boldsymbol{\mu})' \quad \cdots \quad (\mathbf{y}_{t-p+1} - \boldsymbol{\mu})'] \right\} \\ &= \begin{bmatrix} \boldsymbol{\Gamma}_0 & \boldsymbol{\Gamma}_1 & \cdots & \boldsymbol{\Gamma}_{p-1} \\ \boldsymbol{\Gamma}_1' & \boldsymbol{\Gamma}_0 & \cdots & \boldsymbol{\Gamma}_{p-2} \\ \vdots & \vdots & \cdots & \vdots \\ \boldsymbol{\Gamma}_{p-1}' & \boldsymbol{\Gamma}_{p-2}' & \cdots & \boldsymbol{\Gamma}_0 \end{bmatrix}, \end{aligned} \qquad [10.2.12]$$

where Γ_j denotes the jth autocovariance of the original process $\mathbf{y}$. Postmultiplying [10.1.11] by its own transpose and taking expectations gives

$$E[\boldsymbol{\xi}_t\boldsymbol{\xi}_t'] = E[(\mathbf{F}\boldsymbol{\xi}_{t-1} + \mathbf{v}_t)(\mathbf{F}\boldsymbol{\xi}_{t-1} + \mathbf{v}_t)'] = \mathbf{F}E(\boldsymbol{\xi}_{t-1}\boldsymbol{\xi}_{t-1}')\mathbf{F}' + E(\mathbf{v}_t\mathbf{v}_t'),$$

or

$$\boldsymbol{\Sigma} = \mathbf{F}\boldsymbol{\Sigma}\mathbf{F}' + \mathbf{Q}. \qquad [10.2.13]$$

A closed-form solution to [10.2.13] can be obtained in terms of the *vec* operator. If $\mathbf{A}$ is an $(m \times n)$ matrix, then vec($\mathbf{A}$) is an $(mn \times 1)$ column vector, obtained by stacking the columns of $\mathbf{A}$, one below the other, with the columns ordered from left to right. For example, if

$$\mathbf{A} = \begin{bmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \\ a_{31} & a_{32} \end{bmatrix},$$

then

$$\text{vec}(\mathbf{A}) = \begin{bmatrix} a_{11} \\ a_{21} \\ a_{31} \\ a_{12} \\ a_{22} \\ a_{32} \end{bmatrix}. \qquad [10.2.14]$$

Appendix 10.A establishes the following useful result.

Proposition 10.4: Let $\mathbf{A}$, $\mathbf{B}$, *and* $\mathbf{C}$ *be matrices whose dimensions are such that the product* $\mathbf{ABC}$ *exists. Then*

$$\text{vec}(\mathbf{ABC}) = (\mathbf{C}' \otimes \mathbf{A})\cdot\text{vec}(\mathbf{B}) \qquad [10.2.15]$$

where the symbol $\otimes$ *denotes the Kronecker product.*

Thus, if the vec operator is applied to both sides of [10.2.13], the result is

$$\text{vec}(\boldsymbol{\Sigma}) = (\mathbf{F} \otimes \mathbf{F})\cdot\text{vec}(\boldsymbol{\Sigma}) + \text{vec}(\mathbf{Q}) = \mathcal{A}\,\text{vec}(\boldsymbol{\Sigma}) + \text{vec}(\mathbf{Q}), \qquad [10.2.16]$$

where

$$\mathcal{A} \equiv (\mathbf{F} \otimes \mathbf{F}). \qquad [10.2.17]$$

Let $r = np$, so that $\mathbf{F}$ is an $(r \times r)$ matrix and $\mathcal{A}$ is an $(r^2 \times r^2)$ matrix. Equation [10.2.16] has the solution

$$\text{vec}(\boldsymbol{\Sigma}) = [\mathbf{I}_{r^2} - \mathcal{A}]^{-1}\,\text{vec}(\mathbf{Q}), \qquad [10.2.18]$$

provided that the matrix $[\mathbf{I}_{r^2} - \mathcal{A}]$ is nonsingular. This will be true as long as unity is not an eigenvalue of $\mathcal{A}$. But recall that the eigenvalues of $\mathbf{F} \otimes \mathbf{F}$ are all of the form $\lambda_i\lambda_j$, where λ_i and λ_j are eigenvalues of $\mathbf{F}$. Since $|\lambda_i| < 1$ for all i, it follows that all eigenvalues of $\mathcal{A}$ are inside the unit circle, meaning that $[\mathbf{I}_{r^2} - \mathcal{A}]$ is indeed nonsingular.

The first p autocovariance matrices of a $VAR(p)$ process can be calculated

by substituting [10.2.12] into [10.2.18]:

$$\text{vec}\begin{bmatrix} \Gamma_0 & \Gamma_1 & \cdots & \Gamma_{p-1} \\ \Gamma_1' & \Gamma_0 & \cdots & \Gamma_{p-2} \\ \vdots & \vdots & \cdots & \vdots \\ \Gamma_{p-1}' & \Gamma_{p-2}' & \cdots & \Gamma_0 \end{bmatrix} = [\mathbf{I}_{r^2} - \mathcal{A}]^{-1}\,\text{vec}(\mathbf{Q}). \qquad [10.2.19]$$

The jth autocovariance of $\boldsymbol{\xi}$ (denoted $\boldsymbol{\Sigma}_j$) can be found by postmultiplying [10.1.11] by $\boldsymbol{\xi}_{t-j}'$ and taking expectations:

$$E(\boldsymbol{\xi}_t\boldsymbol{\xi}_{t-j}') = \mathbf{F}\cdot E(\boldsymbol{\xi}_{t-1}\boldsymbol{\xi}_{t-j}') + E(\mathbf{v}_t\boldsymbol{\xi}_{t-j}').$$

Thus,

$$\boldsymbol{\Sigma}_j = \mathbf{F}\boldsymbol{\Sigma}_{j-1} \qquad \text{for } j = 1, 2, \ldots, \qquad [10.2.20]$$

or

$$\boldsymbol{\Sigma}_j = \mathbf{F}^j\boldsymbol{\Sigma} \qquad \text{for } j = 1, 2, \ldots. \qquad [10.2.21]$$

The jth autocovariance $\boldsymbol{\Gamma}_j$ of the original process $\mathbf{y}_t$ is given by the first n rows and n columns of [10.2.20]:

$$\boldsymbol{\Gamma}_j = \boldsymbol{\Phi}_1\boldsymbol{\Gamma}_{j-1} + \boldsymbol{\Phi}_2\boldsymbol{\Gamma}_{j-2} + \cdots + \boldsymbol{\Phi}_p\boldsymbol{\Gamma}_{j-p} \qquad \text{for } j = p, p+1, p+2, \ldots. \qquad [10.2.22]$$

10.3. *The Autocovariance-Generating Function for Vector Processes*

Definition of Autocovariance-Generating Function for Vector Processes

Recall that for a covariance-stationary univariate process y_t with absolutely summable autocovariances, the (scalar-valued) autocovariance-generating function $g_Y(z)$ is defined as

$$g_Y(z) \equiv \sum_{j=-\infty}^{\infty} \gamma_j z^j$$

with

$$\gamma_j \equiv E[(y_t - \mu)(y_{t-j} - \mu)]$$

and z a complex scalar. For a covariance-stationary vector process $\mathbf{y}_t$ with an absolutely summable sequence of autocovariance matrices, the analogous matrix-valued autocovariance-generating function $\mathbf{G}_Y(z)$ is defined as

$$\mathbf{G}_Y(z) \equiv \sum_{j=-\infty}^{\infty} \boldsymbol{\Gamma}_j z^j, \qquad [10.3.1]$$

where

$$\boldsymbol{\Gamma}_j \equiv E[(\mathbf{y}_t - \boldsymbol{\mu})(\mathbf{y}_{t-j} - \boldsymbol{\mu})']$$

and z is again a complex scalar.

Autocovariance-Generating Function for a Vector Moving Average Process

For example, for the vector white noise process $\boldsymbol{\varepsilon}_t$ characterized by [10.1.5] and [10.1.6], the autocovariance-generating function is

$$\mathbf{G}_{\varepsilon}(z) = \boldsymbol{\Omega}. \qquad [10.3.2]$$

For the vector $MA(q)$ process of [10.2.3], the univariate expression [3.6.3] for the autocovariance-generating function generalizes to

$$\begin{aligned}\mathbf{G}_{\mathbf{Y}}(z) = {} & (\mathbf{I}_n + \boldsymbol{\Theta}_1 z + \boldsymbol{\Theta}_2 z^2 + \cdots + \boldsymbol{\Theta}_q z^q)\boldsymbol{\Omega} \\ & \times (\mathbf{I}_n + \boldsymbol{\Theta}_1' z^{-1} + \boldsymbol{\Theta}_2' z^{-2} + \cdots + \boldsymbol{\Theta}_q' z^{-q}).\end{aligned} \qquad [10.3.3]$$

This can be verified by noting that the coefficient on z^j in [10.3.3] is equal to $\boldsymbol{\Gamma}_j$ as given in [10.2.5].

For an $MA(\infty)$ process of the form

$$\mathbf{y}_t = \boldsymbol{\mu} + \boldsymbol{\Psi}_0\boldsymbol{\varepsilon}_t + \boldsymbol{\Psi}_1\boldsymbol{\varepsilon}_{t-1} + \boldsymbol{\Psi}_2\boldsymbol{\varepsilon}_{t-2} + \cdots = \boldsymbol{\mu} + \boldsymbol{\Psi}(L)\boldsymbol{\varepsilon}_t,$$

with $\{\boldsymbol{\Psi}_k\}_{k=0}^{\infty}$ absolutely summable, [10.3.3] generalizes to

$$\mathbf{G}_{\mathbf{Y}}(z) = [\boldsymbol{\Psi}(z)]\boldsymbol{\Omega}[\boldsymbol{\Psi}(z^{-1})]'. \qquad [10.3.4]$$

Autocovariance-Generating Function for a Vector Autoregression

Consider the $VAR(1)$ process $\boldsymbol{\xi}_t = \mathbf{F}\boldsymbol{\xi}_{t-1} + \mathbf{v}_t$ with eigenvalues of $\mathbf{F}$ inside the unit circle and with $\boldsymbol{\xi}_t$ an $(r \times 1)$ vector and $E(\mathbf{v}_t\mathbf{v}_t') = \mathbf{Q}$. Equation [10.3.4] implies that the autocovariance-generating function can be expressed as

$$\begin{aligned}\mathbf{G}_{\xi}(z) &= [\mathbf{I}_r - \mathbf{F}z]^{-1}\mathbf{Q}[\mathbf{I}_r - \mathbf{F}'z^{-1}]^{-1} \\ &= [\mathbf{I}_r + \mathbf{F}z + \mathbf{F}^2z^2 + \mathbf{F}^3z^3 + \cdots]\mathbf{Q} \\ &\quad \times [\mathbf{I}_r + (\mathbf{F}')z^{-1} + (\mathbf{F}')^2z^{-2} + (\mathbf{F}')^3z^{-3} + \cdots].\end{aligned} \qquad [10.3.5]$$

Transformations of Vector Processes

The autocovariance-generating function of the sum of two univariate processes that are uncorrelated with each other is equal to the sum of their individual autocovariance-generating functions (equation [4.7.19]). This result readily generalizes to the vector case:

$$\begin{aligned}\mathbf{G}_{\mathbf{X}+\mathbf{W}}(z) &= \sum_{j=-\infty}^{\infty} E[(\mathbf{x}_t + \mathbf{w}_t - \boldsymbol{\mu}_{\mathbf{X}} - \boldsymbol{\mu}_{\mathbf{W}}) \\ &\qquad \times (\mathbf{x}_{t-j} + \mathbf{w}_{t-j} - \boldsymbol{\mu}_{\mathbf{X}} - \boldsymbol{\mu}_{\mathbf{W}})']z^j \\ &= \sum_{j=-\infty}^{\infty} E[(\mathbf{x}_t - \boldsymbol{\mu}_{\mathbf{X}})(\mathbf{x}_{t-j} - \boldsymbol{\mu}_{\mathbf{X}})'z^j] \\ &\qquad + \sum_{j=-\infty}^{\infty} E[(\mathbf{w}_t - \boldsymbol{\mu}_{\mathbf{W}})(\mathbf{w}_{t-j} - \boldsymbol{\mu}_{\mathbf{W}})'z^j] \\ &= \mathbf{G}_{\mathbf{X}}(z) + \mathbf{G}_{\mathbf{W}}(z).\end{aligned}$$

Note also that if an $(r \times 1)$ vector $\boldsymbol{\xi}_t$ is premultiplied by a nonstochastic $(n \times r)$ matrix $\mathbf{H}'$, the effect is to premultiply the autocovariance by $\mathbf{H}'$ and postmultiply by $\mathbf{H}$:

$$E[(\mathbf{H}'\boldsymbol{\xi}_t - \mathbf{H}'\boldsymbol{\mu}_\xi)(\mathbf{H}'\boldsymbol{\xi}_{t-j} - \mathbf{H}'\boldsymbol{\mu}_\xi)'] = \mathbf{H}'E[(\boldsymbol{\xi}_t - \boldsymbol{\mu}_\xi)(\boldsymbol{\xi}_{t-j} - \boldsymbol{\mu}_\xi)']\mathbf{H},$$

implying

$$\mathbf{G}_{\mathbf{H}'\xi}(z) = \mathbf{H}'\mathbf{G}_\xi(z)\mathbf{H}.$$

Putting these results together, consider $\boldsymbol{\xi}_t$ the r-dimensional $VAR(1)$ process $\boldsymbol{\xi}_t = \mathbf{F}\boldsymbol{\xi}_{t-1} + \mathbf{v}_t$ and a new process $\mathbf{u}_t$ given by $\mathbf{u}_t = \mathbf{H}'\boldsymbol{\xi}_t + \mathbf{w}_t$ with $\mathbf{w}_t$ a white noise process that is uncorrelated with $\boldsymbol{\xi}_{t-j}$ for all j. Then

$$\mathbf{G}_\mathbf{U}(z) = \mathbf{H}'\mathbf{G}_\xi(z)\mathbf{H} + \mathbf{G}_\mathbf{W}(z),$$

or, if $\mathbf{R}$ is the variance of $\mathbf{w}_t$,

$$\mathbf{G}_\mathbf{U}(z) = \mathbf{H}'[\mathbf{I}_r - \mathbf{F}z]^{-1}\mathbf{Q}[\mathbf{I}_r - \mathbf{F}'z^{-1}]^{-1}\mathbf{H} + \mathbf{R}. \qquad [10.3.6]$$

More generally, consider an $(n \times 1)$ vector $\mathbf{y}_t$ characterized by

$$\mathbf{y}_t = \boldsymbol{\mu}_\mathbf{Y} + \boldsymbol{\Psi}(L)\boldsymbol{\varepsilon}_t,$$

where $\boldsymbol{\varepsilon}_t$ is a white noise process with variance-covariance matrix given by $\boldsymbol{\Omega}$ and where $\boldsymbol{\Psi}(L) = \sum_{k=0}^{\infty}\boldsymbol{\Psi}_k L^k$ with $\{\boldsymbol{\Psi}_k\}_{k=0}^{\infty}$ absolutely summable. Thus, the autocovariance-generating function for $\mathbf{y}$ is

$$\mathbf{G}_\mathbf{Y}(z) = \boldsymbol{\Psi}(z)\boldsymbol{\Omega}[\boldsymbol{\Psi}(z^{-1})]'. \qquad [10.3.7]$$

Let $\{\mathbf{H}_k\}_{k=-\infty}^{\infty}$ be an absolutely summable sequence of $(r \times n)$ matrices, and suppose that an $(r \times 1)$ vector $\mathbf{x}_t$ is constructed from $\mathbf{y}_t$ according to

$$\mathbf{x}_t = \mathbf{H}(L)\mathbf{y}_t = \sum_{k=-\infty}^{\infty}\mathbf{H}_k\mathbf{y}_{t-k} = \boldsymbol{\mu}_\mathbf{X} + \mathbf{B}(L)\boldsymbol{\varepsilon}_t,$$

where $\boldsymbol{\mu}_\mathbf{X} = \mathbf{H}(1)\boldsymbol{\mu}_\mathbf{Y}$ and $\mathbf{B}(L) = \mathbf{H}(L)\boldsymbol{\Psi}(L)$ as in [10.2.10] and [10.2.11]. Then the autocovariance-generating function for $\mathbf{x}$ can be found from

$$\mathbf{G}_\mathbf{X}(z) = \mathbf{B}(z)\boldsymbol{\Omega}[\mathbf{B}(z^{-1})]' = [\mathbf{H}(z)\boldsymbol{\Psi}(z)]\boldsymbol{\Omega}[\boldsymbol{\Psi}(z^{-1})]'[\mathbf{H}(z^{-1})]'. \qquad [10.3.8]$$

Comparing [10.3.8] with [10.3.7], the effect of applying the filter $\mathbf{H}(L)$ to $\mathbf{y}_t$ is to premultiply the autocovariance-generating function by $\mathbf{H}(z)$ and to postmultiply by the transpose of $\mathbf{H}(z^{-1})$:

$$\mathbf{G}_\mathbf{X}(z) = [\mathbf{H}(z)]\mathbf{G}_\mathbf{Y}(z)[\mathbf{H}(z^{-1})]'. \qquad [10.3.9]$$

10.4. *The Spectrum for Vector Processes*

Let $\mathbf{y}_t$ be an $(n \times 1)$ vector with mean $E(\mathbf{y}_t) = \boldsymbol{\mu}$ and kth autocovariance matrix

$$E[(\mathbf{y}_t - \boldsymbol{\mu})(\mathbf{y}_{t-k} - \boldsymbol{\mu})'] = \boldsymbol{\Gamma}_k. \qquad [10.4.1]$$

If $\{\boldsymbol{\Gamma}_k\}_{k=-\infty}^{\infty}$ is absolutely summable and if z is a complex scalar, the autocovariance-generating function of $\mathbf{y}$ is given by

$$\mathbf{G}_\mathbf{Y}(z) = \sum_{k=-\infty}^{\infty}\boldsymbol{\Gamma}_k z^k. \qquad [10.4.2]$$

The function $\mathbf{G}_Y(z)$ associates an $(n \times n)$ matrix of complex numbers with the complex scalar z. If [10.4.2] is divided by 2π and evaluated at $z = e^{-i\omega}$, where ω is a real scalar and $i = \sqrt{-1}$, the result is the *population spectrum* of the vector $\mathbf{y}$:

$$\mathbf{s}_Y(\omega) = (2\pi)^{-1}\mathbf{G}_Y(e^{-i\omega}) = (2\pi)^{-1} \sum_{k=-\infty}^{\infty} \mathbf{\Gamma}_k e^{-i\omega k}. \qquad [10.4.3]$$

The population spectrum associates an $(n \times n)$ matrix of complex numbers with the real scalar ω.

Identical calculations to those used to establish Proposition 6.1 indicate that when any element of $\mathbf{s}_Y(\omega)$ is multiplied by $e^{i\omega k}$ and the resulting function of ω is integrated from $-\pi$ to π, the result is the corresponding element of the kth autocovariance matrix of $\mathbf{y}$:

$$\int_{-\pi}^{\pi} \mathbf{s}_Y(\omega) e^{i\omega k}\, d\omega = \mathbf{\Gamma}_k. \qquad [10.4.4]$$

Thus, as in the univariate case, the sequence of autocovariances $\{\mathbf{\Gamma}_k\}_{k=-\infty}^{\infty}$ and the function represented by the population spectrum $\mathbf{s}_Y(\omega)$ contain the identical information.

As a special case, when $k = 0$, equation [10.4.4] implies

$$\int_{-\pi}^{\pi} \mathbf{s}_Y(\omega)\, d\omega = \mathbf{\Gamma}_0. \qquad [10.4.5]$$

In other words, the area under the population spectrum is the unconditional variance-covariance matrix of $\mathbf{y}$.

The jth diagonal element of $\mathbf{\Gamma}_k$ is $E(y_{jt} - \mu_j)(y_{j,t-k} - \mu_j)$, the kth autocovariance of y_{jt}. Thus, the jth diagonal element of the multivariate spectrum $\mathbf{s}_Y(\omega)$ is just the univariate spectrum of the scalar y_{jt}. It follows from the properties of the univariate spectrum discussed in Chapter 6 that the diagonal elements of $\mathbf{s}_Y(\omega)$ are real-valued and nonnegative for all ω. However, the same is not true of the off-diagonal elements of $\mathbf{s}_Y(\omega)$—in general, the off-diagonal elements of $\mathbf{s}_Y(\omega)$ will be complex numbers.

To gain further understanding of the multivariate spectrum, we concentrate on the case of $n = 2$ variables, denoted

$$\mathbf{y}_t = \begin{bmatrix} X_t \\ Y_t \end{bmatrix}.$$

The kth autocovariance matrix is then

$$\begin{aligned} \mathbf{\Gamma}_k &= E\begin{bmatrix} (X_t - \mu_X)(X_{t-k} - \mu_X) & (X_t - \mu_X)(Y_{t-k} - \mu_Y) \\ (Y_t - \mu_Y)(X_{t-k} - \mu_X) & (Y_t - \mu_Y)(Y_{t-k} - \mu_Y) \end{bmatrix} \\ &\equiv \begin{bmatrix} \gamma_{XX}^{(k)} & \gamma_{XY}^{(k)} \\ \gamma_{YX}^{(k)} & \gamma_{YY}^{(k)} \end{bmatrix}. \end{aligned} \qquad [10.4.6]$$

Recall from [10.2.2] that $\mathbf{\Gamma}_k' = \mathbf{\Gamma}_{-k}$. Hence,

$$\gamma_{XX}^{(k)} = \gamma_{XX}^{(-k)} \qquad [10.4.7]$$

$$\gamma_{YY}^{(k)} = \gamma_{YY}^{(-k)} \qquad [10.4.8]$$

$$\gamma_{XY}^{(k)} = \gamma_{YX}^{(-k)}. \qquad [10.4.9]$$

For this $n = 2$ case, the population spectrum [10.4.3] would be

$$\mathbf{s}_Y(\omega) = \frac{1}{2\pi}\begin{bmatrix} \sum_{k=-\infty}^{\infty} \gamma_{XX}^{(k)} e^{-i\omega k} & \sum_{k=-\infty}^{\infty} \gamma_{XY}^{(k)} e^{-i\omega k} \\ \sum_{k=-\infty}^{\infty} \gamma_{YX}^{(k)} e^{-i\omega k} & \sum_{k=-\infty}^{\infty} \gamma_{YY}^{(k)} e^{-i\omega k} \end{bmatrix}$$

$$= \frac{1}{2\pi}\begin{bmatrix} \sum_{k=-\infty}^{\infty} \gamma_{XX}^{(k)}\{\cos(\omega k) - i\cdot\sin(\omega k)\} & \sum_{k=-\infty}^{\infty} \gamma_{XY}^{(k)}\{\cos(\omega k) - i\cdot\sin(\omega k)\} \\ \sum_{k=-\infty}^{\infty} \gamma_{YX}^{(k)}\{\cos(\omega k) - i\cdot\sin(\omega k)\} & \sum_{k=-\infty}^{\infty} \gamma_{YY}^{(k)}\{\cos(\omega k) - i\cdot\sin(\omega k)\} \end{bmatrix}. \qquad [10.4.10]$$

Using [10.4.7] and [10.4.8] along with the facts that $\sin(-\omega k) = -\sin(\omega k)$ and $\sin(0) = 0$, the imaginary components disappear from the diagonal terms:

$$\mathbf{s}_Y(\omega) = \frac{1}{2\pi}\begin{bmatrix} \sum_{k=-\infty}^{\infty} \gamma_{XX}^{(k)} \cos(\omega k) & \sum_{k=-\infty}^{\infty} \gamma_{XY}^{(k)}\{\cos(\omega k) - i\cdot\sin(\omega k)\} \\ \sum_{k=-\infty}^{\infty} \gamma_{YX}^{(k)}\{\cos(\omega k) - i\cdot\sin(\omega k)\} & \sum_{k=-\infty}^{\infty} \gamma_{YY}^{(k)} \cos(\omega k) \end{bmatrix}. \qquad [10.4.11]$$

However, since in general $\gamma_{XY}^{(k)} \neq \gamma_{XY}^{(-k)}$, the off-diagonal elements are typically complex numbers.

The Cross Spectrum, Cospectrum, and Quadrature Spectrum

The lower left element of the matrix in [10.4.11] is known as the *population cross spectrum* from X to Y:

$$s_{YX}(\omega) = (2\pi)^{-1} \sum_{k=-\infty}^{\infty} \gamma_{YX}^{(k)}\{\cos(\omega k) - i\cdot\sin(\omega k)\}. \qquad [10.4.12]$$

The cross spectrum can be written in terms of its real and imaginary components as

$$s_{YX}(\omega) = c_{YX}(\omega) + i\cdot q_{YX}(\omega). \qquad [10.4.13]$$

The real component of the cross spectrum is known as the *cospectrum* between X and Y:

$$c_{YX}(\omega) = (2\pi)^{-1} \sum_{k=-\infty}^{\infty} \gamma_{YX}^{(k)} \cos(\omega k). \qquad [10.4.14]$$

One can verify from [10.4.9] and the fact that $\cos(-\omega k) = \cos(\omega k)$ that

$$c_{YX}(\omega) = c_{XY}(\omega). \qquad [10.4.15]$$

The imaginary component of the cross spectrum is known as the *quadrature spectrum* from X to Y:

$$q_{YX}(\omega) = -(2\pi)^{-1} \sum_{k=-\infty}^{\infty} \gamma_{YX}^{(k)} \sin(\omega k). \qquad [10.4.16]$$

One can verify from [10.4.9] and the fact that $\sin(-\omega k) = -\sin(\omega k)$ that the quadrature spectrum from Y to X is the negative of the quadrature spectrum from X to Y:

$$q_{YX}(\omega) = -q_{XY}(\omega).$$

Recalling [10.4.13], these results imply that the off-diagonal elements of $\mathbf{s}_Y(\omega)$ are complex conjugates of each other; in general, the row j, column m element of $\mathbf{s}_Y(\omega)$ is the complex conjugate of the row m, column j element of $\mathbf{s}_Y(\omega)$.

Note that both $c_{YX}(\omega)$ and $q_{YX}(\omega)$ are real-valued periodic functions of ω:

$$c_{YX}(\omega + 2\pi j) = c_{YX}(\omega) \qquad \text{for } j = \pm 1, \pm 2, \ldots$$
$$q_{YX}(\omega + 2\pi j) = q_{YX}(\omega) \qquad \text{for } j = \pm 1, \pm 2, \ldots.$$

It further follows from [10.4.14] that

$$c_{YX}(-\omega) = c_{YX}(\omega),$$

while [10.4.16] implies that

$$q_{YX}(-\omega) = -q_{YX}(\omega). \qquad [10.4.17]$$

Hence, the cospectrum and quadrature spectrum are fully specified by the values they assume as ω ranges between 0 and π.

Result [10.4.5] implies that the cross spectrum integrates to the unconditional covariance between X and Y:

$$\int_{-\pi}^{\pi} s_{YX}(\omega)\, d\omega = E(Y_t - \mu_Y)(X_t - \mu_X).$$

Observe from [10.4.17] that the quadrature spectrum integrates to zero:

$$\int_{-\pi}^{\pi} q_{YX}(\omega)\, d\omega = 0.$$

Hence, the covariance between X and Y can be calculated from the area under the cospectrum between X and Y:

$$\int_{-\pi}^{\pi} c_{YX}(\omega)\, d\omega = E(Y_t - \mu_Y)(X_t - \mu_X). \qquad [10.4.18]$$

The cospectrum between X and Y at frequency ω can thus be interpreted as the portion of the covariance between X and Y that is attributable to cycles with frequency ω. Since the covariance can be positive or negative, the cospectrum can be positive or negative, and indeed, $c_{YX}(\omega)$ may be positive over some frequencies and negative over others.

The Sample Multivariate Periodogram

To gain further understanding of the cospectrum and the quadrature spectrum, let $y_1, y_2, \ldots, y_T$ and $x_1, x_2, \ldots, x_T$ denote samples of T observations on the two variables. If for illustration T is odd, Proposition 6.2 indicates that the value of y_t can be expressed as

$$y_t = \bar{y} + \sum_{j=1}^{M} \{\hat{\alpha}_j \cdot \cos[\omega_j(t-1)] + \hat{\delta}_j \cdot \sin[\omega_j(t-1)]\}, \qquad [10.4.19]$$

where $\bar{y}$ is the sample mean of y, $M = (T-1)/2$, $\omega_j = 2\pi j/T$, and

$$\hat{\alpha}_j = (2/T) \sum_{t=1}^{T} y_t \cdot \cos[\omega_j(t-1)] \qquad [10.4.20]$$

$$\hat{\delta}_j = (2/T) \sum_{t=1}^{T} y_t \cdot \sin[\omega_j(t-1)]. \qquad [10.4.21]$$

An analogous representation for x_t is

$$x_t = \bar{x} + \sum_{j=1}^{M} \{\hat{a}_j \cdot \cos[\omega_j(t-1)] + \hat{d}_j \cdot \sin[\omega_j(t-1)]\} \qquad [10.4.22]$$

$$\hat{a}_j = (2/T) \sum_{t=1}^{T} x_t \cdot \cos[\omega_j(t-1)] \qquad [10.4.23]$$

$$\hat{d}_j = (2/T) \sum_{t=1}^{T} x_t \cdot \sin[\omega_j(t-1)]. \qquad [10.4.24]$$

Recall from [6.2.11] that the periodic regressors in [10.4.19] all have sample mean zero and are mutually orthogonal, while

$$\sum_{t=1}^{T} \cos^2[\omega_j(t-1)] = \sum_{t=1}^{T} \sin^2[\omega_j(t-1)] = T/2. \qquad [10.4.25]$$

Consider the sample covariance between x and y:

$$T^{-1} \sum_{t=1}^{T} (y_t - \bar{y})(x_t - \bar{x}). \qquad [10.4.26]$$

Substituting [10.4.19] and [10.4.22] into [10.4.26] and exploiting the mutual orthogonality of the periodic regressors reveal that

$$\begin{aligned} T^{-1} &\sum_{t=1}^{T} (y_t - \bar{y})(x_t - \bar{x}) \\ &= T^{-1} \sum_{t=1}^{T} \left\{ \sum_{j=1}^{M} \{\hat{\alpha}_j \cdot \cos[\omega_j(t-1)] + \hat{\delta}_j \cdot \sin[\omega_j(t-1)]\} \right. \\ &\quad \left. \times \sum_{j=1}^{M} \{\hat{a}_j \cdot \cos[\omega_j(t-1)] + \hat{d}_j \cdot \sin[\omega_j(t-1)]\} \right\} \qquad [10.4.27] \\ &= T^{-1} \sum_{t=1}^{T} \left\{ \sum_{j=1}^{M} \{\hat{\alpha}_j \hat{a}_j \cdot \cos^2[\omega_j(t-1)] + \hat{\delta}_j \hat{d}_j \cdot \sin^2[\omega_j(t-1)]\} \right\} \\ &= (1/2) \sum_{j=1}^{M} (\hat{\alpha}_j \hat{a}_j + \hat{\delta}_j \hat{d}_j). \end{aligned}$$

Hence, the portion of the sample covariance between x and y that is due to their common dependence on cycles of frequency ω_j is given by

$$(1/2)(\hat{\alpha}_j\hat{a}_j + \hat{\delta}_j\hat{d}_j). \qquad [10.4.28]$$

This magnitude can be related to the sample analog of the cospectrum with calculations similar to those used to establish result (c) of Proposition 6.2. Recall that since

$$\sum_{t=1}^{T} \cos[\omega_j(t-1)] = 0,$$

the magnitude $\hat{\alpha}_j$ in [10.4.20] can alternatively be expressed as

$$\hat{\alpha}_j = (2/T)\sum_{t=1}^{T}(y_t - \bar{y})\cdot\cos[\omega_j(t-1)].$$

Thus,

$$\begin{aligned}
&(\hat{a}_j + i\cdot\hat{d}_j)(\hat{\alpha}_j - i\cdot\hat{\delta}_j)\\
&= (4/T^2)\left\{\sum_{t=1}^{T}(x_t - \bar{x})\cdot\cos[\omega_j(t-1)] + i\cdot\sum_{t=1}^{T}(x_t - \bar{x})\cdot\sin[\omega_j(t-1)]\right\}\\
&\quad\times\left\{\sum_{\tau=1}^{T}(y_\tau - \bar{y})\cdot\cos[\omega_j(\tau-1)] - i\cdot\sum_{\tau=1}^{T}(y_\tau - \bar{y})\cdot\sin[\omega_j(\tau-1)]\right\}\\
&= (4/T^2)\left\{\sum_{t=1}^{T}(x_t - \bar{x})\cdot\exp[i\cdot\omega_j(t-1)]\right\}\left\{\sum_{t=\tau}^{T}(y_\tau - \bar{y})\cdot\exp[-i\cdot\omega_j(\tau-1)]\right\}\\
&= (4/T^2)\left\{\sum_{t=1}^{T}(x_t - \bar{x})(y_t - \bar{y}) + \sum_{t=1}^{T-1}(x_t - \bar{x})(y_{t+1} - \bar{y})\cdot\exp[-i\omega_j]\right.\\
&\quad + \sum_{t=2}^{T}(x_t - \bar{x})(y_{t-1} - \bar{y})\cdot\exp[i\omega_j] + \sum_{t=1}^{T-2}(x_t - \bar{x})(y_{t+2} - \bar{y})\cdot\exp[-2i\omega_j]\\
&\quad + \sum_{t=3}^{T}(x_t - \bar{x})(y_{t-2} - \bar{y})\cdot\exp[2i\omega_j] + \cdots + (x_1 - \bar{x})(y_T - \bar{y})\cdot\exp[-(T-1)i\omega_j]\\
&\quad \left. + (x_T - \bar{x})(y_1 - \bar{y})\cdot\exp[(T-1)i\omega_j]\right\}\\
&= (4/T)\left\{\hat{\gamma}_{yx}^{(0)} + \hat{\gamma}_{yx}^{(1)}\cdot\exp[-i\omega_j] + \hat{\gamma}_{yx}^{(-1)}\cdot\exp[i\omega_j]\right.\\
&\quad + \hat{\gamma}_{yx}^{(2)}\cdot\exp[-2i\omega_j] + \hat{\gamma}_{yx}^{(-2)}\cdot\exp[2i\omega_j] + \cdots\\
&\quad \left. + \hat{\gamma}_{yx}^{(T-1)}\cdot\exp[-(T-1)i\omega_j] + \hat{\gamma}_{yx}^{(-T+1)}\cdot\exp[(T-1)i\omega_j]\right\}, \qquad [10.4.29]
\end{aligned}$$

where $\hat{\gamma}_{yx}^{(k)}$ is the sample covariance between the value of y and the value that x assumed k periods earlier:

$$\hat{\gamma}_{yx}^{(k)} = \begin{cases} (1/T)\displaystyle\sum_{t=1}^{T-k}(x_t - \bar{x})(y_{t+k} - \bar{y}) & \text{for } k = 0, 1, 2, \ldots, T-1\\ (1/T)\displaystyle\sum_{t=-k+1}^{T}(x_t - \bar{x})(y_{t+k} - \bar{y}) & \text{for } k = -1, -2, \ldots, -T+1. \end{cases}$$

[10.4.30]

Result [10.4.29] implies that

$$\begin{aligned}\tfrac{1}{2}(\hat{a}_j + i\cdot\hat{d}_j)(\hat{\alpha}_j - i\cdot\hat{\delta}_j) &= (2/T)\sum_{k=-T+1}^{T-1}\hat{\gamma}_{yx}^{(k)}\cdot\exp[-ki\omega_j] \\ &= (4\pi/T)\cdot\hat{s}_{yx}(\omega_j),\end{aligned} \qquad [10.4.31]$$

where $\hat{s}_{yx}(\omega_j)$ is the *sample cross periodogram* from x to y at frequency ω_j, or the lower left element of the *sample multivariate periodogram*:

$$\hat{\mathbf{s}}_{\mathbf{y}}(\omega) = (2\pi)^{-1}\begin{bmatrix}\sum_{k=-T+1}^{T-1}\hat{\gamma}_{xx}^{(k)}e^{-i\omega k} & \sum_{k=-T+1}^{T-1}\hat{\gamma}_{xy}^{(k)}e^{-i\omega k} \\ \sum_{k=-T+1}^{T-1}\hat{\gamma}_{yx}^{(k)}e^{-i\omega k} & \sum_{k=-T+1}^{T-1}\hat{\gamma}_{yy}^{(k)}e^{-i\omega k}\end{bmatrix} = \begin{bmatrix}\hat{s}_{xx}(\omega) & \hat{s}_{xy}(\omega) \\ \hat{s}_{yx}(\omega) & \hat{s}_{yy}(\omega)\end{bmatrix}.$$

Expression [10.4.31] states that the sample cross periodogram from x to y at frequency ω_j can be expressed as

$$\begin{aligned}\hat{s}_{yx}(\omega_j) &= [T/(8\pi)]\cdot(\hat{a}_j + i\cdot\hat{d}_j)(\hat{\alpha}_j - i\cdot\hat{\delta}_j) \\ &= [T/(8\pi)]\cdot(\hat{a}_j\hat{\alpha}_j + \hat{d}_j\hat{\delta}_j) + i\cdot[T/(8\pi)]\cdot(\hat{d}_j\hat{\alpha}_j - \hat{a}_j\hat{\delta}_j).\end{aligned}$$

The real component is the sample analog of the cospectrum, while the imaginary component is the sample analog of the quadrature spectrum:

$$\hat{s}_{yx}(\omega_j) = \hat{c}_{yx}(\omega_j) + i\cdot\hat{q}_{yx}(\omega_j), \qquad [10.4.32]$$

where

$$\hat{c}_{yx}(\omega_j) = [T/(8\pi)]\cdot(\hat{a}_j\hat{\alpha}_j + \hat{d}_j\hat{\delta}_j) \qquad [10.4.33]$$

$$\hat{q}_{yx}(\omega_j) = [T/(8\pi)]\cdot(\hat{d}_j\hat{\alpha}_j - \hat{a}_j\hat{\delta}_j). \qquad [10.4.34]$$

Comparing [10.4.33] with [10.4.28], the sample cospectrum evaluated at ω_j is proportional to the portion of the sample covariance between y and x that is attributable to cycles with frequency ω_j. The population cospectrum admits an analogous interpretation as the portion of the population covariance between Y and X attributable to cycles with frequency ω based on a multivariate version of the spectral representation theorem.

What interpretation are we to attach to the quadrature spectrum? Consider using the weights in [10.4.22] to construct a new series x_t^* by shifting the phase of each of the periodic functions by a quarter cycle:

$$\begin{aligned}x_t^* = \bar{x} + \sum_{j=1}^{M}\{&\hat{a}_j\cdot\cos[\omega_j(t-1) + (\pi/2)] \\ &+ \hat{d}_j\cdot\sin[\omega_j(t-1) + (\pi/2)]\}.\end{aligned} \qquad [10.4.35]$$

The variable x_t^* is driven by the same cycles as x_t, except that at date $t = 1$ each cycle is one-quarter of the way through rather than just beginning as in the case of x_t.

Since $\sin[\theta + (\pi/2)] = \cos(\theta)$ and since $\cos[\theta + (\pi/2)] = -\sin(\theta)$, the variable x_t^* can alternatively be described as

$$x_t^* = \bar{x} + \sum_{j=1}^{M}\{\hat{d}_j\cdot\cos[\omega_j(t-1)] - \hat{a}_j\cdot\sin[\omega_j(t-1)]\}. \qquad [10.4.36]$$

As in [10.4.27], the sample covariance between y_t and x_t^* is found to be

$$T^{-1} \sum_{t=1}^{T} (y_t - \bar{y})(x_t^* - \bar{x}) = (1/2) \sum_{j=1}^{M} (\hat{a}_j \hat{d}_j - \hat{\delta}_j \hat{a}_j).$$

Comparing this with [10.4.34], the sample quadrature spectrum from x to y at frequency ω_j is proportional to the portion of the sample covariance between x^* and y that is due to cycles of frequency ω_j. Cycles of frequency ω_j may be quite important for both x and y individually (as reflected by large values for $\hat{s}_{xx}(\omega)$ and $\hat{s}_{yy}(\omega)$) yet fail to produce much contemporaneous covariance between the variables because at any given date the two series are in a different phase of the cycle. For example, the variable x may respond to an economic recession sooner than y. The quadrature spectrum looks for evidence of such out-of-phase cycles.

Coherence, Phase, and Gain

The *population coherence* between X and Y is a measure of the degree to which X and Y are jointly influenced by cycles of frequency ω. This measure combines the inferences of the cospectrum and the quadrature spectrum, and is defined as[1]

$$h_{YX}(\omega) = \frac{[c_{YX}(\omega)]^2 + [q_{YX}(\omega)]^2}{s_{YY}(\omega)s_{XX}(\omega)},$$

assuming that $s_{YY}(\omega)$ and $s_{XX}(\omega)$ are nonzero. If $s_{YY}(\omega)$ or $s_{XX}(\omega)$ is zero, the coherence is defined to be zero. It can be shown that $0 \leq h_{YX}(\omega) \leq 1$ for all ω as long as X and Y are covariance-stationary with absolutely summable autocovariance matrices.[2] If $h_{YX}(\omega)$ is large, this indicates that Y and X have important cycles of frequency ω in common.

The cospectrum and quadrature spectrum can alternatively be described in polar coordinate form. In this notation, the population cross spectrum from X to Y is written as

$$s_{YX}(\omega) = c_{YX}(\omega) + i \cdot q_{YX}(\omega) = R(\omega) \cdot \exp[i \cdot \theta(\omega)], \qquad [10.4.37]$$

where

$$R(\omega) = \{[c_{YX}(\omega)]^2 + [q_{YX}(\omega)]^2\}^{1/2} \qquad [10.4.38]$$

and $\theta(\omega)$ represents the radian angle satisfying

$$\frac{\sin[\theta(\omega)]}{\cos[\theta(\omega)]} = \frac{q_{YX}(\omega)}{c_{YX}(\omega)}. \qquad [10.4.39]$$

The function $R(\omega)$ is sometimes described as the *gain* while $\theta(\omega)$ is called the *phase*.[3]

[1]The coherence is sometimes alternatively defined as the square root of this magnitude. The sample coherence based on the unsmoothed periodogram is identically equal to 1.

[2]See, for example, Fuller (1976, p. 156).

[3]The gain is sometimes alternatively defined as $R(\omega)/s_{XX}(\omega)$.

The Population Spectrum for Vector MA and AR Processes

Let $\mathbf{y}_t$ be a vector $MA(\infty)$ process with absolutely summable moving average coefficients:

$$\mathbf{y}_t = \boldsymbol{\mu} + \boldsymbol{\Psi}(L)\boldsymbol{\varepsilon}_t,$$

where

$$E(\boldsymbol{\varepsilon}_t\boldsymbol{\varepsilon}_\tau') = \begin{cases} \boldsymbol{\Omega} & \text{for } t = \tau \\ \mathbf{0} & \text{otherwise.} \end{cases}$$

Substituting [10.3.4] into [10.4.3] reveals that the population spectrum for $\mathbf{y}_t$ can be calculated as

$$\mathbf{s}_Y(\omega) = (2\pi)^{-1}[\boldsymbol{\Psi}(e^{-i\omega})]\boldsymbol{\Omega}[\boldsymbol{\Psi}(e^{i\omega})]'. \qquad [10.4.40]$$

For example, the population spectrum for a stationary $VAR(p)$ as written in [10.1.4] is

$$\begin{aligned}\mathbf{s}_Y(\omega) = (2\pi)^{-1}\{&\mathbf{I}_n - \boldsymbol{\Phi}_1 e^{-i\omega} - \boldsymbol{\Phi}_2 e^{-2i\omega} - \cdots - \boldsymbol{\Phi}_p e^{-pi\omega}\}^{-1}\boldsymbol{\Omega} \\ &\times \{\mathbf{I}_n - \boldsymbol{\Phi}_1' e^{i\omega} - \boldsymbol{\Phi}_2' e^{2i\omega} - \cdots - \boldsymbol{\Phi}_p' e^{pi\omega}\}^{-1}.\end{aligned} \qquad [10.4.41]$$

Estimating the Population Spectrum

If an observed time series $\mathbf{y}_1, \mathbf{y}_2, \ldots, \mathbf{y}_T$ can be reasonably described by a pth-order vector autoregression, one good approach to estimating the population spectrum is to estimate the parameters of the vector autoregression [10.1.4] by *OLS* and then substitute these parameter estimates into equation [10.4.41].

Alternatively, the sample cross periodogram from x to y at frequency $\omega_j = 2\pi j/T$ can be calculated from [10.4.32] to [10.4.34], where $\hat{\alpha}_j$, $\hat{\delta}_j$, $\hat{a}_j$, and $\hat{d}_j$ are as defined in [10.4.20] through [10.4.24]. One would want to smooth these to obtain a more useful estimate of the population cross spectrum. For example, one reasonable estimate of the population cospectrum between X and Y at frequency ω_j would be

$$\hat{c}_{YX}(\omega_j) = \sum_{m=-h}^{h} \left\{\frac{h + 1 - |m|}{(h+1)^2}\right\} \hat{c}_{yx}(\omega_{j+m}),$$

where $\hat{c}_{yx}(\omega_{j+m})$ denotes the estimate in [10.4.33] evaluated at frequency $\omega_{j+m} = 2\pi(j+m)/T$ and h is a bandwidth parameter reflecting how many different frequencies are to be used in estimating the cospectrum at frequency ω_j.

Another approach is to express the smoothing in terms of weighting coefficients κ_k^* to be applied to $\hat{\boldsymbol{\Gamma}}_k$ when the population autocovariances in expression [10.4.3] are replaced by sample autocovariances. Such an estimate would take the form

$$\hat{\mathbf{s}}_Y(\omega) = (2\pi)^{-1}\left\{\hat{\boldsymbol{\Gamma}}_0 + \sum_{k=1}^{T-1} \kappa_k^*[\hat{\boldsymbol{\Gamma}}_k e^{-i\omega k} + \hat{\boldsymbol{\Gamma}}_k' e^{i\omega k}]\right\}$$

where

$$\hat{\Gamma}_k = T^{-1} \sum_{t=k+1}^{T} (\mathbf{y}_t - \bar{\mathbf{y}})(\mathbf{y}_{t-k} - \bar{\mathbf{y}})'$$

$$\bar{\mathbf{y}} = T^{-1} \sum_{t=1}^{T} \mathbf{y}_t.$$

For example, the modified Bartlett estimate of the multivariate spectrum is

$$\hat{\mathbf{s}}_{\mathbf{Y}}(\omega) = (2\pi)^{-1} \left\{ \hat{\Gamma}_0 + \sum_{k=1}^{q} \left[1 - \frac{k}{q+1} \right] [\hat{\Gamma}_k e^{-i\omega k} + \hat{\Gamma}_k' e^{i\omega k}] \right\}. \quad [10.4.42]$$

Filters

Let $\mathbf{x}_t$ be an r-dimensional covariance-stationary process with absolutely summable autocovariances and with $(r \times r)$ population spectrum denoted $\mathbf{s}_{\mathbf{X}}(\omega)$. Let $\{\mathbf{H}_k\}_{k=-\infty}^{\infty}$ be an absolutely summable sequence of $(n \times r)$ matrices, and let $\mathbf{y}_t$ denote the n-dimensional vector process given by

$$\mathbf{y}_t = \mathbf{H}(L)\mathbf{x}_t = \sum_{k=-\infty}^{\infty} \mathbf{H}_k \mathbf{x}_{t-k}.$$

It follows from [10.3.9] that the population spectrum of $\mathbf{y}$ (denoted $\mathbf{s}_{\mathbf{Y}}(\omega)$) is related to that of $\mathbf{x}$ according to

$$\underset{(n\times n)}{\mathbf{s}_{\mathbf{Y}}(\omega)} = \underset{(n\times r)}{[\mathbf{H}(e^{-i\omega})]} \underset{(r\times r)}{\mathbf{s}_{\mathbf{X}}(\omega)} \underset{(r\times n)}{[\mathbf{H}(e^{i\omega})]'}. \quad [10.4.43]$$

As a special case of this result, let X_t be a univariate stationary stochastic process with continuous spectrum $s_X(\omega)$, and let u_t be a second univariate stationary stochastic process with continuous spectrum $s_U(\omega)$, where X_t and u_τ are uncorrelated for all t and τ. Thus, the population spectrum of the vector $\mathbf{x}_t \equiv (X_t, u_t)'$ is given by

$$\mathbf{s}_{\mathbf{X}}(\omega) = \begin{bmatrix} s_{XX}(\omega) & 0 \\ 0 & s_{UU}(\omega) \end{bmatrix}.$$

Define a new series Y_t according to

$$Y_t = \sum_{k=-\infty}^{\infty} h_k X_{t-k} + u_t \equiv h(L)X_t + u_t, \quad [10.4.44]$$

where $\{h_k\}_{k=-\infty}^{\infty}$ is absolutely summable. Note that the vector $\mathbf{y}_t \equiv (X_t, Y_t)'$ is obtained from the original vector $\mathbf{x}_t$ by the filter

$$\mathbf{y}_t = \mathbf{H}(L)\mathbf{x}_t,$$

where

$$\mathbf{H}(L) = \begin{bmatrix} 1 & 0 \\ h(L) & 1 \end{bmatrix}.$$

It follows from [10.4.43] that the spectrum of $\mathbf{y}$ is given by

$$\begin{aligned} \mathbf{s}_Y(\omega) &= \begin{bmatrix} 1 & 0 \\ h(e^{-i\omega}) & 1 \end{bmatrix} \begin{bmatrix} s_{XX}(\omega) & 0 \\ 0 & s_{UU}(\omega) \end{bmatrix} \begin{bmatrix} 1 & h(e^{i\omega}) \\ 0 & 1 \end{bmatrix} \\ &= \begin{bmatrix} s_{XX}(\omega) & s_{XX}(\omega)h(e^{i\omega}) \\ h(e^{-i\omega})s_{XX}(\omega) & h(e^{-i\omega})s_{XX}(\omega)h(e^{i\omega}) + s_{UU}(\omega) \end{bmatrix}, \end{aligned} \qquad [10.4.45]$$

where

$$h(e^{-i\omega}) = \sum_{k=-\infty}^{\infty} h_k e^{-i\omega k}. \qquad [10.4.46]$$

The lower left element of the matrix in [10.4.45] indicates that when Y_t and X_t are related according to [10.4.44], the cross spectrum from X to Y can be calculated by multiplying [10.4.46] by the spectrum of X.

We can also imagine going through these steps in reverse order. Specifically, suppose we are given an observed vector $\mathbf{y}_t = (X_t, Y_t)'$ with absolutely summable autocovariance matrices and with population spectrum given by

$$\mathbf{s}_Y(\omega) = \begin{bmatrix} s_{XX}(\omega) & s_{XY}(\omega) \\ s_{YX}(\omega) & s_{YY}(\omega) \end{bmatrix}. \qquad [10.4.47]$$

Then the linear projection of Y_t on $\{X_{t-k}\}_{k=-\infty}^{\infty}$ exists and is of the form of [10.4.44], where u_t would now be regarded as the population residual associated with the linear projection. The sequence of linear projection coefficients $\{h_k\}_{k=-\infty}^{\infty}$ can be summarized in terms of the function of ω given in [10.4.46]. Comparing the lower left elements of [10.4.47] and [10.4.45], this function must satisfy

$$h(e^{-i\omega})s_{XX}(\omega) = s_{YX}(\omega).$$

In other words, the function $h(e^{-i\omega})$ can be calculated from

$$h(e^{-i\omega}) = \frac{s_{YX}(\omega)}{s_{XX}(\omega)}, \qquad [10.4.48]$$

assuming that $s_{XX}(\omega)$ is not zero. When $s_{XX}(\omega) = 0$, we set $h(e^{-i\omega}) = 0$. This magnitude, the ratio of the cross spectrum from X to Y to the spectrum of X, is known as the *transfer function* from X to Y.

The principles underlying [10.4.4] can further be used to uncover individual transfer function coefficients:

$$h_k = (2\pi)^{-1} \int_{-\pi}^{\pi} h(e^{-i\omega})e^{i\omega k}\, d\omega.$$

In other words, given an observed vector $(X_t, Y_t)'$ with absolutely summable autocovariance matrices and thus with continuous population spectrum of the form of [10.4.47], the coefficient on X_{t-k} in the population linear projection of Y_t on $\{X_{t-k}\}_{k=-\infty}^{\infty}$ can be calculated from

$$h_k = (2\pi)^{-1} \int_{-\pi}^{\pi} \frac{s_{YX}(\omega)}{s_{XX}(\omega)} e^{i\omega k}\, d\omega. \qquad [10.4.49]$$

10.5. *The Sample Mean of a Vector Process*

Variance of the Sample Mean

Suppose we have a sample of size T, $\{\mathbf{y}_1, \mathbf{y}_2, \ldots, \mathbf{y}_T\}$, drawn from an n-dimensional covariance-stationary process with

$$E(\mathbf{y}_t) = \boldsymbol{\mu} \qquad [10.5.1]$$

$$E[(\mathbf{y}_t - \boldsymbol{\mu})(\mathbf{y}_{t-j} - \boldsymbol{\mu})'] = \boldsymbol{\Gamma}_j. \qquad [10.5.2]$$

Consider the properties of the sample mean,

$$\bar{\mathbf{y}}_T = (1/T)\sum_{t=1}^{T} \mathbf{y}_t. \qquad [10.5.3]$$

As in the discussion in Section 7.2 of the sample mean of a scalar process, it is clear that $E(\bar{\mathbf{y}}_T) = \boldsymbol{\mu}$ and

$$\begin{aligned}
&E[(\bar{\mathbf{y}}_T - \boldsymbol{\mu})(\bar{\mathbf{y}}_T - \boldsymbol{\mu})'] \\
&\quad = (1/T^2)E\{(\mathbf{y}_1 - \boldsymbol{\mu})[(\mathbf{y}_1 - \boldsymbol{\mu})' + (\mathbf{y}_2 - \boldsymbol{\mu})' + \cdots + (\mathbf{y}_T - \boldsymbol{\mu})'] \\
&\qquad + (\mathbf{y}_2 - \boldsymbol{\mu})[(\mathbf{y}_1 - \boldsymbol{\mu})' + (\mathbf{y}_2 - \boldsymbol{\mu})' + \cdots + (\mathbf{y}_T - \boldsymbol{\mu})'] \\
&\qquad + (\mathbf{y}_3 - \boldsymbol{\mu})[(\mathbf{y}_1 - \boldsymbol{\mu})' + (\mathbf{y}_2 - \boldsymbol{\mu})' + \cdots + (\mathbf{y}_T - \boldsymbol{\mu})'] \\
&\qquad + \cdots + (\mathbf{y}_T - \boldsymbol{\mu})[(\mathbf{y}_1 - \boldsymbol{\mu})' + (\mathbf{y}_2 - \boldsymbol{\mu})' + \cdots + (\mathbf{y}_T - \boldsymbol{\mu})']\} \\
&\quad = (1/T^2)\{[\boldsymbol{\Gamma}_0 + \boldsymbol{\Gamma}_{-1} + \cdots + \boldsymbol{\Gamma}_{-(T-1)}] \qquad [10.5.4] \\
&\qquad + [\boldsymbol{\Gamma}_1 + \boldsymbol{\Gamma}_0 + \boldsymbol{\Gamma}_{-1} + \cdots + \boldsymbol{\Gamma}_{-(T-2)}] \\
&\qquad + [\boldsymbol{\Gamma}_2 + \boldsymbol{\Gamma}_1 + \boldsymbol{\Gamma}_0 + \boldsymbol{\Gamma}_{-1} + \cdots + \boldsymbol{\Gamma}_{-(T-3)}] \\
&\qquad + \cdots + [\boldsymbol{\Gamma}_{T-1} + \boldsymbol{\Gamma}_{T-2} + \boldsymbol{\Gamma}_{T-3} + \cdots + \boldsymbol{\Gamma}_0]\} \\
&\quad = (1/T^2)\{T\boldsymbol{\Gamma}_0 + (T-1)\boldsymbol{\Gamma}_1 + (T-2)\boldsymbol{\Gamma}_2 + \cdots + \boldsymbol{\Gamma}_{T-1} \\
&\qquad + (T-1)\boldsymbol{\Gamma}_{-1} + (T-2)\boldsymbol{\Gamma}_{-2} + \cdots + \boldsymbol{\Gamma}_{-(T-1)}\}.
\end{aligned}$$

Thus,

$$\begin{aligned}
&T\cdot E[(\bar{\mathbf{y}}_T - \boldsymbol{\mu})(\bar{\mathbf{y}}_T - \boldsymbol{\mu})'] \\
&\quad = \boldsymbol{\Gamma}_0 + [(T-1)/T]\boldsymbol{\Gamma}_1 + [(T-2)/T]\boldsymbol{\Gamma}_2 + \cdots \qquad [10.5.5] \\
&\qquad + [1/T]\boldsymbol{\Gamma}_{T-1} + [(T-1)/T]\boldsymbol{\Gamma}_{-1} + [(T-2)/T]\boldsymbol{\Gamma}_{-2} \\
&\qquad + \cdots + [1/T]\boldsymbol{\Gamma}_{-(T-1)}.
\end{aligned}$$

As in the univariate case, the weights on $\boldsymbol{\Gamma}_k$ for $|k|$ small go to unity as $T \to \infty$, and higher autocovariances go to zero for a covariance-stationary process. Hence, we have the following generalization of Proposition 7.5.

Proposition 10.5: *Let $\mathbf{y}_t$ be a covariance-stationary process with moments given by [10.5.1] and [10.5.2] and with absolutely summable autocovariances. Then the sample mean [10.5.3] satisfies*

(*a*) $\bar{\mathbf{y}}_T \xrightarrow{p} \boldsymbol{\mu}$

(*b*) $\lim_{T\to\infty} \{T\cdot E[(\bar{\mathbf{y}}_T - \boldsymbol{\mu})(\bar{\mathbf{y}}_T - \boldsymbol{\mu})']\} = \sum_{\nu=-\infty}^{\infty} \boldsymbol{\Gamma}_\nu.$

The proof of Proposition 10.5 is virtually identical to that of Proposition 7.5. Consider the following $(n \times n)$ matrix:

$$\sum_{\nu=-\infty}^{\infty} \Gamma_\nu - T \cdot E[(\bar{\mathbf{y}}_T - \boldsymbol{\mu})(\bar{\mathbf{y}}_T - \boldsymbol{\mu})'] = \sum_{|\nu| \geq T} \Gamma_\nu + \sum_{\nu=-(T-1)}^{T-1} (|\nu|/T)\Gamma_\nu, \qquad [10.5.6]$$

where the equality follows from [10.5.5]. Let $\gamma_{ij}^{(\nu)}$ denote the row i, column j element of Γ_ν. The row i, column j element of the matrix in [10.5.6] can then be written

$$\sum_{|\nu| \geq T} \gamma_{ij}^{(\nu)} + \sum_{\nu=-(T-1)}^{T-1} (|\nu|/T)\gamma_{ij}^{(\nu)}.$$

Absolutely summability of $\{\Gamma_\nu\}_{\nu=-\infty}^{\infty}$ implies that for any $\varepsilon > 0$ there exists a q such that

$$\sum_{|\nu| > q} |\gamma_{ij}^{(\nu)}| < \varepsilon/2.$$

Thus,

$$\left| \sum_{|\nu| \geq T} \gamma_{ij}^{(\nu)} + \sum_{\nu=-(T-1)}^{T-1} (|\nu|/T)\gamma_{ij}^{(\nu)} \right| < \varepsilon/2 + \sum_{\nu=-q}^{q} (|\nu|/T)|\gamma_{ij}^{(\nu)}|.$$

This sum can be made less than ε by choosing T sufficiently large. This establishes claim (b) of Proposition 10.5. From this result, $E(\bar{y}_{i,T} - \mu_i)^2 \to 0$ for each i, implying that $\bar{y}_{i,T} \xrightarrow{p} \mu_i$.

Consistent Estimation of T *Times the Variance of the Sample Mean*

Hypothesis tests about the sample mean require an estimate of the matrix in result (b) of Proposition 10.5. Let S represent this matrix:

$$\mathbf{S} \equiv \lim_{T \to \infty} T \cdot E[(\bar{\mathbf{y}}_T - \boldsymbol{\mu})(\bar{\mathbf{y}}_T - \boldsymbol{\mu})']. \qquad [10.5.7]$$

If the data were generated by a vector $MA(q)$ process, then result (b) would imply

$$\mathbf{S} = \sum_{\nu=-q}^{q} \Gamma_\nu. \qquad [10.5.8]$$

A natural estimate then is

$$\hat{\mathbf{S}} = \hat{\Gamma}_0 + \sum_{\nu=1}^{q} (\hat{\Gamma}_\nu + \hat{\Gamma}_\nu'), \qquad [10.5.9]$$

where

$$\hat{\Gamma}_\nu = (1/T) \sum_{t=\nu+1}^{T} (\mathbf{y}_t - \bar{\mathbf{y}})(\mathbf{y}_{t-\nu} - \bar{\mathbf{y}})'. \qquad [10.5.10]$$

As long as $\mathbf{y}_t$ is ergodic for second moments, [10.5.9] gives a consistent estimate of [10.5.8]. Indeed, Hansen (1982) and White (1984, Chapter 6) noted that [10.5.9] gives a consistent estimate of the asymptotic variance of the sample mean for a broad class of processes exhibiting time-dependent heteroskedasticity and autocorrelation. To see why, note that for a process satisfying $E(\mathbf{y}_t) = \boldsymbol{\mu}$ with

time-varying second moments, the variance of the sample mean is given by

$$\begin{aligned} &E[(\bar{\mathbf{y}}_T - \boldsymbol{\mu})(\bar{\mathbf{y}}_T - \boldsymbol{\mu})'] \\ &= E\left[(1/T)\sum_{t=1}^{T}(\mathbf{y}_t - \boldsymbol{\mu})\right]\left[(1/T)\sum_{s=1}^{T}(\mathbf{y}_s - \boldsymbol{\mu})\right]' \qquad [10.5.11] \\ &= (1/T^2)\sum_{t=1}^{T}\sum_{s=1}^{T} E[(\mathbf{y}_t - \boldsymbol{\mu})(\mathbf{y}_s - \boldsymbol{\mu})'] . \end{aligned}$$

Suppose, first, that $E[(\mathbf{y}_t - \boldsymbol{\mu})(\mathbf{y}_s - \boldsymbol{\mu})'] = \mathbf{0}$ for $|t - s| > q$, as was the case for the vector $MA(q)$ process, though we generalize from the $MA(q)$ process to allow $E[(\mathbf{y}_t - \boldsymbol{\mu})(\mathbf{y}_s - \boldsymbol{\mu})']$ to be a function of t for $|t - s| \leq q$. Then [10.5.11] implies

$$\begin{aligned} &T \cdot E[(\bar{\mathbf{y}}_T - \boldsymbol{\mu})(\bar{\mathbf{y}}_T - \boldsymbol{\mu})'] \\ &= (1/T)\sum_{t=1}^{T} E[(\mathbf{y}_t - \boldsymbol{\mu})(\mathbf{y}_t - \boldsymbol{\mu})'] \\ &\quad + (1/T)\sum_{t=2}^{T}\{E[(\mathbf{y}_t - \boldsymbol{\mu})(\mathbf{y}_{t-1} - \boldsymbol{\mu})'] + E[(\mathbf{y}_{t-1} - \boldsymbol{\mu})(\mathbf{y}_t - \boldsymbol{\mu})']\} \\ &\quad + (1/T)\sum_{t=3}^{T}\{E[(\mathbf{y}_t - \boldsymbol{\mu})(\mathbf{y}_{t-2} - \boldsymbol{\mu})'] + E[(\mathbf{y}_{t-2} - \boldsymbol{\mu})(\mathbf{y}_t - \boldsymbol{\mu})']\} + \cdots \\ &\quad + (1/T)\sum_{t=q+1}^{T}\{E[(\mathbf{y}_t - \boldsymbol{\mu})(\mathbf{y}_{t-q} - \boldsymbol{\mu})'] + E[(\mathbf{y}_{t-q} - \boldsymbol{\mu})(\mathbf{y}_t - \boldsymbol{\mu})']\}. \qquad [10.5.12] \end{aligned}$$

The estimate [10.5.9] replaces

$$(1/T)\sum_{t=v+1}^{T} E[(\mathbf{y}_t - \boldsymbol{\mu})(\mathbf{y}_{t-v} - \boldsymbol{\mu})'] \qquad [10.5.13]$$

in [10.5.12] with

$$(1/T)\sum_{t=v+1}^{T}(\mathbf{y}_t - \bar{\mathbf{y}}_T)(\mathbf{y}_{t-v} - \bar{\mathbf{y}}_T)', \qquad [10.5.14]$$

and thus [10.5.9] provides a consistent estimate of the limit of [10.5.12] whenever [10.5.14] converges in probability to [10.5.13]. Hence, the estimator proposed in [10.5.9] can give a consistent estimate of T times the variance of the sample mean in the presence of both heteroskedasticity and autocorrelation up through order q.

More generally, even if $E[(\mathbf{y}_t - \boldsymbol{\mu})(\mathbf{y}_s - \boldsymbol{\mu})']$ is nonzero for all t and s, as long as this matrix goes to zero sufficiently quickly as $|t - s| \to \infty$, then there is still a sense in which $\hat{\mathbf{S}}_T$ in [10.5.9] can provide a consistent estimate of $\mathbf{S}$. Specifically, if, as the sample size T grows, a larger number of sample autocovariances q is used to form the estimate, then $\hat{\mathbf{S}}_T \xrightarrow{p} \mathbf{S}$ (see White, 1984, p. 155).

The Newey-West Estimator

Although [10.5.9] gives a consistent estimate of $\mathbf{S}$, it has the drawback that [10.5.9] need not be positive semidefinite in small samples. If $\hat{\mathbf{S}}$ is not positive semidefinite, then some linear combination of the elements of $\bar{\mathbf{y}}$ is asserted to have a negative variance, a considerable handicap in forming a hypothesis test!

Newey and West (1987) suggested the alternative estimate

$$\tilde{\mathbf{S}} = \hat{\boldsymbol{\Gamma}}_0 + \sum_{v=1}^{q}\left[1 - \frac{v}{q+1}\right](\hat{\boldsymbol{\Gamma}}_v + \hat{\boldsymbol{\Gamma}}_v'), \qquad [10.5.15]$$

where $\hat{\Gamma}_\nu$ is given by [10.5.10]. For example, for $q = 2$,

$$\tilde{S} = \hat{\Gamma}_0 + \tfrac{2}{3}(\hat{\Gamma}_1 + \hat{\Gamma}_1') + \tfrac{1}{3}(\hat{\Gamma}_2 + \hat{\Gamma}_2').$$

Newey and West showed that $\tilde{S}$ is positive semidefinite by construction and has the same consistency properties that were noted for $\hat{S}$, namely, that if q and T both go to infinity with $q/T^{1/4} \to 0$, then $\tilde{S}_T \xrightarrow{p} S$.

Application: Autocorrelation- and Heteroskedasticity-Consistent Standard Errors for Linear Regressions

As an application of using the Newey-West weighting, consider the linear regression model

$$y_t = x_t'\beta + u_t$$

for x_t a $(k \times 1)$ vector of explanatory variables. Recall from equation [8.2.6] that the deviation of the *OLS* estimate b_T from the true value β satisfies

$$\sqrt{T}(b_T - \beta) = \left[(1/T)\sum_{t=1}^{T} x_t x_t'\right]^{-1}\left[(1/\sqrt{T})\sum_{t=1}^{T} x_t u_t\right]. \qquad [10.5.16]$$

In calculating the asymptotic distribution of the *OLS* estimate b_T, we usually assume that the first term in [10.5.16] converges in probability to Q^{-1}:

$$\left[(1/T)\sum_{t=1}^{T} x_t x_t'\right]^{-1} \xrightarrow{p} Q^{-1}. \qquad [10.5.17]$$

The second term in [10.5.16] can be viewed as $\sqrt{T}$ times the sample mean of the $(k \times 1)$ vector $x_t u_t$:

$$\begin{aligned}\left[(1/\sqrt{T})\sum_{t=1}^{T} x_t u_t\right] &= (\sqrt{T})(1/T)\sum_{t=1}^{T} y_t, \\ &= \sqrt{T}\cdot\bar{y}_T,\end{aligned} \qquad [10.5.18]$$

where $y_t \equiv x_t u_t$. Provided that $E(u_t|x_t) = 0$, the vector y_t has mean zero. We can allow for conditional heteroskedasticity, autocorrelation, and time variation in the second moments of y_t, as long as

$$S \equiv \lim_{T\to\infty} T\cdot E(\bar{y}_T\bar{y}_T')$$

exists. Under general conditions,[4] it then turns out that

$$\left[(1/\sqrt{T})\sum_{t=1}^{T} x_t u_t\right] = \sqrt{T}\cdot\bar{y}_T \xrightarrow{L} N(0, S).$$

Substituting this and [10.5.17] into [10.5.16],

$$\sqrt{T}(b_T - \beta) \xrightarrow{L} N(0, Q^{-1}SQ^{-1}). \qquad [10.5.19]$$

In light of the foregoing discussion, we might hope to estimate S by

$$\hat{S}_T = \hat{\Gamma}_{0,T} + \sum_{\nu=1}^{q}\left[1 - \frac{\nu}{q+1}\right](\hat{\Gamma}_{\nu,T} + \hat{\Gamma}_{\nu,T}'). \qquad [10.5.20]$$

[4]See, for example, White (1984, p. 119).

Here,

$$\hat{\Gamma}_{\nu,T} = (1/T) \sum_{t=\nu+1}^{T} (\mathbf{x}_t \hat{u}_{t,T} \hat{u}_{t-\nu,T} \mathbf{x}'_{t-\nu}),$$

$\hat{u}_{t,T}$ is the *OLS* residual for date t in a sample of size T ($\hat{u}_{t,T} = y_t - \mathbf{x}'_t \mathbf{b}_T$), and q is a lag length beyond which we are willing to assume that the correlation between $\mathbf{x}_t u_t$ and $\mathbf{x}_{t-\nu} u_{t-\nu}$ is essentially zero. Clearly, $\mathbf{Q}$ is consistently estimated by $\hat{\mathbf{Q}}_T = (1/T)\Sigma_{t=1}^{T} \mathbf{x}_t \mathbf{x}'_t$. Substituting $\hat{\mathbf{Q}}_T$ and $\hat{\mathbf{S}}_T$ into [10.5.19], the suggestion is to treat the *OLS* estimate $\mathbf{b}_T$ as if

$$\mathbf{b}_T \approx N(\boldsymbol{\beta}, (\hat{\mathbf{V}}_T/T))$$

where

$$\begin{aligned}
\hat{\mathbf{V}}_T &= \hat{\mathbf{Q}}_T^{-1} \hat{\mathbf{S}}_T \hat{\mathbf{Q}}_T^{-1} \\
&= \left[(1/T) \sum_{t=1}^{T} \mathbf{x}_t \mathbf{x}'_t\right]^{-1} (1/T) \left[\sum_{t=1}^{T} \hat{u}_t^2 \mathbf{x}_t \mathbf{x}'_t \right. \\
&\quad \left. + \sum_{\nu=1}^{q} \left[1 - \frac{\nu}{q+1}\right] \sum_{t=\nu+1}^{T} (\mathbf{x}_t \hat{u}_t \hat{u}_{t-\nu} \mathbf{x}'_{t-\nu} + \mathbf{x}_{t-\nu} \hat{u}_{t-\nu} \hat{u}_t \mathbf{x}'_t)\right] \\
&\quad \times \left[(1/T) \sum_{t=1}^{T} \mathbf{x}_t \mathbf{x}'_t\right]^{-1},
\end{aligned}$$

that is, the variance of $\mathbf{b}_T$ is approximated by

$$\begin{aligned}
&(\hat{\mathbf{V}}_T/T) \\
&= \left[\sum_{t=1}^{T} \mathbf{x}_t \mathbf{x}'_t\right]^{-1} \left[\sum_{t=1}^{T} \hat{u}_t^2 \mathbf{x}_t \mathbf{x}'_t \right. \\
&\quad \left. + \sum_{\nu=1}^{q} \left[1 - \frac{\nu}{q+1}\right] \sum_{t=\nu+1}^{T} (\mathbf{x}_t \hat{u}_t \hat{u}_{t-\nu} \mathbf{x}'_{t-\nu} + \mathbf{x}_{t-\nu} \hat{u}_{t-\nu} \hat{u}_t \mathbf{x}'_t)\right] \left[\sum_{t=1}^{T} \mathbf{x}_t \mathbf{x}'_t\right]^{-1}
\end{aligned} \tag{10.5.21}$$

where $\hat{u}_t$ is the *OLS* sample residual. The square root of the row i, column i element of $\hat{\mathbf{V}}_T/T$ is known as a *heteroskedasticity- and autocorrelation-consistent standard error* for the ith element of the estimated *OLS* coefficient vector. The hope is that standard errors based on [10.5.21] will be robust to a variety of forms of heteroskedasticity and autocorrelation of the residuals u_t of the regression.

Spectral-Based Estimators

A number of alternative estimates of $\mathbf{S}$ in [10.5.7] have been suggested in the literature. Notice that as in the univariate case discussed in Section 7.2, if $\mathbf{y}_t$ is covariance-stationary, then $\mathbf{S}$ has the interpretation as the autocovariance-generating function $\mathbf{G}_\mathbf{Y}(z) = \Sigma_{\nu=-\infty}^{\infty} \boldsymbol{\Gamma}_\nu z^\nu$ evaluated at $z = 1$, or, equivalently, as 2π times the population spectrum at frequency zero:

$$\mathbf{S} = \sum_{\nu=-\infty}^{\infty} \boldsymbol{\Gamma}_\nu = 2\pi \mathbf{s}_\mathbf{Y}(0).$$

Indeed, the Newey-West estimator [10.5.15] is numerically identical to 2π times the Bartlett estimate of the multivariate spectrum described in [10.4.42] evaluated at frequency $\omega = 0$. Gallant (1987, p. 533) proposed a similar estimator based on a Parzen kernel,

$$\hat{\mathbf{S}} = \hat{\boldsymbol{\Gamma}}_0 + \sum_{\nu=1}^{q} k[\nu/(q+1)](\hat{\boldsymbol{\Gamma}}_\nu + \hat{\boldsymbol{\Gamma}}'_\nu),$$

where

$$k(z) = \begin{cases} 1 - 6z^2 + 6z^3 & \text{for } 0 \le z \le \frac{1}{2} \\ 2(1 - z)^3 & \text{for } \frac{1}{2} \le z \le 1 \\ 0 & \text{otherwise.} \end{cases}$$

For example, for $q = 2$, we have

$$\hat{\mathbf{S}} = \hat{\boldsymbol{\Gamma}}_0 + \tfrac{5}{9}(\hat{\boldsymbol{\Gamma}}_1 + \hat{\boldsymbol{\Gamma}}_1') + \tfrac{2}{27}(\hat{\boldsymbol{\Gamma}}_2 + \hat{\boldsymbol{\Gamma}}_2').$$

Andrews (1991) examined a number of alternative estimators and found the best results for a quadratic spectral kernel:

$$k(z) = \frac{3}{(6\pi z/5)^2}\left[\frac{\sin(6\pi z/5)}{6\pi z/5} - \cos(6\pi z/5)\right].$$

In contrast to the Newey-West and Gallant estimators, Andrews's suggestion makes use of all $T - 1$ estimated autocovariance estimators:

$$\hat{\mathbf{S}} = \frac{T}{T - k}\left[\hat{\boldsymbol{\Gamma}}_0 + \sum_{\nu=1}^{T-1} k\left(\frac{\nu}{q + 1}\right)(\hat{\boldsymbol{\Gamma}}_\nu + \hat{\boldsymbol{\Gamma}}_\nu')\right]. \qquad [10.5.22]$$

Even though [10.5.22] makes use of all computed autocovariances, there is still a bandwidth parameter q to be chosen for constructing the kernel. For example, for $q = 2$,

$$\hat{\boldsymbol{\Gamma}}_0 + \sum_{\nu=1}^{T-1} k(\nu/3)(\hat{\boldsymbol{\Gamma}}_\nu + \hat{\boldsymbol{\Gamma}}_\nu') = \hat{\boldsymbol{\Gamma}}_0 + 0.85(\hat{\boldsymbol{\Gamma}}_1 + \hat{\boldsymbol{\Gamma}}_1')$$
$$+ 0.50(\hat{\boldsymbol{\Gamma}}_2 + \hat{\boldsymbol{\Gamma}}_2') + 0.14(\hat{\boldsymbol{\Gamma}}_3 + \hat{\boldsymbol{\Gamma}}_3') + \cdots.$$

Andrews recommended multiplying the estimate by $T/(T - k)$, where $\mathbf{y}_t = \mathbf{x}_t\hat{u}_t$ for $\hat{u}_t$ the sample *OLS* residual from a regression with k explanatory variables. Andrews (1991) and Newey and West (1992) also offered some guidance for choosing an optimal value of the lag truncation or bandwidth parameter q for each of the estimators of $\mathbf{S}$ that have been discussed here.

The estimators that have been described will work best when $\mathbf{y}_t$ has a finite moving average representation. Andrews and Monahan (1992) suggested an alternative approach to estimating $\mathbf{S}$ that also takes advantage of any autoregressive structure to the errors. Let $\mathbf{y}_t$ be a zero-mean vector, and let $\mathbf{S}$ be the asymptotic variance of the sample mean of $\mathbf{y}$. For example, if we want to calculate heteroskedasticity- and autocorrelation-consistent standard errors for *OLS* estimation, $\mathbf{y}_t$ would correspond to $\mathbf{x}_t\hat{u}_t$ where $\mathbf{x}_t$ is the vector of explanatory variables for the regression and $\hat{u}_t$ is the *OLS* residual. The first step in estimating $\mathbf{S}$ is to fit a low-order *VAR* for $\mathbf{y}_t$,

$$\mathbf{y}_t = \boldsymbol{\Phi}_1\mathbf{y}_{t-1} + \boldsymbol{\Phi}_2\mathbf{y}_{t-2} + \cdots + \boldsymbol{\Phi}_p\mathbf{y}_{t-p} + \mathbf{v}_t, \qquad [10.5.23]$$

where $\mathbf{v}_t$ is presumed to have some residual autocorrelation not entirely captured by the *VAR*. Note that since $\mathbf{y}_t$ has mean zero, no constant term is included in [10.5.23]. The ith row represented in [10.5.23] can be estimated by an *OLS* regression of the ith element of $\mathbf{y}_t$ on p lags of all the elements of $\mathbf{y}$, though if any eigenvalue of $|\mathbf{I}_n\lambda^p - \hat{\boldsymbol{\Phi}}_1\lambda^{p-1} - \hat{\boldsymbol{\Phi}}_2\lambda^{p-2} - \cdots - \hat{\boldsymbol{\Phi}}_p| = 0$ is too close to the unit circle (say, greater than 0.97 in modulus), Andrews and Monahan (1992, p. 957) recommended altering the *OLS* estimates so as to reduce the largest eigenvalue.

The second step in the Andrews and Monahan procedure is to calculate an estimate $\mathbf{S}^*$ using one of the methods described previously based on the fitted

residuals $\hat{\mathbf{v}}_t$ from [10.5.23]. For example,

$$\hat{\mathbf{S}}_T^* = \hat{\mathbf{\Gamma}}_0^* + \sum_{\nu=1}^{q}\left[1 - \frac{\nu}{q+1}\right](\hat{\mathbf{\Gamma}}_\nu^* + \hat{\mathbf{\Gamma}}_\nu^{*\prime}), \qquad [10.5.24]$$

where

$$\hat{\mathbf{\Gamma}}_\nu^* = (1/T)\sum_{t=\nu+1}^{T}\hat{\mathbf{v}}_t\hat{\mathbf{v}}_{t-\nu}'$$

and where q is a parameter representing the maximal order of autocorrelation assumed for $\mathbf{v}_t$. The matrix $\hat{\mathbf{S}}_T^*$ will be recognized as an estimate of $2\pi\cdot\mathbf{s}_\mathbf{V}(0)$, where $\mathbf{s}_\mathbf{V}(\omega)$ is the spectral density of $\mathbf{v}$:

$$\mathbf{s}_\mathbf{V}(\omega) = (2\pi)^{-1}\sum_{\nu=-\infty}^{\infty}\{E(\mathbf{v}_t\mathbf{v}_{t-\nu}')\}e^{-i\omega\nu}.$$

Notice that the original series $\mathbf{y}_t$ can be obtained from $\mathbf{v}_t$ by applying the following filter:

$$\mathbf{y}_t = [\mathbf{I}_n - \mathbf{\Phi}_1 L - \mathbf{\Phi}_2 L^2 - \cdots - \mathbf{\Phi}_p L^p]^{-1}\mathbf{v}_t.$$

Thus, from [10.4.43], the spectral density of $\mathbf{y}$ is related to the spectral density of $\mathbf{v}$ according to

$$\begin{aligned}\mathbf{s}_\mathbf{Y}(\omega) = {} & \{[\mathbf{I}_n - \mathbf{\Phi}_1 e^{-i\omega} - \mathbf{\Phi}_2 e^{-2i\omega} - \cdots - \mathbf{\Phi}_p e^{-pi\omega}]\}^{-1}\mathbf{s}_\mathbf{V}(\omega)\\ & \times \{[\mathbf{I}_n - \mathbf{\Phi}_1 e^{i\omega} - \mathbf{\Phi}_2 e^{2i\omega} - \cdots - \mathbf{\Phi}_p e^{pi\omega}]'\}^{-1}.\end{aligned}$$

Hence, an estimate of 2π times the spectral density of $\mathbf{y}$ at frequency zero is given by

$$\begin{aligned}\hat{\mathbf{S}}_T = {} & \{[\mathbf{I}_n - \hat{\mathbf{\Phi}}_1 - \hat{\mathbf{\Phi}}_2 - \cdots - \hat{\mathbf{\Phi}}_p]\}^{-1}\hat{\mathbf{S}}_T^*\\ & \times \{[\mathbf{I}_n - \hat{\mathbf{\Phi}}_1 - \hat{\mathbf{\Phi}}_2 - \cdots - \hat{\mathbf{\Phi}}_p]'\}^{-1},\end{aligned} \qquad [10.5.25]$$

where $\hat{\mathbf{S}}_T^*$ is calculated from [10.5.24]. The matrix $\hat{\mathbf{S}}_T$ in [10.5.25] is the Andrews-Monahan (1992) estimate of $\mathbf{S}$, where

$$\mathbf{S} = \lim_{T\to\infty} T\cdot E(\bar{\mathbf{y}}_T\bar{\mathbf{y}}_T').$$

APPENDIX 10.A. *Proofs of Chapter 10 Propositions*

■ **Proof of Proposition 10.1.** The eigenvalues of $\mathbf{F}$ are the values of λ for which the following determinant is zero:

$$\begin{vmatrix} (\mathbf{\Phi}_1 - \lambda\mathbf{I}_n) & \mathbf{\Phi}_2 & \mathbf{\Phi}_3 & \cdots & \mathbf{\Phi}_{p-1} & \mathbf{\Phi}_p \\ \mathbf{I}_n & -\lambda\mathbf{I}_n & \mathbf{0} & \cdots & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{I}_n & -\lambda\mathbf{I}_n & \cdots & \mathbf{0} & \mathbf{0} \\ \vdots & \vdots & \vdots & \cdots & \vdots & \vdots \\ \mathbf{0} & \mathbf{0} & \mathbf{0} & \cdots & \mathbf{I}_n & -\lambda\mathbf{I}_n \end{vmatrix}. \qquad [10.A.1]$$

Multiply each of the final block of n columns by $(1/\lambda)$ and add to the previous block. Multiply each of the n columns of this resulting next-to-final block by $(1/\lambda)$ and add the result to the third-to-last block of columns. Proceeding in this manner reveals [10.A.1] to be the same as

$$\begin{vmatrix} \mathbf{X}_1 & \mathbf{X}_2 \\ \mathbf{0} & -\lambda\mathbf{I}_{n(p-1)} \end{vmatrix}, \qquad [10.A.2]$$

where $\mathbf{X}_1$ denotes the following $(n \times n)$ matrix:

$$\mathbf{X}_1 \equiv (\mathbf{\Phi}_1 - \lambda\mathbf{I}_n) + (\mathbf{\Phi}_2/\lambda) + (\mathbf{\Phi}_3/\lambda^2) + \cdots + (\mathbf{\Phi}_p/\lambda^{p-1})$$

and $\mathbf{X}_2$ is a related $[n \times n(p-1)]$ matrix. Let $\mathbf{S}$ denote the following $(np \times np)$ matrix:

$$\mathbf{S} = \begin{bmatrix} \mathbf{0} & \mathbf{I}_{n(p-1)} \\ \mathbf{I}_n & \mathbf{0} \end{bmatrix},$$

and note that its inverse is given by

$$\mathbf{S}^{-1} = \begin{bmatrix} \mathbf{0} & \mathbf{I}_n \\ \mathbf{I}_{n(p-1)} & \mathbf{0} \end{bmatrix},$$

as may be verified by direct multiplication. Premultiplying a matrix by $\mathbf{S}$ and postmultiplying by $\mathbf{S}^{-1}$ will not change the determinant. Thus, [10.A.2] is equal to

$$\left| \begin{bmatrix} \mathbf{0} & \mathbf{I}_{n(p-1)} \\ \mathbf{I}_n & \mathbf{0} \end{bmatrix} \begin{bmatrix} \mathbf{X}_1 & \mathbf{X}_2 \\ \mathbf{0} & -\lambda\mathbf{I}_{n(p-1)} \end{bmatrix} \begin{bmatrix} \mathbf{0} & \mathbf{I}_n \\ \mathbf{I}_{n(p-1)} & \mathbf{0} \end{bmatrix} \right| = \begin{vmatrix} -\lambda\mathbf{I}_{n(p-1)} & \mathbf{0} \\ \mathbf{X}_2 & \mathbf{X}_1 \end{vmatrix}. \qquad [10.A.3]$$

Applying the formula for calculating a determinant [A.4.5] recursively, [10.A.3] is equal to

$$\begin{aligned}(-\lambda)^{n(p-1)}|\mathbf{X}_1| &= (-\lambda)^{n(p-1)}|\mathbf{\Phi}_1 - \lambda\mathbf{I}_n + (\mathbf{\Phi}_2/\lambda) + (\mathbf{\Phi}_3/\lambda^2) + \cdots + (\mathbf{\Phi}_p/\lambda^{p-1})| \\ &= (-1)^{np}|\mathbf{I}_n\lambda^p - \mathbf{\Phi}_1\lambda^{p-1} - \mathbf{\Phi}_2\lambda^{p-2} - \cdots - \mathbf{\Phi}_p|.\end{aligned}$$

Setting this to zero produces equation [10.1.13]. ■

■ **Proof of Proposition 10.2.** It is helpful to define $z_t(i, l)$ to be the component of y_{it} that reflects the cumulative effects of the lth element of $\boldsymbol{\varepsilon}$:

$$z_t(i, l) = \psi_{il}^{(0)}\varepsilon_{lt} + \psi_{il}^{(1)}\varepsilon_{l,t-1} + \psi_{il}^{(2)}\varepsilon_{l,t-2} + \cdots = \sum_{v=0}^{\infty} \psi_{il}^{(v)}\varepsilon_{l,t-v}, \qquad [10.A.4]$$

where $\psi_{il}^{(v)}$ denotes the row i, column l element of the matrix $\mathbf{\Psi}_v$. The actual value of the ith variable y_{it} is the sum of the contributions of each of the $l = 1, 2, \ldots, n$ components of $\boldsymbol{\varepsilon}$:

$$y_{it} = \mu_i + \sum_{l=1}^{n} z_t(i, l). \qquad [10.A.5]$$

The results of Proposition 10.2 are all established by first demonstrating absolute summability of the moments of $z_t(i, l)$ and then observing that the moments of $\mathbf{y}_t$ are obtained from finite sums of these expressions based on $z_t(i, l)$.

Proof of (a). Consider the random variable $z_t(i, l) \cdot z_{t-s}(j, m)$, where i, l, j, and m represent arbitrary indices between 1 and n and where s is the order of the autocovariance of $\mathbf{y}$ that is being calculated. Note from [10.A.4] that

$$\begin{aligned}E\{z_t(i, l) \cdot z_{t-s}(j, m)\} &= E\left\{\left[\sum_{r=0}^{\infty} \psi_{il}^{(r)}\varepsilon_{l,t-r}\right] \times \left[\sum_{v=0}^{\infty} \psi_{jm}^{(v)}\varepsilon_{m,t-s-v}\right]\right\} \\ &= \sum_{r=0}^{\infty}\sum_{v=0}^{\infty} \{\psi_{il}^{(r)}\psi_{jm}^{(v)}\} \cdot E\{\varepsilon_{l,t-r}\varepsilon_{m,t-s-v}\}.\end{aligned} \qquad [10.A.6]$$

The expectation operator can be moved inside the summation here because

$$\sum_{r=0}^{\infty}\sum_{v=0}^{\infty} |\psi_{il}^{(r)}\psi_{jm}^{(v)}| = \sum_{r=0}^{\infty}\sum_{v=0}^{\infty} |\psi_{il}^{(r)}| \cdot |\psi_{jm}^{(v)}| = \left\{\sum_{r=0}^{\infty} |\psi_{il}^{(r)}|\right\} \times \left\{\sum_{v=0}^{\infty} |\psi_{jm}^{(v)}|\right\} < \infty.$$

Now, the product of ε's in the final term in [10.A.6] can have nonzero expectation only if the ε's have the same date, that is, if $r = s + v$. Thus, although [10.A.6] involves a sum over an infinite number of values of r, only the value at $r = s + v$ contributes to this sum:

$$E\{z_t(i, l) \cdot z_{t-s}(j, m)\} = \sum_{v=0}^{\infty} \{\psi_{il}^{(s+v)}\psi_{jm}^{(v)}\} \cdot E\{\varepsilon_{l,t-s-v}\varepsilon_{m,t-s-v}\} = \sum_{v=0}^{\infty} \psi_{il}^{(s+v)}\psi_{jm}^{(v)}\sigma_{lm}, \qquad [10.A.7]$$

where σ_{lm} represents the covariance between ε_{lt} and ε_{mt} and is given by the row l, column m element of $\mathbf{\Omega}$.

The row i, column j element of $\mathbf{\Gamma}_s$ gives the value of

$$\gamma_{ij}^{(s)} = E(y_{it} - \mu_i)(y_{j,t-s} - \mu_j).$$