

Unit Roots, Cointegration, and Structural Change

G. S. Maddala and In-Moo Kim

Unit Roots, Cointegration, and Structural Change

Time series analysis has undergone many changes in recent years with the advent of unit roots and cointegration. Maddala and Kim present a comprehensive review of these important developments and examine structural change. The volume provides an analysis of unit root tests, problems with unit root testing, estimation of cointegration systems, cointegration tests, and econometric estimation with integrated regressors. The authors also present the Bayesian approach to these problems and bootstrap methods for small-sample inference. The chapters on structural change discuss the problems of unit root tests and cointegration under structural change, outliers and robust methods, the Markov switching model, and Harvey's structural time series model. *Unit Roots, Cointegration, and Structural Change* is a major contribution to **Themes in Modern Econometrics**, of interest both to specialists and graduate and upper-undergraduate students.

G. S. MADDALA is University Eminent Scholar at the Ohio State University and one of the most distinguished econometricians writing today. His many acclaimed publications include *Limited Dependent and Qualitative Variables in Econometrics* (Cambridge, 1983) and *Econometrics* (McGraw-Hill, 1977) and *Introduction to Econometrics* (MacMillan, 1988, 1992).

IN-MOO KIM is Professor of Economics at Sung Kyun Kwan University, Seoul, Korea.

UNIT ROOTS
COINTEGRATION
AND STRUCTURAL CHANGE

G. S. Maddala
The Ohio State University

In-Moo Kim
Sung Kyun Kwan University



CAMBRIDGE
UNIVERSITY PRESS

CAMBRIDGE UNIVERSITY PRESS
Cambridge, New York, Melbourne, Madrid, Cape Town, Singapore, São Paulo

Cambridge University Press
The Edinburgh Building, Cambridge CB2 8RU, UK

Published in the United States of America by Cambridge University Press, New York

www.cambridge.org
Information on this title: www.cambridge.org/9780521582575

© Cambridge University Press 1998

This publication is in copyright. Subject to statutory exception
and to the provisions of relevant collective licensing agreements,
no reproduction of any part may take place without the written
permission of Cambridge University Press.

First published 1998
Sixth printing 2004

A catalogue record for this publication is available from the British Library

ISBN 978-0-521-58257-5 hardback
ISBN 978-0-521-58782-2 paperback

Transferred to digital printing 2007

To my parents

G. S. Maddala

To Jong Han, Jung Youn, and So Youn

In-Moo Kim

Contents

<i>Figures</i>	<i>page</i>	xii
<i>Tables</i>		xiii
<i>Preface</i>		xvii
 <i>Part I Introduction and basic concepts</i>		 1
1 Introduction		3
References		6
2 Basic concepts		8
2.1 Stochastic processes		8
2.2 Some commonly used stationary models		11
2.3 Box-Jenkins methods		17
2.4 Integrated variables and cointegration		20
2.5 Spurious regression		28
2.6 Deterministic trend and stochastic trend		29
2.7 Detrending methods		32
2.8 VAR, ECM, and ADL		34
2.9 Unit root tests		37
2.10 Cointegration tests and ECM		39
2.11 Summary		41
References		42
 <i>Part II Unit roots and cointegration</i>		 45
3 Unit roots		47
3.1 Introduction		47
3.2 Unit roots and Wiener processes		49
3.3 Unit root tests without a deterministic trend		60
3.4 DF test with a linear deterministic trend		65

3.5	Specification of deterministic trends	72
3.6	Unit root tests for a wide class of errors	74
3.7	Sargan–Bhargava and Bhargava tests	82
3.8	Variance ratio tests	86
3.9	Tests for TSP versus DSP	87
3.10	Forecasting from TS versus DS models	89
3.11	Summary and conclusions	92
	References	92
4	Issues in unit root testing	98
4.1	Introduction	98
4.2	Size distortion and low power of unit root tests	100
4.3	Solutions to the problems of size and power	103
4.4	Problem of overdifferencing: MA roots	116
4.5	Tests with stationarity as null	120
4.6	Confirmatory analysis	126
4.7	Frequency of observations and power of unit root tests	129
4.8	Other types of nonstationarity	131
4.9	Panel data unit root tests	133
4.10	Uncertain unit roots and the pre-testing problem	139
4.11	Other unit root tests	140
4.12	Median-unbiased estimation	141
4.13	Summary and conclusions	145
	References	146
5	Estimation of cointegrated systems	155
5.1	Introduction	155
5.2	A general CI system	155
5.3	A two-variable model: Engle–Granger methods	156
5.4	A triangular system	160
5.5	System estimation methods	165
5.6	The identification problem	173
5.7	Finite sample evidence	175
5.8	Forecasting in cointegrated systems	184
5.9	Miscellaneous other problems	187
5.10	Summary and conclusions	191
	References	191
6	Tests for cointegration	198
6.1	Introduction	198
6.2	Single equation methods: residual-based tests	198

6.3	Single equation methods: ECM tests	203
6.4	Tests with cointegration as null	205
6.5	Multiple equation methods	211
6.6	Cointegration tests based on LCCA	222
6.7	Other tests for cointegration	226
6.8	Miscellaneous other problems	228
6.9	Of what use are cointegration tests?	233
6.10	Conclusions	241
	References	242
7	Econometric modeling with integrated regressors	249
7.1	I(1) regressors not cointegrated	249
7.2	I(1) regressors cointegrated	250
7.3	Unbalanced equations	251
7.4	Lagged dependent variables: the ARDL model	252
7.5	Uncertain unit roots	254
7.6	Uncertain unit roots and cointegration	256
7.7	Summary and conclusions	258
	References	258
	<i>Part III Extensions of the basic model</i>	261
8	The Bayesian analysis of stochastic trends	263
8.1	Introduction to Bayesian inference	264
8.2	The posterior distribution of an autoregressive parameter	266
8.3	Bayesian inference on the Nelson–Plosser data	268
8.4	The debate on the appropriate prior	271
8.5	Classical tests versus Bayesian tests	277
8.6	Priors and time units of measurement	277
8.7	On testing point null hypotheses	278
8.8	Further comments on prior distributions	284
8.9	Bayesian inference on cointegrated systems	287
8.10	Bayesian long-run prediction	290
8.11	Conclusion	291
	References	292
9	Fractional unit roots and fractional cointegration	296
9.1	Some definitions	296
9.2	Unit root tests against fractional alternatives	298
9.3	Estimation of ARFIMA models	300
9.4	Estimation of fractionally cointegrated models	302

9.5	Empirical relevance of fractional unit roots	303
9.6	Summary and conclusions	305
	References	306
10	Small sample inference: bootstrap methods	309
10.1	Introduction	309
10.2	A review of the bootstrap approach	309
10.3	The AR(1) model	322
10.4	Bootstrapping unit root tests	325
10.5	The moving block bootstrap and extensions	328
10.6	Issues in bootstrapping cointegrating regressions	332
10.7	Miscellaneous other applications	335
10.8	Conclusions	336
	References	336
11	Cointegrated systems with I(2) variables	342
11.1	Determination of the order of differencing	342
11.2	Cointegration analysis with I(2) and I(1) variables	348
11.3	Empirical applications	355
11.4	Summary and conclusions	358
	References	359
12	Seasonal unit roots and seasonal cointegration	362
12.1	Effect of seasonal adjustment	364
12.2	Seasonal integration	365
12.3	Tests for seasonal unit roots	366
12.4	The unobserved component model	371
12.5	Seasonal cointegration	375
12.6	Estimation of seasonally cointegrated systems	376
12.7	Empirical evidence	378
12.8	Periodic autoregression and periodic integration	379
12.9	Periodic cointegration and seasonal cointegration	381
12.10	Time aggregation and systematic sampling	381
12.11	Conclusion	382
	References	383
	<i>Part IV Structural change</i>	387
13	Structural change, unit roots, and cointegration	389
13.1	Tests for structural change	390
13.2	Tests with known break points	390
13.3	Tests with unknown break points	391

13.4	A summary assessment	398
13.5	Tests for unit roots under structural change	399
13.6	The Bayesian approach	402
13.7	A summary assessment of the empirical work	407
13.8	Effect of structural change on cointegration tests	410
13.9	Tests for structural change in cointegrated relationships	411
13.10	Miscellaneous other issues	414
13.11	Practical conclusions	416
	References	418
14	Outliers and unit roots	425
14.1	Introduction	425
14.2	Different types of outliers in time series models	425
14.3	Effects of outliers on unit root tests	428
14.4	Outlier detection	437
14.5	Robust unit root tests	440
14.6	Robust estimation of cointegrating regressions	445
14.7	Outliers and seasonal unit roots	448
14.8	Conclusions	448
	References	449
15	Regime switching models and structural time series models	454
15.1	The switching regression model	454
15.2	The Markov switching regression model	455
15.3	The Hamilton model	457
15.4	On the usefulness of the MSR model	460
15.5	Extensions of the MSR model	463
15.6	Gradual regime switching models	466
15.7	A model with parameters following a random walk	469
15.8	A general state-space model	470
15.9	Derivation of the Kalman filter	472
15.10	Harvey's structural time series model (1989)	475
15.11	Further comments on structural time series models	477
15.12	Summary and conclusions	479
	References	479
16	Future directions	486
	References	488
<i>Appendix 1</i> A brief guide to asymptotic theory		490
<i>Author index</i>		492
<i>Subject index</i>		500

Figures

2.1	Correlogram of an AR(2) model	16
2.2	Examples of two AR(1) processes with a drift	23
2.3	The variances of x_t and y_t	24
2.4	The autocorrelations of x_t and y_t	25
2.5	Cointegrated and independent I(1) variables	27
2.6	ARIMA(0,1,1) and its components	31
3.1	Random walk and step function	53
8.1	Marginal posterior distributions of ρ when $\rho = 1$	272

Tables

2.1	Regression of integrated variables	32
3.1	Critical values for Dickey–Fuller tests	64
3.2	Asymptotic distributions of the t -ratios for different DGPs	71
3.3	Critical values for the Schmidt–Phillips LM test	85
3.4	Nelson and Plosser’s results	89
4.1	Critical values of DF_{max} statistics	112
4.2	Critical values for the Elliott–Rothenberg–Stock DF–GLS test	114
4.3	Critical values for the Hwang–Schmidt DF–GLS test (t -test)	116
4.4	Critical values for the KPSS test	122
4.5	Quantiles of the LS estimator in an AR(1) model with drift and trend	143
6.1	Critical values for the ADF t -statistic and Z_t	200
6.2	Critical values for the Z_α	200
6.3	Response surface estimates of critical values	201
6.4	Critical values for the Harris and Inder test	210
6.5	Quantiles of the asymptotic distribution of the Johansen’s LR test statistics	213
6.6	Critical values of the LCCA-based tests	224
7.1	Features of regressions among series with various orders of integration	252
8.1	Posterior probabilities for the Nelson–Plosser data	270
12.1	Critical values for seasonal unit roots in quarterly data	369
12.2	Critical values for seasonal unit roots in monthly data	371
13.1	Asymptotic critical values for the diagnostic test	412

Nothing is so powerful as an idea whose time has come.

Victor Hugo

The Gods love the obscure and hate the obvious.

Brihadaranyaka Upanishad

Undue emphasis on niceties is a disease to which persons with mathematical training are especially prone.

G. A. Barnard, "A Comment on E. S. Pearson's Paper,"
Biometrika, 1947, 34, 123-128.

Simplicity, simplicity, simplicity! I say, let your affairs be as two or three, and not a hundred or a thousand. Simplify, simplify.

H. D. Thoreau: *Walden*

Preface

The area of unit roots, cointegration, and structural change has been an area of intense and active research during the past decade. Developments have been proceeding at a fast pace. However, almost all the books are technically oriented and do not bring together the different strands of research in this area. Even if many new developments are going to take place, we thought it is time to provide an overview of this area for the benefit of empirical as well as theoretical researchers. Those who are doing empirical research will benefit from the comprehensive coverage of the book. For those who are doing theoretical research, particularly graduate students starting on their dissertation work, the present book will provide an overview and perspective of this area. It is very easy for graduate students to get lost in the intricate algebraic detail of a particular procedure and lose sight of the general framework their work fits in to.

Given the broad coverage we have aimed at, it is possible that we have missed several papers. This is not because they are not important but because of oversight and/or our inability to cover too many topics.

To keep the book within reasonable length and also to provide accessibility to a broad readership, we have omitted the proofs and derivations throughout. These can be found by interested readers in the papers cited. Parts of the book were used at different times in graduate courses at the University of Florida, the Ohio State University, Caltech, State University of New York at Buffalo, and Sung Kyun Kwan University in Korea.

We would like to thank Adrian Pagan, Peter Phillips, and an anonymous referee for many helpful comments on an earlier draft. Thanks are also due to professor Chul-Hwan Kim of Ajou University, Young Se Kim of Sung Kyun Kwan University, and Marine Carrasso of the Ohio State

University for their helpful comments. Responsibility for any remaining errors is ours. We would also like to thank Patrick McCartan at the Cambridge University Press for his patience in the production of this book.

G. S. Maddala
The Ohio State University, U.S.A.

In-Moo Kim
Sung Kyun Kwan University, Seoul, Korea

Part I

Introduction and basic concepts

This part consists of two chapters. Chapter 1 is just an outline of the book. Chapter 2 introduces the basic concepts: stochastic processes; stationarity, the different kinds of commonly used stationary models (MA, AR, ARMA), Box-Jenkins methods; integrated variables and cointegration; spurious regression; deterministic and stochastic trend; detrending methods; VAR, ECM, ADL models; tests for unit root and cointegration.

All these topics are pursued in subsequent chapters and some of the statements made here (regarding unit root and cointegration tests) are qualified in subsequent chapters. The point here is to explain what all these terms mean.

1

Introduction

During the last decade, the econometric literature on unit roots and cointegration has literally exploded. The statistical theory relating to the first order autoregressive processes where the autoregressive parameter is equal to one (unstable process) and greater than one (explosive process) was developed by Anderson (1959), White (1958, 1959), and Rao (1961) (see Fuller (1985) for a review). However, the econometric literature on unit roots took off after the publication of the paper by Nelson and Plosser (1982) that argued that most macroeconomic series have unit roots and that this is important for the analysis of macroeconomic policies.

Similar is the story on cointegration. Yule (1926) suggested that regressions based on trending time series data can be spurious. This problem of spurious regressions was further pursued by Granger and Newbold (1974) and this also led to the development of the concept of cointegration (loosely speaking, lack of cointegration means spurious regression). Again, the pathbreaking paper by Granger (1981), first presented at a conference at the University of Florida in 1980, did not catch fire until about five years later, and now the literature on cointegration has exploded. As for historical antecedents, Hendry and Morgan (1989) argue that Frisch's concept of *multicollinearity* in 1934 can be viewed as a forerunner of the modern concept of cointegration.

The recent developments on unit roots and cointegration have changed the way time series analysis is conducted. Of course, the publication of the book by Box and Jenkins (1970) changed the methods of time series analysis, but the recent developments have formalized and made systematic the somewhat *ad hoc* methods in Box and Jenkins. Moreover, the asymptotic theory for these models has been developed.

Traditionally, the analysis of time series consisted of a decomposition

of the series into trend, seasonal, and cyclical components. We can write the time series x_t as

$$x_t = T_t + S_t + C_t$$

where T_t = trend, S_t = seasonal, and C_t = the remaining component which we might call the cyclical component. The trend and seasonal were first removed and then the residual was explained by an elaborate model. The trend, which is a long-term component was considered to be beyond the scope of explanation, and so was the seasonal. Attention thus focused entirely on the short-term dynamics as described by C_t .

The earliest approaches to trend removal consisted of regressing the time series on t (if the trend is considered linear) and a polynomial of t (if the trend is considered to be nonlinear). Elaborate methods have been devised for the removal of the seasonal. These are described in Hylleberg (1992), but one common method used in the analysis of seasonally unadjusted data was to use seasonal dummies (see Lovell, 1963).

By contrast to the regression methods for trend removal and seasonal adjustment, the methods suggested in Box and Jenkins (1970) consisted of removing the trend and seasonal by successive differencing. Define

$$\Delta x_t = x_t - x_{t-1}$$

$$\Delta^2 x_t = \Delta \Delta x_t = x_t - 2x_{t-1} + x_{t-2}$$

$$\Delta_4 x_t = x_t - x_{t-4}$$

$$\Delta_{12} x_t = x_t - x_{t-12}$$

A linear trend is removed by considering Δx_t , a quadratic trend by considering $\Delta^2 x_t$. With quarterly data the seasonal is removed by considering $\Delta_4 x_t$. With monthly data the seasonal is removed by considering $\Delta_{12} x_t$.

Thus, there are two approaches to the removal of trend and seasonal:

- (i) regression methods,
- (ii) differencing methods.

During the 1980s there were two major developments. The first was a discussion of these two methods in a systematic way and deriving tests to determine which is more appropriate for a given series. The second major development was the argument that trend and seasonal contain important information and that they are to be *explained* rather than

removed. Methods have been devised to estimate long-run economic relationships. If we are considering two time series, say x_t and y_t , then the trend in x_t may be related to the trend in y_t (common trends) and the seasonal in x_t may be related to the seasonal in y_t (common seasonals). Thus modeling the trend and the seasonal should form an integrated part of the analysis of the time series, rather than a concentration on the short-run dynamics between x_t and y_t .

Two questions also arose in this context. If one is also considering the problem of determining long-run relationships, how long a time series do we need to consider? Is it 20 years or 100 years and is having 240 monthly observations better than having 20 yearly observations? Also, if we are considering a span of 100 years, would the parameters in the estimated relationships be stable over such a long period? This is the problem of structural change. These are the issues that will be covered in this book. In the following chapters these problems are discussed in detail.

The book is divided into four parts.

Part I: Introduction and basic concepts

This consists of this chapter and chapter 2:

Chapter 2 introduces the basic concepts of ARMA models, unit roots and cointegration, spurious regression, Vector Autoregression (VAR), and Error Correction Models (ECM).

Part II: Unit roots and cointegration

This consists of chapters 3 to 7:

Chapter 3 discusses the different unit root tests. It also has an introduction to Wiener processes that will repeatedly be used in the rest of the book.

Chapter 4 discusses issues relating to the power of unit root tests, tests using stationarity as the null, tests for MA unit roots, LM tests for unit roots, and other important problems with unit root testing.

Chapter 5 discusses the different methods of estimation of cointegrated systems.

Chapter 6 discusses different tests for cointegration. These use no cointegration as the null. The chapter also covers tests using cointegration as the null.

Chapter 7 discusses the issues that arise in modeling with integrated regressors.

Part III: Extensions of the basic model and alternative approaches to inference

This consists of chapters 8 to 12:

Chapter 8 is on Bayesian analysis of unit roots and cointegration and the Bayesian approach to model selection.

Chapter 9 is on fractional unit roots and fractional cointegration.

Chapter 10 is on bootstrap methods which are alternatives to asymptotic inference in the preceding chapters.

Chapter 11 extends the analysis of the previous chapters to the case of $I(2)$ variables. It discusses issues of testing $I(2)$ versus $I(1)$ and $I(1)$ versus $I(2)$ as well as modeling systems with $I(2)$, $I(1)$ and $I(0)$ variables.

Chapter 12 is devoted to the analysis of seasonal data, tests for seasonal unit roots, tests for seasonal cointegration, and estimation of seasonally cointegrated systems.

Part IV: Structural change

Chapters 13 to 15 are devoted to analysis of structural change. Specifically, they discuss the effects of structural change on unit root tests and cointegration tests:

Chapter 13 discusses structural change and unit roots.

Chapter 14 is on outlier problems and robust estimation methods. It discusses the effects of different types of outliers on unit root tests, and robust estimation methods in the presence of outliers.

Chapter 15 discusses regime switching models and structural time series models.

Finally chapter 16 presents some avenues for further research.

Throughout the book, emphasis is on the intuition behind the different procedures and their practical usefulness. The algebraic detail is omitted in most cases because interested readers can refer to the original work cited. We have tried to emphasize the basic ideas, so that the book will be of guidance to empirical researchers.

References

- Anderson, T.W. (1959), "On Asymptotic Distributions of Estimates of Parameters of Stochastic Difference Equations," *Annals of Mathematical Statistics*, 30, 676–687.
- Box, G.E.P. and G.M. Jenkins (1970), *Time Series Analysis Forecasting and Control*, Holden-Day, San Francisco.

- Fuller, W.A. (1985), "Nonstationary Autoregressive Time Series," *Handbook of Statistics*, 5, North-Holland Publishing Co., Amsterdam, 1–23.
- Granger, C.W.J. (1981), "Some Properties of Time Series Data and Their Use in Econometric Model Specification," *Journal of Econometrics*, 16, 121–130.
- Granger, C.W.J. and P. Newbold (1974), "Spurious Regression in Econometrics," *Journal of Econometrics*, 2, 111–120.
- Hendry, D.F. and M.S. Morgan (1989), "A Re-Analysis of Confluence Analysis," *Oxford Economic Papers*, 44, 35–52.
- Hylleberg, S. (ed.) (1992), *Modeling Seasonality*, Oxford University Press, Oxford.
- Lovell, M.C. (1963), "Seasonal Adjustment of Economic Time Series and Multiple Regression Analysis," *Journal of American Statistical Association*, 58, 993–1010.
- Nelson, C.R. and C.I. Plosser (1982), "Trends and Random Walks in Macroeconomic Time Series," *Journal of Monetary Economics*, 10, 139–162.
- Rao, M. M. (1961), "Consistency and Limit Distributions of Estimators of Parameters in Explosive Stochastic Difference Equations," *Annals of Mathematical Statistics*, 32, 195–218.
- White, J.S. (1958), "The Limiting Distribution of the Serial Correlation Coefficient in the Explosive Case," *Annals of Mathematical Statistics*, 29, 1188–1197.
- (1959), "The Limiting Distribution of the Serial Correlation Coefficient in the Explosive Case II," *Annals of Mathematical Statistics*, 30, 831–834.
- Yule, G.U. (1926), "Why Do We Sometimes Get Nonsense Correlations Between Time Series? A Study in Sampling and the Nature of Time Series," *Journal of Royal Statistical Society*, 89, 1–64.

2

Basic concepts

The purpose of this chapter is to introduce several terms that will be used repeatedly in the subsequent chapters, and to explain their meaning. The terms included are: stationarity, ARMA models, integrated variables, Box–Jenkins methods, unit roots, cointegration, deterministic and stochastic trends, spurious regression, spurious periodicity and trend, vector autoregression (VAR) models, and error correction models (ECMs).

2.1 Stochastic processes

From a theoretical point of view a time series is a collection of random variables $\{X_t\}$. Such a collection of random variables ordered in time is called a *stochastic process*. The word *stochastic* has a Greek origin and means *pertaining to chance*. If it is a continuous variable, it is customary to denote the random variable by $X(t)$, and if t is a discrete variable, it is customary to denote them by X_t . An example of continuous random variables $X(t)$ is the recording of an electrocardiogram. Examples of discrete random variables X_t are the data of unemployment, money supply, closing stock prices, and so on. We will be considering discrete processes only, and so we shall use the notation X_t or $X(t)$ interchangeably.

The probability structure of the sequence of random variable $\{X_t\}$ is determined by the joint distribution of a stochastic process. The question arises, however, since T (time) is commonly an infinite set, whether we need an infinite dimensional distribution to define the probability structure of the stochastic process. Kolmogorov (1933) showed that when the stochastic process satisfies certain regularity conditions the stochastic process can be described by a finite dimensional distribution. That is, under certain conditions the probabilistic structure of the

stochastic process $\{X_t\}$ is completely specified by the joint distribution $F(X_{t_1}, \dots, X_{t_n})$ for all values of n (a positive integer) and any subset $(t_1, \dots, t_n)$ of T . One of the regularity conditions is the *symmetry*, that reshuffling the ordering of the index does not change the distribution. The other is the *compatibility* that the dimensionality of the joint distribution can be reduced by marginalization.

Since the definition of a stochastic process by the joint distribution is too general, it is customary to define the stochastic process in terms of the first and second moments of the variable, X_t . Given that, for a specific t , X_t is a random variable, we can denote its distribution and density functions by $F(X_t)$ and $f(X_t)$ respectively. The parametric family of densities are determined by the following the first and the second moments:

Mean	$\mu_t = E(X_t)$
Variance	$\sigma_t^2 = \text{var}(X_t)$
Autocovariance	$\gamma_{t_1, t_2} = \text{cov}(X_{t_1}, X_{t_2})$

The distribution of a stochastic process is characterized by the first and the second moments, and they are both functions of t . Note that if X_t follows a normal distribution, the distribution of X_t is completely characterized by the first and the second moments, which is called a *Gaussian process*.

The fact that the unknown parameters μ_t , σ_t^2 , γ_{t_1, t_2} change with t presents us with a difficult problem. There are (finite but still) too many parameters to be estimated. However, we have just a sample of size 1 on each of the random variables. For example, if we say that the unemployment rate at the end of this week is a random variable, we have just one observation on this particular random variable in a week. There is no way of getting another observation, so we have what is called a *single realization*. This feature compels us to specify some highly restrictive models for the statistical structure of the stochastic process. Given a single realization, we need to reduce the number of parameters and the question is how to reduce the number of parameters μ_t , σ_t^2 , γ_{t_1, t_2} .

Reducing the number of parameters to be estimated can be done by imposing certain restrictions. Restrictions come in two forms:

- (i) *stationarity*: restrictions on the *time heterogeneity* of the process,
- (ii) *asymptotic independence*: restrictions on the *memory* of the process.

These two restrictions reduce the numbers of parameters to be estimated and also facilitate the derivation of asymptotic results. We shall discuss stationarity here but omit discussion of asymptotic independence. This can be found in Spanos (1986, chapter 8).

Stationarity A time series is said to be strictly stationary if the joint distribution of $X_{t_1}, \dots, X_{t_n}$ is the same as the joint distribution of $X_{t_1+\tau}, \dots, X_{t_n+\tau}$ for all $t_1, \dots, t_n$, and τ . The distribution of the stationary process remains unchanged when shifted in time by an arbitrary value τ . Thus the parameters which characterize the distribution of the process do not depend on t , but on the *lag* τ . The concept of stationarity is difficult to verify in practice because it is defined in terms of the distribution function. For this reason the concept of stationarity defined in terms of moments is commonly preferred.

A stochastic process $\{X_t, t \in T\}$ is said to be *lth-order stationary* if for any subset $(t_1, t_2, \dots, t_n)$ of T and any τ the joints moments are

$$E(X_{t_1}^{l_1}, \dots, X_{t_n}^{l_n}) = E(X_{t_1+\tau}^{l_1}, \dots, X_{t_n+\tau}^{l_n})$$

where $l_1 + \dots + l_n \leq l$. Let us take positive integers for $l_1, l_2, \dots, l_n$ and l . When $l = 1$, i.e., $l_1 = 1$

$$E(X_t) = E(X_{t+\tau}) = \mu \quad (\text{a constant})$$

the process $\{X_t\}$ is said to be *first-order stationary*. When $l = 2$, the possible cases are $(l_1 = 1, l_2 = 0)$, $(l_1 = 2, l_2 = 0)$, and $(l_1 = 1, l_2 = 1)$. According to three cases the process $\{X_t\}$ has its joint moments as follows

$$E(X_t) = E(X_{t+\tau}) = \mu \quad (\text{a constant})$$

$$E(X_t^2) = E(X_{t+\tau}^2) = \sigma^2 \quad (\text{a constant})$$

$$\text{cov}(X_{t_1}, X_{t_2}) = \text{cov}(X_{t_1+\tau}, X_{t_2+\tau}) = \gamma_{t_1, t_2} = \gamma_\tau$$

where $t_1 - t_2 = \tau$. The mean and variance of X_t are constant and the covariances of X_t depend only on the *lag* or *interval* $\tau = t_1 - t_2$, not on t_1 or t_2 . The process X_t is said to be *second-order stationary*.

Second-order stationarity is also called *weak* or *wide-sense* or *covariance stationarity*. In modeling time series, second-order stationarity is the most commonly used form of stationarity. This is partly due to the fact that for a normal (or Gaussian) stationary process, second-order stationarity is equivalent to strict stationarity. If X_t follow a multivariate normal distribution, since the multivariate normal distribution is

completely characterized by the first and second moments, the two concepts of strict stationarity and weak stationarity are equivalent (recall the case of $n = 2$). For other distributions this is not so.

In order to see how stationarity reduces the number of parameters, let us consider a Gaussian process $\{X_t\}$ and the parameters θ_t for the subset $(t_1, \dots, t_n)$ of T . Without the assumption of stationarity the joint distribution of the process $\{X_t\}$ is characterized by the vector of parameters

$$\theta = [\mu_{t_i}, \text{cov}(X_{t_i}, X_{t_j})] \quad \text{for } i, j = 1, \dots, n$$

which is a $n + \frac{n(n+1)}{2} \times 1$ vector. By imposing stationarity, as we have seen above, the vector of parameters is reduced to

$$\theta = [\mu, \sigma^2, \gamma_\tau]$$

which is a $(n + 1) \times 1$ vector. A sizeable reduction in the number of the unknown parameters is resulted in by imposing stationarity. Note, however, that even in the case of stationarity the number of parameters (especially γ_τ) increases as τ increases, i.e., as the size of T increases. This is the reason why we need more restriction – the memory of the process which is about the meaningful size of τ regardless of the size of T .

2.2 Some commonly used stationary models

We shall now discuss some commonly used stationary processes. We shall denote the autocovariance function by *acvf* and the autocorrelation function by *acf*.

2.2.1 Purely random process

This is a discrete process X_t consisting of a sequence of independent identically distributed (*iid*) random variables. It has a constant mean and constant variance. Its *acvf* is given by

$$\gamma(\tau) = \text{cov}(X_t, X_{t+\tau}) = 0$$

and the *acf* is given by

$$\rho(\tau) = \begin{cases} 1 & \text{if } \tau = 0 \\ 0 & \text{if } \tau \neq 0 \end{cases}$$

A purely random process is also called a *white noise*. A white-noise process is a second-order stationary process and has no memory. If X_t is also assumed to be normal, then the process is strictly stationary.

2.2.2 Moving-average (MA) processes

Suppose that $\{\varepsilon_t\}$ is a purely random process with mean zero and variance σ^2 . Then a process $\{X_t\}$ defined by

$$X_t = \beta_0 \varepsilon_t + \beta_1 \varepsilon_{t-1} + \cdots + \beta_q \varepsilon_{t-q}$$

is called a moving-average process of order q and is denoted by $MA(q)$. Since the ε s are unobserved variables, we scale them so that $\beta_0 = 1$. Since $E(\varepsilon_t) = 0$ for all t , we have $E(X_t) = 0$. And ε_t are independent with a common variance σ^2 . Further, writing out the expressions for X_t and $X_{t-\tau}$ in terms of the ε s and picking up the common terms (since the ε s are independent), we get

$$\gamma(\tau) = \text{cov}(X_t, X_{t-\tau}) = \begin{cases} \sigma^2 \sum_{i=0}^{q-\tau} \beta_i \beta_{i+\tau} & \text{for } \tau = 0, 1, 2, \dots, q \\ 0 & \text{for } \tau > q \end{cases}$$

Also considering $\text{cov}(X_t, X_{t+\tau})$, we get the same expressions as for $\gamma(-\tau)$. Hence $\gamma(-\tau) = \gamma(\tau)$. The *acf* can be obtained by dividing $\gamma(\tau)$ by $\text{var}(X_t)$. For the MA process, $\rho(\tau) = 0$ for $\tau > q$, that is, they are zero for lags greater than the order of the process. Since $\gamma(\tau)$ is independent of t , the $MA(q)$ process is weakly stationary. Note that no restrictions on the β_i are needed to prove the stationarity of the MA process.

To facilitate our notation we shall use the *lag operator* L . It is defined by $L^j X_t = X_{t-j}$ for all j . Thus $LX_t = X_{t-1}$, $L^2 X_t = X_{t-2}$, $L^{-1} X_t = X_{t+1}$, and so on. With this notation the $MA(q)$ process can be written as (since $\beta_0 = 1$)

$$X_t = (1 + \beta_1 L + \beta_2 L^2 + \cdots + \beta_q L^q) \varepsilon_t = \beta(L) \varepsilon_t$$

The polynomial in L has q roots and we can write

$$X_t = (1 - \pi_1 L)(1 - \pi_2 L) \cdots (1 - \pi_q L) \varepsilon_t$$

where $\pi_1, \pi_2, \dots, \pi_q$ are the roots of the equation

$$z^q + \beta_1 z^{q-1} + \cdots + \beta_q = 0$$

After estimating the model we can calculate the residuals from $\varepsilon_t = [\beta(L)]^{-1} X_t$ provided that $[\beta(L)]^{-1}$ converges. This condition is called the *invertibility condition*. The condition for invertibility is that $|\pi_i| < 1$

for all i .¹ This implies that an $\text{MA}(q)$ process can be written as an $\text{AR}(\infty)$ process *uniquely*.

For instance, for the $\text{MA}(2)$ process

$$X_t = (1 + \beta_1 L + \beta_2 L^2) \varepsilon_t \quad (2.1)$$

π_1 and π_2 are roots of the quadratic equation $z^2 + \beta_1 z + \beta_2 = 0$. The condition $|\pi_i| < 1$ gives

$$\left| \frac{-\beta_1 \pm \sqrt{\beta_1^2 - 4\beta_2}}{2} \right| < 1$$

This gives the result that β_1 and β_2 must satisfy

$$\begin{aligned} \beta_1 + \beta_2 &> -1 \\ \beta_2 - \beta_1 &> -1 \\ |\beta_2| &< 1 \end{aligned} \quad (2.2)$$

The last condition is derived from the fact that $\beta_2 = \pi_1 \pi_2$, the product of the roots. The first two conditions are derived from the fact that if $\beta_1^2 - 4\beta_2 > 0$, then $\beta_1^2 - 4\beta_2 < (2 + \beta_1)^2$ or $\beta_1^2 - 4\beta_2 < (2 - \beta_1)^2$. Under the condition (2.2) the $\text{MA}(2)$ process (2.1) can be written as an $\text{AR}(\infty)$ uniquely.

Moving-average processes arise in econometrics mostly through trend elimination methods. One procedure often used for trend elimination is that of successive differencing of the time series X_t . If we have

$$X_t = \alpha_0 + \alpha_1 t + \alpha_2 t^2 + \varepsilon_t$$

where ε_t is a purely random process, successive differencing of X_t will eliminate the trend but the resulting series is a moving-average process that can show a cycle. Thus the trend-eliminated series can show a cycle even when there was none in the original series. This phenomenon of spurious cycles is known as the *Slutsky effect* (Slutsky, 1937).

2.2.3 Autoregressive (AR) processes

Suppose again that ε_t is a purely random process with mean zero and variance σ^2 . Then the process X_t given by

$$X_t = \alpha_1 X_{t-1} + \alpha_2 X_{t-2} + \cdots + \alpha_p X_{t-p} + \varepsilon_t \quad (2.3)$$

¹ An alternative statement often found in books on time series is that the roots of the equation $1 + \beta_1 z + \beta_2 z^2 + \cdots + \beta_q z^q = 0$ all lie outside the unit circle.

is called an autoregressive process of order p and is denoted by $AR(p)$. Since the expression is like a multiple regression equation, it is called *regressive*. However, it is a regression of X_t on its own past values. Hence it is autoregressive.

In terms of the lag operator L , the AR process (2.3) can be written as

$$(1 - \alpha_1 L - \alpha_2 L^2 - \dots - \alpha_p L^p) X_t = \varepsilon_t \quad (2.4)$$

or

$$\begin{aligned} X_t &= \frac{1}{(1 - \alpha_1 L - \alpha_2 L^2 - \dots - \alpha_p L^p)} \varepsilon_t \\ &= \frac{1}{(1 - \pi_1 L)(1 - \pi_2 L) \dots (1 - \pi_p L)} \varepsilon_t \end{aligned}$$

where $\pi_1, \pi_2, \dots, \pi_p$ are the roots of the equation

$$z^p - \alpha_1 z^{p-1} - \dots - \alpha_p = 0$$

The condition is that the expansion of (2.4) is valid and the variance of X_t is finite, that is $|\pi_i| < 1$ for all i .

To find the *acvf*, we could expand (2.3), but the expressions are messy. An alternative procedure is to assume that the process is stationary and see what *acf* are. To do this we multiply equation (2.3) throughout by $X_{t-\tau}$, take expectations of all the terms and divide throughout by $\text{var}(X_t)$, which is assumed finite. This gives us

$$\rho(\tau) = \alpha_1 \rho(\tau - 1) + \dots + \alpha_p \rho(\tau - p)$$

Substituting $\tau = 1, 2, \dots, p$ and noting $\rho(\tau) = \rho(-\tau)$ we get equations to determine the p parameters $\alpha_1, \alpha_2, \dots, \alpha_p$. These equations are known as the *Yule-Walker equations*.

To illustrate these procedures we will consider an $AR(2)$ process

$$X_t = \alpha_1 X_{t-1} + \alpha_2 X_{t-2} + \varepsilon_t$$

π_1 and π_2 are the roots of the equation

$$z^2 - \alpha_1 z - \alpha_2 = 0$$

Thus $|\pi_i| < 1$ implies that

$$\left| \frac{\alpha_1 \pm \sqrt{\alpha_1^2 + 4\alpha_2}}{2} \right| < 1$$

This gives

$$\begin{aligned}\alpha_1 + \alpha_2 &< 1 \\ \alpha_1 - \alpha_2 &> 1 \\ |\alpha_2| &< 1\end{aligned}\tag{2.5}$$

(The conditions are similar to the conditions (2.2) derived for the invertibility of the MA(2) process.)

In the case of the AR(2) process we can also obtain the $\rho(\tau)$ recursively using the Yule–Walker equations. We know that $\rho(0) = 1$ and

$$\begin{aligned}\rho(1) &= \alpha_1\rho(0) + \alpha_2\rho(-1) \\ &= \alpha_1\rho(0) + \alpha_2\rho(1) = \frac{\alpha_1}{1 - \alpha_2}\end{aligned}$$

Thus

$$\begin{aligned}\rho(2) &= \alpha_1\rho(1) + \alpha_2\rho(0) = \frac{\alpha_1^2}{1 - \alpha_2} + \alpha_2 \\ \rho(3) &= \alpha_1\rho(2) + \alpha_2\rho(1) = \frac{\alpha_1(\alpha_1^2 + \alpha_2)}{1 - \alpha_2} + \alpha_1\alpha_2\end{aligned}$$

and so on.

As an example, consider the AR(2) process

$$X_t = 1.0X_{t-1} - 0.5X_{t-2} + \varepsilon_t$$

Here $\alpha_1 = 1.0$ and $\alpha_2 = -0.5$. Note that conditions (2.5) for weak stationarity are satisfied. However, since $\alpha_1^2 + 4\alpha_2 < 0$ the roots are complex and $\rho(\tau)$ will be a sinusoidal function. A convenient method to compute $\rho(\tau)$ is to use the Yule–Walker equations

$$\rho(\tau) = \rho(\tau - 1) - 0.5\rho(\tau - 2)$$

Note that $\rho(0) = 1$ and $\rho(1) = \alpha_1/(1 - \alpha_2) = 0.6666$. We then have the autocorrelations for $\tau = 2, 3, \dots, 13$ recursively. This method can be used whether the roots are real or complex. Figure 2.1 shows a plot of this correlogram.

2.2.4 Autoregressive moving-average (ARMA) processes

We will now discuss models that are combinations of the AR and MA models. These are called ARMA models. An ARMA(p, q) model is defined as

$$X_t = \alpha_1X_{t-1} + \alpha_2X_{t-2} + \dots + \alpha_pX_{t-p} + \varepsilon_t + \beta_1\varepsilon_{t-1} + \dots + \beta_q\varepsilon_{t-q}$$

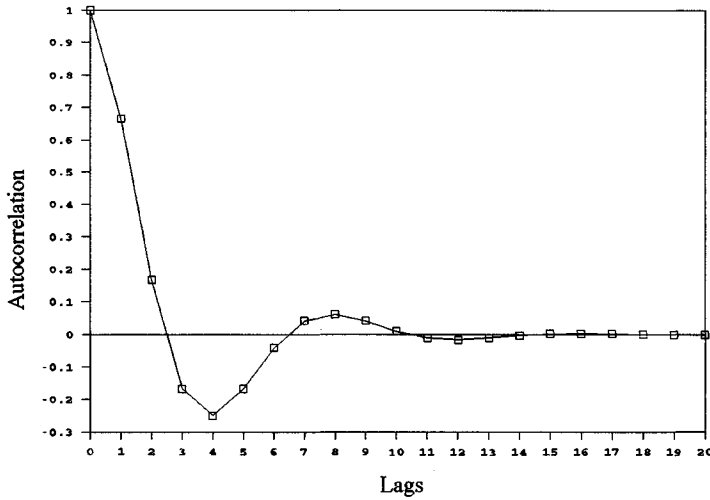


Fig. 2.1. Correlogram of an AR(2) model

where ε_t is a purely random process with mean zero and variance σ^2 . The motivation for these methods is that they lead to parsimonious representations of higher-order AR(∞) or MA(∞) processes.

Using the lag operator L , we can write this as

$$\phi(L)X_t = \theta(L)\varepsilon_t$$

where $\phi(L)$ and $\theta(L)$ are polynomials of orders p and q , respectively, defined as

$$\phi(L) = 1 - \alpha_1 L - \dots - \alpha_p L^p$$

$$\theta(L) = 1 + \beta_1 L + \dots + \beta_q L^q$$

For stationarity we require that the roots of $\phi(L) = 0$ lie outside the unit circle. For the invertibility of the MA component, we require that the roots of $\theta(L)$ lie outside the unit circle. For instance, for the ARMA(2,2) process these conditions are given by equations (2.2) and (2.5). The *acvf* and *acf* of an ARMA model are more complicated than for an AR or MA model.

We will derive the *acf* for the simplest case: the ARMA(1,1) process

$$X_t = \alpha X_{t-1} + \varepsilon_t + \beta \varepsilon_{t-1}$$

In terms of the lag operator L this can be written as

$$(1 - \alpha L)X_t = (1 + \beta L)\varepsilon_t$$

or

$$\begin{aligned} X_t &= \frac{1 + \beta L}{1 - \alpha L} \varepsilon_t \\ &= (1 + \beta L)(1 + \alpha L + \alpha^2 L^2 + \cdots) \varepsilon_t \\ &= [1 + (\alpha + \beta)L + \alpha(\alpha + \beta)L^2 + \alpha^2(\alpha + \beta)L^3 + \cdots] \varepsilon_t \end{aligned}$$

Since $\{\varepsilon_t\}$ is a purely random process with mean zero and variance σ^2 we get

$$\begin{aligned} \text{var}(X_t) &= [1 + (\alpha + \beta)^2 + \alpha^2(\alpha + \beta)^2 + \cdots] \sigma^2 \\ &= \left[1 + \frac{(\alpha + \beta)^2}{1 - \alpha^2} \right] \sigma^2 = \frac{1 + \beta^2 + 2\alpha\beta}{1 - \alpha^2} \sigma^2 \end{aligned}$$

Also

$$\begin{aligned} \text{cov}(X_t, X_{t-1}) &= [(\alpha + \beta) + \alpha(\alpha + \beta)^2 + \alpha^3(\alpha + \beta)^2 + \cdots] \sigma^2 \\ &= \left[\alpha + \beta + \frac{(\alpha + \beta)^2 \alpha}{1 - \alpha^2} \right] \sigma^2 \\ &= \frac{(\alpha + \beta)(1 + \alpha\beta)}{1 - \alpha^2} \sigma^2 \end{aligned}$$

Hence

$$\rho(1) = \frac{\text{cov}(X_t, X_{t-1})}{\text{var}(X_t)} = \frac{(\alpha + \beta)(1 + \alpha\beta)}{1 + \beta^2 + 2\alpha\beta}$$

Successive values of $\rho(\tau)$ can be obtained from the recurrence relation $\rho(\tau) = \alpha\rho(\tau - 1)$ for $\tau \geq 2$. For the AR(1) process $\rho(1) = \alpha$. It can be verified that $\rho(1)$ for ARMA(1,1) process is $>$ or $<$ α depending on whether $\beta > 0$ or < 0 , respectively.

2.3 Box-Jenkins methods

The Box-Jenkins method is one of the most widely used methodologies for the analysis of time series data. The influential work of Box and Jenkins (1970) shifted professional attention away from the *stationary serially correlated deviations from deterministic trend paradigm* toward the *ARIMA(p, d, q) paradigm*. It is popular because of its generality; it can handle any series, stationary or not, with or without seasonal elements, and it has well-documented computer programs. It is perhaps the last factor that contributed most to its popularity. Although Box

and Jenkins have been neither the originators nor the most important contributors in the field of ARMA models (for earlier discussion, see Quenouille, 1957), they have popularized these models and made them readily accessible to everyone, so much so that ARMA models are often referred to as Box–Jenkins models.

The basic steps in the Box–Jenkins methodology consist of the following five steps.

1. Differencing to achieve stationarity How do we conclude whether a time series is stationary or not? We can do this by studying the graph of the correlogram of the series. The correlogram of a stationary series drops off as τ , the number of lags, becomes large, but this is not the case of a nonstationary series. Thus the common procedure is to plot the correlogram of given series y_t and successive differences Δy , $\Delta^2 y$, and so on, and look at the correlograms at each stage. We keep differencing until the correlogram dampens.

2. Identification of a tentative model Once we have used the differencing procedure to get a stationary time series, we examine the correlogram to decide on the appropriate orders of the AR and MA components. The correlogram of a MA process is zero after a point; that of an AR process declines geometrically. The correlograms of ARMA processes show different patterns (but all dampen after a while). Based on these, one arrives at a tentative ARMA model. This step involves more of a judgmental procedure than the use of any clear-cut rules.

3. Estimation of the model The next step is the estimation of the tentative ARMA model identified in step 2. The estimation of AR models is straightforward. We estimate them by ordinary least squares (OLS) by minimizing the error sum of squares, $\Sigma \varepsilon_t^2$. In the case of MA models, we cannot write the error sum of squares $\Sigma \varepsilon_t^2$ as simply a function of the observed y s and the parameters as in the model. What we can do is to write down the covariance matrix of the moving-average error and, assuming normality, use the maximum likelihood method of estimation. An alternative procedure suggested by Box and Jenkins is the *grid-search procedure*. In this procedure we compute $\hat{\varepsilon}_t$ by successive substitution for each value of the MA parameters and choose the set of values of the parameters that minimizes the error sum of squares $\Sigma \hat{\varepsilon}_t^2$. For ARMA models, again the problem is with the MA component, either use ML

methods or use the grid-search procedure for the MA component. Ansley (1979) provides an algorithm for the exact likelihood of the mixed autoregressive moving-average process.

4. Diagnostic checking When an AR, MA, or ARMA model has been fitted to a given time series, it is advisable to check that the model does really give an adequate description of the data. There are two criteria often used that reflect the closeness of fit and the number of parameters estimated. One is the *Akaike Information Criterion* (AIC) and the other is *Schwartz Bayesian Information Criterion* (BIC). If p is the total number of parameters estimated, we have

$$AIC(p) = n \log \hat{\sigma}^2 + 2p$$

and

$$BIC(p) = n \log \hat{\sigma}^2 + p \log n$$

Here n is the sample size. If $RSS = \sum \hat{\varepsilon}_t^2$ is the residual sum of squares, then $\hat{\sigma}^2 = RSS/(n - p)$. If we are considering several ARMA models, we choose the one with the lowest AIC or BIC. The two criteria can lead to different conclusions.

In addition, we have to check the serial correlation pattern of the residuals. Box and Pierce (1970) suggest looking at not just the first-order autocorrelation but autocorrelations of all orders of the residuals. They suggest calculating $Q = N \sum_{\tau=1}^m \rho_\tau^2$, where ρ_τ is the autocorrelation of lag τ and N is the number of observations in the series. If the model fitted is appropriate, they argue that Q has an asymptotic χ^2 distribution with $m - p - q$ degrees of freedom, where p and q are, respectively, the orders of the AR and MA components. Although the Q statistics are quite widely used by those using time series programs (there is no need to list here the hundreds of papers, books, and programs that still use them), they are not appropriate in autoregressive models (or models with lagged dependent variables). The arguments against their use, that the OLS estimates of the coefficients of lagged dependent variables are inconsistent under the presence of serially correlated errors, are exactly the same as those against the use of the DW statistic in such a situation. Despite this inappropriateness of the Q statistic, the discussion of the Q statistic in the time series literature all concentrate only on the “low power” of the Q statistics. For further discussion of modifications of the Q statistics (these modifications are also inappropriate) and some alt-

ernative Lagrangian multiplier (LM) test statistics, see Maddala (1992, pp. 540–542).

5. Forecasting Suppose that we have estimated the model with n observations. We want to forecast x_{n+k} . This is called a k -periods ahead forecast. First we need to write out the expression for x_{n+k} . And then replace all future values x_{n+j} ($0 < j < k$) by their forecasts and ε_{n+j} ($j > 0$) by zero (since its expected value is zero). We also replace all ε_{n-j} ($j \geq 0$) by the predicted residuals. For example, from the ARMA(2,2) model

$$x_t = \alpha_1 x_{t-1} + \alpha_2 x_{t-2} + \varepsilon_t + \beta_1 \varepsilon_{t-1} + \beta_2 \varepsilon_{t-2}$$

we have the expression for the k -period ahead forecast

$$x_{n+k} = \alpha_1 x_{n+k-1} + \alpha_2 x_{n+k-2} + \varepsilon_{n+k} + \beta_1 \varepsilon_{n+k-1} + \beta_2 \varepsilon_{n+k-2}$$

When $k = 2$, we replace x_n by the observed x_n and x_{n+1} by the forecasts. The forecasted unknown error ε_{n+2} and ε_{n+1} are replaced by zero, while ε_n is replaced by the predicted residual

$$\hat{\varepsilon}_t = \hat{z}_t - \hat{\alpha}_1 \hat{z}_{t-1} - \hat{\alpha}_2 \hat{z}_{t-2}$$

where $z_t = (1 - \hat{\alpha}_1 L - \hat{\alpha}_2 L^2)^{-1} \varepsilon_t$.

2.4 Integrated variables and cointegration

In time series analysis we do not confine ourselves to the analysis of stationary time series. In fact, most of the time series we encounter are *nonstationary*.

Consider the following processes

$$\begin{aligned} x_t &= \rho x_{t-1} + u_t, & |\rho| < 1 \\ y_t &= y_{t-1} + v_t \end{aligned}$$

The error terms u_t and v_t are assumed to be normally independently identically distributed with mean zero and unit variance, $u_t, v_t \sim iin(0, 1)$, i.e., a purely random processes. Both x_t and y_t are AR(1) models. The difference between two models is that y_t is a special case of an x_t process when $\rho = 1$ and is called a *random walk* model. It is also referred to as an AR(1) model with a *unit root* since the root of the AR(1) equation is 1 (or *unit*). When we consider the statistical behavior of the two processes by investigating the mean (the first moment), the

variance and autocovariance (the second moments), they are completely different. Although the two processes belong to the same AR(1) class, x_t is a stationary process, while y_t is a nonstationary process.

The two stochastic processes can be expressed as the sum of the initial observation and the errors by successive substitution

$$\begin{aligned} x_t &= \rho x_{t-1} + u_t = \rho(\rho x_{t-2} + u_{t-1}) + u_t = \dots \\ &= \rho^t x_0 + u_t + \rho u_{t-1} + \dots + \rho^{t-1} u_1 \\ &= \rho^t x_0 + \sum_{i=0}^{t-1} \rho^i u_{t-i} \end{aligned}$$

Similarly, in the unit root case

$$y_t = y_0 + \sum_{i=0}^{t-1} v_{t-i}$$

Since both series are expressed as the sum of the initial observation and the errors, it can be said that the autoregressive model has been transformed to the moving-average form.

Suppose that the initial observations are zero, $x_0 = 0$ and $y_0 = 0$. The means of the two processes are

$$E(x_t) = 0 \quad \text{and} \quad E(y_t) = 0$$

and the variances are

$$\text{var}(x_t) = \sum_{i=0}^{t-1} \rho^{2i} \text{var}(u_{t-i}) \rightarrow \frac{1}{1 - \rho^2}$$

where $\rightarrow$ means *converge to asymptotically* (as $t \rightarrow \infty$) and

$$\text{var}(y_t) = \sum_{i=0}^{t-1} \text{var}(v_{t-i}) = t$$

The autocovariances of the two series are

$$\begin{aligned} \gamma_\tau^x &= E(x_t x_{t+\tau}) = E \left[\left(\sum_{i=0}^{t-1} \rho^i u_{t-i} \right) \left(\sum_{i=0}^{t+\tau-1} \rho^i u_{t+\tau-i} \right) \right] \\ &= \sum_{i=0}^{t+\tau-1} \rho^i \rho^{\tau+i} \end{aligned}$$

and

$$\gamma_\tau^y = E(y_t y_{t+\tau}) = E \left[\left(\sum_{i=0}^{t-1} v_{t-i} \right) \left(\sum_{i=0}^{t+\tau-1} v_{t+\tau-i} \right) \right] = (t - \tau)$$

since the errors are assumed to be *iin* and $cov(u_t, u_s) = 0, t \neq s$. The means of x_t and y_t are the same, but the variances (including autocovariances) are different. The important thing to note is that the variances and the autocovariance of y_t are functions of t , while those of x_t converge to a constant asymptotically. Thus as t increases the variance of y_t increases, while the variance of x_t converges to a constant.

The above example shows that the two processes x_t and y_t have different statistical properties. The variance of the stationary stochastic process x_t converges to a constant, while the variance of the random walk process y_t increases as t increases. Now if we add a constant to the AR(1) model, then the means of two processes also behave differently. Consider the AR(1) processes with a constant (or drift) as follows

$$\begin{aligned}x_t &= \alpha + \rho x_{t-1} + u_t, \quad |\rho| < 1 \\y_t &= \alpha + y_{t-1} + v_t\end{aligned}$$

The successive substitution yields

$$x_t = \rho^t x_0 + \alpha \sum_{i=0}^t \rho^i + \sum_{i=0}^t \rho^i u_{t-i}$$

and

$$y_t = y_0 + \alpha t + \sum_{i=0}^{t-1} v_{t-i} \quad (2.6)$$

Note that y_t series contains a (deterministic) trend t . If the initial observations are zero, $x_0 = 0$ and $y_0 = 0$, then the means of two processes are

$$\begin{aligned}E(x_t) &\rightarrow \frac{\alpha}{1 - \rho} \\E(y_t) &= \alpha t\end{aligned}$$

but the variances and the autocovariances are the same as those derived from the AR(1) model without the constant. By adding a constant to the AR(1) processes, the means of two processes as well as the variances are different. Both the mean and variances of y_t are time varying, while those of x_t are constant.

To illustrate these properties, we generate the 150 observations of x_t and y_t with $\alpha = 0.5$ and $\rho = 0.9$. The innovations u_t and v_t are generated by using a pseudo-random number generator. Figure 2.2 show the typical shape of the two stochastic processes. If we slice the time domain with some windows, say $\tau = 30$, we can find that the process

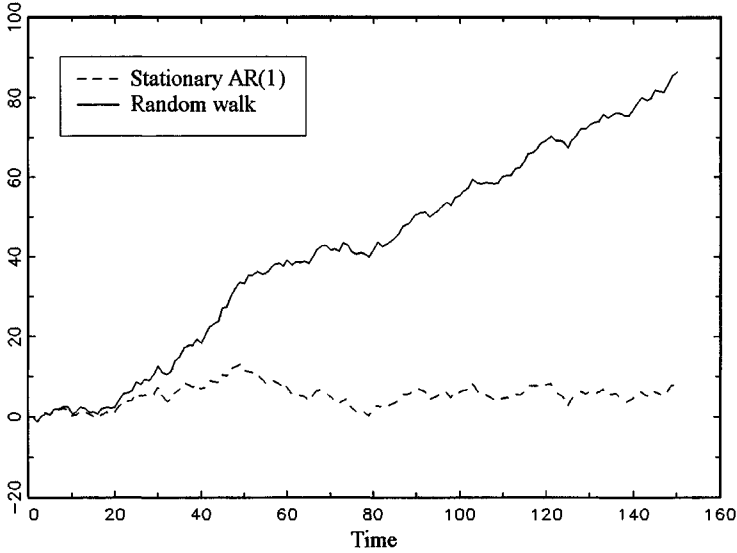


Fig. 2.2. Examples of two AR(1) processes with a drift

x_t passes the mean of the x_t process ($E(x_t) = 0$) at least once, while the process y_t does not. The process x_t has a force to converge toward the mean (that is, it is *mean-reverting*) and randomly fluctuates around the mean (no systematic changes). On the other hand, the process y_t increases systematically as $t \rightarrow \infty$ (sometimes systematically decreases) and there is no force to move it toward its mean.

The variances of the two series are computed after the first 20 periods. Figure 2.3 illustrates the variances of the two series after the first 20 periods. The variance of y_t increases as t increases, while the variance of x_t converges to a constant ($1/(1 - \rho) = 10$) after $t > 70$.

Figure 2.4 shows the correlograms of x_t and y_t . As we have seen

$$\gamma_\tau^x \rightarrow \frac{\rho^\tau}{1 - \rho^2} = \rho^\tau \gamma_0^x$$

thus we can expect

$$\rho_\tau^x = \frac{\gamma_\tau^x}{\gamma_0^x} = \rho^\tau \rightarrow 0$$

as $\tau \rightarrow \infty$ since $|\rho| < 1$. For a nonstationary series y_t , since

$$\rho_\tau^y = \frac{(t - \tau)}{t}$$

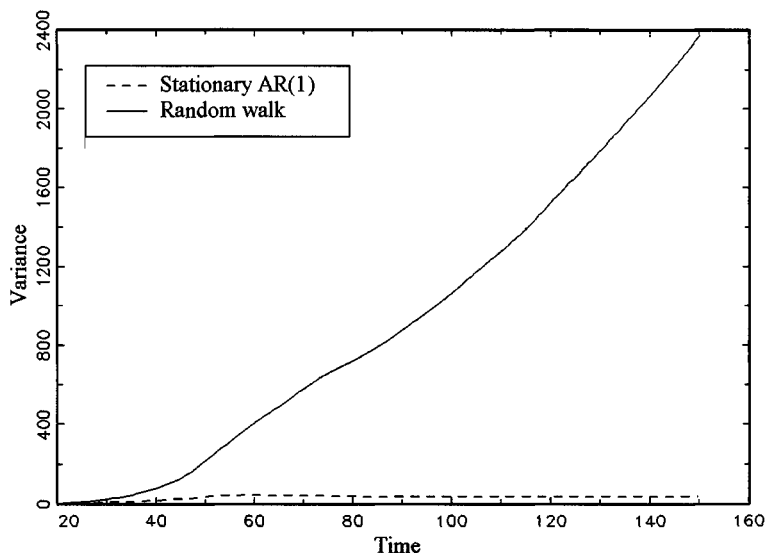


Fig. 2.3. The variances of x_t and y_t

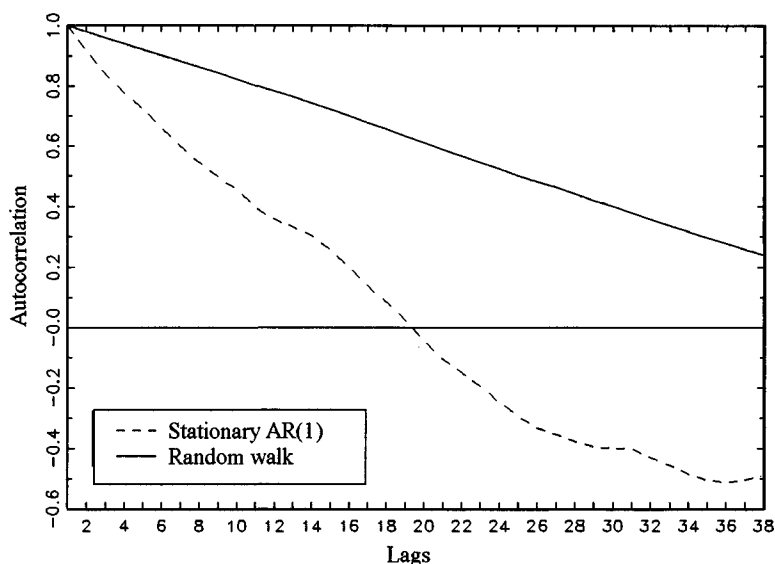
the values of ρ_τ^y will not come down to zero except for a very large value of the lag.

Since the variance of the nonstationary series is not constant over time (not covariance stationary or just nonstationary), the conventional asymptotic theory cannot be applied for these series. One of the easiest ways to analyze those series is to make those series stationary by differencing. In our example, the random walk series y_t can be transformed to a stationary series by differencing once

$$\Delta y_t = y_t - y_{t-1} = (1 - L)y_t = \varepsilon_t$$

where L is a lag operator. Since the error ε_t is assumed to be independently normal, the first difference of y_t is stationary. The variance of Δy_t is constant over the sample period.

When the nonstationary series can be transformed to the stationary series by differencing once, the series is said to be *integrated of order 1* and is denoted by $I(1)$. If the series needs to be differenced k times to be stationary, then the series is said to be $I(k)$. In our example, y_t is $I(1)$ variable, since the series needs to be differenced once to be stationary, while x_t is the $I(0)$ variable. The $I(k)$ series ($k \neq 0$) is also called a *difference-stationary process* (DSP). When $\Delta^d X_t$ is a stationary series

Fig. 2.4. The autocorrelations of x_t and y_t

that can be represented by an $\text{ARMA}(p, q)$ model, we say that X_t is an autoregressive integrated moving-average (ARIMA) process. Since the number of differences is equal to the order of integration, X_t is denoted as $\text{ARIMA}(p, d, q)$ process.

Another important class is the *trend-stationary process* (TSP). Consider the series

$$z_t = \alpha + \delta t + e_t \quad (2.7)$$

The mean of z_t is $E(z_t) = \alpha + \delta t$ and is not constant over the sample period, while the variance of z_t is $\text{var}(z_t) = \sigma^2$ and constant. Although the mean of z_t is not constant over the period, it can be forecasted perfectly whenever we know the value of t and the parameters α and δ . In this sense it is stationary around the deterministic trend t and z_t can be transformed to stationarity by regressing it on time. Note that both the DSP model equation (2.6) and the TSP model equation (2.7) exhibit a linear trend, but the appropriate method of eliminating the trend differs.

Most econometric analysis is based on the variance and covariance among the variables. For example, the OLS estimator from the regression y_t on x_t is the ratio of the covariance between y_t and x_t to the

variance of x_t . Thus if the variances of the variables behave differently, the conventional asymptotic theory cannot be applicable. When the order of integration is different, the variance of each process behaves differently. For example, if y_t is an $I(0)$ variable and x_t is $I(1)$, the OLS estimator from the regression y_t on x_t converges to zero asymptotically, since the denominator of the OLS estimator, the variance of x_t , increases as t increases, and thus it dominates the numerator, the covariance between x_t and y_t . That is, the OLS estimator does not have an asymptotic distribution. (It is degenerate with the conventional normalization of $\sqrt{T}$.) We need to have the normalization of T rather than that of $\sqrt{T}$.

Cointegration

An important property of $I(1)$ variables is that there can be linear combinations of these variables that are $I(0)$. If this is so then these variables are said to be *cointegrated*. The concept of *cointegration* was introduced by Granger (1981). Suppose that we consider two variables y_t and x_t that are $I(1)$. Then y_t and x_t are said to be cointegrated if there exists a β such that $y_t - \beta x_t$ is $I(0)$. This is denoted by saying that y_t and x_t are $CI(1,1)$. More generally, if y_t is $I(d)$ and x_t is $I(d)$, then y_t and x_t are $CI(d,b)$ if $y_t - \beta x_t$ is $I(d - b)$ with $b > 0$. What this mean is that the regression equation

$$y_t = \beta x_t + u_t$$

makes sense because y_t and x_t do not drift too far apart from each other over time. Thus, there is a long-run equilibrium relationship between them. If y_t and x_t are not cointegrated, that is $y_t - \beta x_t = u_t$ is also $I(1)$, then y_t and x_t would drift apart from each other over time. In this case the relationship between y_t and x_t that we obtain by regressing y_t and x_t would be *spurious*. We shall discuss spurious regression in the next section.

To fix idea let us consider a simple system of equations

$$\begin{aligned} x_t &= u_t, & u_t &= u_{t-1} + \varepsilon_{1t} \\ y_t + \alpha x_t &= v_t, & v_t &= \rho v_{t-1} + \varepsilon_{2t} \end{aligned}$$

In this system of equation, x_t and y_t are $I(1)$ variables regardless of the value of ρ . If $\rho = 1$, then the linear combination $y_t + \alpha x_t \sim I(1)$, thus x_t and y_t are two independent random walks. If $|\rho| < 1$, then $y_t + \alpha x_t \sim I(0)$, thus x_t and y_t are cointegrated. In order to see how two independent random walk variables and two cointegrated variables drift apart from each other over time, we draw the typical shapes of

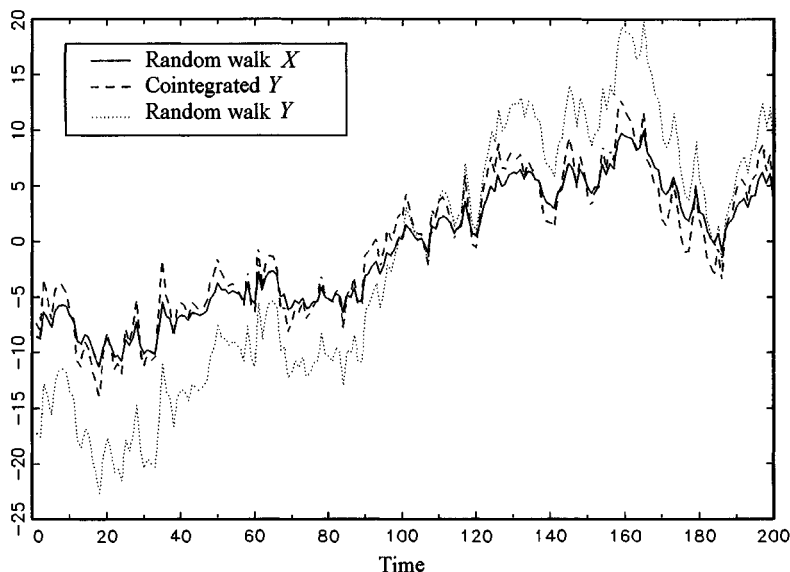


Fig. 2.5. Cointegrated and independent I(1) variables

two cases based on the 100 observations of x_t and y_t with $\alpha = -1$ and $\rho = 0.8, 1$. The errors u_t and v_t are generated from $iin(0, 1)$. Figure 2.5 shows the typical shapes of three stochastic processes, where x_t is a I(1) variable and y_t is an independent I(1) variable (when $\rho = 1$) and a cointegrated I(1) variable (when $\rho = 0.8$ and $\alpha = -1$). Two cointegrated I(1) variables x_t and y_t (solid line and dashed line) show some tendency for the two series not to drift too far apart (or move together), while two independent I(1) variables x_t and y_t (solid line and dotted line) do not have such tendency.

The concept of cointegration can be extended for the higher order of integrated variables. We can have I(2) variables that are cointegrated to produce an I(1) variable. When dealing with I(2) variables, different types of cointegration can occur. Firstly, linear combinations of I(2) variables can be I(1) or even I(0) and, secondly, some linear combinations of I(1) variables, can cointegrate with first-differences of the I(2) variables to produce an I(0) variable. In the modeling of demand for money equations, if $m_t = \log$ of nominal money, $y_t = \log$ of income, and $p_t = \log$ of prices, it has been found that m_t and p_t are possibly I(2) and real money ($m_t - p_t$) and velocity ($m_t - p_t - y_t$) are possibly I(1).

2.5 Spurious regression

Consider two uncorrelated random walk processes

$$\begin{aligned}y_t &= y_{t-1} + u_t, & u_t &\sim iin(0, \sigma_u^2) \\x_t &= x_{t-1} + v_t, & v_t &\sim iin(0, \sigma_v^2)\end{aligned}$$

where u_t and v_t are assumed to be serially uncorrelated as well as mutually uncorrelated. And consider the regression

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

Since y_t and x_t are uncorrelated random walk processes, we would expect that the R^2 from this regression would tend to zero. However, this is not the case. The parameter β_1 detects correlation and Yule (1926) showed long ago that spurious correlation can persist even in large samples in nonstationary time series. If the two time series are growing over time, they can be correlated even if the increments in each series are uncorrelated. Thus, we have to be cautious when interpreting regressions with $I(1)$ variables.

This point was also illustrated in Granger and Newbold (1974) who present some examples with artificially generated data where the errors u_t and v_t were generated independently so that there was no relationship between y_t and x_t , but the correlations between y_t and y_{t-1} , and x_t and x_{t-1} were high. The regression of y on x gave a high R^2 but a low Durbin–Watson (DW) statistic. When the regression was run in first differences, the R^2 was close to zero and the DW statistic was close to 2, thus demonstrating that there was no relationship between y and x and that the R^2 obtained was spurious.

Phillips (1986) shows that with two independent random walk processes with *no drift* (like the ones we have considered) the least squares regression

$$y_t = \hat{\beta}_0 + \hat{\beta}_1 x_t + \hat{\varepsilon}_t \tag{2.8}$$

leads to a divergent estimator $\hat{\beta}_0$ and convergence of $\hat{\beta}_1$ to a *random variable*. By showing the limiting distributions of the OLS estimator and test statistics, including t and DW statistics and R^2 , Phillips explains the same results found by the simulation of Granger and Newbold (1974).

Entorf (1992) shows that the results are altered if we consider independent random walk processes *with drifts*. Suppose that we have

$$\begin{aligned}y_t &= \alpha_y + y_{t-1} + u_t \\x_t &= \alpha_x + x_{t-1} + v_t\end{aligned}$$

then Entorf shows that in the regression (2.8)

$$\hat{\beta}_1 \xrightarrow{p} \frac{\alpha_y}{\alpha_x}$$

where $\xrightarrow{p}$ denotes convergence in probability. Thus $\hat{\beta}_1$ converges to a *constant* rather than a random variable. Entorf also finds that $\hat{\beta}_0$ has a divergent distribution. Thus, the results from spurious regression depend on whether we consider random walk processes with drifts or with no drifts.

2.6 Deterministic trend and stochastic trend

As we have seen, integrated variables exhibit a systematic variation. But the variation is hardly predictable, though the variation is systematic. This type of variation is called a *stochastic trend*. On the other hand, trends which are completely predictable (if we know the coefficient of time) are known as *deterministic trends*. The specification of a deterministic trend can be any functional form of time. For example, the deterministic trend DT_t can be any of following:

$$DT_t = 0(\text{zero}), \quad c(\text{constant}), \quad \alpha + \beta t(\text{linear form})$$

or

$$\sum_{i=0}^p \beta_i t^i \quad (\text{polynomial time trend})$$

or

$$\begin{cases} \alpha_0 + \beta_0 t, & t = 1, \dots, m \\ \alpha_1 + \beta_1 t, & t = m + 1, \dots, T \end{cases} \quad (\text{segmented trend})$$

To fix ideas about the deterministic trend and the stochastic trend, let us consider the following ARIMA(0,1,1) model with a drift (a constant term)

$$\Delta y_t = \alpha + e_t + \gamma e_{t-1}$$

where e_t is assumed to be *iid* errors. Let $y_0 = e_0 = 0$ so that y_t can be written by successive substitution as

$$\begin{aligned} y_t &= \alpha + y_{t-1} + e_t + \gamma e_{t-1} \\ &= \alpha + (\alpha + y_{t-2} + e_{t-1} + \gamma e_{t-2}) + e_t + \gamma e_{t-1} \\ &= \alpha t + \sum_{i=1}^t e_i + \gamma \sum_{j=1}^{t-1} e_j \end{aligned}$$

$$= \alpha t + (1 + \gamma) \sum_{i=1}^t e_i - \gamma e_t$$

Letting

$$\begin{aligned} DT_t &= \alpha t \\ ST_t &= (1 + \gamma) \sum_{i=1}^t e_i \\ C_t &= -\gamma e_t \end{aligned}$$

we can rewrite y_t as

$$y_t = DT_t + Z_t = DT_t + ST_t + C_t \quad (2.9)$$

Here DT_t is a deterministic trend in y_t , and Z_t is the noise function or stochastic component of y_t . The noise function Z_t can be decomposed as the sum of the stochastic trend ST_t and the cyclical component C_t . The cyclical component is assumed to be a mean-zero stationary process. The stochastic trend incorporates all random shocks (e_1 to e_t) that have permanent effects on the level of y_t . The sum of the deterministic trend DT_t and the stochastic trend ST_t is the overall trend and the permanent component of y_t .

If we denote the permanent component of y_t as y_t^p , then y_t can be rewritten as

$$y_t = y_t^p + C_t$$

where the permanent component of y_t is $y_t^p = DT_t + ST_t$, the sum of the deterministic trend and the stochastic trend. It can be shown that the permanent component of y_t is a random walk with drift such that

$$\begin{aligned} y_t^p &= \alpha t + (1 + \gamma) \sum_{i=1}^t e_i \\ &= \alpha + \alpha(t-1) + (1 + \gamma) \sum_{i=1}^{t-1} e_i + (1 + \gamma)e_t \\ &= \alpha + y_{t-1}^p + (1 + \gamma)e_t \end{aligned}$$

To see the typical shape of each component we generate the ARIMA (0,1,1) series by setting $\alpha = 0.008$ and $\gamma = 0.3$. The innovations e_t are generated from a normal distribution of $(0, 0.053^2)$. Figure 2.6 shows the typical shape of series and each component. The solid line is the graph of the generated ARIMA(0,1,1) series. The dashed line is the deterministic trend $DT_t = 0.008t$, the short-dashed line is the stochastic

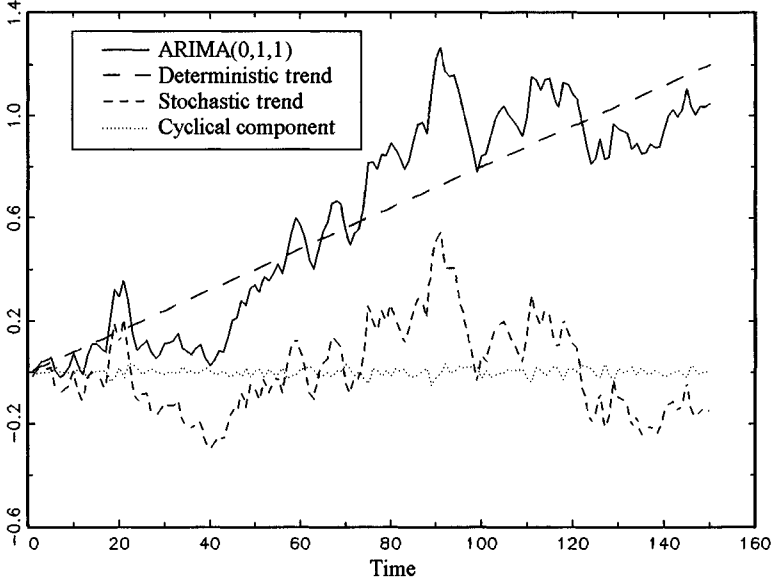


Fig. 2.6. ARIMA(0,1,1) and its components

trend $ST_t = 1.3 \sum_{i=1}^t e_t$, and the dotted line is the cyclical component $C_t = -0.3e_t$. The sum of two trend and the cyclical component is the generated ARIMA(0,1,1) series.

The decomposition of the noise function Z_t into ST_t and C_t as we have seen in equation (2.9) can be extended to any ARIMA($p, 1, q$) model. Beveridge and Nelson (1981) show that any ARIMA($p, 1, q$) model can be represented as a stochastic trend plus a stationary component. The noise function Z_t is assumed to be described by an autoregressive-moving-average process

$$A(L)Z_t = B(L)e_t$$

where $A(L)$ and $B(L)$ are polynomials in the lag operator L of order p and q , respectively, and e_t is assumed to be a sequence of *iid* errors. Suppose that the polynomial $A(L)$ has a unit root, we can write

$$A(L) = (1 - L)A^*(L)$$

where $A^*(L)$ has roots strictly outside the unit circle. The first-difference of the noise function is

$$(1 - L)A^*(L)Z_t = A^*(L)\Delta Z_t = B(L)e_t$$

Table 2.1. *Regression of integrated variables*

$x_t \setminus y_t$	Deterministic	Stochastic
Deterministic	Regression valid	Spurious regression
Stochastic	Spurious regression	Spurious regression unless y_t and x_t are cointegrated

and

$$\begin{aligned}\Delta Z_t &= A^*(L)^{-1}B(L)e_t \\ &= \psi(L)e_t = [\psi(1) + (1-L)\psi^*(L)]e_t\end{aligned}\quad (2.10)$$

where $\psi^*(L) = (1-L)^{-1}[\psi(L) - \psi(1)]$. Applying the operator $(1-L)^{-1}$ for both sides of (2.10) yields

$$\begin{aligned}Z_t &= \psi(1) \sum_{i=1}^t e_i + \psi^*(L)e_t \\ &= ST_t + C_t\end{aligned}$$

Thus, any ARIMA($p, 1, q$) model, here the noise function Z_t , can be decomposed into a stochastic component ST_t and a cyclical component C_t .

In the previous sections we talked of spurious regressions arising from trending variables. Whether such regressions are spurious or not depends on the type of trend – whether it is deterministic or stochastic. Table 2.1 shows when the regressions of y_t on x_t are valid. In the case of random walk processes with drifts, considered by Entorf, note that both y_t and x_t can be expressed as a function of time plus a random walk process with no drift.

2.7 Detrending methods

The method used for detrending depends on whether the time series is TSP (trend-stationary process or a process that is stationary around a trend) or DSP (difference-stationary process or a process that is stationary in first-differences). We shall be discussing problems of distinguishing between TSP and DSP in the next two chapters. The empirical evidence on this is mixed. One summary conclusion that emerges from

the voluminous empirical work is that the evidence in favor of deterministic trends is stronger for real variables than for nominal variables.

There have been several papers that have studied the consequences of *underdifferencing* or *overdifferencing*. If the time series is DSP and we treat it as TSP, this is a case of underdifferencing. If the time series is TSP, but we treat it as DSP, we have a case of overdifferencing. However, the serial correlation properties of the resulting errors from the misspecified processes need to be considered. For instance, if the regression relationship is correctly specified in first differences, i.e.

$$\Delta y_t = \beta \Delta x_t + e_t$$

this implies that

$$y_t = \alpha + \beta x_t + u_t$$

where $u_t = e_t + e_{t-1} + \dots$ is serially correlated and nonstationary. On the other hand, if the regression relationship is correctly specified in levels, i.e.

$$y_t = \alpha + \beta x_t + v_t$$

this implies that

$$\Delta y_t = \beta \Delta x_t + v_t - v_{t-1}$$

The errors follow a noninvertible moving-average process. The only question is whether OLS estimation of this equation with first-order MA errors leads us astray.

Plosser and Schwert argue that even if the MA coefficient is somewhat underestimated, the sampling distribution of β does not lead frequently to incorrect conclusions, and hence “the cost of overdifferencing may not be large when care is taken to analyze the properties of regression disturbances.” (1978, p. 643)

Nelson and Kang (1984) list several ways in which investigators would be led to misleading results if they estimate underdifferenced relationships. But these results hold if we do not correct for serial correlation in the errors. Their results on pages 79–80 show that the consequences of underdifferencing are not as serious if serial correlation in the errors is taken into account. Plosser and Schwert (1978, p. 638) argue that “the real issue is not differencing but an appropriate appreciation of the role of the error term in regression models.” This point is discussed further in McCallum (1993), with reference to regression models with lagged dependent variables. Note that in all this discussion we have been considering DSP without trend and TSP with a linear trend.

There is also some discussion of detrending on the periodic properties of the detrended series. Nelson and Kang (1981) argue that if the true process is DSP (with errors not exhibiting any cycles) and trend removal is done by regression on time (treating it as TSP) then the detrended series exhibits *spurious periodicity*.

2.8 VAR, ECM, and ADL

The vector autoregressive (VAR) model is just a multiple time series generalization of the AR model. The multiple time series generalization of the ARMA model is the VARMA model but we shall not consider it here. The VAR model has been popularized by Sims (1980) and it also forms a starting point for the analysis of cointegrating regression. In matrix notation, the VAR model for k variables can be written as

$$Y_t = A_1 Y_{t-1} + \dots + A_p Y_{t-p} + U_t$$

where $Y_t' = (y_{1t}, y_{2t}, \dots, y_{kt})$ and $A_1, A_2, \dots, A_p$ are $k \times k$ matrices and U_t is a k -dimensional vector of errors. With $E(u_t) = 0$ and Σ is positive definite. More compactly the VAR model can be represented as

$$Y_t = A(L)Y_t + U_t$$

where L is the lag operator. Since the VAR model is nothing but the stacked form of stationary AR(p) models and the regressors are the same for all the equations, the estimation of the VAR model is straightforward. The maximum likelihood estimator (MLE) reduces to the OLS estimator for each equation in the VAR model. The MLE of Σ is also provided by the OLS residuals $\hat{u}_t$ giving $\hat{\Sigma} = E(\hat{u}_t \hat{u}_t')$.

The above results apply only for the unrestricted VAR model. In practice it has been found that the unrestricted VAR model gives very erratic estimates (because of high multicollinearity among the explanatory variables) and several restricted versions have been suggested. Also, when some of the variables in Y_t are I(1), then one needs to use them in first-differences. If some of these, I(1) variables are cointegrated, then this imposes further restrictions on the parameters of the VAR model. These problem will be discussed in subsequent chapters.

The error correction model (ECM), first introduced into the econometric literature by Sargan (1964) and popularized by Davidson *et al.* (1978) has been a viable alternative to the VAR model. For some time in the 1980s, the American econometricians were all estimating VARs and the European econometricians were all estimating ECMs. There

are several interpretations of the error correction model and these are discussed in Algoskoufis and Smith (1991). The main characteristic of ECMs as compared with the VARs is the notion of an equilibrium long-run relationship and the introduction of past disequilibrium as explanatory variables in the dynamic behavior of current variables. The recent revival in the popularity of the ECMs has been based on the demonstration by Granger and Weiss (1983) that if two variables are integrated of order 1, and are cointegrated, they can be modeled as having been generated by an ECM.

The ECM links the realized value y_t to its target value $y_t^* = \beta' z_t$. In its simplest form, it can be written as

$$\Delta y_t = \lambda_1 \Delta y_t^* + \lambda_2 (y_{t-1}^* - y_{t-1})$$

where $\lambda_1 > 0, \lambda_2 > 0$. The last term represents past disequilibrium. The partial adjustment model is given by

$$\Delta y_t = \lambda (y_t^* - y_{t-1}) = \lambda \Delta y_t^* + \lambda (y_{t-1}^* - y_{t-1})$$

Thus the partial adjustment model corresponds to the ECM with $\lambda_1 = \lambda_2$.

Another class of models often considered is the *autoregressive distributed lag* (ADL) model discussed in Hendry, Pagan, and Sargan (1984). A general ADL model with p regressors, m lags in y , and n lags in each of the p regressors is denoted by $\text{ADL}(m, n; p)$. It is given by

$$y_t = \alpha_0 + \sum_{i=1}^m \alpha_i y_{t-i} + \sum_{j=1}^p \sum_{i=0}^n \beta_{ji} x_{jt-i} + \varepsilon_t$$

Such models are also called *dynamic linear regression* (DLR) models.

Consider the simplest $\text{ADL}(1, 1; 1)$ model

$$y_t = \alpha_0 + \alpha_1 y_{t-1} + \beta_0 x_t + \beta_1 x_{t-1} + \varepsilon_t \quad (2.11)$$

where it is assumed that $\varepsilon_t \sim iid(0, \sigma^2)$ and $|\alpha_1| < 1$. We shall show the connection between this and the ECM model. In long-run equilibrium $y_t = y_{t-1}$ and $x_t = x_{t-1}$, then we can write

$$(1 - \alpha_1)y = \alpha_0 + (\beta_0 + \beta_1)x$$

Thus, the long-run response to y of a change in x is given by

$$k = \frac{\beta_0 + \beta_1}{1 - \alpha_1}$$

Now write the equation (2.11) as

$$y_t - y_{t-1} = \alpha_0 + (\alpha_1 - 1)y_{t-1} + \beta_0(x_t - x_{t-1}) + (\beta_0 + \beta_1)x_{t-1} + \varepsilon_t$$

Writing $\beta_0 + \beta_1 = k(1 - \alpha_1)$ we get

$$\Delta y_t = \alpha_0 + (\alpha_1 - 1)(y_{t-1} - kx_{t-1}) + \beta_0 \Delta x_t + \varepsilon_t \quad (2.12)$$

Note that $(y_{t-1} - kx_{t-1})$ is a last periods disequilibrium. Thus (2.12) is the ECM that is implied by the ADL(1,1;1) model.

If we write (2.12) as

$$\Delta y_t = \alpha_0 + \beta_0 \Delta x_t + \gamma_1 y_{t-1} - \sigma_1 x_{t-1} + \varepsilon_t \quad (2.13)$$

then clearly an estimate of the long-run response k is given by

$$\hat{k} = \hat{\sigma}_1 / \hat{\gamma}_1$$

Equation (2.13) is known as the Bardsen (1989) transformation.

An alternative transformation of the same equation is that of Bewley (1979) which is as follows: define $\lambda = (1 - \alpha_1)^{-1}$ so that the long-run response $k = \lambda(\beta_0 + \beta_1)$. Now write the equation (2.11) as follows

$$(y_t - \alpha_1 y_t) = \alpha_0 - \alpha_1(y_t - y_{t-1}) + \beta_0 x_t + \beta_1 x_{t-1} + \varepsilon_t$$

or

$$(1 - \alpha_1)y_t = \alpha_0 - \alpha_1 \Delta y_t + (\beta_0 + \beta_1)x_t - \beta_1 \Delta x_t + \varepsilon_t$$

or

$$y_t = \lambda \alpha_0 - \lambda \alpha_1 \Delta y_t + \lambda(\beta_0 + \beta_1)x_t - \lambda \beta_1 \Delta x_t + \lambda \varepsilon_t \quad (2.14)$$

In this equation the coefficient of x_t is the long-run response. However, since Δy_t is correlated with ε_t (because it involves y_t) OLS estimation of this equation gives inconsistent estimates. Hence the equation has to be estimated by instrumental variable (IV) methods. Note that (2.11), (2.12), (2.13), and (2.14) are all equivalent transformations of the same equation, and hence should give the same estimates of the long-run parameter k . That (2.11) and (2.13) give the same estimates is fairly obvious given that both the equations can be estimated by OLS and the parameters are simple linear transforms. What is not obvious is (the result is proved in Bewley (1979, pp. 69–73) and Wickens and Breusch (1988)) that the IV estimation of (2.14) gives the same estimates and the same standard errors as the OLS estimation of (2.11) provided we use the explanatory variables in (2.11), that is, (y_{t-1}, x_t, x_{t-1}) as instrumental variables.

One other point to note is that we can write equation (2.12) as

$$\Delta y_t = \alpha_0 + (\alpha_1 - 1)(y_{t-1} - x_{t-1}) + \beta_0 \Delta x_t + (\alpha_1 - 1)(1 - k)x_{t-1} + \varepsilon_t$$

or

$$\begin{aligned} \Delta y_t &= \alpha_0 + (\alpha_1 - 1)(y_{t-1} - x_{t-1}) + \beta_0 \Delta x_t \\ &\quad + (\alpha_1 + \beta_0 + \beta_1 - 1)x_{t-1} + \varepsilon_t \end{aligned} \quad (2.15)$$

Note that the estimate of the error correction term, namely $(\alpha_1 - 1)$, is the same in both (2.12) and (2.15). What we have done is, we assumed the long-run response in the error correction term to be equal to 1, and added x_{t-1} as an explanatory variable. This simple but important result will be used when we come to cointegration tests based on ECM. The generalization of these results to a general ADL($m, n; p$) model are straightforward except for some extra notation and will not be pursued here.¹

2.9 Unit root tests

As discussed earlier, the question of whether to detrend or to difference a time series prior to further analysis depends on whether the time series is trend-stationary (TSP) or difference-stationary (DSP). If the series is trend-stationary, the data generating process (DGP) for y_t can be written as

$$y_t = \gamma_0 + \gamma_1 t + e_t$$

where t is time and e_t is a stationary ARMA process. If it is difference-stationary, the DGP for y_t can be written as

$$y_t = \alpha_0 + y_{t-1} + e_t$$

where e_t is again a stationary ARMA process. If the e_t s are serially uncorrelated, then this is a random walk with drift α_0 .

Following Bhargava (1986), we can nest these two models in the following model

$$\begin{aligned} y_t &= \gamma_0 + \gamma_1 t + u_t \\ u_t &= \rho u_{t-1} + e_t \end{aligned}$$

so that

$$y_t = \gamma_0 + \gamma_1 t + \rho[y_{t-1} - \gamma_0 - \gamma_1(t-1)] + e_t \quad (2.16)$$

¹ For details see chapter 2 of Banerjee *et al.* (1993).

where e_t is a stationary process. If $|\rho| < 1$, y_t is trend-stationary. If $|\rho| = 1$, y_t is difference-stationary. Equation (2.16) can be written as

$$y_t = \beta_0 + \beta_1 t + \rho y_{t-1} + e_t \quad (2.17)$$

or

$$\Delta y_t = \beta_0 + \beta_1 t + (\rho - 1)y_{t-1} + e_t \quad (2.18)$$

where $\beta_0 \equiv \gamma_0(1 - \rho) + \gamma_1\rho$ and $\beta_1 \equiv \gamma_1(1 - \rho)$. Note that if $\rho = 1$, then $\beta_1 \equiv 0$. If we start with the model

$$\begin{aligned} y_t &= \gamma_0 + u_t \\ u_t &= \rho u_{t-1} + e_t \end{aligned}$$

then we get

$$y_t = \gamma_0(1 - \rho) + \rho y_{t-1} + e_t$$

or

$$y_t = \beta_0 + \rho y_{t-1} + e_t$$

with $\beta_0 = 0$ if $\rho = 1$. If we have a quadratic trend, then equation (2.16) would be

$$y_t = \gamma_0 + \gamma_1 t + \gamma_2 t^2 + \rho[y_{t-1} - \gamma_0 - \gamma_1(t-1) - \gamma_2(t-1)^2] + e_t$$

which can be written as

$$y_t = \beta_0 + \beta_1 t + \beta_2 t^2 + \rho y_{t-1} + e_t \quad (2.19)$$

where

$$\begin{aligned} \beta_0 &= \gamma_0(1 - \rho) + (\gamma_1 - \gamma_2)\rho \\ \beta_1 &= \gamma_1(1 - \rho) + 2\gamma_2\rho \end{aligned}$$

and

$$\beta_2 = \gamma_2(1 - \rho)$$

Thus, if $\rho = 1$, then $\beta_2 \equiv 0$. We shall discuss the problems caused by this in later chapters.

It is customary to test the hypothesis $\rho = 1$ against the one-sided alternative $|\rho| < 1$. This is called a *unit root test*. One cannot, however, use the usual t -test to test $\rho = 1$ in equation (2.17) because under the null hypothesis, y_t is $I(1)$, and hence the t -statistic does not have an asymptotic normal distribution. The relevant asymptotic distribution, based on Wiener processes, will be discussed in chapter 3.

Equation (2.16) and (2.17) have been derived by considering the issue of testing whether a series is difference-stationary or trend-stationary. By contrast the tests for unit roots developed by Dickey (1976), Fuller (1976), and Dickey and Fuller (1979), commonly known as Dickey–Fuller tests are based on a simple autoregression with or without a constant or time trend. They are based on testing $\rho = 1$ in the equations

$$y_t = \rho y_{t-1} + e_t \quad (2.20)$$

$$y_t = \beta_0 + \rho y_{t-1} + e_t \quad (2.21)$$

$$y_t = \beta_0 + \beta_1 t + \rho y_{t-1} + e_t \quad (2.22)$$

However, as we noted earlier, the Bhargava formulation implies that in equation (2.21), $\beta_0 = 0$ if $\rho = 1$ and in equation (2.22), $\beta_1 = 0$ if $\rho = 1$. No such restrictions are imposed in the formulations in (2.21) and (2.22). As a consequence, the parameters in equations (2.21) and (2.22) have different interpretations under the null and the alternative (see Schmidt and Phillips, 1992). For instance, in equation (2.21), under the null hypothesis $\rho = 1$; β_0 represents the coefficient of trend, whereas under the alternative y_t is stationary around the level $\beta_0/(1 - \rho)$ (see the discussion in section 1.6 earlier). Similarly, by successive substitution, we can show that in equation (2.22), under the null hypothesis $\rho = 1$, the parameters β_0 and β_1 represent coefficients of t and t^2 in a quadratic trend, whereas under the alternative they represent the level and the coefficient of t in a linear trend. Because of these problems, the Bhargava-type formulation is preferred in unit root testing.

2.10 Cointegration tests and ECM

As we mentioned earlier (section 2.4), if a linear combination of $I(1)$ variables is stationary or $I(0)$, then the variables are said to be cointegrated. Suppose we have a set of k variables y_t which are all $I(1)$ and $\beta' y_t = u_t$ is $I(0)$, then β is said to be a *cointegrated vector* and the equation $\beta' y_t = u_t$ is called the *cointegrating regression*. Note that the elements of the cointegrating vector can be zero, but not all of them. If there are two such vectors β_1 and β_2 so that $\beta_1' y_t = u_{1t}$ and $\beta_2' y_t = u_{2t}$ are both $I(0)$, then any linear combination of these vectors is also a cointegrating vector because linear combinations of $I(0)$ variables are $I(0)$. There is, thus, an identification problem. Unless we bring in some extraneous information, we cannot identify the long-

run equilibrium relationship. These problems will be illustrated in later chapters.

We shall discuss here the relationship between cointegration and ECM. To fix ideas, consider the case of two variables y_{1t} and y_{2t} that are both $I(1)$. In this case if we write $y_{1t} = \beta y_{2t} + u_t$ and u_t is $I(0)$ then y_{1t} and y_{2t} are cointegrated and β is the cointegration vector (for the case of two variables, scalar). In the two variable case, if there is cointegration, we can show that β is unique. Because, if we have $y_{1t} = \gamma y_{2t} + v_t$ where v_t is also $I(0)$, by subtraction we have $(\beta - \gamma)y_{2t} + u_t - v_t$ is $I(0)$. But $u_t - v_t$ is $I(0)$ which means $(\beta - \gamma)y_{2t}$ is $I(0)$. This is not possible since y_{2t} is $I(1)$. Hence we should have $\beta = \gamma$. But, in the case of more than two variables, the cointegration vector is no longer unique.

How do we know whether y_{1t} and y_{2t} are cointegrated? We can test whether the error in the cointegrating regression is $I(1)$. Thus, the hypothesis that u_t has a unit root is a hypothesis that there is no cointegration. Tests for cointegration thus, have no cointegration as the null hypothesis. On the other hand, if the null hypothesis is that there is cointegration, this has to be based on stationarity as the null hypothesis for u_t . These issues will be discussed in the chapter on tests for cointegration.

The earliest cointegration test is the one suggested in Engle and Granger (1987) which consists of estimating the cointegrating regression by OLS, obtaining the residuals $\hat{u}_t$ and applying unit root tests for $\hat{u}_t$. Since $\hat{u}_t$ are themselves estimates, new critical values need to be tabulated. Several extensions of this test have been proposed and critical values have been tabulated for these tests. Since they are all based on $\hat{u}_t$, they are called *residual-based tests*.

Another set of tests are those that are based on the ECM model. These tests are based on what is known as the *Granger representation theorem* which says that if a set of $I(1)$ variables are cointegrated, they can be regarded as being generated by an ECM.

To illustrate these ideas, consider a simple two-equation model used in Engle and Granger (1987)

$$y_{1t} + \beta y_{2t} = u_{1t}, \quad u_{1t} = u_{1,t-1} + e_{1t} \quad (2.23)$$

$$y_{1t} + \alpha y_{2t} = u_{2t}, \quad u_{2t} = \rho u_{2,t-1} + e_{2t}, \quad |\rho| < 1 \quad (2.24)$$

where e_{1t} and e_{2t} are possibly correlated white-noise errors. The model is internally consistent only if $\alpha \neq \beta$. The reason for this constraint

is that if $\alpha = \beta$, it is impossible to find any values of y_{1t} and y_{2t} that satisfy both equations. The reduced forms for y_{1t} and y_{2t} are

$$\begin{aligned} y_{1t} &= \frac{\alpha}{\alpha - \beta} u_{1t} - \frac{\beta}{\alpha - \beta} u_{2t} \\ y_{2t} &= -\frac{1}{\alpha - \beta} u_{1t} + \frac{1}{\alpha - \beta} u_{2t} \end{aligned}$$

These equations also make clear the fact that both y_{1t} and y_{2t} are driven by a common I(1) variable u_{1t} . This is known as the *common trend representation* of the cointegrated system. Also equation (2.24) states that a linear combination of y_{1t} and y_{2t} is stationary. Hence y_{1t} and y_{2t} are cointegrated with a cointegration coefficient α . Note that if $\rho = 1$, then u_{2t} is also I(1) and hence there is no cointegration. The null hypothesis $\rho = 1$ is thus a test of the hypothesis of no cointegration. The Engle–Granger test is thus based on equation (2.24).

There is also an *autoregressive representation* for this system. Equation (2.23) and (2.24) can be written as

$$\begin{aligned} \Delta y_{1t} &= \beta \theta y_{1,t-1} + \alpha \beta \theta y_{2,t-1} + \eta_{1t} \\ \Delta y_{2t} &= -\theta y_{1,t-1} - \alpha \theta y_{2,t-1} + \eta_{2t} \end{aligned}$$

where $\theta = (1 - \rho)/(\alpha - \beta)$ and η_{1t} and η_{2t} are linear combinations of e_{1t} and e_{2t} . If we write $z_t = y_{1t} + \alpha y_{2t}$, then equation (2.24) implies

$$z_t = \rho z_{t-1} + e_{2t}$$

or

$$\Delta z_t = (\rho - 1) z_{t-1} + e_{2t}$$

or

$$\Delta y_{1t} = -\alpha \Delta y_{2t} + (\rho - 1) z_{t-1} + e_{2t} \quad (2.25)$$

This is in the form of an ECM where z_{t-1} represents past disequilibrium.

2.11 Summary

This chapter provides an introduction to the basic models that will be discussed in detail in later chapters. Besides an introduction to ARMA models and Box–Jenkins methods, the chapter introduces trend and difference stationarity (TSP and DSP), unit roots and cointegration, vector autoregressive (VAR) models, and error correction models (ECMs).

The next two chapters will discuss problems concerning unit roots and the subsequent two chapters will discuss cointegration.

References

- Algoskoufis, G. and R. Smith (1991), "On Error Correction Models: Specification, Interpretation, Estimation," *Journal of Economic Surveys*, 5, 97–128.
- Ansley, C.F. (1979), "An Algorithm for the Exact Likelihood of Mixed Autoregressive Moving Average Process," *Biometrika*, 66, 59–65.
- Banerjee, A., J.J. Dolado, J.W. Galbraith, and D.F. Hendry (1993), *Cointegration, Error Correction, and the Econometric Analysis of Non-Stationary Data*, Oxford University Press, Oxford.
- Bardsen, G. (1989), "The Estimation of Long-Run Coefficients from Error-Correction Models," *Oxford Bulletin of Economics and Statistics*, 51, 345–350.
- Beveridge, S. and C.R. Nelson (1981), "A New Approach to Decomposition of Economic Time Series into Permanent and Transitory Components with Particular Attention to Measurement of the 'Business Cycle'," *Journal of Monetary Economics*, 7, 151–174.
- Bewley, R. (1979), "The Direct Estimation of the Equilibrium Response in a Linear Dynamic Model," *Economic Letters*, 3, 357–361.
- Bhargava, A. (1986), "On the Theory of Testing for Unit Roots in Observed Time Series," *Review of Economic Studies*, 53, 137–160.
- Box, G.E.P. and G.M. Jenkins (1970), *Time Series Analysis Forecasting and Control*, Holden-Day, San Francisco.
- Box, G.E.P. and D.A. Pierce (1970), "Distribution of Residual Autocorrelations in Autoregressive Integrated Moving Average Time Series Models," *Journal of the American Statistical Association*, 65, 1509–1526.
- Davidson, J.E.H., D.F. Hendry, F. Srba, and S. Yeo (1978), "Econometric modeling of the Aggregate Time-Series Relationship Between Consumer's Expenditure and Income in the United Kingdom," *Economic Journal*, 88, 661–692.
- Dickey, D.A. (1976), "Estimation and Hypothesis Testing for Nonstationary Time Series," Ph.D. dissertation, Iowa State University.
- Dickey, D.A. and W.A. Fuller (1979), "Distribution of the Estimators for Autoregressive Time Series with a Unit Root," *Journal of the American Statistical Association*, 74, 427–431.
- Engle, R.F. and C.W.J. Granger (1987), "Co-integration and Error

- Correction: Representation, Estimation and Testing," *Econometrica*, 55, 251–276.
- Entorf, H. (1992), "Random Walk with Drift, Simultaneous Errors, and Small Samples: Simulating the Bird's Eye View," Institut National de la Statistique et des Etudes Economiques.
- Fuller, W.A. (1976), *Introduction to Statistical Time Series*, John Wiley, New York.
- Granger, C.W.J. (1981), "Some Properties of Time Series Data and Their Use in Econometric Model Specification," *Journal of Econometrics*, 16, 121–130.
- Granger, C.W.J. and P. Newbold (1974), "Spurious Regression in Econometrics," *Journal of Econometrics*, 2, 111–120.
- Granger, C.W.J. and A.A. Weiss (1983), "Time-Series Analysis of Error-Correction Models," in S. Karlin, T. Amemiya, and L.A. Goodman (eds.), *Studies in Econometrics, Time Series and Multivariate Statistics*, Academic Press, New York.
- Hendry, D.F., A.R. Pagan, and J.D. Sargan (1984), "Dynamic Specification," in Z. Griliches and M.D. Intriligator (eds.), *Handbook of Econometrics II*, North-Holland, Amsterdam, 1023–1100.
- Kolmogorov, A. (1933), *Foundations of the Theory of Probability*, published in 1950 by Chelsea, New York.
- Maddala, G.S. (1992), *Introduction to Econometrics*, 2nd ed, Macmillan, New York.
- McCallum, B.T. (1993), "Unit Roots in Macroeconomic Time Series: Some Critical Issues," *Economic Quarterly*, Federal Reserve Bank of Richmond, 79, 13–43.
- Nelson, C.R. and H. Kang (1981), "Spurious Periodicity in Inappropriately Detrended Time Series," *Journal of Monetary Economics*, 10, 139–162.
- (1984), "Pitfalls in the Use of Time as an Explanatory Variable in Regression," *Journal of Business and Economics*, 2, 73–82.
- Phillips, P.C.B. (1986), "Understanding Spurious Regression in Econometrics," *Journal of Econometrics*, 33, 311–340.
- Plosser, C.I. and W.G. Schwert (1978), "Money, Income and Sunspots: Measuring Economic Relationships and the Effects of Differencing," *Journal of Monetary Economics*, 4, 637–660.
- Quenouille, M. (1957), *The Analysis of Multiple Time Series*, Charles Griffin, London.
- Sargan, J.D. (1964), "Wages and Prices in the United Kingdom: A Study in Econometric Methodology," in P.E. Hart, G. Mills,

- and J.K. Whitaker (eds.), *Econometric Analysis for National Economic Planning*, Butterworth, London; reprinted in D.F. Hendry and K.F. Wallis (eds.), *Econometrics and Quantitative Economics*, Basil Blackwell, Oxford, 1984.
- Schmidt, P. and P.C.B. Phillips (1992), "LM Test for a Unit Root in the Presence of Deterministic Trends," *Oxford Bulletin of Economics and Statistics*, 54, 257–287.
- Sims, C. (1980), "Macroeconomics and Reality," *Econometrica*, 48, 1–48.
- Slutsky, E. (1937), "The Summation of Random Causes as the Source of Cyclic Processes," *Econometrica*, 5, 105.
- Spanos, A. (1986), *Statistical Foundations of Econometric modeling*, Cambridge University Press, Cambridge.
- Wickens, M.R. and T.S. Breusch (1988), "Dynamic Specification, the Long Run and the Estimation of Transformed Regression Models," *Economic Journal*, 98, 189–205.
- Yule, G.U. (1926), "Why Do We Sometimes Get Nonsense Correlations Between Time Series? A Study in Sampling and the Nature of Time Series," *Journal of the Royal Statistical Society*, 89, 1–64.

Part II

Unit roots and cointegration

This part contains five chapters that form the core material that needs to be understood to follow the rest of the book.

Chapter 3 gives a brief introduction to Wiener processes. We do not go into these in great detail because we do not go into details of the derivations of asymptotic distributions. Those interested in these can refer to the source material. (Many empirical researchers do not need the derivations.) We next discuss the importance of scaling factors in the derivation of asymptotic distributions. Next we discuss the Dickey–Fuller (DF) distribution and the DF tests, the ADF test and the problem of selection of lag length (a problem that needs special attention). Next we discuss the Phillips–Perron (PP) tests, Sargan–Bhargava tests, variance ratio tests, and finally forecasting problems.

Although often used, the ADF and PP tests are useless in practice and should not be used. Some useful modifications of these tests are discussed in chapter 4. The material covered in this chapter forms the basis of all the modifications discussed in the next chapter.

Chapter 4 considers several issues in unit root testing. The reason why there are so many unit root tests is that there is no uniformly powerful test for the unit root hypothesis. We discuss several of the tests for completeness. Some of them are not worth considering but they are all *promoted* by the respective authors and the Nelson–Plosser data set is used as a guinea pig for every new test suggested.

We first discuss the problems of size distortions and low power of unit root tests and then some solutions to these problems. We also discuss tests for stationarity as null, the oft-quoted KPSS test. We do not recommend its use – it has the same low power problems as the ADF and PP tests. It is discussed here because it is often referred to – as useful for *confirmatory analysis* in conjunction with the ADF and PP

tests. But we feel that such confirmatory analysis is an illusion (with two tests that lack power). Some useful modifications of the ADF, PP, and KPSS tests are discussed and these should be preferred. This chapter also discusses in detail panel data unit root tests. The Levin-Lin test is very often used (in fact over used) and we discuss useful modifications of this test.

Chapter 5 is on estimation of cointegrated systems and inference on the estimated coefficients. We first discuss the two-variable model and the normalization issue (not usually discussed). Second, we discuss a triangular system, the FM-OLS method, and several other methods involving lags and leads. Third, discuss system estimation methods – the Johansen procedure (widely used) and the Box–Tiao method (rarely used). Fourth, we discuss identification problems since the system methods identify only the cointegration space but not the individual cointegrating vectors without more prior information. Fifth, we discuss several Monte Carlo studies that have been conducted to study the different estimation methods but very few unambiguous conclusions emerge except that the Engle–Granger two-step method should be avoided. Finally we discuss several miscellaneous issues such as forecasting from cointegrated systems, threshold cointegration, and so on.

Chapter 6 on tests for cointegration is complementary to chapter 5. It discusses the commonly used residual-based tests (these should be avoided), ECM-based tests, tests with cointegration as null, and tests associated with system-based estimation methods (the Johansen and Box–Tiao methods). The chapter also has a sceptical note on the use of cointegration tests. One important problem discussed (often completely ignored) is the pre-testing problem – cointegration itself depends on preliminary tests for unit roots and thus there is a question of the appropriate significance levels to use in cointegration tests.

Chapter 7, the final chapter in this part, is on inferential procedures in regression models involving $I(1)$ and $I(0)$ regressors, unbalanced equations, the problem of uncertain unit roots, and testing procedures under uncertainty about unit roots and cointegration. Many of these problems are not commonly discussed in books.

3

Unit roots

3.1 Introduction

In the previous chapter we discussed the econometric problems that might occur when we run a regression with different orders of integrated variables. To avoid this problem, the first thing we have to do is to identify the correct order of the integration of each variable. In the context of ARIMA modeling, this identification is equivalent to determining the parameter d in the ARIMA(p, d, q) model. The Box–Jenkins approach discussed in the previous chapter suggested the use of visual inspection of correlograms for determining the parameter d . The recent development of unit root tests is nothing but the use of formal statistical tests in place of the visual inspection of the correlogram.

The idea that the parameter d is equal to the number of unit roots led Dickey and Fuller to replace the subjective inspection of the sample autocorrelation function with a formal test of the unit root null hypothesis. This test, known as the standard Dickey–Fuller (DF) test, is based on independently and identically distributed (*iid*) errors. For a wide class of errors which allows some heterogeneity and serial correlations in errors, two approaches have been proposed to modify the standard DF test. One is the parametric approach which suggests the augmented Dickey–Fuller (ADF) test. The nonparametric approach leads to the Phillips–Perron tests. The DF tests were applied to many US macroeconomic time series by Nelson and Plosser (1982). The evidence presented by them showed that most macroeconomic variables are well described by the ARIMA models with one unit root. This was believed to have important implications for understanding the sources and nature of the business cycle although most macroeconomists now think that unit root tests do not shed much light on the sources and nature of the business

cycle. Since then an enormous number of studies on the unit root have appeared. The results of Nelson and Plosser have been re-examined using different methods and given different conclusions. In fact every time a new method has been suggested, it has been tried on the Nelson–Plosser data.

One thing we have to note is the importance of the specification of the deterministic trend in the unit root testing procedures. If a time series has a unit root, then it shows a systematic pattern in its movement (though the movement cannot be predictable) which is named as *stochastic trend* (see the section 2.6). Many macroeconomic variables are more fruitfully modeled with both the deterministic trend and stochastic trend. When we incorporate the misspecified deterministic trend into the model which is also believed to have a stochastic trend, the systematic movement by the stochastic trend may be shadowed or overemphasized by the misspecified deterministic trend. Thus it is important to determine the appropriate form of deterministic trends *a priori* before conducting the unit root tests. The critical values for the unit root tests also differ according to the different form of deterministic trends. Because of this fact, the unit root testing procedures are not as simple as conventional testing procedures and require the use of appropriate tables of critical values.

In most econometric applications a linear specification of the deterministic trend is used and coefficients of the linear trend are assumed to be constant over the sample period. Recently the importance of using a flexible specification for the deterministic component has been emphasized. It is motivated by the view that the linear time trend hypothesis is inappropriate for modeling the deterministic component of an economic time series. From an econometric perspective, the appropriate representation of nonstationarity in macroeconomic time series is a vital issue since misspecification of a random walk as a stationary process evolving around a deterministic trend has major effects on the statistical analysis of the data. Recent studies (Perron, 1989; Hamilton, 1989; Phillips, 1992, 1994), suggesting that the linear trend might not be an appropriate specification for some macroeconomic variables, will be discussed later. In this chapter we confine ourselves to a discussion of the cases of a zero intercept, a constant, and a linear trend.

In the following sections we shall first discuss Wiener processes which are extensively used in the derivation of the several asymptotic distributions in the subsequent chapters. The discussion here will be brief and concerns the derivation of the most basic distributions. Since our

purpose is to present the basic results we shall not go into the details of the derivation of the asymptotic distributions in this book. These derivations can be found in the papers cited.

3.2 Unit roots and Wiener processes

A time series y_t is said to have an autoregressive unit root if we can write it as

$$y_t = DT_t + z_t$$

$$z_t = \rho z_{t-1} + e_t$$

and

$$\rho = 1$$

where e_t is stationary and DT_t is a deterministic component. With $\rho = 1$, Δz_t is stationary and Δy_t is stationary around the change in the deterministic part. In this case y_t is said to be integrated of order 1 and is denoted as $I(1)$. Stationary series are said to be $I(0)$. When y_t is stationary and is thus $I(0)$, and DT_t is a linear trend, then

$$y_t = DT_t + e_t$$

and the difference of y_t is

$$\Delta y_t = \alpha + e_t - e_{t-1}$$

where $\alpha = DT_t - DT_{t-1}$ is constant. Thus Δy_t has a moving-average unit root. MA unit roots arise if a stationary series differenced. This is known as *overdifferencing*. Tests using $I(1)$ as the null hypothesis are based on the unit autoregressive root as the null. Tests using $I(0)$ as the null hypothesis are based on the unit moving-average root as the null. In this chapter we shall consider the case of tests for the autoregressive unit root, since these are the ones most common. Tests for the moving-average unit root will be discussed in the next chapter.

As we have seen in chapter 2, if there is a unit root (from now on, unit root means autoregressive unit root), that is $\rho = 1$, the variance of the process y_t increases over time. This implies that since the OLS estimator of the autoregressive parameter $\hat{\rho}$ is the ratio of the covariance of y_t and y_{t-1} to the variance of y_t and $\text{var}(y_t) \rightarrow \infty$ as $t \rightarrow \infty$, the OLS estimator $\hat{\rho}$ no longer has an asymptotic normal distribution when the true autoregressive parameter $\rho = 1$. Thus the t - or F -type test

statistics based on the OLS estimator do not follow the conventional t - or F -distribution.

In order to test the unit root null against the stationary alternative hypothesis, first we need to derive the distribution of the OLS estimator of the autoregressive parameter under the unit root null hypothesis $H_0 : \rho = 1$. This distribution turns out to be a function of *Wiener processes* and hence we shall first give a brief introduction to the Wiener processes.

3.2.1 Wiener processes: some definitions

The Wiener process has been used in physics to describe the motion of a particle that is subject to a large number of molecular shocks. It is also referred to as a *Brownian motion*. Consider a small interval of length Δt , and define ΔW as the change in W during Δt . Then the behavior of a variable, W which follows a Wiener process can be described by two relations:

- (i) ΔW is related to Δt by the relation

$$\Delta W = \varepsilon \sqrt{\Delta t}$$

where $\varepsilon \sim N(0, 1)$. Hence $E(\Delta W) = 0, \text{var}(\Delta W) = \Delta t$.

- (ii) The values of ΔW for any two different short intervals are independent. That is, the ε s are independently normally distributed.

Consider next the value of W during a long time period T . This can be denoted by $W(T) - W(0)$. This can be regarded as the sum of n small intervals of length Δt . Hence $T = n\Delta t$ and

$$W(T) - W(0) = \sum_{i=1}^n \varepsilon_i \sqrt{\Delta t}$$

Since the ε_i are $IN(0, 1)$, by assuming $W(0) = 0$, we have $E(W(T)) = 0$ and $\text{var}(W(T)) = n \cdot \Delta t = T$. Now Let us consider the relationship between a random walk and $W(T)$. Let $\varepsilon_t \sim IN(0, 1)$ and $S_T = \sum_{t=1}^T \varepsilon_t$. Then S_T is a random walk, since $S_T = S_{t-1} + \varepsilon_T$, or an $I(1)$ process with independent increments. We have $E(S_T) = 0$ and $\text{var}(S_T) = T$, as with the variable $W(T)$.

As we will show later, the index t in the increasing range $[0, T]$ can be mapped into an index r in the fixed range $[0, 1]$ and the Wiener process is defined over this interval as $W(r)$. If $W(r)$ is a Wiener process, then for fixed r

$$W(r) \sim N(0, r), \quad 0 \leq r \leq 1$$

Thus the Wiener process is like a continuous random walk defined on the interval $[0,1]$.

In differential equation form, the Wiener process is defined by

$$dW = \varepsilon\sqrt{dt}$$

This is the basic Wiener process with a drift rate of 0 and variance rate of 1. The drift is said to be zero because $E(W(T)) = 0$. The variance rate is said to be 1, because $\text{var}(W(T)) = 1$ multiplied by T . A *generalized Wiener process* x is defined by

$$\Delta x = a\Delta t + b\Delta W$$

or in differential equation form

$$dx = a dt + b dW$$

Substituting $\Delta W = \varepsilon\sqrt{\Delta t}$ we get

$$\Delta x = a\Delta t + b\varepsilon\sqrt{\Delta t}$$

Hence

$$\Delta x \sim N(a\Delta t, b^2\Delta t)$$

As before, if we consider the behavior of x during a long period T and defining $n = T/\Delta t$ we get

$$x(T) \sim N(aT, b^2T)$$

Again, mapping the index t in the increasing range $[0, T]$ to an index r in the fixed range $[0,1]$ we can see the generalized Wiener process

$$x(r) \sim N(ar, b^2r), \quad 0 \leq r \leq 1$$

A generalized Wiener process, in which a and b are functions of the underlying value of x at time t , is the *Ito process*. It is defined by

$$dx = a(x, t)dt + b(x, t)dW$$

where W is the basic Wiener process. Both the drift and variance of the Ito process change over time.

Another process that is often used is the *geometric Brownian motion*. The variable x is said to have a geometric Brownian motion if

$$\frac{\Delta x}{x} = a\Delta t + b\Delta W$$

where W is the basic Wiener process. These processes are used in the description of the movement of stock prices. However, in the sequel we shall be using only the basic Wiener process.

As mentioned earlier, let $\varepsilon_t \sim IN(0, 1)$ and $S_T = \sum_{t=1}^T \varepsilon_t$, then $S_T \sim N(0, T)$ is an $I(1)$ process with independent increments. The distribution of the standardized sums is closely related to the Wiener process and this relationship is known as the *functional central limit theorem* (FCLT). In the study of the asymptotic theory of integrated processes, the distribution of the standardized sums $S_T/\sqrt{T}$ plays an important role. Just as the *central limit theorem* (CLT) plays an important role in the asymptotic theory of stationary processes, the *functional central limit theorem* (FCLT) plays an important role in the asymptotic theory of integrated processes.

The FCLT is based on the mapping of the increasing interval from 0 to T to a fixed interval $[0, 1]$. Divide the interval $[0, 1]$ into $T + 1$ parts of $0, 1/T, 2/T, \dots, 1$. Define a new index r in $[0, 1]$ corresponding to the time index in $[0, T]$ by the relation

$$\frac{j-1}{T} \leq r < \frac{j}{T}, \quad j = 1, \dots, T$$

Next we construct the step function from the standardized sum S_t by defining

$$R_T(r) = \frac{1}{\sqrt{T}} S_{[Tr]}$$

where $[Tr]$ denotes the integer part of Tr , e.g., if $T = 100$ and $r = 0.791$, then $[Tr] = 79$. Since r is the real number in $[0, 1]$, so is Tr .

Figures 3.1 shows how the standardized random walk can be approximated by the step function $R_T(r)$. The first graph in figure 3.1 shows this mapping for the sample size of 10. The other graphs in figures 3.1 show the mapping for sample sizes of 50, 100, and 200, respectively and indicate that as $T \rightarrow \infty$, the step function approaches to the random walk rapidly. In other words, as $T \rightarrow \infty$, $R_T(r)$ becomes increasingly dense on $[0, 1]$ and it converges weakly to the standard Wiener process under some regularity conditions.

This is the idea behind the functional central limit theorem. The result

$$R_T(r) \Rightarrow W(r)$$

is known as *Donsker's theorem* (Donsker, 1951). Here $\Rightarrow$ is used to signify weak convergence of the associated probability measure. For details

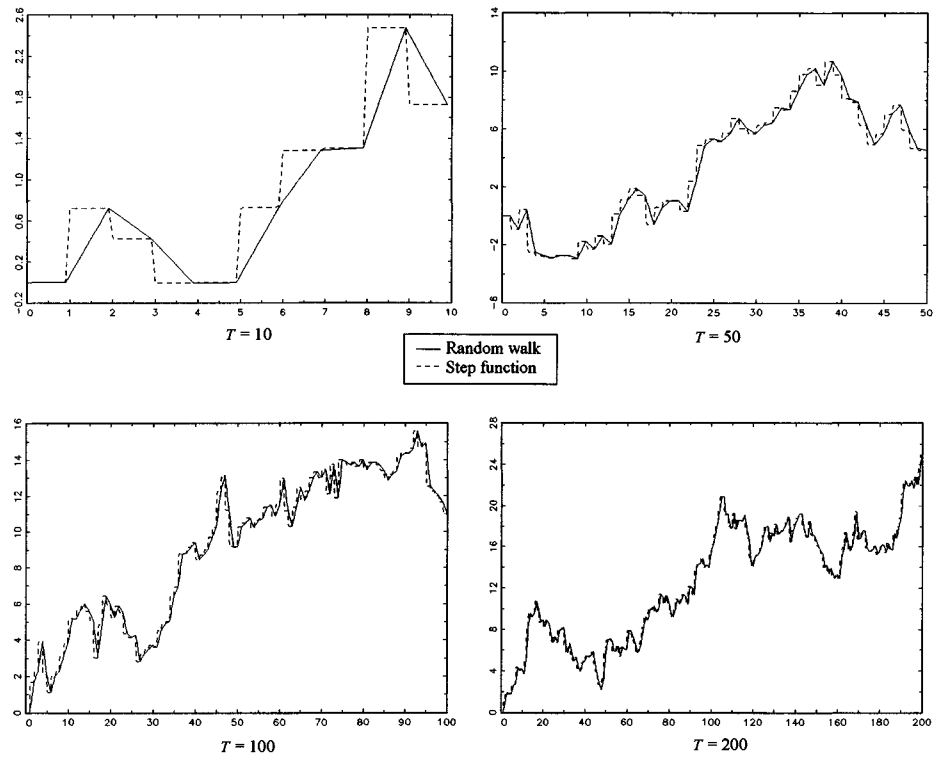


Fig. 3.1. Random walk and step function

and proof see Billingsley (1968). For a wide class of weakly dependent and heterogenously distributed errors, some general results have been established by McLeish (1975) and Herrndorf (1984).

The other important theorem used to derive the asymptotics for unit roots is the *continuous mapping theorem* (CMT). If $f(\cdot)$ is a continuous functional on $[0,1]$, then a consequence of the FCLT is that

$$f(R_T(r)) \Rightarrow f(W(r))$$

For details and proof see Billingsley (1968) and Hall and Heyde (1980).

3.2.2 Some basic results on Wiener processes

We shall now give some basic results on the use of Donsker's theorem and continuous mapping theorem (CMT) in the derivation of the distribution associated with $I(1)$ processes.

Suppose that y_t is a random walk and thus $y_t = y_{t-1} + e_t$ where e_t are $iid(0, 1)$. Assuming the variance to be 1 has the advantage that we can drop the scaling parameter. We assume $y_0 = 0$ for simplicity. We shall show some results in the form of a lemma. (In the following unless otherwise noted, all summations are over $t = 1$ and T .)

Lemma 3.1

- (i) $\frac{\bar{y}}{\sqrt{T}} \Rightarrow \int_0^1 W(r)dr$ where $\bar{y} = \sum_{t=1}^T y_t/T$
- (ii) $T^{-2} \sum y_t^2 \Rightarrow \int_0^1 [W(r)]^2 dr$
- (iii) $T^{-\frac{5}{2}} \sum ty_t \Rightarrow \int_0^1 rW(r)dr$
- (iv) $T^{-\frac{3}{2}} \sum t\varepsilon_t \Rightarrow \int_0^1 r dW(r)$
- (v) $T^{-1} \sum_{t=2}^T y_{t-1}\varepsilon_t \Rightarrow \int_0^1 W(r)dW(r)$

Note the following things in the mapping

$$\begin{aligned} \sum &\Rightarrow \int \\ t &\Rightarrow r \\ e_t &\Rightarrow dW(r) \\ y_t &\Rightarrow W(r) \end{aligned}$$

Proofs (i) Consider

$$\begin{aligned} R_T(r) &= \frac{y_{[Tr]}}{\sqrt{T}} = \frac{y_{i-1}}{\sqrt{T}} \quad \text{for} \quad \frac{i-1}{T} \leq r < \frac{i}{T} \\ R_T(1) &= \frac{y_T}{\sqrt{T}} \end{aligned}$$

where $R_T(r)$ is a step function with steps $y_i/\sqrt{T}$ at i/T and is constant between steps. Hence

$$\begin{aligned} \int_0^1 R_T(r) dr &= \sum_{i=1}^T \int_{(i-1)/T}^{i/T} R_T(r) dr \\ &= \sum_{i=1}^T \frac{y_{i-1}}{\sqrt{T}} \cdot \int_{(i-1)/T}^{i/T} dr \\ &= \frac{1}{T} \sum_{i=1}^T \frac{y_{i-1}}{\sqrt{T}} \end{aligned}$$

Note that

$$\int_{(i-1)/T}^{i/T} dr = [r]_{(i-1)/T}^{i/T} = \frac{i}{T} - \frac{i-1}{T} = \frac{1}{T}$$

Since $\sum_{i=1}^T y_{i-1} \simeq \sum_{i=1}^T y_i$ we have

$$\int_0^1 R_T(r) dr = \frac{\bar{y}}{\sqrt{T}}$$

Using Donsker's theorem we have

$$R_T(r) \Rightarrow W(r)$$

Hence we get

$$\frac{\bar{y}}{\sqrt{T}} \Rightarrow \int_0^1 W(r) dr$$

(ii) We can show as before that

$$\int_0^1 [R_T(r)]^2 dr = \frac{1}{T} \left(\frac{\sum y_{i-1}^2}{T} \right)$$

Since $\sum_{i=1}^T y_{i-1}^2 \simeq \sum_{i=1}^T y_i^2$ and using the CMT we get

$$\frac{\sum_{i=1}^T y_i^2}{T^2} \Rightarrow \int_0^1 [W(r)]^2 dr$$

Proofs of results (iii), (iv), and (v) follow along similar lines and are omitted.

3.2.3 Relationships with normal distributions

Let e_t be $IN(0, 1)$. Define $y_t = \sum_{i=1}^T e_i$. Then $y_T \sim N(0, T)$ and hence

$$\frac{y_T}{\sqrt{T}} \Rightarrow W(1)$$

Note that $W(1) = N(0, 1)$. We shall now show the relation between Wiener processes and normal distributions in the form of a lemma as follows:

Lemma 3.2

- (i) $\int_0^1 W(r)dr \sim N(0, \frac{1}{3})$
- (ii) $\int_0^1 r dW(r) \sim N(0, \frac{1}{3})$
- (iii) $\int_0^1 W(r)dW(r) = \frac{1}{2}[W(1)^2 - 1] \sim \frac{1}{2}[\chi^2(1) - 1]$
- (iv) $\int_0^1 (r - a)W(r)dr \sim N(0, \lambda), \quad \lambda = \frac{1}{60}(8 - 25a + 20a^2)$
- (v) $\int_0^1 rW(r)dr \sim N(0, \frac{2}{15}) \quad \text{if } a = 0.$
- (vi) *If $W(r)$ and $V(r)$ are two independent Wiener processes, then*

$$\left[\int_0^1 [W(r)]^2 dr \right]^{-1/2} \left[\int_0^1 W(r)dV(r) \right] \sim N(0, 1)$$

Proof (i) We note that

$$\begin{aligned} \sum_1^T t &= \frac{T(T+1)}{2}, \\ \sum_1^T t^2 &= \frac{T(T+1)(2T+1)}{6} \end{aligned}$$

If y_t is a random walk, then

$$y_t = y_{t-1} + e_t, \quad e_t \sim IN(0, 1), \quad t = 1, \dots, T$$

and $y_0 = 0$. Consider $T^{-\frac{1}{2}}\bar{y} = T^{-\frac{3}{2}} \sum y_t$ and $y_t = e_1 + e_2 + \dots + e_t$. Thus, when we consider $\sum y_t$, we have $\sum y_t = Te_1 + (T-1)e_2 + \dots + e_T$. Hence $\text{var}(\sum y_t) = \sum_1^T t^2$. Since the e s are $IN(0, 1)$, we have

$$T^{-\frac{3}{2}} \sum y_t \sim N(0, V)$$

where

$$V = T^{-3} \sum_1^T t^2 = T^{-3} \frac{T(T+1)(2T+1)}{6} \approx \frac{1}{3}$$

Earlier we showed that $T^{-\frac{3}{2}} \bar{y} \Rightarrow \int_0^1 W(r) dr$. Hence $\int_0^1 W(r) dr \approx N(0, \frac{1}{3})$.

(ii) Consider next $T^{-\frac{3}{2}} \sum t e_t$. Again this is $N(0, V)$, since e_t are $IN(0, 1)$ and

$$V = T^{-3} \sum_1^T t^2 = T^{-3} \frac{T(T+1)(2T+1)}{6} \approx \frac{1}{3}$$

Thus

$$\int_0^1 r dW(r) \approx N(0, \frac{1}{3})$$

(iii) Since

$$\sum y_t^2 = \sum (y_{t-1} + e_t)^2 = \sum y_{t-1}^2 + \sum e_t^2 + 2 \sum y_{t-1} e_t$$

we have

$$\sum y_{t-1} e_t = \frac{1}{2} \left[\sum y_t^2 - \sum y_{t-1}^2 - \sum e_t^2 \right] = \frac{1}{2} \left[y_T^2 - \sum e_t^2 \right]$$

Note that

$$y_T = e_1 + e_2 + \cdots + e_T \sim N(0, T)$$

Hence

$$\frac{y_T}{\sqrt{T}} \sim N(0, 1)$$

and

$$\frac{y_T^2}{T} \sim \chi_1^2$$

Also

$$\frac{1}{T} \sum e_t^2 \rightarrow \text{var}(e_t) = 1$$

Hence

$$\frac{1}{T} \sum y_{t-1} e_t \Rightarrow \frac{1}{2} [\chi_1^2 - 1]$$

Since we showed that

$$\frac{1}{T} \sum y_{t-1} e_t \Rightarrow \int_0^1 W(r) dW(r)$$

result (iii) follows. Proofs of results (iv), (v), and (vi) are omitted here.

Note on (iii) In the general case where e_t are not independent, but

$$\text{var}(e_t) = \sigma_e^2$$

and

$$\lim_{T \rightarrow \infty} E(T^{-1}y_T^2) = \sigma^2 (> 0)$$

we have

$$\frac{y_T}{\sqrt{T}} \sim N(0, \sigma)$$

and

$$\frac{y_T^2}{T} \sim \sigma \chi_1^2$$

In this case we have

$$\frac{1}{T} \sum y_{t-1}e_t \Rightarrow \frac{\sigma^2}{2} \left[\chi_1^2 - \frac{\sigma_e^2}{\sigma^2} \right] = \frac{\sigma^2}{2} [\chi_1^2 - 1] + \left(\frac{\sigma^2 - \sigma_e^2}{2} \right)$$

This equation is the basis of the Phillips-Perron test that will be discussed later. Note that if $e_t \sim iid(0, \sigma_e^2)$, then $\sigma^2 = \sigma_e^2$ and we have

$$\frac{1}{T} \sum y_{t-1}e_t \Rightarrow \frac{\sigma_e^2}{2} [\chi_1^2 - 1]$$

When we consider a multiple (n -vector) time series and $n > 1$, Phillips (1988b) shows that for $n > 1$, as $T \rightarrow \infty$,

$$T^{-1} \sum_1^T y_{t-1}e'_t \Rightarrow \int_0^1 B(r)dB(r)' + \Omega_1$$

where $B(r)$ is vector Brownian motion with covariance matrix Ω and

$$\Omega_1 = \lim_{T \rightarrow \infty} T^{-1} \sum_1^T E(y_{t-1}e'_t) = \sum_{k=2}^{\infty} E(e_1e'_k)$$

For the proof, see Phillips (1988b).

3.2.4 Scaling factors in asymptotic distributions

When we discuss the asymptotic distributions of regression coefficients where the regressors are trending variables, it is important to remember the scaling factors. For instance in the usual regression model

$$y_t = \beta x_t + u_t, \quad u_t \sim iid(0, \sigma^2) \quad t = 1, 2, \dots, T$$

we know that $\text{var}(\hat{\beta}) = \sigma^2 / \sum x_t^2$. The usual assumption is that

$$\frac{1}{T} \sum x_t^2 \rightarrow \text{a constant}$$

or

$$\sum x_t^2 = O(T)$$

In this case we consider the asymptotic distribution of $\sqrt{T}(\hat{\beta} - \beta)$, which had a scaling factor $\sqrt{T}$. Then the asymptotic distribution has a finite variance.

What if $x_t = t$? As we have seen above

$$\sum_1^T t^2 = \frac{T(T+1)(2T+1)}{6}$$

or

$$\sum t^2 = O(T^3)$$

In this case we have to consider the asymptotic distribution of $T^{3/2}(\hat{\beta} - \beta)$ which will have a finite variance. If $x_t = t^2$, we consider the asymptotic distribution of $T^{5/2}(\hat{\beta} - \beta)$ and so on. For example, in the regression

$$y_t = \alpha + \beta t + \gamma t^2 + \delta t^3 + u_t$$

we have to consider the asymptotic distributions of

$$\sqrt{T}(\hat{\alpha} - \alpha), \quad T\sqrt{T}(\hat{\beta} - \beta), \quad T^2\sqrt{T}(\hat{\gamma} - \gamma), \quad \text{and} \quad T^3\sqrt{T}(\hat{\delta} - \delta)$$

Note that we have

$$\sum_1^T 1 = T = O(T)$$

$$\sum_1^T t = \frac{T(T+1)}{2} = O(T^2)$$

$$\sum_1^T t^2 = \frac{T(T+1)(2T+1)}{6} = O(T^3)$$

$$\sum_1^T t^3 = \left[\frac{T(T+1)}{2} \right]^2 = O(T^4)$$

In general

$$\sum_1^T t^j = O(T^{j+1})$$

If $\sum x_t^2 = O(T^{j+1})$, then the appropriate scaling factor is $T^{(2j-1)/2}$ to get the asymptotic distribution of corresponding coefficient of x_t . In the subsequent chapters we will use the appropriate normalization rules when considering asymptotic distributions.

3.3 Unit root tests without a deterministic trend

3.3.1 Some historical notes

Most of the papers about the asymptotic behavior of the OLS estimator $\hat{\rho}$ have been concerned with the AR(1) model without the deterministic trend

$$y_t = \rho y_{t-1} + e_t \quad (3.1)$$

where $\{e_t\}$ is a sequence of independent normal random variables with mean zero and variance σ^2 , i.e., $e_t \sim IN(0, \sigma^2)$. We may rewrite (3.1) as

$$y_t = \rho^t y_0 + \sum_{i=0}^{t-1} \rho^i e_{t-i}$$

The distribution of the successive y s is not uniquely determined by that of the e s alone. Since y_t depends on y_0 , the distribution of y_0 must also be specified.

Three distributions have been commonly proposed for y_0 as follows (White, 1958; Phillips, 1987a):

- (i) y_0 = a constant with probability one.
- (ii) y_0 has a certain specified distribution.
- (iii) $y_0 = y_T$ where T is the sample size.

Distribution (iii) is the so-called circularity condition that is used mainly as a mathematical device to simplify distribution theory (Anderson, 1942). Distribution (ii) permits the greatest flexibility that allows for a nonstationary series and includes (i) as a special case (Phillips, 1987a). The assumption (i) is the commonly used condition (Dickey and Fuller, 1979; Evans and Savin, 1981). It can be shown that the asymptotic distribution of normalized $\hat{\rho}$ does not depend upon y_0 if $|\rho| \leq 1$ (White, 1958). If $|\rho| > 1$, the asymptotic distribution depends upon y_0 , and as

y_0/σ increases the distribution of $\hat{\rho}$ becomes more bell-shaped (Evans and Savin, 1981). Needless to say, the finite sample distribution of $\hat{\rho}$ depends upon y_0 .

Given n observations, the maximum likelihood estimator of ρ is the least squares estimator

$$\hat{\rho} = \left(\sum_{t=1}^n y_{t-1}^2 \right)^{-1} \sum_{t=1}^n y_t y_{t-1} \quad (3.2)$$

Mann and Wald (1943) first proved that if $|\rho| < 1$, then

$$\frac{\sqrt{T}(\hat{\rho} - \rho)}{\sqrt{1 - \rho^2}} \Rightarrow N(0, 1)$$

under the *iid* assumption of the errors. Rubin (1950) showed that $\hat{\rho}$ is a consistent estimator for all values of ρ . White (1958) has shown that the limiting distribution of $|\rho|^T(\hat{\rho} - \rho)/(\rho^2 - 1)$ is a Cauchy distribution by inverting the limiting joint-moment generating function when $|\rho| > 1$. On the other hand, when $\rho = 1$, White was unable to invert the limiting joint-moment generating function to obtain the limiting distribution. However, based on Donsker's theorem, he represented the limiting distribution of $T(\hat{\rho} - 1)/\sqrt{2}$ as that of the ratio of two integrals defined on the Wiener process. Later M. M. Rao (1961) obtained an expression of the limiting distribution of $T(\hat{\rho} - 1)/\sqrt{2}$ but it is not of an easily recognizable form. However, the standardization of $\hat{\rho}$ as $T(\hat{\rho} - \rho)/\sqrt{2}$ by White (1958) and M. M. Rao (1961) has been proved incorrect and the correct standardization of $\hat{\rho}$ is $T(\hat{\rho} - \rho)$ (Phillips, 1987a). Fuller (1976) and Dickey and Fuller (1979) derived the asymptotic distribution of $\hat{\rho}$ under the assumptions of the *iid* errors and assumption of $y_0 = 0$. Phillips (1987a) derived it under the more general assumption about the errors and assumption (ii) of y_0 .

3.3.2 The Dickey-Fuller distribution

Using the results on Wiener processes in the previous sections, we shall derive the distribution of $\hat{\rho}$ under the null hypothesis of a unit root, that is, $\rho = 1$. From the equation (3.2), we have

$$T(\hat{\rho} - 1) = \frac{T^{-1} \sum_{t=1}^T y_{t-1} e_t}{T^{-2} \sum_{t=1}^T y_{t-1}^2} \quad (3.3)$$

As we have seen in the section 3.2.2, the numerator in (3.3) converges to a random variable

$$T^{-1} \sum_{t=1}^T y_{t-1} e_t \Rightarrow \int_0^1 W(r) dW(r) \quad (3.4)$$

Interpreting (3.4) heuristically, averaging corresponds to integrating (analogous to taking expectations). We construct the step function from $y_{t-1}/\sqrt{T}$ which converges weakly to the Wiener process $W(r)$. And then $e_t\sqrt{T}(=\Delta y_t\sqrt{T})$ is the innovation over which averaging occurs and corresponds to $dW(r)$. The denominator has a nonzero asymptotic variance and

$$T^{-2} \sum_{t=1}^T y_{t-1}^2 \Rightarrow \int_0^1 W(r)^2 dr \quad (3.5)$$

By the continuous mapping theorem, the limiting distribution of the OLS estimator $\hat{\rho}$ when $\rho = 1$ is

$$T(\hat{\rho} - 1) \Rightarrow \frac{\int_0^1 W(r) dW(r)}{\int_0^1 W(r)^2 dr} \quad (3.6)$$

This is known as Dickey–Fuller distribution. Although Fuller (1976) and Dickey and Fuller (1979) did not provide the limiting distribution of $\hat{\rho}$ with functionals of Wiener processes, they first derived its distribution and made the tables of this distribution. The functional form of Wiener processes is due to Phillips (1987a, 1988b).

We can use the distribution (3.6) for testing the unit root null hypothesis $H_0 : \rho = 1$, that is

$$K = T(\hat{\rho} - 1)$$

or we can normalize it with the standard error of the OLS estimator and construct the t -statistic testing $\rho = 1$. The test statistic is

$$t_{\hat{\rho}} = \frac{(\hat{\rho} - 1)}{SE(\hat{\rho})}$$

The asymptotic distribution of the t -type test statistic is given by

$$t_{\hat{\rho}} \Rightarrow \frac{\int_0^1 W(r) dW(r)}{[\int_0^1 W(r)^2 dr]^{1/2}} = \frac{\frac{1}{2}[W(1)^2 - 1]}{[\int_0^1 W(r)^2 dr]^{1/2}} \quad (3.7)$$

The proof follows from the fact that

$$t_{\hat{\rho}} = \frac{T(\hat{\rho} - 1)}{\sqrt{T(1 - \hat{\rho}^2)}}$$

The numerator is given by equation (3.6). The denominator can be written as

$$\begin{aligned} \left[\frac{T^{-1} \sum y_t^2 (1 - \hat{\rho}^2)}{T^{-2} \sum y_t^2} \right]^{1/2} &= \left[\frac{T^{-1} \sum \hat{e}_t^2}{T^{-2} \sum y_t^2} \right]^{1/2} \\ &\Rightarrow \frac{1}{(T^{-2} \sum y_t^2)^{1/2}} \Rightarrow \frac{1}{\left(\int_0^1 W(r)^2 dr \right)^{1/2}} \end{aligned}$$

Hence the distribution of the t -type statistic is given by (3.7). This distribution is known as the Dickey–Fuller t -distribution. The test procedure using K or $t_{\hat{\rho}}$ for the unit root null hypothesis is called the Dickey–Fuller (DF) test.

The asymptotic distribution of the t -statistic under $\rho = 1$ is not the standard t -distribution, and so conventional critical values are not valid. The numerator of the distribution is skewed to the right, being a $\chi^2(1)$ minus its expectation, but the ratio is skewed to the left. In practice, most of the t -statistic outcomes will be negative and often very different from -2.0 even for large sample sizes. Thus, using the conventional critical values can lead to considerable overrejection of the null hypothesis of a unit root. Fuller (1976) provided the critical values of these statistics. Table 3.1 shows some of those critical values.

The following example was given in Dickey and Fuller (1979). Gould and Nelson (1974) investigated the stochastic structure of the velocity of money using the yearly observations from 1869 through 1960 given in Friedman and Schwartz (1963). They considered the following models

$$y_t - y_1 = 1.0044 (y_{t-1} - y_1) + e_t, \quad \hat{\sigma}^2 = 0.0052 \quad (3.8) \\ (0.0094)$$

$$y_t = 0.0141 + 0.970 y_{t-1} + e_t, \quad \hat{\sigma}^2 = 0.0050 \quad (3.9) \\ (0.0176) \quad (0.0199)$$

Model (3.8) assumes that it is known that no intercept enters the model if y_1 is subtracted from all observations. Model (3.9) permits an intercept in the model. The test statistics for the unit root null are for model (3.8)

$$K = 91(1.0044 - 1) = 0.4004$$

$$t = (1.0044 - 1)/0.0094 = 0.4681$$

Table 3.1. *Critical values for Dickey-Fuller tests*

Sample size	<i>K</i> -test		<i>t</i> -test		<i>F</i> -test (a)		<i>F</i> -test (b)	
	1%	5%	1%	5%	1%	5%	1%	5%
AR(1) without a drift								
25	-11.9	-7.7	-2.62	-1.95				
100	-13.3	-7.9	-2.60	-1.95				
250	-13.6	-8.0	-2.58	-1.95				
500	-13.7	-8.0	-2.58	-1.95				
∞	-13.8	-8.1	-2.58	-1.95				
AR(1) with a drift								
25	-17.2	-12.5	-3.75	-3.00	7.88	5.18		
50	-18.9	-13.3	-3.58	-2.93	7.06	4.86		
100	-19.8	-13.7	-3.51	-2.89	6.70	4.71		
250	-20.3	-14.0	-3.46	-2.88	6.52	4.63		
500	-20.5	-14.0	-3.44	-2.87	6.47	4.61		
∞	-20.7	-14.1	-3.43	-2.86	6.43	4.59		
AR(1) with a drift and a time trend								
25	-22.5	-17.9	-4.38	-3.60	10.61	7.24	8.21	5.68
50	-25.7	-19.8	-4.15	-3.50	9.31	6.73	7.02	5.13
100	-27.4	-20.7	-4.04	-3.45	8.73	6.49	6.50	4.88
250	-28.4	-21.3	-3.99	-3.43	8.43	6.34	6.22	4.75
500	-28.9	-21.5	-3.98	-3.42	8.34	6.30	6.15	4.71
∞	-29.5	-21.8	-3.96	-3.41	8.27	6.25	6.09	4.68

Notes: *F*-test (a) is for $H_0 : \alpha = 0, \rho = 1$ in $y_t = \alpha + \rho y_t + u_t$
and $H_0 : \delta = 0, \rho = 1$ in $y_t = \alpha + \delta t + \rho y_t + u_t$.

F test (b) is for $H_0 : \alpha = 0, \delta = 0, \rho = 1$ in $y_t = \alpha + \delta t + \rho y_t + u_t$.

Sources: Fuller (1976, p. 371) for the *K*-test and p. 373 for the *t*-test.

Dickey and Fuller (1981, p. 1063) for the *F*-test.

More detailed percentiles at the levels:

1%, 2.5%, 5%, 10%, 50%, 90%, 95%, 97.5%, and 99%

for the *K*-test and the *t*-test can be found in Fuller (1996,
pp. 641 and 642, respectively).

and for model (3.9)

$$K = 92(0.9702 - 1) = -2.742$$

$$t = (0.9702 - 1)/0.0094 = -1.50$$

Thus using table 3.1 the null hypothesis of a unit root cannot be rejected at the 5 percent level. They concluded that the logarithm of velocity is consistent with the random walk model.

3.3.3 Computation of critical values

It is worth noting that different methods have been used to compute the critical values of the limiting distributions in the presence of a unit root, since the standard DF test requires different critical values for the different specifications of deterministic trend even with time invariant parameters.

Evans and Savin (1981, 1984) provide the exact finite sample distribution calculated by the method due to Imhof (1961) or the method due to Pan (1968). Although the effects of a nonzero value of y_0 vanish asymptotically, they are substantial in finite samples as demonstrated by Evans and Savin (1981) (for $\hat{\rho}$) and Nankervis and Savin (1985) (for $t_{\hat{\rho}}$ in case of *iid* errors). Under the assumption of the nonzero y_0 , the actual size of the test is well below the (asymptotic) nominal size.

For calculating the limiting distribution, several methods have been developed. The most commonly used method is the simulation method; that is, the integral functions in the limiting distribution are approximated by functions of partial sums of independent normal random variables. See appendix B of Zivot and Andrews (1992). Phillips (1977) and Satchell (1984) considered an Edgeworth approximation but this performs poorly. Nabeya and Tanaka (1990) proposed Fredholm approximation for computing the limiting distribution, which is based on the relationship between the kernel of the characteristic function and the Fredholm determinant. Perron (1991) suggested the continuous time approximation which is based on the Ornstein–Uhlenbeck diffusion process. The approximations proposed by Nabeya and Tanaka (1990) and Perron (1991) are based on the assumption $\rho = 1 - c/T$ and $\exp(-c/T)$, respectively, where c is a fixed constant. This is known as a *local to unity* process. As noted by Chan and Wei (1987), this is similar to that in approximating binomial distributions by Poisson distributions when the number of trials is large but the probability of success is near 0. These methods are useful to study the power of a test of unit root against near unit root where c is small. More detailed discussion of all these issues is in Tanaka (1996).

3.4 DF test with a linear deterministic trend

In the previous section we discussed the distribution of $\hat{\rho}$ under the hypothesis $\rho = 1$ when both the data generating process (DGP) and the estimated equation had no drift (constant term) or trend. We shall

now discuss models with drift and trend. Here, the important thing to remember is the difference between the maintained null hypothesis (or the true DGP) and the estimating regressions that are used to test the null hypothesis.

3.4.1 Models with drift

We can consider two possible cases where y_t are believed to be generated by the following two DGPs:

- (i) $y_t = y_{t-1} + e_t$
- (ii) $y_t = \alpha + y_{t-1} + e_t$

For the given data y_t when we estimate the regression

$$y_t = \alpha + \rho y_{t-1} + e_t \quad (3.10)$$

The asymptotic distribution of $\hat{\rho}$ will depend on whether the DGP is (i) or (ii).

(i) Let us consider the first DGP. Dickey and Fuller (1979) derived the asymptotic distribution of the standardized OLS estimates. The asymptotic distribution can be expressed in terms of Wiener processes as follows

$$T(\hat{\rho} - 1) \Rightarrow \frac{\int W_* dW}{\int W_*^2 dr} \quad (3.11)$$

where $W_* = W(r) - \int W(r) dr$ is the “demeaned” Brownian motion. The asymptotic distribution of the OLS estimates $\hat{\rho}$ in (3.11) is not the same as the asymptotic distribution in (3.6). The corresponding t -statistic has the asymptotic distribution

$$t_{\hat{\rho}} \Rightarrow \frac{\int W_* dW}{(\int W_*^2 dr)^{\frac{1}{2}}} \quad (3.12)$$

The t -statistic for the constant term also follows the nonstandard asymptotic distribution, a functional of the Wiener processes.

When a constant term is used in the estimated regression and y_t is generated by the pure random walk (without a drift), then a different table of critical values must be used. The critical values of $T(\hat{\rho} - 1)$ and $t_{\hat{\rho}}$ are tabulated by Fuller (1976) and those for $t_{\hat{\alpha}}$ are given in Dickey and Fuller (1981). Part of the critical values are given in table 3.1.

Although the test statistics are for the unit root null hypothesis, a maintained hypothesis on which the asymptotic distribution was derived

is that the true value of α is zero. Thus naturally we could consider the joint hypothesis of $H_0 : \alpha = 0$ and $\rho = 0$. Dickey and Fuller (1981) derived the asymptotic distribution of the F -statistics for the joint hypothesis and tabulated the critical values by Monte Carlo methods. Part of the critical values are given in table 3.1.

(ii) Now consider the second DGP, that is, a random walk with a drift. Recall that the random walk with drift can be written as

$$y_t = y_0 + \alpha t + \sum_{i=1}^t e_i$$

Since y_t has a deterministic trend inherently, y_t is dominated by the deterministic time trend asymptotically. Thus in the denominator of the OLS estimator of $\hat{\rho}$, $\sum_{t=1}^T y_{t-1}^2$, the regressor y_{t-1} is asymptotically dominated by the time trend $\alpha(t-1)$. In large samples it is as if the explanatory variable y_{t-1} were replaced by the time trend $\alpha(t-1)$. As a result, the asymptotic distributions of the OLS estimate are asymptotically normal and the t - and F -statistics follow the standard t - and F -distributions.

The above results have been studied by West(1988) who shows the asymptotic normality of unit root test statistics. However, as we will argue later, this result is of limited usefulness in the context of unit root tests, although it illustrates the importance of the use of scaling factors. The asymptotic normality result is as follows: under the unit root null $H_0 : \rho = 1$, y_t has a linear trend, and assuming $y_0 = 0$ for simplicity we have

$$y_t = \alpha t + \sum_{j=1}^t e_j = \alpha t + S_t$$

As discussed earlier, we need to scale the coefficients properly. Thus we consider the distribution of $\sqrt{T}(\hat{\alpha} - \alpha)$ and $T^{3/2}(\hat{\rho} - 1)$. We have

$$\begin{bmatrix} T^{1/2}(\hat{\alpha} - \alpha) \\ T^{3/2}(\hat{\rho} - 1) \end{bmatrix} = A^{-1}b$$

where

$$A = \begin{bmatrix} 1 & T^{-2} \sum y_{t-1} \\ T^{-2} \sum y_{t-1} & T^{-3} \sum y_{t-1}^2 \end{bmatrix}$$

and

$$b = \begin{bmatrix} T^{-1/2} \sum e_t \\ T^{-3/2} \sum y_{t-1} e_t \end{bmatrix}$$

Consider the vector b . Since

$$\sum y_{t-1}e_t = \sum \alpha(t-1)e_t + \sum e_t S_{t-1}$$

by normalizing $T^{-3/2}$ we have from the first term of the right-hand side of equation

$$\text{var}(T^{-3/2} \sum te_t) = \frac{\sigma_e^2}{3}$$

and from the second term of the right-hand side of equation

$$T^{-3/2} \sum e_t S_{t-1} \rightarrow 0 \quad \text{in probability}$$

As noted in the section 3.2.2

$$T^{-1} \sum e_t S_{t-1} \Rightarrow \frac{\sigma_e^2}{2} [\chi^2(1) - 1]$$

Thus

$$T^{-3/2} \sum y_{t-1}e_t \Rightarrow N\left(0, \frac{\alpha^2 \sigma_e^2}{3}\right)$$

Hence

$$b \Rightarrow N(0, \sigma_e^2 B)$$

where

$$B = \begin{bmatrix} 1 & \alpha/2 \\ \alpha/2 & \alpha^2/3 \end{bmatrix}$$

The scaling factors are such that all terms with the stochastic component S_t can be ignored asymptotically.

Consider the matrix A

$$T^{-3} \sum y_{t-1}^2 = T^{-3} \sum [\alpha(t-1) + S_{t-1}]^2$$

Again the scaling factor T^{-3} ensures that

$$T^{-3} \sum S_{t-1}^2 \rightarrow 0 \quad \text{in probability}$$

and

$$T^{-3} \sum (t-1)S_{t-1} \rightarrow 0 \quad \text{in probability}$$

All we are left with is

$$T^{-3} \sum \alpha^2(t-1)^2 \rightarrow \frac{\alpha^2}{3}$$

Similarly

$$T^{-2} \sum y_{t-1} = T^{-2} \sum [\alpha(t-1) + S_{t-1}] \rightarrow \frac{\alpha}{2}$$

Thus the matrix $A \rightarrow B$ and since $b \Rightarrow N(0, \sigma_u^2 B)$ we have

$$\begin{bmatrix} T^{1/2}(\hat{\alpha} - \alpha) \\ T^{3/2}(\hat{\rho} - 1) \end{bmatrix} = A^{-1}b \Rightarrow N(0, \sigma_u^2 B^{-1})$$

This gives us the result

$$T^{3/2}(\hat{\rho} - 1) \Rightarrow N\left(0, \frac{12\sigma_e^2}{\alpha^2}\right)$$

Note that this result holds only if $\alpha \neq 0$.

When the true DGP (not estimating regression) is

$$y_t = \alpha + \delta t + \rho y_{t-1} + e_t$$

again we have, under $\rho = 1$

$$y_t = \alpha t + \delta t^2 + S_t$$

and the deterministic term $\alpha t + \delta t^2$ dominates S_t in the derivation of the asymptotic distributions. In this case the scale factors for $\hat{\alpha}$, $\hat{\delta}$, and $\hat{\rho}$ are $T^{-1/2}$, $T^{-3/2}$, and $T^{-5/2}$ respectively. The matrix B is now

$$B = \begin{bmatrix} 1 & 1/2 & \beta/3 \\ 1/2 & 1/3 & \beta/4 \\ \beta/3 & \beta/4 & \beta^2/5 \end{bmatrix}$$

and hence

$$T^{5/2}(\hat{\rho} - 1) \Rightarrow N\left(0, \frac{180\sigma_e^2}{\alpha^2}\right)$$

The reason why this result on asymptotic normality is not useful is that the behavior of y_t under the null and the behavior under the alternative are different. Consider the case

$$y_t = \alpha + \rho y_{t-1} + e_t$$

If $\alpha \neq 0$, then under $H_0 : \rho = 1$, y_t has a linear trend whereas under $H_1 : |\rho| < 1$ it has no trend. Similarly, if we consider the model

$$y_t = \alpha + \delta t + \rho y_{t-1} + e_t$$

If $\delta \neq 0$, then, under $H_0 : \rho = 1$, y_t has a quadratic trend and, under $H_1 : |\rho| < 1$, y_t has a linear trend. This problem does not arise with the approach suggested by Bhargava (1986) whose formulation we have discussed earlier in chapter 2 (section 2.9).

3.4.2 Models with linear trend

Now consider the more general estimating regression (not DGP) with a constant term and a linear deterministic trend

$$y_t = \alpha + \delta t + \rho y_{t-1} + e_t$$

In this case, even though y_{t-1} would be asymptotically equivalent to a time trend when $\alpha \neq 0$, since the deterministic trend is incorporated as a separate regressor into the estimating regression, the drift term in the maintained DGP α turns out not to affect the asymptotic distribution. Thus the asymptotic distribution of the OLS estimator $\hat{\rho}$ is invariant to the value of α and the maintained hypothesis, $H_0 : \rho = 1$ and $\delta = 0$.

Again we consider two DGPs:

- (i) Random walk with drift only.
- (ii) Random walk with linear trend.

(i) When the DGP is a random walk with drift, that is, $\rho = 1$ and $\delta = 0$, the asymptotic distributions of $\hat{\rho}$ and the corresponding t -statistics are as follows

$$T(\hat{\rho} - 1) \Rightarrow \frac{\int W(r)dW + A}{D} \quad (3.13)$$

and

$$t_{\hat{\rho}} \Rightarrow \frac{\int W(r)dW + A}{D^{\frac{1}{2}}} \quad (3.14)$$

where

$$\begin{aligned} D &= \int W(r)^2 dr - 12 \left(\int rW(r)dr \right)^2 \\ &\quad + 12 \int W(r)dr \int rW(r)dr - 4 \left(\int W(r)dr \right)^2 \\ A &= 12 \left[\int rW(r)dr - \frac{1}{2} \int W(r)dr \right] \\ &\quad \times \left[\int W(r)dr - \frac{1}{2} W(1) \right] - W(1) \int W(r)dr \end{aligned}$$

The t -statistics for the OLS estimator $\hat{\alpha}$ and $\hat{\delta}$ also follow the functionals of the Wiener processes. Fuller (1976) tabulated the critical values. Part of those critical values are given in table 3.1.

It is natural to consider the F -test of the joint hypotheses $H_0 : \rho = 1$ and $\delta = 0$ and $H_0 : \rho = 1, \alpha = 0$ and $\delta = 0$. Dickey and Fuller (1981) provide the asymptotic distributions of the F -statistics and tabulated

Table 3.2. *Asymptotic distributions of the t-ratios for different DGPs*

Regression	DGP	Asymptotic distribution
No deterministic trend ($\alpha = \delta = 0$)	$\alpha = \delta = 0$	A function of Wiener processes (equation 3.6)
With a drift ($\alpha \neq 0, \delta = 0$)	$\alpha = \delta = 0$	A function of Wiener processes (equation 3.12)
	$\alpha \neq 0, \delta = 0$	Normal distribution
With a linear trend ($\alpha \neq 0, \delta \neq 0$)	$\alpha = \delta = 0$	A function of Wiener processes (equation 3.14)
	$\alpha \neq 0, \delta \neq 0$	Normal distribution

the critical values. Part of the critical values are given in table 3.1.

(ii) If the true DGP also has a linear trend, as shown earlier, the test statistic of $\hat{\rho}$ has an asymptotic normal distribution.

For different DGPs, the asymptotic distributions of the t -ratio for the unit root null from estimating regressions are summarized in table 3.2.

3.4.3 An illustrative example

The following example is given in Dickey and Fuller (1981). They applied the above procedure for the logarithm of the quarterly Federal Reserve Board Production Index 1950:I to 1970:IV (110 observations). They assume that the series is adequately represented by the model

$$y_t = \alpha + \delta t + \rho y_{t-1} + \beta \Delta y_{t-1} + e_t$$

where e_t are assumed to be $iin(0, \sigma^2)$. The OLS estimates are

$$y_t = 0.52 + 0.0012 t - 0.119 y_{t-1} + 0.498 \Delta y_{t-1}, \quad RSS = 0.056448$$

$$(0.15) \quad (0.00034) \quad (0.033) \quad (0.081)$$

$$y_t = 0.0054 + 0.447 \Delta y_{t-1}, \quad RSS = 0.063211$$

$$(0.0025) \quad (0.083)$$

$$y_t = 0.511 \Delta y_{t-1}, \quad RSS = 0.065966$$

$$(0.079)$$

where RSS is the residual sum of squares. The number in parentheses are the standard errors. The t -statistic for the unit root null is

$$t_{\hat{\rho}} = -0.119/0.033 = -3.61$$

This leads to rejection of the unit root hypothesis at the 5 percent level, since $t_{\hat{\rho}} < -3.45$. The F -statistic for the null hypothesis $H_0 : \alpha = 0, \delta = 0$ and $\rho = 1$ is computed

$$F = (0.065966 - 0.056448)/3(0.056448/106) = 5.96$$

Since $F > 4.88$ in table 3.1, one can reject the unit root without a drift at the 5 percent level. On the other hand, the F -statistic for the null hypothesis $H_0 : \delta = 0$ and $\rho = 1$ is computed

$$F = (0.063211 - 0.056448)/2(0.056448/106) = 6.34$$

Since $F < 6.49$ in table 3.1, one cannot reject the unit root with possible drift at the 5 percent level. Thus there is some evidence to support the unit root null hypothesis with possible drift at the 5 percent significant level.

3.5 Specification of deterministic trends

It is hard to believe that the pure AR(1) model without the deterministic trend describes well most of the macroeconomic variables. Almost all macroeconomic variables usually show some tendency to increase over time. For example GNP and consumption increase over time. This suggests that it may be appropriate to incorporate the linear trend term into the model. When we include the linear trend term into the model, we can classify the time series into two important classes which imply the different methods of eliminating the trend. These classes are the *trend-stationary process* (TSP) and the *difference-stationary process* (DSP) In order to identify which class macroeconomic series belong to, Nelson and Plosser (1982) apply the above Dickey–Fuller tests for identifying the classes of the macroeconomic series.

Besides the importance of the presence of the deterministic trend in macroeconomic series, the functional form (or specification) of the deterministic trend plays an essential role in the unit root testing procedure and it is also closely related to the power and size of the unit root tests. If we omit a trend variable, which is in the true DGP, from the

estimating regression, then the power of the t -test goes to zero as the sample size increases (Perron, 1988). Thus it is important to include as many deterministic regressors as there are deterministic components in the trend function of the DGP. Otherwise the test will at best lose finite sample power or at worst have power that goes to zero as the sample size increases. On the other hand, it is desirable not to include extraneous deterministic regressors. The power of a test decreases as additional deterministic regressors are included.

Campbell and Perron (1991) argue that “the proper handling of deterministic trends is a vital prerequisite for dealing with unit roots.” Perron (1988) proposed a sequential testing strategy and argued that a proper testing strategy should start from the most general trend specification and test down to more restricted specifications. However, in the more general models which allow more than a simple linear trend, such a sequential testing procedure cannot yet be applied given that the distribution theory for the relevant statistics has not been derived.

Cochrane (1991a, p. 202) argued that Campbell and Perron’s (1991) advice is correct and sensible, but expressed some reservations:

One never knows the deterministic trends with great precision before analysis begins. Economic theory does not give any guidance. ... “proper handling” of deterministic trend is an impossible task. To a humble *macroeconomist* it would seem that an edifice of asymptotic distribution theory that depends crucially on *unknowable* quantities must be pretty useless in practice.

Recent research has demonstrated the importance of using a flexible specification for the deterministic component. It is motivated by the view that the linear time trend hypothesis is inappropriate for modeling the deterministic component of an economic time series. Various models for the flexible deterministic trend have been proposed such as the structural break model, the Markov switching model, and the Bayes updating models. These will be discussed in later chapters.

Before discussing the application of the unit root tests for discriminating the DSP and the TSP classes, it is important to consider the possibility of the serial correlation of errors. The Dickey–Fuller testing procedures, discussed so far, assume *iid* errors and $y_0 = 0$. These are strong assumptions for the real world. In the next section we shall discuss some modified testing procedures for unit roots under the assumption of a wide class of errors where the serial correlation and some heteroskedasticity of errors are allowed.

3.6 Unit root tests for a wide class of errors

In general, the asymptotic distribution of the statistics we have considered so far, are not invariant to other parameters and depend on the other regression parameters and the variances of errors. The most important noninvariance shared by all regression models (discussed in the previous section) is that the limiting (as well as finite sample) distributions of all the statistics considered depend upon the correlation structure of the errors. More precisely, the limiting distributions depend upon the ratio σ^2/σ_e^2 , where

$$\sigma^2 = \lim_{T \rightarrow \infty} T^{-1} E\left(\left(\sum_{j=1}^T e_j\right)^2\right)$$

is the *variance of the sum of errors* and

$$\sigma_e^2 = \lim_{T \rightarrow \infty} T^{-1} \sum_{j=1}^T E(e_j^2)$$

is the *variance of errors*.

To derive the asymptotic distributions, Dickey and Fuller (1979, 1981) assumed that the errors e_t were $iid(0, \sigma^2)$. It is easy to verify that this is a sufficient condition for $\sigma^2 = \sigma_e^2$ (see section 3.2.2); it is not, however, a necessary condition. This equivalence also holds, for example, with errors that are *martingale difference sequences* under mild additional conditions. Therefore, the limiting distributions obtained by Dickey and Fuller are also valid in the presence of some heterogeneity in the error sequence provided the errors are martingale differences. In general, however, they cease to be appropriate when the errors are nonorthogonal (or serially correlated) and thus $\sigma^2 \neq \sigma_e^2$.

When the errors e_t are correlated, there is a need to either change the estimation method (adopt another regression model) or modify the statistics to obtain consistent estimators and statistics. Dickey and Fuller (1979) and Said and Dickey (1984) use the first approach of changing the estimating regressions using the parametric approach. Phillips (1987a) and Phillips and Perron (1988) follow the second approach of modifying the statistics using a nonparametric approach.

3.6.1 Changing the estimating equations: the ADF test

Consider the case in which the series of first-differences $\{\Delta y_t\}$ has a stationary $AR(p)$ representation with p known; i.e., $(1 - L)a(L)y_t = e_t$,

with a p th-order polynomial in L . Then, we can test the null hypothesis of a unit root by estimating an autoregression of Δy_t on its own lags and y_{t-1} using OLS

$$\Delta y_t = \alpha + \rho y_{t-1} + \sum_{j=1}^p \beta_j \Delta y_{t-j} + e_t \quad (3.15)$$

This is known as the *augmented* Dickey–Fuller (ADF) regression. If there is a unit root, under $H_0 : \rho = 1$, the t -statistic for the unit root null hypothesis follows the same DF distribution in the equation 3.12. The same principle holds for the F -statistics.

The reason behind this result is that in a regression of $I(1)$ variable on $I(1)$ and $I(0)$ variables, the asymptotic distribution of the coefficient of $I(1)$ and $I(0)$ variables are independent. To see this, consider the DGP

$$y_t = \rho y_{t-1} + e_t, \quad e_t \sim iid(0, 1)$$

Under the null of unit root $H_0 : \rho = 1$, y_t is $I(1)$ and $\Delta y_t = e_t$ is $I(0)$. Consider the estimating regression

$$y_t = \rho y_{t-1} + \beta \Delta y_{t-1} + e_t$$

We will show that the asymptotic distributions of $\hat{\rho}$ and $\hat{\beta}$ are independent. We have to normalize the coefficients and write

$$y_t = (T\rho) \left(\frac{y_{t-1}}{T} \right) + (\beta\sqrt{T}) \left(\frac{\Delta y_{t-1}}{\sqrt{T}} \right) + e_t$$

Then

$$\begin{bmatrix} T\hat{\rho} \\ \sqrt{T}\hat{\beta} \end{bmatrix} = \begin{bmatrix} \frac{1}{T^2} \sum y_{t-1}^2 & \frac{1}{T\sqrt{T}} \sum y_{t-1}e_{t-1} \\ \frac{1}{T\sqrt{T}} \sum y_{t-1}e_{t-1} & \frac{1}{T} \sum e_{t-1}^2 \end{bmatrix}^{-1} \begin{bmatrix} \cdot \\ \cdot \end{bmatrix}$$

We have

$$\begin{aligned} \frac{1}{T^2} \sum y_{t-1}^2 &\Rightarrow \int_0^1 W(r)dr \\ \frac{1}{T} \sum y_{t-1}e_{t-1} &\Rightarrow \int_0^1 W(r)dW(r) \\ \frac{1}{T} \sum e_{t-1}^2 &\rightarrow var(e_{t-1}) = 1 \end{aligned}$$

Hence

$$\frac{1}{T\sqrt{T}} \sum y_{t-1}e_{t-1} \Rightarrow \frac{1}{\sqrt{T}} \int_0^1 W(r)dW(r) \rightarrow 0$$

Thus the asymptotic distribution of $T\hat{\rho}$ and $\sqrt{T}\hat{\beta}$ are independent. This result can be used for showing that the distribution of $\hat{\rho}$ in the ADF regression is the Dickey–Fuller distribution. Also the asymptotic distribution of $\sqrt{T}(\hat{\beta} - \beta)$ is normal.

This fact has been extended by Said and Dickey (1984) to the more general case in which, under the null hypothesis, the series of first-differences are of the general ARMA(p, q) form with p and q unknown. They showed that a regression model, such as (3.15), is still valid for testing the unit root null under the presence of the serial correlations of errors, if the number of lags of Δy_t introduced as regressors increases with the sample size at a controlled rate $T^{\frac{1}{3}}$. Essentially the moving-average terms are being approximated by including enough autoregressive terms.

Consider the model with the serial correlation in errors described by

$$y_t = \rho y_{t-1} + z_t$$

and

$$z_t = \alpha z_{t-1} + e_t + \beta e_{t-1}$$

where it is assumed that $|\alpha| < 1$, $|\beta| < 1$, $y_0 = 0$ and $\{e_t\}$ is a sequence of *iid* random variables. If $|\rho| < 1$, then y_t is a stationary ARIMA(2,0,1). If $\rho = 1$ the series is an ARIMA(1,1,1) process. We consider $\rho = 1$ as a null hypothesis to be tested. Notice that

$$e_t = \sum_{j=0}^{\infty} (-\beta)^j (z_{t-j} - \alpha z_{t-j-1})$$

and it follows that

$$y_t - y_{t-1} = (\rho - 1)y_{t-1} + (\alpha + \beta)(z_{t-1} - \beta z_{t-2} + \beta^2 z_{t-3} - \cdots) + e_t$$

Under the null hypothesis that $\rho = 1$ we see that $z_t = y_t - y_{t-1}$. This motivates us to estimate the coefficients by regressing the first-difference Δy_t on $y_{t-1}, \Delta y_{t-1}, \dots, \Delta y_{t-k}$ where k is a suitably chosen integer. To get consistent estimates of the coefficients it is necessary to let k be a function T . We shall assume that $T^{1/3}k \rightarrow 0$ and that there exist $c > 0, r > 0$ such that $ck > T^{1/r}$. In this case, the limiting distribution of the t -statistics of the coefficient on the lagged dependent variable y_{t-1} has the same Dickey–Fuller t -distribution as when the errors are *iid*.

3.6.2 Choice of lag-length in the ADF test

It has been observed that the size and power properties of the ADF test are sensitive to the number of lagged terms (k) used. See, for instance, Schwert (1989) and Agiakoglou and Newbold (1992). Several guidelines have been suggested for the choice of k . Ng and Perron (1995) examine these in detail.

The guidelines are:

- (i) Rule for fixing k at an arbitrary level independent of T

Ng and Perron do a detailed simulation study of this rule. For a model with MA(1) parameter θ , they find that for $\theta = -0.8$, and sample size $T = 100$, the size is 0.283 instead of 0.05 when $k = 4$. The size is closer to the nominal size when k is increased to 10. However, the size increases with sample size for negative values of θ . For instance, with $\theta = -0.8$ and $k = 3$, the exact size increases from 0.455 to 0.598 as T increases from 100 to 500. The power, of course, rises with sample size. As for AR errors, the size is close to the nominal size if $k \geq$ the true lag order of the AR processes, but overparameterization results in loss of power. Power also rises with sample size, although size distortions are qualitatively the same. Overall, choosing a fixed k is not desirable.

- (ii) Rule for fixing k as a function of T

A rule commonly used is the one suggested by Schwert (1989) which is to choose

$$k = \text{Int}\{c(T/100)^{1/d}\}$$

Schwert suggest $c = 12$ and $d = 4$. The problem with such a rule is that it need not be optimal for all p and q in the ARMA(p, q).

- (iii) Information based rules

The information criteria suggest choosing k to minimize an objective function that trades off parsimony against reduction in sum of squares. The objective function is of the form

$$I_k = \log \hat{\sigma}_k^2 + k \frac{C_T}{T}$$

The Akaike information criterion (AIC) chooses $C_T = 2$. The Schwarz Bayesian information criterion (BIC) chooses $C_T = \log T$. These are the most commonly used information criteria in econometrics. (In chapter 8 we shall discuss the predictive information criterion (PIC) suggested by Peter Phillips.) Ng and Perron argue that both AIC and BIC are asymptotically the same with

ARMA (p, q) models and that both of them choose k proportional to $\log T$.

(iv) Sequential rules

Hall (1994) discusses two sequential rules in the context of pure autoregressions. The first rule, called *general to specific rule*, is to start with a large value of k (k_{max}), test the significance of the last coefficient and reduce k iteratively until a significant statistic is encountered. The other rule, called the *specific to general rule*, is to start with a small k and increase k successively until a nonsignificant coefficient is encountered. Hall showed that the specific to general approach is not generally asymptotically valid. He also found its performance to be inferior to that based on the general to specific approach in ARMA models.

Ng and Perron (1995) compare AIC, BIC, and Hall's general to specific approach through a Monte Carlo study using $T = 100$ and both AR and MA processes. The major conclusions are:

- (i) Both AIC and BIC choose very small values of k (e.g., $k = 3$). This results in high size distortions, especially with MA errors.
- (ii) Hall's criterion tends to choose higher values of k . The higher the k_{max} is, the higher is the chosen value of k . This results in the size being at the nominal level, but of course with a loss of power.

What this study suggests is that Hall's general to specific methods is preferable to the others. DeJong *et al.* (1992) show that increasing k typically results in a modest decrease in power but a substantial decrease in size distortions. If this is the case the information criteria are at a disadvantage because they result in a choice of very small values of k . (For some opposite evidence arguing in favor of BIC compared with Hall's method, see Stock (1994, p. 2781). However, he also says that one should use either the Schwarz criterion or Hall's sequential method.)

3.6.3 Modifications of the test statistic: Phillips–Perron test

As we have seen in section 3.2.2, when the errors are not *iid*, we have

$$T(\hat{\rho} - 1) \Rightarrow \left[\int_0^1 W(r)^2 dr \right]^{-1} \left[\frac{1}{2} [\chi_1^2 - 1] + \lambda \sigma^{-2} \right]$$

where

$$\lambda = \frac{\sigma^2 - \sigma_e^2}{2}$$

Hence we can consider the modified test statistics

$$Z_{\hat{\rho}} = T(\hat{\rho} - 1) - \lambda \left[T^{-2} \sum_{t=2}^T y_{t-1}^2 \right]^{-1}$$

Since

$$T^{-2} \sum y_{t-1}^2 \Rightarrow \sigma^2 \int_0^1 W(r)^2 dr$$

the adjusted test statistic has the Dickey–Fuller distribution.

The test statistics proposed by Phillips (1987a) and Phillips and Perron (1988) are these modified test statistics, using the estimators s_e^2 and s_{Tl}^2 , which are consistent estimators of σ_e^2 and σ^2 , respectively. The consistent estimator of σ_e^2 is given by

$$s_e^2 = T^{-1} \sum_{t=1}^T e_t^2.$$

Consistent estimation of $\sigma^2 = \lim_{T \rightarrow \infty} T^{-1} E((\sum_{j=1}^T e_j)^2)$ is more difficult. The problem is essentially equivalent to the consistent estimation of an asymptotic covariance matrix in the presence of weakly dependent and heterogeneously distributed observations. This has been examined by White and Domowitz (1984) and White (1984).

Under the stronger moment condition, Phillips (1987a) proved that

$$s_{Tl}^2 = T^{-1} \sum_{t=1}^T e_t^2 + 2T^{-1} \sum_{\tau=1}^l \sum_{t=\tau+1}^T e_t e_{t-\tau}$$

is a consistent estimator of σ^2 . However, s_{Tl}^2 is not constrained to be nonnegative. When there are large negative sample serial covariances, s_{Tl}^2 can take on negative values. Newey and West (1987) have suggested a modification to variance estimators such as $\tilde{s}_{Tl}^2$ which ensures that they are nonnegative. The modification suggested is

$$\tilde{s}_{Tl}^2 = T^{-1} \sum_{t=1}^T e_t^2 + 2T^{-1} \sum_{\tau=1}^l w_{\tau l} \sum_{t=\tau+1}^T e_t e_{t-\tau}$$

where

$$w_{\tau l} = 1 - \frac{\tau}{l+1}$$

They showed that $\tilde{s}_{Tl}^2$ is a consistent estimator of σ^2 under the same conditions as s_{Tl}^2 and that it is nonnegative by construction. Rather than

using the first-differences $e_t = y_t - y_{t-1}$ we may also use the residuals $\hat{e}_t = y_t - \hat{\rho}y_{t-1}$ from the least squares regression, since $\hat{\rho} \rightarrow 1$ as $T \rightarrow \infty$.

Inevitably the choice of l will be an empirical matter. If we allow the number of estimated autocovariances to increase as $T \rightarrow \infty$ but control the rate of increase so that $l = O(T^{1/4})$, then s_{Tl}^2 yields a consistent estimator of σ^2 . White and Domowitz (1984) and Perron (1988) provide some guidelines for the selection of l . A preliminary investigation of the sample autocorrelations of $e_t = y_t - y_{t-1}$ will help in selecting an appropriate choice of l . Since the sample autocorrelations of first-differenced economic time series usually decay quickly it is likely that in moderate sample sizes quite a small value of l will be chosen. Another way to approach the problem is to check the sensitivity of the results to various values of l .

Phillips (1987a) and Phillips and Perron (1988) proposed the nonparametric test statistics for the unit root null by using consistent estimates of variances as follows (for simplicity of notation, we shall denote s_{Tl}^2 by s^2):

(i) AR(1) without a drift

$$\begin{aligned} Z_\rho &= T(\hat{\rho} - 1) - \frac{1}{2} \frac{(s^2 - s_e^2)}{T^{-2} \sum_1^T y_{t-1}^2} \\ Z_t &= \frac{s_e}{s} t \hat{\rho} - \frac{1}{2} \frac{(s^2 - s_e^2)}{s(T^{-2} \sum_1^T y_{t-1}^2)^{1/2}} \end{aligned}$$

(ii) AR(1) with a drift

$$\begin{aligned} Z_\rho &= T(\hat{\rho} - 1) - \frac{1}{2} \frac{(s^2 - s_e^2)}{T^{-2} \sum_1^T (y_{t-1} - \bar{y}_{-1})^2} \\ Z_t &= \frac{s_e}{s} t \hat{\rho} - \frac{1}{2} \frac{(s^2 - s_e^2)}{s[T^{-2} \sum_1^T (y_{t-1} - \bar{y}_{-1})^2]^{1/2}} \end{aligned}$$

where $\bar{y}_{-1} = \sum_1^{T-1} y_t / (T - 1)$.

(iii) AR(1) with a drift and a linear trend

$$\begin{aligned} Z_\rho &= T(\hat{\rho} - 1) - \frac{T^6}{24D_X} (s^2 - s_e^2) \\ Z_t &= \frac{s_e}{s} t \hat{\rho} - \frac{T^3(s^2 - s_e^2)}{4\sqrt{3}D_X^{1/2}s} \end{aligned}$$

where $D_X = \det(X'X)$ and the regressors are $X = (1, t, y_{t-1})$.

The limiting distributions of Z_ρ and Z_t are identical to those of $K = T(\hat{\rho} - 1)$ and the t -statistics, respectively, when $\sigma^2 = \sigma_e^2$. Thus the asymptotic critical values of the tests are the same as the asymptotic critical values tabulated by Fuller (1976) which are given in table 3.1.

Note that Phillips uses α for ρ and hence these tests are commonly referred to as the Z_α and Z_t tests. To maintain consistency of notation with our previous discussion, we are defining them as the Z_ρ and Z_t tests. The procedure of correcting test statistics has been widely used by Phillips as well as several others in many other tests with nonstationary time series.

3.6.4 A comparison of the two approaches

If the moving-average terms are important, the number of extra lags of Δy_t needed as regressors in the autoregressive correction may be quite large. Furthermore, as shown by Schwert (1989), using Monte Carlo simulations, the exact size of the test may be far from the nominal size if the order of the autoregressive correction is not increased as the sample size increases. An adjustment is essential because the effect of the correlation structure of the residuals on the shape of the distribution becomes more precise. This point highlights the fact that with important MA components in the structure of the series $\{y_t\}$ a large number of nuisance parameters may be needed in the estimation. Since we furthermore lose one effective observation for each extra lag of Δy_t introduced, this implies that the Said and Dickey approach may have substantially lower power when MA terms are more important than if the errors were *iid*.

On the other hand, several simulation studies have shown that Phillips and Perron's nonparametric tests have serious size distortions in finite samples when the data generating process has a predominance of negative autocorrelations in first-differences (Schwert, 1989; Phillips and Perron, 1988; DeJong *et al.*, 1992). This has often been taken to suggest that the Phillips and Perron's nonparametric tests may be less reliable than the ADF tests when there is a predominance of negative autocorrelations in first-differences. However, it should be noted that Perron and Ng (1996) suggested useful modifications of the PP tests, that solve this problem. These are discussed in section 4.3 of chapter 4.

3.7 Sargan–Bhargava and Bhargava tests

The preceding tests for unit roots are in the framework of the DF tests. We shall discuss here some alternative approaches. Sargan and Bhargava (1983) suggest tests in the Durbin–Watson framework. They generalize the Durbin–Watson statistic (used for testing serial correlation) to the problem of testing the hypothesis that the residuals from the least squares regression follow a random walk. They also discuss the Berenblutt–Webb statistic (which they argue is the preferred alternative) and suggest computing the significance points using the Imhof routine. This statistic is given by

$$g = \frac{\hat{e}'\hat{e}}{\tilde{u}'\tilde{u}}$$

where $\hat{e}$ and $\tilde{u}$ are respectively the least squares residuals from the first-difference equation and the levels equation respectively.

Bhargava uses the framework we discussed earlier in section 2.9 of chapter 2, which, as argued there, is a better approach to unit root testing than the Dickey–Fuller approach. He suggests some tests for unit roots along the lines of the tests developed in Sargan and Bhargava. He suggests four test statistics R_1 , R_2 , N_1 , and N_2 , but since the last two are slight modification of the R_1 and R_2 we shall discuss only these here. The statistics are defined as

$$R_1 = \frac{\sum_{t=2}^T (y_t - y_{t-1})^2}{\sum_{t=1}^T (y_t - \bar{y})^2}$$

where $\bar{y} = \sum_{i=1}^T y_i / T$. The simple random walk hypothesis is rejected for large values of R_1 . This statistic is the modified DW statistic in Sargan and Bhargava.

One statistic that is related to R_1 is the statistic MSB (discussed in Stock, 1990) defined as

$$\text{MSB} = T^{-2} \frac{\sum y_{t-1}^2}{s^2}$$

where s^2 is an estimate of the variance of y_t . For s^2 Bhargava used $T^{-1} \sum_{t=1}^T (y_t - \bar{y})^2$. But it is customary to use the Newey–West estimator for s^2 (discussed earlier in the context of Phillips–Perron tests). A better estimator suggested by Perron and Ng is discussed in section 4.3 of the next chapter. Perron and Ng used this in their study of the MSB statistic.

Since the sum of squares of an $I(1)$ variables is $O_p(T^2)$ but that of an

I(0) series is $O_p(T)$ the MSB statistic can be used to test for a unit root. For an I(0) series MSB is related to the PP tests by the relation

$$Z_t = \text{MSB} \cdot Z_\rho$$

Note that MSB tends to zero. The unit root hypothesis is rejected in favor of stationarity when MSB is smaller than some appropriate critical value. Critical values with y_t demeaned and detrended are given in Stock (1990).

The second test statistic R_2 suggested by Bhargava, is defined as

$$R_2 = \frac{A}{B}$$

where

$$A = \sum_{t=2}^T (y_t - y_{t-1})^2 - \frac{(y_T - y_1)^2}{T-1}$$

$$B = \frac{1}{(T-1)^2} \sum_{t=1}^T [(T-1)y_t - (t-1)y_T - (T-t)y_1 - (T-1)(\bar{y} - (y_1 + y_T)/2)]^2$$

Schmidt and Phillips (1992) derive Bhargava's R_2 test statistic by considering an LM (Lagrangian Multiplier) test or score test for ρ in the model

$$y_t = \gamma_0 + \gamma_1 t + u_t \quad (3.16)$$

$$u_t = \rho u_{t-1} + e_t$$

As noted in section 2.9 of chapter 2, this model is equivalent to

$$y_t = \beta_0 + \beta_1 t + \rho y_{t-1} + e_t \quad (3.17)$$

where

$$\beta_0 = \gamma_0(1 - \rho) + \gamma_1 \rho$$

and

$$\beta_1 = \gamma_1(1 - \rho)$$

This equation can also be written as

$$\Delta y_t = \beta_0 + \beta_1 t + (\rho - 1)y_{t-1} + e_t$$

or

$$\Delta y_t = \beta_0 + \beta_1 t + \phi y_{t-1} + e_t \quad (3.18)$$

Let $\hat{S}_{t-1}$ be the residual from a regression of y_{t-1} on an intercept and t . Then the estimator of ϕ from equation (3.18) is the same as an estimate of ϕ from the equation

$$\Delta y_t = \text{intercept} + \phi \hat{S}_{t-1} + e_t, \quad t = 1, 2, \dots, T \quad (3.19)$$

If the least squares estimator of ϕ from the equation (3.19) is $\hat{\phi}$, then the usual coefficient test based on $T\hat{\phi}$ and the corresponding t -statistic have the DF distribution (3.11) and (3.12) under the hypothesis $\phi = 0$.

To derive the LM statistic, Schmidt and Phillips (referred to as SP hereafter) assume that the e_t are $IN(0, \sigma^2)$ in (3.16). Under the hypothesis $\rho = 1$ we can write

$$\begin{aligned} y_t &= (\gamma_0 + u_0) + \gamma_1 t + e_1 + \dots + e_t \\ &= \delta + \gamma_1 t + e_1 + \dots + e_t \end{aligned}$$

where $\delta = \gamma_0 + u_0$. And

$$\Delta y_t = \gamma_1 + e_t$$

Hence the estimator of γ_1 which we shall refer to as $\tilde{\gamma}_1$ is given by

$$\tilde{\gamma}_1 = \text{mean of } \Delta y_t = \frac{y_T - y_1}{T - 1}$$

Also

$$\tilde{\delta} = y_1 - \tilde{\gamma}_1$$

Let

$$\tilde{S}_t = y_t - \tilde{\delta} - \tilde{\gamma}_1 t, \quad t = 1, 2, \dots, T$$

These are the residuals using estimates of the parameters in (3.16) under the hypothesis $\rho = 1$.

Note that

$$\begin{aligned} \tilde{S}_t &= y_t - y_1 - (t - 1)\overline{\Delta y_t} \\ &= \frac{1}{T - 1} [(T - 1)y_t - (t - 1)y_T - (T - t)y_1] \end{aligned}$$

Hence $S_0 = S_T = 0$ and

$$\bar{S} = \frac{1}{T} \sum_{t=1}^T \tilde{S}_t = \bar{y} - \frac{y_1 + y_T}{2}$$

Table 3.3. *Critical values for the Schmidt–Phillips LM test*

Sample size	coeff. test		t-test	
	1%	5%	1%	5%
25	−20.4	−15.7	−3.90	−3.18
50	−22.8	−17.0	−3.73	−3.11
100	−23.8	−17.5	−3.63	−3.06
500	−25.3	−18.1	−3.59	−3.04

Source: Schmidt and Phillips (1992).

From these results SP note that the denominator B in Bhargava's R_2 statistic mentioned earlier can be written as $\sum(\tilde{S}_t - \bar{S})^2$.

SP show that (we shall omit the details) the LM test statistic is obtained from the estimator $\tilde{\phi}$ of ϕ in the equation

$$\Delta y_t = \text{intercept} + \phi \tilde{S}_{t-1} + \text{error} \quad (3.20)$$

Then

$$\tilde{\rho} = T\tilde{\phi}$$

is the statistic for the coefficient test and $\tilde{\tau}$ the statistic for the corresponding t -test of the hypothesis $\phi = 0$ in (3.20). SP note that the difference between the LM test statistics and the DF test statistics is that the DF tests use $\hat{S}_{t-1}$ and the LM tests use $\tilde{S}_{t-1}$ as a regressor in the same equation (see equations 3.19 and 3.20). Since $\hat{S}_t$ is obtained from a regression of y_t on a time trend, if y_t is $I(1)$, this is a spurious regression and SP argue that for this reason the LM tests can be expected to be more powerful than the DF tests. This they confirm by a Monte Carlo study. Except when $u_0^* = u_0/\sigma_e$ is large in absolute value, their experiments reveal that the tests based on $\tilde{\tau}$ and $\tilde{\rho}$ are more powerful than the DF tests.

SP derive the asymptotic distributions of $\tilde{\rho}$ and $\tilde{\tau}$. The finite sample distributions are complicated but they tabulate the percentiles by simulation. They are given in table 3.3. The lower tail critical values are smaller than the corresponding lower tail values of the DF tests (tabulated in table 3.1). SP also provide tables of critical values for the $\tilde{\tau}$ and $\tilde{\rho}$ tests under polynomial trends of order 2, 3, and 4 (table 1B of the SP paper). For the case where e_t are not *iid*, they suggest applying the same type of corrections to the LM statistics as those we discussed earlier for the PP tests.

Finally SP show that $\tilde{\rho}$ is related approximately to Bhargava's R_2 test statistic mentioned earlier by the relation

$$\tilde{\rho} = -\frac{T}{2}R_2$$

Hence in subsequent literature the LM tests are often referred to BSP (Bhargava–Schmidt–Phillips) tests. Bhargava derived his tests as best invariant tests and this guarantees invariance with respect to γ_0, γ_1 , and σ_e in (3.16). Since $\tilde{S}_t$ depends on only e_t , the LM test is also invariant to γ_0, γ_1 , and σ_e . Under the alternative hypothesis the distribution of $\tilde{\tau}$ and $\tilde{\rho}$ is still invariant to γ_0, γ_1 , and σ_e but depends on $u_0^* = u_0/\sigma_e$.

3.8 Variance ratio tests

The measurement of the degree of persistence in a time series is another way of evaluating the presence of a unit root. Cochrane (1988) proposes a variance ratio (VR) statistic for this purpose. The VR is the variance of the k th difference of a time series divided by k -times the variance of its first-difference. Thus

$$VR_k = \frac{V_k}{V_1}$$

where $V_k = \text{var}(y_t - y_{t-k})/k$. It is straightforward to show that VR_k can be written as the weighted sum of the correlations r_j between Δy_t and Δy_{t-j} .

$$VR_k = 1 + 2 \sum_{j=1}^k \left[1 - \frac{j}{k+1} \right] r_j$$

Cochrane uses the following estimator for the variance ratio

$$V\hat{R}_k = \frac{k^{-1}\text{var}(y_t - y_{t-k})}{\text{var}(y_t - y_{t-1})} \left(\frac{T}{T-k+1} \right)$$

The second term is a bias correction factor. Anderson (1971) showed that as $T \rightarrow \infty$, $k \rightarrow \infty$, and $k/T \rightarrow 0$, this estimator has a limiting normal distribution with mean VR_k and variance $4kVR_k^2/3T$.

Lo and MacKinlay (1988) derived the asymptotic distribution of $V\hat{R}_k$ for the case of a random walk null and normal *iid* innovations under the assumption that k is fixed and T goes to infinity. They show that the distribution is normal with mean 1 and variance $2(2k-1)(k-1)/3kT$. Note that when the null model is a random walk, VR_k is 1 and with k large, this expression agrees with the one derived by Anderson (1971).

There are several values of k for which the variance ratio test can be used. In practice it is customary to consider VR_k for different values of k and consider a model rejected when at least some of the VR statistics provide evidence against it. This method of testing penalizes the null hypothesis and the correct critical values of the tests are not known because the tests are correlated.

Cecchetti and Lam (1994) examine (i) the accuracy of the asymptotic distributions of the VR test statistics in small samples and (ii) the consequences of the use of multiple tests for different values of k . They argue that there are substantial size distortions with the use of the asymptotic approximations and more size distortions with the use of the sequential testing procedure with different values of k . They suggest using a joint test based on all the VR_k statistics and obtaining the critical values of these test statistics using Monte Carlo methods.

3.9 Tests for TSP versus DSP

Ever since the publication of the paper by Nelson and Plosser (1982), macroeconomists have been interested in unit roots in time series. Nelson and Plosser found that they could not reject the null of an autoregressive unit root in 13 out of 14 US macroeconomic time series, in some cases spanning over 100 years. The existence of a unit root was interpreted as having important implications for the theory of business cycles and the persistence of the effect of real shocks to the economy. Though some economists like Cochrane (1991b) argued that the evidence on unit roots is empirically ambiguous and also irrelevant to the question of the persistence of the effect of real shocks, the literature on unit roots kept proliferating and more and more tests for unit roots have been appearing.

Nelson and Plosser started with the problem of discrimination between the trend-stationary process (TSP) and the difference-stationary process (DSP) models

$$\text{TSP: } y_t = \alpha + \delta t + u_t$$

$$\text{DSP: } \Delta y_t = \alpha + u_t$$

where u_t is stationary. This is a model selection problem. They, however, approach it as a hypothesis testing problem, a test for a nested hypothesis.

To test the hypothesis that a time series belongs to the DSP class

against the alternative hypothesis that it belongs to the TSP class, Nelson and Plosser employed the unit root tests described in the previous section. They started with the TSP model and the assumption of the first-order serial correlation in errors

$$y_t = \beta_0 + \beta_1 t + u_t$$

$$u_t = \rho u_{t-1} + e_t$$

This formulation for unit root testing has been advocated by Bhargava (1986). This gives

$$y_t = \beta_0 + \beta_1 t + \rho [y_{t-1} - \beta_0 - \beta_1(t-1)] + e_t$$

Thus the reduced form for this model can be written as

$$y_t = \alpha + \delta t + \rho y_{t-1} + e_t$$

where e_t is assumed to be $iin(0, \sigma^2)$ and $\delta = 0$ if $\rho = 1$. They test the unit root null hypothesis $H_0 : \rho = 1$ and $\delta = 0$. (Recall that the value of constant α does not affect the asymptotic distribution of the OLS estimates $\hat{\rho}$ and $\hat{\delta}$ when the estimating regression includes t as an independent variable (see section 3.4). If the null hypothesis is rejected, then y_t belongs to the TSP class. If the unit root null cannot be rejected, then it belongs to the DSP class. After observing the sample autocorrelations of the first-differences of US historical data, Nelson and Plosser assume moving-average errors (the presence of the serial correlation in errors) and incorporate the additional regressors Δy_{t-k} into the model to correct the serial correlations in the errors. In other words they employed the ADF approach.

Nelson and Plosser analyzed a set of 14 US macroeconomic time series. They used the logarithm of the series, except for the interest rate series, instead of level based on the fact that (Nelson and Plosser, 1982 p.141):

The tendency of economic time series to exhibit variation that increases in mean and dispersion in proportion to absolute level, motivates the transformation to natural logs and the assumption that trends are linear in the transformed data.

Table 3.3 shows part of their results. They found that all 14 series, except the unemployment rate, were characterized by the DSP class.

It is often argued that the way we estimate the equation, whether in levels or in first-differences, depends on whether the series is TSP or DSP respectively. If it is TSP and we use first-differences, then we have *overdifferencing*. If it is DSP and we estimate the equation in levels,

Table 3.4. *Nelson and Plosser's results*

Series	Sample size (T)	Lag length (k)	$t_{\hat{\rho}}$
Real GNP	62	2	-2.99
Nominal GNP	62	2	-2.32
Real per capita GNP	62	2	-3.04
Industrial production	111	6	-2.53
Employment	81	3	-2.66
Unemployment rate	81	4	-3.55*
GNP deflator	82	2	-2.52
Consumer prices	111	4	-1.97
Wages	71	3	-2.09
Real wages	71	2	-3.04
Money stock	82	2	-3.08
Velocity	102	1	-1.66
Interest rate	71	3	0.686
Common stock prices	100	3	-2.05

Note: The values of $t_{\hat{\rho}}$ are smaller than the critical value -3.45 at the 5% significance level.

Source: Nelson and Plosser (1982, table 5, p. 151).

then we have *underdifferencing*. There has been some debate in the literature on the overdifferencing versus underdifferencing issue, arguing that the former is a less serious error than the latter. However, it is more important to take the serial correlation structure in both the models into account, and when this is done the issue of over- versus underdifferencing becomes a nonissue. See McCallum (1993), who suggests estimating the equation in both levels and first-differences and choosing the one that requires the smaller amount of correction to remove autocorrelation of the residuals, and illustrates this with two samples.

3.10 Forecasting from TS versus DS models

One of the main issues that has been discussed in the literature on unit roots is the use of unit root tests in deciding whether to use the trend-stationary (TS) model or the difference-stationary (DS) model for forecasting. Campbell and Perron (1991) argue that unit root test procedures can be practically useful for forecasting even in small samples where they have only a limited ability to distinguish between TS and DS alternatives. They provide evidence for this through a Monte Carlo study. They generated 5,000 samples of length 100 from the ARMA(1,1)

process

$$x_t = \phi x_{t-1} + u_t + \theta u_{t-1}$$

For $\phi = 1.0$ they consider the cases $\theta = (0.98, 0.95, 0.90, 0.80, 0.50)$. For $\theta = 0$ they consider the cases $\phi = (1.0, 0.98, 0.95, 0.90, 0.80, 0.50)$. For each case they consider one-period ahead and 20-period ahead forecasts from an autoregressive model in levels and an autoregressive model in differences. In the former case a linear trend is estimated whereas in the latter case the mean of the differenced data is estimated. For each model a lag length k is chosen with the maximum lag fixed at $k_{max} = 6$ (The optimal k is decided by starting with $k = 6$ and working down until the last coefficient in the $AR(k)$ model is significant.) For each sample and forecast horizon, out of sample mean squared errors of forecasts were calculated using 25 draws of the DGP.

Campbell and Perron find that stationary forecasting models are superior for all processes with $\phi = 1$ and $\theta \leq -0.9$, while unit root forecasting models are superior for all processes with $\theta = 0$ and $\phi > 0.9$ (one-period ahead forecast) or 0.95 (20-period ahead forecast). They also consider a mixed strategy which is a pre-testing method: use the levels model if the Said–Dickey and Phillips–Perron tests reject the unit root at the 5 percent level and use the difference model if it does not. The performance of this mixed strategy was found to be closer to that of the better of the two individual strategies.

In his comment on the Campbell and Perron paper, Cochrane (1991a) argues that the Monte Carlo study does not throw any light on the unit root question. He presents his own Monte Carlo study which shows that for the ARMA(1,1) model with $\phi = 0.95$ and $\theta = 0.5$, the AR(1) in differences provides better one-period ahead forecasts whereas the AR(1) in levels provides the better forecasts for 20- and 50-period horizons. Over the long-run stationarity asserts itself. He used only one lag to separate the issues of choice of lag length from the unit root question.

However, the forecasting issue is separate from that of whether the unit root tests can distinguish between the TS and DS processes. The main question is: for most macroeconomic time series, is it better to forecast using the trend-stationary (TS) model or the difference-stationary (DS) model? One common presumption has been that it is better to use the DS model because it is better in characterizing the uncertainty associated with forecasting these series over long horizons (see Dickey, Bell, and Miller (1986) and Phillips (1991) for this view). On the other hand, Meese and Geweke (1984) studied the accuracy of forecasts gen-

erated by a wide range of autocorrelated procedures for 150 quarterly and monthly macroeconomic time series, and found that forecasts based on detrended data were, on average, more accurate than those based on differenced data.

Interval forecasts

Meese and Geweke (1984) considered only point forecasts but not the standard errors of the forecasts. Sampson (1991), for instance, shows that without parameter uncertainty, forecast variance for the unit root model grows with the forecast horizon, whereas for the trend-stationary model, it tends to a constant. On the other hand, with the estimated parameters, forecast variance grows with the square of the variance for both models. However, the leading terms in the expressions for the forecast variances, as we shall presently see, are roughly proportional to $12n^2/N^3$ for the trend-stationary model and $4n^2/N$ for the unit root model, where n is the forecast horizon and N is the length of the time series on which the estimation is based. Thus the leading terms are in the ratio $1/N^2$ and hence the forecast variances are much higher for the unit root model than the trend-stationary model. In Sampson's empirical example $N = 41$ and $n = 60$. This is too long a forecast horizon to be of any practical interest. Usually the forecast horizon beyond which predictability ceases occurs at relatively small values such as $n = 8 \sim 10$ periods.

Sampson considered a general stationary process for the errors but to simplify matters we shall consider *iid* errors. Consider the two models

$$\begin{aligned} \text{TS} \quad y_t &= \mu + \beta t + u_t, \quad u_t \sim in(0, \sigma_u^2) \\ \text{DS} \quad \Delta y_t &= \mu + v_t, \quad v_t \sim in(0, \sigma_v^2) \end{aligned}$$

For the TS model we have

$$\begin{bmatrix} N^{1/2}(\hat{\mu} - \mu) \\ N^{3/2}(\hat{\beta} - \beta) \end{bmatrix} \sim N \left[0, \sigma_u^2 \begin{pmatrix} 4 & -6 \\ -6 & 12 \end{pmatrix} \right]$$

The rate of convergence of $\hat{\beta}$ to β is faster in the TS model than in the unit root model. The n -period ahead forecast error is given by

$$(\mu - \hat{\mu}) + (N + n)(\hat{\beta} - \beta) + u_{N+n}$$

The variance of this is

$$\sigma_u^2 \left[1 + \frac{4}{N} - \frac{12(N+n)}{N^2} + \frac{12(N+n)^2}{N^3} \right]$$

For the unit root model, the variance is a complicated expression.

3.11 Summary and conclusions

This chapter discusses the basic unit root tests: the Dickey–Fuller (DF), the augmented Dickey–Fuller (ADF) and the Phillips–Perron (PP) tests. Some other tests are also discussed: Sargan–Bhargava tests and variance ratio tests.

Although often used, the DF, ADF, and PP tests lack power against meaningful alternatives and should not be used any more. In the next chapter we shall discuss useful modifications of these tests. As for variance ratio tests, which are also commonly used, using the asymptotic distributions results in substantial size distortions in small samples, it is best to obtain small sample critical values using Monte Carlo methods. If more than one variance ratio test statistic is used, it is best to apply a joint test with critical values obtained by Monte Carlo methods.

We discussed the Sargan–Bhargava tests here because they form the basis of other tests discussed in subsequent chapters. Finally we also review the controversy surrounding the issue of forecasting from detrended versus differenced data.

References

- Agiakoglou, C. and P. Newbold (1992), “Empirical Evidence on Dickey–Fuller type tests,” *Journal of Time Series Analysis*, 13, 471–483.
- Anderson, R.L. (1942), “Distribution of the Serial Correlation Coefficient,” *Annals of Mathematical Statistics*, 13, 1–13.
- Anderson, T.W. (1971), *The Statistical Analysis of Time Series*, Wiley, New York.
- Ansley, C.F. (1979), “An Algorithm for the Exact Likelihood of Mixed Autoregressive Moving Average Process,” *Biometrika*, 66, 59–65.
- Bhargava, A. (1986), “On the Theory of Testing for Unit Roots in Observed Time Series,” *Review of Economic Studies*, 53, 369–384.
- Billingsley, P. (1968), *Convergence of Probability Measures*, John Wiley, New York.
- Box, G.E.P. and G.M. Jenkins (1970), *Time Series Analysis Forecasting and Control*, Holden–Day, San Francisco.
- Box, G.E.P. and D.A. Pierce (1970), “Distribution of Residual Autocorrelations in Autoregressive Integrated Moving Average Time

- Series Models," *Journal of the American Statistical Association*, 65, 1509–1526.
- Campbell, J.Y. and P. Perron (1991), "Pitfalls and Opportunities: What Macroeconomists Should Know about Unit Roots," in O.J. Blanchard and S. Fisher (eds.), *NBER Macroeconomics Annual*, The MIT Press, 141–201.
- Cecchetti, S.G. and P.S. Lam (1994), "Variance-Ratio Tests: Small-Sample Properties with an Application to International Output Data," *Journal of Business and Economic Statistics*, 12, 177–186.
- Chan, N.H. and C.Z. Wei (1987), "Asymptotic Inference for Nearly Nonstationary AR(1) Processes," *Annals of Statistics*, 15, 1050–1063.
- Cochrane, J.H. (1988), "How Big is the Random Walk in GNP?" *Journal of Political Economy*, 96, 893–920.
- (1991a), "Comment on Campbell and Perron," in O.J. Blanchard and S. Fisher (eds.), *NBER Macroeconomics Annual*, The MIT Press, 201–210.
- (1991b), "A Critique of the Application of Unit Root Tests," *Journal of Economic Dynamics and Control*, 15, 275–284.
- Davidson, J.E.H. (1981), "Problems with the estimation of moving-average processes," *Journal of Econometrics*, 19, 295–310.
- DeJong, D.N., J.C. Nankervis, N.E. Savin, and C.H. Whiteman (1992), "The Power Problems of Unit Root Tests for Time Series with Autoregressive Errors," *Journal of Econometrics*, 53, 323–343.
- Dickey, D.A. and W.A. Fuller (1979), "Distribution of the estimators for autoregressive time series with a unit root," *Journal of the American Statistical Association*, 74, 427–431.
- (1981), "Likelihood Ratio Statistics for Autoregressive Time Series With a Unit Root," *Econometrica*, 49, 1057–1072.
- Dickey, D.A., W.R. Bell, and R.B. Miller (1986), "Unit Roots in Time Series Models: Tests and Implications," *American Statistician*, 40, 12–26.
- Donsker, M.D. (1951), "An Invariance Principle for Certain Probability Limit Theorems," *Memoirs of the American Mathematical Society*, 6, 1–12.
- Evans, G.B.A. and N. E. Savin (1981), "Testing for Unit Roots: 1," *Econometrica*, 49, 753–779.
- (1984), "Testing for Unit Roots: 2," *Econometrica*, 52, 1241–1269.

- Friedman M. and A.J. Schwartz (1963), *A Monetary History of the United States 1867-1960*, Princeton University Press, Princeton, NJ.
- Fuller, W. A. (1976), *Introduction to Statistical Time Series*, New York, Wiley.
- (1996), *Introduction to Statistical Time Series*, 2nd edn, New York, Wiley.
- Gould J.P. and C.R. Nelson(1974), "The Stochastic Structure of the Velocity of Money," *American Economic Review*, 64, 405-417.
- Hall, A. (1992), "Testing for a Unit Root in Time Series Using Instrumental Variable Estimation with pretest Data-Based Model Selection," *Journal of Econometrics*, 54, 223-250.
- (1994), "Testing for a Unit Root in Time Series with pretest Data-Based Model Selection," *Journal of Business and Economic Statistics*, 12, 461-470.
- Hall, P. and C.C. Heyde (1980), *Martingale Limit Theory and Its Applications*, Academic Press, New York.
- Hamilton, J.D.(1989), "A New Approach to the Economic Analysis of Nonstationary Time Series and the Business Cycle," *Econometrica*, 57, 357-384.
- Herrndorf, N. (1984), "A Functional Central Limit Theorem for Weakly Dependent Sequences of Random Variables," *Annals of Probability*, 12, 141-153.
- Imhof, J. P. (1961), "Computing the Distribution of Quadratic Forms in Normal Variables," *Biometrika*, 48, 419-426.
- Lo, A. W. and A.C. MacKinlay (1988), "Stock Market Prices Do Not Follow Random Walks: Evidence From Simple Specification Tests," *Review of Financial Studies*, 1, 41-66.
- Maddala, G.S. (1992), *Introduction to Econometrics*, 2nd edn, Macmillan, New York.
- Mann, H.B. and A. Wald (1943), "On Stochastic Limit and Order Relationships," *Annals of Mathematical Statistics*, 14, 217-277.
- McCallum, B.T. (1993), "Unit Roots in Macroeconomic Time Series: Some Critical Issues," *Economic Quarterly*, Federal Reserve Bank of Richmond, 79, 13-43.
- McLeish, D.L. (1975), "Invariance Principles for Dependent Variables," *Zeitschrift für Wahrscheinlichkeitstheorie und Verwandte Gebiete*, 32, 165-178.
- Meese, R. and J. Geweke (1984), "A Comparison of Autoregressive

- Univariate Forecasting Procedures for Macroeconomic Time Series," *Journal of Business and Economic Statistics*, 2, 191–200.
- Nabeya, S. and K. Tanaka (1990), "A General Approach to the Limiting Distribution for Estimators in Time Series Regression With Nonstable Autoregressive Errors," *Econometrica*, 58, 145–163.
- Nankervis, J.C. and N.E. Savin (1985), "Testing the Autoregressive Parameter With the t -statistic," *Journal of Econometrics*, 27, 143–161.
- Nelson, C. R. and C. I. Plosser (1982), "Trends and Random Walks in Macroeconomic Time Series," *Journal of Monetary Economics*, 10, 139–162.
- Newey, W.K. and K.D. West (1987), "A Simple Positive Semi-Definite Heteroskedasticity and Autocorrelation-Consistent Covariance Matrix," *Econometrica*, 55, 703–708.
- Ng, S. and P. Perron (1995), "Unit Root Tests in ARMA Models With Data-Dependent Methods for the Selection of the Truncation Lag," *Journal of American Statistical Association*, 90, 268–281.
- Pan, J. J. (1968), "Distribution of the non-circular serial correlation coefficient," American Mathematical Society and Institute of Mathematical Statistics, *Selected Translations in Probability and Statistics*, 7, 281–292.
- Perron, P. (1988), "Trends and Random Walks in Macroeconomic Time Series: Further Evidence from a New Approach," *Journal of Economic Dynamics and Control*, 12, 297–332.
- (1989), "The Great Crash, the Oil Price Shock, and the Unit Root Hypothesis," *Econometrica*, 57, 1361–1401.
- (1991), "A Continuous Time Approximation to the Stationary First-Order Autoregressive Model," *Econometric Theory*, 7, 236–252.
- Perron, P. and S. Ng (1996), "Useful Modifications to Some Unit Root Tests with Dependent Errors and Their Local Asymptotic Properties," *Review of Economic Studies*, 63, 435–465.
- Phillips, P. C. B. (1977), "Approximations to Some Finite Sample Distributions Associated with a First-order Stochastic Difference Equation," *Econometrica*, 45, 463–485.
- (1987a), "Time Series Regression with a Unit Root," *Econometrica*, 55, 277–301.
- (1987b), "Towards a Unified Asymptotic Theory for Autoregression," *Biometrika*, 74, 535–547.
- (1988a), "Regression Theory for Near-Integrated Time Series," *Econometrica*, 56, 1021–1043.

- (1988b), "Weak Convergence to the Matrix Stochastic Integral $\int_0^1 BdB'$," *Journal of Multivariate Analysis*, 24, 252–264.
- (1991), "To Criticize the Critics: An Objective Bayesian Analysis of Stochastic Trends," *Journal of Applied Econometrics*, 6, 333–364.
- (1992), "Bayesian Model Selection and Prediction with Empirical Applications," Cowles Foundation Discussion Paper No. 1023.
- (1994), "Bayes Models and Forecasts of Australian Macroeconomic Time Series," in C. Hargreaves (ed.), *Nonstationary Time Series Analysis and Cointegration*, Oxford University Press, Oxford, chapter 3.
- Phillips, P.C.B. and P. Perron (1988), "Testing for a Unit Root in Time Series Regression," *Biometrika*, 75, 335–346.
- Quenouille, M. (1957), *The Analysis of Multiple Time Series*, Charles Griffin, London.
- Rao, M.M. (1961), "Consistency and Limit Distributions of Estimators of Parameters in Explosive Stochastic Difference Equations," *Annals of Mathematical Statistics*, 32, 195–218.
- Rubin, H. (1950), "Consistency of maximum likelihood estimates in the explosive case," in T. C. Koopmans (ed.) *Statistical Inference in Dynamic Economic Models*, Wiley, New York, 356–364.
- Said, S.E. and D.A. Dickey (1984), "Testing for Unit Roots in Autoregressive-Moving Average Models of Unknown Order," *Biometrika*, 71, 599–607.
- Sampson, M. (1991), "The Effect of Parameter Uncertainty on Forecast Variances and Confidence Intervals for Unit Root and Trend Stationary Time-Series Models," *Journal of Applied Econometrics*, 6, 67–76.
- Sargan, J.D. and A. Bhargava (1983), "Testing Residuals from Least Squares Regression for Being Generated by the Gaussian Random Walk," *Econometrica*, 51, 153–174.
- Satchell, S. E. (1984), "Approximation to the Finite Sample Distribution for Nonstable First Order Stochastic Difference Equations," *Econometrica*, 52, 1271–1289.
- Schmidt, P. and P.C.B. Phillips (1992), "LM Tests for a Unit Root in the Presence of Deterministic Trends," *Oxford Bulletin of Economics and Statistics*, 54, 257–287.
- Schwert, G.W. (1989), "Tests for Unit Roots: A Monte Carlo Investigation," *Journal of Business and Economic Statistics*, 7, 147–159.
- Stock, J.H. (1990), "A Class of Tests for Integration and Cointegration," manuscript, Harvard University.

- (1994), "Unit Roots, Structural Breaks, and Trends," in R.F. Engle and D.L. McFadden (eds.), *Handbook of Econometrics*, vol. IV, North-Holland, Amsterdam.
- Tanaka, K. (1996), *Time Series Analysis: Non-Stationary and Non-Invertible Distribution Theory*, John Wiley, New York.
- West, K.D. (1988), "Asymptotic Normality, When Regressors Have a Unit Root," *Econometrica*, 56, 1397–1417.
- White, H. (1984), *Asymptotic Theory for Econometricians*, Academic Press, New York.
- White, H. and I. Domowitz (1984), "Nonlinear Regression with Dependent Observations," *Econometrica*, 52, 143–162.
- White, J.S. (1958), "The Limiting Distribution of the Serial Correlation Coefficient in the Explosive Case," *Annals of Mathematical Statistics*, 29, 1188–1197.
- Zivot, E. and D.W.K. Andrews (1992), "Further Evidence on the Great Crash, the Oil Price Shock, and the Unit Root Hypothesis," *Journal of Business and Economic Statistics*, 10, 251–270.

4

Issues in unit root testing

4.1 Introduction

In the preceding chapter we introduced some tests (ADF and Phillips–Perron tests) routinely used in testing for unit roots. There are, however, several problems with these tests – particularly their low power. In this chapter we discuss these problems and some solutions and alternatives that have been suggested, but are not wide spread in practice. The main issues that need to be discussed are:

- (i) The low powers of unit root tests: what can be done to improve this?
- (ii) The problem of overdifferencing and moving-average (MA) unit roots.
- (iii) Use of instrumental variable (IV) methods (IV tests).
- (iv) Using stationarity as null – rather than unit root as null.
- (v) Devising more powerful tests.
- (vi) Does frequency of observation matter (monthly versus quarterly data, quarterly versus annual data, and so on)?
- (vii) Can we increase the sample size and power by using panel data (panel data unit root tests)?
- (viii) What are the consequences of uncertainty about unit roots?
- (ix) What are the appropriate significance levels to use on unit root tests? Isn't a unit root test a pre-test?
- (x) Is $I(1)$ the only type of nonstationarity? What about tests for $I(0)$ versus $I(d)$ with $d > 0$ but $d \neq 1$.
- (xi) What about a multivariate approach to unit root testing instead of the often used univariate approach?

These are the problems we shall be discussing in this chapter.

These issues will not be discussed in sequence but the following sections address these issues. The sections are organized around the theme of the poor power of unit root tests, and what to do about the problem.

We shall discuss a large number of unit root tests just so that we are not guilty of omitting them, although we do not believe that many of them are useful. The large number of unit root tests is a consequence of the fact that there is no uniformly powerful test of the unit root hypothesis (see Stock, 1994b). We shall, therefore, give an overview of the chapter and a clarification of the tests discussed. Then we shall outline what tests we think are important and what are not.

In section 4.2 we discuss in detail the evidence on size distortions and low power of the commonly used unit root tests such as the ADF and PP (Phillips–Perron) tests. This gives a diagnosis of the problem. The next question is what to do about it. This is discussed in section 4.3.

One of the major problems noted is the presence of MA components in the errors which cannot be taken into account without very long autoregressions. Hence, in section 4.3.1 we discuss unit root tests in ARMA models. Another solution is to use instrumental variables. Hence, in sections 4.3.2 and 4.3.3 we discuss unit root tests based on instrumental variable regressions.

The remaining parts of section 4.3 are devoted to modifications of the PP tests (by Perron and Ng) and ADF test (by Elliott, Rothenberg, and Stock), tests based on weighted symmetric estimators (Fuller), and tests based on reverse and forward regressions (Leybourne).

All these tests provide improvements over the PP and ADF tests. There is, as yet, no comprehensive study comparing all these tests. But our guess is that such studies will not provide any clearcut evidence in favor of one or the other. The only clearcut evidence that would emerge (which is already there in the several separate studies comparing them with the existing tests) is that they clearly dominate the ADF and PP tests. Hence, ADF and PP tests, still often used, should be discarded in favor of these tests.

Given that the major problem with tests based on autoregressions are high MA roots, tests have been developed for MA unit roots. These are discussed in section 4.4. Another issue that has been raised is: why do we assume the unit root as the null hypothesis? This problem does not arise in the Bayesian approach (see chapter 8) where the null and alternative are on an equal footing. Tests using stationary as null are discussed in section 4.5. These tests, however, have the same type of problems as the ADF and PP tests. They have also been suggested as

useful for *confirmatory analysis* that is in conjunction with the ADF or PP tests. This issue is discussed in section 4.6 where questions have been raised about the usefulness of such confirmatory analysis.

The rest of the chapter is devoted to discussions of other tests that have been suggested to *solve* the poor power problem of unit root tests, by increasing the number of observations, by increasing the frequency of observations (section 4.7), and using panel data (section 4.9). Some other problems discussed are: uncertain unit roots and the pre-testing problem (section 4.10) and unbiased estimation as a solution to unit root testing (section 4.12).

The chapter thus discusses several unit root tests but they follow a logical order as outlined above. We also discuss some other issues raised in unit root testing.

4.2 Size distortion and low power of unit root tests

Schwert (1989) first presented Monte Carlo evidence to point out the size distortion problems of the commonly used unit root tests. He argued that the distribution of the Dickey–Fuller tests is far different from the distribution reported by Dickey and Fuller if the underlying distribution contains a moving-average (MA) component. He also suggests that the Phillips–Perron (PP) tests suffer from size distortions when the MA parameter is large, which is the case with many economic time series as noted by Schwert (1989). The test with the least size distortion is the Said and Dickey (1984) high-order autoregression *t*-test. Schwert argues that it is important to consider the correct specification of the ARIMA process before testing for the presence of unit roots. Whereas Schwert complained about the size distortion of unit root tests, DeJong *et al.* complained about the low power of unit root tests. DeJong *et al.* (1992a) argued that the unit root tests have low power against plausible trend-stationary alternatives. DeJong *et al.* (1992b) consider Monte Carlo experiments to study the size and power of the ADF test and the PP tests and argue that the ADF test displays size distortions in the presence of negatively correlated MA errors and the PP tests suffer from serious size distortions with plausibly correlated MA or AR error structures. They argue that the PP tests have very low power (generally less than 0.10) against trend-stationary alternatives but the ADF test has power approaching one-third and thus is likely to be more useful in practice. They conclude that tests with higher power need to be

developed. Similar problems about size distortions and low power were noticed by Agiakoglou and Newbold (1992).

There is one other aspect to these size distortions and low power problems. The Beveridge and Nelson (1981) decomposition (discussed in chapter 2) showed that a series with a unit root can be decomposed into a pure random walk with drift and a stationary component. A trend-stationary series can also have trivially such decomposition with the variance of the random walk component zero. The effect of the variances of the errors in the random walk and stationary components on the ADF and PP tests was studied by Liu and Praschnik (1993) who show that the size distortions are sensitive to the ratio of these variances but that, of the tests they considered, the ADF test was the least sensitive.

Cochrane (1991) uses the Beveridge–Nelson decomposition to argue that since the random walk component can have arbitrarily small variance, tests for unit roots or trend stationarity have arbitrarily low power in small samples. Also, Cochrane shows that there are unit root processes whose likelihood function and autocorrelation functions are arbitrarily close to those of any given stationary process and vice versa. Cochrane, thus, argues that inference on unit roots or trend stationarity can be fragile.

The poor power problems is not unique to the unit root tests. Cochrane argues that any test of the hypothesis $\theta = \theta_0$ has arbitrarily low power against alternatives $\theta_0 - \varepsilon$ in small samples, but in many cases the difference between θ_0 and $\theta_0 - \varepsilon$ would not be considered important from the statistical or economic perspective. But the low power problem is particularly disturbing in the unit root case because of the discontinuity of the distribution theory near the unit root. Cochrane complained that:

the results of unit root tests do not necessarily answer one important question namely: which distribution theory provides a better small sample approximation? (Cochrane, 1991, p. 283)

He points out that the borderline cases are not a technical curiosity but occur in a lot of empirical examples.

Mention must be made of a recent paper by Gonzalo and Lee (1996), who complain about the repetition of the phrase “lack of power unit root tests.” They show numerically that the lack of power and size distortions of the Dickey–Fuller tests for unit roots are similar to and in many situations even smaller than the lack of power and size distortions of the standard student t -tests for stationary roots in an autoregressive

model. But arguments like this miss the important point. There is no discontinuity of inference in the latter case but there is in the case of unit root tests. Thus, the consequences of lack of power are vastly different in the two cases.

Blough (1992) follows the same arguments of Cochrane but also presents Monte Carlo results. He formulates the unit root hypothesis not as a point hypothesis but as an interval hypothesis. He starts from the model

$$y_t = m_t + \delta y_{1t} + (1 - \delta)y_{2t}, \quad 0 \leq \delta \leq 1$$

where m_t is a deterministic process, y_{1t} is a random walk, and y_{2t} is a purely nondeterministic stationary process, i.e.

$$\Delta y_{1t} = \varepsilon_{1t} \quad \text{and} \quad y_{2t} = a(L)\varepsilon_{2t}$$

The errors $(\varepsilon_{1t}, \varepsilon_{2t})$ are *iid* with mean zeros and covariances

$$\begin{bmatrix} \sigma_1^2 & \sigma_{12} \\ \sigma_{12} & \sigma_2^2 \end{bmatrix}$$

and $a(L)$ is a square summable lag polynomial of possibly infinite order with $a(0) = 1$.

It is possible to construct a variety of processes by varying δ . Call these $y_t(\delta)$. For any $\delta > 0$ the asymptotic properties of statistics generated from $y_t(\delta)$ are dominated by the random walk component y_{1t} . However, for finite samples, $y_t(\delta)$ will behave like y_{2t} when δ is small.

Blough defines the *generic* null hypothesis of a unit root as the hypothesis that $\delta > 0$. Thus, for a generic null of unit root versus generic stationary alternative, we have

$$H_0 : 0 < \delta \leq 1$$

$$H_1 : \delta = 0$$

For this test he shows that the power against any stationary alternative can be no greater than the level of the test.

Blough argues from this that the explanation of unit root tests by the standard power/level criteria should proceed with caution. Unit root tests are often used to provide information about the distributions of statistics for subsequent inference. The asymptotic distributions of such statistics depend on whether the underlying processes have unit roots. But their finite sample distributions must be nearly the same for a given stationary process and for a unit root process nearby. Hence the diagnostic value of unit root tests lies not in establishing a unit root

but in establishing what asymptotic theory provides good finite sample approximation. As one example Blough provides limited Monte Carlo evidence that suggests that probabilities of spurious regression match up quite well with the rejection probabilities of low-order ADF tests. In this example the interest is not in rejection of acceptance of a unit root *per se* but in the spurious regression.

We shall discuss this issue of using unit root tests for subsequent inference in a later section of this chapter.

4.3 Solutions to the problems of size and power

There have been several solutions to the problems of size distortion and low power of the ADF and PP tests mentioned in the previous section. Some of these are modifications of the ADF and PP tests and others are new tests. We shall review them first and then outline some guidelines for choice.

4.3.1 LR tests in ARMA models

Since the criticism of the tests based on AR models has been that they have size distortions and low power against meaningful stationary alternatives in the presence of large MA components, it would be desirable to look at tests for unit roots in ARMA models. Yap and Reinsel (1995) develop likelihood ratio (LR) tests for a unit root in an ARMA model and present limited simulation evidence on their performance in the context of an ARMA (1,1) model. Since Schwert (1987) argues that many economic time series can be characterized by an ARMA (1,1) model, this evidence is of practical interest.

Consider the ARMA model

$$y_t = \sum_{j=1}^p \phi_j y_{t-j} + \varepsilon_t - \sum_{j=1}^q \theta_j \varepsilon_{t-j}$$

where $\varepsilon_t \sim iid(0, \sigma_\varepsilon^2)$. Consider the case where y_t is stationary and $\phi(L) = 0$ has one unit root and all other roots are outside the unit circle. We can write this equation in first differences as

$$\Delta y_t = c y_{t-1} + \sum_{j=1}^{p-1} \phi_j \Delta y_{t-j} + \varepsilon_t - \sum_{j=1}^q \theta_j \varepsilon_{t-j}$$

where $c = -(1 - \sum_{j=1}^p \phi_j)$.

In the presence of a unit root, $c = 0$. Let S denote the residual sum of squares (RSS) in the Gaussian estimation of the unconstrained model and S_0 denote the RSS under the restricted model $c = 0$. Under $H_0 : c = 0$, Yap and Reinsel show that the LR test statistics is asymptotically distributed as

$$\frac{T(S_0 - S)}{S} \Rightarrow \frac{\int_0^1 W(r)dW(r)}{\int_0^1 W(r)^2 dr}$$

which is the Dickey–Fuller distribution. This implies that the addition of MA terms to a unit root nonstationary AR model does not alter the asymptotic distribution of the LR test statistics for the hypothesis $H_0 : c = 0$ that one would obtain in a pure AR model.

Yap and Reinsel discuss other Wald type statistics for testing the hypothesis $H_0 : c = 0$ and show that their asymptotic distribution is the same as that of the LR test statistic. This does not mean that the finite sample distributions are the same. To investigate this they consider a limited Monte Carlo study of an ARMA (1,1) model. Their results suggest that overall the LR test performs better than the other tests (in terms of size and power properties) in the ARMA (1,1) model when a substantial MA component (θ_1 positive) is present.

4.3.2 IV tests in ARMA models

One other set of unit root tests that have been suggested in the context of ARMA models is the instrumental variable (IV) tests starting with the paper by Hall (1989). Hall suggests an IV estimation in an ARMA (1, q) model and Pantula and Hall (1991) extend this to an ARMA (p , q) model. Hall suggests using y_{t-k} ($k > q$) as an instrumental variable. The IV estimator is therefore defined as

$$\hat{\rho} = \frac{\sum y_t y_{t-k}}{\sum y_{t-1} y_{t-k}}$$

Based on this Hall defines the coefficient test (K -test) and the t -test corresponding to the similar test statistics in the Dickey–Fuller regression. He proves that the asymptotic distribution of these statistics are the corresponding DF distributions.

In the case of the t -statistic the factor $\sum y_{t-1} y_{t-k}/T^2$ need not be positive. In this case Hall's t -statistic is not defined. This can happen if (i) the autoregressive coefficient is small, (ii) k is large, or (iii) T is

small. Li (1995) defines a modified t -statistic by noting that

$$\begin{aligned} \frac{1}{T^2} \sum_{t=k}^T y_{t-k} y_{t-1} &= \frac{1}{T^2} \sum_{t=k}^T y_{t-1}^2 \\ &+ \frac{1}{T^2} \sum_{t=k}^T y_{t-1} (u_{t-1} + u_{t-2} + \cdots + u_{t-k}) \end{aligned}$$

The second term converges in probability to zero. Hence Li defines a modified t -statistics by substituting $\sum_{t=k}^T y_{t-1}^2$ for $\sum_{t=k}^T y_{t-k} y_{t-1}$. Li also investigates the powers of both the coefficient test and the modified t -test (which was not done in Hall's paper) and finds that the power functions of the IV tests are not monotonic functions of the autoregressive parameter when there is a negative moving-average component.

The tests developed in Hall (1989) and Pantula and Hall (1991) follow the Dickey-Fuller framework. Lee and Schmidt (1994) use the Bhargava approach to develop IV unit root tests (see section 3.7 of chapter 3 for the difference in the two approaches).

For the model with trend, define the residuals

$$\tilde{u}_t = (y_t - y_1) - \frac{t-1}{T-1} (y_T - y_1)$$

Then the Bhargava test, discussed in Schmidt and Phillips (1992) as an LM test and denoted as the BSP test depends on the regression

$$\Delta \tilde{u}_t = \phi \tilde{u}_{t-1} + \text{error}, \quad t = 2, \dots, T$$

The IV test developed in Lee and Schmidt uses $\tilde{u}_{t-k}$ as an instrumental variable where $k > q$ is the order of the MA component as in Hall. They derive the asymptotic distributions of the coefficient test and t -test. They also suggest a modified coefficient, the modification being similar to the one in Li that we discussed earlier viz substituting $\sum \tilde{u}_{t-1}^2$ for $\sum \tilde{u}_{t-1} \tilde{u}_{t-k}$. Thus the test statistic used is

$$T \frac{\sum \tilde{u}_{t-k} \Delta \tilde{u}_t}{\sum \tilde{u}_{t-1}^2}$$

Lee and Schmidt conduct detailed Monte Carlo studies to compare the size and power of the BSP coefficient and t -tests in the autoregression with no MA error, and the IV tests and corrected IV tests developed in their paper. They conclude that the modified IV test has good power and is the best choice. The simulation results were all for $q = 1$ and $k = 2$, using larger values of k resulted in a loss of power.

In practice one major problem with all these IV tests is the choice of k . Clearly overestimating k results in a loss of power. The optimal k is $q + 1$. The IV methods avoid the need to estimate the MA parameters but one needs to have information on the orders of the processes. Hall (1995) discusses methods of determining the orders of p and q in the ARMA (p, q) model using the autocovariances of the residuals from the IV regression. There is the problem that the properties of these autocovariances themselves depend on whether the true process is stationary or has a unit root, and there is the dilemma that these are being used as a basis of model selection prior to performing a unit root. Hall suggests on the basis of simulation evidence that it is better to use the IV residuals assuming the process is stationary. Detailed results from Hall are omitted here. Hall and Hassett (1990) gives an earlier discussion of this issue.

The importance of correctly specifying the order of p and q in the IV tests has been emphasized in Pantula and Hall (1991) who show that if the model is overspecified, then the empirical levels are higher than the nominal levels in moderate-sized samples. On the other hand, if the model is underspecified, the IV estimators are inconsistent.

4.3.3 Other IV based tests – Durbin–Hausman tests

Choi (1992a) suggests some tests based on what he calls pseudo-IV estimators. The IV estimator is called a pseudo-IV estimator because it is based on using the current value y_t rather than the lagged values y_{t-k} , used in the previous discussion as instruments. Thus, in the model

$$y_t = \alpha y_{t-1} + u_t$$

the OLS estimator is $\sum y_{t-1}y_t / \sum y_{t-1}^2$. The pseudo-IV estimator that Choi considers is

$$\hat{\alpha}_{IV} = \frac{\sum y_t^2}{\sum y_{t-1}y_t}$$

This is the estimator we would get if we considered the reverse regression

$$y_{t-1} = \alpha^* y_t + u_t^*$$

and used OLS. We shall see later that this is related to tests based on symmetric estimators and reverse Dickey–Fuller tests.

Choi considers the difference $(\hat{\alpha}_{IV} - \hat{\alpha}_{OLS})$ standardized by the usual variance estimators of $\hat{\alpha}_{OLS}$. He calls this the Durbin–Hausman test for

a unit root and derives the asymptotic distributions of similarly defined statistics in the three DF-type regression models

$$y_t = \alpha y_{t-1} + u_t$$

$$y_t = \mu + \alpha y_{t-1} + u_t$$

$$y_t = \mu + \beta t + \alpha y_{t-1} + u_t$$

and labels them respectively as DHS₁, DHS₂, and DHS₃.

The Durbin–Hausman test (or Hausman’s specification error test) is based on the difference between two estimators both of which are consistent under the null and one of which is not consistent under the alternative. It also requires the difference to diverge under the alternative for test consistency. The variance of the difference of the two estimators is equal to the differences in the variances if one of the two estimators is efficient. This follows from the result in Rao (1973) that the covariance between an efficient estimator and the difference between an efficient and inefficient estimator is zero. Also, as is well known, the Hausman specification test is not the same as a hypothesis test.

The test statistics that Choi considers are not standardized by using the correct variance of $(\alpha_{IV} - \alpha_{OLS})$. They are standardized by using the arbitrary variance, $var(\alpha_{OLS})$. Thus they cannot be called Durbin–Hausman tests. This may explain the lack of widespread use of these tests.

Choi argues that his DHS tests display better power properties in finite samples than the Dickey–Fuller tests, especially DHS₂ and DHS₃. Also the limiting distributions are easier to derive, and serial correlation corrections are easier to make. These advantages might be outweighed by the fact that the tests are not really Durbin–Hausman tests because of the arbitrary normalization factors used.

4.3.4 Modifications of the Phillips–Perron (PP) tests

We have discussed likelihood ratio based tests and IV tests in the presence of MA errors. We shall discuss one other alternative: modification of the PP tests.

The PP tests were originally designed to take care of MA errors as well. However, as discussed earlier, they suffer from serious size distortions when there are negative MA errors. The DF test does not have such serious size distortions but it is less powerful than the PP tests. (Of course, in all these cases one can consider the *size adjusted power* or

the power when the test is performed using whatever critical value that makes the size correct).

Perron and Ng (1996) suggest modifications of the PP tests to correct this problem. They use methods suggested by Stock (1990) to derive modifications of the Z_ρ and Z_t statistics (see chapter 3, section 3.6.3). The modified Z_ρ and Z_t statistics are

$$MZ_\rho = Z_\rho + \frac{T}{2}(\hat{\rho} - 1)^2$$

Convergence of $\hat{\rho}$ at rate T ensures that Z_ρ and MZ_ρ are asymptotically equivalent. Defining

$$MSB = (T^{-2} \sum y_{t-1}^2 / s^2)^{1/2}$$

they note that $Z_t = MSB \cdot Z_\rho$. Hence they define the modified Z_t statistics by

$$MZ_t = MSB \cdot MZ_\rho$$

If we write the model as

$$y_t = \rho y_{t-1} + u_t$$

then for the computation of the PP test statistics we need estimates of two error variances $\sigma_u^2 = \text{var}(u_t)$ and $\sigma^2 = \lim_{T \rightarrow \infty} T^{-1} E(S_T^2)$ where $S_T = \sum_{j=1}^T u_j$. For an estimate of σ_u^2 they use

$$S_u^2 = T^{-1} \sum u_t^2$$

For the estimate of σ^2 they suggest using an autoregressive estimator defined as

$$S_{AR}^2 = \frac{S_{e_k}^2}{(1 - \hat{b}(1))^2}$$

where

$$S_{e_k}^2 = T^{-1} \sum_{t=k+1}^T \hat{e}_{tk}, \quad \hat{b}(1) = \sum_{j=1}^k \hat{b}_j$$

and $\hat{b}_j$ and $\{\hat{e}_{tk}\}$ are obtained from the autoregression

$$\Delta y_t = b_0 y_{t-1} + \sum_{j=1}^k b_j \Delta y_{t-j} + e_{tk}$$

Perron and Ng argue that this autoregressive spectral density estimator works best and that the modifications of the PP tests suggested still

have problems with other estimators of σ^2 . They use local asymptotic analysis to explain why the other estimators yield no improvement even with the modified statistics.¹

In addition to the MZ_ρ and MZ_t statistics, Perron and Ng also investigate the size and power properties of the MSB statistic defined earlier, but with the above estimate of σ^2 . Critical values for the demeaned and detrended case for this statistic were taken from Stock (1990). Perron and Ng show that the modified PP tests are able to maintain good power while correcting the size distortion problems in the presence of negative MA errors common in most macroeconomic time series.

The importance of these modifications of the PP tests from the practical point of view is that they do not have to specify the order of the MA component as the IV tests discussed earlier. Also, MA can arise in time series models with additive outliers (outlier problems in time series are discussed in chapter 14), and the advantage of the modified PP tests is that one does not need to identify the outliers before applying the unit root tests.

4.3.5 Forward and reverse Dickey–Fuller regressions

Leybourne (1995) suggests tests based on forward and reverse Dickey–Fuller regressions. Since these are related to the tests by Choi (1992a) discussed earlier, we shall go through these in detail. The results in Leybourne are more transparent than in Choi and also, as argued earlier, the modification of the tests in Choi, being based on the Durbin–Hausman tests, is rather questionable.

Consider the following DF regression

$$y_t = \alpha + \rho y_{t-1} + u_t, \quad t = 2, \dots, T$$

Let $\hat{\rho}$ be the estimate of ρ and RSS the residual sum of squares from a least squares estimation of this equation. Then

$$\hat{\sigma}^2 = \frac{RSS}{T-3}$$

and the DF t -statistic which we denote by DF_f can be written as (using the expression from the estimation of a linear regression)

$$DF_f = (\hat{\rho} - 1)[\text{var}(\hat{\rho})]^{-1/2}$$

¹ Local asymptotic power involves investigating power for alternatives $\rho_T = 1 + c/T$. For details and references to work in this area, see Stock (1994b, pp. 2270–2271).

$$= \hat{\sigma}^{-1} \frac{\text{cov}(y_t, y_{t-1}) - \text{var}(y_{t-1})}{[\text{var}(y_{t-1})]^{1/2}}$$

Now consider the reverse DF regression. To avoid confusion, define $z_t = y_{T+1-t}$ ($t = 0, 1, 2, \dots, T+1$), i.e., $z_1 = y_T$, $z_2 = y_{T-1}$ and so on, and the reverse regression

$$z_t = \alpha^* + \rho^* z_{t-1} + v_t$$

The DF t -statistic from the reverse regression is defined DF_r by the same expression as DF_f with z_t replacing y_t and $\hat{\sigma}^*$ replacing $\hat{\sigma}$.

Noting that $\text{var}(z_t) = \text{var}(y_{t-1})$, $\text{var}(z_{t-1}) = \text{var}(y_t)$, and $\text{cov}(z_t, z_{t-1}) = \text{cov}(y_t, y_{t-1})$, it can be seen that $\hat{\rho}^*$ from the reverse DF regression is the same as the pseudo-IV estimator that Choi considers. However, its motivation from the reverse DF regression is more transparent than from the IV perspective.

Leybourne shows after some algebra that the difference between the forward and reverse DF_f statistic is given by

$$DF_r = DF_f - \lambda$$

where

$$\lambda = \hat{\sigma}^{-1} [\text{var}(y_{t-1})]^{-1/2} [\text{var}(y_t) - \text{var}(y_{t-1})] = \hat{\sigma}^{-1} Q$$

$$Q = (y_T^2 - y_1^2) + (T-1)^{-1} \left\{ (y_T - y_1)^2 - 2(y_T - y_1) \sum_{t=2}^T y_t \right\}$$

We might note that the difference Choi considers is the difference between the coefficient statistics, not the t -statistics. Thus λ is an end-effect measure that depends on $y_T - y_1$. If y_t is $I(1)$, as under the null of a unit root, this difference will diverge. Note that a requirement for the Durbin-Hausman test is that the difference diverge under the *alternative* (see Choi, 1992a, p291). Here the difference diverges under the *null*.

Under the unit root null

$$y_t = \alpha + y_{t-1} + u_t$$

Using this and simplifying the expression in Q Leybourne shows that

$$\lambda \Rightarrow D^{-1/2} \left[W(1)^2 - 2W(1) \int_0^1 W(r) dr \right]$$

where

$$D = \int_0^1 W(r)^2 dr - \left(\int_0^1 W(r) dr \right)^2$$

Since

$$DF_f \Rightarrow D^{-1/2} \left[\frac{W(1)^2}{2} - \frac{1}{2} - W(1) \int_0^1 W(r) dr \right]$$

we get

$$DF_r \Rightarrow D^{-1/2} \left[\frac{W(1)^2}{2} - \frac{1}{2} + W(1) \int_0^1 W(r) dr \right]$$

Under the stationary alternative, DF_f and DF_r have the same distribution which is the distribution for DF_f .

The above expressions are for the case of a model with drift. In a model with time trend the expressions carry through but they involve demeaned and detrended Brownian motions. Again $DF_r = DF_f - \lambda$. In the case where the error follows an $AR(p)$ process, we have to consider the ADF_f and ADF_r statistics. These can be shown to have the same distributions (under the null of a unit root) as the DF_f and DF_r distributions respectively.

Having derived the asymptotic distribution of $DF_f - DF_r$, it is not clear why Leybourne did not consider a test based on this. Instead he considers a test statistic DF_{max} which is defined as

$$DF_{max} = \text{Max}(DF_f, DF_r)$$

The asymptotic distribution of this statistic is very difficult to derive. Leybourne obtains critical values of this test statistic by simulation using a model with normally distributed errors. The critical values do not vary much with sample size (see table 4.1).

Leybourne investigates the power of the DF_{max} test and finds that there is a 15 percent increase in power relative to DF_f test. He also investigates the robustness of the results to obtain other error distributions and also gives an empirical application based on the Nelson–Plosser data arguing that the DF_{max} test shows less evidence of a unit root than the DF_f test.

4.3.6 Weighted symmetric estimators

Consider the process

$$y_t = \alpha + \rho y_{t-1} + u_t, \quad t = 2, \dots, T$$

Table 4.1. *Critical values of DF_{max} statistics*

Nominal size	Sample size					
	Model with drift			Model with drift and trend		
	25	50	200	25	50	200
1.00	-2.15	-2.14	-2.13	-2.89	-2.87	-2.53
0.05	-2.50	-2.48	-2.44	-3.26	-3.22	-3.12
0.01	-3.25	-3.17	-3.06	-3.99	-3.84	-3.72

Source: Leybourne (1995, p. 565).

The idea behind the symmetric estimators is that if a normal stationary process satisfies this equation, then it also satisfies the equation

$$y_t = \alpha + \rho y_{t+1} + u_t$$

This symmetry leads one to consider an estimator of ρ that minimizes

$$Q(\rho) = \sum w_t (y_t - \rho y_{t-1})^2 + \sum (1 - w_t) (y_t - \rho y_{t+1})^2 \quad (4.1)$$

where $y_t = y_t - T^{-1} \sum_{j=1}^T y_j$. For $w_t = 1$ we get the OLS estimator. For $w_t = 0.5$ we get the simple symmetric estimator considered in Dickey, Hasza, and Fuller (1984). For $w_t = (t - 1)/T$ we get the weighted symmetric (WS) estimator suggested in Park and Fuller (1995). A detailed discussion of WS estimators in nonstationary models can be found in chapter 10 of Fuller (1996).

For the model with drift (but no trend), with y_t defined as deviations from the mean, the WS estimator of ρ is

$$\hat{\rho}_{ws} = \frac{\sum_{t=2}^T y_t y_{t-1}}{\sum_{t=2}^{T-1} y_t^2 + T^{-1} \sum_{t=1}^T y_t^2}$$

and the corresponding t -statistic is

$$\hat{t}_{ws} = \hat{\sigma}_{ws}^{-1} (\hat{\rho}_{ws} - 1) \left(\sum_{t=2}^{T-1} y_t^2 + T^{-1} \sum_{t=1}^T y_t^2 \right)^{1/2}$$

where $\sigma_{ws}^2 = Q(\hat{\rho}_{ws})/(T - 2)$ and $Q(\cdot)$ is defined in (4.1). The WS estimator can exceed 1 in some models, particularly those for which the ML estimator is < 1 but very close to 1. The symmetric estimator treats the beginning and last observation the same way. Note that under the null of a unit root $y_T - y_1$ diverges.

The limiting distributions of the WS estimators are given by

$$T(\hat{\rho}_{ws} - 1) \Rightarrow D^{-1}N$$

$$\hat{t}_{ws} \Rightarrow D^{-1/2}N$$

where

$$D = \int_0^1 W(r)^2 dr - \left[\int_0^1 W(r) dr \right]^2$$

the expression we get in the DF distributions and

$$N = \frac{W(1)^2 - 1}{2} - W(1) \int_0^1 W(r)^2 dr - \int_0^1 W(r)^2 dr + 2 \left[\int_0^1 W(r) dr \right]^2$$

The expressions for the case of a model with linear trend are more complicated. They are given in Fuller (1996, p. 571) and will not be reproduced here.

Pantula *et al.* (1994) present Monte Carlo evidence in the case of the model with drift (but no trend) to show that the WS estimators are considerably more powerful than the DF statistics and that they have the best power properties among the estimators they investigated.

4.3.7 DF-GLS test

Elliott, Rothenberg, and Stock (1996), which we shall denote by ERS, first derive the asymptotic power envelope for point optimal tests of a unit root in the autoregressive representation of a Gaussian time series under various specification of the trend. King (1987) defines a point optimal test as a test that optimizes power at a predetermined point under the alternative. These tests are second best when uniformly most powerful (UMP) tests do not exist. We shall show later why UMP tests do not exist for the unit root hypothesis.

After developing the asymptotic power envelope, ERS propose a family of tests whose power functions are tangent to the power envelope at one point and are never too far below the envelope. They call this test $P_T(0.5)$. They then suggest the Dickey-Fuller GLS (DF-GLS) test as one that has the limiting power function close to that of the $P_T(0.5)$ test.

Table 4.2. *Critical values for the Elliott–Rothenberg–Stock DF–GLS test*

Sample size	Critical values		
	1%	5%	10%
50	−3.77	−3.19	−2.89
100	−3.58	−3.03	−2.74
200	−3.46	−2.93	−2.64
∞	−3.48	−2.89	−2.57

Note: The model is with linear trend and $\bar{c} = -13.5$.

Source: Elliott, Rothenberg, and Stock (1996, table 1, p. 825).

Let y_t be the process we consider. The DF–GLS t -test is performed by testing the hypothesis $a_0 = 0$ in the regression

$$\Delta y_t^d = a_0 y_{t-1}^d + a_1 \Delta y_{t-1}^d + \cdots + a_p \Delta y_{t-p}^d + \text{error}$$

where y_t^d is the locally detrended series y_t . The local detrending depends on whether we consider a model with drift only or a linear trend. The latter is one most commonly used. In this case we have

$$y_t^d = y_t - \hat{\beta}_0 - \hat{\beta}_1 t$$

where $(\hat{\beta}_0, \hat{\beta}_1)$ are obtained by regressing $\bar{y}$ on $\bar{z}$ where

$$\bar{y} = [y_1, (1 - \bar{\alpha}L)y_2, \dots, (1 - \bar{\alpha}L)y_T]$$

$$\bar{z} = [z_1, (1 - \bar{\alpha}L)z_2, \dots, (1 - \bar{\alpha}L)z_T]$$

and

$$z_t = (1, t)', \quad \bar{\alpha} = 1 + \frac{\bar{c}}{T}$$

The $\bar{\alpha}$ that produces the asymptotic power depends on the significance level used ε . This is very inconvenient but ERS argue that fixing $\bar{c} = -7$ in the model with drift and $\bar{c} = -13.5$ in the linear trend case, the limiting power function of the P_T tests and the DF t -test applied to the locally trended data are within 0.01 of the power envelope for $0.01 < \varepsilon < 0.10$. They provide critical values for this DF–GLS test. They are given in table 4.2.

Hwang and Schmidt (1996) suggest another type of DF–GLS test.

They start with the Bhargava-type specification (see section 2.9 of chapter 2)

$$y_t = \gamma_0 + \gamma_1 t + u_t$$

$$u_t = \alpha u_{t-1} + e_t$$

which gives

$$\Delta y_t = \beta_0 + \beta_1 t + (\alpha - 1)y_{t-1} + e_t$$

Consider two residuals from this equation $\tilde{e}_t$ obtained by estimating this equation with $\alpha = 1$ and e_t^* obtained by estimating this equation using a value $\alpha = \alpha^*$.

The Bhargava test which is derived in Schmidt and Phillips as an LM test (see the section 3.3 of chapter 3) is based on the regression

$$\Delta \tilde{e}_t = \phi \tilde{e}_{t-1} + \text{error}$$

and testing $\phi = 0$. They suggest choosing a value of α so as to optimize the power at a predetermined point for the alternative (as in point optimal tests). They suggest a value $\alpha^* = 0.85$ for annual data. Note that this implies a half life of four years. (A half life is defined as n where $(\alpha^*)^n = 0.50$ – it is the time taken to achieve a 50 percent adjustment to shocks.) Hence with quarterly data if we take $\alpha^* = 0.95$, this implies a half life of around 14 quarters or 3.5 years.

Hwang and Schmidt present (using simulations) the critical values for this test for α^* in the range (0, 1) and sample sizes 50 – 500. They also present the powers of this test for different true values of α . The tables are too numerous to be discussed here. Some important numbers for practical use with the DF type t -test are given in table 4.3. The critical values are for $\alpha^* = 0.85$ (for use with annual data) and $\alpha^* = 0.95$ (for use with quarterly data).

Hwang and Schmidt argue that their GLS test is more powerful than the DF test or the Bhargava–Schmidt–Phillips LM test. There is no comparison with the ERS test. The limitation of the Hwang and Schmidt procedure is that, it does not have an asymptotic justification. It is only relevant for the exact same model used to generate the small sample critical values. The procedure is not of general applicability.

Table 4.3. *Critical values for the Hwang-Schmidt DF-GLS test (t-test)*

Sample size	Critical values					
	$\alpha^* = 0.85$			$\alpha^* = 0.95$		
	1%	5%	10%	1%	5%	10%
50	-3.85	-3.26	-2.95	-3.57	-2.97	-2.68
100	-3.91	-3.32	-3.03	-3.62	-3.07	-2.81
200	-3.93	-3.67	-3.09	-3.73	-3.20	-2.93
500	-3.93	-3.39	-3.11	-3.82	-3.32	-2.04

Source: Hwang and Schmidt (1996).

4.4 Problem of overdifferencing: MA roots

Let us consider the simple moving-average model

$$y_t = d_t + u_t, \quad \Delta u_t = (1 - \theta L)e_t$$

where e_t is $I(0)$ and d_t is a deterministic component. The case $|\theta| = 1$ is known as the noninvertible MA process. If $\theta = 1$ we have a unit MA root and $u_t = e_t + u_0 - e_0$. In this case with $u_0 = e_0$, u_t is an $I(0)$ process. If $|\theta| < 1$, then $(1 - \theta L)$ is invertible and u_t is $I(1)$. Thus a test of the hypothesis $\theta = 1$ versus $|\theta| < 1$ is a test of the null hypothesis that y_t is $I(0)$ versus y_t is $I(1)$.

MA unit roots can occur if a stationary series is differenced, i.e., there is *overdifferencing*. For this reason the occurrence of a maximum of the likelihood function near or at $\theta = 1$ can be interpreted as evidence of overdifferencing (see, for instance, Plosser and Schwert, 1977). However, Sargan and Bhargava (1983) point out that the occurrence of a maximum of the likelihood function at $\theta = 1$ is insubstantial evidence for overdifferencing because the likelihood function can have a local maximum at $\theta = 1$ with relatively high probability, even when the true θ is less than 1. Kang (1975) showed that the likelihood function is stationary at $\theta = 1$ and simulation studies on the behavior of the ML estimator of θ reveal that the global maximum can occur at $\theta = 1$ even if the true θ is in the invertible region. Anderson and Takemura (1986) also proved that the probability that a noninvertible process is estimated when the true process is invertible and is $O(T^{-n})$ where n is a positive integer. This problem is known as the *pile-up* problem in the literature on MA models.

Sargan and Bhargava prove that the ML estimate of a noninvertible

MA(q) model is T -consistent, showing the same speed of consistency as the least squares estimator of a unit root AR process. They also show that the limiting probability that the MLE is exactly 1 when $\theta = 1$ is 0.6575. One other point made by them is that if $\theta = 1$, the MLE $\hat{\theta}$ is not normally distributed and hence the LR test cannot be applied to test the hypothesis $\theta = 1$. On the other hand, the MLE of an invertible MA process is normally distributed.

Shephard (1993) establishes an approximate sampling theory for the MLE of an MA process which will be accurate even in a strictly noninvertible case. He does this by considering a *local level model* which is an unobserved components model represented as follows

$$\begin{aligned} y_t &= \mu_t + \varepsilon_t, \quad \varepsilon_t \sim iid(0, \sigma^2) \\ \mu_t &= \mu_{t-1} + \eta_t, \quad \eta_t \sim iid(0, q\sigma^2) \end{aligned}$$

ε_t and η_s are independent for all t and s . Interest centers in the parameter q which is also called the *signal-noise ratio*. This is an important model studied (among several others) by Leybourne and McCabe (1989) and Harvey (1989). For this model, even when the random component of y_t is dominant (i.e. $q \neq 0$), there is a nonzero probability that the estimated value of q using ML is zero. Shephard and Harvey (1990) suggest assuming a diffuse prior for μ_0 and integrating it out.

It should be noted that the unobserved components (UC) model is very much related to the MA root model. They are alternative parameterizations of the same model. See Stock (1994b, p. 2789) who says:

In general for suitable choices of initial conditions, all MA models have UC representation and all UC models have MA representation.

This correspondence is useful because the work on UC models can be used to study tests for MA unit roots.

Choi and Yu (1997) consider a method of testing the unit root MA hypothesis by using the framework for unit root tests based on AR models. This is done by aggregating the series y_t as follows: define $S_t = \sum_{i=1}^t y_i$ with $y_1 = 0$, then $S_t = \alpha S_{t-1} + y_t$ with $\alpha = 1$. Hence a test of the hypothesis $\theta = 1$ (which as discussed earlier implies that y_t is $I(0)$) is equivalent to the joint hypothesis $\alpha = 1$ and y_t is $I(0)$. Thus the unit root testing framework for AR models can be used on this aggregated equation. Choi and Yu argue that this is an equivalent way of testing for a unit MA root but the advantage is that the likelihood function for the AR(1) model with fixed initial conditions is easier to handle than the likelihood function for the unconditional MA(1) model.

Noting this analogy, any of the unit root tests discussed earlier can be used to test $\alpha = 1$ on S_t . However, Choi and Yu, develop what they call SBDH tests (Sargan–Bhargava–Durbin–Hausman tests – although as we noted earlier the Durbin–Hausman terminology has some objections).

For the unit root hypothesis, the LM test statistic is (see chapter 3)

$$\text{LM} = \left[\frac{\sum_{t=2}^T S_{t-1} \Delta S_t}{T \hat{\sigma}^2} \right]^2$$

where $\hat{\sigma}^2 = \sum_{t=2}^T y_t^2 / T$. The SBDH statistic they suggest is

$$\text{SBDH} = \frac{\sum_{t=2}^T S_t^2}{\hat{\sigma}^2 T^2}$$

To allow for more general error structures, they modify these test statistics along the lines of Phillips and Perron (1988) so that the asymptotic distributions of these tests are free of nuisance parameters. The modified test statistics are

$$\text{LM}^0 = \left[\frac{\sum_{t=2}^T \Delta S_t S_{t-1}}{\hat{\sigma}_\tau^2 - (1 - \lambda)/2} \right]^2$$

and

$$\text{SBDH}^0 = \frac{\sum_{t=1}^T S_t^2}{\hat{\sigma}_\tau^2 T^2}$$

where

$$\lambda = \frac{\hat{\sigma}_y^2}{\hat{\sigma}_\tau^2}, \quad \hat{\sigma}_y^2 = \frac{\sum y_t^2}{T}$$

$$\hat{\sigma}_\tau^2 = \sum_{n=-\tau}^{\tau} \hat{c}(n) k(n/\tau), \quad \hat{c}(n) = \frac{\sum_{t=2}^{T-N} \Delta S_t \Delta S_{t+n}}{T}$$

and $k(n/\tau)$ is a lag window. $\hat{\sigma}_\tau^2/2\pi$ is the spectral density estimate of y_t at the zero frequency. These tests can, however, be used for zero mean processes only.

Choi and Yu show that under the null the asymptotic distribution of LM^0 and SBDH^0 are given by

$$\text{LM}^0 \Rightarrow \left[\int_0^1 W(r) dW(r) dr \right]^2$$

and

$$\text{SBDH}^0 \Rightarrow \int_0^1 W(r)^2 dr$$

They tabulate the empirical size and power of these tests by simulation. They also extend these tests to models with time trends and to tests for the hypotheses

$$I(1) \text{ versus } I(k), \quad k = 2$$

and

$$I(2) \text{ versus } I(k), \quad k = 3$$

The asymptotic distributions of the test statistics for these latter tests are more complicated. Using these tests one can test the null hypotheses $I(0)$, $I(1)$, and $I(2)$ sequentially.

Note, however, that there have been some objections raised in the literature on this procedure of *testing up* and it has been argued that *testing down* is a better approach (see Dickey and Pantula, 1987). This is similar to the issue of specific to general versus general to specific approach to dynamic econometric modeling, the latter advocated in the so-called LSE methodology.

Breitung (1994) suggests three simple tests of the MA unit root hypothesis. The first test makes use of the fact that in models with MA unit roots, the spectral density is close to zero at frequency zero. Thus, the MA unit root hypothesis can be tested by testing the significance of the spectral density at frequency zero. Breitung finds that this test is less powerful than the other tests for the MA unit root.

The second test is based on a comparison of the variance of the integrated series under the null and the alternative. Let $z_t = \sum_{j=1}^t y_j$ be the partial sum process for y_t . If y_t has an MA unit root, the variances of y_t and z_t are of the same order of magnitude, whereas if y_t does not have a MA unit root then the variance of the partial sum diverges at a rate of T . Thus, the variance difference (VD) statistic is based on the difference between the variance of the original and the integrated series. To integrate the process Breitung considers a slightly different concept than the partial sum method. For details, see Breitung (1994).

The third statistic he considers is the Tanaka (1990) statistic. This statistic essentially depends on a variance comparison of the double integrated series with a single integrated series.

Breitung finds that the powers of the three tests: spectral density test, VD test, and Tanaka's test, have powers converging at the rate of $T^{1/4}$, $T^{3/4}$, and T respectively. Results from his Monte Carlo experiments suggest that the Tanaka test performs best if the alternative is close to the null whereas the VD statistic performs better for more

substantial violations of the null. Breitung (1994, p. 361), however, suggests some useful modifications of the Tanaka test statistic.

We shall see in the next section how the Tanaka test statistic is related to other tests using stationary as null. Since, as argued at the beginning of this section, under the hypothesis of an MA unit root the series is stationary whereas it is nonstationary if the MA root θ is less than 1 in absolute value, *tests for the MA unit root as null and tests for stationarity as null are related*. We shall therefore look at tests of stationarity as null.

4.5 Tests with stationarity as null

There have been several tests for stationarity as null, although these are not as numerous as tests using unit AR root as null. Some of these are: Tanaka (1990), Park (1990), Kwiatkowski, Phillips, Schmidt, and Shin (1992), Saikkonen and Luukkonen (1993), Choi (1994), Leybourne and McCabe (1994), and Arellano and Pantula (1995). Most of these papers also present an analogy between tests for an MA unit root and tests for stationary as null, although the models they start with are slightly different from the ones we discussed in the preceding section. We shall now discuss these tests, though not in their chronological order.

4.5.1 KPSS test

Kwiatkowski, Phillips, Schmidt, and Shin (1992), which is often referred to as KPSS, start with the model

$$y_t = \delta t + \zeta_t + \varepsilon_t$$

where ε_t is a stationary process and ζ_t is a random walk given by

$$\zeta_t = \zeta_{t-1} + u_t, \quad u_t \sim iid(0, \sigma_u^2)$$

The null hypothesis of stationarity is formulated as

$$H_0 : \sigma_u^2 = 0 \text{ or } \zeta_t \text{ is a constant}$$

This is a special case of a test for parameter constancy against the alternative that the parameters follow a random walk. This problem was discussed by Nabeya and Tanaka (1988) for a regression model

$$\begin{aligned} y_t &= \beta_t x_t + \gamma' z_t + \varepsilon_t \\ \beta_t &= \beta_{t-1} + u_t \end{aligned}$$

Thus the KPSS model is a special case with $x_t = 1$ and $z_t = t$.

The Nabeya–Tanaka test statistic for this hypothesis is given by

$$LM = \frac{\sum_{t=1}^T S_t^2}{\hat{\sigma}_e^2}$$

where e_t are the residuals from the regression of y_t on a constant and a time trend, $\hat{\sigma}_e^2$ is the residual variance from this regression (residual sum of squares divided by T) and S_t is the partial sum of e_t defined by

$$S_t = \sum_{i=1}^T e_i, \quad t = 1, 2, \dots, T$$

For testing the null of level stationarity instead of trend stationarity the test is constructed the same way except that e_t is obtained as the residual from a regression of y_t on an intercept only. The test is an upper tail test.

The asymptotic distribution of the LM test statistic has been derived in Nabeya and Tanaka. However, this is valid only if the errors are *iid*. KPSS consider the case of a general error process and hence modify the test statistic as in Phillips (1987) and Phillips and Perron (1988). They then derive the asymptotic distribution of the modified statistic and tabulate the critical values by simulation.

When the errors are *iid* the denominator of the LM statistic $\hat{\sigma}_e^2$ converges to σ^2 . However, when the errors are not *iid* the appropriate denominator of the test statistic is an estimate of σ^2 not σ_e^2 , where σ^2 is the *long-run variance* defined by

$$\sigma^2 = \lim_{T \rightarrow \infty} T^{-1} E(S_T^2)$$

A consistent estimator of σ^2 is $\hat{s}_{Tl}^2$ which is given by

$$\hat{s}_{Tl}^2 = T^{-1} \sum_{t=1}^T e_t^2 + 2T^{-1} \sum_{\tau=1}^l w_{\tau l} \sum_{t=\tau+1}^T e_t e_{t-\tau}$$

Here $w_{\tau l}$ is an optimal weighting function that corresponds to the choice of a spectral window. KPSS use the Bartlett window, as suggested by Newey and West (1987)

$$w_{\tau l} = 1 - \frac{\tau}{l+1}$$

For consistency of $\hat{s}_{Tl}^2$, it is necessary that $l \rightarrow \infty$ as $T \rightarrow \infty$. The rate $l = o(T^{1/2})$ is usually satisfactory under both the null and the alternative (see section 3.6.3 for the Newey–West estimator).

Table 4.4. *Critical values for the KPSS test*

Tests	Critical values		
	0.10	0.05	0.01
Test A: Intercept only	0.347	0.463	0.739
Test B: Linear trend	0.119	0.146	0.216

Source: Kwiatkowski, Phillips, Schmidt,
and Shin (1992, table 1, p. 166).

Note, however, that the PP tests have been found to have low power (see section 4.2) and that the problem could be corrected by changing this spectral window (see section 4.3 for a discussion). Anyway, we shall proceed with the way KPSS formulated the test. The critical values for the KPSS test are given in table 4.4.

We shall comment later on the size distortions and power properties of the KPSS test.

4.5.2 *Leybourne and McCabe test*

Leybourne and McCabe (1994) suggest a modification of the KPSS test which can be viewed as an analogue of the ADF test whereas the KPSS test is an analogue of the Phillips–Perron test. They argue that inference from the KPSS test can be very sensitive to the value of the lag l that is used in the computation of s_{Tl}^2 , and that the alternative they suggest is more robust to lag specification (the lag being the order of the AR as in the ADF test) and also more powerful than the KPSS test. They show that their test is consistent to order T whereas the KPSS test is consistent to order T/l or $T^{2/3}$ since l is of order $T^{1/3}$. Also, in the test they suggest the null and alternative hypotheses are based on more realistic models than are encountered in practice.

The test suggested by Leybourne and McCabe is based on a generalization of the local level (or unobserved components) model (Harvey, 1989, chapter 2). The null hypothesis is that the process is a stationary $\text{ARIMA}(p, 0, 0)$ process and the alternative hypothesis is that it is an $\text{ARIMA}(p, 1, 1)$ process with a positive $\text{MA}(1)$ coefficient. The test involves choosing p but they argue that inference using a value of $p > p_0$ (the true value) does not have much of an effect on the inference.

The model is

$$\Phi(L)y_t = \alpha_t + \beta t + \varepsilon_t$$

$$\alpha_t = \alpha_{t-1} + \eta_t, \quad \alpha_0 = \alpha, \quad t = 1, 2, \dots, T$$

where $\varepsilon_t \sim iid(0, \sigma_\varepsilon^2)$, $\eta_t \sim iid(0, \sigma_\eta^2)$, ε_t and η_t are independent, and $\Phi(L) = 1 - \phi_1 L - \phi_2 L^2 - \dots - \phi_p L^p$ is a p th order AR polynomial with roots outside the unit circle. They call this the *structural model*. This model can be shown to be second-order equivalent to the ARIMA($p, 1, 1$) process, which will be referred to as the *reduced-form model*

$$\Phi(L)(1 - L)y_t = \beta + (1 - \theta L)u_t, \quad 0 < \theta < 1$$

where $u_t \sim iid(0, \sigma_u^2)$ and $\sigma_u^2 = \sigma_\varepsilon^2 / \theta$. Define $\lambda = \sigma_\eta^2 / \sigma_\varepsilon^2$. Then θ is related to σ_η^2 by the relation

$$\theta = \frac{\lambda + 2 - (\lambda^2 + 4\lambda)^{1/2}}{2}$$

Thus if $\sigma_\eta^2 = 0$, then $\theta = 1$, and the model collapses to a stationary AR(p) process. If σ_η^2 is very close to zero, then θ is very close to 1, and y_t is almost stationary.

The test for stationarity in this model is

$$H_0 : \sigma_\eta^2 = 0 \text{ versus } H_1 : \sigma_\eta^2 > 0$$

Under H_1 the model is an ARIMA model with an MA characteristic that closely resembles many economic time series, as argued by Schwert (1987). This is a more realistic alternative than a pure random walk with *iid* errors. In fact, the condition $\theta > 0$ excludes such processes. Another feature of the model is that p , the order of the AR component, is > 0 . This ensures that the short-run and long-run effects of η_t are different which is a realistic assumption.

Leybourne and McCabe deal with the structural model because the inferential procedures fall in the category of a linear model with an intercept that follows a random walk. The earlier work by Leybourne and McCabe (1989) suggests a test statistic of the form $\varepsilon'V\varepsilon$ where ε is a $T \times 1$ vector with the t th element ε_t and V is a $T \times T$ matrix with the (i, j) th element equal to the minimum of i and j . The procedure they suggest for constructing $\hat{\varepsilon}_t$ consists of the following steps:

- (i) Estimate the ARIMA($p, 1, 1$) model by ML

$$\Delta y_t = \beta + \sum_{i=1}^p \phi_i \Delta y_{t-i} + u_t - \theta u_{t-1}$$

to get $\hat{\phi}_i$.

(ii) Construct y_t^* as

$$y_t^* = y_t - \sum_{i=1}^p \hat{\phi}_i y_{t-i}$$

(iii) Regress y_t^* on an intercept and a time trend to get the residuals $\hat{\varepsilon}_t$.

Then the test statistic is

$$\hat{s}_\beta = \hat{\sigma}_\varepsilon^{-2} T^{-2} \hat{\varepsilon}' V \hat{\varepsilon}$$

where $\hat{\sigma}_\varepsilon^2 = \hat{\varepsilon}' \hat{\varepsilon} / T$ is a consistent estimator of σ_ε^2 .

In the case of $\beta = 0$, $\hat{\varepsilon}_t$ is a residual from a regression of y_t^* on intercept alone. The corresponding test statistic is denoted as $\hat{s}_\alpha$. Leybourne and McCabe show that the asymptotic distribution of $\hat{s}_\alpha$ and $\hat{s}_\beta$ are the same as the corresponding statistics derived by KPSS and thus one can use the critical values from the KPSS paper (tabulated earlier). The important difference between the two tests is that the test by Leybourne and McCabe accounts for autocorrelation in a parametric fashion by including the lagged term in y_t in the initial model specification. Thus, the test is an analogue of the ADF test. The KPSS test, on the other hand, starts with the basic model and then modifies the test statistic nonparametrically by changing the estimate of σ^2 to $\tilde{s}_{Tl}^2$, the same way that the PP tests adjust the DF tests.

Leybourne and McCabe apply their tests to make inferences on unit roots using the 11 US macroeconomic time series studied by Schwert (1987). In three cases (PI, PPI, and Wage), the KPSS test did not reject stationarity, whereas the test by Leybourne and McCabe showed unambiguously that these series contained unit roots.

4.5.3 Some other tests

Park (1990) suggests tests based on the idea that if a variable follows a unit root process, then OLS standard errors are usually inappropriate and tend to indicate that unrelated variables have a statistically significant relationship. His test statistics add one or more spurious variables to the regression of y_t on a constant and trend, and test the significance of these spurious variables. His J test is defined by

$$J = \frac{RSS_1 - RSS_2}{\tilde{s}_{Tl}^2}$$

where RSS_1 is the residual sum of squares from a regression of y_t on a constant and trend, RSS_2 is the residual sum of squares with the superfluous regressors added, and $\hat{s}_{Tl}^2$ is the Newey–West estimator, defined earlier in the KPSS test. Under the null hypothesis, this statistic has a χ^2 distribution with d.f. equal to the number of superfluous regressors. The choice of the superfluous regressors is arbitrary. Some of the candidates are polynomial trends and pseudo-random walks. The power of the test appears to vary greatly but using two or more variables seem to give better discriminatory power than using just one.

Amano and van Norden (1992) compare the size and power of the KPSS test and Park's test and find that, as with conventional unit root tests, these tests also suffer from size distortion and loss of power for certain data generation processes. Both tests also appear sensitive to the specification of the truncation parameter.

Bierens and Guo (1993) suggest some tests which are based on the Cauchy distribution. They argue that these tests have better asymptotic properties than Park's test. These test statistics, however, have not been used in practice.

Choi (1994) provides a test for stationarity within the framework of testing for an MA unit root. We have discussed the tests proposed by Choi and Yu (1997) in the previous section and so will not go through this test in detail.

Arellano and Pantula (1995) also suggest a test based on testing for an MA unit root. They note the *pile-up problem* that we discussed in the previous section and also the fact that the distribution of the MLE of the moving-average parameter θ when $\theta = -1$ is not known. They also note that when analyzing simulated data with noninvertible MA processes that the SAS program gave ML estimates that were unstable and the standard errors were unrealistic. Arellano and Pantula, instead consider the one-step Gauss–Newton estimator of θ starting with an initial value $\theta = -1$, and derive the asymptotic distribution of the test statistics suggested. Since there is no comparison with the tests we have described in the econometric literature and also the illustrative example is not of interest to us, we shall not go into details of these tests here.

4.5.4 Some general comments

There is no comparison of all these tests. However, it appears that the most fruitful approach is to start with the unobserved components

model as in Leybourne and McCabe. These tests for stationarity have been suggested as useful for two purposes:

- (i) for confirmatory analysis – this we discuss in the next section and
- (ii) as tests where it makes sense to test stationarity as null (e.g., in discussion of convergence in the literature on growth).

As for the problem of deciding between $I(0)$ and $I(1)$ the most fruitful approach is to analyze it in the Bayesian framework. This we shall discuss in chapter 8.

4.6 Confirmatory analysis

It has been suggested (see, e.g., KPSS, p. 176 and Choi, 1994, p. 721) that the tests using stationarity as null can be used for confirmatory analysis, i.e., to confirm our conclusions about unit roots. However, if both tests fail to reject the respective nulls or both reject the respective nulls, we do not have a confirmation. The situation is similar to the tests of nonnested hypotheses.

There is also the problem of which tests to use for each of the hypotheses: unit root null and stationarity as null. As we noted in the previous section, the KPSS and Leybourne–McCabe tests gave conflicting results on the test for stationarity as null for three of 11 series.

In spite of these limitations, it is generally agreed that using both tests together is better than using either test alone. Amano and van Norden (1992) perform a Monte Carlo study and conclude that the joint testing approach gives the most reliable results when the joint test indicates that the data are stationary or that the data have a unit root for small samples and large k (the truncation lag). They used the KPSS–PP combination but the conclusions were basically unaltered by the KPSS–ADF combination. Note that the PP test they considered was the original PP test, and not the one with modifications noted in section 4.4 earlier.

Henricsson and Lundbäck (1995) apply this confirmatory analysis to a study of purchasing power parity (PPP) theory and conclude that the absence of PPP cannot be rejected (using the KPSS test) and the presence of PPP can be rejected (using the DF test). This, they argue, is confirmation that PPP does not hold and that the rejection of PPP using the DF test is not a consequence of the low power of unit root tests. We shall discuss this issue of testing PPP later when we discuss panel data unit root tests.

KPSS (p. 175) also present results from confirmatory analysis applied

to 14 time series (the Nelson–Plosser data analyzed by several investigators). A rough summary of these findings is that for 12 of the 14 series (except unemployment rate and industrial production) we cannot reject the null hypothesis of a unit root. With the KPSS test, we cannot reject the null of trend stationarity at the usual critical levels for six series: real per capita GNP, employment, unemployment rate, GNP deflator, wages, and money. We might note that some of these results are strange because many other studies have actually argued (as we will note in a subsequent chapter) that money and prices are $I(2)$. Also, Leybourne and McCabe (although they investigate a different series, that analyzed by Schwert) came up with the conclusion that monetary base, CPI, and wages show clear evidence of unit roots (the null of stationarity is rejected).

In any case KPSS conclude on the basis of the KPSS test and the Dickey–Fuller tests that there is confirmation in the case of four series: unemployment which is stationary and consumer prices, real wages, velocity, and stock prices which all have unit roots. Three more series: real GNP, nominal GNP, and the interest rate, probably have unit roots, though the evidence against the trend-stationary hypothesis is only marginally significant. (The result on interest rates is puzzling. As John Cochrane remarked: “Interest rates now are the same as in Babylonian days. How can there be a unit root in interest rates?”)

Finally, for six series – real per capita GNP, employment, unemployment rate, GNP deflator, wages, and money – it is not possible to reject either the unit root hypothesis or the trend-stationary hypothesis. KPSS conclude that the data are not sufficiently informative to distinguish between these hypotheses. In the case of one series, industrial production, there is evidence against both the hypotheses and so KPSS argue that it is not clear what to conclude. Many of the conclusions from the confirmatory analysis of KPSS are contradictory to commonsense. What this suggests is that it is important to use better tests of both the hypotheses in confirmatory analysis.

Burke (1994) does a detailed Monte Carlo study to determine the usefulness of CDA (confirmatory data analysis). He investigates several important issues through simulation studies, choosing model structures similar to macroeconomic time series in the US. For the unit root test he selects the ADF test with lag selection based on the AIC criterion. For the stationary tests he uses the KPSS test and the SBDH tests in Choi and Yu that we discussed in section 4.4. (He investigated Park’s

test but finds it less useful for CDA.) The major issues investigated in the study are:

- (i) How often does confirmation occur?
- (ii) How many of these confirmations are correct?
- (iii) How do these results change if the size of the tests is increased from 5 percent to 10 percent.

Regarding issue (iii) he arrives at the conclusions that using the 10 percent significant level gives better results than using the 5 percent significance level. It is not possible to summarize the numerous results in Burke's paper. The important conclusions are:

- (i) Joint rejections are relatively infrequent, but joint nonrejections are far more common.
- (ii) Even if confirmation occurs this may not be correct.

For instance, for a stationary model, the stationarity test gave correct inference on 70 percent of occasions, the unit root test on only 43 percent of occasions. Using CDA there was confirmation on only 50 percent of cases and there was a 61 percent chance of the confirmation to be correct. Under nonstationarity, the situation was better. In the corresponding case to that described earlier, the stationarity test gave correct inference in 71 percent of occasions, the unit root test on 81 percent. CDA allowed inference in 73 percent of cases, with a 91 percent chance that this inference is correct. In an experiment with trend- and difference-stationary models considered by Rudebusch (1992), Burke found that, when the true model is trend stationary, the proportion of confirmations is 50–60 percent and about half of these are correct. When the true model is difference-stationary, the proportion of confirmations is 60–65 percent of which about 82 percent are correct.

The overall conclusion is that if the true model is stationary, the proportion of correct confirmations is low. It is thus, more important to consider better unit root tests and stationary tests (as discussed in section 4.3–4.5) than to use confirmatory analysis with defective tests.

The arbitrary choice of significance levels for the respective tests is also a point of concern. This problem can be mitigated with the Bayesian approach discussed in chapter 9.

4.7 Frequency of observations and power of unit root tests

Are unit root tests based on quarterly data more powerful than those based on the corresponding yearly data? Shiller and Perron (1985) and Perron (1989) find using Monte Carlo experiments that over a substantial range of values, power depends more on the span of the data rather than on the number of observations (see Shiller and Perron, p. 381). Perron (1991) showed analytically that with fixed alternatives tests for a unit root are only consistent when the time span rises with the number of observations. Diebold *et al.* (1991) also argue that long data spans are important for identifying mean reversion in slowly decaying processes.

When we consider time aggregation and skip sampling problems we need to keep in mind the distinction between *flow* data and *stock* data. If we are considering skip sampling at intervals of m periods, with a stock variable y_t , and sample values y_t^* , we have

$$y_t^* = y_t \text{ for } t = m, 2m, 3m, \dots$$

In the case of a flow variable, however, it is a problem of time aggregation and we have

$$y_t^* = (1 + L + L^2 + \dots + L^{m-1})y_t$$

Interest rates and currency rates fall in the former category whereas consumption and GNP fall in the latter category. Shiller and Perron (1985) and Perron (1989) discuss the skip sampling problem with stock data. On the other hand, Choi (1992b) considers the problem of time aggregation that is relevant for flow data. He finds, in his simulation study, that using the data generated by aggregating subinterval data results in lower powers of unit root tests. Thus, using quarterly data is better than using annual data, and using monthly data is better than using quarterly data. He also finds that for the aggregated data, the PP tests are more powerful than the ADF test. Note that time aggregation produces a MA residual in the aggregated series. For example, if the quarterly model is

$$y_t = \rho y_{t-1} + u_t, \quad t = 1, 2, \dots, T$$

then the yearly model is

$$x_s = \rho^4 x_{s-1} + v_s, \quad s = 1, 2, \dots, T/4$$

where

$$x_s = y_{4,s} + y_{4,s-1} + y_{4,s-2} + y_{4,s-3}$$

and v_s is a moving-average of u_t .

Choi starts with a quarterly AR(1) model to generate the data on y_t . He then generates the yearly data using time aggregation and applies unit root tests to the quarterly data and yearly data. The sampling frequency is thus fixed.

Ng (1995) also does a Monte Carlo study starting from a continuous time model. She studies the effect of varying the sampling frequency as well as the total time span S of the data. She finds that with flow data:

- (i) Increasing the frequency of observation while keeping S fixed increases power but at a diminishing rate.
- (ii) Power decreases if the frequency of observation is increased but S is decreased.
- (iii) Power increases if S is increased even if the total number of observation is kept constant (by reducing the frequency of observation).

Since time aggregation results in MA errors one would think that IV tests would be appropriate but the power functions of the IV test are not monotonic as S increases. With the ADF test and Phillips Z_α and Z_t tests power increases with increases in S .

Overall the conclusion on time aggregation with flow data is that increasing the time span S increases the power of the commonly used ADF and PP tests (though not of the IV tests) but increasing the frequency of observation with S fixed also increases power, i.e., using quarterly rather than yearly data does help in increasing power. With skip sampling the results in Shiller and Perron and Perron suggest that increasing the frequency of observation with the time span S fixed does not increase the power of unit root tests. However, Choi and Chung (1995) re-examine these studies with the ADF and PP tests (not examined earlier by Shiller and Perron and Perron). They also consider the null of a unit root and a unit root with drift (not considered earlier) and also relax the assumption of taking the critical value as zero. The results now show that with the ADF tests, using data with high sampling frequency can provide significant improvement in finite sample power. But in the case of PP tests sampling frequencies do not significantly affect the power of the tests. For low frequency data, the PP test appears to be more powerful than the ADF test.

Thus, with skip sampling too, there is an improvement in power by using the frequency of observations if one uses the ADF test. Thus, using quarterly data is better than using yearly data, and monthly data than quarterly data. As an empirical example, Papell (1997) finds that

the evidence against the unit root in real exchange rate data is stronger for monthly than for quarterly data. The data are on exchange rates (not flow data) and thus fall into the category of skip sampling.

4.8 Other types of nonstationarity

The preceding discussion was in terms of $I(d)$ models where d was assumed to take on the values 1 or 0, the model with $d = 1$ being the model with persistence of shocks. However, persistence of shocks can also be modeled with $d > 0$ but < 1 . This problem of *fractional unit roots* is discussed in chapter 9.

Another cause for nonstationarity is the *randomness of the coefficients* of autoregressions. Some references in this area are Cramer (1991), Nicholls and Quinn (1982), McCabe and Tremayne (1995), and Leybourne, McCabe, and Tremayne (1996).

There has also been some discussion about some economic variables being $I(2)$, that is, having *double unit roots*. Examples of this are money supply and prices. Hence questions arise as to how to test for more than one unit root. There are two approaches: the first, a *bottom-up* approach is the one discussed in Choi and Yu (1997) suggests testing sequentially: $I(0)$ versus $I(1)$ and $I(1)$ versus $I(2)$ if the first hypothesis is rejected. The second approach suggested in Dickey and Pantula (1987) is a *top-down* approach. First test $I(2)$ versus $I(1)$ and if the hypothesis of $I(2)$ is rejected, then test $I(1)$ versus $I(0)$. This procedure is appropriate for tests of nonstationarity as null, whereas the bottom-up approach is appropriate for tests of stationarity as null. As discussed earlier, we can also think of using both the tests together for *confirmatory analysis* but we shall not pursue this here.

Dickey and Pantula consider the case of three unit roots which is rarely the case with economic variables. But their procedure can be easily adapted for tests of double unit roots. For completeness, we shall review their procedure here.

The standard DF unit root tests assume that there is at most one unit root in the series. Dickey and Pantula (1987) proposed t^* - and F -tests that compare a null hypothesis of k unit roots with an alternative of $k - 1$ unit roots. Consider

$$y_t = \sum_{j=1}^p \beta_j y_{t-j} + e_t$$

where $\{e_t\}$ is a sequence of *iid* random variables with mean 0 and vari-

ance $\sigma^2 = 1$ and $y_{-p+1} = \dots = y_0 = 0$. They considered the AR model with $p = 3$ and three possible unit roots. Let m_1, m_2 , and m_3 denote the roots of the characteristic equation

$$m^3 - \beta_1 m^2 - \beta_2 m - \beta_3 = 0$$

The above AR(3) model can be written as

$$x_t = \theta_1 y_{t-1} + \theta_2 z_{t-1} + \theta_3 w_{t-1} + e_t$$

where $z_t = y_t - y_{t-1}$, $w_t = z_t - z_{t-1}$ and $x_t = w_t - w_{t-1}$. Note that z_t, w_t , and x_t are the first-, second-, and third-differences of the process y_t , respectively. Consider the hypotheses

(i) no unit root

$$H_0 : |m_1| < 1$$

or

$$H_0 : \theta_1 < 0 \text{ and some restrictions on } \theta_2 \text{ and } \theta_3$$

For example, the restrictions on θ_2 and θ_3 are

$$-12 < \theta_2 + 2\theta_1 < 0, -2 < \theta_3 < 0$$

(ii) one unit root

$$H_1 : m_1 = 1, |m_2| < 1$$

or

$$H_0 : \theta_1 = 0, \theta_2 < 0 \text{ and some restrictions on } \theta_2 \text{ and } \theta_3$$

The restrictions on θ_2 and θ_3 are $0 < 4 + \theta_2 + 2\theta_3, -2 < \theta_3 < 0$.

(iii) two unit roots

$$H_2 : m_1 = m_2 = 1, |m_3| < 1$$

or

$$H_0 : \theta_1 = \theta_2 = 0, \theta_3 < 0$$

(iv) three unit roots

$$H_3 : m_1 = m_2 = m_3 = 1$$

or

$$H_0 : \theta_1 = \theta_2 = \theta_3 = 0$$

Dickey and Pantula showed that applying F -statistics for testing from the higher number of unit roots to the lower number of unit roots is valid, while the reverse order of applying F -statistics is not valid. Based on the asymptotic distributions of F -statistics derived by Pantula (1986), they suggest testing the hypotheses sequentially in the order H_3, H_2 , and H_1 :

- (i) If H_3 is rejected by the F -test, then go to step (ii); otherwise conclude that H_3 is true.
- (ii) If H_2 is rejected by the F -test, then go to step (iii); otherwise conclude that H_2 is true.
- (iii) If H_1 is rejected by the F -test, then conclude that H_0 is true; otherwise conclude that H_1 is true.

The empirical percentiles of the asymptotic distributions of the F -statistics can be found in Pantula (1986).

Dickey and Pantula proposed alternative testing procedures based on t -test statistics. However, the t -statistics from the regression of x_t on y_{t-1}, z_{t-1} , and w_{t-1} has different asymptotic distributions depending on the number of unit roots present. And thus they argue that a sequential procedure based on these statistics is not consistent. They suggest to use alternative t^* -statistics:

- (i) for H_3 against H_2 , obtain the t^* -statistic from the regression x_t on w_{t-1} ,
- (ii) for H_2 against H_1 , use the t^* -statistic in the regression x_t on z_{t-1} and w_{t-1} for testing the coefficient of z_{t-1} is 0,
- (iii) for H_1 against H_0 , use the t^* -statistic in the regression x_t on y_{t-1}, z_{t-1} and w_{t-1} for testing the coefficient of y_{t-1} is 0.

Their results of a Monte Carlo power study show that the procedure based on t^* is more powerful than that based on F in most cases.

4.9 Panel data unit root tests

The principle motivation behind panel data unit root tests is to increase the power of unit root tests by increasing the sample size. An alternative route of increasing the sample size by using long time series data, it is argued, causes problems arising from structural changes. However, it is not clear whether this is more of a problem than cross-sectional heterogeneity, a problem with the use of panel data.

It is often argued that the commonly used unit root tests such as ADF,

Z_α , and Z_t are not very powerful, and that using panel data you get a more powerful test. For instance, in the case of tests for the purchasing power parity (PPP) hypothesis, one can test the hypothesis that the real exchange rate has a unit root. A nonrejection of this hypothesis implies that PPP does not hold, even in the long run. Several studies fail to reject the null of unit root, whereas panel data unit root tests as in Oh (1996) and Wu (1996) reject the null. However, an alternative solution to the low power problem is to use tests that have been demonstrated to have better power. Cheung and Lai (1995) use the test developed by Elliott *et al.* (1996) and find greater support for the PPP hypothesis. In the following sections, we shall review the different panel data unit root tests and then argue that they test a different hypothesis and thus, do not really solve the problem of low power which is the basic motivation in the first place. We shall also suggest a simpler test due to Fisher (1932).

4.9.1 The different panel data unit root tests

Breitung and Meyer (1994) (to be denoted BM) suggest a simple panel data unit root test valid for fixed T and $N \rightarrow \infty$. They consider the model

$$y_{i,t} = \alpha y_{i,t-1} + (1 - \alpha)\mu_i + \varepsilon_{i,t}$$

$$\varepsilon_{i,t} \sim iin(0, \sigma^2), \quad i = 1, 2, \dots, N, t = 1, 2, \dots, T$$

Regressing $y_{i,t}$ on $y_{i,t-1}$ ignoring μ_i gives an estimate $\hat{\alpha}$ with asymptotic bias

$$plim_{N \rightarrow \infty}(\hat{\alpha} - \alpha) = \frac{(1 - \alpha)S^2}{S^2 + \sigma^2/(1 - \alpha^2)}$$

where $S^2 = N^{-1} \sum \mu_i^2$. When $\alpha = 1$, the bias vanishes. Hence the unit root hypothesis can be tested using the t -statistic for $H_0 : \alpha = 1$. However, since under the alternative $|\alpha| < 1$ the OLS estimate $\hat{\alpha}$ is biased, the test leads to a loss of power. BM suggest estimating the equation

$$(y_{i,t} - y_{i,0}) = \alpha(y_{i,t-1} - y_{i,0}) + u_{i,t}$$

Denoting the estimator by $\tilde{\alpha}$ they show that

$$plim_{N \rightarrow \infty} \tilde{\alpha} = \frac{1}{2}(\alpha + 1)$$

Again, under the unit root hypothesis the bias disappears. BM suggest other modifications to handle higher-order autoregressions, time trends, and time effects.

As explained in Levin and Lin (1993a, 1993b), the approach taken by BM cannot be used to analyze the influence of individual specific effects or serial correlation on the appropriate critical values at which to evaluate the t -tests. Also, it cannot be extended to allow for heterogeneous error distributions. For these reasons there are not many empirical applications of these tests.

Quah (1994) considers the simple model

$$y_{i,t} = \alpha y_{i,t-1} + \varepsilon_{i,t}, \quad \varepsilon_{it} \sim iid(0, \sigma^2)$$

and considering testing $\alpha = 1$. Quah (1992) considers the model

$$y_{i,t} = \mu + \alpha y_{i,t-1} + \varepsilon_{i,t}, \quad \varepsilon_{it} \sim iid(0, \sigma^2)$$

and testing $\alpha = 1$. These cases are not very interesting from the empirical point of view. As pointed out by Levin and Lin, Quah's methodology cannot be extended to the case of individual and time specific effects.

Levin and Lin (1993a) conduct an exhaustive study and develop unit root tests for the model

$$y_{i,t} = \alpha y_{i,t-1} + \delta_0 + \delta_1 t + \eta_i + v_t + \varepsilon_{i,t}, \quad \varepsilon_{it} \sim iid(0, \sigma^2)$$

The model incorporates a time trend as well as individual and time specific effects. Initially, they assume the *iid* errors, but they showed that under the serial correlation of errors the test statistics have the same limiting distributions by the inclusion of lagged first-differences of $y_{i,t}$ like the parametric correction of the ADF test.

Levin and Lin consider the following six models

- (i) $y_{i,t} = \alpha y_{i,t-1} + \varepsilon_{i,t}, \quad H_0 : \alpha = 1$
- (ii) $y_{i,t} = \alpha y_{i,t-1} + \delta_0 + \varepsilon_{i,t}, \quad H_0 : \alpha = 1, \delta_0 = 0$
- (iii) $y_{i,t} = \alpha y_{i,t-1} + \delta_0 + \delta_1 t + \varepsilon_{i,t}, \quad H_0 : \alpha = 1, \delta_1 = 0$
- (iv) $y_{i,t} = \alpha y_{i,t-1} + v_t + \varepsilon_{i,t}, \quad H_0 : \alpha = 1$
- (v) $y_{i,t} = \alpha y_{i,t-1} + \eta_i + \varepsilon_{i,t}, \quad H_0 : \alpha = 1, \eta_i = 0 \text{ for all } i$
- (vi) $y_{i,t} = \alpha y_{i,t-1} + \eta_{i,0} + \eta_{i,1} t + \varepsilon_{i,t}$

For models (i) to (iv), they show that

$$T\sqrt{N}(\hat{\alpha} - 1) \Rightarrow N(0, 2)$$

$$t_\alpha \Rightarrow N(0, 1)$$

For model (v), if $\sqrt{N}/T \rightarrow 0$, then

$$T\sqrt{N}(\hat{\alpha} - 1) + 3\sqrt{N} \Rightarrow N(0, 10.2)$$

$$\sqrt{1.25}t_{\alpha} + \sqrt{1.875N} \Rightarrow N(0, 845/112)$$

In model (vi), both intercept and time trend vary with individuals.

In the empirical applications, Oh(1996) uses only models (i) to (v). Wu (1996) uses the complete model with trend, and individual and time specific effects but uses the distributions derived for model (v). Pappell (1997) uses model (v) with lagged first-differences of $y_{i,t}$ added but computes his own exact finite sample critical values using Monte Carlo methods and finds them 3–15 percent higher than those tabulated in Levin and Lin (1993a).

Levin and Lin argue that in contrast to the standard distributions of unit root test statistics for a single time series, the panel test statistics have limiting normal distributions. However, the convergence rates are faster as $T \rightarrow \infty$ (superconsistency) than as $N \rightarrow \infty$. The paper by Levin and Lin (1993b) provides some new results on panel data unit root tests. However, since the models discussed are the same, we shall not go into details.

The major limitation of the Levin–Lin tests is that α is the same for all observations. Thus, if we denote by α_i the value of α for the i th cross-section unit then the Levin–Lin test specifies the null H_0 and alternative H_1 as

$$H_0 : \alpha_1 = \alpha_2 = \cdots = \alpha_N = \alpha = 1$$

$$H_1 : \alpha_1 = \alpha_2 = \cdots = \alpha_N = \alpha < 1$$

Im, Pesaran, and Shin (1996, to be denoted IPS) relax the assumption that $\alpha_1 = \alpha_2 = \cdots = \alpha_N$ under H_1 . The basic idea of the test is simple. Take the model (iv) in Levin and Lin and substitute α_i for α . Essentially what we have is a model with a linear trend for each of the N cross-section units. Thus, instead of pooling the data, we use separate unit root tests for the N cross-section units. Consider the t -test for each cross-section unit based on T observations. Let $t_i, i = 1, 2, \dots, N$ denote the t -statistics for testing unit roots, and let $E(t_i) = \mu$ and $\text{var}(t_i) = \sigma^2$. Then

$$\sqrt{N} \frac{t - \mu}{\sigma} \Rightarrow N(0, 1)$$

The problem is computing μ and σ^2 . This they do by Monte Carlo methods and tabulate them for ready references (table A, B, C of their paper).

The important thing to note is that the IPS test is a way of combining the evidence on the unit root hypothesis from the N unit root tests performed on the N cross-section units. Note that implicit in the test is the assumption that T is the same for all cross-section units and hence $E(t_i)$ is the same for all i . Thus, we are considering only balanced panel data.

R. A. Fisher (1932) suggested a method of combining the evidence from several independent tests. Suppose there are N unit root tests as in IPS. Let P_i be the observed significance level (p -value) for the i th test. The $(-2 \sum \log P_i)$ has a χ^2 distribution with d.f. $2N$. This is an exact distribution. This test often known as the P_λ test mentioned in Maddala (1977, section 4.12) has not received much attention.

The advantage of P_λ test is that it does not require a balanced panel as in the case of the IPS test. The disadvantage is the the p -value has to be derived by Monte Carlo methods. Another advantage is that it can also be carried out for any unit root test. The IPS test is easy to use because there are ready tables available in the paper by Im *et al.* (1996) for $E(t_i)$ and $var(t_i)$. However, these are valid only for the ADF test.

The paper by Maddala and Liu (1996) presents the results on the different panel data unit root tests based on real exchange rate data for 131 countries divided into six regional groups. The results are different for the different tests. In general the Fisher test rejects the unit root null more often than the others. It is difficult to judge which of these tests is better, without a Monte Carlo study comparing their performance.

Maddala and Wu (1996) compare the powers of the LL, IPS, and Fisher tests. They find that in a variety of situations the Fisher test is more powerful than the IPS test which in turn is more powerful than the LL tests.

One drawback of both the IPS and Fisher tests is that they depend on the assumption that there is no cross-country correlation among the errors. This assumption is almost always violated in practice. Maddala and Wu find that the Fisher test is more robust than the IPS test to the violation of this assumption. They suggest bootstrap methods (see chapter 10) to obtain the critical values for the Fisher test.

4.9.2 The hypotheses of interest

Whether it is the use of panel data unit root tests in investigations of PPP or growth and convergence, the first question is: what are the hypotheses of interest? Consider the case of PPP and test of the long-run validity of the PPP hypothesis in the form of testing for a unit root in the real exchange rate. One may be interested in testing whether the hypothesis of the PPP holds for the Japanese Yen against US dollar (JY/US) exchange rate. In this case, what is of relevance are the data on this exchange rate. It is no use to be told that we reject the validity of the PPP even in the long run for the JY/US but that if we throw in a large number of countries and use the panel data unit root test, we do not reject the PPP hypothesis for the JY/US exchange rate. The hypotheses under consideration in the two cases are different. This is the major problem with the studies such as those of Oh (1996), McDonald (1996), and Wu (1996).

On the other hand, one may be interested not in the validity of the PPP for any particular exchange rate but as a general hypothesis. One has the results of unit root tests for a number of exchange rates and the null hypothesis of a unit root is rejected at different significance levels for the different exchange rates. What one needs is a summary picture of the conflicting evidence. The question is how does one do it? The panel data unit root test of Levin and Lin does not help us answer this question, the IPS test does, but under the limitations mentioned earlier.

A more important issue is not a test of the hypothesis of the validity of the PPP but an estimate of the time it takes for deviations from the PPP to correct themselves. We may be interested in this for each particular exchange rate or as a general summary for all the exchange rates. For this purpose the appropriate procedure is to use the panel data to get improved estimates for the autoregressive parameter in the equation for each of the real exchange rates. This issue has been discussed in Maddala (1991) and Maddala and Hu (1995). The procedure is based on Bayesian and empirical Bayesian methods discussed in the statistical literature.

Again, one can talk about the time it takes for deviations from PPP to correct themselves – for each exchange rate or for a set of exchange rates. The question is: what is a summary statistic? The answer is that it is not a simple average of the time for each exchange rate but a weighted average, or alternatively a simple average of the shrinkage estimates obtained from the panel data. In any case panel data unit

root tests do not provide answers to all these problems. Thus, although theoretically appealing, they are of limited value in practice.

4.10 Uncertain unit roots and the pre-testing problem

The unit root tests that we have discussed are always a prelude to further inference. In this researchers have to face the problem of uncertainty about the unit root. As Cochrane (1991) argued the important question is not whether there exists a unit root and to classify time series into the unit root or no unit root categories but to outline the appropriate inferential procedures (confidence interval statements and so on). The confidence intervals are different under the unit root and no unit root cases. Under the unit root case we use the nonstandard distributions and in the no unit root case we use the normal distributions. Which of these should researchers apply? This problem does not arise in the Bayesian framework, as we shall discuss in chapter 8. See also Stock (1994a), who suggests a Bayesian-type approach. For confidence interval statement, see Stock (1991).

Watson (1994, p. 2568) discusses a conservative inference which he argues is widely used in applications. The procedure is as follows: let W be the statistic of interest. Let C_U and C_N be the critical values under the unit root (U) and no unit root (N) hypotheses, i.e., $P(W > C_U|U) = P(W > C_N|N) = \alpha$. Then reject the unit root hypothesis if $W > \max(C_U, C_N)$. Do not reject it if $W < \min(C_U, C_N)$. However, a problem arises if $\min(C_U, C_N) < W < \max(C_U, C_N)$. In this case he says you have to look at the data and see which hypothesis is plausible. By construction, the size of the test is $\leq \alpha$.

This procedure, however, is just another procedure to classify the time series into the unit root or no unit root categories as the procedures we discussed earlier. It should therefore be compared with say the confirmatory analysis we discussed earlier. It does not address the problem of taking the uncertainty about unit roots into account in further inference like confidence interval statements. Under this procedure we use the standard normal distribution if we reject the unit root hypothesis and the nonstandard distribution if we do not. The pre-testing problem still remains.

The properties of this pre-test estimator from a decision theoretic point of view are discussed in Lahiri and Paul (1996). Based on the analysis of risk functions, Lahiri and Paul conclude that OLS is a better choice for almost all economic time series following the AR(1) model,

unless we have a strong belief that the series behaves like stock price series, i.e., the autoregressive parameter ρ is equal to 1. In this case nonstandard inference is used. The pre-test estimator was never found to be optimal.

In the statistical literature, there is considerable discussion of the pre-testing problem but the only consensus that has come out all this is that one should not use the usual 5 percent significance levels but one should use higher (say 25 percent) significance levels in pre-tests. A good discussion of this is in the series of papers in *Communications in Statistics, A: Theory and Methods*. 1976, under the title: "For What Use Are Tests of Hypotheses and Tests of Significance?"

The pre-testing problem is all the more important in the discussion of estimation and tests of cointegration relationship (chapters 5 and 6). They all start with the hypothesis that the variables under consideration are $I(1)$. We shall discuss the consequences of the violation of this assumption in the forthcoming chapters. Given that all this analysis depends on a unit root pre-test, the question of what significant levels to use for unit root tests, if these tests are a prelude to cointegration analysis, is a very important one for which there is no definite answer yet.

4.11 Other unit root tests

Besides the commonly used ADF and PP tests, there are many other unit root tests in the literature, but these have not been used much in practice.

Burridge and Guerre (1996) observe that the number of level crossings in a variable would be larger if there was a unit root than there would be otherwise. They use this to derive a unit root test. However, they find that this nonparametric test is more sensitive to the shape of the distribution of the errors than the DF tests.

Bierens (1993) devises unit root tests based on higher-order sample autocorrelations. Box and Jenkins observe that for a stationary process the sample autocorrelation function tails off quickly. In practice, it happens quite often that unit root tests do not reject the null hypothesis of a unit root although the sample autocorrelation function tails off rather quickly. Bierens argues that the Box-Jenkins approach to determining the degree of differencing on the basis of the behavior of the sample autocorrelation function is misleading.

Abadir (1995b) notes that nonstationarity causes the limiting distri-

butions of the Wald (W) and Lagrange multiplier (LM) statistics to become different from each other. He uses this divergence to devise a new test for unit roots.

None of these tests have found much applicability. Note that they all test the unit root hypothesis by looking at different consequences of unit roots. The variance ratio tests discussed earlier in section 3.8 also fall in this category but they have been used often in the literature on macroeconomics and finance.

4.12 Median-unbiased estimation

Consider the AR(1) model with drift and trend as follows

$$y_t = \mu + \beta t + \alpha y_{t-1} + \varepsilon_t \quad (4.2)$$

where ε_t are assumed to be $iidN(0, \sigma^2)$. The least squares (LS) estimator in this model exhibits substantial biases, especially when the AR parameter α approaches 1. For the AR parameter α the bias of the LS estimator tends to be downward and for the coefficient on the time trend β the bias tends to be upward. Those biases are quite large as the AR parameter gets close to 1.

Numerous papers deal with this bias of the LS estimator in the AR processes. For stationary AR processes without a time trend Hurwicz (1950), Marriott and Pope (1954), and Kendall (1954) established the mean-bias of the LS estimator. Based on these results Orcutt and Winokur Jr. (1969) constructed approximately mean-unbiased estimates of the AR parameter in stationary models. Le Breton and Pham (1989) calculated the exact and asymptotic biases of the LS estimator in stationary, unit root, and explosive AR(1) models without intercepts.

When the parameter space is bounded or when the distributions of estimators are skewed and/or kurtotic, the concept of median-unbiasedness is often more useful than that of mean-unbiasedness. In the classical normal linear regression model with fixed regressors the LS estimator is median-unbiased. But, in the autoregressive model, the LS estimator is not median-unbiased. As we mentioned in chapter 3, the LS estimators of the AR models with unit roots have asymmetric distributions. In this case there is no unambiguous measure of the center of the distribution. The median may be a preferred measure to the mean because the median is less sensitive to the tails of the distribution. In this sense Rudebusch (1992) and Andrews (1993) proposed median-unbiased estimators of the AR parameters. Since Rudebusch's (1992) procedure is

open to the question of the existence and uniqueness of his estimator, we will discuss Andrews' (1993) procedure for the AR(1) model and its extension to the AR(p) model by Andrews and Chen (1994).

Andrews (1993) proposed a bias correction procedure for the LS estimator in the AR(1) model allowing stationary and unit root cases. Since the LS estimator is downwardly biased, the probability that the LS estimator $\hat{\alpha}_{LS}$ underestimates the true AR parameter is more than half. Andrews proposed to use the corrected LS estimator, *median-unbiased estimator*, which satisfies the condition that the probability of underestimation equals the probability of overestimation.

Andrews (1993) considered three models: AR(1) without drift, AR(1) with drift, and AR(1) with drift and trend. Since in the AR(1) models the distribution of $\hat{\alpha}_{LS}$ depends only on α and is monotone in α (a proof of this property is given in appendix A of Andrews, 1993), one can compute the quantiles of the distribution of the LS estimator according to the values of α . Then the median-unbiased estimator is the value of α at which the LS estimate is equal to a median (0.5 quantile of the distribution of the LS estimator).

Andrews (1993) provides the 0.05, 0.5, 0.95 quantiles of the distribution of the LS estimator according to the values of $\alpha \in (-1, 1]$ for three different AR(1) models. To compute the quantiles of the distribution of the LS estimator he used the Imhof (1961) algorithm. To understand his procedure let us discuss how to use his table. For the AR(1) model with drift and trend, the quantile of the LS estimator with the sample size 60 ($T + 1 = 60$) are given as table 4.5. To get the median-unbiased estimator we use the second column which shows the 0.5 quantiles of the LS estimator (medians) corresponding to the different values of α . If the LS estimate $\hat{\alpha}_{LS}$ is greater than 0.853, then the median-unbiased estimate $\hat{\alpha}_{MU}$ is 1. If $\hat{\alpha}_{LS}$ is less than -0.997, then $\hat{\alpha}_{MU}$ is -1. For any case where the value of $\hat{\alpha}_{LS}$ is between -0.997 and 0.853, the corresponding value of α from table 4.5 is the value of $\hat{\alpha}_{MU}$. Suppose that with the sample size of 60 we have the LS estimate of α as 0.8, then the 0.5 quantile value which corresponds to 0.8 is given as 0.9. Thus $\hat{\alpha}_{MU} = 0.9$.

Andrews (1993) also provides the 0.05 and 0.95 quantiles of the distribution of the LS estimates of three models which can be used to construct the confidence intervals for the parameter α . For example consider constructing a 90 percent two-sided confidence interval for α in the AR(1) model with drift and trend. As in the above example, when $\hat{\alpha}_{LS} = 0.8$, the lower bound for the 90 percent CI can be found in

Table 4.5. *Quantiles of the LS estimator in an AR(1) model with drift and trend*

α	Quantiles ($T + 1 = 60$)		
	0.05	0.5	0.95
-0.999	-1.010	-0.997	-0.945
$\vdots$	$\vdots$	$\vdots$	$\vdots$
0.70	0.417	0.624	0.773
0.80	0.515	0.715	0.846
0.85	0.562	0.758	0.880
0.90	0.607	0.799	0.912
$\vdots$	$\vdots$	$\vdots$	$\vdots$
1.0	0.666	0.853	0.956

Source: Andrews (1983, table III).

0.95 quantiles of the $\hat{\alpha}_{LS}$. When we look at the 0.95 quantiles, we can find that $\alpha = 0.74$ yields the value of the 0.95 quantile of 0.8 (utilizing a linear interpolation). Thus the lower bound for the 90 percent CI is 0.74. For the upper bound we use the 0.05 quantiles column. The 0.05 quantile column shows that if $\hat{\alpha}_{LS} = 0.8$, then the upper bound should be 1. Thus the median-unbiased 90 percent CI for α is (0.74, 1).

Andrews showed that the median unbiased estimator can be used to compute the impulse response functions and to construct unbiased model selection procedures for determining whether the series belongs to the TSP or the DSP class. Andrews also showed that the median-unbiased estimates of α and the corresponding exact confidence intervals for α are quite robust to the nonnormal distribution of ε_t .

Andrews applied the median-unbiased correction procedure for the eight real exchange rate series for the sample period 73:01 to 88:07 which has been analyzed in Schotman and van Dijk (1991a) and for two series, the velocity of money and industrial production, of the Nelson and Plosser data. The AR(1) model with drift is used for the real exchange rate series and the AR(1) with drift and trend is used for two series from the Nelson and Plosser (1982) data. The six real exchange rate series are of the US dollar against the currencies of France (FR), West Germany (WG), Japan (JP), Canada (CA), United Kingdom (UK), and The Netherlands (NL). The two series are of the German Dmark against the currencies of France and The Netherlands. Andrews finds that for

each of the US dollar exchange rates except the UK/US, the median-unbiased estimate of α is 1.00. For the UK/US it is 0.995. For the FR/WG and NL/WG exchange rates, the median-unbiased estimates are 0.968 and 0.965, respectively. He finds that the magnitudes of the differences between the LS estimates and the median-unbiased estimates are small (between 0.017 and 0.022), but they are large in terms of their implications for the persistence of the time series. He also finds that the residuals $(y_t - \hat{\alpha}_{MU}y_{t-1})$ for the eight real exchange rate series reveal evidence of nonnormality for some of the series. But in spite of nonnormal errors the median-unbiased estimates of α were unchanged except for the FR/WG and NL/WG series for which the estimates differed by at most 0.001 and 0.002. The confidence intervals were unchanged.

Andrews and Chen (1994) extended the exact median-unbiased correction method of Andrews (1993) to the $AR(p)$ model. In the $AR(1)$ case the distribution of $\hat{\alpha}_{LS}$ depends only on α and is monotone in α . But in the $AR(p)$ case for $p > 1$, the distribution of $\hat{\alpha}_{LS}$ depends on α and some nuisance parameters, i.e., the coefficients of the high-order term y_{t-j} for $1 < j \leq p$. Andrews and Chen proposed an iterative procedure that jointly estimates α and the nuisance parameters and yields an approximately median-unbiased estimate of α . Once they obtained the approximately median-unbiased estimator of α , the rest of the procedures is the same as those in Andrews (1993). The method for obtaining a median-unbiased estimator of α is extended to generate approximate confidence intervals for α and to an approximately unbiased model selection procedure for determining whether a data series belongs to TSP class or DSP class.

Andrews and Chen applied their procedure to the Nelson and Plosser data set. They found that for three series out of 14 the median-unbiased estimates are equal to 1.0, for seven series the estimates are 0.96 or large, and for three series the estimates are 0.89 or large. These results indicate that the median-unbiased estimates of α show less persistence than the results of the ADF tests performed by Nelson and Plosser (1982) (as we have seen the ADF tests suffered from low power), but show considerably more persistence than the results by the Bayesian estimation methods. (The Bayesian estimation methods will be discussed in chapter 8.) When they use the extended Nelson and Plosser data compiled by Schotman and van Dijk (1991b), eight out of the 14 series, including all of the nominal variables except money stock, have the median-unbiased estimates equal to 1.0.

Abadir (1995a) proposed a minimum mean square error (MSE) esti-

mator for the unit root case which is almost unbiased, while the maximum likelihood estimator is not. He showed that the bias of the minimum MSE estimator is virtually zero because of its asymptotic order of T^{-2} unlike the MLE's T^{-1} bias. However, since Abadir's conclusions are based on the AR(1) model without drift, it is open to the question whether his conclusions hold for AR(p) processes with drift and trend.

4.13 Summary and conclusions

The purpose of this chapter is to discuss several problems with the unit root tests mentioned in the previous chapter, and to outline modifications of those unit root tests. We discuss several other unit root tests – many of which we do not recommend, but they are all discussed for the sake of completeness and also because they are often used. The reason there are so many unit root tests is that there is no uniformly powerful test for the unit root hypothesis (as discussed in Stock, 1994b).

We first start with the size distortions and poor power problem of unit root tests. We then discuss solutions to these problem. The solutions discussed are unit root tests for ARMA models (IV tests), modifications of the PP (Phillips–Perron) tests, modifications of the ADF test (DF–GLS test), and tests based on weighted symmetric estimators. Which of these should be preferred is a question for which there is no clear-cut answer, but it is better to use one of these than the ADF and PP tests discussed in the previous chapter. It is time now to completely discard the ADF and PP tests (they are still used often in applied work!).

We next discuss tests for MA unit root and tests with stationarity as the null. In the latter category a test often referred to is the KPSS test. We do not recommend it. It has the same poor power properties as the ADF test. It should be avoided. The stationarity tests, it is claimed, can be used in conjunction with the ADF test for *confirmatory analysis*. But this is an illusion (as shown by Monte Carlo studies).

We next discuss problems of time aggregation and median-unbiased estimation as an alternative to unit root testing. Finally, we discuss panel data unit root tests which have been suggested as a solution to the poor power problem. But these tests are for an entirely different hypothesis. So it is meaningless to assert that they *solve* the problem of poor power of unit root tests. They solve a different problem. We discuss the Levin–Lin test (very widely used) and point out its limitations. We also discuss the alternative Im–Pesaran–Shin (IPS) test and suggest an

alternative: the Fisher test. In all these cases it is important to ask the basic question: what is the point in the panel data unit root tests?

In fact, it is important to ask the question (rarely asked): why are we interested in testing for unit roots? Much of this chapter (as is customary) is devoted to the question "*How to use unit root tests?*" rather than "*Why unit root tests?*"

One answer is you need the unit root tests as a prelude to cointegration analysis (discussed in the next two chapters). But if this is the case, then the unit root tests are pre-tests and it is not clear what significance levels should be used. If the pre-testing literature is of any guidance we can say that the 1 percent or 5 percent significance levels should *not* be used. Much higher significance levels (say 25 percent) are appropriate.

The other argument is that if there is a unit root then inference is non-standard. Confidence intervals constructed with the usual asymptotic normal distribution are not valid. But just because you fail to reject the unit root null at the 1 percent or 5 percent significance level, does not mean that the unit root null is valid. There is a nonzero probability that it is not valid. Hence it is not clear that nonstandard inference is any more appropriate than standard inference. Note that this problem does not arise in Bayesian inference (discussed later in chapter 8) because the posterior probabilities of the unit root model and stationary model are taken into account. As Cochrane (1991, p. 283) remarked, "the results of unit root tests do not necessarily answer the important question, namely which distribution theory provides a better small sample approximation." The problem is, neither does by itself. One needs to combine both.

In summary, it is high time we asked the question: *Why all this unit root testing* rather than keep suggesting more and more unit root tests and use the Nelson–Plosser data as a *guinea pig* for every new unit root test suggested.

References

- Abadir, K.M. (1995a), "Unbiased Estimation as a Solution to Testing for Random Walks," *Economics Letters*, 47, 263–268.
- (1995b), "A New Test for Nonstationary Against the Stable Alternative," *Econometric Theory*, 11, 81–104.
- Agiakoglou, C. and P. Newbold (1992), "Empirical Evidence on Dickey–Fuller type tests," *Journal of Time Series Analysis*, 13, 471–483.

- Amano, R.A. and S. van Norden (1992), "Unit Root Tests and the Burden of Proof," Working paper no. 92-7, Bank of Canada.
- Anderson, T.W. and A. Takemura (1986), "Why do Non-invertible Estimated Moving Averages Occur?" *Journal of Time Series Analysis*, 7, 235–254.
- Andrews, D.W.K. (1993), "Exactly Median-Unbiased Estimation of First Order Autoregressive / Unit Root Models," *Econometrica*, 61, 139–165.
- Andrews, D. W. K. and D. Chen (1994), "Approximately Median-Unbiased Estimation of Autoregressive Models," *Journal of Business and Economic Statistics*, 12, 187–204.
- Arellano, C. and S. G. Pantula (1995), "Testing for Trend Stationarity versus Difference Stationarity," *Journal of Time Series Analysis*, 16, 147–164.
- Beveridge, S. and C.R. Nelson (1981), "A New Approach to Decomposition of Economic Time Series into Permanent and Transitory Components with Particular Attention to Measurement of the 'Business Cycle'," *Journal of Monetary Economics*, 7, 151–174.
- Bierens, H.J. (1993), "Higher Order Sample Autocorrelations and the Unit Root Hypothesis," *Journal of Econometrics*, 57, 137–160.
- Bierens H.J. and S. Guo (1993), "Testing Stationarity and Trend Stationarity against the Unit Root Hypothesis," *Econometric Reviews*, 12, 1–32.
- Blough, S.R. (1992), "The Relationship Between Power and Level for Generic Unit Root Tests in Finite Samples," *Journal of Applied Econometrics*, 7, 295–308.
- Breitung, J. (1994), "Some Simple Tests of the Moving-Average Unit Root Hypothesis," *Journal of Time Series Analysis*, 15, 351–370.
- Breitung, J. and W. Meyer (1994), "Testing for Unit Roots in Panel Data: Are Wages on Different Bargaining Levels Cointegrated?" *Applied Economics*, 26, 353–361.
- Burke, S.P. (1994), "Confirmatory Data Analysis: The Joint Application of Stationarity and Unit Root Tests," Discussion paper No.20, University of Reading, UK.
- Burridge, P. and E. Guerre (1996), "The Limit Distribution of Level Crossings of a Random Walk, and a Simple Unit Root Test," *Econometric Theory*, 12, 705–723.
- Cheung, Y. and K. Lai (1995), "Lag Order and Critical Values of the

- Augmented Dickey-Fuller Test," *Journal of Business and Economic Statistics*, 13, 277–280.
- Choi, I. (1992a), "Durbin-Hausman Tests for a Unit Root," *Oxford Bulletin of Economics and Statistics*, 54, 289–304.
- (1992b), "Effects of Data Aggregation on the Power of Tests for a Unit Root," *Economics Letters*, 40, 397–401.
- (1994), "Residual-Based Tests for the Null of Stationary with Application to US Macroeconomic Time Series," *Econometric Theory*, 10, 720–746.
- (1995), "Sampling Frequency and the Power of Tests for a Unit Root: A Simulation Study," *Economics Letters*, 49, 131–136.
- Choi, I. and B. Chung (1995), "Sampling Frequency and the Power of Tests for a Unit Root: A Simulation Study," *Economics Letters*, 49, 131–136.
- Choi, I. and B.C. Yu (1997), "A General Framework for Testing $I(m)$ against $I(m+k)$," *Journal of Economic Theory and Econometrics*, 3, 103–138.
- Cochrane, J.H. (1991), "A Critique of the Application of Unit Root Tests," *Journal of Economic Dynamics and Control*, 15, 275–284.
- Cramer, H. (1991), "On Some Classes of Non-Stationary Processes," *Proceedings of the Fourth Berkeley Symposium on Mathematical Statistics and Probability*, vol. II, University of California Press, Berkeley.
- DeJong, D.N., J.C. Nankervis, N.E. Savin, and C.H. Whiteman (1992a), "The Power Problems of Unit Root Tests in Time Series with Autoregressive Errors," *Journal of Econometrics*, 53, 323–343.
- (1992b), "Integration versus Trend Stationarity in Time Series," *Econometrica*, 60, 423–433.
- Dickey, D.A., D.P. Hasza, and W.A. Fuller (1984), "Testing for Unit Roots in Seasonal Time Series," *Journal of American Statistical Association*, 79, 355–367.
- Dickey D.A. and S.G. Pantula (1987), "Determining the Order of Differencing in Autoregressive Processes," *Journal of Business and Economic Statistics*, 5, 455–461.
- Diebold, F.X. and G. Rudebusch (1991b), "On the Power of Dickey-Fuller Tests Against Fractional Alternatives," *Economics Letters*, 35, 155–160.

- Elliott, G., T.J. Rothenberg, and J.H. Stock (1996), "Efficient Tests for an Autoregressive Unit Root," *Econometrica*, 64, 813–836.
- Fisher, R.A. (1932), *Statistical Methods for Research Workers*, 4th ed, Oliver and Boyd, Edinburgh.
- Fuller, W.A. (1996), *Introduction to Statistical Time Series*, 2nd ed, Wiley, New York.
- Gonzalo, J. and T.H. Lee (1996), "Relative Power of t -Type Tests for Stationary and Unit Root Processes," *Journal of Time Series Analysis*, 17, 37–47.
- Hall A. (1989), "Testing for a Unit Root in the Presence of Moving Average Errors," *Biometrika*, 76, 49–56.
- (1995), "Residual Autocorrelation and Unit Root Tests Based on Instrumental Variable Estimators from Time Series Regression Models," *Journal of Time Series Analysis*, 16, 555–569.
- Hall, A. and K. Hassett (1990), "Instrument Choice and Tests for a Unit Root," *Economics Letters*, 35, 161–165.
- Harvey, A.C. (1989), *Forecasting, Structural Time Series Models and the Kalman Filter*, Cambridge University Press, Cambridge.
- Henricsson, R. and E. Lundbäck (1995), "Testing the Presence and the Absence of PPP: Results for Fixed and Flexible Regimes," *Applied Economics*, 27, 635–641.
- Hurwicz, L. (1950), "Least-squares Bias in Time Series," in T.C. Coopmans (ed.), *Statistical Inference in Dynamic Economic Models*, Wiley, New York.
- Hwang, J. and P. Schmidt (1996), "Alternative Methods of Detrending and the Power of Unit Root Tests," *Journal of Econometrics*, 71, 227–248.
- Im, K.S., M.H. Pesaran, and Y. Shin (1996), "Testing for Unit Roots in Heterogeneous Panels," Discussion paper, University of Cambridge.
- Imhof, J.P. (1961), "Computing the Distribution of Quadratic Forms in Normal Variates," *Biometrika*, 48, 419–426.
- Kang, K.M. (1975), "A Comparison of Estimators of Moving Average Processes," unpublished manuscript, Australian Bureau of Census.
- Kendall, M.G. (1954), "Note on Bias in the Estimation of Autocorrelation," *Biometrika*, 41, 403–404.
- King, M.L. (1987), "Towards a Theory of Point Optimal Testing," *Econometric Reviews*, 6, 169–218.

- Kwiatkowski D., P.C.B. Phillips, P. Schmidt, and Y. Shin (1992), "Testing the Null Hypothesis of Stationary against the Alternative of a Unit Root," *Journal of Econometrics*, 54, 159–178.
- Lahiri, K. and M. Paul (1996), "Relative Performance of Pre-Test Estimators in the Presence of a Unit Root," *1996 Proceedings of the Section on Bayesian Statistical Science*, American Statistical Association, Washington DC, 238–243.
- Le Breton, A. and D.T. Pham (1989), "On the Bias of the Least Squares Estimator for the First Order Autoregressive Process," *Annals of the Institute of Statistical Mathematics*, 41, 555–563.
- Lee, J. and P. Schmidt (1994), "Unit Root Tests Based on Instrumental Variables Estimation," *International Economic Review*, 35, 449–462.
- Levin, A. and C-F. Lin (1993a), "Unit Root Test in Panel Data: Asymptotic and Finite Sample Properties," Discussion paper 93-23, UCSD.
- (1993b), "Unit Root Test in Panel Data: New Results," Discussion paper 93-56, UCSD.
- Leybourne, S.J. (1995), "Testing for Unit Roots Using Forward and Reverse Dickey-Fuller Regressions," *Oxford Bulletin of Economics and Statistics*, 57, 559–571.
- Leybourne S.J. and B.P.M. McCabe (1989), "On the Distribution of Some Test Statistics for Parameter Constancy," *Biometrika*, 76, 169–177.
- (1994), "A Consistent Test for a Unit Root," *Journal of Business and Economic Statistics*, 12, 157–166.
- Leybourne, S.J., B.P.M. McCabe, and A.R. Tremayne (1996), "Can Time Series be Differenced to Stationarity?" *Journal of Business and Economic Statistics*, 14, 435–446.
- Li, H. (1995), "The Power of Modified Instrumental Variables Tests for Unit Roots," *Biometrika*, 82, 660–666.
- Liu, P.C. and J. Praschnik (1993), "The Size of the Nonstationary Component and Its Effect on Tests for Unit Roots," *Statistical Papers*, 34, 83–88.
- Maddala, G.S. (1977), *Econometrics*, McGraw-Hill, New York.
- (1991), "To Pool or Not to Pool: That is the Question," *Journal of Quantitative Economics*, 7, 255–264.
- Maddala, G.S. and W. Hu (1995), "The Pooling Problem," in L. Matyas and P. Sevestre (eds.), *Econometrics of Panel Data*, 2nd ed, Kluwer, New York.

- Maddala, G.S. and P.C. Liu (1996), "Panel Data Unit Root Tests: What do They Test?" Paper presented at the International Conference on Panel Data, Amsterdam.
- Maddala, G.S. and S. Wu (1996), "A Comparative Study of Panel Data Unit Root Tests and a Simplified Test," Paper presented at the Econometric Society Meetings, New Orleans, Jan. 1997.
- Marriott, F.H.C. and J.A. Pope (1954), "Bias in the Estimation of Autocorrelations," *Biometrika*, 41, 390–402.
- McCabe, B.P.M. and A.R. Tremayne (1995), "Testing a Time Series for Difference Stationarity," *Annals of Statistics*, 23, 1015–1028.
- McDonald, R. (1996), "Panel Unit Root Tests and Real Exchange Rates," *Economics Letters*, 50, 7–11.
- Nabeya, S. and K. Tanaka (1988), "Asymptotic Theory of a Test for the Constancy of Regression Coefficients Against the Random Walk Alternative," *Annals of Statistics*, 16, 218–235.
- Nelson, C.R. and C.I. Plosser (1982), "Trends and Random Walks in Macroeconomic Time Series," *Journal of Monetary Economics*, 10, 139–162.
- Ng, S. (1995), "Testing for Unit Roots in Flow Data Sampled at Different Frequencies," *Economics Letters*, 47, 237–242.
- Nicholls, D.F. and B.G. Quinn (1982), *Random Coefficient Autoregressive Models: An Introduction*, Springer Verlag, New York.
- Oh, K.Y. (1996), "Purchasing Power Parity and Unit Root Tests Using Panel Data," *Journal of International Money and Finance*, 15, 405–418.
- Orcutt, G.H. and H.S. Winokur, Jr. (1969), "First Order Autoregression: Inference, Estimation, and Prediction," *Econometrica*, 37, 1–14.
- Pantula, S.G. (1986), "Determining the Order of Differencing in Autoregressive Processes," unpublished manuscript, Dept. of Statistics, North Carolina State University.
- (1991), "Asymptotic Distributions of Unit Root Tests When the Process is Nearly Stationary," *Journal of Business and Economic Statistics*, 9, 63–71.
- Pantula S.G. and A. Hall (1991), "Testing for Unit Roots in Autoregressive Moving Average Models," *Journal of Econometrics*, 48, 325–353.
- Pantula, S. G., G. Gonzalez-Farias, and W. A. Fuller (1994), "A Comparison of Unit-Root Test Criteria," *Journal of Business and Economic Statistics*, 12, 449–459.

- Papell, D.H. (1997), "Searching for Stationarity: Purchasing Power Parity under the Current Float," *Journal of International Economics*, Forthcoming.
- Park, H.J. and W.A. Fuller (1995), "Alternative Estimators and Unit Root Tests for the Autoregressive Process," *Journal of Time Series Analysis*, 16, 415–429.
- Park, J.Y. (1990), "Testing for Unit Roots and Cointegration by Variable Addition," in T.B. Fomby and G.F. Rodes (eds.), *Advances in Econometrics*, vol. 8, JAI Press, 107–133.
- Perron, P. (1989), "Testing for a Random Walk: A Simulation Experiment of Power When the Sampling Interval Is Varied," in B. Raj (ed.), *Advances in Econometrics and modeling*, Kluwer Academic Publishers.
- (1991), "Test Consistency with Varying Sampling Frequency," *Econometric Theory*, 7, 341–368.
- Perron, P. and S. Ng (1996), "Useful Modifications to Some Unit Root Tests with Dependent Errors and Their Local Asymptotic Properties," *Review of Economic Studies*, 63, 435–465.
- Phillips, P.C.B. (1987), "Time Series Regression with a Unit Root," *Econometrica*, 55, 277–301.
- Phillips, P.C.B. and P. Perron (1988), "Testing for a Unit Root in Time Series Regression," *Biometrika*, 75, 335–346.
- Plosser, C.I. and G.W. Schwert (1977), "Estimation of a Non-invertible Moving Average Process: The Case of Over-differencing," *Journal of Econometrics*, 6, 199–224.
- Quah, D. (1992), "International Patterns of Growth; I : Persistence in Cross Country Disparities," LSE working paper.
- (1994), "Exploiting Cross Section Variation for Unit Root Inference in Dynamic Data," *Economics Letters*, 44, 9–19.
- Rao, C.R. (1973), *Linear Statistical Inference and Its Applications*, 2nd ed, Wiley, New York.
- Rudebusch, G. (1992), "Trends and Random Walks in Macroeconomic Time Series: A Re-examination," *International Economic Review*, 33, 661–680.
- Said, S.E. and D.A. Dickey (1984), "Testing for Unit Roots in Autoregressive-Moving Average Models of Unknown Order," *Biometrika*, 71, 599–607.
- Saikkonen, P. and R. Luukkonen (1993), "Testing for a Moving Average Unit Root in ARIMA Models," *Journal of the American Statistical Association*, 88, 596–601.

- Sargan, J.D. and A. Bhargava (1983), "Maximum Likelihood Estimation of Regression Models with First Order Moving Average Errors When the Root Lies on the Unit Circle," *Econometrica*, 51, 799–820.
- Schmidt, P. and P.C.B. Phillips (1992), "LM Tests for a Unit Root in the Presence of Deterministic Trends," *Oxford Bulletin of Economics and Statistics*, 54, 257–287.
- Schotman, P. and H.K. van Dijk (1991a), "A Bayesian Analysis of the Unit Root in Real Exchange Rates," *Journal of Econometrics*, 49, 195–238.
- (1991b), "A Bayesian Routes to Unit Roots," *Journal of Applied Econometrics*, 6, 387–401.
- Schwert, G.W. (1987), "Effects of Model Specification on Tests for Unit Roots," *Journal of Monetary Economics*, 20, 73–103.
- (1989), "Tests for Unit Roots: A Monte Carlo Investigation," *Journal of Business and Economic Statistics*, 7, 147–159.
- Shephard, N. (1993), "Distribution of the ML Estimator of an MA(1) Model and a Local Level Model," *Econometric Theory*, 9, 377–401.
- Shephard, N. and A.C. Harvey(1990), "On the Probability of Estimating a Deterministic Component in the Local Level Model," *Journal of Time Series Analysis*, 11, 339–347.
- Shiller, R. and P. Perron (1985), "Testing the Random Walk Hypothesis: Power versus Frequency of Observation," *Economics Letters*, 18, 381–386.
- Stock, J.H. (1990), "A Class of Tests for Integration and Cointegration," manuscript, Harvard University.
- (1991), "Confidence Intervals of the Largest Autoregressive Root in Macroeconomic Time Series," *Journal of Monetary Economics*, 28, 435–460.
- (1994a), "Deciding Between I(1) and I(0)," *Journal of Econometrics*, 63, 105–131.
- (1994b), "Unit Roots, Structural Breaks, and Trends," in R.F. Engle and D.L. McFadden (eds.), *Handbook of Econometrics*, vol. IV, Elsevier, Amsterdam.
- Tanaka K. (1990), "Testing for a Moving Average Unit Root," *Econometric Theory*, 6, 433–444.
- Watson, M.W. (1994), "Vector Autoregressions and Cointegration," in R.F. Engle and D.L. McFadden (eds.), *Handbook of Econometrics*, vol. IV, Elsevier, North-Holland.

- Wu, Y. (1996), "Are Real Exchange Rates Stationary? Evidence from a Panel Data Set," *Journal of Money, Credit, and Banking*, 28, 54–63.
- Yap, S.F. and G.C. Reinsel (1995), "Results on Estimation and Testing for a Unit Root in the Non-stationary ARMA Model," *Journal of Time Series Analysis*, 16, 339–353.

5

Estimation of cointegrated systems

5.1 Introduction

In chapter 2 we introduced the concept of cointegration. In this chapter we shall discuss different estimation methods for cointegrated systems. In the next chapter we shall discuss tests for cointegration.

The methods of estimation fall into two categories:

- (i) single equation methods,
- (ii) system methods.

This classification is similar to the classification that we usually adopt in the estimation of simultaneous equations models. In the single equation methods we are interested in estimating only one particular cointegrating (CI) vector. In the system methods, we also determine the number of cointegrating vectors. In the following sections we shall discuss:

- (i) the different methods of estimation when the number of CI vectors is known,
- (ii) estimation of the number of CI vectors,
- (iii) inference on the coefficients (elements) of the CI vectors.

For the last problem we also face identification problems, the nature of which is somewhat different from the identification problems we encounter in the usual simultaneous equation models.

5.2 A general CI system

Consider a VAR (vector autoregressive) model

$$Y_t = A_1 Y_{t-1} + \cdots + A_k Y_{t-k} + U_t, \quad t = 1, 2, \dots, T \quad (5.1)$$

where Y_t is an n -vector of $I(1)$ variables.

We shall consider some subcases of (5.1) before we discuss the general VAR model. These are:

Case (i) $n = 2$ and no dynamics. We have only two $I(1)$ variables and we can write the CI equation as

$$y_{1t} = \beta y_{2t} + u_t$$

where u_t is $I(0)$. This is *normalized* with respect to y_{1t} . We shall discuss the issue of normalization later.

Case (ii) $n > 2$ but the model is a triangular system as discussed, *inter alia* in Phillips (1991)

$$y_{1t} = \beta_2' y_{2t} + u_{1t}$$

$$\Delta y_{2t} = u_{2t}$$

where y_{2t} is a vector of all the $I(1)$ variables other than y_{1t} and u_{2t} is a vector of the same dimension. u_{1t} is $I(0)$ because we are considering a CI relationship and u_{2t} is $I(0)$ because y_{2t} is $I(1)$.

We shall discuss the several single equation estimation methods with reference to this equation and then move on to the system methods. However, since it is easier to highlight the main problems with reference to the simplest methods, we start with a two-variable model first.

5.3 A two-variable model: Engle–Granger methods

Consider two variables y_{1t} and y_{2t} which are both $I(1)$ and the cointegration relationship

$$y_{1t} = \beta y_{2t} + u_t \tag{5.2}$$

where u_t is $I(0)$. This two-variable model, with some modifications, is often used in practice (e.g., in discussions of the tests of purchasing power parity (PPP) theory and the Fisher effect on interest rates). Hence it is important to consider it in detail.

There are several issues related to this simple model. They are:

- (i) Uniqueness of β . This has been discussed in section 2.10 of chapter 2. As noted there this uniqueness does not carry over to models with more than two variables. In these models the CI vector need not be unique.

- (ii) Superconsistency of $\hat{\beta}$. Stock (1987) shows that the OLS estimator $\hat{\beta}$ of β converges to its true value at the rate T (superconsistency) instead of the usual rate $\sqrt{T}$ (consistency). This is because $\sum y_{2t}^2$ is $O_p(T^2)$ if y_{2t} is $I(1)$. This superconsistency properly also holds in the case of several $I(1)$ variables, where we consider the OLS estimation of the CI vector.
- (iii) Although $\hat{\beta}$ is superconsistent, Banerjee *et al.* (1986) and Banerjee *et al.* (1993, pp. 214–230) show through Monte Carlo studies that there can be substantial small sample biases and suggest estimating the long-run parameter β by estimating a dynamic regression rather than the static regression (5.2). We shall discuss details of this along with the small sample properties of the other estimators.
- (iv) As noted in section 2.10 of chapter 2, the fact that y_{1t} and y_{2t} are cointegrated implies that they are related by an ECM. The ECM measures short-run dynamics whereas the CI relationship (5.2) gives the long-run relationship. The Engle–Granger *two-step procedure* involves estimation of β by ordinary least squares using the static regression (5.2), substituting this for β in the ECM and then estimating the ECM by least squares. The superconsistency of $\hat{\beta}$ assures that the two-step estimators of the parameters of the ECM have the *same asymptotic distribution* as the ones that one would get if β is known.
- (v) Another issue is that of normalization. Note that equation (5.2) is normalized with respect to y_{1t} . If we normalize the equation with respect to y_{2t} and write

$$y_{2t} = \alpha y_{1t} + v_t$$

then the least squares estimator $\hat{\alpha}$ of α will also be superconsistent and since $\alpha = 1/\beta$, we can get a superconsistent estimator of β as $1/\hat{\alpha}$. If the R^2 from the equation is very close to 1, then we will have $\hat{\beta} \simeq 1/\hat{\alpha}$. Hendry (1986) argues that this is the case with most economic relationships. However, this will not be the case in many instances. Ng and Perron (1997) study the normalization problem in two-variable models. They show that the least squares estimator can have very poor finite sample properties when normalized in one direction but can be well behaved when normalized in the other. This occurs when one of the variables is a weak random walk or is nearly stationary. They suggest to use as regressand, the variable that is *less integrated*. As a practical

matter they suggest ranking the variables by the spectral density at frequency zero of the first-differenced series.

As an empirical illustration they consider the estimation of the Fisher equation relating interest rates to expected inflation rates. The Fisher equation is defined as

$$(1 - \tau)i = \pi^e + r$$

where r is the real interest rate, i is the nominal interest rate, τ is the marginal tax rate, and π^e is the expected inflation rate (unobserved). If we use the observed inflation rate π for π^e , this introduces an errors in variables bias, but if i and π are $I(1)$ and they are cointegrated, we can get consistent estimates of the parameters.

Ng and Perron do several unit root tests. They conclude that the interest rate series is unambiguously $I(1)$, but the evidence on π is mixed. They report results using π and i as regressands and argue that the results from a regression of π on i is the one to be trusted.

- (vi) Although the least squares estimator $\hat{\beta}$ is superconsistent, its asymptotic distribution depends on nuisance parameters arising from endogeneity of the regressor and serial correlation in errors. To see this, we shall write equation (5.2) as

$$\begin{aligned} y_{1t} &= \beta y_{2t} + u_{1t} \\ \Delta y_{2t} &= u_{2t} \end{aligned} \tag{5.3}$$

where u_{1t} and u_{2t} are $I(0)$ since y_{2t} is $I(1)$. Then endogeneity can be modeled by assuming that

$$\begin{aligned} \text{cov}(u_{1t}, u_{2t}) &= \sigma_{12} \text{ if } t = s \\ &= 0 \text{ otherwise} \end{aligned}$$

Serial correlation can be introduced in u_{1t} and u_{2t} . The alternative estimation methods that we will be discussing in the following sections differ in the way these two problems of endogeneity of the regressors and serial correlation in the errors are handled.

- (vii) If u_{1t} and u_{2t} in (5.3) are independent and thus $\sigma_{12} = 0$, there is no simultaneity. If u_{1t} are first-order autocorrelated with correlation ρ , the DW statistic computed from the residuals of (5.3) converges to the usual $2(1 - \rho)$ as with regressions of stationary

variables. Of course, if $\rho = 1$, then y_{1t} and y_{2t} are not cointegrated and the DW statistic should be close to zero. This is the reasoning behind the use of the DW statistic as a test for cointegration, as suggested by Sargan and Bhargava (1983). We shall discuss this test along with other tests for cointegration in the next chapter.

- (viii) In the case where there is no endogeneity and no serial correlation in the errors u_{1t} and u_{2t} , the t -statistic for testing the hypothesis $\beta = 0$ has the standard normal distribution asymptotically.
- (ix) The asymptotic distribution of $\hat{\beta}$ is also normal if in (5.3) we assume that Δy_{2t} has a drift, that is $\Delta y_{2t} = \mu + u_{2t}$, and we assume no endogeneity and no serial correlation in the errors u_{1t} and u_{2t} . In this case

$$y_{2t} = t\mu + \sum_{s=1}^t u_{2s} = t\mu + S_t \quad (\text{say})$$

As we discussed in section 3.4 of chapter 3, $t\mu$ will dominate the term S_t . Also

$$\hat{\beta} - \beta = \frac{\sum y_{2t} u_{1t}}{\sum y_{2t}^2}$$

Now $\sum y_{2t}^2$ will be dominated by $\mu^2 \sum t^2$ and thus will be $O_p(T^3 \mu^2 / 3)$ (see section 3.2). Hence

$$plim T^{-3} \sum y_{2t}^2 = \frac{\mu^2}{3}$$

Also, $\sum y_{2t} u_{1t}$ will be dominated by $\mu \sum t u_{1t}$ (see section 3.2) and

$$T^{-3/2} \sum t u_{1t} \Rightarrow \int_0^1 r dW(r) \Rightarrow N(0, \frac{\sigma_1^2}{3})$$

where $\sigma_1^2 = \text{var}(u_{1t})$. Hence we get the result

$$T^{3/2}(\hat{\beta} - \beta) \Rightarrow N(0, \frac{\sigma_1^2}{\mu^2})$$

The result here is similar to the asymptotic normality of unit root test statistics we discussed in section 3.4.1 and is subject to the same criticism noted there.

5.4 A triangular system

Consider now a system consisting of more than two $I(1)$ variables. In this case we can have more than one CI vector. The single equation methods, however, assume that there is only one CI vector and concentrate on the estimation of one CI vector. We shall discuss these methods in terms of the triangular system

$$y_{1t} = \beta' y_{2t} + u_{1t}$$

$$\Delta y_{2t} = u_{2t}$$

where y_{2t} is the set of all the $I(1)$ variables other than y_{1t} . We assume that each element of y_{2t} has one unit root and that there are no CI relationships among the regressors y_{2t} . Denote $u'_t = (u_{1t}, u_{2t})$. We assume that $\{u_t\}$ is strictly stationary with mean 0 and contemporaneous covariance matrix

$$\Sigma = \begin{bmatrix} \sigma_{11} & \sigma'_{21} \\ \sigma_{21} & \Sigma_{22} \end{bmatrix}$$

It is assumed that $\Sigma > 0$.

The cumulative sums of u_t follow a multivariate random walk. Following the same arguments as in section 3.2 of chapter 3, these partial sums have a limiting Brownian motion or Wiener process $W(r)$ which we shall partition as

$$W = \begin{bmatrix} W_1 \\ W_2 \end{bmatrix}$$

corresponding to the partition of u_t into u_{1t} and u_{2t} . The covariance matrix of this is called the long-run covariance matrix which we denote by Ω .

$$\Omega = \lim_{T \rightarrow \infty} \frac{1}{T} \sum_{t=1}^T \sum_{s=1}^T E(u_t u'_s)$$

This is the sum of all covariances backwards and forwards of u_t and u_s . It can be decomposed into a contemporaneous variance and sums of autocovariances. Thus

$$\Omega = \Sigma + \Lambda + \Lambda'$$

where $\Sigma = E(u_0 u'_0)$, $\Lambda = \sum_{t=1}^{\infty} E(u_0 u'_t)$ and $\Lambda' = \sum_{t=1}^{\infty} E(u_t u'_0)$. The least squares estimator $\hat{\beta}$ of β is given by

$$\hat{\beta} = (Y_2' Y_2)^{-1} Y_2' y_1$$

where y_1 is the vector of observations on y_{1t} and Y_2 is the matrix of

observations on y_{2t} . The least squares estimator $\hat{\beta}$ is superconsistent but its asymptotic distribution depends on nuisance parameters arising from the endogeneity of the regressors and serial correlation in the errors. Specifically, the asymptotic distribution is given by (see Park and Phillips, 1988)

$$T(\hat{\beta} - \beta) \Rightarrow \left(\int_0^1 W_2 W_2' \right) \left(\left[\int_0^1 W_2 d(W_{1\cdot 2} + \omega_{12} \Omega_{22}^{-1} W_2) \right] + \delta \right)$$

where $W_{1\cdot 2} = W_1 - \omega_{12} \Omega_{22}^{-1} W_2$, Ω is partitioned as

$$\Omega = \begin{bmatrix} \omega_{11} & \omega'_{21} \\ \omega_{21} & \Omega_{22} \end{bmatrix}$$

and

$$\delta = \sum_{t=0}^{\infty} E(u_{1t} u_{20})$$

There are two different sets of procedures to deal with the endogeneity and serial correlation problems. The procedures are similar to those discussed in chapter 3 in connection with the PP and ADF unit root tests. The PP tests modify the test statistic and the ADF test modifies the estimating equation. Similarly, the fully modified OLS (FM-OLS) of Phillips and Hansen (1990) applies nonparametric corrections to the OLS estimator $\hat{\beta}$. The other methods like Saikkonen's dynamic OLS (DOLS), Phillips and Loretan's nonlinear least squares (NLS), and Stock and Watson's dynamic GLS (DGLS), add leads and lags to the estimating equation. We shall discuss the FM-OLS method first and then the other methods.

5.4.1 The FM-OLS method

The FM-OLS procedure (Phillips and Hansen, 1990) eliminates the nuisance parameters in the following way. First modify y_{1t} using the transformation

$$\hat{y}_{1t}^+ = y_{1t} - \hat{\omega}_{12} \hat{\Omega}_{11} \Delta y_{2t}$$

and the error u_{1t} also by

$$\hat{u}_{1t}^+ = u_{1t} - \hat{\omega}_{12} \hat{\Omega}_{11} \Delta y_{2t}$$

This is a correction for endogeneity. Next we construct a serial correlation correction term $\hat{\delta}^+$ which is a consistent estimator of

$$\delta^+ = \sum_{k=0}^{\infty} (u_{1k}^+ u'_{21})$$

where $u_{1t}^+ = u_{1t} - \omega_{12}\Omega_{11}\Delta y_{2t}$. The FM-OLS estimator combines these two corrections to the least squares estimator and is given by

$$\tilde{\beta} = (Y_2'Y_2)^{-1}(Y_2'\hat{y}_1^+ - T\hat{\delta}^+)$$

As for $\hat{\Omega}$ they suggest the Newey–West estimator we discussed in chapter 3 in connection with the Phillips–Perron tests. However, as commented in chapter 4, the modifications suggested by Perron and Ng to the PP tests can also be implemented in this case.

There is also the question of normalization we discussed earlier with the two-variable model. Ng and Perron (1997) argue that the normalization problems apply to the FM-OLS method as well and that the guidelines using less integrated variables as regressands would seem appropriate.

Phillips and Hansen (1990) present simulation evidence to argue that the FM-OLS performs well. However, we shall discuss this after going through the other single equation and system estimation methods, where it is argued that some other estimators perform better.

5.4.2 Adjusting estimating equations

The FM-OLS method is similar in spirit to the PP unit root tests in that it starts with the OLS estimator and applies corrections to it to take care of the endogeneity and serial correlation problems. The other methods are similar to the ADF test in that they modify the estimating equations. The methods we shall discuss are those suggested by Hendry; Saikkonen; Phillips and Loretan; Stock and Watson; and Inder. We shall discuss these in turn.

Single equation ECM (SEECM): Hendry's dynamic regression method

In the case of the Engle–Granger two-step procedure we discussed earlier, Banerjee *et al.* (1986) present simulation results and argue that ignoring lagged terms in a static equation like (5.2) may lead to substantial biases in the estimation of β in finite samples. They propose

to estimate the long-run parameters by an unrestricted error correction model (ECM) incorporating all the dynamics. Hendry has advocated this method in several papers. Broadly speaking the method involves adding lags of Δy_{1t} and Δy_{2t} to equation (5.2) and estimating the resulting equation. The idea is to start with a sufficiently large number of these lags and progressively to simplify it. Hendry's PC-GIVE computer program can be used to implement this. Ericsson *et al.* (1990) describe the computer program and also Hendry's methodology.

Saikkonen's dynamic ordinary least squares (DOLS)

This estimator was suggested by Saikkonen (1991), but the term DOLS was used by Stock and Watson (1993) who generalize it to systems with higher orders of integration. Saikkonen's method involves estimating

$$y_{1t} = \beta' y_{2t} + \sum_{j=-k_1}^{k_2} b_j \Delta y_{2,t-j} + v_t \quad (5.4)$$

where k_1 and k_2 are selected to increase at an appropriate rate with T . The procedure involves adding leads and lags of Δy_{2t} but not of y_{1t} .

Phillips and Loretan's nonlinear least squares (NLS)

Phillips and Loretan (1991) suggest a dynamic equation of the form

$$y_{1t} = \beta' y_{2t} + \sum_{j=1}^k b_j \Delta y_{2,t-j} + \sum_{j=1}^l \gamma_j (y_{1,t-j} - \beta' y_{2,t-j}) + v_t$$

The procedure is in the spirit of the ECM and included lagged values of the equilibrium error. By comparison, Hendry's method which involves adding lags of Δy_{1t} and Δy_{2t} with no restrictions on the coefficients and merely designed to mop up the dynamics is more *ad hoc*. However, since β occurs in a nonlinear way, the equation has to be estimated by nonlinear least squares methods. Phillips and Loretan (1991, pp. 427–428) report that in their simulation study, the nonlinear formulation did create some problems in estimation.

Stock and Watson's dynamic generalized least squares (DGLS)

Stock and Watson (1993) suggest correcting for serial correlation in Saikkonen's method by using GLS. They suggest getting the estimated residuals using Saikkonen's method to construct the covariance matrix of the errors and then estimating the equation by GLS. Both Saikkonen's

and Stock and Watson's procedures thus start with the OLS estimation of the model (5.4).

Inder's fully parametric least squares (FPLS)

Inder (1995) starts with the VAR model in equation (5.1) and derives the implied single equation

$$y_{1t} = \beta' y_{2t} + A(L)\Delta y_{1t} + B(L)\Delta y_{2t} + v_t \quad (5.5)$$

where $A(L)$ and $B(L)$ are polynomials in the lag operator L . Then after suitable truncation of the lag polynomials $A(L)$ and $B(L)$ his suggestion is to regress y_{1t} on y_{2t} , leads and lags of Δy_{2t} , and lags of Δy_{1t} . The leads and lags of Δy_{2t} eliminate any effect of endogeneity and the lags of Δy_{1t} capture remaining autocorrelation in the stationary component of the regression. Inder shows how this is linearization of the nonlinear equation of Phillips and Loretan which is

$$y_{1t} = \beta' y_{2t} + A_1(L)(y_{1t} - \beta' y_{2t}) + B_1(L)\Delta y_{2t} + v_t$$

There are two problems yet to be resolved. The first is that the presence of Δy_{1t} in equation (5.5) means that equation (5.5) cannot be estimated by ordinary least squares. Hence, one has to use the instrumental variable method suggested in Bewley (1979). The other problem is that of choosing the lag length. This problem is common to all the procedures discussed above.

Selection of the lag length

In all these procedures that we have discussed, there is the problem of the selection of the appropriate lag length in the truncation of the lag polynomials $A(L)$ and $B(L)$. Inder suggests starting with a model with only the current values of Δy_{1t} and Δy_{2t} , regressing the residuals $\hat{v}_t$ on two extra lags of Δy_{1t} and Δy_{2t} and two leads of Δy_{2t} , and checking TR^2 using a χ^2 distribution for its significance. This is the LM test for the joint significance of the extra regressors. If this is rejected, increase one of the lag and lead truncation values by one and repeat until the diagnostic test is not rejected. However, this procedure of going to a more general model from a restricted model often termed *specific to general* has been often criticized. See the discussion in section 3.6.2 of chapter 3. As discussed there, Hall's *general to specific* method is preferable to other rules for selection of the truncation lag. Although there is no corresponding discussion of lag selection as in the case of the ADF test, we conjecture that the same rules apply in the estimation of the CI regressions. As

noted earlier, the above methods are in the spirit of the ADF test for unit roots, i.e., modifying the estimating equation adding lags (and leads).

5.5 System estimation methods

We have discussed the single equation methods, where we assume the existence of one CI vector, which is of interest. Since all these methods are associated with least squares estimation, there are the normalization problems, which we talked about earlier. Ng and Perron (1997) argue that their rules derived for the two-variable model, apply to the FM-OLS method as well, but do not comment on their applicability to the other methods discussed earlier by Hendry, Saikkonen, Phillips, and Loretan, and others.

By contrast in the system estimation methods the problem of normalization does not appear (except where noted later) and also the number of CI vectors is not fixed *a priori* but determined in the course of estimation.

The system estimation methods we shall discuss are:

- (i) Johansen's procedure (this is the most commonly used),
- (ii) Box-Tiao procedure,
- (iii) Methods depending on principal components.

5.5.1 The Johansen procedure

Johansen's procedure applies maximum likelihood to the VAR model, assuming that the errors are Gaussian.

- (i) Start

$$Y_t = A_1 Y_{t-1} + \cdots + A_k Y_{t-k} + U_t, \quad t = 1, \dots, T$$

where Y_t is an n -vector of $I(1)$ variables.

- (ii) Write it as

$$\Delta Y_t = B_1 Y_{t-1} + B_2 \Delta Y_{t-1} + \cdots + B_k \Delta Y_{t-k+1} + U_t$$

where $B_1 = -I + \sum_{i=1}^k A_i$ and $B_j = -\sum_{i=j}^k A_i$ for $j = 2, \dots, k$.

- (iii) Since $\Delta Y_t, \dots, \Delta Y_{t-k+1}$ are all $I(0)$ but Y_{t-1} is $I(1)$, in order that this equation be consistent, B_1 should not be of full rank. Let its rank be r . Write

$$B_1 = \alpha\beta'$$

where α is an $n \times r$ matrix and β' is an $r \times n$ matrix. Then $\beta'Y_{t-1}$ are the r cointegrated variables, β' is the matrix of coefficients of the cointegrating vectors and α has the interpretation of the matrix of error correction terms.

- (iv) Since our interest is in α and β' we eliminate $B_2, \dots, B_k$ first. To do this we proceed as follows. Regress ΔY_t on $\Delta Y_{t-1}, \dots, \Delta Y_{t-k+1}$. Get the residuals. Call them R_{0t} . Regress Y_{t-1} on these same variables. Get the residuals. Call them R_{1t} . Now our regression equation is reduced to

$$R_{0t} = \alpha\beta'R_{1t} + u_t$$

This is a multivariate regression problem. Define

$$\begin{bmatrix} S_{00} & S_{01} \\ S_{10} & S_{11} \end{bmatrix}$$

as the matrix of sums of squares and sums of products of R_{0t} and R_{1t} . (Each of these matrices is of order $n \times n$.) Johansen (1991) shows that the asymptotic variance of $\beta'R_{1t}$ is $\beta'\Sigma_{11}\beta$, the asymptotic variance of R_{0t} is Σ_{00} and the asymptotic covariance matrix of $\beta'R_{1t}$ and R_{0t} is $\beta'\Sigma_{10}$ where Σ_{00}, Σ_{10} , and Σ_{11} are the population counterparts of S_{00}, S_{10} , and S_{11} .

- (v) We shall maximize the likelihood function with respect to α holding β constant and then maximize with respect to β in the second step. We get

$$\hat{\alpha}' = (\beta'S_{11}\beta)^{-1}\beta'S_{10}$$

Note that $\hat{\alpha}'$ is an $r \times n$ matrix and the conditional maximum of the likelihood function is given by

$$[L(\beta)]^{-2/T} = |S_{00} - S_{01}\beta(\beta'S_{11}\beta)^{-1}\beta'S_{10}|$$

Maximization of the likelihood function with respect to β implies minimization of this determinant with respect to β .

- (vi) We shall use the identity

$$|C - B'A^{-1}B| = \frac{|A - BC^{-1}B'| \cdot |C|}{|A|}$$

with $C = S_{00}, A = \beta'S_{11}\beta$, and $B = \beta'S_{10}$. We then have to minimize

$$\frac{|\beta'S_{11}\beta - \beta'S_{10}S_{00}^{-1}S_{01}\beta| \cdot |S_{00}|}{|\beta'S_{11}\beta|}$$

But

$$\min_X \frac{|X'(A_1 - A_2)X|}{|X'A_1X|}$$

is given by the maximum characteristic root of the equation $|A_2 - \lambda A_1| = 0$. Thus, substituting $A_1 = S_{11}$ and $A_2 = S_{10}S_{00}^{-1}S_{01}$ we get the maximum of the likelihood function by solving the eigenvalue problem

$$|S_{10}S_{00}^{-1}S_{01} - \lambda S_{11}| = 0$$

or finding the eigenvalue of

$$|S_{11}^{-1}S_{10}S_{00}^{-1}S_{01} - \lambda I| = 0 \quad (5.6)$$

But the roots of this equation are the r canonical correlations between R_{1t} and R_{0t} . That is we seek those linear combinations of Y_{t-1} that are highly correlated with linear combinations of ΔY_t after conditioning on the lagged variables $\Delta Y_{t-1}, \dots, \Delta Y_{t-k+1}$.

- (vii) Note that if the eigenvalues of A are λ_i , the eigenvalues of $(I - A)$ are $(1 - \lambda_i)$. Hence if λ_i are the canonical correlations given by solving equation (5.6), then $(1 - \lambda_i)$ are the eigenvalues of $(I - S_{11}^{-1}S_{10}S_{00}^{-1}S_{01})$.
- (viii) Since the value of the determinant of a matrix is equal to the product of its eigenvalues, we have

$$\prod_{i=1}^n (1 - \lambda_i) = |I - S_{11}^{-1}S_{10}S_{00}^{-1}S_{01}| = \frac{|S_{11} - S_{10}S_{00}^{-1}S_{01}|}{|S_{11}|}$$

Again using the determinant identity in (vi) we get this equal to

$$\frac{|S_{00} - S_{01}S_{11}^{-1}S_{10}|}{|S_{00}|}$$

Hence

$$L_{max}^{-2/T} = |S_{00}| \cdot \prod_{i=1}^n (1 - \lambda_i) \quad (5.7)$$

Note that this result corresponds to the result for the normal multiple regression model

$$L_{max}^{-2/T} = \text{constant} \cdot (y'y)(1 - R^2)$$

$y'y$ is replaced by the generalized variance $|S_{00}|$ and $1 - R^2$ is replaced by

$$\prod_{i=1}^n (1 - \lambda_i)$$

where λ_i are the canonical correlations.

- (ix) To determine the number of CI vectors Johansen suggests two tests: the trace test and the maximum eigenvalue test. We shall discuss these in the next chapter.

The number of cointegrating relationships in the system, r , is chosen in the procedure by LR tests described in chapter 6, which make use of the fact that if there are r cointegrating vectors, then the $(n-r)$ smallest eigenvalues of equation (5.6) are zero. The corresponding r eigenvectors are chosen as cointegrating vectors. This way of imposition of $n-r$ restrictions in the system yields an asymptotically efficient and optimal estimator of the cointegrating vectors.

The Johansen procedure is called *reduced rank regression*. This was introduced by Anderson (1951) in the context of independent variables and has been applied by Ahn and Reinsel (1988) for stationary processes and Ahn and Reinsel (1990) for nonstationary processes. By reduced rank regression of x_t and y_t corrected for z_t we mean the following regression: regress x_t on z_t and y_t on z_t to form residuals u_t and v_t respectively and then regress u_t on v_t . This is the procedure used in step (iv) described earlier.

There are two very minor differences between the Johansen and Ahn and Reinsel approaches. The first is in the normalization used. The second is in the computational algorithms used. Johansen used partial canonical correlation analysis, while Ahn and Reinsel suggest an algorithm based on iterated least squares. Both methods yield the same final results. Because the differences are minor, we shall not elaborate on the Ahn and Reinsel method here.

5.5.2 The Johansen procedure with trends

In the preceding discussion, the model had no constant term or trend. This is an unrealistic assumption in practice. But before we discuss the Johansen procedure with trends, we shall clarify the difference between stochastic and deterministic cointegration.

Stochastic versus deterministic cointegration

The presence of deterministic trends in the $I(1)$ variables, has led to two definitions of cointegration: *stochastic cointegration* and *deterministic cointegration* (see Ogaki and Park, 1997 and Campbell and Perron, 1991).

A vector of $I(1)$ variables Y_t are said to be *stochastically cointegrated*

with cointegrating rank r , if there are r linearly independent combinations of the variables Y_t that are $I(0)$. These combinations may have nonzero deterministic trends.

The variable Y_t are said to be *deterministically cointegrated* with cointegrating rank r , if the r combinations of Y_t that are $I(0)$ are stationary with no deterministic trends. That is, under deterministic cointegration, the same cointegrating vectors that eliminate stochastic nonstationarity and also eliminate deterministic nonstationarity. The Engle and Granger (1987) concept of cointegration is that of deterministic cointegration.

Johansen (1992a) and Perron and Campbell (1993) have modified the original Johansen method to accommodate the possibility of stochastic cointegration and trend stationarity. Though these modifications are slightly different from each other, the modifications are simply to include time trend as an additional regressor in the Johansen estimation method. The significance tables for the test statistics, however, are different. These differences will be discussed in chapter 6 (section 6.5.1).

5.5.3 The Box–Tiao method

The Box–Tiao (1977) procedure (to be referred to as BT) is an alternative way of deriving cointegrating vectors using canonical correlation analysis. Bossaerts (1988) first proposed the BT procedure as an alternative to the Engle–Granger methods of estimating cointegrating vectors and applied the procedure to investigate common nonstationary components of US stock prices.

BT noted that if one finds the canonical correlations between y_t and y_{t-1} , the most predictable components would be nearly or actually nonstationary and the least predictable components would reflect stable stationary processes. But what we consider in the BT procedure is canonical correlations in the levels and hence Bewley and Yang (1995) call this method *levels canonical correlation analysis*. LCCA.

The idea in Johansen method is to find linear combinations of Y_{t-1} that are most highly correlated with ΔY_t (which is $I(0)$) on the argument that $I(0)$ and $I(1)$ variables are uncorrelated. The idea behind the BT method is to find linear combinations of Y_{t-1} that are least correlated with Y_t (which is $I(1)$). Because the dependent variable in the Johansen procedure is $I(0)$ and the dependent variable in the BT method is $I(1)$, the distribution theory is more complicated in the BT procedure than in Johansen's method. However, this distribution theory is developed in

Bewley and Yang (1995) and this should result in more usage of the BT procedure as modified by Bewley *et al.* (1994).

Bewley *et al.* (1994) modify the original BT procedure to allow for deterministic trends and other variables explaining short-run dynamics. As in the Johansen procedure, these variables have to be factored out first by estimating regressions of Y_t and Y_{t-1} on these variables, and the residuals computed. If we call these residuals R_{0t} and R_{1t} then the equation to be estimated is

$$R_{0t} = BR_{1t} + u_t \quad (5.8)$$

What we are concerned with is the canonical correlations between R_{0t} and R_{1t} . These are obtained as the eigenvalues of the same equation as (5.6) except that R_{0t} and R_{1t} (and hence S_{00} and S_{01}) are defined differently in the BT method than in the Johansen method. In the BT method the residuals R_{0t} are obtained from a regression of Y_t on the deterministic components and short-run dynamics. In the Johansen method R_{0t} are obtained from a regression of ΔY_t on these variables.

Also, in the Johansen procedure we pick the larger eigenvalues, and consider the corresponding eigenvectors as cointegrating vectors. In the BT procedure we pick the smallest eigenvalues and consider the corresponding eigenvectors as cointegrating vectors.

The asymptotic theory for the BT procedure has only recently been developed in Bewley and Yang (1995). Until now this has been a major drawback of the BT procedure. However, some Monte Carlo studies compared the performance of the BT and Johansen's methods. Gonzalo (1994) compared the two. He concluded that the asymptotic distributions of non-ML estimators contain terms that create systematic biases in finite samples and hence concluded that the Johansen estimator is better than the others he considered (including the BT estimator).

Bewley *et al.* (1994) extend this comparison between the BT and Johansen estimators taking into account kurtosis and dispersion of the small sample distributions. They find that the BT estimator performs better than the Johansen estimator (i.e., it exhibits less kurtosis and less dispersion) when the adjustment to equilibrium is slow, the sample size is small, and the disturbances are not highly correlated (this correlation is a measure of the endogeneity of the I(1) regressor in the cointegrating regression). The relative advantage of the BT estimator declines as the sample size increases, and as the correlation in the disturbances (degree of endogeneity) increases in absolute value. Bewley *et al.* argue that

there are many interesting cases in which the performance of the BT estimator is relatively good.

The asymptotic theory of the BT estimator, derived in Bewley and Yang (1995) leads to tests for cointegration similar to those in the Johansen procedure. We shall discuss these in chapter 6 and present a comparison of the two testing procedures.

5.5.4 Principal components and cointegration

Consider a set of n variables $y_1, y_2, \dots, y_n$ with covariance matrix Σ . Consider a linear combination $d = c'y$ of the y s. We want to determine c so that it has maximum variance. Obviously we need some constraint on the c s. A constraint often imposed is $c'c = 1$. The variance of d is then $c'\Sigma c$. Thus, the problem is to maximize $c'\Sigma c$ subject to $c'c = 1$ or to maximize

$$c'\Sigma c - \lambda(c'c - 1)$$

On differentiating with respect to c and equating the derivatives to zero we get

$$\Sigma c - \lambda c = 0 \quad \text{or} \quad (\Sigma - \lambda I)c = 0 \quad \text{or} \quad |\Sigma - \lambda I| = 0$$

Thus, the problem is to find the eigenvalues of Σ . Order the eigenvalues as $\lambda_1 > \lambda_2 > \dots > \lambda_n$. Let the corresponding eigenvectors be $c_1, c_2, \dots, c_n$. Then $d_1 = c'_1 y, d_2 = c'_2 y, \dots, d_n = c'_n y$ are the principal components. We have $\text{var}(d_1) = \lambda_1, \text{var}(d_2) = \lambda_2, \dots$, etc. where d_1 is the first principal component and has the highest variance and d_2 is the second principal component and has the second highest variance, and so on. Also, the principal components are orthogonal, given the orthogonal properties of the c s. Also note that the generalized variance $|\Sigma| = \lambda_1 \lambda_2 \dots \lambda_n$.

The relationship of principal components to cointegration is now clear. Since the variance of a CI relationship is smaller compared to that of an $I(1)$ variable, the principal components corresponding to the smaller eigenvalues give the CI vectors and those corresponding to the larger eigenvalues give the common stochastic trends. The principal component approach was first used by Stock and Watson (1988) in the context of testing for common trends. Its asymptotic properties have not been published except in Gonzalo (1994) who derived the asymptotic distribution for a particular data generating process. Harris (1997) works out the asymptotic distributions of the principal components estimator. He

shows that it is consistent but asymptotically inefficient and suggests a modified principal components estimator. The modifications are similar to those used in FM-OLS, except that the adjustments made are to the data (and not to the estimator). The adjusted data are used to compute the principal components. Harris argues that the results from the Johansen procedure are sensitive to the number of lagged terms used in the VAR model. The principal components method, by contrast, does not suffer from this problem.

We shall not go into the details of Harris' paper. But given that the canonical correlation and principal components methods are commonly in use in multivariate analysis and that the idea behind principal components is an intuitively appealing one for cointegration, it is quite likely that the principal components method will gain more acceptance in the future.

5.5.5 Stock and Watson method

If there are n variables and there are k cointegrating relationships, this means that there are k linear combinations that are $I(0)$ and $n - k$ linear combinations that are $I(1)$. These are the common stochastic trends.

The Stock and Watson (1988) procedure is:

- (i) Pick the autoregressive order p .
- (ii) Compute the eigenvectors of $\sum y_t y_t'$, that is, do a *principal component analysis* of y_t .
- (iii) Using the m principal components with largest variance, that is, the largest eigenvalues, fit a vector autoregression to the differences. If P_t is the vector of m principal components, then estimate

$$\Delta P_t = A_1 \Delta P_{t-1} + \dots + A_{p-1} \Delta P_{t-p+1} + \varepsilon_t$$

Let $\hat{A}_1, \dots, \hat{A}_{p-1}$ be the estimates of the coefficient matrices.

- (iv) Compute

$$\hat{F}_t = P_t - \hat{A}_1 P_{t-1} - \dots - \hat{A}_{p-1} P_{t-p}$$

- (v) Regress ΔF_t on F_{t-1} getting the coefficient matrix B .
- (vi) Compute the eigenvalues of B , normalize, and compare with the tables in Stock and Watson (1988). Rejecting the null of m common trends in favor of $(m - q)$ common trends increases the number of cointegrating vectors by q .

5.5.6 Some problems with the Johansen procedure

The Johansen procedure depends on the assumption that the errors are independent normal. The procedure is very sensitive to this assumption. When the errors are not independent normal (see Huang and Yang, 1996), it has been found that the Johansen method has a greater probability (than least squares methods) of rejecting the null of no cointegration even when there are no cointegrating relations.

Another limitation discussed by Phillips (1994) and mentioned earlier is that the Johansen procedure produces more outliers than the other procedures.

The study by Gonzalo and Lee referred to in the next chapter also shows that the Johansen LR tests tend to find spurious cointegration in some situations of practical interest (see section 6.5.3 for other limitations of the Johansen procedure).

In summary, the major drawbacks of the Johansen procedure are: extreme sensitivity to departures from the underlying distributions of the error terms, tendency to find spurious cointegration, and high variance and high probability of producing outliers. By comparison, the least squares methods are more robust. This is not a new conclusion. The same was observed in the case of simultaneous equations estimation methods regarding FIML and least squares procedures.

5.6 The identification problem

The word *identification* has somewhat different meanings in the statistical and econometric literature. For instance Box and Jenkins talk of identification of the appropriate orders of p and q in $\text{ARMA}(p, q)$ models, or p , d , and q in $\text{ARIMA}(p, d, q)$ models. This is different from the terminology in econometrics. In the following we shall use the word identification as is customary in econometrics.

It is a well known fact that linear combinations of CI vectors are also cointegrated. Thus, when the Johansen procedure shows the existence of r CI vectors, one needs to impose some conditions arising from economic theory to give meaning to the CI vectors. Since the CI vectors constitute long-run relationships, these conditions would help identify the long-run relationship.

Johansen (1995) and Johansen and Juselius (1994) argue that there are two sets of conditions to identify the long-run and short-run relationships. Intuitively, this makes sense because there can be several

short-run dynamic paths to reach the same long-run equilibrium and one needs some identifying conditions to choose among these different paths.

The basic starting point of the whole literature on cointegration was that economic theory had very little to say about short-run economic behavior and all economic theories are about long-run behavior. If this is the case we can get very little help from economic theory to identify short-run relationships. However, if economic theory does provide information to identify the short-run relationships, then it is intuitively clear that this information would also identify the long-run relationships.

Hsiao (1997), on the other hand, argues that there is only one set of conditions that identify both the short-run and long-run relationships and these conditions are the same as those for dynamic simultaneous equation models. But if there exists a set of prior restrictions to identify the short-run relationships, then this same set of restrictions is sufficient to identify the corresponding long-run relationships. On the other hand, if there exist a set of prior restrictions to identify the long-run relationships, these are not sufficient to identify the short-run relationships.

The main reasons why there is a seeming conflict between the approaches of Johansen and Juselius (hereafter denoted as JJ) and Hsiao are as follows: Hsiao assumes that all the relevant economic information is in the structural equation system and so we can incorporate all our knowledge of the system in restrictions on this structural equation system. Since the structural equations define both the short-run and long-run behavior, the restrictions for identification are the same for both the short-run and long-run behavior. In the discussion by Hsiao the number of CI vectors is known.

The Johansen method, on the other hand, starts with a VAR model in the $I(1)$ variables and first determines the number of CI vectors. The approach is a-theoretical. Cointegration is a purely statistical concept and the CI vectors need not have any economic meaning. That is why JJ (1994, p. 8) distinguish between three concepts of identification:

- (i) *generic identification* which is related to a linear statistical model,
- (ii) *empirical identification* which is related to the estimated parameter values,
- (iii) *economic identification* which is related to the economic interpretability of the estimated coefficients of an empirically identified structure.

JJ, thus do not introduce economic theory until much later in the anal-

ysis. The question is: which of these two approach is more useful in practice. As argued earlier, economic theories are supposed to be about long-run behavior and there is very little that can be said about short-run behavior. In fact the VAR approach arose from Sims' criticism that the identification conditions imposed under the Cowles foundation approach are *ad hoc*.

Even if one adopts the Cowles foundation structural approach, the JJ long-run identification conditions are still useful, because they tell us which structural relationships, if any, can be estimated using the super-consistency property of CI relationships (or which structural parameter estimates converge at a faster rate). If we are primarily interested in the long-run behavior, then it does not matter if some short-run parameters are not identified. Thus a condition for identification of the long-run relationship is useful.

Some other discussions of the identification problems are in Campbell and Shiller (1988), Boswijk (1992), and Wickens (1996).

Campbell and Shiller make the important point that error correction models (those implied by cointegrating systems) do not necessarily reflect partial adjustment. They can also arise because one variable forecasts another. Boswijk has a detailed discussion of identification problems taking into account the relationship between exogeneity and cointegration. Wickens argues that if the VECM corresponds to a complete structural system then the long-run structural and reduced-form coefficients can be shown to be linear transformations of the cointegrating vectors. But as mentioned earlier, the Johansen procedure starts with a VECM with no structural interpretation. Identifying information is introduced much later – after determining the number of CI vectors.

5.7 Finite sample evidence

There have been many Monte Carlo studies that investigate the finite sample performance of the estimates obtained by the different estimation methods discussed in the previous sections. Practitioners in this area are faced with two questions:

- (i) What estimation method should we use?
- (ii) Given the choice of the estimation method, what significance levels, p -values, etc. should we use for inference on the estimated coefficients?

The Monte Carlo studies to be reviewed here are concerned with question (i). Issues concerning question (ii) will be discussed in the next chapter and in chapter 10 on bootstrap methods.

We shall first review some of the Monte Carlo studies and then outline some points on which there is a consensus.

Banerjee, Dolado, Hendry, and Smith (1986)

This is the first Monte Carlo study that investigates the relevance for econometric practice of asymptotic theory in this area. The main problem considered in this paper is whether the Engle–Granger methods of estimating the cointegrating vector through a static regression gives good results. Another problem they investigate is that of contrasting inferences drawn from static and dynamic regressions.

The DGP they consider is: the same as that in Engle–Granger (1987) which is as follows

$$\begin{aligned}x_t + y_t &= v_t, & v_t(1 - \rho_1 L) &= e_{1t} \\ y_t + 2x_t &= u_t, & u_t(1 - \rho_2 L) &= e_{2t}\end{aligned}$$

and

$$\begin{pmatrix} e_{1t} \\ e_{2t} \end{pmatrix} \sim iin \left[\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} \sigma_1^2 & 0 \\ 0 & \sigma_2^2 \end{pmatrix} \right]$$

The reduced form consists of the two regressions

$$\begin{aligned}y_t &= \frac{2e_{1t}}{1 - \rho_1 L} - \frac{e_{2t}}{1 - \rho_2 L} \\ x_t &= \frac{e_{2t}}{1 - \rho_2 L} - \frac{e_{1t}}{1 - \rho_1 L}\end{aligned}$$

If λ is the coefficient of x_t in a static regression of y_t on x_t , then

$$plim \hat{\lambda} = -1 - \frac{1}{\theta}$$

where

$$\theta = 1 + \frac{\sigma_2^2(1 - \rho_1)^2}{\sigma_1^2(1 - \rho_2)^2}$$

Note that as $\rho_1 \rightarrow 1$, $plim \hat{\lambda} = -2$ and as $\rho_2 \rightarrow 1$, $plim \hat{\lambda} = -1$. If $|\rho_1| < 1$ and $|\rho_2| < 1$, both x_t and y_t are $I(0)$ variables. If $\rho_1 = 1$ and $\rho_2 = 1$, both x_t and y_t are $I(1)$ variables, but not cointegrated. If $\rho_1 = 1$ and $|\rho_2| < 1$, then both variables are $I(1)$ and y and x are cointegrated. We have in this case $plim \hat{\lambda} = -2$.

The parameter values Banerjee *et al.* consider in their Monte Carlo study are

$$\begin{aligned} T &= (33, 66, 99, 150, 199) \\ s &= \sigma_1/\sigma_2 = (16, 8, 4, 2, 1, 0.5) \\ \rho_2 &= (0.6, 0.8, 0.9) \end{aligned}$$

Although $\text{plim } \hat{\lambda} = -2$ and $\hat{\lambda}$ is superconsistent, these authors find that for small T and large values of ρ_2 and s , the biases are very large. For example if $T = 33, s = 2, \rho_2 = 0.9$, the average bias is 0.589. If $T = 66$, the bias reduces to 0.468.

Banerjee *et al.* next consider dynamic models and compare the estimates of the cointegrating coefficient from static and dynamic regressions. The main conclusion is that the biases in the estimates of the long-run parameter are larger in the static model than in the dynamic model.

These points are pursued in Banerjee *et al.* (1993, chapter 7) where they plot the biases in the estimates of the long-run parameter. The conclusion is that the biases increase with ρ_2 and s . They conclude that some nonparametric corrections as in Phillips and Hansen (1990) and in Phillips and Loretan (1991), or additional dynamics as in Hendry's ECM are necessary to correct the biases in the estimates from static regressions. The issue then is, which of these is better? The following studies try to answer this question.

Phillips and Hansen (1990)

The Monte Carlo study of Banerjee *et al.* (1986) found that superconsistency of OLS in cointegrating regressions was misleading in small samples. The study implies that asymptotic theory seems to provide a poor approximation in sample sizes typical of economic data. Phillips and Hansen (1990) argue that this is not true. Superconsistency in itself does not give any information on the sampling distribution. The asymptotic distribution theory, they argue, is useful even in small samples in choosing between different estimators and test statistics. They illustrate this through a simulation study. They compare the OLS, the FM-OLS and Hendry's dynamic regression (see section 5.4).

The DGP they used was the triangular system

$$\begin{aligned} y_{1t} &= \beta y_{2t} + u_{1t} \\ \Delta y_{2t} &= u_{2t} \end{aligned}$$

$$\begin{bmatrix} u_{1t} \\ u_{2t} \end{bmatrix} = u_t = e_t + \theta e_{t-1}, \quad e_t \sim iin(0, \Sigma)$$

$$\theta = \begin{bmatrix} 0.3 & -0.4 \\ \theta_{21} & 0.6 \end{bmatrix} \quad \Sigma = \begin{bmatrix} 1 & \sigma_{21} \\ \sigma_{21} & 1 \end{bmatrix}$$

They set $\beta = 2, T = 50$ and let θ_{21} and σ_{21} vary. The parameters chosen were $(0.8, 0.4, 0)$ for θ_{21} and $(-0.8, -0.4, 0.4, 0.8)$ for σ_{21} . They report means and standard deviations of $\hat{\beta} - \beta$ and $t_{\hat{\beta}}$. All simulations used 30,000 replications. Their conclusion is that overall, the FM-OLS and ECM methods work well and improve over the OLS. In general, OLS is the most biased estimator. The FM-OLS method shows a small but persistent bias and seems generally preferable to Hendry's method.

Phillips and Hansen note that the relatively inferior behavior of the dynamic regression method may have been caused by an insufficient number of lagged variables. Banerjee *et al.* (1993, pp. 248–251) pursue this argument and show that the superiority of the FM-OLS method is due to the particular DGP used in this study and depends on some parameter values. In general no clear ranking emerges.

Hansen and Phillips (1990)

The DGP used in this Monte Carlo study is the same as that used by Banerjee *et al.* (1986) and Engle and Granger (1987). The purpose of the study is to explore the small sample properties of OLS and IV estimators in cointegrating regressions. The authors consider the unadjusted, bias corrected, and fully modified versions of the OLS and IV estimators in order to evaluate the asymptotic theory developed in Phillips and Hansen (1990).

The fully modified estimators have some advantages over the unadjusted and bias corrected estimators. Bias corrections are complicated by the fact that they depend on preliminary coefficient estimates which are themselves biased.

The DGP used in this exercise was decomposed to highlight the problems into three categories: signal-noise ratio, serial correlation, and endogeneity. The DGP was

$$\begin{aligned} y_t - 2x_t &= u_t, & (1 - \rho L)u_t &= e_{1t} \\ -y_t + 3x_t &= v_t, & (1 - L)v_t &= e_{2t} \end{aligned}$$

$$\begin{pmatrix} e_{1t} \\ e_{2t} \end{pmatrix} \sim iin \left[\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 1 & \sigma\theta \\ \sigma\theta & \sigma^2 \end{pmatrix} \right]$$

$$|\rho| < 1, \quad |\theta| < 1$$

The signal–noise ratio is given by σ (we shall comment on this later), serial correlation by ρ and endogeneity by θ . The general conclusion of the study was that the signal–noise ratio (σ) appeared to be the critical factor, not the degree of long-run endogeneity. If σ is very high, the bias problem is negligible and OLS works well. In this case the fully modified (FM) estimates will permit inference in a conventional fashion. If σ is low, the FM-OLS method does not perform well and IV methods are necessary.

Phillips and Loretan (1991)

Phillips and Loretan report results of simulations for the following procedures:

- (i) OLS,
- (ii) FM-OLS,
- (iii) Linear ECM or Hendry type estimator,
- (iv) Their own estimator (PL) based on nonlinear regression that includes leads of Δy_{2t} .

(For a discussion of these estimators, see section 5.4.2.)

The DGP they use is identical to that used in Phillips and Hansen (1990) described earlier. They set $\beta = 2.0$ and $T = 50$ and allow θ_{21} and σ_{21} vary. They consider the values $(0.8, 0.4, 0, -0.8)$ for θ_{21} and $(-0.85, -0.5, 0.5)$ for σ_{21} . The model has only one regressor and the time series dynamics are generated by MA(1).

The major finding of the Phillips and Loretan simulation study is that FM-OLS and single equation ECM estimates are both substantially better than OLS. The performance of the ECM estimator is itself substantially improved by (i) adding more lags, (ii) including leads of the differences of the regressor variable when these are not strongly exogenous, and (iii) most importantly by formulating the ECM nonlinearly in the parameters. However, Phillips and Loretan find that there are substantial size distortions in inferences based on ECM estimates.

Inder (1993)

Inder argues that the favorable results obtained by Phillips and Hansen (1990) and Phillips and Loretan (1991) for the FM-OLS method are due to the particular DGP they used in their simulation. From his Monte Carlo study he shows that estimates which include the dynamics (like the

single equation ECM) are much more reliable, even if the dynamic structure is overspecified. Also even if the t -statistics based on the FM-OLS method are asymptotically valid, they do not have good finite sample properties.

Inder's DGP includes a lagged dependent variable. He argues that in this case FM-OLS yields almost no improvement over OLS, and that the semi-parametric correction is insufficient to remove the autocorrelation in the errors when the DGP includes a lagged dependent variable. These problems do not arise with the unrestricted ECM (Hendry type) whose performance is uniformly better. This argument is pursued in Inder (1995) where he argues for a linear model with leads and lags and shows how it performs better than the nonlinear model in Phillips and Loretan (1991). These papers argue for Hendry style dynamic regression models. In all these methods there is the question of selection of lag length as discussed in section 5.4.2.

Entorf (1992)

This paper illustrates the finite sample problems of regression models with $I(1)$ variables. In particular, it emphasizes the role of drifts.

Consider the cointegrated system as follows

$$\begin{aligned} y_t &= \alpha + \beta x_t + u_t \\ \Delta y_t &= \gamma_y + e_t, \quad e_t \sim iid(0, \sigma_y^2) \\ \Delta x_t &= \gamma_x + v_t, \quad v_t \sim iid(0, \sigma_x^2) \end{aligned}$$

Then, if it is estimated by OLS, as $T \rightarrow \infty$

$$\hat{\beta} \rightarrow \frac{\gamma_y}{\gamma_x}$$

and

$$\begin{aligned} T^{-1/2} \hat{\alpha} &\Rightarrow \sigma_y \int_0^1 W(r) dr - \frac{\gamma_y}{\gamma_x} \sigma_x \int_0^1 V(r) dr \\ &\sim N \left[0, \frac{\sigma_y^2 + \sigma_x^2 (\gamma_x / \gamma_y)^2}{3} \right] \end{aligned}$$

where $W(r)$ and $V(r)$ are independent standard Wiener processes. The last expression is derived using the result (see chapter 3)

$$\int_0^1 W(r) dr \sim N \left(0, \frac{1}{3} \right)$$

In the presence of drifts we can write

$$y_t = y_0 + \gamma_y t + e_t + e_{t-1} + \cdots + e_1$$

$$x_t = x_0 + \gamma_x t + v_t + v_{t-1} + \cdots + v_1$$

Thus it is clear that $\beta = \gamma_x/\gamma_y$ and since $\hat{\beta}$ can be shown to be consistent we get the required result. Entorf shows that if a time trend is included, that is, if the estimated equation is

$$y_t = \hat{\alpha} + \hat{\beta}x_t + \hat{\delta} + \hat{u}_t$$

then

$$\hat{\delta} \rightarrow \gamma_y$$

and $\hat{\beta}$ converges to a random variable as in the case of a spurious regression with no drifts.

Entorf investigates the rate of convergence of the OLS estimators to those predicted by asymptotic theory. He takes $\sigma_y = \sigma_x = 1$ and $\beta = \gamma_y/\gamma_x = 2.0$. He considers $\gamma_x = 0.05, 0.25, 0.75$ with the corresponding values of $\gamma_y = 0.10, 0.50, 1.50$. For the smallest drift the asymptotic results are not valid even for samples of size 170 or higher. For the largest value of drift the asymptotic results are valid for samples as small as 30. Thus, the magnitude of drifts plays an important role.

The behavior of the DW statistic has been commonly suggested as a guide to ensure that a spurious regression has not been estimated. Entorf investigates the small sample distribution of the DW statistic. He finds that for more than five regressors the 95 percent fractile of the DW distribution is larger than two. Thus even regressions with DW values of about two do not necessarily ensure that we do not estimate spurious regressions.

Gonzalo (1994)

Gonzalo considers five estimators from:

- (i) Engle–Granger static regression model,
- (ii) the nonlinear least squares method by Phillips and Loretan,
- (iii) the Johansen procedure,
- (iv) principal components method,
- (v) the Box–Tiao method (see section 5.5.3) due to Bossaerts (1988) and Bewley and Orden (1994).

He derives the asymptotic distributions for these estimators. From the Monte Carlo study he shows that the Johansen procedure performs better than the others even when the errors are not normally distributed or when the dynamics are unknown and we overparameterize by including additional lags in the ECM. Needless to say, these conclusions are not in agreement with many other studies we have reviewed.

It should be noted that Gonzalo considered, in his Monte Carlo study, the procedure in Johansen (1988) which does not have a constant term. The method in Johansen (1991) which includes a constant is more empirically relevant. Also, Gonzalo was the first to derive the asymptotic distribution of the principal component estimator but in a limited model. The paper by Harris (1997) does this for a more general model.

Bewley *et al.* (1994)

These authors compare, by Monte Carlo studies the performance of the Box–Tiao method compared to the Johansen method and conclude that in a variety of situations, the distributions of the Box–Tiao estimator are less dispersed than those of the Johansen estimator.

Cappuccio and Lubian (1994)

These authors argue that consistent estimation of the long-run covariance matrix via VAR prewhitening results in a considerable improvement in the performance of the FM-OLS estimator and it performs better than the Johansen estimator. They find that even in samples of 200, the empirical distribution of the Johansen estimator is affected by the Cauchy like factor noted by Phillips (1994) when the error correction mechanism exhibits strong serial correlation.

Eitrheim (1992)

This is an ambitious study that examines the effects on the Johansen estimator of several model misspecifications such as:

- (i) wrong order of the VAR model,
- (ii) nonnormality of the errors,
- (iii) time aggregation and skip sampling,
- (iv) measurement errors.

The study concludes that provided the cointegration rank r is correctly specified, the Johansen estimator is robust to these specification errors. The problems considered in this study are very extensive and more detailed Monte Carlo studies are needed to examine each of the issues.

Hargreaves (1994)

In his Monte Carlo study with a four equation system with one cointegrating vector and two cointegrating vectors, Hargreaves (1994) finds that:

- (i) The Johansen estimator had the best median bias but the worst variation.
- (ii) Suppose we estimate a single cointegrating vector when in fact there were two cointegrating vectors. If the cointegrating vector we are estimating has a much lower variance than the other cointegrating vector which is ignored, then one will do fairly well. In the opposite case one will do badly.
- (iii) If one is unsure of the cointegrating dimensionality and is trying to estimate any one cointegrating vector, then the FM-OLS estimator is the best.
- (iv) The Johansen estimator is the best only if the model is well specified without highly autocorrelated cointegrating errors.

Phillips (1994)

The paper by Phillips is not a Monte Carlo study but it is helpful in understanding the results of several Monte Carlo studies that report that the Johansen estimator, although almost unbiased, has a very high variance. Phillips derives same exact finite sample distributions and characterizes the tail behavior of ML estimators of the cointegrating coefficients in VECM models. He shows that the distribution of the Johansen estimator has Cauchy-like tails and hence has no finite moments of integer order. The ML estimation in Phillips' triangular representation, by contrast, has a distribution that has matrix t -distribution type tails with finite integer moments of order $T - n + r$ where T is the sample size, n the number of variables in the system, and r is the dimension of the cointegration space.

These results of Phillips help to explain the fact that several Monte Carlo studies found more extreme outliers with the Johansen method than with other asymptotically efficient methods.

Summary of the Monte Carlo studies

It can be easily seen that there are only a few points on which there is an agreement in all these Monte Carlo studies. The unambiguous conclusions we can draw are that estimation of the long-run parameters

by static regressions is to be avoided and that the Johansen estimator (though showing less bias than the other estimators) exhibits large variation.

Among single equation methods, where estimation of a single CI vector is of concern, on balance a linear model with leads and lags appears to be the best choice. Among systems methods, where we also estimate the number of CI vectors, there are two methods available: the Johansen method and the Box-Tiao (BT) method. The BT estimator performs better (has less dispersion and less kurtosis) when the adjustment to equilibrium is slow, sample size is small, and disturbances are not highly correlated (the correlation is a measure of the endogeneity of the $I(1)$ regressors in the CI regression). When these conditions are not met, the Johansen method should be used.

There is one issue that makes it difficult to compare the results from the different studies and from different parameter combinations in the same study. Li and Maddala (1995) suggest that in the choice of the parameters in the DGP for Monte Carlo studies, attention should be paid to the signal-noise ratio (SNR). There is the question of what should be an appropriate measure of SNR in cointegrated systems and this is an issue that needs careful study. Earlier, we noted that Hansen and Phillips (1990) argued that SNR was a critical factor in their Monte Carlo study.

Li and Maddala (1995) suggest defining SNR as the ratio of the innovation of $I(1)$ errors to the variance of the $I(0)$ regression errors. They present SNRs for a few of the Monte Carlo studies discussed earlier and show that there is a large variation in SNR between the studies. More detailed study of the role of SNR in comparing the results of the different Monte Carlo studies is needed.

5.8 Forecasting in cointegrated systems

In section 3.10 of chapter 3, we discussed issues related to forecasting from single equations. We shall now turn to multiple equations, the VAR system. One important contribution of cointegration tests is in the modeling of VAR systems, whether they should be in levels or first-differences or both, with some restrictions. For this purpose the cointegration relationships need not have any economic interpretation. They are of value in determining the restrictions of the VAR system which should be of value in forecasting. We shall now discuss whether and when the cointegration restrictions improve the forecasting.

If a set of nonstationary variables satisfies a cointegration relation, simple first-differencing of all the variables can lead to econometric problems. In the general VAR system with n variables, if all the variables are stationary, using an unrestricted VAR in levels is appropriate. If the variables are all $I(1)$ but no cointegration relation exists, then application of an unrestricted VAR in first differences is appropriate. If, in addition, there are r cointegrating relationships, then we need to model the system as a VAR in the r stationary combinations and $(n - r)$ differences of the $I(1)$ variables.

Engle and Yoo (1987) compare the mean squared forecast errors for cointegrated system with the forecasts generated by unrestricted VAR. This may not be a valid comparison because it is well known that unrestricted VARs do not give good forecasts. What we need is a comparison of forecasts from the cointegrated system (which is essentially a restricted VAR) with those from other restricted VARs such as Bayesian VARs, or alternative forms of error correction models (besides those that result from the cointegration relationships). It is, however, interesting to note that in the Engle and Yoo study, for forecast horizons of up to five, the unrestricted VAR did better than the Engle-Granger two step method applied to the ECM. It is only for longer horizon forecasts that the latter was better. Note that the VAR in levels is not appropriate if all the variables are $I(1)$.

Engle and Yoo consider a two-variable system with the autoregressive representation

$$\begin{bmatrix} 1 - 0.6L & -0.8L \\ -0.1L & 1 - 0.8L \end{bmatrix} \begin{bmatrix} x_t \\ y_t \end{bmatrix} = e_t, \quad e_t \sim iid \left[0, \begin{pmatrix} 100 & 0 \\ 0 & 100 \end{pmatrix} \right]$$

The corresponding ECM representation is

$$(1 - L) \begin{pmatrix} x_t \\ y_t \end{pmatrix} = \begin{pmatrix} -0.4 \\ 0.1 \end{pmatrix} (1 - \alpha) \begin{pmatrix} x_{t-1} \\ y_{t-1} \end{pmatrix} + e_t$$

with $\alpha = 2$. The long-run relationship is $x = 2y$. One hundred replications were computed in each case with 100 observations used for fitting the model and 20 observations used for post-sample forecasting.

Given that one is not often interested in forecasting over long horizons, the Engle and Yoo study actually suggests that one does not gain much by imposing the cointegration restrictions on the VAR. Actually, the Engle-Yoo study uses the information that there is one cointegrating relationship. In practice we would first conduct a test whether cointe-

gration exists or not. Also one should compare the forecasts from the VAR in first-differences (no cointegration).

Clements and Hendry (1995) do another Monte Carlo study with the Engle–Yoo bivariate DGP. Since the conclusions in these studies are sensitive to the particular DGP used, they try several DGPs. They also design an efficient Monte Carlo experiment using antithetic variates and control variates (for a discussion of these methods, see Hendry, 1984). The sample sizes were 50 and 100 and the forecast horizons were 1, 5, 10, and 20 periods ahead. In addition to the unrestricted VAR (UV) and the Engle–Granger two-step method (EG), Clements and Hendry also consider the ML method (that incorporates the restrictions in the ECM) and also VAR in differences (DV). Clements and Hendry evaluate the different forecasting methods for two sets of variables. The first is the original set of $I(1)$ variables, say $x_t = (x_{1t}, x_{2t})$. The other is a set of $I(0)$ variables, say $w_t = (w_{1t}, w_{2t})$ where $w_{1t} = x_{1t} + \alpha x_{2t}$ is the cointegrating equation and $w_{2t} = \Delta x_{2t}$.

Engle and Yoo conjectured that the ML method (which imposes cross-equation constraints) would dominate the UV and EG methods. Clements and Hendry find that this is the case only if we consider x_t (not w_t). Engle and Yoo also conjecture that the DV method (which is clearly misspecified in the presence of cointegration) performs worse than the UV method. Clements and Hendry find that this is not the case if we are forecasting x_t . The DV method is in fact almost as good as the ML, and it outperforms the UV method at long horizons. However, when it comes to forecasting w_t , the DV method is much worse than the UV method.

Given that we are not usually interested in long forecast horizons, the following conclusions emerge:

- (i) The Engle–Yoo study compares UV and EG methods. For forecast horizons of five or less, UV is better.
- (ii) Clements and Hendry consider the UV, EG, ML, and DV methods. For shorter ($n < 10$) horizons and forecasting x_t , the ranking in decreasing order is ML, EG, UV, and DV. For forecast horizons of ten or higher, the ranking is ML, EG, DV, and UV. For forecasting w_t , ML and UV are about the same, the EG method is slightly worse than UV, and the DV method is substantially worse than UV.

All this suggests that the DV method is not often appropriate, and, except for long horizons, one might just as well use the UV method (un-

restricted VAR). This means that one need not impose the cointegration restrictions on the VAR models for the purpose of forecasting.

Some other studies make the opposite argument. In a Monte Carlo study, Ohanian (1990) reports that even wrongly imposing unit roots produces better forecasts than models that do not impose unit roots. In an empirical study LeSage (1990) argues that even in the absence of cointegration, the ECM and Bayesian VAR models with error correction terms added produce good forecasts, especially at longer horizons (eight to twelve months ahead). Thus he argues in favor of wrongly imposing cointegration restrictions.

Reinsel and Ahn (1992) consider forecasting in a Monte Carlo experiment of a four-variate AR(2) model with two unit roots (and thus two cointegration relationships). As a measure of forecast performance, they use the trace of the prediction mean square error matrix estimated from post-sample prediction errors. They consider predictions with $d = 0, 1, 2, 3$, and 4 imposed where d is the number of unit roots. The extreme cases $d = 0$ and $d = 4$ gave very poor forecasting performance. In cases $d = 1$ and $d = 3$, the forecasting performance was fairly good (compared with the correct model $d = 2$) with the model $d = 1$ providing slightly better short-run forecasts relative to the model $d = 3$ which provides better long-run forecasts. Thus imposing too much cointegration and too little cointegration is bad. But moderate departures are alright.

The preceding discussion gives a survey of the different methods of forecasting from cointegrated systems. However, none of these papers give an overall set of guidelines to be used by time series practitioners. Such guidelines with an empirical illustration are given in Joutz, Madala, and Trost (1995). This involves checking the data for the presence of unit roots and cointegration, estimating the implied long-run relationships by dynamic models and then specifying and testing the VAR model.

5.9 Miscellaneous other problems

5.9.1 Models with exogenous variables

In the Johansen procedure and other procedures discussed in the earlier sections, we considered *closed-form models* where all variables depended on one another (or in standard econometric terminology, all variables are endogenous). In fact this is one of the problems in the

different approaches to identification discussed in section 5.6). In a lot of economic models, however, certain variables can be treated as weakly exogenous for the estimation of long-run relationships among other variables. The estimation of the long-run relationships can be conducted conditional on these variables. When it is known that certain variables are weakly exogenous, neglecting this information can result in a loss of power. This problem has been discussed in Johansen (1992b) and Urbain (1992, 1993). It is shown that if a given variable is weakly exogenous for the long-run parameters (i.e., the parameters α and β in the notation of section 5.5.1), the cointegrating vectors of interest must not appear in the generating model of that variable, i.e., the variable must not be error correcting. Tests for weak exogeneity, as well as Wald, LR, and LM tests for testing cointegration in these models are discussed in the next chapter. Kleibergen (1995) uses these procedures for the estimation of import demand models and shows how treating oil price as weakly exogenous decreases the long-run memory of the inflation rate.

5.9.2 Cointegration and Granger causality

“When I say a word, it is supposed to mean
exactly what I want it to mean,
nothing more, nothing less.”

The mad hatter in *Alice in Wonderland*.

Suppose that there are two time series x_t and y_t . When the past and present values of y_t provide some useful information to forecast x_{t+1} at time t , it is said that y_t Granger causes x_t . Granger causality has been used in the context of rational expectations, definition of super exogeneity, and econometric modeling strategy. Many testing procedures for Granger causality have been proposed; in practice a commonly used testing procedure for Granger causality is testing the significance of the coefficients of lagged y_t , which are used as the explanatory variables for x_t in the regression context.

A better term for Granger causality is *precedence*. But like the Mad Hatter in *Alice in Wonderland*. Granger has chosen the word *causality* for precedence. Although he has amply made it clear in several papers what exactly the term means, some of the literature on Granger causality has used results from causality tests to infer causality, as it is commonly understood. An example that drives home what exactly

Granger causality implies is the statement: *The weatherman's prediction about rain (Granger) causes the rain.*

Many of the studies conducted on Granger causality were conducted in a strictly bivariate framework, even though Granger himself warned that omission of relevant other variables could result in spurious causality (see Granger, 1969, p. 429). We fully concur with the summary assessment by Adrian Pagan (1989) of the work on Granger causality:

There was a lot of high powered analysis of this topic, but I came away from a reading of it with the feeling that it was one of the most unfortunate turnings for econometrics in the last two decades, and it has probably generated more nonsense results than anything else during that time.

Is there any relationship between Granger causality and cointegration? The answer to this question has been discussed in Granger (1988). As we have discussed, cointegration is concerned with long-run equilibrium. On the other hand, Granger causality is concerned with short-run forecastability. These two different concepts can be considered in an error correction model (ECM). Suppose that x_t, y_t are both $I(1)$ variables and they are cointegrated such that $z_t = x_t - \beta y_t$ is $I(0)$. When $x - \beta y = 0$ can be considered as a long-run (steady-state) equilibrium, z_t can be interpreted as the extent to which the system is out of equilibrium. As we have seen in section 2.8, this cointegrated system can be written in the form of ECM as follows

$$\begin{aligned}\Delta x_t &= \gamma_1 z_{t-1} + \text{lagged } (\Delta x_t, \Delta y_t) + \varepsilon_{1t} \\ \Delta y_t &= \gamma_2 z_{t-1} + \text{lagged } (\Delta x_t, \Delta y_t) + \varepsilon_{2t}\end{aligned}$$

where one of $\gamma_1, \gamma_2 \neq 0$ and $\varepsilon_{1t}, \varepsilon_{2t}$ are finite-order moving-averages. In the ECM, we can find that Δx_t or Δy_t (or both) must be Granger caused by z_{t-1} which is itself a function of x_{t-1}, y_{t-1} . Thus, either x_{t+1} is Granger caused by y_t or y_{t+1} by x_t . This implies that for a pair of series to have an attainable equilibrium, there must be some Granger causation between them to provide the necessary dynamics. Note that if the lagged Δx_t and Δy_t have nonzero coefficients then there is causality in both directions.

Of what use is this relationship between Granger causality and cointegration? Granger shows that this does not change causality tests. However, when estimating cointegrated relationships, it would be of use in the normalization issue (discussed earlier in section 5.3).

5.9.3 Threshold cointegration

In the previous section, we assumed that if two $I(0)$ variables y_t and x_t are cointegrated with cointegrating coefficient β , then the disequilibrium error $z_t = y_t - \beta x_t$ is stationary or *mean-reverting*. The literature on cointegration and the corresponding error correction models is that this adjustment to long-run equilibrium is continuous. However, in many instances, due to transaction costs or institutional constraints (for instance, dividend adjustments, or exchange rate behavior in the presence of target zones) this adjustment is not continuous, but takes place only if the equilibrium error z_t reaches a certain threshold θ . For instance, z_t could be a random walk if $|z_t| \leq \theta$, but z_t could follow a stationary autoregressive process if $|z_t| > \theta$. Such models are called *threshold cointegration models* and are analyzed in Balke and Fomby (1997) (although no empirical illustration is provided).

Since globally the system is a cointegrated system, Balke and Fomby suggest first estimating the cointegration relationship between y_t and x_t using Engle–Granger procedure (i.e., using OLS) and then testing for the existence of threshold behavior in the estimated equilibrium error $\hat{z}_t = y_t + \hat{\beta}x_t$. Balke and Fomby present Monte Carlo evidence to argue that this procedure works well.

However, considering the large biases in the estimation of long-run parameters by using static regression (as discussed in section 5.7) it might be desirable to use a dynamic regression at the first stage to get estimates of the cointegrating parameter β and the equilibrium error z_t . Thus, further modifications to the Balke–Fomby method need to be investigated.

5.9.4 Aggregation in cointegrated systems

When we talk of aggregation, we have to distinguish between temporal aggregation and cross-sectional aggregation. Temporal aggregation occurs when a variable is generated say, over a month but is only observed quarterly or yearly. The question is what model will temporally aggregated data obey. The answer will depend on whether the variable is a stock or flow. Cross-sectional aggregation occurs when we have a number of micro variables and we aggregate them to get a macro variable.

Granger (1990) discusses the effects of aggregation on cointegration. He shows that integration and cointegration are not affected by temporal aggregation. As for cross-sectional aggregation, its effect on cointegra-

tion is complex. It is possible to have cointegration at the aggregate level and not at the disaggregate level and it is possible to have cointegration at the micro level but not at the macro level. There are special conditions, as investigated by Gonzalo (1993) under which cointegration at the micro level implies cointegration at the macro level and vice versa. These conditions depend on some common factor assumptions rather than the homogeneity of the micro units.

5.10 Summary and conclusions

This chapter discussed several single equation methods and system methods for the estimation of CI relationships. The advantage with the system methods is that they also determine the number of CI relationships. In practice this creates an identification problem because linear combinations of CI vectors are themselves cointegrated. This necessitates some identification conditions to be imposed so as to give meaningful economic interpretation to the estimated cointegrated vectors. The issues in this regard, relate to the stage of analysis at which these extra identifying conditions are imposed.

This chapter also presents a review of several Monte Carlo studies done on the different single equation and system methods for the estimation of CI vectors. Although very few unambiguous conclusions follow from these studies, we can say that, in general, among single equation methods, an equation with leads and lags is to be preferred. The choice between the two system methods (Johansen and Box–Tiao) depends on a number of factors. In practice, only the Johansen method has been widely used.

Of critical importance (in both the single equation and system methods) is the selection of lag length. This has not received as much attention as necessary. It is customary to choose the lag length using some information criteria (AIC or BIC), or sequentially increasing the number of lags (as done in Inder, 1995) but a preferable approach is to use a general to specific modeling strategy starting with a large number of lags, as described in section 3.6.2 of an earlier chapter.

References

- Ahn, S.K., and G.C. Reinsel (1988), “Nested Reduced-Rank Autoregressive Models for Multiple Time Series,” *Journal of the American Statistical Association*, 83, 849–856.

- (1990), "Estimation for Partially Non-Stationary Autoregressive Models," *Journal of the American Statistical Association*, 85, 813–823.
- Algoskoufis, G. and R. Smith (1991), "On Error Correction Models: Specification, Interpretation, Estimation," *Journal of Economic Surveys*, 5, 97–128.
- Anderson, T.W. (1951), "Estimating Linear Restrictions on Regression Coefficients for Multivariate Normal Distributions," *Annals of Mathematical Statistics*, 22, 327–351.
- Balke, N.S. and T.B. Fomby (1997), "Threshold Cointegration," *International Economic Review*, 38, 627–646.
- Banerjee, A., J.J. Dolado, D.F. Hendry and G.W. Smith (1986), "Exploring Equilibrium Relationships in Econometrics Through Static Models: Some Monte Carlo Evidence," *Oxford Bulletin of Economics and Statistics*, 48, pp. 253–278.
- Banerjee, A., J.J. Dolado, J.W. Galbraith, and D.F. Hendry (1993), *Cointegration, Error Correction, and the Econometric Analysis of Non-Stationary Data*, Oxford University Press, Oxford.
- Bewley, R. (1979), "The Direct Estimation of the Equilibrium Response in a Linear Dynamic Model," *Economic Letters*, 3, 357–361.
- Bewley, R. and D. Orden (1994), "Alternative Methods for Estimating Long-run Responses with Applications to Australian Import Demand," *Econometric Reviews*, 13, 179–204.
- Bewley, R. and M. Yang (1995), "Tests for Cointegration Based on Canonical Correlation Analysis," *Journal of American Statistical Association*, 90, 990–996.
- Bewley, R., D. Orden, M. Yang, and L.A. Fisher (1994), "Comparison of Box–Tiao and Johansen Canonical Estimators of Cointegrating Vectors in VEC(1) Models," *Journal of Econometrics*, 64, 3–27.
- Bossaerts, P. (1988), "Common Nonstationary Components of Asset Prices," *Journal of Economic Dynamics and Control*, 12, 347–364.
- Boswijk, H.P. (1992), *Cointegration, Identification and Exogeneity*, Tinbergen Institute Thesis Publishers, Amsterdam.
- Box, G.E.P. and G.C. Tiao (1977), "A Canonical Analysis of Multiple Time Series," *Biometrika*, 64, 355–365.
- Campbell, J.Y. and P. Perron (1991), "Pitfalls and Opportunities: What Macroeconomists Should Know about Unit Roots," in

- O.J. Blanchard and S. Fisher (eds.), *NBER Macroeconomics Annual*, The MIT Press, 141–201.
- Campbell, J.Y. and R. Shiller (1988), “Interpreting Cointegrated Models,” Working paper No. 2568, NBER.
- Cappuccio, N. and D. Lubian (1994), “Estimating Long-Run Equilibria in Finite Samples,” manuscript, University of Padova, Italy (revision of paper presented at Econometric Society Meetings, Uppsala, Sweden, 1993).
- Clements, M.P. and D.F. Hendry (1995), “Forecasting in Cointegrated Systems,” *Journal of Applied Econometrics*, 10, 127–146.
- Eitrheim, O. (1992), “Inference in Small Cointegrated Systems: Some Monte Carlo Results,” Presented at the Econometric Society Meeting in Brussels.
- Elliott, G. (1995), “On the Robustness of Cointegration Methods When Regressors Almost Have Unit Roots,” Discussion paper 95-18, UCSD.
- Elliott, G. and J.H. Stock (1994), “Inference in Time Series Regression When the order of Integration of a Regression is Unknown,” *Econometric Theory*, 10, 672–700.
- Engle, R.F. and C.W.J. Granger (1987), “Cointegration and error correction: representations, estimation and testing,” *Econometrica*, 55, 252–276.
- Engle, R.F. and B.S. Yoo (1987), “Forecasting and Testing in Cointegrated Systems,” *Journal of Econometrics*, 35, 143–159.
- (1991), “Cointegrated Economic Time Series: An Overview with New Results,” in R.F. Engle and C.W.J. Granger (eds.), *Long Run Economic Relationships: Reading in Cointegration*, Oxford University Press, Oxford, chapter 12.
- Entorf, H. (1992), “Random Walks with Drift, Simultaneous Equation Errors, and Small Samples: Simulating the Birds-Eye View,” INSEE Discussion paper No. 9222.
- Ericsson, N.R., J. Campos, and H.A. Tran (1990), “PC GIVE and David Hendry’s Econometric Methodology,” *Revista de Econometria*, 10, 7–47.
- Escribano, A. and D. Pena (1994), “Cointegration and Common Factors,” *Journal of Time Series Analysis*, 15, 577–586.
- Gonzalo, J. (1993), “Cointegration and Aggregation,” *Ricerche Economiche*, 47, 281–291.
- (1994), “Five Alternative Methods of Estimating Long Run Equilibrium Relationships,” *Journal of Econometrics*, 60, 203–233

- Granger, C.W.J. (1969), "Investigating Causal Relations by Econometric Models and Cross-Spectral Methods," *Econometrica*, 37, 424–438.
- (1988), "Some Recent Developments in a Concept of Causality," *Journal of Econometrics*, 39, 199–211.
- (1990), "Aggregation of Time Series Variables: A Survey," in T. Barker and M.W. Pesaran (eds.), *Disaggregation in Economic modeling*, Ratledge, New York.
- Hansen, B.E. and P.C.B. Phillips (1990), "Estimation and Inference in Models of Cointegration: A Simulation Study," in T.B. Fomby and G.F. Rhodes (eds.), *Advances in Econometrics: Cointegration, Unit Roots and Spurious Regression*, vol. 8, JAI Press.
- Hargreaves, C. (1994), "A Review of Methods of Estimating Cointegrating Relationships," in C. Hargreaves (ed.), *Nonstationary Time Series Analysis and Cointegration*, Oxford University Press, Oxford, chapter 4.
- Harris, D. (1997), "Principal Components Analysis of Cointegrated Time Series," *Econometric Theory*, 13, 529–557.
- Harris, R. (1995), *Using Cointegration Analysis in Econometric modeling*, Prentice Hall, Harvester Wheatsheaf, England.
- Hendry, D.F. (1984), "Monte Carlo Experimentation in Econometrics," in Z. Griliches and M.D. Intrilligator (eds.), *Handbook of Econometrics*, vol. II, Elsevier, Amsterdam.
- (1986), "Econometric modeling with Cointegrated Variables: An Overview," *Oxford Bulletin of Economics and Statistics*, 48, 201–239.
- Hsiao, C. (1997), "Cointegration and Dynamic Simultaneous Equations Model," *Econometrica*, 65, 647–670.
- Huang, B.N. and C.W. Yang (1996), "Long-run Purchasing Power Parity Revisited: A Monte Carlo Simulation," *Applied Economics*, 28, 967–974.
- Inder, B. (1993), "Estimating Long-Run Relationships in Economics," *Journal of Econometrics*, 57, 53–68.
- (1995), "Finite Sample Arguments for Appropriate Estimation of Cointegrating Vectors," Monash University.
- Johansen, S. (1988), "Statistical Analysis of Cointegration Vectors," *Journal of Economic Dynamics and Control*, 12, 231–254.
- (1990), "Maximum Likelihood Estimation and Inference on Cointegration—with Applications to the Demand for Money," *Oxford Bulletin of Economics and Statistics*, 52, 169–210.

- (1991), "Estimation and Hypothesis Testing of Cointegration Vectors in Gaussian Vector Autoregressive Models," *Econometrica*, 59, 1551–1580.
- (1992a), "Determination of Cointegration Rank in the Presence of a Linear Trend," *Oxford Bulletin of Economics and Statistics*, 54, 383–397.
- (1992b), "Cointegration in Partial Systems and the Efficiency of Single Equation Analysis," *Journal of Econometrics*, 50, 389–402.
- (1994), "The Role of the Constant and Linear Terms in Cointegration Analysis of Non-Stationary Variables," *Econometric Reviews*, 13, 205–229.
- (1995), *Likelihood Based Inference on Cointegration in the Vector Autoregressive Model*, Oxford university press, Oxford.
- Johansen, S. and K. Juselius (1990), "Maximum Likelihood Estimation and Inference on Cointegration—with Applications to the Demand for Money," *Oxford Bulletin of Economics and Statistics*, 52, 169–210.
- (1994), "Identification of the Long-run and Short-run Structure: an Application to the IS-LM Model," *Journal of Econometrics*, 63, 7–37.
- Joutz, F., G.S. Maddala, and R.P. Trost (1995), "An Integrated Bayesian Vector Autoregression and Error Correction Model for Forecasting Electricity Consumption and Prices," *Journal of Forecasting*, 14, 287–310.
- Kleibergen, F. (1995), "The Influence of Exogenous Variables on Long-Run Parameters: An Application to Price Level and Import Demand Models," Report TI 95-44, Tinbergen Institute, Rotterdam.
- LeSage, J.P. (1990), "A Comparison of the Forecasting Ability of ECM and VAR Models," *The Review of Economics and Statistics*, 72, 664–671.
- Li, H. and G.S. Maddala (1995), "Estimating Long-run Relationship in Economics: Some Comments and Further Results," Manuscript, Ohio State University.
- Ng, S. and P. Perron (1997), "Estimation and Inference in Nearly Unbalanced Nearly Cointegrated Systems," *Journal of Econometrics*, 79, 53–81.
- Ogaki, M. and J. Park (1997), "A Cointegration Approach to Estimating Preference Parameters," *Journal of Econometrics*, Forthcoming.

- Ohanian, L.E. (1990), "Evaluating Trends in Aggregate Time Series: A Monte Carlo Based Forecasting Approach," in T.B. Fomby and G.F. Rhodes (eds.), *Advances in Econometrics Series: Cointegration, Spurious Regression, and Unit Roots*, vol. 8, JAI Press, Greenwich, 323–342.
- Pagan, A.R. (1989), "20 Years after: Econometrics 1966–1986," in B. Cornet and H. Tulkens (eds.), *Contributions to Operations Research and Econometrics*, The MIT Press, Cambridge, MA.
- Park, J.Y. and P.C.B. Phillips (1988), "Statistical Inference in Regressions with Integrated Processes: Part I," *Econometric Theory*, 4, 468–497.
- Perron, P. and J.Y. Campbell (1993), "A Note on Johansen's Cointegration Procedure When Trends Are Present," *Empirical Economics*, 18, 777–789.
- Phillips, P.C.B. (1991), "Optimal Inference in Cointegrated Systems," *Econometrica*, 59, 283–306.
- (1994), "Some Exact Distribution Theory for Maximum Likelihood Estimators of Cointegrating Coefficients in Error Correction Models," *Econometrica*, 62, 73–93.
- Phillips, P.C.B. and B.E. Hansen (1990) "Statistical Inference in Instrumental Variables Regression with I(1) Processes," *Review of Economic Studies*, 57, 99–125.
- Phillips, P.C.B. and M. Loretan (1991), "Estimating Long-Run Economic Equilibria," *Review of Economic Studies*, 58, 407–436.
- Reinsel, G.C. and S.K. Ahn (1992), "Vector Autoregressive Models with Unit Roots and Reduced Rank Structure: Estimation, Likelihood Ratio Test, and Forecasting," *Journal of Time Series Analysis*, 13, 353–375.
- Saikkonen, P. (1991), "Asymptotically efficient estimation of cointegrating regressions," *Econometric Theory*, 7, 1–21.
- Sargan, J.D. and A. Bhargava (1983), "Testing Residuals from Least Squares Regression for Being Generated by the Gaussian Random Walk," *Econometrica*, 51, 153–174.
- Stock, J.H. (1987), "Asymptotic Properties of Least Squares Estimators of Cointegrating Vectors," *Econometrica*, 55, 1035–1056.
- Stock, J.H. and M.W. Watson (1988), "Testing for Common Trends," *Journal of the American Statistical Association*, 83, 1097–1107.
- (1993), "A Simple Estimator of Cointegrating Vectors in Higher Order Integrated Systems," *Econometrica*, 61, 783–820.

- Urbain, J.D. (1992), "On Weak Exogeneity in Error Correction Models," *Oxford Bulletin of Economics and Statistics*, 54, 187–208.
- (1993), "Exogeneity in Error Correction Models," *Lecture Notes in Economics and Mathematical Systems*, Springer Verlag, Berlin.
- Watson, M.W. (1994), "Vector Autoregressions and Cointegration," in R.F. Engle and D.L. McFadden (eds.), *Handbook of Econometrics*, vol. IV, Elsevier, Amsterdam.
- Wickens, M.R. (1996), "Interpreting Cointegrating Vectors and Common Stochastic Trends," *Journal of Econometrics*, 74, 255–271.
- Wickens, M.R. and T.S. Breusch (1988), "Dynamic Specification, the Long-Run and the Estimation of Transformed Regression Models," *Economic Journal*, 98, 189–205.
- Xiao, Z. and P.C.B. Phillips (1996), "Efficient De-trending in Cointegrating Regression," Manuscript, Cowles foundation, Yale University.

6

Tests for cointegration

6.1 Introduction

In the previous chapter we discussed methods of estimation in cointegration systems. In this chapter we shall discuss tests for cointegration. Corresponding to the single equation and multiple equation methods of estimation considered in the previous chapter, we have tests for cointegration in single equation and system frameworks. Also, corresponding to the tests for unit roots discussed in chapter 4 where we considered the case of unit root as null and stationarity (no unit root) as null, we have tests for cointegration with no cointegration as null and cointegration as null.

There is one additional factor to be considered with regard to tests for cointegration in systems of multiple equations. Here we are interested in finding how many cointegrating relationships there are among the variables. In other words, we are interested in testing hypotheses about the rank of the cointegration space. In the following sections we shall start with tests in a single equation framework and then consider system methods.

6.2 Single equation methods: residual-based tests

The residual-based tests were the earlier tests for cointegration and were discussed in Engle and Granger (1987). Consider the set of $(k + 1)$ variables y_t which are $I(1)$. If there exists a vector θ such that $\theta'y_t$ is $I(0)$, then θ is the cointegrating vector. Since θ is determined only up to a multiplicative constant, we shall normalize the first variable in y_t to have coefficient 1. Thus, if we write θ as $(1, -\beta')'$ and partition y_t conformably into (y_{1t}, y_{2t}) , then $y_{1t} - \beta'y_{2t}$ is the cointegrating relationship. The

residual-based tests consider the equation

$$y_{1t} = \beta' y_{2t} + u_t \quad (6.1)$$

If u_t has a unit root, then $y_{1t} - \beta' y_{2t}$ is not a cointegrating relationship. Thus, a test for a unit root in u_t is a test that the variables y_t are not cointegrated. In practice β is not known and (6.1) is estimated, say by OLS, and unit root tests are applied to the estimated residual $\hat{u}_t$.

It is common to apply the ADF test and Phillips–Perron Z_α and Z_t tests to $\hat{u}_t$. In practice any of the tests discussed in chapters 3 and 4 can be applied. However, the critical values are not the same because we are applying the tests to $\hat{u}_t$, not u_t . The critical values will depend on

- (i) the number of regressors in y_{2t} in equation (6.1) and
- (ii) whether a constant and/or a time trend is included in (6.1).

6.2.1 Critical values for residual-based tests

Engle and Yoo (1987) and Phillips and Ouliaris (1990) have tabulated critical values for the ADF t -statistic. (The critical values for the Z_t are the same as the ADF t -statistic.) Phillips and Ouliaris also tabulated critical values for Z_α . Since the asymptotic distributions differ according to different trend variables in cointegrating regression, the critical values consist of three parts corresponding to the different trend variables in the cointegrating regressions as follows:

- (a) $y_{1t} = \beta' y_{2t} + u_t$
- (b) $y_{1t} = \alpha + \beta' y_{2t} + u_t$
- (c) $y_{1t} = \alpha + \delta t + \beta' y_{2t} + u_t$

Tables 6.1 and 6.2 report part of tables I and II in Phillips and Ouliaris (1990), omitting minus signs for simplicity.

Critical values in tables 6.1 and 6.2 are based on asymptotic distributions. For small sample sizes (and for all the sample sizes) MacKinnon (1991) provides an approximation formula for computing critical values for all sample sizes T . MacKinnon estimated response surface regressions which approximate the critical values remarkably well. He found a functional form for the response surface regressions after considerable experiments as follows

$$C_k(p, T_k) = \beta_\infty + \beta_1 T_k^{-1} + \beta_2 T_k^{-2} + e_k$$

where $C_k(p, T_k)$ is the critical values for a test at the p percent level with

Table 6.1. *Critical values for the ADF t -statistic and Z_t*

N	Regression (a)			Regression (b)			Regression (c)		
	1%	5%	10%	1%	5%	10%	1%	5%	10%
1	3.39	2.76	2.45	3.96	3.37	3.07	4.36	3.8	3.52
2	3.84	3.27	2.99	4.31	3.77	3.45	4.65	4.16	3.84
3	4.30	3.74	3.44	4.73	4.11	3.83	5.04	4.49	4.20
4	4.67	4.13	3.81	5.07	4.45	4.16	5.36	4.74	4.46
5	4.99	4.40	4.14	5.28	4.71	4.43	5.58	5.03	4.73

Source: Phillips and Ouliaris (1990, p. 190).

Table 6.2. *Critical values for the Z_α*

N	Regression (a)			Regression (b)			Regression (c)		
	1%	5%	10%	1%	5%	10%	1%	5%	10%
1	22.83	15.64	12.54	28.32	20.49	17.04	35.42	27.09	23.19
2	29.27	21.48	18.18	34.17	26.09	22.19	40.34	32.22	27.78
3	36.16	27.85	23.92	41.13	32.06	27.58	47.36	37.73	33.16
4	42.87	33.48	28.85	47.51	37.15	32.74	53.61	42.46	37.74
5	48.52	38.09	33.80	52.17	41.94	37.01	58.16	47.38	42.32

Source: Phillips and Ouliaris (1990, pp. 189–190).

the sample size T_k and β_s are parameters to be estimated. The response surface regressions were estimated by feasible GLS. Table 6.3 shows the GLS estimates of β_s for three cases (a), (b), and (c). The estimates of β_∞ provide asymptotic critical values directly, while values for any finite T can be computed by using the estimates of all three parameters. Note that k denotes the number of variables included in a cointegrating regression, while N in tables 6.1 and 6.2 denotes the number of regressors included in a cointegrating regression. Thus, $k = N + 1$. The estimates for $k = 1$ provide the critical values of the DF tests for a unit root. For example, at $p = 5$ percent level the critical values can be approximated by

$$C(T) = -3.4126 - 4.039T^{-1} - 17.83T^{-2}$$

If $T \rightarrow \infty$, $C = -3.4126$. If $T = 100$, $C = -3.4547$. Similarly $C = -3.5005$ when $T = 50$. These approximated critical values are remarkably close to the corresponding critical values in table 3.1 of chapter 3

Table 6.3. *Response surface estimates of critical values*

k	Regression	Size	β_∞	β_1	β_2
1	(a)	1%	-2.5658	-1.960	-10.04
		5%	-1.9393	-0.398	0.0
		10%	-1.6156	-0.181	0.0
	(b)	1%	-3.4336	-5.999	-29.25
		5%	-2.8621	-2.738	-8.36
		10%	-2.5671	-1.438	-4.48
	(c)	1%	-3.9638	-8.353	-47.44
		5%	-3.4126	-4.039	-17.83
		10%	-3.1279	-2.418	-7.58
2	(b)	1%	-3.9001	-10.534	-30.03
		5%	-3.3377	-5.967	-8.98
		10%	-3.0462	-4.069	-5.73
	(c)	1%	-4.3266	-15.531	-34.03
		5%	-3.7809	-9.421	-15.06
		10%	-3.4959	-7.203	-4.01
3	(b)	1%	-4.2981	-13.790	-46.37
		5%	-3.7429	-8.352	-13.41
		10%	-3.4518	-6.241	-2.79
	(c)	1%	-4.6676	-18.492	-49.35
		5%	-4.1193	-12.024	-13.13
		10%	-3.8344	-9.188	-4.85
4	(b)	1%	-4.6493	-17.188	-59.20
		5%	-4.1000	-10.745	-21.57
		10%	-3.8110	-8.317	-5.19
	(c)	1%	-4.9695	-22.504	-50.22
		5%	-4.4294	-14.501	-19.54
		10%	-4.1474	-11.165	-9.88
5	(b)	1%	-4.9587	-22.140	-37.29
		5%	-4.4185	-13.641	-21.16
		10%	-4.1327	-10.638	-5.48
	(c)	1%	-5.2497	-26.606	-49.56
		5%	-4.7154	-17.432	-16.50
		10%	-4.4345	-13.654	-5.77
6	(b)	1%	-5.2400	-26.278	-41.65
		5%	-4.7048	-17.120	-11.17
		10%	-4.4242	-13.347	0.0
	(c)	1%	-5.5127	-30.735	-52.50
		5%	-4.9767	-20.883	-9.05
		10%	-4.6999	-16.445	0.0

Source: MacKinnon (1991).

(which is computed by a Monte Carlo method), -3.41 , -3.45 , and -3.50 , respectively.

The critical values of the ADF tests for cointegration can also be computed by the response surface estimates. For example, the regression (c) with $k = 2$ and $T = 500$, the critical value at the 5 percent level is approximated by $C = -3.799$ which is very close to the corresponding critical value of table 6.1 ($N=1$ at the 5 percent level), -3.8 .

By using these critical values, one can reject the null hypothesis of no cointegration if the computed value of the statistic is smaller than the appropriate critical value. For example, for a regression with a constant term and two explanatory variables ($N = 2$), we reject the null of no cointegration at the 5 percent level if the computed value of ADF t -statistic and $\hat{Z}_t$ is less than -3.77 or the computed value of $\hat{Z}_\alpha$ is less than -26.09 .

MacKinnon (1994) expands the response surface methodology to calculate both asymptotic and finite sample critical values and p -values for some commonly used unit root and cointegration tests. (The methodology can be extended to Johansen's tests and ECM tests discussed in later sections.) The programs are available on internet. MacKinnon says that

interested readers should ftp to qed.econ.queensu.ca and then go to the directory **pub/uroot**. The directory contains the source codes for **urcdist**, a compiled version for DOS-based PCs with at least 4MB memory and a numeric coprocessor and 12 zipped files containing the estimated response surface coefficients. The **urcdist** program is run interactively and prompts the user for input.

6.2.2 Other residual-based tests

Besides these three tests, various residual-based tests have been proposed. Engle and Granger (1987) suggested to use the Durbin–Watson statistic from the cointegrating regression. However, Campbell and Perron (1991) argued that it should not be used as the basis of a test of the null hypothesis of no cointegration versus the alternative hypothesis of cointegration (Campbell and Perron, 1991, rule 19, p. 177). The reason is that under the null hypothesis of no cointegration, the asymptotic distribution of the Durbin–Watson statistic depends on nuisance parameters such as the correlations among the first-differences of the variables included in the regression.

Hansen (1990) suggested working with the estimated residuals from a

Cochrane–Orcutt (C–O) version of the static regression that allows for AR(1) errors

$$y_{1t} = \beta' y_{2t} + u_t, \quad u_t = \rho u_{t-1} + e_t$$

He showed that under the null hypothesis of no cointegration, the C–O estimate of the cointegrating vector converges to a constant, not to a random variable as does the OLS estimator, and also has the usual Dickey–Fuller distribution. Hansen reported some Monte Carlo comparisons which indicate that C–O tests are uniformly more powerful than Stock–Watson, Johansen, and Phillips–Ouliaris tests. He also shows that the C–O tests can be extended to allow for more general serial correlation in the residual u_t .

6.3 Single equation methods: ECM tests

The residual-based tests discussed in the previous section have been found to have low power, although they are very often used. See Kremers *et al.* (1992), Zivot (1994), and Banerjee (1995). The problem is with the estimation of the static regression (6.1) in the first step. Kremers *et al.* argue that the low power of the residual-based tests is due to ignoring equation dynamics in (6.1) and concentrating on error dynamics (which they call imposing a common factor restriction). This problem is now well known in the estimation of dynamic models. This problem in its most elementary version is as follows: Consider the two models:

$$\text{Model 1 : } y_t = \beta x_t + u_t, \quad u_t = \rho u_{t-1} + e_t$$

$$\text{Model 2 : } y_t = \alpha_1 y_{t-1} + \alpha_2 x_t + \alpha_3 x_{t-1} + e_t$$

Model 1 has no equation dynamics but has error dynamics. Model 2 is one with equation dynamics but no error dynamics. But model 1 is the same as model 2 with the restriction $\alpha_1 \alpha_2 + \alpha_3 = 0$. If this restriction (known as common factor restriction) is violated, then model 1 is invalid and the observed error dynamics is spurious – it is a consequence of ignoring equation dynamics. This argument also carries over to cointegrating systems.

To avoid this problem Kremers *et al.* suggest using ECM tests instead of the residual-based tests. The ECM tests are based on the error correction model implied by the cointegrated system. The Granger representation theorem says that there is an error correction representation for every cointegration relationship. For simplicity, let us assume that

y_{2t} in equation (6.1) is of dimension 1 and hence β is a scalar. The ECM representation is then

$$\begin{aligned}\Delta y_{1t} &= \alpha_1(y_{1,t-1} - \beta y_{2,t-1}) + \text{lagged}(\Delta y_{1t}, \Delta y_{2t}) + \varepsilon_{1t} \\ \Delta y_{2t} &= \alpha_2(y_{1,t-1} - \beta y_{2,t-1}) + \text{lagged}(\Delta y_{1t}, \Delta y_{2t}) + \varepsilon_{2t}\end{aligned}\quad (6.2)$$

with α_1 and/or α_2 nonzero. If $\alpha_1 = \alpha_2 = 0$, then we have no cointegration. One can apply tests for cointegration without estimating β , as follows: write (6.2) as

$$\begin{aligned}\Delta y_{1t} &= \alpha_1(y_{1,t-1} + y_{2,t-1}) + \text{lagged}(\Delta y_{1t}, \Delta y_{2t}) + \gamma_1 y_{2,t-1} + \varepsilon_{1t} \\ \Delta y_{2t} &= \alpha_2(y_{1,t-1} + y_{2,t-1}) + \text{lagged}(\Delta y_{1t}, \Delta y_{2t}) + \gamma_2 y_{2,t-1} + \varepsilon_{2t}\end{aligned}\quad (6.3)$$

where $\gamma_1 = -\alpha_1(\beta + 1)$ and $\gamma_2 = -\alpha_2(\beta + 1)$.

Kremers *et al.* consider a conditional ECM for y_{1t} and a marginal unit root process for y_{2t} . This is also the system adopted in Banerjee *et al.* (1996). In our notation the DGP they consider is

$$\begin{aligned}\Delta y_{1t} &= \gamma' \Delta y_{2t} + \alpha_1(y_{1,t-1} - \beta' y_{2,t-1}) + u_{1t} \\ \Delta y_{2t} &= u_{2t}\end{aligned}\quad (6.4)$$

where

$$\begin{pmatrix} u_{1t} \\ u_{2t} \end{pmatrix} \sim iid \left[\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} \sigma_1^2 & 0 \\ 0 & \Sigma_{22} \end{pmatrix} \right]$$

The F -test for $\alpha_1 = 0$ is a test for the null of no cointegration. Kremers *et al.* actually consider a case where the cointegrating vector β is known. Their purpose, however, is to demonstrate how the residual-based tests are based on the common factor restriction and ignore equation dynamics, and how the ECM-based tests are more powerful than the residual-based tests if the common factor restriction does not hold.

The paper by Kremers *et al.* is based on the restrictive assumption of a known cointegrating vector. This is relaxed in the paper by Banerjee *et al.* (1996). They write equation (6.4) as

$$\Delta y_{1t} = \gamma' \Delta y_{2t} + \alpha_1 y_{1,t-1} + \theta' y_{2,t-1} + u_{1t} \quad (6.5)$$

where $\theta = -\alpha_1 \beta$. The noncointegrating restriction $\alpha_1 = 0$ implies $\theta = 0$. They suggest estimating equation (6.5) by OLS and testing that the coefficient of $y_{1,t-1}$ is equal to zero. The ECM test is thus based on $\hat{\alpha}_1$ or its t -ratio. Banerjee *et al.* derive the asymptotic distributions of $T\hat{\alpha}_1$ and the t -statistic based on $\hat{\alpha}_1$, and suggest using the t -test. They also tabulate critical values for models with a constant and with a time

trend. However, note that $\alpha_1 = 0$ also implies $\theta = 0$ in (6.5) and this is not taken into account in the test by Banerjee *et al.* (1996).

The ECM-based test derived in Banerjee *et al.* is based on the assumption that y_{2t} are strictly exogenous (note the specification of the covariance matrix for u_{1t} and u_{2t}). Other tests based on exogeneity assumptions are discussed in Boswijk (1992) and Boswijk and Franses (1992). In spite of the fact that the residual-based tests have low power, the ECM tests have not been found to be used in many applications. This is because of the availability of system based tests (which are discussed in section 6.5 later) and also the nonavailability of ready tables of critical values for use with the ECM tests.

Zivot (1994) considers a more general framework that relaxes the restrictive assumptions in Kremers *et al.* and Banerjee *et al.* For an unknown cointegrating vector he suggests a two-step ECM test, which is implemented as follows: using the static regression (6.1) get an estimator $\hat{\beta}$ for β . This estimator, as is well-known, is superconsistent. Next substitute $\hat{\beta}$ for β in equation (6.4), and apply the ECM test by testing the significance of $\hat{\alpha}_1$ using a standard t -test.

Zivot considers another ECM test which is a one-step test and is similar to the ECM test considered in Boswijk (1992, 1994). Write (6.4) as

$$\Delta y_{1t} = \gamma' \Delta y_{2t} + \Pi' y_{t-1} + u_{1t}$$

where $\Pi' = \gamma(1, -\beta')$ and $y_t' = (y_{1t}, y_{2t})$. The no cointegration restriction $\gamma = 0$ implies $\Pi = 0$. Now estimate this equation by OLS and test the significance of $\hat{\Pi}$ using a standard Wald test. Zivot performs Monte Carlo experiments comparing the power of these ECM tests with the power of the Engle–Granger residual-based ADF t -test, and finds them more powerful.

6.4 Tests with cointegration as null

The tests considered in the previous sections are for the null hypothesis of no cointegration. These are based on tests for a unit root hypothesis in the residuals of the cointegrating regression. In chapter 4 we discussed unit root tests with stationarity as the null hypothesis. Correspondingly there are tests with cointegration as the null, although they have not been widely used, nor does anyone talk of using them for confirmatory analysis (as we discussed in chapter 4). We shall consider four tests here: (i) the Leybourne and McCabe test which is based on an unobserved

components model, (ii) the Park and Choi test which is based on testing the significance of superfluous regressors, (iii) the Shin test which is a residual-based test, and (iv) Harris and Inder test.

We shall review these tests here. There is as yet no study comparing the power of these tests. So no practical conclusions can be offered as to their relative merits. Nor is there any study investigating the power of these tests. So we do not even know whether their power properties are poor as for the unit root tests discussed in chapter 4. Harris and Inder (1994) compare the power of their test with that of the DF but note that the power comparison is between that for two tests which have their null and alternative reversed (and thus invalid). They do not compare the power of their test with that of the other tests for cointegration as null, which is the valid comparison. Nor has anyone made a study (as done by Burke quoted in chapter 4) on using these tests in conjunction with tests for no cointegration, for confirmatory analysis.

6.4.1 Leybourne and McCabe (1993)

Leybourne and McCabe (1993) proposed the test for the null hypothesis of cointegration against the alternative hypothesis of no cointegration. They consider the cointegrating regression

$$y_{1t} = \beta' y_{2t} + u_t$$

where the error u_t can be written

$$\begin{aligned} u_t &= a_t + \varepsilon_t, & \varepsilon_t &\sim IN(0, \sigma^2) \\ a_t &= a_{t-1} + \eta_t, & \eta_t &\sim IN(0, \sigma_\eta^2) \end{aligned}$$

Then the null hypothesis of cointegration is

$$H_0 : \sigma_\eta^2 = 0$$

and the alternative hypothesis of no cointegration is

$$H_1 : \sigma_\eta^2 > 0$$

Leybourne and McCabe suggest the (conditionally) locally best invariant (LBI) test of H_0 against H_1 based on the statistic

$$T^{-2} \hat{\sigma}_\varepsilon^2 \hat{\varepsilon}' V \hat{\varepsilon}$$

where $\hat{\varepsilon}$ is the vector of OLS residuals from the cointegrating regression under H_0 , V is a $T \times T$ matrix with its ij th element equal to the

minimum of i and $j(i, j = 1, \dots, T)$. For the estimate of the variance, $\hat{\sigma}_\varepsilon^2$, they proposed to use the Newey–West type estimator considered in Phillips and Perron (1988) (see chapter 3). They derived the asymptotic distribution of the LBI test statistics, which is a function of Wiener processes and depends only on the number of variables in y_{2t} and σ_ε^2 .

Leybourne and McCabe performed a small Monte Carlo experiment which showed that the test has approximately the correct sizes under H_0 for both samples of $T=100$ and 200 . However, comparisons of the relative robustness or power of the LBI test with other tests for the null of no cointegration, are not attempted because of the differing natures of the null and alternative hypotheses.

6.4.2 Tests using superfluous regressors

Park and Choi (1988) and Park (1990) propose tests for cointegration which have the advantage that they can be formulated either with a null of no cointegration or with a null of cointegration. Their basic idea is to introduce *superfluous* regressors into the cointegrating regression and test for the coefficients of the added superfluous regressors.

Consider the following version of the cointegrating regression with added regressors

$$y_{1t} = \beta' y_{2t} + \gamma_1' s_{1t} + \gamma_2' s_{2t} + e_t \quad (6.6)$$

The added variable s_{1t} is a vector of q nonstationary deterministic functions that are of a higher order than the variables in the cointegrating regression. For example, if 1 and t are included in the cointegrating regression, s_{1t} could include the regressors $\{t^2, t^3, \dots, t^{q+1}\}$. The vector s_{2t} is a p -element vector containing variables that are integrated of order one.

Testing procedures are based on the comparison between the residual sum of squares (RSS) from the original cointegrating regression and the RSS from the cointegrating regression with superfluous regressors s_{1t} and s_{2t} . Park (1990) proposed the J_1 test statistic for the null of cointegration

$$J_1 = \frac{RSS_1 - RSS_2}{w}$$

where RSS_1 is the RSS from the regression after nonparametric transformation for independence of nuisance parameters and RSS_2 is the RSS from the transformed regression with superfluous regressors. The

denominator is the normalizing variance corresponding to the test statistics. Park shows that under the null hypothesis of cointegration, the J_1 statistic has limiting χ^2 distribution with the degree of freedom equal to the number of the superfluous regressors. But under the hypothesis of no cointegration, it diverges to infinity. Thus when J_1 is large, the test rejects the null hypothesis of cointegration. Asymptotic critical values can be found in his paper.

For a choice of the superfluous regressors, s_t , it is required that the stochastic components of the regressors $\{s_t\}$ and $\{y_{2t}\}$ be not cointegrated. Park suggests that the most irrelevant time trend would be the best choice. Likewise, any $I(1)$ process which is completely independent of the processes in the model, as for instance, a computer-generated pseudo random walk would serve well. Park also illustrated the case where the lagged dependent variable is added as a regressor.

6.4.3 Residual-based test

Shin (1994) extends the KPSS test which tests the null of stationarity against the alternative of unit root (see chapter 4) to a multivariate context of testing the null of cointegration against the alternative of no cointegration. He extensively uses the parametric correction procedure of estimation of cointegrating regression (see section 5.4.2 of chapter 5 for details). When the elements of the error vector in the cointegrating regression are contemporaneously correlated, which causes the endogeneity problem in single equation estimation methods, as we have seen in section 5.4.2, the effect of this correlations can be eliminated by adding Δy_{2t} to the regressor set. The parametrically corrected cointegrating regression is

$$y_{1t} = \beta' y_{2t} + \sum_{j=-k_1}^{k_2} b_j \Delta y_{2,t-j} + e_t \quad (6.7)$$

(For reader's convenience, this equation is duplicated from the equation (5.4) in section 5.4.2.)

Based on the residual $\hat{e}_t$ from this equation, Shin proposed the following test statistic for the null of cointegration

$$C = T^{-2} \sum S_t^2 / \hat{\sigma}_e^2$$

where $S_t = \sum_{i=1}^t \hat{e}_i^2$ and $\hat{\sigma}_e^2$ are semiparametric consistent estimators of the long-run variance of e_t (see chapter 3). The asymptotic distribution

of the test statistic C is shown to be

$$C \Rightarrow \int_0^1 Q^2$$

where

$$Q = W_1 - \left(\int_0^r W_2' \right) \left(\int_0^1 W_2 W_2' \right)^{-1} \left(\int_0^1 W_2 dW_1 \right)$$

The standard Brownian motions W_1 and W_2 are independent and correspond to scalar variable y_{1t} and m -vector variable y_{2t} , respectively. The limiting distributions are affected by the inclusion of deterministic components in the cointegrating equation. The asymptotics carry in a similar fashion with the only modification being that the standard Brownian motions are substituted by demeaned and detrended Brownian motions (see chapter 3 and also Park and Phillips, 1988). The asymptotic distributions depend on m , the dimension of the cointegrated system. Critical values are given for $m=1$ to 5 in Shin (1994, pp. 100–101).

6.4.4 Harris and Inder test

Harris and Inder (1994) also suggest a test for the null of cointegration. They use nonparametric correction procedure for estimation of cointegrating regression, while Shin uses the parametric correction procedure. But both the tests are the same in the sense that they use the same test statistics, the LM test statistic. Their procedure is as follows: consider the problem whether the $I(1)$ variables y_t and x_t are cointegrated. First, apply OLS to the regression

$$y_t = \mu_0 + \beta_0' x_t + u_t$$

and obtain residual $\hat{u}_t$. Form $\hat{z}_t = [\hat{u}_t, \Delta x_t']'$ and calculate the estimated covariance matrices

$$\begin{aligned} \hat{\Omega} &= \begin{bmatrix} \hat{\omega}_{11} & \hat{\omega}_{12} \\ \hat{\omega}_{21} & \hat{\Omega}_{22} \end{bmatrix} \\ &= \frac{1}{T} \left[\sum_{t=1}^T \hat{z}_t \hat{z}_t' + \sum_{k=1}^l w(k, l) \sum_{t=k+1}^T (\hat{z}_{t-k} \hat{z}_t' + \hat{z}_t \hat{z}_{t-k}') \right] \end{aligned}$$

and

$$\hat{\Delta} = \begin{bmatrix} \hat{\delta}_{11} & \hat{\delta}_{12} \\ \hat{\delta}_{21} & \hat{\Delta}_{22} \end{bmatrix} = \frac{1}{T} \sum_{k=0}^l \sum_{t=k+1}^T \hat{z}_t \hat{z}_{t-k}'$$

Table 6.4. *Critical values for the Harris and Inder test*

N	10%	5%
1	0.2335	0.3202
2	0.1617	0.2177
3	0.1203	0.1590
4	0.0929	0.1204
5	0.0764	0.0972

Source: Harris and Inder (1994).

where $w(k, l) = 1 - k/(l + 1)$. Then calculate

$$y_t^+ = y_t - \hat{\omega}_{12} \hat{\Omega}_{22}^{-1} \Delta x_t$$

and

$$\delta_t^+ = \hat{\delta}_{21} - \hat{\Delta}_{22} \hat{\Omega}_{22}^{-1} \hat{\omega}_{21}$$

and re-estimate the cointegrating regression using the fully modified estimator

$$\hat{\beta}^+ = (X'X)^{-1}(X'y^+ - e_k T \hat{\delta}_{21}^+)$$

where $e_k = [0, I_k]'$. Obtain the residuals $\hat{u}_t^+$ from this regression and construct the test statistic

$$S_2^+ = \frac{T^{-2} \sum_{t=1}^T (K_t^+)^2}{\hat{\omega}_{1.2}^2}$$

where

$$K_t^+ = \sum_{i=1}^t \hat{u}_i^+$$

and

$$\hat{\omega}_{1.2}^2 = \hat{\omega}_{11} - \hat{\omega}_{12} \hat{\Omega}_{22}^{-1} \hat{\omega}_{21}$$

The critical values for this test, which depend only on the number of regressors n (excluding the constant) provided by Harris and Inder are given in table 6.4.

In practice it might be desirable to obtain the critical values by Monte Carlo methods (since the critical values of table 6.4 are based on the asymptotic distribution).

6.5 Multiple equation methods

The results from single equation methods depend on what variable is used for normalization of the cointegration relationship (see section 5.5.3). Also the single equation methods do not enable us to test how many cointegration relationships there are. For this reason it is common practice to use system methods. The most popular of these system methods is the Johansen method which is based on canonical correlation methods. (We shall discuss limitations of this method in the next section.) The modification of the Box–Tiao procedure by Bewley and Orden discussed in chapter 5 (section 5.5.3) is also based on canonical correlations. We shall discuss the relative merits of the two procedures in the next section, after discussing two methods based on principal components and other methods.

6.5.1 Methods based on canonical correlations: Johansen Tests

The Johansen procedure has been discussed in chapter 5. The procedure leads to two test statistics for cointegration. The first, called the *trace test*, tests the hypothesis that there are at most r cointegrating vectors. The second called the *maximum eigenvalue test*, tests the hypothesis that there are $r + 1$ cointegrating vectors versus the hypothesis that there are r cointegrating vectors. Johansen and Juselius (1990) suggest that the maximum eigenvalue test may be better.

Trace test

Result (5.7) in section 5.5.3 shows that the maximum of the likelihood is given by

$$-2 \log L_{\max} \propto T \sum_{i=1}^n \ln(1 - \lambda_i)$$

where λ_i are the roots of the determinant equation (5.6). The LR test statistic for the hypothesis of at most r cointegrating vectors is

$$\lambda_{\text{trace}} = -T \sum_{i=r+1}^n \ln(1 - \hat{\lambda}_i)$$

where $\hat{\lambda}_{r+1}, \dots, \hat{\lambda}_n$ are the $(n-r)$ smallest eigenvalues of the determinant equation (5.6) presented in chapter 5. The asymptotic distribution of

this statistic is given by the trace of the stochastic matrix

$$\int_0^1 (dW)W' \left(\int_0^1 WW' dr \right)^{-1} \int_0^1 W(dW)' \quad (6.8)$$

where W is an $(n - r)$ dimensional Brownian motion. In case there are a constant and/or a trend term in the VAR model we start with, (6.8) is changed to

$$\int_0^1 (dW)\tilde{W}' \left(\int_0^1 \tilde{W}\tilde{W}' dr \right)^{-1} \int_0^1 \tilde{W}(dW)' \quad (6.9)$$

where $\tilde{W}$ is the demeaned or detrended Brownian motion.

Maximum eigenvalue test

To test the null hypothesis of r cointegrating vectors versus the alternative of $(r + 1)$ cointegrating vectors the LR test statistic is

$$\lambda_{max} = -T \ln(1 - \hat{\lambda}_{r+1})$$

The asymptotic distributions of this statistic are given by the maximum eigenvalue of the stochastic matrix (6.8) and (6.9) according to the different specifications of deterministic trends in the VAR model.

Detailed tables of critical values for these tests are provided in Osterwald-Lenum (1992) who tabulates them for systems of order $n \leq 11$, and number of cointegrating vectors 0 to n (0 implying all linear combinations are $I(1)$ and n implying that all the variables are stationary). He also tabulates them for different cases of constant and/or trend terms included. The paper gives the 50, 80, 90, 95, 97.5, and 99 percent quantiles, as well as the mean and variance of the test statistics.

As discussed in chapter 5 (section 5.5.2), Johansen (1992) and Perron and Campbell (1993) extended the Johansen tests to include trends (and thus cover the case of stochastic cointegration). The tests considered earlier are tests for deterministic cointegration (see section 5.5.2 for a discussion of deterministic and stochastic cointegration).

Table 6.5 provides part of critical values of the Johansen's LR tests in Osterwald-Lenum (1992). As we have seen in equations (6.8) and (6.9) the asymptotic distributions of test statistics are different according to the different specifications of deterministic trends in the VAR model. Moreover, when we construct the ECM from the VAR model, the deterministic terms in the ECM may differ from those in the VAR model. When there are deterministic cointegration relationships among variables, deterministic trend terms in the VAR model will not be present

Table 6.5. *Quantiles of the asymptotic distribution of the Johansen's LR test statistics*

$n - r$	Case I			Case II			Case III		
	90%	95%	99%	90%	95%	99%	90%	95%	99%
Maximum eigenvalue test									
1	2.86	3.84	6.51	2.69	3.76	6.65	7.52	9.24	12.97
2	9.52	11.44	15.69	12.07	14.07	18.63	13.75	15.67	20.20
3	15.59	17.89	22.99	18.60	20.97	25.52	19.77	22.00	26.81
4	21.58	23.80	28.82	24.73	27.07	32.24	25.56	28.14	33.24
5	27.62	30.04	35.17	30.90	33.46	38.77	31.66	34.40	39.79
6	33.62	36.36	41.00	36.76	39.37	45.10	37.45	40.30	46.82
7	38.98	41.51	47.15	42.32	45.28	51.57	43.25	46.45	51.91
8	44.99	47.99	53.90	48.33	51.42	57.69	48.91	52.00	57.95
9	50.65	53.69	59.78	53.98	57.12	62.80	54.35	57.42	63.71
10	56.09	59.06	65.21	59.62	62.81	69.09	60.25	63.57	69.94
11	61.96	65.30	72.36	65.38	68.83	75.95	66.02	69.74	76.63
Trace test									
1	2.86	3.84	6.51	2.69	3.76	6.65	7.52	9.24	12.97
2	10.47	12.53	16.31	13.33	15.41	20.04	17.85	19.96	24.60
3	21.63	24.31	29.75	26.79	29.68	35.65	32.00	34.91	41.07
4	36.58	39.89	45.58	43.95	47.21	54.46	49.65	53.12	60.16
5	55.44	59.46	66.52	64.84	68.52	76.07	71.86	76.07	84.45
6	78.36	82.49	90.45	89.48	94.15	103.18	97.18	102.14	111.01
7	104.77	109.99	119.80	118.50	124.24	133.57	126.58	131.70	143.09
8	135.24	141.20	152.32	150.53	156.00	168.36	159.48	165.58	177.20
9	169.45	175.77	187.31	186.39	192.89	204.95	196.37	202.92	215.74
10	206.05	212.67	226.40	225.85	233.13	247.18	236.54	244.15	257.68
11	248.45	255.27	269.81	269.96	277.71	293.44	282.45	291.40	307.64

Source: Osterwald-Lenum (1992).

in the ECM. On the other hand, if there are stochastic cointegration relationships, deterministic trend terms appear in the ECM. But in this case there are two possibilities. One is the case that deterministic trend enters only via the error correction term (so there is no separate drift term in the ECM). The other is the case that a drift term enters as an independent term in the ECM. Thus the critical values in table 6.5 should be used according to the following cases (in the following, D_t is a column vector of zero-mean seasonal dummies and $\varepsilon_t \sim N(0, \Lambda)$):

Case I: There is no drift in the ECM model.

$$\Delta X_t = \sum_{i=1}^{k-1} \Gamma_i \Delta X_{t-i} + \alpha \beta' X_{t-1} + \Psi D_t + \varepsilon_t$$

This is the case of the deterministic cointegration.

Case II: There is a separate drift in the the ECM model.

$$\Delta X_t = \sum_{i=1}^{k-1} \Gamma_i \Delta X_{t-i} + \alpha \beta' X_{t-1} + \mu + \Psi D_t + \varepsilon_t$$

This is the case of the stochastic cointegration. A (separate) drift term in the ECM implies that the first-differenced variables in the ECM do not have the same mean. And thus the level variables may have different growth patterns (though they have common stochastic growth pattern).

Case III: There is no (separate) drift in the ECM, but a constant enters only via the error correction term.

$$\Delta X_t = \sum_{i=1}^{k-1} \Gamma_i \Delta X_{t-i} + \alpha(\beta', \beta_0)(X'_{t-1}, 1)' + \Psi D_t + \varepsilon_t$$

This model implies that the first-differenced variables in the ECM have a common mean. For example, models of the term structure of interest rates are expected to have such a property.

Besides these three cases, critical values for the ECM with a trend are also given by Osterwald-Lenum.

The question of what deterministic variables should be included in the ECM is closely related to the null hypothesis of unit root tests. As we have seen in chapter 3, when the null hypothesis of unit root tests (or DGP) is a random walk with t^k trend term, one should include at least t^{k+1} trend term in the estimating regression. Thus when one considers the ECM with the presence of drift and/or trend, one should check whether the appropriate specification of the deterministic trends has been used in the regression for unit root tests.

6.5.2 Finite sample evidence on the Johansen tests

There have been several Monte Carlo studies on the finite sample evidence on the Johansen tests. We shall review these here and then present a summary assessment of these studies.

Podivinsky (1990)

Podivinsky (1990) investigates the empirical sizes of Johansen's test by Monte Carlo simulation. He finds that the tabulated critical values based on the asymptotic distribution may be inappropriate when applied to sample sizes of 100 or smaller. The second problem investigated is that

of underspecifying or overspecifying the number of variables in the system. The author investigates the powers of the Johansen test, the ADF test, and the CRDW (cointegrating regression Durbin–Watson) test in these cases. The main conclusion is that, all the tests can be misleading if too few potential variables are included in the analysis (see Maddala, 1992, p. 601, for a discussion of the consequences of omitted variables in tests for cointegration). On the other hand, overspecification of the number of variables in the system has less serious consequences, particularly when combined with Johansen’s procedure for testing linear restrictions on the elements of a cointegrating vector.

Boswijk and Franses (1992)

Boswijk and Franses (1992) investigate the effects of dynamic specification on the size and power of three cointegration tests: (i) the residual-based ADF test proposed by Engle and Granger, (ii) the ECM-based Wald test by Boswijk, and (iii) Johansen’s test. They find:

- (i) An insufficient lag length p can lead to substantial size distortions of the null hypothesis. On the other hand, overparameterization results in a loss of power.
- (ii) The overrejection of the null of no cointegration suggests the possibility of spurious cointegration. In a way the idea of cointegration arose as a means of safeguarding against spurious regressions. This Monte Carlo study suggests that this attempt is only partly successful.
- (iii) The Wald test is more powerful than ADF test. It is also more powerful than Johansen’s test in this case because the conditioning variable is weakly exogenous. Johansen’s LR test does not exploit this feature.

Eitrheim (1992)

This is an extensive Monte Carlo study on Johansen’s trace and maximum eigenvalue tests and the ML method of estimating the cointegrating vectors. The author also discusses the effects of several model misspecifications: (i) wrong order in the VAR model, (ii) ignoring nonnormality (e.g., if the errors are of the ARCH type or are serially correlated), (iii) effect of temporal aggregation or systematic sampling on inference about cointegration, (iv) measurement errors: the nonstationary latent variables are measured with errors. Several aspects of the test statis-

tics and the estimators of the short-run and long-run parameters were studied but some of the main conclusions are:

- (i) The trace test and maximum eigenvalue test have, as expected, low power against near cointegration alternatives.
- (ii) The Monte Carlo results corroborate the theoretical analysis in Johansen (1991) with respect to the proposed sequence of tests starting with $H_0 : r = 0, H_1 : r \leq 1$ and stopping at the first nonrejection where r is the number of cointegrating vectors. The Monte Carlo results indicate that in the near cointegration case and in small samples, r may be underestimated.

Reimers (1992)

This is a Monte Carlo study of the Johansen LR test. The main conclusion is that the λ_{max} test statistic should be corrected for the number of estimated parameters to obtain satisfactory size properties in finite samples. This is accomplished by multiplying the test statistic by a factor $(T - np)/T$ where T is the number of observations, n is the number of variables in the VAR system, and p is the lag length of the VAR. This suggestion has been implemented by Gregory (1994) in his Monte Carlo study.

Cheung and Lai (1993)

Cheung and Lai examine the finite sample sizes of Johansen's (1991) LR tests for cointegration and their robustness to lag length specification and nonnormal errors. (The 1991 paper by Johansen includes a constant term, but the 1988 paper does not. Gonzalo (see chapter 5, section 5.7) considered Johansen's 1988 paper.) Cheung and Lai use response surface methodology similar to that of MacKinnon (see section 6.2). Unlike the study by Reimers they adjust the critical values rather than the test statistics. Their findings are:

- (i) The response surface estimation shows that the finite sample bias of Johansen's tests is a positive function of $T/(T - np)$ where T is the sample size, n is the number of variables and p is the number of lags in the VAR. Since $T/(T - np)$ is > 1 for any finite T , the tests are always biased toward finding cointegration too often if asymptotic critical values are used.
- (ii) If a low-order VAR model is used both the trace and maximum

eigenvalue tests are seriously biased toward spuriously finding cointegration.

- (iii) They find the AIC and BIC criteria not satisfactory in the selection of the lag length. But this is not surprising. See the discussion in chapter 3 (section 3.6.2) on the selection of lag length.
- (iv) Regarding nonnormality, skewness in innovations has a statistically significant effect on the test sizes of both the Johansen tests, but less so for the trace test. On the other hand, the trace test appears to be more robust to excess kurtosis (in the innovations) than the maximum eigenvalue test. Overall, the trace test is found to be more robust to nonnormality than the maximum eigenvalue test.

Toda (1994, 1995)

Toda (1995) investigates the performance of the Johansen (1991) tests (i.e., including a constant term rather than the tests in Johansen (1988) that do not include a constant term). These tests are for deterministic cointegration. Toda (1994) investigates tests for stochastic cointegration and the trace test statistic only. In his 1995 paper, Toda finds that the test performance is very sensitive to how close to unity the stationary roots of the process are. Even 100 observations are not sufficient to detect the true cointegration rank if a stationary root is close to unity, say 0.8 or higher. In this situation the cointegration rank is underestimated. Toda also finds that the test performance is very sensitive to the correlation between the innovations that drive the stationary and non-stationary components of the cointegrated process. If this correlation is zero, the tests may not perform well, even if the root of the stationary component is not close to unity. Toda suggests that a sample size of at least 300 is needed to see satisfactory performance of the Johansen test. However, such conclusions about the sample size depend on the error variances in the system (see the discussion on signal-noise ratios in chapter 5, section 5.7).

Toda (1994) investigates tests for stochastic cointegration as discussed in Johansen (1992) and Perron and Campbell (1993). In this paper Toda considers both the trace and maximum eigenvalue test statistics. His conclusions are similar to those in Toda (1995) for deterministic cointegration discussed earlier, except that the tests from the modified procedures perform worse than the tests for deterministic cointegration. Among the trace tests and maximum eigenvalue tests, neither dominate

the other except that the latter are better when the power of the tests is low. This study is rather discouraging with regard to Johansen-type tests for stochastic cointegration.

A summary assessment of the Monte Carlo studies

In addition to these studies that concentrate on the Johansen tests, there are other studies that analyze the Johansen tests along with other single equation tests (see, e.g., Gregory, 1994 and, Haug, 1996). They both come to the conclusion that no test dominates the others and hence it is important for empirical researchers to report the results from several tests for cointegration. However, this raises the issue of what to conclude in case of conflicting results from the several tests. Combination of tests results (as in the Fisher test discussed in chapter 4 in connection of panel data unit root tests) is not possible since the tests are not independent.

Gregory concludes from his Monte Carlo study that the ADF, Z_α , Z_t tests appear to be most reliable in terms of size and power. The LR tests of Johansen (as well as the Stock and Watson test) tend to overreject the null. (Note that these two sets of tests are from different categories as we shall discuss later.) The paper by Haug comes to different conclusions. It concludes that the Johansen and Juselius λ_{max} test and the ADF test reveal least size distortion. First of all, the two tests are not comparable – the λ_{max} test is for the number of cointegrating vectors and the ADF test is a single equation test for testing the existence of a cointegration relationship. Anyway the conclusion that the λ_{max} test exhibits the least size distortion is in conflict with the conclusion by Gregory and several others.

A major drawback of these studies is that when considering single equation tests like the ADF, Z_α , Z_t , etc., the ECM-based tests discussed in section 6.3 are completely excluded. The ECM-based tests have been shown to have superior power properties than the ADF, Z_α , and Z_t tests. Thus, any conclusions about the latter tests are almost useless. We also remarked in section 6.2 that any of the unit root tests can be used to test for cointegration. Thus, given more powerful unit root tests (such as the DF–GLS and modified Phillips–Perron tests discussed in chapter 4) one should no longer use the ADF, Z_α , and Z_t tests for cointegration. Of course, when using the new tests for unit roots, one cannot use the significance levels tabulated, because the residuals used in the cointegration tests are estimated errors. However, it is always best to get the appropriate significance levels by Monte Carlo methods

(or bootstrap techniques as described in chapter 10) rather than from the asymptotic distributions.

Furthermore, it is important to separate out single equation tests and system tests when reporting the results from Monte Carlo studies. Gregory does this but Haug does not. It is important to bear in mind that when talking about cointegration tests, we are referring to three categories of tests:

- (i) Single equation tests – these test whether a given set of $I(1)$ variables are cointegrated.
- (ii) Multiple equation tests such as Johansen's λ_{max} and trace tests – these are tests for the number of cointegrating vectors.
- (iii) Tests for significance of the coefficients of the estimated cointegrating vector – these tests apply both in the single equation as well as the multiple equation contexts.

The Monte Carlo studies should separate out the tests in these three categories.

Xiao and Phillips (1996) extend the Elliot–Rothenberg–Stock procedure we discussed in chapter 4 to the case of cointegration but they do this for tests in category (ii) (which is, of course, the more difficult problem). What we were mentioning earlier is to use that test for tests in category (i) instead of the ADF, Z_α , and Z_t tests for cointegration.

The Monte Carlo studies on the Johansen tests (tests in category (ii)) discussed earlier emphasize different characteristics of these tests. There are no conflicts in the conclusions. We can summarize the results as follows:

- (i) Tabulated critical values based on asymptotic distributions may be inappropriate if sample sizes are 100 or smaller (Podivinsky).
- (ii) All tests can be misleading if too few variables are included (Podivinsky).
- (iii) Insufficient lag length can lead to substantial size distortions. Overspecification, on the other hand, leads to loss of power (Boswijk and Franses).
- (iv) The Johansen λ_{max} test statistic should be corrected for the number of estimated parameters. This is done by multiplying it by $(T - np)/T$, where T is the number of observations, n is the number of variables, and p is the number of lags.

- (v) If a low-order VAR model is used both the trace and λ_{max} statistics are seriously biased toward spuriously finding cointegration.

There appears to be scope for improvement with the Johansen tests by using efficient detrending procedures, as discussed in Xiao and Phillips (1996). More discussion of tests in category (iii) will be provided in chapter 10.

6.5.3 Further comments on the Johansen tests

The Johansen procedure is the most commonly used system method in cointegration analysis. It has been built into several econometric software packages widely used (e.g., Microfit, PC-GIVE, E-VIEWS) and has a separate software package CATS in RATS. The Monte Carlo studies reviewed in the previous section discuss some problems in this procedure. The main problems are sensitivity to misspecification of the lag length, and substantial size distortions in the tests for the second and subsequent cointegrating vectors when the ratio of data points to the number parameters is small (of the order of 5 or less). See Toda (1994) and Ho and Sørensen (1996). We shall discuss a few more of the problems noted in the literature.

The Johansen procedure identifies the cointegration space and not the cointegrating vectors because linear combinations of cointegrating vectors are themselves cointegrated. Thus, some further identifying conditions need to be imposed. We have discussed this problem in chapter 5 (section 5.6).

Gonzalo and Lee (1995) show analytically and numerically that in some situations of practical interest the Johansen LR tests tend to find spurious cointegration with probability approaching one asymptotically. They consider two situations in which this occurs: the first is one where the variables have long memory and they are not pure $I(1)$ processes, but they are difficult to differentiate from $I(1)$ processes with standard unit root tests (we consider these in chapter 9). The second corresponds to $I(1)$ variables whose VAR representation has a singular or near singular covariance matrix. They argue that the Engle–Granger is more robust in these situations than the Johansen LR tests and recommend using both tests in order to avoid or discover a pitfall. However, in practice, it is almost always the case that the two procedures give different results, but it has been noted that the difference is not necessarily due to this particular pitfall. In fact, as argued in section 6.3 on ECM-based tests,

the Engle–Granger tests have been found to have very low power. The appropriate course of action is to apply tests for these situations and devise powerful cointegration tests to take account of these problems.

The study by Gonzalo and Lee considers the problems with the Johansen procedure if the variables are not $I(1)$ because of long memory. The problems actually arise for all cointegration tests, not just the Johansen tests. Elliott (1995a) raises another issue: the problem that the variables considered are not $I(1)$ but close to unit root processes. Again as he clearly indicates this is a problem with all cointegration tests, not just the Johansen tests, because they all apply unit root tests (which have low power) in the first step and then assuming that the variables are $I(1)$ proceed with the cointegration tests.

Elliott (1995a) shows, analytically, using local to unity asymptotic distribution theory that the estimates of the cointegrating vector when there is a large root in the explanatory variables which is not equal to one but close to one, remain consistent, but the tests for the coefficients of the cointegration vector tend to overreject the true null hypotheses. Elliott shows that the extent of this overrejection depends on the extent of the departure of the largest root in the regressor from one and on the degree of simultaneity between the errors in the equation of interest and the explanatory variables. For any deviation of the largest root in the explanatory variable from one, the size of the test approaches one as the degree of simultaneity increases. Elliott argues that this can be a very important problem in empirical work. For plausible parameter values he shows that the actual size can be 50 percent when the nominal size is 5 percent. For many economic variables the confidence intervals for the largest root include the unit root but they do include many stationary roots as well. For instance, for the US GNP quarterly data from 1948:I to 1992:I, a 95 percent confidence interval for the largest root is $[0.885, 1.03]$. Since the value 1.0 is in this interval we would not reject the unit root hypothesis. But we would not reject many stationary alternatives either. Elliott and Stock (1994) show that pre-tests to distinguish $I(0)$ from $I(1)$ and then proceeding in subsequent analysis is to use standard normal asymptotic theory if $I(0)$ is accepted and nonstandard theory if $I(1)$ is accepted results in nominal sizes of up to 30 percent for a nominal test of 10 percent. These problems and solutions to this problem that they propose are discussed in section 7.5 of the next chapter. The problem with cointegration tests is this two-step procedure that is universally followed.

We shall not go into the asymptotic distributions that Elliott derives. Elliott considers pre-tests in reference to the model

$$y_{1t} = \alpha_1 + \rho y_{1,t-1} + \varepsilon_{1t}$$

$$y_{2t} = \alpha_2 + \gamma y_{1,t-1} + \varepsilon_{2t}$$

The problem is testing hypotheses about γ when $\rho = 1 + c/T$ is close to unity but not equal to 1.

This is the model analyzed by Elliott and Stock (1994), which we discuss in greater detail in section 7.5 of the next chapter. For this model he derives:

- (i) the asymptotic distribution of $T(\hat{\gamma} - \gamma)$,
- (ii) the asymptotic distribution of the t -statistic $t_{\hat{\gamma}=\gamma}$,
- (iii) the asymptotic sizes of the tests

for all the efficient methods of estimation (discussed in chapter 5). The conclusions have been stated earlier. Elliott shows that the problem can be seen as an omitted variable bias problem. In the absence of any dynamics in the errors, the estimation of γ involves estimating the equation (by taking conditional expectation of ε_{2t} given ε_{1t})

$$y_{2t} = \alpha + \gamma y_{1,t-1} + \phi(1 - L)y_{1t} - \phi(\rho - 1)y_{1,t-1} + v_t$$

where $\phi = \text{cov}(\varepsilon_{1t}, \varepsilon_{2t})/\text{var}(\varepsilon_{1t})$. If we impose $\rho = 1$ when it is not, we are omitting the last term in this equation and this is the source of the bias. This depends on ϕ and $(\rho - 1)$. If $\phi \simeq 0$, then even if ρ is not close to 1, the problem is not severe. Elliott argues that in practice the problem is not so severe if the simultaneous equation bias is not large. Nevertheless, it is always the best to use a more efficient method of estimation than OLS.

The problems discussed by Elliott refer not to tests for cointegration directly (which we have been discussing) but to tests on the coefficients of regressions when there is uncertainty whether they are cointegrating regressions. In the example considered by Elliott it is not certain whether γ is a cointegrating vector or not.

6.6 Cointegration tests based on LCCA

As described in chapter 5 (section 5.5.3) the Box-Tiao method *levels canonical correlation analysis* (which Bewley terms LCCA) has been used to compute cointegration vectors by Bossaerts (1988) and Bewley

and Orden (1994). Bossaerts suggested that the canonical variates obtained by this method could be directly tested for cointegration. If z_t is the canonical variate corresponding to the minimum eigenvalue, then one estimates

$$z_t = \rho z_{t-1} + \varepsilon_t \quad (6.10)$$

and tests the hypothesis $\rho = 1$. Bossaerts conjectured that the significance tables in Fuller (1976) could be used for this test, and Bewley and Orden conjectured that the critical values depended on the number of variables in the system.

Bewley and Yang (1995) to be referred to as BY examine these conjectures and develop four tests for cointegration based on the levels canonical cointegration. As discussed in chapter 5 (section 5.5.3) the eigenvalues are obtained from equation (5.6) as in the Johansen method except that, as noted there, the definition of R_{0t} and R_{1t} (hence S_{00} and S_{01}) are defined from the levels regression. The four tests developed by BY are:

- (i) The Dickey–Fuller type t -test based on $\hat{t}_\rho$.
- (ii) The Dickey–Fuller type coefficient test based on $T(\hat{\rho} - 1)$. Both of these are for the equation (6.10).
- (iii) The minimum eigenvalue test based on

$$\lambda_{min} = T(1 - \hat{\lambda}_1)$$

where $\hat{\lambda}_1$ is the minimum eigenvalue.

- (iv) The trace test based on $T \sum_{i=1}^{n-r} (1 - \hat{\lambda}_i)$.

Tests (iii) and (iv) are the analogues of Johansen's maximum eigenvalue and trace tests, respectively. However, the hypotheses they test are different.

BY show that the test statistics have nonstandard asymptotic distributions and provide critical values for each of the tests for models up to six variables. These tables also show that Bossaerts conjecture that one could use the Dickey–Fuller tables is not correct. The asymptotic distributions and hence the critical values depend on the number of variables in the system.

BY perform a Monte Carlo study to examine the small sample performance of these tests. The model used is the bivariate model

$$\begin{aligned} y_t - bx_t &= u_t & u_t &= \rho u_{t-1} + \varepsilon_{1t}, & |\rho| < 1 \\ x_t - ay_t &= v_t & v_t &= v_{t-1} + \varepsilon_{2t} \end{aligned}$$

Table 6.6. *Critical values of the LCCA-based tests*

<i>n</i>	<i>T</i>	$\hat{t}_\rho$		$T(\hat{\rho} - 1)$		$T(1 - \hat{\lambda}_1)$	
		1%	5%	1%	5%	1%	5%
2	75	4.26	3.65	27.85	21.64	45.03	36.50
	150	4.16	3.59	29.46	22.53	52.34	41.01
	250	4.13	3.57	30.21	22.93	56.15	43.11
4	75	5.49	4.90	42.62	35.89	60.62	54.09
	150	5.29	4.74	46.34	38.43	77.81	66.32
	250	5.21	4.68	47.96	39.52	86.04	72.14
6	75	6.58	5.96	54.75	48.21	69.19	64.87
	150	6.22	5.70	61.24	53.08	96.73	86.64
	250	6.11	5.60	64.36	55.19	111.35	97.38

Source: Bewley and Yang (1995, table 1, p. 992).

where ε_{1t} and ε_{2t} have zero means and covariance matrix

$$\begin{bmatrix} 1 & \theta\sigma \\ \theta\sigma & \sigma^2 \end{bmatrix}$$

The crucial parameter in the system are ρ and θ . BY observe that for this model, Johansen’s maximal root is (asymptotically) $(1-\rho)/[2-(1+\rho)\theta^2]$. Hence as $\theta \rightarrow 1$, the power of Johansen’s λ_{max} test statistic, $T \ln(1 - \lambda_{max})$ approaches unity. On the other hand, the LCCA minimal root can be expressed asymptotically as $\lambda_{min} = \rho^2$. This does not involve θ and hence the power of the LCCA-based test is insensitive to θ in finite samples.

BY find that for approximately $|\theta| < 0.5$ each of the LCCA tests has greater power than the Johansen tests or the Engle–Granger test. If $|\theta| > 0.5$, then the Johansen test dominates. Also the LCCA trace test is as powerful as the λ_{min} test for θ close to zero, but its power deteriorates rapidly as $\theta \rightarrow 1$. On the basis of the simulation experiment BY argue that the LCCA trace test is not recommended. But each of other tests perform well compared with the two Johansen tests and the Engle–Granger test except when the disturbances generating the cointegrating equation are highly correlated with those generating the common trends.

BY present critical values for the four tests at significance levels 1 percent, 5 percent, and 10 percent for sample sizes $T = 75, 100, 150, 200, 250$, and 500, and the number of variables $n = 2, 3, 4, 5, 6$ (table 1, p. 992 of their paper). Part of their table is presented in table 6.6. We have omitted the trace test because BY said it was not recommended.

Note, however, that these critical values in BY are for the model with no constant term as in Johansen (1988). Yang and Bewley (1996), however, provide critical values for the tests in the case of drift. Johansen has different tests for models with constant term (Johansen, 1991) and models with trend (Johansen, 1992). The tables from Osterwald-Lenum presented earlier are for these three cases. Thus, when referring to Johansen tests, it is important to keep in mind which of the Johansen tests one is talking about.

Another difference between the Johansen tests and the tests in BY is that in the Johansen procedure the tests determine the number of cointegrating vectors. The tests in BY test one eigenvector at a time sequentially. First the eigenvector corresponding to the smallest eigenvalue λ_1 is tested for a unit root. Then the dimension of the system is reduced and the roots are recalculated and the procedure is repeated.

Bewley and Yang (1996) pursue this comparison further and investigate the properties of the LCAA-based tests by comparing them with Johansen's LR tests, examining size distortions in tests for the second and subsequent cointegrating relationships and the impact of misspecifying the lag length in each of these situations. Since the LCCA-based tests have also been found to suffer from size distortions, they propose some simple adjustments to improve their small sample performance. The main conclusions of this study are:

- (i) The LCAA-based tests are less sensitive to misspecification of the lag length than the Johansen LR tests.
- (ii) The size distortion problem in the LCAA tests can be mitigated by applying a correction factor $[(T - p)/T]^{0.75}$ where T is the sample size and p is the lag length. Note that a factor of $[(T - p)/T]$ was suggested by Reinsel and Ahn that was also used in the Reimers (1992) and Cheung and Lai (1993) papers reviewed in the previous section..
- (iii) In an empirical example in a *typical macroeconomic setting* where all evidence from exploratory data analysis points to the existence of cointegrating relationships, the LR tests reject cointegration whereas the LCCA-based tests detect cointegrated relationships. As noted earlier, in situations where the degree of correlation between the innovations in the cointegrated relationship and innovations in the process driving the stochastic trend are weakly correlated, the LCCA-based tests are more powerful than the

Johansen tests. This appears to be the case in many practical situations.

We shall not go into more details of this study.

6.7 Other tests for cointegration

In the preceding sections we discussed residual-based tests and tests based on canonical correlations. We shall now discuss some other tests.

6.7.1 Tests based on principal components

The principal component method is another multivariate technique that depends on eigenvalues of a matrix like the canonical correlation method. Simply stated it identifies the linear combination of a set of n variables $y = (y_1, \dots, y_n)$ that has the highest (or lowest) variance subject to a normalization condition. Let the covariance matrix of y be Σ . Then variance of a linear combination $c'y$ is $c'\Sigma c$. Maximizing this with respect to c subject to a normalization rule $c'c = 1$ gives c as the eigenvector of $|\Sigma - cI| = 0$. If $\lambda_1, \lambda_2, \dots, \lambda_n$ are the eigenvalues of Σ and $c_1, c_2, \dots, c_n$ are the corresponding eigenvectors, then the c_i are mutually orthogonal and $\text{var}(c'_i y) = \lambda_i$. If we order the λ_i in descending order, $\lambda_1 > \lambda_2 > \dots > \lambda_n$, then $c'_1 y$ is the linear combination that has the highest variance. $c'_2 y$ is the linear combination orthogonal to this with the next highest variance, and so on. $c'_n y$ is the linear combination with the largest variance.

The idea in using the principal component approach to cointegration is that $I(1)$ variables have much higher variance than $I(0)$ variables. Thus, the principal components corresponding to the lowest λ_i give the cointegrating vectors and those corresponding to the largest λ_i give the common stochastic trends. The procedure of finding the minimum roots of a covariance matrix is similar to the LCCA method we discussed in the preceding section.

Stock and Watson (1988), Phillips and Ouliaris (1988), and Harris (1997) use the principal component approach to derive tests for cointegration. The last paper is the most recent and the most comprehensive one. But we shall leave the details to interested readers because going through the paper will involve intricate detail.

6.7.2 Other tests

Some other tests suggested in the literature are those by Saikkonen (1992), Kleibergen and van Dijk (1994), and Horvath and Watson (1995). Saikkonen's tests are system tests. He derives the tests using an autoregressive approximation to the DGP (data generating process). He derives Wald tests for cointegration that follow the usual χ^2 distributions under the null, assuming the number of cointegrating vectors and the normalization rules to be known. He also derives tests for these assumptions. These tests by Saikkonen have not been used in application. Even Gregory (1994) has Saikkonen's paper in the references but does not include them in his Monte Carlo study. Hence we shall not discuss them in detail here. But the other two have different approaches. They are both system tests starting with the VAR system as in the Johansen procedure. The Horvath and Watson method is based on some prior information on some of the cointegrating vectors.

Consider the LR matrix B_1 in the VAR model we discussed in section 5.5.1 (chapter 5) in connection with the Johansen procedure. The Johansen procedure first determines the rank of the cointegration space and introduces any identifying information later. The framework considered by Kleibergen and van Dijk (to be referred to as KVD) differs from Johansen's and considers the identification of the cointegration space and that of the different cointegrating vectors in a single step using parametric identifying restrictions. For conformity with the KVD notation, we shall use Π for the LR matrix B_1 . Also, we shall write $\Pi = \beta\alpha$ (instead of $B_1 = \alpha\beta'$ used in chapter 5).

First partition Y_t into two components Y_{1t} and Y_{2t} of dimensions r and $n - r$. Then perform a triangular decomposition of the matrix Π . The parametric restriction imposed on Π to ensure identification are

$$\beta^* = \begin{bmatrix} I_r & 0 \\ -\beta_2 & I_{n-r} \end{bmatrix}$$

and

$$\alpha^* = \begin{bmatrix} \alpha_{11} & \alpha_{12} \\ 0 & \alpha_{22} \end{bmatrix}$$

so that

$$\Pi = \begin{bmatrix} \alpha_{11} & \alpha_{12} \\ -\beta_2\alpha_{11} & -\beta_2\alpha_{12} + \alpha_{22} \end{bmatrix}$$

where β_2 is $(n - r) \times r$, α_{11} is $r \times r$, α_{12} is $r \times (n - r)$, and α_{22} is $(n - r) \times (n - r)$. The number of cointegrating vectors can be tested by

testing the significance of α_{22} for different values of r , the cointegrating rank. If $\alpha_{22} = 0$, then

$$\beta^* = \beta = (I_r - \beta_2')'$$

and the error correction coefficients are given by

$$\alpha^* = \alpha = (\alpha_{11}, \alpha_{12})$$

If $\alpha_{22} \neq 0$, then β is no longer interpretable as a cointegrating vector. Thus, testing for cointegration is performed by testing whether $\alpha_{22} = 0$ for different values of r . Partition Π into

$$\Pi = \begin{bmatrix} \Pi_{11} & \Pi_{12} \\ \Pi_{21} & \Pi_{22} \end{bmatrix}$$

If Π_{11} is of full rank, then α and β are exactly identified. We have the relationships: $\hat{\alpha}_{11} = \hat{\Pi}_{11}$, $\hat{\alpha}_{12} = \hat{\Pi}_{12}$, $\hat{\beta}_2 = -\hat{\Pi}_{21}\hat{\Pi}_{11}^{-1}$, and $\hat{\alpha}_{22} = -\hat{\Pi}_{21}\hat{\Pi}_{11}^{-1}\hat{\Pi}_{12} + \hat{\Pi}_{22}$. KVD construct the Wald tests to test $\alpha_{22} = 0$. We shall not go into the details of this and refer readers to the KVD paper.

The Horvath and Watson paper is concerned with the rank test for cointegration rank when some of the CI vectors are known. This sort of situation often arises in the case of foreign exchange markets.

There have been very few applications of the KVD testing procedure and the Horvath and Watson test. The KVD procedure is illustrated in Kleibergen *et al.* (1995). The Horvath and Watson procedure is illustrated in Edison *et al.* (1997). Edison *et al.* find the critical values from the asymptotic distributions given in Horvath and Watson paper to be too low compared with the critical values they derive by Monte Carlo methods. This suggests that in all the applications of the several tests we have discussed, it is better to derive the critical values by Monte Carlo methods rather than from the asymptotic distributions.

6.8 Miscellaneous other problems

6.8.1 Effect of time disaggregation

In chapter 4 (section 4.7) we discussed the effects of time aggregation on the power of unit root tests. We shall now discuss this problem with reference to cointegration tests.

Hooker (1993) first argued through Monte Carlo experiments that temporal disaggregation (using monthly versus quarterly data or quarterly versus yearly data) yields significant increases in the power of ADF

cointegration tests and that this result is contrary to the findings regarding ADF unit root tests that showed increases to power by increasing the time span of the data and not by increasing the number of observations by time disaggregation. Lahiri and Mamingi (1995), however, argue that their Monte Carlo experiments yielded no increase in power by time disaggregation, thus contradicting Hooker's results and reiterating the conclusions reached regarding the unit root tests. Hakkio and Rush (1991), which we shall discuss in section 6.9, also present Monte Carlo evidence that time disaggregation does not improve the power of residual-based cointegration tests.

In yet another study, Hu (1996) investigates the effect of skip sampling and time disaggregation on the power of Johansen's trace and maximum eigenvalue tests. She finds that although there are power gains when switching to high frequency data to gain more observations with a fixed time span, the power gains are much more significant when data with a longer time span are used. It also confirms the findings reported in chapter 4 (section 4.7) that the length of the sample time span is more important than the number of observations within a fixed time span.

The study by Hu considers both the cases of skip sampling and time aggregation. Also it studies the more powerful Johansen tests than the residual-based tests studied earlier. It also provides critical values for skip sampling and time aggregation for the Johansen tests.

Although the gains in power by increasing the number of observations with a fixed time span are not as great as those coming from longer time spans, it is not true that one should not use high frequency data. In practice data over long time spans may not always be available and even if they are, there is the problem of structural changes (discussed later in chapter 13). Thus, in practice one should use the highest frequency data available (although there are the problems of seasonal roots which we discuss in chapter 12). Further discussion of time aggregation problems is in chapter 12 (section 12.10).

6.8.2 The pre-testing problem

The pre-testing problem arises in the discussion of cointegration tests at several stages. First, one has to check for unit roots before starting cointegration tests. Thus, the unit root tests are pre-tests and it is not clear what effect this has on the significance levels used for the subsequent cointegration tests. Also, it is customary to use the *conventional* 5 percent and 1 percent significance levels for the unit root tests. (Who

started this *convention* is a good question. It is R.A. Fisher who suggested it in an obscure paper in 1923 and since he is the *father* of modern statistics, it has been blindly followed ever since.)

There is a substantial statistical literature on the pre-testing problem and although there is no definite answer to the question of what significance levels should be used for pre-tests, there is a consensus that the 5 percent and 1 percent significance levels are the *wrong* ones to use. The significance levels used should be much higher, say 25 percent (this is the rule suggested by the statistician I.J. Good). If this is done, we would reject the unit root null more often than we do at the 5 percent level, and then a discussion of cointegration is out of the question.

A further problem is that just because we failed to reject the null of unit root at the 5 percent level does not mean that we have a unit root process. There is a nontrivial probability that it is a stationary process. Also, the root can be close to one but not exactly one. As argued by Elliott (1995a), and as discussed in section 6.5.3 earlier, this has the consequence that the tests on the coefficients of the cointegrating vector have substantial size distortions.

However, the problem is not just one of size distortion of the tests. There is the problem of bias in the estimated coefficients as well. The problem of pre-test bias has been investigated by Kim (1995), in a typical triangular cointegration system. He follows a decision theoretic approach and evaluates the different estimators by comparing their risk functions. He considers three estimation methods: OLS, 2SLS, and FIML, and three different cases: stationary but near-integrated, unit root, and explosive but near-integrated series. His conclusions are that the FIML estimator of a cointegrating vector shows high risk in an explosive but near-integrated series, but it turns out to be the best estimator in stationary but near-integrated and unit root cases. The OLS estimator shows high risk in all three cases. The 2SLS estimator shows consistently low risks for all three cases. (A decision theoretic approach was also used by Lahiri and Paul to evaluate pre-test estimators in AR(1) models (see section 4.10 of chapter 4).)

The pre-testing problem does not arise in a Bayesian context because the respective posterior probabilities of unit roots and stationarity are taken into account. The Bayesian approach is discussed in chapter 8.

The second problem where pre-testing occurs is in the context of system procedures (as in the Johansen method or the Box-Tiao method). First, we use the tests to determine the number of cointegrating vectors and then we use tests on the significance of the coefficients of the esti-

mated cointegrating vectors. This is a two-step testing procedure. The significance levels used at the second stage are affected by the significance levels used in the first stage.

Elliott (1995b) suggests a one-step procedure as an alternative to solve the pre-testing problem caused by the two-step procedure. He suggests testing the coefficients of the implied error correction model. However, his Monte Carlo studies indicate that this one-step procedure is not always better than the two-step procedure particularly when there is strong prior information on the number of cointegrating vectors (or when the first test strongly favors one number for the cointegrating vectors). He, therefore, suggests using both the one-step and two-step procedures, but this creates problems when they give conflicting results.

The issues relating to pre-testing biases in cointegrating tests are far from resolved. A Bayesian investigation of this problem would be very fruitful but this has not been done yet.

6.8.3 Using cointegration concepts to test for unit roots

Tests for unit roots usually precede tests for cointegration. So it looks paradoxical that one should use the concepts of cointegration to derive a test for unit roots. However, this is what is done in Hansen (1995). He derives a unit root test that is related to tests for cointegration.

It is often said that “unit root tests have low power,” but this may be due to the fact that tests for unit roots depend on univariate time series and do not take account of information in related series. Hansen shows that taking account of this information results in considerable increase in the power of unit root tests. He suggests a *covariate-adjusted* Dickey–Fuller (CADF) t -statistic and tabulates its critical values. A GAUSS program that calculates the test statistics and critical values is available from Bruce Hansen.

The test that Hansen derives is based on the t -ratio of y_{t-1} in a transformed equation of the AR(1) model

$$\Delta y_t = \gamma y_{t-1} + u_t$$

It is assumed that there is a set variables x_t which is $I(1)$ so that Δx_t is $I(0)$. Assume that $(\Delta x_t, u_t)$ are *iid* with mean zero and $E(\Delta x_t^2) = \sigma_x^2$, $cov(\Delta x_t, u_t) = \sigma_{xu}$. Then we can write the AR(1) model as

$$\Delta y_t = \gamma y_{t-1} + b' \Delta x_t + e_t$$

where $e_t = u_t - b' \Delta x_t$ and $b = \sigma_{xu} / \sigma_x^2$ is the regression coefficient of u_t

on Δx_t . Note that γ retains the same meaning in this equation as in the AR(1) model. However, $\text{var}(e_t) < \text{var}(u_t)$ unless $\sigma_{xu} = 0$ in which case the two variances are equal. Hence, Hansen argues that the parameter γ can be more precisely estimated (at least in large samples) from the transformed equation. Note, however, that the variance of $\hat{\gamma}$ from the transformed equation is $\text{var}(e_t)/\text{var}(\tilde{y}_{t-1})$ where $\tilde{y}_{t-1}$ is the residual from a regression of y_{t-1} and Δx_t . Thus, $\text{var}(\tilde{y}_{t-1}) < \text{var}(y_{t-1})$. Hence it is not exactly obvious that the variance of γ from the transformed equation is smaller. Hansen argues that this transformed equation is similar to that considered by Kremers *et al.* (1992) discussed in section 6.3, but those authors consider it as a test for cointegration and not for univariate unit roots.

The derivation of the test statistic for γ in the transformed equation (particularly for leads and lags in the errors) is complicated but Hansen shows that the t -statistic has a distribution that is a convex combination of the DF distribution and the standard normal, depending on a parameter ρ^2 that can be computed from the data. Hansen provides 1 percent, 5 percent, and 10 percent critical values of the test statistic for values of ρ^2 between 0 and 1, for the cases: standard, demeaned, and detrended. His GAUSS computer program gives the details.

Hansen shows that there is a considerable gain in the power using the CADF t -statistics when testing for unit roots. As is usual he illustrates it using the Nelson–Plosser data (the common *guinea pig* for every new unit root test).

The idea of using covariates in unit root testing is a good one, but the fact that only I(1) variates can be considered is very restrictive. These variables are judged to be I(1) by some earlier tests that are based on single-variable unit root tests. Also, there is the issue of *why unit root tests* are not included in this whole discussion. Why not jointly test for unit roots in y_t and x_t instead of a test for unit roots conditional on x_t being I(1). The panel data unit root tests discussed in section 4.9 fall into this category of joint tests. Thus, there are several other issues in the use of covariates in unit root testing, although the argument about considering other variates makes sense. The basic issue is how best to use the information in y_t and x_t to analyze whatever problem you are interested in. Starting from a VAR is the common route.

The use of unit root tests using a single variable does not make economic sense in many situations and can sometimes give misleading results. An example of this is testing the purchasing power parity (PPP) theory based on unit root testing in real exchange rates. Steigerwald

(1996) illustrates this point. He shows how considering nominal exchange rates and prices as jointly endogenous and incorporating dynamics in the equations produces dramatically different results compared to those based on tests for unit roots on real exchange rates.

6.9 Of what use are cointegration tests?

We have discussed several tests for cointegration. However, several Monte Carlo studies about the power of these tests show that in general these tests are not powerful. Also as noted by Elliott (1995a) there is the problem of substantial size distortions if the variables under consideration are not really $I(1)$ but have roots close to unity.

Campbell and Perron (1991) argue that it may be difficult to distinguish processes that exhibit cointegration from those that do not, and more so to estimate precisely the exact number of cointegrating relationships. In fact if the goal of cointegration tests is to uncover the true long-run relationships, this argument is very disconcerting. In some applications the goal is not to uncover the *true number* of cointegrating relationships per se, but rather to have a useful guide in imposing restrictions on VAR models and ECM models that may lead to more efficient estimation and improve forecasting performance. For this purpose the testing procedures described in this chapter can serve a useful purpose, although recent studies show that the VECM model, the VAR with cointegrating restrictions does not provide much evidence of improvement in forecasting. See chapter 5 (section 5.8).

Another issue is how to interpret the results of cointegration analysis. If there is only one cointegration relationship, then it may be easy to interpret it as a long-run relationship. However, if we find the number of cointegrating vectors to be greater than one, this creates problems of interpretation, and we need to bring in more economic theory. We discussed the identification problems in chapter 5 (section 5.6).

In spite of the limitations mentioned above, many empirical studies take cointegration very seriously and aim to test economic theories using cointegration techniques. We shall review some of these studies and see what has been accomplished. We shall start with the work on (see Maddala, 1991)

- (i) tests of the *market efficiency hypothesis* (MEH),
- (ii) tests of the long-run demand for money,
- (iii) tests of the purchasing power parity (PPP) theory.

6.9.1 Testing the MEH

If the prices in two markets are cointegrated this implies that it would be possible to forecast one from the other. This, in turn, implies that the markets are not efficient. The MEH thus implies absence of cointegration (or the non-rejection of the no-cointegration null).

In the case of the foreign exchange markets, the MEH for the spot rates implies that the spot rates for the different currencies should not be cointegrated. The MEH applied to the forward market implies that the forward rates for the different currencies are not cointegrated and that each of the spot rates is cointegrated with its own forward rate. MacDonald and Taylor (1989) consider monthly data on the log of the exchange rates for the period January 1973 to December 1985 for nine currencies (Australian dollar, British pound, Canadian dollar, Danish krone, Dutch guilder, French franc, German mark, Italian lira, and Japanese yen). They use the ADF test and find that the logs of all the spot rates are $I(1)$ and that there is no strong evidence of cointegration among the different spot rates (in pairwise cointegration tests). Hakkio and Rush (1989) consider German and UK spot and forward rates for the period July 1975 to October 1986 (on a monthly basis). They find that neither the two spot rates nor the two forward rates are cointegrated, thus, suggesting market efficiency. In addition both the German future spot and current forward rates, as well as the UK future spot and current forward rates appeared to be cointegrated. This again suggests market efficiency.

Both the papers, however, dealt with only bivariate tests of cointegration. It is intuitively obvious that they suffer from the usual omitted variable problem. Suppose that x , y , and z are three $I(1)$ variables which are cointegrated so that $x = \beta_1 y + \beta_2 z + u$ where u is $I(0)$. If now we run the regression $x = \beta_1 y + v$, then, since $v = \beta_2 z + u$ is $I(1)$, we will observe that x and y are not cointegrated. Thus, a bivariate cointegration test will not reject the hypothesis of no-cointegration. Such bivariate tests do not tell us whether markets are efficient or not. The Johansen (1988) procedure that tests for all possible cointegrating relationships among a set of variables is the appropriate one to use to test the MEH. Baillie and Bollerslev (1989) examined the logarithms of daily exchange rates for seven currencies for the period March 1, 1980 to January 28, 1985 and found that both the Engle and Granger two-step cointegration test and the Johansen test showed evidence of cointegration. The data were on the exchange rates versus the US dol-

lar for the UK, West Germany, France, Italy, Switzerland, Japan, and Canada. The rejection of the hypothesis of no-cointegration is partly due to the fact that they had a large sample of 1,245 (daily) observations.

Sephton and Larson (1991) provide convincing evidence that the issue of whether or not a set of spot rates are cointegrated depends on the time period considered. They show that the results from the Johansen procedure are sensitive to the time period considered. However, it is not just the Johansen procedure that gives such results. The instability can be observed with even the Engle-Granger two-step procedure. It is not at all surprising that tests for market efficiency are sensitive to structural changes because it has been noted that tests for unit roots are sensitive to structural changes. (This is discussed in chapter 13.) Note that in tests for cointegration the null hypothesis is one of no-cointegration. Since the significance levels used are the same (5 percent) no matter what the sample size is, the null hypothesis (of no-cointegration) is likely to be rejected if the number of observations is very large. This is just a statistical artifact. Bayesians have always objected to the idea of using a constant significance level irrespective of the sample size. Baillie and Bollerslev have 1,245 observations covering a period of five years. If we were to be analyzing the same period with monthly data, we would have only 60 observations, and the hypothesis of no-cointegration may not be rejected at the usual significance levels.

This point has been noted by Layton (1993). He argues that MacDonald and Taylor (1989) and Baillie and Bollerslev (1989) came to different conclusions due to the different sampling frequencies used. Since Baillie and Bollerslev used a multivariate cointegration testing framework while MacDonald and Taylor employed bivariate tests, one possible explanation for the difference in results could rest on the issue of technique. However, Layton argues that the contradictory results are due to the fact that Baillie and Bollerslev used daily data and MacDonald and Taylor used monthly data. Layton argues that foreign exchange markets display well-known varying degree of temporally correlated volatility depending on how frequently the data are sampled. He substantiates his arguments by considering the data on the Australian dollar sampled daily, weakly, and monthly. He tests the data for ARCH effects and finds that the evidence for cointegration becomes weaker as the ARCH effect gets stronger.

6.9.2 Testing long-run equilibrium relationships: the long-run demand for money

One other application of cointegration is to test the existence of long-run relationships. One argument sometimes made is that cointegration is about long-run economic relationships, and one needs really long time series (not in the number of observations but in time span) to use cointegration techniques (see, for instance, Hakkio and Rush (1991) for such an argument). This is not a meaningful argument for several reasons. If the variables are nonstationary, then existence of a long-run equilibrium economic relationship implies cointegration. But not all cointegrating relationships need have meaning in the sense of long-run economic relationships. Cointegration is a purely statistical concept. It is part of *a-theoretical* econometrics. Cointegrating relationships need not have any economic interpretation. Whether or not a series is long enough to establish a cointegrating relationship depends purely on the statistical properties of the series and whether the cointegrating relation has any economic interpretation can never be answered without any relevance to the question being asked.

How long the long run is depends on the speed of adjustment of the particular markets considered. This was the conventional wisdom in old style econometrics before cointegration. There is no reason to discard this. This is also an additional reason why the analysis of short-run dynamics is very important because that is where we get estimates of speeds of adjustment. For financial markets with rapid speed of adjustment, the long run is indeed short. For goods markets the speeds of adjustment are perhaps slow for some commodities and fast for others (there is not much empirical evidence on which commodities adjust fast which do not). Since the speeds of adjustment in the goods markets are the ones that are relevant for testing the PPP theory, it is likely that we would need a long time series (certainly not 100 years!) to apply the cointegration techniques. For testing the MEH certainly the data considered in the relevant studies are long enough for cointegration techniques to be applicable. Even for testing the PPP theory, a time series of 10–15 years can be considered adequate assuming some plausible speeds of adjustment in the traded goods markets. Thus, contrary to the claims in Hakkio and Rush (1991) that we have to wait some years to have data of adequate time span to use cointegration techniques, the problem with testing the PPP is not the lack of a long series of data, but rather the lack of a meaningful hypothesis to start with. For almost all

the problems analyzed by using cointegration, the data are long enough for cointegration techniques to be applicable.

Another argument why a long time span does not help in estimating cointegrating vectors is that a long time series is more likely to have quite a few structural changes, and under parameter changes, a single cointegrating relationship cannot be considered meaningful. There is abundant evidence to show that ignoring structural breaks leads to misleading inference on both unit roots and tests for cointegrating relationships. (This is discussed in chapter 13.)

What sort of results have been obtained on long-run economic relationships using cointegration techniques? The range of applications has been really wide, from the long-run demand for money, PPP theory and the permanent income hypothesis to US military expenditures and the dollar. For instance, Grilli and Beltratti (1989) use quarterly data for the period 1951:I to 1986:III and cointegration techniques to establish a significant relationship between US military expenditures (real) and the real exchange rate.

We shall consider two sets of studies – one on the long-run demand for money and the other on PPP theory. The studies discussed are not the most interesting in the respective areas, but we have chosen them because they illustrate some problems in the use of multivariate cointegration techniques. Given the limitations of bivariate cointegration relationships mentioned earlier, we shall consider only multivariate cointegration techniques. (The issue is similar to the use of simple versus multiple regression.) A commonly used method for this purpose is the Johansen method (1988). As discussed in chapter 5, this is a flexible technique that allows one to test for the number of cointegrating vectors, get the ML estimates of the coefficients, and so on.

Johansen and Juselius (1990) estimate demand for money functions for Denmark and Finland using quarterly data. For the Danish data the sample was 1974:I to 1987:III (55 observations). For the Finnish data the sample was 1958:I to 1984:III (67 observations). For the Danish data the order of cointegration was one which simplified the interpretation of the cointegrating vector as a long-run demand for money function. But for the Finnish data there were three cointegrating vectors and this caused problems of interpretation.

The study on demand for money for UK by Cuthbertson and Taylor (1990) – to be referred to as CT – presented similar problems. The variables they considered were: the log of nominal M_3 (m), the log of GDP deflator (p), the log of real GDP (y), three month treasury bill rate

(RB), and the own rate on money (RM) which is the maximum of the seven-day deposit rate and the rate on high interest checking accounts. They found that m and p were $I(2)$ but $m-p$ was $I(1)$ as were y , RB , and RM . Hence, subsequent analysis was conducted in terms of these four variables. The Johansen procedure suggested two cointegrating vectors. The hypothesis of income homogeneity for the two vectors could not be rejected, but the joint hypothesis of income homogeneity and equal and opposite signs for the interest rates (arguing that it is the interest rate differential that is the determinant of the demand for money) was rejected. The two cointegrating vectors were

$$(m - p)_t = y_t - 0.14RB_t + 0.034RM_t$$

and

$$(m - p)_t = y_t - 0.36RB_t + 0.34RM_t$$

CT, however, rejected the second vector saying that the interest elasticities were too high to be plausible and argued that the first equation represents the long-run demand for money.

Whenever we have more than one cointegrating vector, we have serious problems of interpretation. In the above example, since any linear combination of two cointegrating vectors is also cointegrating, there are an infinite number of elasticities one can generate. Thus, identification of the demand for money function requires some extraneous information. This is not surprising, because, as mentioned earlier, cointegration techniques are purely statistical in nature, and they belong to the area of a-theoretical econometrics, much the same way as VAR models do.

An alternative route is not to ask how many cointegrating relations there are (and what they are) but to test whether a prespecified relation derived on the basis of economic theory is a cointegrating relationship. One such candidate in the above example is

$$(m - p)_t = y_t - \beta(RB_t - RM_t)$$

Looking at the second cointegrating vector that CT estimated, it appears that one would not reject the hypothesis of cointegration. CT claim that this hypothesis was rejected, but this is surprising. One possible explanation is that the restriction was tested for both the vectors jointly, which is not a meaningful hypothesis to test.

This example on the demand for money by CT highlights the problems one faces when there are more than one cointegrating vector. It also illustrates the arbitrary way a cointegrating vector was picked up

and exhibited as the long-run demand for money. Podivinsky (1990) suggests two more problems with the application of the Johansen procedure: the first, that his tabulated critical values may be inappropriate when applied to sample size of 100 or smaller; second, that the tests can be misleading in the presence of omitted variables. Hence, it is important to include all the plausible variables in the analysis from the start.

6.9.3 Cointegration and the purchasing power parity (PPP) theory

The literature on testing the PPP theory is enormous, and in recent years many of these studies have used the cointegration methods. All this work, however, is like looking for a black cat in a dark room in which no cat (black or white) exists. There has been discussion of whether PPP holds during some time periods (e.g., in the 1920s, in the nineteenth century), or some countries (Latin American countries with high inflation rates), or in one of the exchange rate regimes (fixed exchange rate periods as opposed to flexible exchange rate periods), and so on. One argument for devoting so much attention to this theory is that it is one of the equations in almost all models of exchange rate behavior, and it continues to be used even in spite of overwhelming evidence against it.

The PPP constrains the exchange rate between two countries to be proportional to the ratio of the price levels in the two countries. The PPP theory is closely related to the *law of one price* which states that the price of a commodity is the same everywhere in the world. Apart from the fact that commodities with the same name are not the same everywhere, and transportation costs also account for differences in prices, even arbitrage that should bring prices to equality is not instantaneous. It takes time for economic agents to react to price differences. Moreover, the law of one price does not apply to nontraded goods. Finally, even if PPP held exactly for each of the traded commodities, the PPP theory can be violated because of differences in the weights given to the different commodities in the construction of price indexes in the two countries. Thus, PPP cannot be considered to be a sensible hypothesis. As Leamer (1991) argues, testing a sharp hypothesis like the PPP at a constant significance level is meaningless to start with. The more interesting question is not whether PPP holds exactly in the long run, whatever the term long run means (100 years?) but as to how fast arbitrage eliminates price differences, and for which commodities. The

interesting questions are how far do exchange rates deviate from the price ratio and for how long. As a theory of the determination of exchange rates, the price ratio is only one of the determining variables and the question is how important is it compared to other variables, such as money supplies, interest rate differentials, and so on.

The empirical work on PPP using the Johansen procedure illustrates the problems of obtaining cointegrating relationships when some important variables are omitted from the analysis. Johansen and Juselius (1992) – to be referred to as JJ – analyze the exchange rate for UK using quarterly observations from 1972:I to 1987:III. The variables they considered were: the log of wholesale price index for UK (p_1), the log of trade-weighted foreign price index (p_2), the log of the effective exchange rate (e_{12}), three month treasury bill rate in UK (i_1), and three month Euro-dollar interest rate (i_2). They test the PPP which says

$$p_1 - p_2 - e_{12} = 0$$

and the uncovered interest parity (UIP) which says

$$i_{1,t} - i_{2,t} = \Delta e_{12,t+1}$$

JJ find that there are two cointegrating vectors, one of which can be identified as $(0, 0, 0, -1, 1)$ indicating that the interest rate difference is stationary. The hypothesis that the vector $(1, -1, -1, 0, 0)$ is a cointegrating vector is rejected; thus, the PPP theory is rejected. But a vector of the form $(1, -1, -1, a, b)$ turns out to be a cointegrating vector. This can be viewed as describing a modified PPP that includes interest rates. This they interpret as saying that evidence for PPP can be found only if we allow for the interaction between the goods and asset markets.

A similar study by Juselius (1991) considers the exchange rate between Denmark and Germany for the period 1972:I to 1987:III. This study finds three cointegrating vectors and support for both PPP and UIP (because of pegged exchange rates within certain ranges until 1983). Liu (1992) studies the PPP for nine Latin American countries arguing that because of the high inflation rates, exchange rates for these countries versus the US dollar are more likely to be influenced by the relative price level. He tests a weaker version of the PPP theory, i.e., $e_{12} = \beta_0 + \beta_1 p_1 + \beta_2 p_2$ with β_1 negative and β_2 positive (but not necessarily unity in magnitude). He finds support for this theory for almost all the countries considered when the Johansen procedure is used. The interesting thing is that the estimates of the cointegrating vectors were quite different when they were estimated by OLS as compared to the Johansen ML procedure.

There has been some criticism of the Johansen procedure in studies with exchange rate data. The first criticism is that it assumes homoskedastic errors, which is rarely the case with exchange rates because of volatility clustering. Franses *et al.* (1992) discuss GARCH errors but only in relation to Engle–Granger tests of cointegration. The second problem with the Johansen procedure is that it assumes that the errors are orthogonal across equations. This is also an assumption that is not likely to be valid with exchange rate data. Abuaf and Jorion (1990) argue the importance of relaxing this assumption in their study of PPP (they do not use cointegration methods). For these reasons, Moore and Copeland (1995) suggest using the Phillips–Hansen method (in preference to the Johansen method) arguing that this method is valid under a wide range of distributional assumptions of the errors. They show that the conclusions are sensitive to the estimation method used. They also argue that in this case there is no obvious advantage to be gained by the fact that the Johansen procedure can accommodate more than one cointegrating vector.

There have also been several arguments given about using panel data methods to increase the power of cointegration tests. We discussed the problems of panel data unit root tests in chapter 4 (section 4.9). Pedroni (1995) has advanced similar arguments in favor of panel cointegration methods. Much of the criticism of panel unit root tests also carries over to panel cointegration tests.

A more fruitful approach to the PPP theory is to analyze the determinants of the deviations from PPP. Koedijk and Schotman (1989) characterize the pattern of deviations from PPP through the use of principal component analysis.

6.10 Conclusions

This chapter discusses tests for cointegration. It would be useful to distinguish between three types of cointegration tests:

- (i) Single equation tests – tests to determine whether a number of variables are cointegrated.
- (ii) Multiple equation tests – tests to determine the number of cointegrating vectors. Johansen’s trace test and maximum eigenvalue test fall in this category.
- (iii) Tests for significance of the coefficients of the estimated cointe-

grating vector. These are discussed in greater detail in chapter 10 on bootstrap methods.

Some Monte Carlo studies combine the different tests but it is important to separate out the results into these different categories because the results are comparable only for tests within a category.

We have discussed several tests in this chapter. Many are listed here for completeness. Although tests associated with the Johansen procedure are discussed very often, tests based on levels canonical correlation analysis (LCCA) are rarely discussed. These are discussed in section 6.6.

A number of cointegration tests have been discussed. Not all are useful. There have been many Monte Carlo studies comparing the size and power characteristics of these tests. Not many definitive conclusions emerge from the Monte Carlo studies. The conclusions that we can draw are summarized in section 6.5.2 and need not be repeated here. The Monte Carlo studies do not cover the ECM-based tests discussed in section 6.3.

An important issue that arises in connection with cointegration tests is the pre-testing problem. This has not received any attention until very recently and there is no satisfactory solution as yet. This is an important issue that needs further study.

References

- Abuaf, N. and P. Jorion (1990), "Purchasing Power Parity in the Long Run," *Journal of Finance*, 45, 157–174.
- Baillie, R.T. and T. Bollerslev (1989), "Common Stochastic Trends in a System of Exchange Rates," *Journal of Finance*, 44, 167–181.
- Banerjee, A. (1995), "Dynamic Specification and Testing for Unit Roots and Cointegration," in K. D. Hoover (eds.), *Macroeconometrics*, Kluwer, Boston, chapter 12.
- Banerjee, A., J.J. Donald, and R. Mestre (1996), "ECM Tests for Cointegration in a Single Equation Framework," Manuscript, Oxford University.
- Bewley, R. and D. Orden (1994), "Alternative Methods for Estimating Long-run Responses with Applications to Australian Import Demand," *Econometric Reviews*, 13, 179–204.
- Bewley, R. and M. Yang (1995), "Tests for Cointegration Based on

- Canonical Correlation Analysis," *Journal of American Statistical Association*, 90, 990–996.
- (1996), "On the Size and Power of Systems Tests for Cointegration," Manuscript, University of New South Wales, Sydney.
- Bossaerts, P. (1988), "Common Nonstationary Components of Asset Prices," *Journal of Economic Dynamics and Control*, 12, 347–364.
- Boswijk, H.P. (1992), "Cointegration, Identification, and Exogeneity," Ph.D. Dissertation, Tinbergen Institute, No. 37, Tinbergen Institute Research Series.
- (1994), "Testing for an Unstable Root in Conditional and Structural Error Correction Models," *Journal of Econometrics*, 63, 37–60.
- Boswijk, H.P. and P.H. Franses (1992), "Dynamic Specification and Cointegration," *Oxford Bulletin of Economics and Statistics*, 54, 369–381.
- Campbell, J.Y. and P. Perron (1991), "Pitfalls and Opportunities: What Macroeconomists Should Know About Unit Roots," in O.J. Blanchard and S. Fisher (eds.), *Macroeconomics Annual*, The MIT Press, Cambridge, MA, 141–201.
- Cheung, Y. and K. Lai (1993), "Finite Sample Sizes of Johansen's Likelihood Ratio Tests for Cointegration," *Oxford Bulletin of Economics and Statistics*, 55, 313–328.
- Cuthbertson, K. and M.P. Taylor (1990), "Money Demand, Expectations and the Forward Looking Model," *Journal of Policy Modeling*, 12, 289–315.
- Edison, H.J., J.E. Gagnon, and W.R. Melick (1997), "Understanding the Empirical Literature on Purchasing Power Parity: the Post-Bretton Woods Era," *Journal of International Money and Finance*, 16, 1–17.
- Eitrheim, O. (1992), "Inference in Small Cointegrated Systems: Some Monte Carlo Results," Presented at the Econometric Society Meeting in Brussels.
- Elliott, G. (1995a), "On the Robustness of Cointegration Methods When Regressors Almost Have Unit Roots," Discussion paper 95-18, Department of Economics, University of California at San Diego.
- (1995b), "Tests for the Correct Specification of Cointegrating Vectors and the Error Correction Model," Discussion paper 95-42, Department of Economics, University of California at San Diego.

- Elliott, G. and J.H. Stock (1994), "Inference in Time-Series Regression When the Order of Integration of a Regressor is Known," *Econometric Theory*, 10, 672–700.
- Engle, R. and C.W.J. Granger (1987), "Cointegration and error correction: representations, estimation and testing," *Econometrica*, 55, 252–276.
- Engle, R.F. and B.S. Yoo (1987), "Forecasting and Testing in Co-Integrated Systems," *Journal of Econometrics*, 35, 143–159.
- Franses, P.H., P. Kofman, and J. Moser (1992), "GARCH Effects on a Test of Cointegration," Working paper TI 93-76, Tinbergen Institute.
- Fuller, W. A. (1976), *Introduction to Statistical Time Series*, New York, Wiley.
- Gonzalo, J. and T.H. Lee (1995), "Pitfalls in Testing for Long-Run Relationships," Working paper No. 38, Department of Economics, Boston University.
- Gregory, A.W. (1994), "Testing for Cointegration in Linear Quadratic Models," *Journal of Business and Economic Statistics*, 12, 347–360.
- Grilli, V. and A. Beltratti (1989), "US Military Expenditures and the Dollar," *Economic Enquiry*, 27, 737–44.
- Hakkio, C.S. and M. Rush (1989), "Market Efficiency and Cointegration: An Application to the Sterling and Deutschmark Exchange Rates," *Journal of International Money and Finance*, 8, 75–88.
- (1991), "Cointegration: How Short is the Long Run," *Journal of International Money and Finance*, 9, 75–88.
- Hansen, B.E. (1990), "A Powerful Simple Test for Cointegration Using Cochrane-Orcutt," Working paper No. 230, University of Rochester.
- (1995), "Rethinking the Univariate Approach to Unit Root Testing," *Econometric Theory*, 11, 1148–1171.
- Harris, D. (1997), "Principal Components Analysis of Cointegrated Time Series," *Econometric Theory*, 13, 529–557.
- Harris, D. and B. Inder (1994), "A Test of the Null of Cointegration," in C. Hargreaves (ed.), *Non-Stationary Time Series Analysis and Cointegration*, Oxford University Press, Oxford, 133–152.
- Haug, A.A (1996), "Tests for Cointegration: A Monte Carlo Comparison," *Journal of Econometrics*, 71, 89–115.

- Ho, M.S. and B.E. Sørensen (1996), "Finding Cointegration Rank in High Dimensional Systems Using the Johansen Test: An Illustration Using Data Based Monte Carlo Simulation," *The Review of Economics and Statistics*, 78, 31–43.
- Hooker, M.A. (1993), "Testing for Cointegration: Power versus Frequency of Observation," *Economics Letters*, 41, 359–362.
- Horvath, M.T.K. and M.W. Watson (1995), "Testing for Cointegration When Some of the Cointegrating Vectors Are Known," *Econometric Theory*, 11, 984–1014.
- Hu, W. (1996), "Time Aggregation and Skip Sampling in Cointegration Tests," *Statistical Papers*, 37, 225–234.
- Johansen, S. (1988), "Statistical Analysis of Cointegration Vectors," *Journal of Economic Dynamics and Control*, 12, 231–254.
- (1991), "Estimation and Hypothesis Testing of Cointegration Vectors in Gaussian Vector Autoregressive Models," *Econometrica*, 59, 1551–1580.
- (1992), "Determination of Cointegration Rank in the Presence of a Linear Trend," *Oxford Bulletin of Economics and Statistics*, 54, 383–397.
- Johansen, S. and K. Juselius (1990), "Maximum Likelihood Estimation and Inference on Cointegration – with Applications to the Demand for Money," *Oxford Bulletin of Economics and Statistics*, 52, 169–210.
- (1992), "Testing Structural Hypotheses in a Multivariate Cointegration Analysis of the PPP and UIP for UK," *Journal of Econometrics*, 53, 211–244.
- Juselius, K. (1991), "Long-Run Relations in a Well Defined Statistical Model for the Data Generating Process: Cointegration Analysis of the PPP and UIP Relations between Denmark and Germany," in J. Gruber (ed.), *Econometric Decision Models: New Methods of modeling and Application*, Springer Verlag.
- Kim I.M. (1995), "Pretest Bias in a Cointegrated System," *The Korean Economic Review*, Sung Kyun Kwan University, 22, 115–140.
- Kleibergen, F. and H.K. van Dijk (1994), "Direct Cointegration Testing in Error Correction Models," *Journal of Econometrics*, 63, 61–103.
- Kleibergen, F., J.D. Urbain, and H.K. van Dijk (1995), "A Cointegration Study of Aggregate Imports Using Likelihood Based Testing Principles," Tinbergen Institute paper No. TI 95-169.

- Koedijk, K. and P. Schotman (1989), "Dominant Real Exchange Rate Movements," *Journal of International Money and Finance*, 8, 517–531.
- Kremers, J.J., N.R. Ericsson, and J.J. Dolado (1992), "The Power of Co-integration Tests," *Oxford Bulletin of Economics and Statistics*, 54, 325–348.
- Lahiri, K. and N. Mamingi (1995), "Testing for Cointegration: Power versus Frequency of Observation: Another View," *Economics Letters*, 49, 121–124.
- Layton, A.P. (1993), "A Suggested Reconciliation of Recent Cointegration Testing of Foreign Exchange Market Spot Rates," *Applied Financial Economics*, 3, 353–356.
- Leamer, E. (1991), "The Interplay of Theory and Data in the Study of International Trade," in M. Nerlove (ed.), *Issues in Contemporary Economics*, volume 2: Macroeconomics and Econometrics, Macmillan, IEA.
- Leybourne, S.J. and B.P.M. McCabe (1993), "A Simple Test for Cointegration," *Oxford Bulletin of Economics and Statistics*, 55, 97–103.
- Liu, P.C. (1992), "Purchasing Power Parity in the Latin American Countries: A Cointegration Analysis," *Weltwirtschaftliches Archiv*, 128, 662–680.
- MacDonald, R. and M.P. Taylor (1989), "Foreign Exchange Market Efficiency and Cointegration," *Economics Letters*, 29, 63–68.
- MacKinnon, J.G. (1991), "Critical Values for Co-Integration Tests," in R.F. Engle and C.W.J. Granger (eds.), *Long-Run Economic Relationships*, Oxford University Press, 267–276.
- (1994), "Approximate Asymptotic Distribution Functions for Unit Root and Cointegration Tests," *Journal of Business and Economic Statistics*, 12, 167–176.
- (1995), "Numerical Distribution Functions for Unit Root and Cointegration Tests," Discussion paper No. 918, Queens University.
- Maddala, G.S. (1991), "On the Use of Cointegration Tests," Paper presented at the Bank of Korea, Seoul, August, 1991.
- (1992), *Introduction to Econometrics*, 2nd ed., Macmillan, New York.
- Moore, M.J., and L.S. Copeland, (1995), "A Comparison of Johansen and Phillips–Hansen Cointegration Tests of Forward Market Efficiency: Baillie and Bollerslev Revisited," *Economics Letters*, 47, 131–135.

- Osterwald-Lenum, M. (1992), "A Note with Fractiles of the Asymptotic Distribution of the Maximum Likelihood Cointegration Rank Test Statistics: Four Cases," *Oxford Bulletin of Economics and Statistics*, 54, 461–72.
- Park, J.Y. (1990), "Testing for Unit Roots and Cointegration by Variable Addition," in T.B. Fomby and G.F. Rodes (eds.), *Advances in Econometrics*, vol. 8, JAI Press, 107–133.
- Park, J.Y. and B. Choi (1988), "A New Approach to Testing for a Unit Root," CAE Working paper No. 88-23, Cornell University.
- Park, J.Y. and P.C.B. Phillips (1988), "Statistical Inference in Regressions with Integrated Processes: Part I," *Econometric Theory*, 4, 468–497.
- Pedroni, P. (1995), "Panel Cointegration: Asymptotic and Finite Sample Properties of Pooled Time Series Tests with an Application to the PPP Hypothesis," Manuscript, Indiana University.
- Perron, P. and J.Y. Campbell (1993), "A Note on Johansen's Cointegration Procedure When Trends Are Present," *Empirical Economics*, 18, 777–789.
- Phillips, P.C.B. (1994), "Some Exact Distribution Theory for Maximum Likelihood Estimators of Cointegrating Coefficients in Error Correction Models," *Econometrica*, 62, 73–93.
- Phillips, P.C.B. and S. Ouliaris (1988), "Testing for Cointegration Using Principal Components Methods," *Journal of Economic Dynamics and Control*, 12, 205–230.
- (1990), "Asymptotic Properties of Residual Based Tests for Cointegration," *Econometrica*, 58, 165–193.
- Phillips, P.C.B. and P. Perron (1988), "Testing for a Unit Root in Time Series Regression," *Biometrika*, 75, 335–346.
- Podivinsky, J.M. (1990), "Testing Misspecified Cointegrating Relationships," Working Paper No. 19, Monash University, Australia.
- Reimers, H.E. (1992), "Comparisons of Tests for Multivariate Cointegration," *Statistical Papers*, 33, 335–359.
- Saikkonen, P., (1992), "Estimation and Testing of Cointegrated Systems by an Autoregressive Approximation," *Econometric Theory*, 8, 1–27.
- Sephton, P.S. and H.K. Larson (1991), "Tests of Exchange Market Efficiency: Fragile Evidence from Cointegration Tests," *Journal of International Money and Finance*, 10, 561–570.

- Shin, Y. (1994), "A Residual-Based Test of the Null of Cointegration Against the Alternative of No-Cointegration," *Econometric Theory*, 10, 91–115.
- Steigerwald, D.G. (1996), "Purchasing Power Parity, Unit Roots and Dynamic Structure," *Journal of Empirical Finance*, 2, 343–357.
- Stock, J.H. and M.W. Watson (1988), "Testing for Common Trends," *Journal of the American Statistical Association*, 83, 1097–1107.
- Toda, H.Y. (1994), "Finite Sample Properties of Likelihood Ratio Tests for Cointegrating Ranks When Linear Trends are Present," *Review of Economics and Statistics*, 76, 66–79.
- (1995), "Finite Sample Performance of Likelihood Ratio Tests for Cointegration Ranks in Vector Autoregression," *Econometric Theory*, 11, 1015–1032.
- Xiao, Z. and P.C.B. Phillips (1996), "Efficient Detrending in Cointegrating Regression," Manuscript, Cowles Foundation, Yale University.
- Yang, M. and R. Bewley (1996), "On Cointegration Tests for VAR Models with Drift," *Economics Letters*, 51, 45–50.
- Zivot, E. (1994), "Single Equation Conditional Error Correction Model Based Tests for Cointegration," Discussion paper 94-12, Department of Economics, University of Washington, Seattle.

7

Econometric modeling with integrated regressors

In the previous chapter we discussed estimation methods for cointegrated systems. In the models considered, the variables were all known to be $I(1)$. Also, in the Phillips triangular system, the number of cointegrating vectors is known. In the Johansen method, on the other hand, the number of cointegrating vectors is estimated (it is not assumed known).

In this chapter we will consider estimation of models with a mixture of $I(1)$ and $I(0)$ variables as well as problems of inference when there is uncertainty about the unit roots, that is, when we do not know whether a variable is $I(1)$ or $I(0)$. First we will deal with the case where we know the classification of the variables into the $I(1)$ and $I(0)$ categories. Next, we consider problems where this classification is unknown. Problems with $I(2)$ variables are treated in chapter 11.

The asymptotic theory relevant for the discussion of the problems discussed here is rather lengthy and hence we shall summarize only the main results. Details on the asymptotic theory can be found in the papers cited.

7.1 $I(1)$ regressors not cointegrated

For simplicity of exposition, we shall start with a model with one $I(1)$ and $I(0)$ regressor. The model is

$$y_t = \alpha + \beta x_t + \gamma z_t + \varepsilon_t \quad (7.1)$$

where $\varepsilon_t \sim I(0)$, $z_t \sim I(0)$, $x_t \sim I(1)$ and z_t and ε_t are uncorrelated. This model has been studied in Park and Phillips (1988, 1989) and Sims *et al.* (1990). The main results are as follows:

- (i) Note that $\gamma z_t + \varepsilon_t$ is $I(0)$. Hence β can be consistently estimated ignoring z_t . Let $\hat{\beta}$ be the estimator of β ignoring z_t in (7.1).
- (ii) Suppose we estimate (7.1) by OLS. Let $(\tilde{\beta}, \tilde{\gamma})$ be the estimators of β and γ . Then the distributions of $\tilde{\beta}$ and $\tilde{\gamma}$ are independent. This follows from the differences in the normalization factors and has been discussed in section 3.6.1 of chapter 3.
- (iii) The distribution of $\tilde{\beta}$ is of the same form as that of $\hat{\beta}$. However, it is not true that the asymptotic distributions of $T(\hat{\beta} - \beta)$ and $T(\tilde{\beta} - \beta)$ are the same because the errors are different.

The fact that the asymptotic distribution of the coefficients of the $I(1)$ and $I(0)$ regressors are independent and that the distribution theory for the coefficients of the $I(0)$ regressors is standard, can be used conveniently to determine the lag length in an autoregressive model. For example, suppose we consider the AR model

$$y_t = \beta_1 y_{t-1} + \beta_2 y_{t-2} + \beta_3 y_{t-3} + \varepsilon_t$$

where y_t is $I(1)$ and we are interested in testing the hypothesis $\beta_3 = 0$, i.e., the lag length is 2 rather than 3. We can write this equation as

$$y_t = \beta_1 y_{t-1} + (\beta_2 + \beta_3) y_{t-2} - \beta_3 \Delta y_{t-2} + \varepsilon_t$$

Now Δy_{t-2} is an $I(0)$ variable and hence the distribution of $\sqrt{T}(\hat{\beta}_3 - \beta_3)$ will follow the standard normal distribution theory. Hence hypothesis tests of the Wald type based on the asymptotic normal distribution can be used in testing the hypothesis $\beta_3 = 0$ by estimating a regression equation regressing y_t on y_{t-1} , y_{t-2} , and Δy_{t-2} . This sort of reformulation can be done for testing any one lag coefficient in a model with k lags.

7.2 $I(1)$ regressors cointegrated

We shall next consider the case where the $I(1)$ regressors are cointegrated. Consider again for simplicity

$$y = \alpha + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \gamma z_t + \varepsilon_t \quad (7.2)$$

As before, we assume z_t and ε_t are independent and $I(0)$. Suppose x_2 and x_3 are cointegrated, i.e., $x_2 + \delta x_3 = w$ is $I(0)$. We can write equation (7.2) as

$$y = \alpha + \beta_1 x_1 + (\beta_3 - \beta_2 \delta) x_3 + \beta_2 w + \gamma z_t + \varepsilon_t \quad (7.3)$$

It would appear that all the parameters in (7.3) can be consistently estimated. The regression of x_2 on x_3 gives an estimator $\hat{\delta}$ of δ which is

superconsistent and hence the estimators of the parameters in (7.3) with $\hat{\delta}$ substituted for δ would give consistent estimators of all the parameters.

However, as discussed in Wooldridge (1992) it is not true that all the parameters in (7.3) can be consistently estimated unless w is uncorrelated with ε_t . In general, the error term ε_t has to be uncorrelated with all stationary linear combinations of the regressors.

As for tests for cointegration when some regressors are cointegrated, note that we apply tests for CI using equation (7.3) rather than (7.2).

7.3 Unbalanced equations

An unbalanced equation is one in which the regressand is not of the same order of integration as the regressors or any linear combination of the regressors. A requirement in order to obtain a meaningful estimation with integrated variables is balance in the orders of integration of the variables on the left-hand side and right-hand side of the regression equation. Consider a regression where the dependent variable is $I(0)$ and the explanatory variables are $I(1)$. Pagan and Wickens (1989, p. 1002) argue that the disturbance will also be $I(1)$ if there is only one regressor, and that to achieve an $I(0)$ disturbance, there must be at least two $I(1)$ regressors.

Banerjee *et al.* (1993) study the features of regression among series with various orders of integration. Table 7.1 shows some of their results based on the model

$$y_t = \alpha + \beta x_t + u_t, \quad u_t \sim iin(0, 1)$$

The last column of table 7.1 indicates the probability of rejecting the null hypothesis of $\beta = 0$ at the conventional significance level. Case (ii) is spurious regression discussed in section 2.5 of chapter 2. But spurious regression can also arise in a regression of an $I(2)$ variable on another independent $I(2)$ variable as in case (iii) or regression of $I(2)$ on $I(1)$ and vice versa as in cases (vi) and (vii). When an $I(0)$ series is regressed on an $I(1)$ series as in case (iv), Banerjee *et al.* argue that the only way in which OLS can make the regression consistent and minimize the sum of squares is to drive the coefficient of the $I(1)$ variable to zero.

In the Monte Carlo simulations reported in Banerjee *et al.*, they obtain the histogram for the regression coefficient of an $I(2)$ series regressed on an $I(1)$ series, and show that the empirical distribution of this regression coefficient is long-tailed and peaked, i.e., distinctly nonnormal. They

Table 7.1. *Features of regressions among series with various orders of integration*

Case	y	x	$\Pr[t(\beta = 0) \geq 2]$
(i)	I(0)	I(0)	0.0493
(ii)	I(1)	I(1)	0.7570
(iii)	I(2)	I(2)	0.9406
(iv)	I(0)	I(1)	0.0458
(v)	I(1)	I(0)	0.0486
(vi)	I(2)	I(1)	0.8530
(vii)	I(1)	I(2)	0.8444

Source: Banerjee *et al.* (1993, table 3.1, p. 79).

also report the results of several other Monte Carlo studies on unbalanced equations (pp. 164–8). Marmol (1996) derives analytically the distributions of the regression coefficients, the Durbin–Watson statistic and the correlation coefficient. He shows that the Monte Carlo results of Banerjee *et al.* (1993) agree with his analyzed results. Haldrup (1994) also contains some asymptotic results on regression with I(2) and I(1) variables.

Should one estimate unbalanced regressions? Of course not, if it can be avoided. But if it has to be done, one has to be careful in their interpretation and use appropriate critical values. Banerjee *et al.* (1993, p. 166) say:

The fact that a regression is unbalanced may not be a matter of concern, for example, ADF statistics are computed from models that are unbalanced. They are nonetheless valid tools of inference as long as the correct critical values are used.

Perhaps Banerjee *et al.* are more optimistic than others, regarding unbalanced equations.

7.4 Lagged dependent variables: the ARDL model

Models with lagged dependent variables and serially correlated errors have a long history in econometrics. The problem of biases in the OLS estimation of lagged dependent variable models with serially correlated errors have been studied extensively. The extension of these models to the case of I(1) regressors can be analyzed within the framework of regressions with cointegrated regressors (discussed in the previous

sections) for which asymptotic theory has been developed in Park and Phillips (1989) in a general framework. However, from the practical point of view it would be interesting to look at the expressions for the asymptotic biases in the least squares estimates.

Maekawa *et al.* (1996) investigate the asymptotic properties of the OLS estimator, the DW test statistic, and Durbin's h -test statistic in the following model

$$\begin{aligned} y_t &= \alpha y_{t-1} + \beta z_t + u_t, \quad |\alpha| < 1 \\ u_t &= \rho u_{t-1} + v_t, \quad |\rho| < 1 \\ z_t &= z_{t-1} + \varepsilon_t \end{aligned}$$

where $v_t \sim IN(0, \sigma_v^2)$, $\varepsilon_t \sim IN(0, \sigma_\varepsilon^2)$ and they are independent. This model was considered earlier by Maddala and Rao (1973) with stationary z_t , i.e., z_t was specified by

$$z_t = \lambda z_{t-1} + \varepsilon_t, \quad |\lambda| < 1$$

Thus the expressions for the asymptotic biases derived there could be expected to hold with $\lambda = 1$. Maekawa *et al.* find that this is indeed the case. Define

$$P = \frac{\rho \sigma_v^2}{(1 - \rho^2)(1 - \alpha \rho)}$$

and

$$Q = \frac{1}{(1 - \alpha^2)} \left[\frac{(1 + \alpha \rho) \sigma_v^2}{(1 - \rho^2)(1 - \alpha \rho)} + \frac{\beta \sigma_\varepsilon^2}{(1 - \alpha)^2} \right]$$

Then they show that

$$\begin{aligned} \text{plim } \hat{\alpha} &= \alpha + \gamma \\ \text{plim } \hat{\beta} &= \beta + \delta \\ \text{plim } \hat{\rho} &= \rho - \gamma \\ \text{plim DW} &= 2(1 - \rho + \gamma) \end{aligned}$$

where

$$\gamma = \frac{P}{Q} \quad \text{and} \quad \delta = -\frac{\beta}{1 - \alpha} \gamma$$

These expressions agree with those derived in Maddala and Rao (1973) substituting $\lambda = 1$ ($\theta = 1$ in their notation). If $\rho = 0$, then $\hat{\alpha}$, $\hat{\beta}$, and $\hat{\rho}$ are consistent. Also, $DW \rightarrow 2$. Under the assumption $\rho = 0$ the distributions of $\sqrt{T}(\hat{\alpha} - \alpha)$ and $\sqrt{T}(\hat{\beta} - \beta)$ are asymptotically normal.

Maekawa *et al.* show that in this model, if $\rho \neq 0$ then Durbin's two-step procedure produces consistent estimates. An alternative procedure is to use the Bewley and Wickens-Breusch procedure discussed in chapter 2.

7.5 Uncertain unit roots

Throughout the preceding discussion we assumed that we knew whether a variable belonged to the $I(0)$ or $I(1)$ category. However, in practice there is some uncertainty on this issue. Elliott and Stock (1994) examine this issue. Let x_t be a variable about which there is some uncertainty whether it is $I(1)$. We are ultimately interested in the coefficient of x_{t-1} in a regression of y_t on x_{t-1} . The motivation for this problem is that in the finance literature, it is often considered with y_t denoting excess stock returns and x_t denoting dividend yields. A common empirical observation is that lagged dividend yield is a significant explanatory variable for excess stock returns. However, there are some questions about the validity of this inference if x_t is $I(1)$.

The model that Elliott and Stock consider is

$$\begin{aligned}x_t &= \mu_1 + \rho x_{t-1} + \eta_{1t} \\ y_t &= \mu_2 + \gamma x_{t-1} + \eta_{2t}\end{aligned}$$

Let $\delta = \text{corr}(\eta_{1t}, \eta_{2t})$. This is a measure of endogeneity of the regressor x_{t-1} . There are two approaches to follow:

- (i) Just proceed with testing $\gamma = 0$ ignoring the unit root problem.
- (ii) Apply a pre-test for a unit root in x_t . If the unit root null is rejected, then the $I(0)$ normal distribution is used. If the unit root null is not rejected, then the $I(1)$ distribution (under endogeneity) is used.

With procedure (i), as theory predicts, there were no appreciable size distortions when $\delta = 0$, even at high values of ρ . However, when δ was high and ρ was high, there were substantial size distortions. For example, when $\delta = -0.9$, $\rho = 0.95$ and $T=50$, they found rejection rates under 1 percent in the left tail and 22 percent in the right tail when using a standard normal (5 percent, 95 percent) critical values.

As for procedure (ii), Elliott and Stock consider a first-stage one-sided DF t -test to test for unit root in x_t . At the second stage they used an equal tailed 10 percent nominal significance level, i.e., (5 percent, 95

percent) quantiles of the $I(0)$ and $I(1)$ distributions depending on the outcome of the DF t -test. With $\delta = 0$ both distributions are normal so that the same critical values (± 1.645) were used whether or not the DF t -test rejected the unit root null. In this case, as expected, there were no size distortions, e.g., for $\delta = -0.9$ and $\rho = 0.9$, the empirical size was 30 percent even for $T = 100$. Elliott and Stock suggest a Bayesian solution to this pre-testing problem based on a Bayesian mixture approximation. We shall discuss this in chapter 9 when we talk of the Bayesian approach to unit roots.

What this suggests is that both the procedures (i) and (ii) result in substantial size distortions, when δ is high. The fact that the true significance levels are higher than the nominal levels means that if as was observed γ was found to be significant at the 5 percent level, it is not significant at the 5 percent level but at a much higher level (say even 30 percent). Thus the evidence on the predictive ability of lagged dividends in predicting excess stock returns is weaker than the values suggested by the significance of the usual t -values.

Cavanagh, Elliott, and Stock (1995) investigate further the magnitudes of size distortions in the second-stage test of the hypothesis $\gamma = 0$ for different values of c in the local to unity model $\rho = 1 + c/T$. They first derive the asymptotic sizes of the pre-test procedure as a function of c and δ . These asymptotic sizes for different values of c and δ are tabulated in table 1 of their paper. For δ small, say $\delta \leq 0.4$ in the demeaned case and $\delta \leq 0.3$ in the detrended case, the asymptotic size distortions of the two-step procedure are small and are possibly negligible for empirical work. However for $\delta = 0.9$ in the detrended case, the rejection rate for the two-stage test with nominal size of 5 percent is 37 percent when $c = 20$ and the maximal rejection rate (empirical size) over all values of c is 64 percent. Cavanagh *et al.* argue that these size distortions also persist when alternatives to the DF t -test are used. The alternatives they consider are those suggested by Stock and by Phillips and Ploberger for choosing between $I(0)$ and $I(1)$ specifications, which depend on Bayesian methods. These are discussed in chapter 9.

Cavanagh *et al.* discuss alternatives to the two-step procedure that exhibit large size distortions. They suggest several bounds tests which provide conservative critical values and confidence intervals. Their summary finding is that the bounds procedures correct the size distortions but at a cost of substantial power loss. Of the different bounds tests they considered the *sup-bounds procedure* was the best. This procedure is as follows: denote by t_γ the t -test statistic for testing $\gamma = 0$. For

ease of notation we shall denote it by t . Let $d_{t,c,\alpha}$ be the asymptotic critical value for the test $\gamma = 0$ at significance level α , for the value c in $\rho = 1 + c/T$. Let

$$d_{\alpha}(L) = \inf_c d_{t,c,\alpha}$$

and

$$d_{\alpha}(M) = \sup_c d_{t,c,\alpha}$$

(L denotes lower limit and M denotes maximum). If $t \notin (d_{\alpha}(L), d_{1-\alpha}(M))$ a test of $\gamma = 0$ with asymptotic level 2α will reject the null regardless of the value c . Values of t within the conservative acceptance region $(d_{\alpha}(L), d_{1-\alpha}(M))$ but outside the acceptance region $(d_{t,\alpha}(L), d_{t,1-\alpha}(M))$ constitute an indetermined region. Cavanagh *et al.* tabulate (in table 2 of their paper) the upper and lower bounds $(d_{\alpha}(L), d_{1-\alpha}(M))$ for selected values of ρ and asymptotic significance levels $2\alpha = 5$ percent and 10 percent.

7.6 Uncertain unit roots and cointegration

In sections 7.1 and 7.2 we discussed models where there was certainty about the cointegrating relationships among the $I(1)$ variables. Kitamura and Phillips (1997) develop instrumental variable methods that are applicable where there is uncertainty about the order of integration in the variables, as well as uncertainty regarding the number and location of the cointegrating relationships. They suggest FM-GMM and FM-GIVE estimators. These are generalizations of the FM-IV estimator to the GMM (generalized methods of moments) and the GIVE (generalized instrumental variable estimator) used with stationary regressors in econometric work. The FM-OLS and FM-IV estimators in the Phillips and Hansen paper (see chapter 5) were developed for the case where all the variables were $I(1)$. Phillips (1991, 1995) proved that the FM procedure can be applied to models with cointegrated regressors, models with $I(1)$ and $I(0)$ regressors, and even to models with only stationary regressors, without losing the good asymptotic properties of the FM method. This idea is utilized in the development of the FM-GMM and FM-GIVE methods. We shall not reproduce the detailed expressions for these methods here. A practical guide to the formulas for empirical implementation is described in detail in the paper by Kitamura and Phillips. The reason why instrumental variables are introduced is to allow for endogeneity of the regressors (especially the stationary regressors).

Kitamura and Phillips (1995) present results of a simulation study to show the practical usefulness of the FM-GMM and FM-GIVE estimators. The asymptotic distributions of the estimators for the coefficients of the stationary regressors are standard normal distributions, and the asymptotic distributions of the estimators of the coefficients of the non-stationary regressors are mixed normal distributions.

However, although in the estimation stage one can ignore the issues of unit roots and cointegration relationships among the regressors, when it comes to hypothesis testing we still need to know which asymptotic distributions to use – the standard normal or the mixed normal. The uncertainty about the unit roots becomes important again at the hypothesis testing stage. This problem is discussed in Kitamura (1994), who develop a test statistic which has a standard normal distribution no matter where unit roots are located. The test statistic is always asymptotically χ^2 distributed with degrees of freedom equal to the number of restrictions. Kitamura also presents results of a simulation study supporting the theoretical results.

The bounds procedures by Cavanagh *et al.*, discussed in the previous section, were also designed to make (conservative) inferences in the presence of uncertainty about unit roots. However, these authors noted substantial loss of power by using these bounds tests. In the case of the vector autoregression model, Phillips (1995) also develops conservative bounds procedures. He considers the FM-VAR estimation method and shows that its optimal properties do not depend on the number and location of unit roots or cointegration relationships among the $I(1)$ variables. He shows that when testing hypotheses using the Wald statistics, if q is the number of restrictions tested then the χ_q^2 distribution is an upper bounds variable and therefore the usual χ^2 critical value can be used to construct tests that have conservative size.

The tests that Kitamura constructs are not based on bounds procedures. His test involves the difference of the restricted minimized value of the GMM objective function and the unrestricted minimized value of the GMM objective function using the score vectors obtained from the FM-GMM procedure as weights. It is thus a likelihood ratio type statistic but not exactly. (It depends on the difference rather than the ratio of two quantities.) The details of Kitamura's test are very intricate and will not be discussed here. Readers are referred to Kitamura's paper. Kitamura also has a computer program to calculate the test statistic.

One other paper that considers the problem of inference under uncertainty about unit roots and cointegration relationships is the paper

by Toda and Yamamoto (1995). They consider the problem of hypothesis tests involving restrictions on the parameters of a VAR when there is uncertainty about the number of unit roots and cointegration relationships. They suggest estimating the VAR model in levels, choosing the lag length k by using the usual procedures for determining the lag length. Having determined the lag length k , they suggest estimating a $(k + d_{max})$ th order VAR where d_{max} is the maximal order of integration that we suspect might occur in the process. The coefficient matrices of the last d_{max} lagged vectors in the model are ignored and we can test linear or nonlinear restrictions on the first k coefficient matrices using standard asymptotic theory.

We, thus, have different procedures for making inference under uncertainty about unit roots and cointegrating relationships in the context of a VAR model. Practical applications of these procedures and comparative studies of the different procedures will be forthcoming in the near future.

7.7 Summary and conclusions

This chapter deals with problems in regressions with I(1) and I(0) variables, and unbalanced equations. We next discuss the role of lagged dependent variables in a regression with I(1) regressors.

We also discuss the important problems of uncertain unit roots and estimation and testing procedures under uncertainty about unit roots and cointegration. The usual procedure is to apply a unit root test and if the null is not rejected at the conventional 1 percent or 5 percent significance levels, then to assume (with certainty) that a unit root exists. The uncertainty about unit roots is completely ignored. Thus, discussion of the consequences of uncertainty about unit roots and cointegration is very important

References

- Banerjee, A., J.J. Dolado, J.W. Galbraith, and D.F. Hendry (1993), *Cointegration, Error Correction, and the Econometric Analysis of Nonstationary Data*, Oxford University Press, Oxford.
- Cavanagh, C.L., G. Elliott, and J.H. Stock (1995), "Inference in Models with Nearly Nonstationary Regressors," *Econometric Theory*, 11, 1131–1147.

- Elliott, G. and J.H. Stock (1994), "Inference in Time-Series Regression When the Order of Integration of a Regressor is Known," *Econometric Theory*, 10, 672–700.
- Haldrup, N. (1994), "The Asymptotics of Single-Equation Cointegration Regression with I(1) and I(2) Variables," *Journal of Econometrics*, 63, 153–181.
- Kitamura, Y. (1994), "Hypothesis Testing in Models with Possibly Non-Stationary Processes," Manuscript, University of Minnesota.
- Kitamura, Y. and P.C.B. Phillips (1995), "Efficient IV Estimation in Non-Stationary Regression: An Overview and Simulation Study," *Econometric Theory*, 11, 1095–1130.
- (1997), "Fully Modified IV, GIVE, and GMM Estimation with Possibly Non-Stationary Regressors and Instruments," *Journal of Econometrics*, 80, 85–124.
- Maddala, G.S. and A.S. Rao (1973), "Tests for Serial Correlation in Regression Models with Lagged Dependent Variables and Serially Correlated Errors," *Econometrica*, 47, 761–774.
- Maekawa, K., T. Yamamoto, Y. Takanehi, and M. Katanaka (1996), "Estimation in Dynamic Regression with an Integrated Process," *Journal of Statistical Planning and Inference*, 49, 279–303.
- Marmol, F. (1996), "Nonsense Regressions Between Integrated Processes of Different Orders," *Oxford Bulletin of Economics and Statistics*, 58, 525–536.
- Pagan, A.R. and M.R. Wickens (1989), "A Survey of Some Recent Econometric Methods," *The Economic Journal*, 99, 962–1025.
- Park, J.Y. and P.C.B. Phillips (1988), "Statistical Inference in Regressions with Integrated Processes: Part I," *Econometric Theory*, 4, 468–497.
- (1989), "Statistical Inference in Regressions with Integrated Processes: Part II," *Econometric Theory*, 5, 95–131.
- Phillips, P.C.B. (1991), "Time Series Regression with Cointegrated Regressors," Yale University.
- (1995), "Fully Modified Least Squares and Vector Autoregression," *Econometrica*, 63, 1023–1078.
- Sims, C.A., J.H. Stock, and M.W. Watson (1990), "Inference in Linear Time Series Models with Some Unit Roots," *Econometrica*, 58, 113–144.

- Toda, H.Y. and T. Yamamoto (1995), "Statistical Inference in Vector Autoregressions with Possibly Integrated Processes," *Journal of Econometrics*, 66, 225–250.
- Wooldridge, J.M. (1992), "Notes on Regression with Difference-Stationary Data," Manuscript, Michigan State University.

Part III

Extensions of the basic model

This part consists of five chapters that consider various extensions of the basic models discussed in part II.

Chapter 8 discusses the Bayesian approach to unit roots and cointegration. In chapters 4 and 7 we mentioned the problems created by uncertain unit roots. The Bayesian approach where the posterior probabilities of a unit root and stationarity are taken into account, is a natural solution to this problem. The chapter discusses the different priors and their merits and defects in the analysis of unit roots and cointegration. It also discusses the Bayesian model selection approach.

Chapter 9 is on fractional unit roots and fractional cointegration. These models are alternative ways of modeling persistence in economic time series. The chapter discusses different estimation methods and the empirical relevance of fractional integration.

Chapter 10 is on the bootstrap approach. Almost all the literature on unit roots and cointegration is asymptotic. However, it has also been found that the estimators suggested are biased, and the tests exhibit substantial size distortions. The bootstrap method helps in correcting these two problems. This chapter discusses the different methods of bootstrap data generation and bootstrap-based tests.

Chapter 11 is on cointegrated system with $I(2)$ variables. It has been observed that some variables like prices are more appropriately characterized as $I(2)$ processes rather than $I(1)$ processes. This chapter discusses problems of analysis with $I(2)$ variables and mixtures of $I(1)$ and $I(2)$ variables.

Chapter 12 is on seasonal unit roots and seasonal cointegration. Most economic time series at the quarterly and monthly level exhibit seasonal patterns. Thus it is important to extend the unit roots and cointegration concepts discussed in part II to seasonal data. It is also true that seasonal

patterns are not constant over time. In this case one considers the periodic models and extends the concepts of unit roots and cointegration to periodic models. The chapter also discusses an alternative model for changing seasonal pattern – the evolving seasonal model.

8

The Bayesian analysis of stochastic trends

The recent intensive research in unit root econometrics has focused on two implications. One is the economic implication about the effect of shocks on the economy. In models with unit roots, shocks have persistent effects that last forever, while in the case of traditional trend-stationary models, shocks only have a temporary effect. The other, statistical implication is that the presence of unit roots in the data greatly complicates statistical inference. As we have discussed in the previous chapters, the OLS estimators and the corresponding statistics have nonstandard asymptotic distributions under the presence of unit roots, while they have the standard distributions without unit roots.

The Bayesian approach has cast considerable doubt on the statistical implications of the unit root in economic time series. Sims (1988) argued that because the asymptotic distribution theory changes discontinuously between the stationary and the unit root cases, classical hypothesis testing based on asymptotic theory cannot deliver reasonable procedures for inference based on the discontinuous asymptotic theory. He argued that the simple *flat prior* Bayesian theory is both a more convenient and a logically sounder starting place for inference than classical hypothesis testing. In reply to Sims' criticism, Phillips (1991) argued that the reason for the discrepancy between the results obtained by Bayesian and classical methods on the existence of unit roots in US macroeconomic time series is due to the use of the flat prior in the Bayesian analysis and that the flat prior is not a satisfactory representation of uninformativeness. He suggested an *ignorance prior* (which is Jeffreys' prior), which showed that most US macroeconomic time series have a unit root under the Bayesian analysis as well. The special issue of the *Journal of Applied Econometrics* (1991, volume 6, number 4) is devoted entirely to this debate. There have subsequently been other special issues of

journals devoted to this topic. (See the issue on Bayesian inference in *Econometric Theory*, 1994, volume 10 and also *Journal of Econometrics*, 1995, volume 69, number 1.)

In this chapter we shall begin with a review of Bayesian methods for time series analysis. After reviewing the main points of the debate in 1991, we consider the recent developments in Bayesian inference on unit roots and cointegration as well as on the model selection approach.

8.1 Introduction to Bayesian inference

Most researchers use classical methods not because these methods are superior to Bayesian methods, but because most of the econometric literature is in the classical tradition and because, the classical methods were the first ones (in many cases) and the only ones they were taught.

The basis of all Bayesian inference is Bayes theorem: posterior probability varies with prior probability times likelihood. It is extended in the context of distributions as: posterior distribution $p(\theta|y)$ varies with prior distribution $\pi(\theta)$ times the likelihood function $L(y|\theta)$

$$p(\theta|y) \propto \pi(\theta)L(y|\theta)$$

Bayesian methods take the data as given, but regard the true parameter as random, while classical methods view the true parameter of interest as unknown and fixed, and study the behavior of its estimator in repeated samples. All Bayesian inference is based on the given sample and the posterior distribution. If many parameters are involved, one has to take the joint posterior distribution of all these parameters and integrate out all but the one of interest to get the marginal distribution of this parameter. Then inferences about this parameter can be made from this marginal distribution.

Given a particular model, the sample information is given in the likelihood function $L(y|\theta)$. In Bayesian methods we have an additional ingredient: the prior distribution $\pi(\theta)$. There can be various specifications of the prior distribution about θ , while the form of the likelihood function depends on the probability distribution underlying the data. In practice, however, only a few candidates are actually useful and used. The most commonly used forms of prior distribution are a *conjugate prior* and a *diffuse prior*. A conjugate prior is a prior that when combined with the likelihood function yields a posterior distribution that has the same functional form as the prior. The advantage of this prior is that every time a new sample is observed, the revision of opinions

about θ can be performed by the same analytical procedure. Raiffa and Schlaifer (1961) contains a detailed discussion of conjugate priors.

When there is not any prior notion about θ , what is suggested is a diffuse or noninformative prior. When we know nothing *a priori* about the parameters of the model, we let the variances of the prior distribution of θ increase without limit. The prior distribution is uniform over an unspecified range and is given by

$$\pi(\theta) \propto \text{a constant}$$

This prior is also called an *improper prior* or a *flat prior* because it does not integrate to 1.

Jeffreys (1961) suggested that for location parameters $p(\theta)$ be taken proportional to a constant and for scale parameters, that the prior be taken as proportional to their inverses. For instance, for a sample from a normal population $N(\mu, \sigma^2)$ the diffuse prior is

$$\pi(\mu, \sigma) \propto \frac{1}{\sigma} \quad (8.1)$$

In essence it is $\log \sigma$ that has a uniform distribution. This prior is invariant to transformations of the form $\phi = \sigma^n$, since $d\sigma = n\sigma^{n-1}$ and thus

$$\frac{d\phi}{\phi} \propto \frac{d\sigma}{\sigma}$$

This invariance is important because some parameterize the model in terms of the standard deviation and others in terms of the variance σ^2 .

Jeffreys also suggested a general invariant prior in a multiparameter case. He suggests (Jeffreys, 1961, p. 179) that the pdf for the parameter vector be taken as

$$p(\theta) \propto |I(\theta)|^{1/2} \quad (8.2)$$

where

$$I(\theta) = E \left[-\frac{\partial^2 \log L}{\partial \theta \partial \theta'} \right]$$

is Fisher's information matrix for θ . In this case the prior pdf will be invariant in the sense that if we parameterize the model in terms of $\eta = F(\theta)$ with a one to one correspondence between η and θ , and take our prior for η as

$$\pi(\eta) \propto |I(\eta)|^{-1/2}$$

then the posterior distribution of η will be consistent with the posterior

distribution of θ starting with the prior (8.2). We shall omit the proof and other properties of invariant prior distributions here (see Zellner, 1971, pp. 48–50).

Often the results obtained by Bayesian methods with the use of diffuse priors coincide with the results obtained by the classical procedures (Jeffreys, 1961; Lindley, 1965; Zellner, 1971). However, they differ sharply in the case of nonstationary time series, which is the source of the debate between the two groups. The difference between the two groups is closely related to the concept of a noninformative prior in the analysis of time series, i.e., which prior is noninformative in the analysis of time series. Sims (1988) used a flat prior as a noninformative prior, as did Zellner (1971, pp. 186–188). On the other hand, Phillips (1991) argued that a flat prior in the analysis of time series is informative and proposed an ignorance prior or Jeffreys' invariant prior as a noninformative prior for time series models. The basic argument of Phillips is that the autoregressive parameter ρ in the equation $y_t = \rho y_{t-1} + u_t$ is not on the same footing as β in the regression model $y_t = \beta x_t + u_t$ because the sample is more informative about ρ in the different ranges. When $|\rho|$ is large, the data would be more informative about ρ . Thus in treating all values of ρ as equally likely, the flat prior unwittingly carries information that downweights large values of ρ . Flat priors, instead of being noninformative are in fact informative by effectively downplaying the possibility of unit root and explosive alternatives.

8.2 The posterior distribution of an autoregressive parameter

Consider the simple AR(1) model

$$y_t = \rho y_{t-1} + \varepsilon_t$$

Conditioning on the initial value y_0 , the Gaussian likelihood follows the density

$$L(y|\rho, \sigma, y_0) = (2\pi)^{-T/2} \sigma^{-T} \exp \left(-\frac{\sum_{t=1}^T (y_t - \rho y_{t-1})^2}{2\sigma^2} \right)$$

Assume a flat prior for (ρ, σ)

$$\pi(\rho, \sigma) \propto \frac{1}{\sigma}, \quad -1 < \rho < 1, \quad \sigma > 0$$

The joint posterior distribution for (ρ, σ) is then given by

$$p(\rho, \sigma | y, y_0) \propto \sigma^{-T-1} \exp \left(-\frac{\sum_{t=1}^T (y_t - \rho y_{t-1})^2}{2\sigma^2} \right)$$

Let $\hat{\rho}$ be the OLS estimator of ρ so that $\hat{\rho} = \sum y_t y_{t-1} / \sum y_{t-1}^2$. Then we can write

$$\sum (y_t - \rho y_{t-1})^2 = R + (\rho - \hat{\rho})^2 Q$$

where $Q = \sum y_{t-1}^2$ and $R = \sum \hat{\varepsilon}_t^2$ is the residual sum of squares with $\hat{\varepsilon}_t = y_t - \hat{\rho} y_{t-1}$. The joint posterior distribution can be written as

$$p(\rho, \sigma | y, y_0) \propto \sigma^{-T-1} \exp \left[-\frac{R + (\rho - \hat{\rho})^2 Q}{2\sigma^2} \right]$$

The marginal posteriors are obtained after integrating out the other parameters and are given by

$$p(\rho, | y, y_0) \propto (R + (\rho - \hat{\rho})^2 Q)^{-T/2}$$

$$p(\sigma | y, y_0) \propto \sigma^{-T} \exp \left(-\frac{R}{2\sigma^2} \right)$$

The marginal posterior distribution for ρ is a univariate t -distribution, ρ is symmetrically distributed about the OLS estimate, and the variance of ρ is $R/(T-3)Q$, which decreases as Q increases. The marginal posterior distribution for σ is an inverted gamma-2 distribution (Raiffa and Schlaifer, 1961; Zellner, 1971).

This framework has been used widely in the analysis of time series. Thornber (1967) and Zellner (1971) used this framework and emphasized its applicability for stationary and nonstationary cases. Geweke (1988) used the same approach in a cross-country applied study but used a restricted domain in addition to the flat prior. Sims (1988) and Sims and Uhlig (1991) also use this framework, although in the latter paper the model is even simpler because σ is assumed to be known for computational convenience. Schotman and van Dijk (1991b) employ a similar approach in studying real exchange rate data. They modify a flat prior so that ρ has a flat prior in a proper subset of the stationary interval and they assign a discrete prior probability mass to $\rho = 1$ (values of ρ in the explosive range being excluded).

The Bayesian methods focus on the posterior distribution of the true parameter by taking the data and estimator as given, while the classical methods concentrate on the behavior of the estimators in repeated samples viewing the true parameter of interest as unknown and fixed. Thus,

under the presence of unit roots, while the classical methods are based on the asymmetric nonstandard distribution, for example the Dickey–Fuller distribution (see chapter 3) of the OLS estimators of autoregressive parameters, the Bayesian methods use the symmetric standard (posterior) distribution of true autoregressive parameters. The difference between the two approaches is well described by comparing the two distributions, $p(\rho|\hat{\rho} = 1)$ and $p(\hat{\rho}|\rho = 1)$ (Sims and Uhlig, 1991).

Sims and Uhlig (1991) compute the joint posterior distribution for ρ and $\hat{\rho}$ under a flat prior by Monte Carlo methods. For computational simplicity they assume that $\sigma^2 = 1$ and known. They show the asymmetric sampling distribution of $\hat{\rho}$ and the symmetric posterior distribution of ρ . Because of the asymmetry of the distribution of $\hat{\rho}$, Sims and Uhlig argue that the classical p -values, which the classical unit root tests rely on, are quite misleading. Suppose we observed $\hat{\rho} = 1$ and tried comparing the p -values of the null hypotheses $\rho = 0.98$ and $\rho = 1.02$ by classical procedures. Sims and Uhlig compute the p -values by computing the area under the curve to the right of the observed $\hat{\rho}$. They show that the p -value for $\rho = 0.98$ given an observed $\hat{\rho} = 1$ is 0.033, while the p -value for $\rho = 1.02$ given $\hat{\rho} = 1.02$ is 0.245. Thus we can reject $H_0 : \rho = 0.98$ at the 5 percent level, while easily accepting $H_0 : \rho = 1.02$. They conclude that the classical methods based on the asymmetric distribution of the OLS estimator $\hat{\rho}$ can mislead, giving too much credence to large ρ values and the simple flat-prior Bayesian method is a more convenient and a logically sounder starting place for inference than classical hypothesis testing for unit roots.

8.3 Bayesian inference on the Nelson–Plosser data

The above Bayesian methods have been applied to the Nelson and Plosser data by DeJong and Whiteman (1991) and Phillips (1991) and the results are strikingly different from those obtained by Nelson and Plosser (1982).

We shall begin to extend the above analysis to a model with a linear deterministic trend and a richer dynamic structure such as

$$y_t = \alpha + \delta t + \Phi(L)y_t + \varepsilon_t, \quad \varepsilon_t \sim iin(0, \sigma^2)$$

where $\Phi(L) = \sum_{i=1}^k \phi_i L^i$. This formulation includes the empirical specifications used in Nelson and Plosser (1982), where $k \leq 6$, and the model used by DeJong and Whiteman (1991), where $k = 3$. It is convenient to

employ the following alternative parameterization

$$y_t = \alpha + \delta t + \rho y_{t-1} + \sum_{i=1}^{k-1} \beta_i \Delta y_{t-i} + \varepsilon_t$$

where $\rho = \sum_{i=1}^k \phi_i$. Let V be the matrix of observations on $(1, t, \Delta y_{t-1}, \dots, \Delta y_{t-k+1})$ and let $\gamma = (\alpha, \delta, \beta_1, \dots, \beta_{k-1})$ be the corresponding vector of parameters. Then the joint posterior density for (ρ, σ, γ) corresponding to the flat prior $\pi(\rho, \gamma, \sigma) \propto 1/\sigma$ is

$$p(\rho, \gamma, \sigma | y, y_0) \propto \sigma^{-T-1} \times \exp \left(-\frac{\sum_{t=1}^T (y_t - \alpha - \delta t - \rho y_{t-1} - \sum_{i=1}^{k-1} \beta_i \Delta y_{t-i})^2}{2\sigma^2} \right)$$

Integrating out both γ and σ directly from the joint posterior distribution density yields the marginal posterior density for ρ

$$p(\rho | y, y_0) \propto (R_V + (\rho - \hat{\rho})^2 Q_V)^{-(T-k-1)/2}$$

where $R_V = \sum_{t=1}^T \hat{\varepsilon}_t^2$ is the residual sum of squares with

$$\hat{\varepsilon}_t = y_t - \hat{\alpha} - \hat{\delta}t - \hat{\rho}y_{t-1} - \sum_{i=1}^{k-1} \hat{\beta}_i \Delta y_{t-i}$$

$$Q_V = y'_{-1}(I - V(V'V)^{-1}V')y_{-1},$$

and y_{-1} represent the observation vector of (y_{t-1}) .

DeJong and Whiteman (1991) applied the above procedure with $k = 3$ to the same data used by Nelson and Plosser (1982). They focused on the posterior distribution of the dominant root, which is the largest autoregressive root, defined such that

$$\Lambda \equiv \max_j |\lambda_j|$$

where

$$\lambda(z) \equiv 1 - \Phi(z) = \prod_{j=1}^3 (1 - \lambda_j z)$$

In this setting, trend stationarity (TS) corresponds to $\Lambda < 1$, and difference stationarity (DS) is the special case $\Lambda = 1$. They employ a truncated flat prior on the autoregressive coefficients such that

$$\pi(\theta) = g(\theta) i_A(\theta)$$

Table 8.1. *Posterior probabilities for the Nelson–Plosser data*

Series	T	$p(\Lambda \geq 0.975)$	$p_F(\rho \geq 0.975)$	$p_J(\rho \geq 0.975)$
Real GNP	62	0.003	0.005	0.019
Nominal GNP	62	0.020	0.063	0.141
Real per capita GNP	62	0.003	0.004	0.016
Industrial production	111	0.001	0.003	0.192
Employment	81	0.004	0.014	0.060
Unemployment rate	81	0.002	0.000	0.087
GNP deflator	82	0.010	0.029	0.062
Consumer prices	111	0.196	0.528	0.652
Wages	71	0.018	0.046	0.100
Real wages	71	0.003	0.005	0.021
Money stock	82	0.005	0.025	0.044
Velocity	102	0.592	0.204	0.642
Interest rate	71	0.617	0.892	0.998
Stock prices	100	0.040	0.059	0.278

Note: p_F and p_J denote results from a flat prior and from Jeffreys' prior.

Sources: DeJong and Whiteman (1991, p. 231) and Phillips (1991, p. 358).

where $g(\theta) = 1/\sigma$ is a diffuse prior, $i_A(\theta)$ is an indicator function for $\theta \subseteq A$ and the set A is given by

$$A = \{(\delta, \Lambda) : 0 \leq \delta < 0.016, 0.55 \leq \Lambda < 1.055\}$$

With this truncated prior the integrals required for obtaining the posterior distributions cannot be evaluated analytically. Thus they adopted the Monte Carlo integration developed by Kloek and van Dijk (1978) and Geweke (1989).

The second column of table 8.1 shows part of their results. Based on the estimated posterior probabilities of the dominant root, they infer that evidence in support of a stochastic trend is present for only two series (velocity and bond yields) and they deem the evidence to be marginal in the case of a third series (consumer prices).

Sowell (1991) criticized the methodology of DeJong and Whiteman (DJW) arguing that

- (i) in DJW, analysis is restricted to AR(3) models, while it is not in Nelson and Plosser (1982),
- (ii) there is no formal justification for the arbitrary rule that the posterior mass associated with values of $(\Lambda \geq 0.975)$ is support for a unit root hypothesis,

- (iii) the question of the correct model for the (deterministic) trend behavior remains unanswered.

Regarding the criticism (ii), Phillips (1991) provides the posterior probabilities under the flat prior of ρ rather than the dominant root Λ . The fourth column of table 8.3 shows that the results of Phillips (1991) support a similar inference, although they are slightly different especially for consumer prices.

8.4 The debate on the appropriate prior

Sims (1988) suggested the simple flat-prior Bayesian theory for testing the unit root hypothesis. As mentioned earlier, Phillips (1991) raised the important point that the autoregressive time series model is different from the usual regression model and the flat prior used in the latter is not applicable to the time series model. However, the question is about the appropriateness of Jeffreys' prior that Phillips suggests (in fact any priors in general) for the time series models. In the debate in the *Journal of Applied Econometrics* (1991, volume 6, number 4) the main points raised were as follows (the debate is long but a few points stand out):

Phillips (1991) argued that flat priors ignore the way in which the coefficients influence the amount of information contained in the sample and suggested the use of Jeffreys' prior in equation (8.2). Phillips called this an *ignorance prior*. With the simple AR(1) model, the ignorance prior is given by

$$\pi(\rho, \sigma) \propto (1/\sigma) I_{\rho\rho}^{1/2}$$

with

$$\begin{aligned} I_{\rho\rho} &= \frac{T}{1-\rho^2} - \frac{1}{1-\rho^2} \frac{1-\rho^{2T}}{1-\rho^2} + \left(\frac{y_0}{\sigma}\right)^2 \frac{1-\rho^{2T}}{1-\rho^2}, \quad \rho \neq 1 \\ &= \frac{T(T-1)}{2} + T \left(\frac{y_0}{\sigma}\right)^2, \quad \rho = 1 \end{aligned}$$

The first issue raised was the appropriateness of Jeffreys' prior. The first objection is that the prior depends on the sample size T . It was argued by Leamer (1991) and Kim and Maddala (1991) that it favors high values of ρ (see figure 8.1). However, it was pointed out that this is the case with the simple autoregressive model

$$y_t = \rho y_{t-1} + \varepsilon_t$$

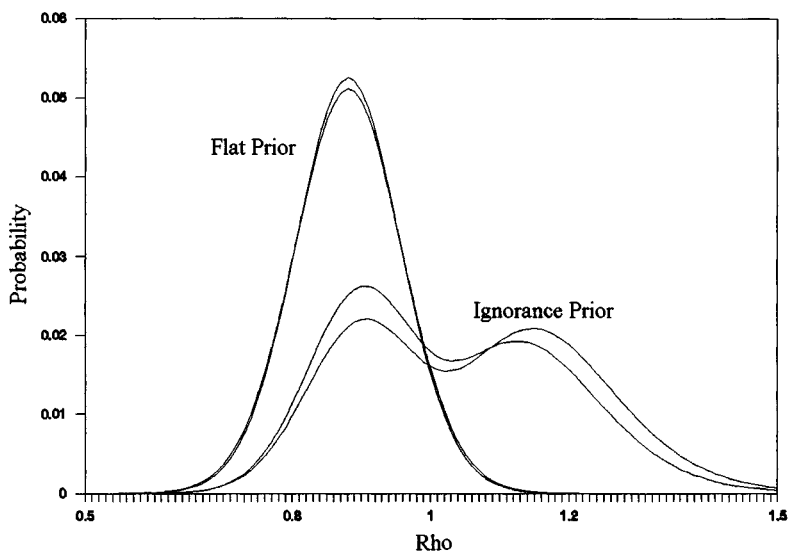


Fig. 8.1. Marginal posterior distributions of ρ when $\rho = 1$

but not in the model with a trend and intercept

$$y_t = \mu + \beta t + \rho y_{t-1} + \varepsilon_t$$

This was shown through Monte Carlo studies by Phillips and also analytically by Schotman and van Dijk (SVD) who argue (1991a, p. 388):

Paradoxically the ignorance prior proposed by Phillips downweights the unit root hypothesis relative to a flat prior in a model with trend and intercept.

Thus, the conclusions that follow from Jeffreys' prior are very sensitive to the way the model is formulated. One issue raised by SVD is that of alternative parameterization of the unit root model, the structural form

$$y_t = \gamma + \delta t + u_t, \quad u_t = \rho u_{t-1} + \varepsilon_t$$

and the reduced form

$$y_t = \mu + \beta t + \rho y_{t-1} + \varepsilon_t$$

with $\varepsilon_t \sim iid(0, \sigma^2)$. The parameters in the two equations are related as

$$\mu = \gamma(1 - \rho) + \delta\rho, \quad \beta = \delta(1 - \rho)$$

For the structural form, Jeffreys' prior is proportional to $|I_{\theta\theta}|^{1/2}$, where

$$I_{\theta\theta} = \begin{pmatrix} \bar{X}'\bar{X}/\sigma^2 & 0 & 0 \\ 0 & I_{\rho\rho} & 0 \\ 0 & 0 & 2T/\sigma^2 \end{pmatrix} \quad (8.3)$$

with

$$\bar{X} = (\bar{x}_1, \dots, \bar{x}_T)', \quad \bar{x}_t = x_t - \rho x_{t-1}$$

$$I_{\rho\rho} = \alpha_0(\rho) + \frac{(y_0 - \gamma)^2}{\sigma^2} \frac{1 - \rho^{2T}}{1 - \rho^2}$$

$$\alpha_0(\rho) = \frac{T}{1 - \rho^2} \cdot \frac{1 - \rho^{2T}}{1 - \rho^2}$$

Ignoring the term due to the initial condition y_0 by assuming $y_0 = \gamma$ and using (8.3), Jeffreys' prior then becomes

$$P(\rho, \gamma, \delta, \sigma) \propto (\det(I_{\theta\theta}))^{1/2} \propto \sigma^{-3} (1 - \rho)^2 \alpha_0(\rho)^{1/2} \quad (8.4)$$

The last proportionality sign can be verified by direct calculation of the determinant of the matrix $\bar{X}'\bar{X}$. Equation (8.4) has the convenient property that the priors on all elements of θ are independent. Note that the prior (8.4) does not depend on γ and δ . These parameters have flat priors. Also, the prior (8.4) drops to zero as $\rho \rightarrow 1$. SVD plot Jeffreys' priors for the autoregressive model with (i) no constant and no trend, (ii) constant only, and (iii) constant and time trend. In case (i) Jeffreys' prior gives considerable weight to values of $\rho > 1$. In cases (ii) and (iii) the prior has value zero at $\rho = 1$. They argue that this perhaps explains the bias toward stationarity in models with a fitted intercept and time trend.

Phillips, in his reply to the comment by SVD, argued that the problems with Jeffreys' priors noted by SVD arose from the fact that the reduced form of the unit root structural model has one less parameter under the null and this produces degeneracy. However, when we condition the likelihood function on the initial value, this degeneracy does not arise. To see this argument, let us consider the model with no trend. The structural model is

$$y_t = \mu + u_t, \quad u_t = \rho u_{t-1} + \varepsilon_t, \quad \varepsilon_t \sim iin(0, \sigma^2)$$

The reduced form is

$$y_t = \alpha + \rho y_{t-1} + \varepsilon_t$$

with $\alpha = \mu(1 - \rho)$ so that when $\rho = 1$, μ disappears from the reduced form. One way to avoid this degeneracy is to write the model conditional on u_0 . We then have the reduced form

$$y_t = \begin{cases} \mu + \rho u_0 + e_1, & t = 1 \\ \mu(1 - \rho) + \rho y_{t-1} + \varepsilon_t, & t = 2, \dots, T \end{cases}$$

Now when $\rho = 1$, μ does not disappear from the model. It is retained in the first observation and contributes to the level of the series. This seemingly innocuous inclusion of the initial value has an enormous impact on the Bayesian analysis of the unit root hypothesis.

We now write down the log-likelihood function

$$L(\rho, \mu, \sigma | \text{data}, u_0) \propto -T \log(\sigma) - \frac{(y_1 - \mu - \rho u_0)^2}{2} - \frac{\sum (y_t - \mu(1 - \rho) - \rho y_{t-1})^2}{2}$$

The information matrix involves u_0 and is of the form

$$I_{\theta\theta} = \begin{bmatrix} A & B & 0 \\ B & C & 0 \\ 0 & 0 & D \end{bmatrix}$$

where

$$\begin{aligned} A &= [1 + (T - 1)(1 - \rho)^2] / \sigma^2 \\ B &= u_0(1 + \rho - \rho^T) / \sigma^2 \\ C &= \frac{u_0^2}{\sigma^2} \left(\frac{1 - \rho^{2T}}{1 - \rho^2} \right) + \alpha_0(\rho) \\ D &= (T - 2) / \sigma^2 \end{aligned}$$

where $\alpha_0(\rho)$ is defined earlier. If we assume $u_0 = 0$, in which case $y_0 = \mu$, we get Jeffreys' prior as

$$\pi(\rho, \mu, \sigma | u_0 = 0) \propto \sigma^{-2} [\alpha_1(\rho) \alpha_0(\rho)]^{1/2}$$

where

$$\alpha_1(\rho) = 1 + (T - 1)(1 - \rho)^2$$

Note that SVD had $(1 - \rho)^2$ instead of $\alpha_1(\rho)$. As before, the joint Jeffreys' prior does not depend on μ , but one can think of the prior for μ as

$$\pi(\mu) \propto \sigma^{-1} [\alpha_1(\rho)]^{1/2}$$

with the priors for σ and ρ as

$$\pi(\sigma) \propto \frac{1}{\sigma} \quad \text{and} \quad \pi(\rho) \propto [\alpha_0(\rho)]^{1/2}$$

The prior for μ conditional on σ and ρ is flat but not degenerate at $\rho = 1$. The above analysis assumes $u_0 = 0$. Since u_0 is not known, we can consider the distribution of u_0 and get the unconditional likelihood function. We can assume $u_0 \sim N(0, \sigma^2)$ which implies that the process started at time 0, or assume that the process started s periods before and then derive the distribution of u_0 . These alternatives are discussed in Uhlig (1994a). Further analysis of the test of the unit root hypothesis using Jeffreys' priors and conditioning on u_0 , as well as alternative distributions of u_0 , can be found in Zivot (1994). The important point to note is that conditioning on the initial value plays a very important role in the Bayesian inference of the autoregressive model.

Lubrano (1995) also emphasizes the importance of initial observations. He shows that with a structural model and an adequate treatment of the first observation, a Bayesian unit root test can produce results which are more or less in accordance with classical results. Lubrano uses the structural model

$$\begin{aligned} y_t &= \mu + \delta t + u_t \\ u_t &= \rho u_{t-1} + \varepsilon_t, \quad \varepsilon_t \sim iid(0, \sigma^2) \end{aligned}$$

as in Schotman and van Dijk (1991a).

Many authors consider y_0 fixed and define the likelihood function conditional of y_0 . Lubrano considers a random initial condition with the distribution of y_0 given by

$$y_0 \sim N\left(\mu, \frac{\sigma^2}{1 - \rho^2}\right)$$

that is the equilibrium distribution of u_t under $|\rho| < 1$. Zellner (1971, p. 88) shows that in the general stationary case it does not matter whether y_0 is treated as fixed or random, but Lubrano shows that it makes a difference in the unit root case, when there is a constant term in the model (for details refer to his paper).

This initial distribution is not valid for $|\rho| > 1$. To allow for the possibility of small values of $|\rho|$ greater than 1, Lubrano modifies the distribution of y_0 as

$$y_0 \sim N\left(\mu, \frac{\sigma^2}{g(1 - \rho^2)}\right)$$

where

$$g(1 - \rho^2) = \begin{cases} 0 & \text{if } |\rho| > \sqrt{1+v} \\ (1 - \rho^2 + v)^2/4v & \text{if } \sqrt{1-v} \leq |\rho| \leq \sqrt{1+v} \\ (1 - \rho^2) & \text{if } |\rho| < \sqrt{1-v} \end{cases}$$

and v is a small positive quantity. He suggests $v = 0.5$ so that the interval for ρ is in $[-1.225, 1.2251]$ which is wide enough.

As for the prior, one can consider a mass point at $\rho = 1$ as done in Schotman and van Dijk (1991a) and compare the posterior odds. The other alternative Lubrano considers is a diffuse prior which is obtained as a limit of an informative prior. For this he considers a Beta distribution

$$\pi(\rho|v) \propto (\sqrt{1+v} + \rho)^{p-1}(\sqrt{1+v} - \rho)^{q-1}, \quad p > 0, q > 0$$

This is an informative prior in a finite range. Setting $p = 1$ and $q = 0$ he obtains the non-informative limit

$$\pi(\rho|v) \propto (\sqrt{1+v} - \rho)^{-1}$$

Its shape is similar to that of Jeffreys' prior but it does not depend on T . This prior allows for explosive values of ρ in the case $1 < \rho < \sqrt{1+v}$, but does not depend on T . Finally, Lubrano compares his results on posterior odds with those from Phillips' prior (Jeffreys' prior), the SVD prior, and with those from the ADF test using the extended Nelson-Plosser data (up to 1988), and shows that his results are more in line with those of the classical results. Of the 14 series, the ADF test rejects the unit root null in five series, with the posterior odds, if we reject the unit root null if the posterior probability $p(\rho \geq 1|y)$ is < 0.05 , the Lubrano prior rejects the unit root null in six series, Phillips' prior rejects it in seven series, and the SVD prior rejects it in eight series (see table 1, p. 99 of Lubrano's paper).

Apart from the issue of initial observations, there is another issue that arises when estimating models with several lags. The usual models estimated include lagged values of Δy_t so that the model is

$$y_t = \mu + \beta t + \rho y_{t-1} + \sum_{i=1}^{k-1} \phi_i \Delta y_{t-i} + u_t$$

The use of Jeffreys' prior discussed earlier does not take into account the correlation between the long-run dynamics (captured by ρ) and the short-run dynamics (captured by ϕ_i). Zivot and Phillips (1994) show that the correct application of Jeffreys' principle to this model yields a prior which is difficult to work with numerically and a resulting posterior

which is improper. They suggest a variant of Phillips' prior (1991) which yields a proper posterior and also allows for interaction between the long-run and short-run parameters. However, this is an approximation to the correct prior and still ignores the same off-diagonal elements of the information matrix. It also depends on some arbitrary constant ε and they recommend in practice to try different values of ε . Because of these limitations we shall omit the details here.

8.5 Classical tests versus Bayesian tests

Of course, in problems like this, there is the question of the comparability of classical significance levels and Bayesian posterior odds. The problem is whether a classical p -value can be viewed as a posterior probability of the null hypothesis. In some cases where one is testing a one-sided null hypothesis against a one-sided alternative hypothesis such an interpretation is possible (see Berger, 1985, pp. 147–148, Casella and Berger, 1987, and the references there). This interpretation is not possible for tests of point null against one-sided or two-sided alternatives (see Berger, 1985, pp. 148–151, Berger and Sellke, 1987, and the references there). However, some approximate equivalence is discussed in Hodges (1992). On the other hand, Andrews (1994) derives some new results and discusses the relationship of his results to earlier ones. He shows that for certain priors, the Bayesian posterior odds test is equivalent in large samples to classical Wald, LM, and LR tests for some significance levels and vice versa. This result is more general for the purpose of the discussion here. For our purpose the earlier references are sufficient.

8.6 Priors and time units of measurement

There were also some general questions raised about the use of priors in general. Leamer argues that one should not be stuck with a rigid prior irrespective of the frequency of the time series and the variable being considered. Leamer (1991, p. 371) argues:

Do we have the same ideas about ρ if y is nominal GNP, real wages, the saving rate, the Dow Jones average ...? I don't. Do you have the same ideas about ρ if the time unit is centuries, decades, years, days and seconds. I don't...

Sims (1991, p. 425) discusses the last point regarding analysis of the same problem at different frequencies. Suppose we are analyzing the same problem with monthly data and annual data. If a univariate AR

specification is exactly correct at the monthly interval with a coefficient ρ , it remains correct for the annual data with coefficient ρ^{12} . If the prior pdf applies to $\phi = \rho^{12}$ for the annual data, the implied prior for ρ for the monthly data is

$$g(\rho) = \rho^{11} f(\phi) = \rho^{11} f(\rho^{12})$$

Thus, if we have a flat prior for ϕ for annual data, our corresponding prior for monthly data is ρ^{11} , a convex upward-sloping function of ρ . A prior $\exp(-\phi)$ at the annual level, implies that the prior for monthly data is proportional to $\rho^{11} \exp(-\rho^{12})$. Sims plots this prior. It starts rising from $\rho = 0.6$, reaches its peak at $\rho = 1$, but falls off to zero for $\rho = 1.2$. There is one prior that remains invariant to the choice of the time unit of measurement – that is a flat prior on $\log(\rho)$ over $\rho > 0$ (i.e., a prior pdf of $1/\rho$).

8.7 On testing point null hypotheses

Given that the unit root null hypothesis $\rho = 1$ is a point null hypothesis, we cannot use a continuous prior density for ρ , since any such prior will give $\rho = 1$, a prior probability zero, and hence a posterior probability zero. The problems associated with the testing of point null hypotheses in the Bayesian framework have been reviewed in Berger and Delampady (1987). DeJong and Whiteman (1991) use intervals like $P(\Lambda < 0.975)$ where Λ is the largest root, in their posterior probability calculations. Their procedure has been criticized on the grounds that they are not really testing the unit root null and that many hypotheses in economics (like the market efficiency hypothesis and the permanent income hypothesis) are point null hypotheses. Thus, it is important to consider a Bayesian approach to the unit root null (see Schotman and van Dijk, 1991a, pp. 392–393, for this argument). Whether any economic theories really hinge on testing precise hypotheses is a debatable issue, but it is often argued that one needs a nonzero prior for the null to do posterior odds calculations for the unit root hypothesis. Phillips and Ploberger (1994) argue that this is not the case and this approach has been followed up by Phillips (1995). As we shall argue the approach in Phillips and Ploberger (1994) is some sort of *predictive approach*. An earlier discussion of the Bayesian predictive approach is in Geisser and Eddy (1979).

8.7.1 The posterior odds approach

Using posterior distribution

In the classical approach the null hypothesis and the alternative hypothesis are not on the same footing. The null hypothesis is on a pedestal and it is rejected only if there is overwhelming evidence against it. This is usually implicit in the use of the 5 percent and 1 percent significance levels. In the Bayesian approach the null and alternative hypotheses are on the same footing. There are several ways of comparing hypotheses using Bayesian methods, but the most common one is the posterior odds ratio. If we are interested in two hypotheses H_1 and H_2 , then the posterior odds ratio K_{12} based on data y is given by

$$K_{12} = \frac{P(H_1|y)}{P(H_2|y)} = \frac{P(y|H_1) \cdot P(H_1)}{P(y|H_2) \cdot P(H_2)}$$

If θ_1 and θ_2 are the relevant parameters under H_1 and H_2 respectively, then, since

$$P(y|H_i) = \int P(y|\theta_i, H_i) \cdot P(\theta_i|H_i) d\theta_i$$

the posterior odds ratio is calculated as

$$K_{12} = \frac{P(H_1) \int P(\theta_1|y, H_1) d\theta_1}{P(H_2) \int P(\theta_2|y, H_2) d\theta_2}$$

where $P(\theta_i|y, H_i)$, $i = 1, 2$, is the posterior distribution of θ_i and is given by

$$P(\theta_i|y, H_i) \propto P(\theta_i|H_i) \cdot L(y|\theta_i, H_i)$$

Thus, the posterior odds ratio depends on the prior probabilities for the hypotheses, the prior distributions for the parameters θ_i under the respective hypotheses, and the likelihood functions under the two hypotheses. Although the decision of which hypothesis to accept should be based on an explicit loss function, the common procedure is to accept H_1 if $K_{12} > 1$ and to accept H_2 otherwise.

The posterior odds approach, although straightforward, has implications on the testing of point null hypotheses. Implied in the prior odds for the two hypotheses is an assumption about the prior probability for the null. For instance, suppose we are interested in testing the hypothesis $\rho = 0$ in the model

$$y_t = \beta x_t + u_t, \quad u_t = \rho u_{t-1} + \varepsilon_t, \quad \varepsilon_t \sim \text{iin}(0, \sigma^2)$$

Let

H_1 : Model with no serial correlation in the errors.

H_2 : Model with serial correlation in the errors.

If we specify the prior odds $P(H_1)/P(H_2) = 1$ arguing that we do not know which model is correct, implicitly we are specifying a prior probability $P(\rho = 0) = 1/2$. Also note that the prior probabilities for the parameters and the prior probabilities for the hypotheses are interrelated.

Sims (1988) proposed the posterior odds ratio test, which follows the lines of Leamer's (1978) general version of the idea embodied in the Schwarz's Bayesian information criterion (BIC). Since the Schwarz criterion's asymptotics depend on the distribution of the estimate converging at the same rate for all true parameter values in a neighborhood of the null hypothesis, it is not applicable directly for choosing the model between the stationary and the unit roots models. Sims shows that the criterion would be

$$\tau = 2 \log \left(\frac{1 - \alpha}{\alpha} \right) - \log(\hat{\sigma}_\rho^2) + 2 \log(1 - 2^{-1/s})$$

where s is the number of periods per year (e.g., 12 for monthly data) and α is the prior probability (weight) on the interval $(0,1)$, probability $1 - \alpha$ on $\rho = 1$. Since the first and the last terms of τ are constant for a given data and prior, a small τ would be an evidence against the unit root. Based on this fact, if $t^2 > \tau$, we reject the unit root hypothesis, otherwise, we do not. Sims suggests $\alpha = 0.8$ at which the odds (between the stationarity and unit root) are approximately even.

Using predictive distribution

Another approach to model choice is through predictive distributions rather than posterior distributions. A discussion of predictive methods for model selection from the classical point of view can be found in the paper by Wei (1992). In the Bayesian framework, the predictive density of future observations y^* , is defined as $P(y^* | \text{data } y)$. Using the laws of probability, we see that

$$P(y^* | y) = \int L(y^* | \theta, y) P(\theta | y) d\theta$$

where $L(y^* | \theta, y)$ is the likelihood function of the future observations and $P(\theta | y)$ is the posterior density of θ .

The predictive odds ratios for the two models are given by

$$K_{12} = \frac{P(y^*|y, H_1) \cdot P(H_1)}{P(y^*|y, H_2) \cdot P(H_2)}$$

But

$$P(y^*|y, H_i) = \int P(y^*|y, \theta_i, H_i) \cdot P(\theta_i|y, H_i) d\theta_i$$

Dropping H_i in the conditioning variables, we get

$$K_{12} = \frac{\int P(y^*|y, \theta_1) \cdot P(\theta_1|y) d\theta_1}{\int P(y^*|y, \theta_2) \cdot P(\theta_2|y) d\theta_2} \cdot \frac{P(H_1)}{P(H_2)}$$

where $P(\theta_i|y)$ is the posterior distribution of θ_i given y (under H_i).

The predictive density in the Bayesian language has been re-interpreted in classical language as a *Bayes model* or *Bayesian frame of reference* and used to derive classical tests for unit roots by Phillips and Ploberger (1994) and by Phillips (1994a, 1994b, 1995) for model selection and prediction. In essence this work amounts to the use of predictive densities and predictive posterior odds and the differences between the two approaches are minor.

The fundamental arbitrariness of the posterior odds approach with noninformative priors is discussed in Leamer (1978, section 4.5). On the other hand, the predictive odds approach does not suffer from this problem. Initial noninformative priors of the form $P(\beta, \sigma^2) \propto \sigma^{-2}$ are applied to each model over periods 1 to m ($m < n$) to produce *proper* posterior densities for (β, σ^2) . These are then used as priors to evaluate *proper* predictive densities for periods $(m + 1)$ to n . The ratio of two such predictive densities is unambiguously defined.

8.7.2 Model selection approach

Predictive approach

Geisser and Eddy (1979) put forth two model selection criteria based on predictive sample re-use (PSR) methods. These are, however, suitable for independently distributed variates such that x_j has density $f(x_j|\theta_k, M_k)$ where M_k is the k th model. The first criterion, termed PSR quasi-likelihood (PSRQL) selects the model that maximizes

$$\hat{L}_k = \prod_{j=1}^N f(x_j|\hat{\theta}_{k(j)}, M_k)$$

where $\hat{\theta}_{k(j)}$ is the MLE of θ_k with x_j omitted. The second criterion, termed PSR quasi-Bayes (PSRQB) selects the model that maximizes

$$L_k = \prod_{j=1}^N f_p(x_j|x_{(j)}, M_k)$$

where $x_{(j)}$ denotes that x_j has been deleted and f_p denotes the predictive density, which is calculated as

$$f_p(x_j|x_{(j)}, M_k) = \int f(x_j|\theta_k, M_k) dP(\theta_k|x_{(j)}, M_k)$$

where $P(\theta_k|x_{(j)}, M_k)$ is the posterior density of θ_k based on $x_{(j)}$ and usually a diffuse prior on θ_k (see Geisser and Eddy for details).

Posterior information criterion (PIC)

Given a number of regression models, the Bayesian information criterion (BIC) chooses the model that minimizes

$$\log \hat{\sigma}_k^2 + k \frac{\log(n)}{n}$$

where $\hat{\sigma}_k^2$ is the estimate of σ^2 the error variance with k regressors and n is the sample size.

Wei (1992) suggests the Fisher information criterion (FIC) for model choice. This suggests choosing the model that minimizes

$$FIC_k = n\hat{\sigma}_k^2 + \hat{\sigma}_K^2 \ln |A_k|$$

where $A_k = X'X$ for the model with k parameters and $\hat{\sigma}_k^2$ and $\hat{\sigma}_K^2$ are variance estimates of σ^2 based on the model with k parameters and the full set of K parameters respectively. Compared with AIC and BIC that have a penalty term that is based on the dimension of the selected model, the FIC has a term that is proportional to the logarithm of the statistical information in the model with k parameters. The redundant information, by introducing a spurious variable, is used to represent its penalty. The stimulation studies by Wei (1992), however, are for models with exogenous regressors. He shows that FIC performs very well in selecting the true model, but the performance of BIC is very close to that of FIC.

Phillips and Ploberger, in a number of unpublished papers, the published versions of which are Phillips and Ploberger (1994, 1996) and Phillips (1994a, 1994b, 1995), suggested an alternative criterion for

model choice: the *posterior information criterion* (PIC). This minimizes

$$c_{1K} |A_k / \hat{\sigma}_K^2|^{1/2} \exp[-(1/2\hat{\sigma}_K) \hat{\beta}_k' A_k \hat{\beta}_k]$$

where c_{1K} is a constant depending on K , the maximum number of regressors, $\hat{\sigma}_K^2$ is the estimate of the error variance from the regression model with the largest number of regressors, K , $A_k = (X_k' X_k)$ where X_k is the matrix of observations on the k regressors, and $\hat{\beta}_k$ is the estimate of the coefficient vector with k explanatory variables.

Unlike the BIC which depends on the number of regressors k and the residual variance $\hat{\sigma}_k$, the PIC depends explicitly on the data matrix X_k . This is an advantage of the PIC. The motivation behind the PIC is complicated and described in the papers by Phillips and Ploberger. However, the procedure is not completely Bayesian because of the way the residual variance is estimated (from the largest model). Richard in his comment on Phillips (1995) objects to this and suggests using $\hat{\sigma}_k^2$ for the k th model and constructing Bayesian posterior odds. When viewed in the light of Bayesian posterior odds, the limitations of PIC as a criterion of model choice are clear. For more detailed discussion of this point, see the paper by Phillips (1995), the comments, and Phillips' reply.

However, Phillips discusses another criterion, PICF, which is an extension of PIC based on predictive densities. As argued earlier predictive posterior odds do not suffer from the same problems as posterior odds with noninformative priors. But the predictive odds have to be calculated as discussed earlier, using a subsample to get *proper* posterior distributions and then using these as priors for computing predictive densities of the *remaining* observations. One can do this sequentially for the k th observation ($k = m + 1, m + 2, \dots, n$), using the predictive odds.

Phillips uses his PIC and PICF criteria to do such sequential model choice where you are not confined to a single model for the whole period. This can also be done using the Bayesian predictive odds sequentially.

Koop (1994) uses a *Bayesian likelihood ratio* (BLR) criterion for choosing between $I(0)$ and $I(1)$ models, which is also due to Phillips and Ploberger. This criterion is described in section 2 of Phillips' paper, "The Long-run Australian Consumption Function Reexamined: An Empirical Exercise in Bayesian Inference" (September, 1991). It's motivation is the same as that for PIC.

For the model

$$y_t = \mu + \beta t + \rho y_{t-1} + \sum_{i=1}^{k-1} \phi_i \Delta y_{t-i} + u_t$$

where $u_t \sim IN(0, \sigma^2)$, the BLR criterion is

$$BLR = \frac{\exp[-\frac{1}{2}\hat{\sigma}^{-2}(\hat{\rho} - 1)^2 y'_{-1} Q y_{-1}]}{\hat{\sigma}^{-2}(\hat{\rho} - 1)^2 y'_{-1} Q y_{-1}}$$

where $\hat{\sigma}^2$ is the OLS estimator of σ^2 , $\hat{\rho}$ is the OLS estimator of ρ , y_{-1} is the vector of y_t lagged once, and $Q = I - X(X'X)^{-1}X'$ with X as the matrix of observations on the explanatory variables $1, t, \Delta y_{t-1}, \dots, \Delta y_{t-k+1}$. With equal prior odds, we decide in favor of the hypothesis $\rho = 1$ if $BLR < 1$.

8.8 Further comments on prior distributions

In the preceding sections we discussed the controversy over the flat prior and Jeffreys' prior in the case of time series models, the importance of taking account of the initial conditions in Bayesian inference, and the posterior odds approach to the problem of testing the unit root hypothesis. We shall now discuss further results on priors and on Bayesian model selection methods.

The paper by Kass and Wasserman (1996) gives a critical survey (with an annotated bibliography) of the different methods of generating prior distributions for Bayesian inference. Section 3.5.1 of their paper discusses the problem of priors for autoregressive models, using the Berger–Bernardo method, which is a method of generating noninformative priors. See Berger and Bernardo (1992). This method leads to the prior

$$\pi(\rho) \propto (\sqrt{1 - \rho^2})^{-1} \quad \text{if } |\rho| < 1$$

To allow for the possibility of explosive roots Berger and Yang (1994) suggested the prior

$$\pi(\rho) = \begin{cases} \left[2\phi\sqrt{1 - \rho^2}\right]^{-1} & \text{if } |\rho| < 1 \\ \left[2\phi|\rho|\sqrt{1 - \rho^2}\right]^{-1} & \text{if } |\rho| > 1 \end{cases}$$

This is the Berger–Bernardo prior in the range $|\rho|$ and the prior outside this range is obtained by mapping $\rho \rightarrow 1/\rho$. This is actually a proper prior integrating to 1. It gives equal probability to $|\rho| < 1$ and $|\rho| > 1$.

Another class of priors is the *maximal data information priors* (MDIP) suggested by Zellner, starting with Zellner (1971). See Zellner (1996) for a historical review and an account of recent developments, including Zellner and Min (1993). The basic idea is to choose the prior that maximizes the difference in the information provided by the data and the information provided by the prior density using Shannon's measure of information. Define

$$I(\theta) = - \int f(y|\theta) \log f(y|\theta) dy$$

This is the information in $f(y|\theta)$. Zellner and Min (1993) suggest choosing the prior that maximizes the difference

$$G = \int I(\theta)\pi(\theta)d\theta - \int \pi(\theta) \log \pi(\theta)d\theta$$

The first term is the average information in $f(y|\theta)$ and the second term is the information in the prior density. This method yields the prior

$$\pi(\theta) \propto \exp[I(\theta)]$$

This prior, however, is not invariant to reparameterization. For the simple AR(1) model

$$y_t = \rho y_{t-1} + \varepsilon_t, \quad \varepsilon_t \sim IN(0, \sigma^2)$$

the MDIP prior is given by

$$\pi(\rho, \sigma) \propto \frac{1}{\sigma} (1 - \rho^2)^{1/2}$$

which converges to zero as $\rho \rightarrow 1$.

Uhlig (1994b) discusses the behavior of the different priors. Since the paper by Uhlig summarizes a number of important issues regarding the Bayesian approach to unit roots and cointegration, we shall quote some points made.

- (i) Consider a VAR model in Y_t , the regression parameters being denoted by B and the covariance matrix of errors by Σ . Denote Σ^{-1} by H . Then given the data Y_t , the conditional likelihood function as a function of B and H is proportional to a Normal-Wishart density function. This is true regardless of whether there are unit roots, cointegrating vectors, explosive roots or not. (The Normal-Wishart distribution specifies that the precision matrix H follows a Wishart distribution and that conditional on H , the

matrix B in its column vectorized form $\text{vec}(B)$ follows a multivariate normal distribution. See Zellner (1971) for the use of the Normal-Wishart distribution in Bayesian inference.) If we condition on H as well, the likelihood function is proportional to a normal distribution. Hence conventional t and F statistics and their p -values are meaningful in summarizing the shape of the likelihood function regardless of whether there are unit roots or not.

- (ii) Whereas the conditional likelihood function is viewed as a function of the data given the parameters may not be standard (the classical perspective), the conditional likelihood function is viewed as a function of the parameters given the data is standard (the Bayesian perspective). This is the central message of Sims and Uhlig (1991).
- (iii) If the prior is given by a Normal-Wishart density, the posterior is also given by a Normal-Wishart density (see Zellner, 1971). This is the reason why the Normal-Wishart prior is popular. However, in contrast with the case of exogenous regressors, the Normal-Wishart prior has the disadvantage that in AR models it can be informative about some properties of the models. This was pointed out by Phillips (1991). The argument is that some parameter regions are packed *denser* than the others. It makes sense to reparameterize the model so that the parameter values are *evenly packed*. This amounts to a prior proportional to $|I|^{1/2}$ which is Jeffreys' prior. However, this prior also has some problems which have been discussed in section 8.4 earlier.
- (iv) For the univariate AR(1) model, Uhlig plots the different priors. This shows that the major differences are in the explosive region $|\rho| > 1$. Hence he concludes that conditional on non-explosive roots, the differences between the prior suggested by Phillips (1991) and a flat prior is small and will usually not matter in practical applications. The difference becomes large once explosive roots are taken seriously. If explosive roots are taken seriously, then it is best to report the results from several priors discussed earlier.
- (v) Pre-testing for unit roots or trend stationarity and proceeding as if one is sure about the conclusions of the pre-test can be misleading in calculating the uncertainty of n -step ahead forecasts and also in answering macroeconomic questions, in general. The Bayesian methods take the uncertainty about the presence of a

unit root into account. But the tails of the predictive densities can be sensitive to the prior treatment of explosive roots.

8.9 Bayesian inference on cointegrated systems

There have been many procedures for Bayesian tests for cointegration and Bayesian analysis of cointegrated systems. First, we shall discuss tests for cointegration.

Bayesian tests for cointegration

The simplest of the Bayesian tests for cointegration is to apply Bayesian unit root tests (discussed earlier) to the residuals from an estimated cointegrating regression. This procedure is not fully Bayesian because the estimation of the cointegrating regression is not done by Bayesian methods. The Bayesian methods are brought in at the end.

Koop (1994) presents an alternative approach based on the number of nonstationary roots in a VAR system.¹ In a cointegrated system with k variables which are $I(1)$, if there are r cointegrating vectors, then this leaves $(k - r)$ series that are unit root series. Thus, cointegration is present if the number of unit roots in the VAR is less than the total number of unit roots in the univariate series. Koop defines cointegration in a different way so that the problems of testing a point null hypothesis (discussed earlier) do not arise. He defines cointegration to be present if the number of *nonstationary* roots driving the VAR is less than the number of nonstationary roots in the univariate series.

Koop's procedure runs as follows: let λ^U be the number of nonstationary roots in the univariate models and λ^M be the number of nonstationary roots in the multivariate model (VAR model). Suppose that the Bayesian analysis gives

$$P(\lambda^U = m|y) = 0.90, \quad P(\lambda^U = m - 1|y) = 0.10$$

and

$$P(\lambda^M = m|y) = 0.05, \quad P(\lambda^M = m - 1|y) = 0.25, \quad P(\lambda^M = m - 2|y) = 0.70$$

Then there is cointegration if $\lambda^M < \lambda^U$. Hence the probability that there is cointegration

$$P(\lambda^M < \lambda^U|y) = (0.90)(0.25 + 0.70) + (0.10)(0.70) = 0.925$$

¹ Koop (1991) also discusses a Bayesian approach to cointegration. Since this paper is based on unit root tests that do not take into the developments in Bayesian unit root testing since 1991, we shall omit the discussion of this paper here.

i.e., there is 92.5 percent probability that there is cointegration. The posterior probabilities $P(\lambda^U|y)$ can be calculated using the univariate methods described earlier. For calculating the posterior probabilities $P(\lambda^M|y)$, Koop starts with the Geisser prior discussed in Zellner (1971, pp. 224–233 and 396–399)

$$P(\phi, \Sigma) \propto |\Sigma|^{-(m+1)/2}$$

For this prior the posterior density $P(\phi|y)$ is a generalized student t . Koop does n repeated draws from this posterior density and calculates the number of eigenvalues greater than or equal to 1 in absolute value. The frequency distribution gives the probabilities $P(\lambda^M = m)$ as n the number of draws increases. Using this procedure, Koop finds that there are no common trends in stock prices or exchange rates across countries, but that for any given country, spot and forward exchange rates are cointegrated.

There is, however, one major problem with Koop's procedure. This is the use of the Geisser (*noninformative*) prior for the VAR model. This prior was suggested for the standard multivariate regression model. The VAR model is **not** a standard regression model. Earlier we mentioned the criticism of Phillips against the use of flat priors, appropriate for the standard regression model, for the AR(1) model, arguing that the AR(1) model is **not** a standard regression model. The same criticism applies here in the multivariate case. The problem of appropriate priors for the VAR model remains to be studied.

Bayesian analysis of cointegration

There are three papers concerned with the Bayesian estimation of cointegrated systems: Kleibergen and van Dijk (1994), Bauwens and Lubrano (1996), and Geweke (1996). Since the Bayesian analysis is notation heavy and involves a lot of integration (analytical as well as numerical) we shall not go into the details here. Readers interested in this area can refer to the original articles. Of more concern to us is the priors used and the appropriate priors for cointegration analysis. The discussion on priors for unit root models in the preceding sections is relevant here. An important point we mentioned earlier is to note that the VAR model used in cointegration analysis is not the usual multivariate model with exogenous regressors and hence priors appropriate for this latter model are not appropriate for the cointegration model.

Geweke (1996) considers the model

$$Y = XA + ZB + E$$

where Y, X , and Z are respectively $n \times L$, $n \times p$, and $n \times k$ matrices, A and B are respectively $p \times L$ and $k \times L$ matrices of parameters, and E is an $n \times L$ matrix of errors. Z is supposed to be of full rank but X is supposed to be of reduced rank q ($q < p$ or L).

Geweke develops Bayesian methods for the analysis of this model and also gives procedures for determining q . Since this method is not directly concerned with the cointegration model, we shall not discuss it in detail. The main issue we are concerned with is the identification of the cointegrating vectors in the Johansen procedure.

As for the priors, Geweke uses the inverted Wishart prior for Σ , the covariance matrix of the errors E s, and normal priors $N(0, \Sigma^{-2})$ for each of the elements of A . This is the Normal-Wishart prior that we discussed in section 8.8 (Uhlig's comments). It is a diffuse prior which Kleibergen and van Dijk criticize.

The paper by Kleibergen and van Dijk (1994) argues that diffuse priors should not be used in cointegration models. The resulting posteriors may be nonintegrable. They suggest the use of Jeffreys' priors for which this problem does not arise. But Jeffreys' prior need not be unique. They suggest four different versions of Jeffreys' prior, only some of which would be useful in practice.

The paper by Bauwens and Lubrano (1996) considers the identification problem in terms of the representation considered by Johansen. The matrix B , in section 5.5.1 is denoted by Π in the Bauwens and Lubrano paper. As done there they parameterize it as

$$\Pi = \alpha\beta'$$

where α and β are $n \times r$ matrices of rank r . Π is identified but α and β are not because $(\alpha H^{-1})(H\beta')$ is also equal to Π , where H is any $r \times r$ matrix of rank r .

The first result in the Bauwens and Lubrano paper is that $(2nr - r^2)$ elements of (α, β) are identified. As Π is identified and Π has n^2 elements, the resulting $(n - r)^2$ elements are determined by the rank condition $\text{rank}(\Pi) = r$.

The matrices α and β must be subject to a total of r^2 restrictions but they have $2nr$ elements. Bauwens and Lubrano use r normalization restrictions on β . This leaves $r(r - 1)$ restrictions. They do this by imposing linear restrictions on β . To do the Bayesian analysis, they note

that the model is a linear model conditional on β . Hence they say that the traditional noninformative prior (the Geisser prior discussed in Zellner, 1971) can be used. This results in the overall prior $|\Omega|^{-(n+1)/2}P(\beta)$ where Ω is the covariance matrix of the errors.

However, note that this prior for a multivariate regression (with exogenous regressors) is not valid for VAR models. We shall omit the details of the subsequent Bayesian analysis which can be found in the Bauwens and Lubrano paper.

The important thing to note is the use of the Geisser prior which is not valid for a VAR model. The argument is the same as that of Phillips (1991) as against the flat prior for the AR(1) model. The appropriate priors for VAR models and hence for cointegration models is still an open issue. There is a long discussion of appropriate priors for the AR(1) model which we reviewed earlier. There is no corresponding detailed discussion in the case of VAR models.

One other issue is: what new insights are being gained by the use of Bayesian analysis? This is not entirely clear in these papers. Of course, they are all interested in tackling the computational problems first. But at some point it is important to get to the issue of what more we are learning. The usual argument is that Bayesian analysis gives us finite sample results.

One important area where Bayesian analysis should be of help is in the pre-testing problem, which as we discussed in chapter 6 arises often in the analysis of cointegrated systems. There is no work on this yet.

We have given only a brief review of the studies on Bayesian analysis of cointegrated systems. Our purpose is to see what priors are being used and point out their limitations. The computational problems are not easy and the papers have made important contributions in this respect. But more attention needs to be devoted to the priors.

8.10 Bayesian long-run prediction

The Bayesian approach is specially useful in studying the uncertainty about forecasts and impulse response functions. Uhlig (1994b) demonstrates convincingly that pre-testing for unit roots or trend stationarity and then proceeding as if one is sure about the conclusion of this pre-test can be misleading with regard to n -step ahead forecasts. By contrast, in the Bayesian approach, the uncertainty about the underlying coefficients is taken into account in computing the predictive density.

Uhlig also demonstrates that the predictive density and, in particular,

its tails can be sensitive to prior treatment of explosive roots. Hence sensitivity analysis should be performed with alternative priors if explosive roots are to be taken seriously.

Koop *et al.* (1995) also show how Bayesian long-run forecasts can be very sensitive to apparently innocuous assumptions made in the prior. In the case of impulse response functions, the problems are more serious. Koop *et al.* (1994) show that apart from this sensitivity, impulse responses at some forecast horizons may have posterior density functions that have no finite moments. Hence it is important that the entire posterior density be studied rather than just the posterior mean and posterior variance.

8.11 Conclusion

The initial Bayesian analysis of the unit root model was by Sims (1988) who used a flat prior for the autoregressive model. Subsequently Phillips criticized this and suggested Jeffreys' prior, which itself has some problems. We discuss in detail the problems with the different priors suggested (section 8.2 to 8.4). Further comments on priors for unit root models along with the clarification of several issues by Uhlig are discussed in section 8.8.

One advantage of the Bayesian approach is the symmetric way the null and the alternative hypotheses are treated (compared with the classical approach where the null hypothesis is on a pedestal). The posterior odds approach is useful for comparing models, but there are problems with it if uninformative priors are used. We also discuss how the predictive odds approach does not have these problems.

The unit root hypothesis is a point null hypothesis. The problems created by this and the solutions offered have been discussed in this chapter. One has no problem evaluating $H_0 : |\rho| < 1$ versus $H_1 : |\rho| \geq 1$ because the posterior probabilities of these regions can be computed once the posterior distribution of ρ has been determined. But the posterior probability of $\rho = 1$ is zero with any continuous posterior distribution for ρ .

We also discuss the criteria for model selection suggested by Phillips and Ploberger: PIC, PICF, and BLR. A complete Bayesian approach is that of predictive odds ratios, which can also be used for model selection using sequential data. This would enable us to allow for different models for different spans of the data rather than consider a single model over the entire time span.

Finally, we consider some issues concerning the Bayesian analysis of cointegrated systems. One major conclusion is that one should not use diffuse priors for the Bayesian analysis of cointegrated systems. Other priors need to be investigated. Also, there is the issue that needs to be investigated as to the additional insights the Bayesian analysis gives that we do not get from classical analysis.

References

- Andrews, D.W.K. (1994), "The Large Sample Correspondence Between Hypothesis Tests and Bayesian Posterior Odds Tests," *Econometrica*, 62, 1207–1232.
- Bauwens, L. and M. Lubrano (1996), "Identification Restrictions and Posterior Densities in Cointegrated Gaussian VAR Systems," in T.B. Fomby (ed.), *Advances in Econometrics*, vol. 11B, JAI press, Greenwich, CT.
- Berger, J.O. (1985), *Statistical Decision Theory and Bayesian Analysis*, 2nd edn, Springer Verlag, New York.
- Berger, J.O. and J.M. Bernardo (1992), "On the Development of the Reference Prior Method," in J. M. Bernardo, J.O. Berger, A.P. Dawid, and A.F.M. Smith (eds.), *Bayesian Statistics 4: Proceedings of the Fourth Valencia International Meeting*, Clarendon Press, 35–60.
- Berger, J.O. and M. Delampady (1987), "Testing Precise Hypotheses," *Statistical Science*, 2, 317–352.
- Berger, J.O. and T. Sellke (1987), "Testing a Point Null Hypothesis: The Irreconcilability of p values and Evidence," *Journal of the American Statistical Association*, 82, 112–122.
- Berger, J.O. and R. Yang (1994), "Non-Informative Priors and Bayesian Testing for the AR(1) Model," *Econometric Theory*, 10, 461–482.
- Casella, G. and J.O. Berger (1987), "Reconciling Bayesian and Frequentist Evidence in the One-Sided Testing Problem," *Journal of the American Statistical Association*, 82, 106–111.
- DeJong, D.N. and C.H. Whiteman (1991), "Reconsidering Trends and Random Walks in Macroeconomic Time Series," *Journal of Monetary Economics*, 28, 221–254.
- Geisser, S. and W.F. Eddy (1979), "A Predictive Approach to Model Selection," *Journal of the American Statistical Association*, 74, 153–160.

- Geweke, J. (1988), "The Secular and Cyclical Behavior of Real GDP in Nineteen OECD Countries, 1957–1983," *Journal of Business and Economic Statistics*, 6, 479–486.
- (1989), "Bayesian Inference in Econometric Models Using Monte Carlo Integration," *Econometrica*, 57, 1317–1339.
- (1996), "Bayesian Reduced Rank Regression in Econometrics," *Journal of Econometrics*, 75, 127–146.
- Hodges, J. (1992), "Who Knows What Alternative Lurks in Significance Tests?" in J.M. Bernardo *et al.* (eds.), *Bayesian Statistics*, vol. 3, Oxford University Press, Oxford.
- Jeffreys, H. (1961), *Theory of Probability*, 3rd ed, Oxford University Press, London.
- Kass, R.E. and L. Wasserman (1996), "The Selection of Prior Distributions by Formal Rules," *Journal of the American Statistical Association*, 91, 1343–1370.
- Kim, I.M. and G.S. Maddala (1991), "Flat Priors versus Ignorance Priors in the Analysis of the AR(1) Model," *Journal of Applied Econometrics*, 6, 375–380.
- Kleibergen, F. and H.K. van Dijk (1994), "On the Shape of the Likelihood/Posterior in Cointegration Models," *Econometric Theory*, 10, 514–551.
- Kloek, T. and H.K. van Dijk (1978), "Bayesian Estimates of Equation System Parameters: An Application of Integration by Monte Carlo," *Econometrica*, 46, 1–19.
- Koop, G. (1991), "Cointegration Tests in Present Value Relationships," *Journal of Econometrics*, 49, 105–139.
- (1994), "An Objective Bayesian Analysis of Common Stochastic Trends in International Stock Prices and Exchange Rates," *Journal of Empirical Finance*, 1, 343–364.
- Koop, K., J. Oseiwalski, and M. Steel (1994), "Posterior Properties of Long-Run Impulse Responses," *Journal of Business and Economic Statistics*, 12, 489–492.
- (1995), "Bayesian Long-Run Prediction in Time Series Models," *Journal of Econometrics*, 69, 61–80.
- Leamer, E. (1978), *Specification Searches*, Wiley, New York.
- (1991), "Comment on 'To Criticize the Critics'," *Journal of Applied Econometrics*, 6, 371–373.
- Lindley, D.V. (1965), *Introduction to Probability and Statistics from a Bayesian Viewpoint*, 2 vols., Cambridge University Press, Cambridge.

- Lubrano, M. (1995), "Testing for Unit Roots in a Bayesian Framework," *Journal of Econometrics*, 69, 81–109.
- Nelson, C.R. and C.I. Plosser (1982), "Trends and Random Walks in Macroeconomic Time Series," *Journal of Monetary Economics*, 10, 139–162.
- Phillips, P.C.B. (1991), "To Criticize the Critics: An Objective Bayesian Analysis of Stochastic Trends," *Journal of Applied Econometrics*, 6, 333–364.
- (1994a), "Bayes Models and Forecasts of Australian Macroeconomic Time Series," in C. Hargreaves (ed.), *Non-Stationary Time Series Analysis and Cointegration*, Oxford University Press, 53–86.
- (1994b), "Model Determination and Macroeconomic Activity," Fisher-Schultz Lecture, European Meetings of the Econometric Society, Maastricht.
- (1995), "Bayesian Model Selection and Prediction with Empirical Applications," with comments by F.C. Palm and J.F. Richard and Reply by P.C.B. Phillips, *Journal of Econometrics*, 69, 289–365.
- Phillips, P.C.B. and W. Ploberger (1994), "Posterior Odds Testing for a Unit Root with Data-Based Model Selection," *Econometric Theory*, 10, 774–808.
- (1996), "An Asymptotic Theory of Bayesian Inference for Time Series," *Econometrica*, 64, 381–412.
- Raiffa, H. and R. Schlaifer (1961), *Applied Statistical Decision Theory*, Graduate School of Business Administration, Harvard University.
- Schotman, P.C. and H.K. van Dijk (1991a), "On Bayesian Routes to Unit Roots," *Journal of Applied Econometrics*, 6, 387–401.
- (1991b), "A Bayesian Analysis of the Unit Root in Real Exchange Rates," *Journal of Econometrics*, 49, 195–238.
- Sims, C.A. (1988), "Bayesian Skepticism on Unit Root Econometrics," *Journal of Economic Dynamics and Control*, 12, 463–474.
- (1991), "Comments on To Criticize the Critics by P.C.B. Phillips," *Journal of Applied Econometrics*, 6, 423–434.
- Sims, C.A. and H. Uhlig (1991), "Understanding Unit Rooters: A Helicopter Tour," *Econometrica*, 59, 1591–1599.
- Sowell, F. (1991), "On DeJong and Whiteman's Bayesian Inference for the Unit Root Model," *Journal of Monetary Economics*, 28, 255–263.

- Thornber, H. (1967), "Finite Sample Monte Carlo Studies: An Autoregressive Illustration," *Journal of the American Statistical Association*, 62, 801–808.
- Uhlig, H. (1994a), "A Note on Jeffreys' Prior When Using the Exact Likelihood Function," *Econometric Theory*, 10, 633–644.
- (1994b), "What Macroeconomists Should Know About Unit Roots: A Bayesian Perspective," *Econometric Theory*, 10, 645–671.
- Wei, C.Z. (1992), "On Predictive Least Squares Principles," *Annals of Statistics*, 20, 1–42.
- Zellner, A. (1971), *An Introduction to Bayesian Inference in Econometrics*, Wiley, New York.
- (1996), "Past and Recent Results on Maximal Data Information Priors," Technical Report, Graduate School of Business, University of Chicago.
- Zellner, A. and C. Min (1993), "Bayesian Analysis, Model Selection and Prediction," in W.T. Grandy Jr. and P.W. Millioli (eds.), *Physics and Probability: Essays in Honor of E.T. James*, Cambridge University Press, Cambridge, 195–206.
- Zivot, E. (1994), "A Bayesian Analysis of the Unit Root Hypothesis within an Unobserved Components Model," *Econometric Theory*, 10, 552–578.
- Zivot, E. and P.C.B. Phillips (1994), "A Bayesian Analysis of Trend Determination in Economic Time Series," *Econometric Reviews*, 13, 291–336.

9

Fractional unit roots and fractional cointegration

In the preceding chapters we assumed that the time series is $I(d)$ with d being 0 or 1 or some greater integer. The model with $d = 1$ corresponds to a model with persistence of shocks. However, persistence or long memory can also be modeled with $d > 0$ but < 1 . The introduction of autoregressive fractionally integrated moving average (ARFIMA) model by Granger and Joyeux (1980) and Hosking (1981) allows us to replace the discrete choice of unit root versus no unit root with a continuous parameter estimate of the long memory component of time series. Likewise, if two series y_t and x_t are $I(d)$, they are said to cointegrated if $y_t + \beta x_t$ is $I(b)$ with $b < d$, and we allow integer values for b and d . Typically $d = 1$ and $b = 0$. In the case of fractional cointegration, suppose that y_t and x_t are both $I(d)$ and $y_t + \beta x_t$ is $I(b)$ with $b < d$. If $b = 0$, then the long memory components in y_t and x_t are common and we say that y_t and x_t are fractionally cointegrated. If $b > 0$, then $y_t + \beta x_t$ has a long memory component left in it.

In this chapter we shall discuss the issues related to these problems. The paper by Granger (1980) gives a motivation behind the $I(d)$ process. Granger shows that the sum of a large number of stationary $AR(1)$ processes with random parameters can result in a process with long memory. Thus it makes sense to consider $I(d)$ processes while analyzing aggregate data.

9.1 Some definitions

A time series x_t follows an ARFIMA(p, d, q) process if

$$\Phi(L)(1-L)^d x_t = \mu + \Theta(L)\varepsilon_t, \quad \varepsilon_t \sim iid(0, \sigma^2)$$

where $\Phi(L) = 1 - \phi_1 L - \dots - \phi_p L^p$, $\Theta(L) = 1 - \theta_1 L - \dots - \theta_q L^q$, and μ can be any deterministic function of time. If x_t is trend-stationary, then $d = 0$. If x_t is difference-stationary, then $d = 1$ (or some greater integer). The ARFIMA model generalizes the ARIMA model by allowing the differencing parameter d to take on any real value, rather than restricting it to the integer domain. The fractional differencing term $(1 - L)^d$ can be written as an infinite order MA process using the binomial expansion

$$\begin{aligned} (1 - L)^d &= 1 - dL + \frac{d(d-1)}{2!} L^2 - \frac{d(d-1)(d-2)}{3!} L^3 + \dots \\ &= \sum_{k=0}^{\infty} \frac{\Gamma(k-d)L^k}{\Gamma(-d)\Gamma(k+1)} \end{aligned}$$

where the *gamma function* is defined as

$$\Gamma(g) = \int_0^{\infty} x^{g-1} e^{-x} dx$$

In order to see the role of the differencing parameter d on the stationarity and memory of the process x_t (see chapter 2 for basic concepts), we need to know the covariances of the ARFIMA process x_t . Some algebra provides the corresponding autocovariance functions γ_{τ} and autocorrelation functions ρ_{τ} in the form of

$$\begin{aligned} \gamma_0 &= \frac{\sigma^2 \Gamma(1-2d)}{\Gamma(1-d)\Gamma(1-d)} \\ \gamma_{\tau} &= \frac{\sigma^2 \Gamma(1-2d)\Gamma(\tau+d)}{\Gamma(d)\Gamma(1-d)\Gamma(\tau+1-d)}, \quad \tau = \pm 1, \pm 2, \dots \end{aligned}$$

and

$$\rho_{\tau} = \frac{\Gamma(1-d)\Gamma(\tau+d)}{\Gamma(d)\Gamma(\tau+1-d)}$$

The process x_t is both stationary and invertible if the roots of $\Phi(L)$ and $\Theta(L)$ are outside the unit circle and $d < |0.5|$. But, since as $\tau \rightarrow \infty$, $\rho_{\tau} \propto \tau^{2d-1}$ (Hosking, 1981), the autocorrelations do not have a finite sum, that is, the ARFIMA processes with $0 < d < 0.5$ displays long memory. In other words, these processes exhibit more persistence with the autocorrelation function decaying much slower than for the corresponding I(0) series with similar ARMA parameters.

For $0.5 < d < 1.0$, the process x_t is nonstationary with a noninvertible ARMA representation, since the variance of the process γ_0 does not have a finite sum. However, even though the series is nonstationary, as we can see in Hosking's formula, the autocorrelation function still decays to zero.

This implies that the memory of the process is finite and that given a shock the process tends to revert to its mean, that is, the process is mean-reverting. For $d > 1.0$ the process is not mean-averting, and a shock to the process causes the process to deviate away from its starting point. Thus the memory property of an ARFIMA process depends crucially on the value of d and its autocorrelation functions decays at a much slower rate than those for the corresponding $I(0)$ series.

9.2 Unit root tests against fractional alternatives

The low power of the standard unit root tests under the presence of MA errors and the near unit root case has been well documented (see chapter 4). When the DGP is a fractionally integrated process, what will be the effect of ARFIMA alternatives on the standard unit root tests? Sowell (1990) derived the asymptotic distribution of the Dickey–Fuller test statistics in the case of fractionally integrated processes and conjectured that in finite samples it would be difficult to discriminate between unit root processes $I(1)$ and fractionally integrated processes $I(d)$ with $d < 1$. Diebold and Rudebusch (1991b) confirm this in a Monte Carlo study. Hassler and Wolter (1994) provide analytical arguments as well as Monte Carlo evidence and argue that the ADF test is even less powerful when the alternative is a fractionally integrated process and that the Phillips–Perron tests behave similarly.

In line with Sowell (1990) and Diebold and Rudebusch (1991b), Smith and Yadav (1994) investigate the performance of various unit root tests against ARFIMA alternatives by Monte Carlo experiments. They investigate the power of the ADF test (see chapter 3 for the ADF test), Cochrane’s variance ratio test (Cochrane, 1988, see section 3.8), and different versions of variance ratio tests proposed by Lo and MacKinlay (1988). The ADF test is based on a regression

$$\Delta y_t = \alpha + \rho y_{t-1} + \sum_{j=1}^q \beta_j \Delta y_{t-j} + \varepsilon_t, \quad \varepsilon_t \sim iid(0, \sigma^2)$$

where the lagged terms $\Delta y_{t-1}, \dots, \Delta y_{t-q}$ were included to whiten the serial correlations of errors (see chapter 3 for details). The variance ratio test proposed by Cochrane (1988) uses a weighted average of the sample autocorrelation coefficients (for details, see section 3.8). The null hypothesis of the unit root is that the variance ratio is one. Lo and MacKinlay (1988) suggest using different versions of the variance ratios,

under homoskedastic errors

$$z(q) = \sqrt{T} M_q \left(\frac{2(2q-1)(q-1)}{3q} \right)^{-1/2}$$

where

$$M_q = \frac{(T-1) \sum_{i=q}^T (y_i - y_{i-q} - q\hat{\alpha})^2}{m \sum_{i=1}^T (y_i - y_{i-q} - \hat{\alpha})^2}$$

with $m = q(T-q+1)(1-q/T)$ and under heteroskedastic errors

$$z(q)^* = \frac{\sqrt{T} M_q}{\sqrt{\hat{\delta}}}$$

where

$$\begin{aligned} \hat{\delta} &= \sum_{j=1}^{q-1} \left(\frac{2(q-j)}{q} \right) \\ &\times \frac{T \sum_{i=j+1}^T (y_i - y_{i-1} - \hat{\alpha})^2 (y_{i-j} - y_{i-j-1} - \hat{\alpha})^2}{\sum_{i=1}^T (y_i - y_{i-q} - \hat{\alpha})^2} \end{aligned}$$

Lo and MacKinlay (1988) show that the asymptotic distributions of both variance ratios, $z(q)$ and $z(q)^*$ follow standard normal distributions under the unit root null, $H_0 : z(q)(= z(q)^*) = 1$. Note that since the above asymptotic result also requires $T/q \rightarrow \infty$, Lo and MacKinlay suggested to use the empirical critical values in their paper.

Smith and Yadav (1994) investigate the power of these tests by generating time series with a range of integration parameters, $d = 0.7, 0.8, 0.9, 1, 1.05, 1.1, 1.3$ and $y_0 = 0$. They found that the Cochrane test and the two tests of Lo and MacKinlay perform similarly, with high power against fractional alternatives. The performance of the ADF test was the worst. Whether or not the errors are normally distributed has little effect on the performance of the test. When ARCH errors are introduced, there was a marked fall in power for all the tests. Especially, the Cochrane test and the Lo and MacKinlay (homoskedastic) test show substantial size bias under the presence of ARCH errors.

There have been other studies discussing the power of other unit root tests against fractional alternatives. For instance, Lee and Schmidt (1996) discuss the KPSS test (see chapter 4 for the KPSS test). They show that the KPSS test is consistent against stationary long memory alternatives, that is $I(d)$ processes for $|d| < 0.5$, and, hence, it can be used to distinguish short memory and long memory stationary processes.

Their results, however, show that a rather large sample size (e.g., $T = 1000$) is necessary to distinguish reliably between a long memory process and a short memory process.

Crato and de Lima (1996) consider several MA unit root tests (see chapter 4 for a discussion of MA unit root tests). They characterize the asymptotic behavior under fractionally integrated processes (as Sowell did for the DF test). By means of simulation they also compared the finite sample power of the MA unit root tests and contrast them with AR unit root tests. Their conclusion is that, overall, overdifferencing tests (MA unit root tests) performed worse than underdifferencing tests (AR unit root tests). However, the MA unit root tests seemed to recognize better the nonstationarity of ARFIMA models with $d > 0.5$.

There is as yet no discussion of the power of the modified PP tests and the DF-GLS test (discussed in chapter 4) against fractional alternatives. The studies we have discussed consider the sensitivity of the unit root tests to fractional alternatives. They do not derive new tests in the presence of fractional alternatives. This is done in the papers by Robinson (1994) and Gil-Alana and Robinson (1997). We shall not go into the details of these tests here except to note that this is the proper approach to fractional unit roots rather than the investigation of how sensitive the existing unit root tests are to fractional alternatives. (These papers are not referred to in the survey paper by Baillie, 1996.)

9.3 Estimation of ARFIMA models

The estimation procedures suggested for ARFIMA models fall into three categories:

- (i) two step procedures (Geweke and Porter-Hudak, 1983),
- (ii) approximate ML (Li and McLeod, 1986 and Fox and Taqqu, 1986),
- (iii) exact ML (Sowell, 1992a).

For the asymptotic properties of the exact and approximate ML estimates, see Beran (1994a, 1994b).

Geweke and Poter-Hudak (1983, hereafter denoted by GPH) proposed a semi-nonparametric procedure to estimate a differencing parameter d . The procedure is a two-step procedure. First estimate d . Then use this and estimate the AR and MA components. This two-step procedure has been widely used in practice. GPH (1983) showed that for frequencies near zero, d can be consistently estimated from the least squares

regression

$$\ln(I(w_j)) = c - d \ln(4 \sin^2(w_j/2)) + \eta_j, \quad j = 1, \dots, n$$

where $w_j = 2\pi j/T$, $n = g(T) < T$, and $I(w_j)$ is the periodogram of X at frequency w_j defined by

$$I(w) = \frac{1}{2\pi T} \left| \sum_{t=1}^T e^{itw} (X_t - \bar{X}) \right|^2 \quad (9.1)$$

With a proper choice of n , the asymptotic distribution of $\hat{d}$ depends on neither the order of the ARMA component nor the distribution of the error term of the ARFIMA process. It is suggested to set $n = \sqrt{T}$ and to use the known variance of η_j , $\pi^2/6$, to compute the estimated variance of $\hat{d}$.

The joint ML estimation procedures of the order of fractional integration and the ARMA parameters were proposed in a time-domain context by Li and McLeod (1986) and Sowell (1992a) and in a frequency-domain context by Fox and Taquq (1986). Assuming normality, the likelihood function of the ARFIMA model is given by

$$L = (2\pi)^{-T/2} |\Sigma|^{-1/2} \exp(-X' \Sigma^{-1} X/2) \quad (9.2)$$

where Σ is the $T \times T$ covariance matrix of X and is a function of d , θ_i s, and ϕ_i s, and σ^2 . Li and McLeod (1986) showed that the ML estimator, which is obtained by maximizing (9.2) with respect to the parameter vector $\beta = (d, \theta_i, \phi_i)$ and σ^2 , is consistent and asymptotically normal.

Fox and Taquq (1986) suggested a frequency-domain method to estimate ARFIMA models. The frequency-domain estimators can be obtained by minimizing the following variance with respect to the parameter vector $\beta = (d, \theta_i, \phi_i)$

$$\sigma_T^2(\beta) = \sum_{k=1}^{T-1} I\left(\frac{2\pi k}{T}\right) \left[f\left(\frac{2\pi k}{T}, \beta\right) \right]^{-1}$$

where $I(\lambda)$ is the periodogram defined in (9.1) and $f(\lambda, \beta)$ is proportional to the spectral density of X at frequency λ . It can be shown that, for ARFIMA models

$$f(\lambda, \beta) = |1 - e^{-i\lambda}|^{-2d} \frac{|\Theta(e^{-i\lambda})|^2}{|\Phi(e^{-i\lambda})|^2}$$

Fox and Taquq (1986) showed that the distribution of the frequency-domain estimator $\hat{\beta}$ can be approximated by

$$\sqrt{T}(\hat{\beta} - \beta) \rightarrow N(0, Q)$$

where

$$Q = 2\sigma_T^2(\hat{\beta}) \left(\frac{\partial^2 \sigma_T^2(\beta)}{\partial \beta \partial \beta'} \right)^{-1}$$

The properties of these estimators were discussed by Yajima (1985, 1988).

Diebold and Rudebusch (1989) applied the two-step estimation procedure of GPH to US macroeconomic series and found post-war real output series have fractional roots. Sowell (1992b) applied the ML procedure to the same data and showed that the two-step estimation procedure in Diebold and Rudebusch (1989) might bias upwards the estimate of the fractional differencing parameter.

9.4 Estimation of fractionally cointegrated models

Fractionally cointegrated models are discussed in Dueker and Startz (1995) and Baillie and Bollerslev (1994). In the usual framework of cointegration we have two series y_t and x_t which are $I(1)$ and the linear combination of y_t and x_t which is $I(0)$. What if this linear combination is $I(d)$ where $0 < d < 1$. This is the case considered by Baillie and Bollerslev (1994) which we shall discuss in the next section. But is this the case of fractional cointegration?

Dueker and Startz (1995) argue that a broader definition of fractional cointegration is that there exists an $I(d - b)$ linear combination of $I(d)$ series with $b \geq 0$. Under this definition a continuous measure of cointegration can provide more information than the $I(1)/I(0)$ framework.

Previous studies on fractional cointegration have only considered the fractional integration of the residuals from a cointegrating regression. Thus, they were conditional on $d = 1$. For instance Baillie and Bollerslev (1994) perform a battery of tests that fail to reject unit roots in the nominal exchange rates they study. They estimate a cointegrating vector by OLS and then estimate the fractional order of integration of the cointegrating residuals. They get an estimate of $d - b = .89$ which is significantly less than 1. But this test is conditional on $d = 1$. Cheung and Lai (1993) also test that the hypothesis test ($d - b$) is less than unity, conditional on $d = 1$.

Dueker and Startz argue that it is important to estimate d and $(d - b)$ simultaneously. They consider a bivariate ARFIMA model and adopt the ML procedure for multivariate ARFIMA models described in Sowell (1989) to estimate d and $(d - b)$ simultaneously. They apply this to an example of the long-run relationship between three month and one year treasury bill rates. They find that they can reject both the hypotheses of nocointegration and unit cointegration.

9.5 Empirical relevance of fractional unit roots

Several empirical studies on the dynamic properties of the US macroeconomic time series showed that these series are fractionally cointegrated. See Diebold and Rudebusch (1989, 1991a) on output and consumption, Porter-Hudak (1990) on money supply, and Shea (1991) and Backus and Zin (1993) on interest rates. Many of these studies use the GPH approach. However, as argued by Sowell (1992a) and Smith *et al.* (1993), the ML method is more efficient than the GPH approach. Crato and Rothman (1994) who use the ML approach seem to get higher estimates of d . In the following sections we shall examine briefly the evidence on fractional integration in foreign exchange rates, and in stock prices. For more references and detail, see Baillie (1996).

9.5.1 Exchange rates dynamics

Cheung (1993) estimated ARFIMA (p, d, q) models for five nominal dollar spot rates – British pound (BP), Deutsche mark (DM), Swiss franc (SF), French franc (FF), and Japanese yen (JY) – which are end-of-week exchange rates from January 1974 to December 1989 taken from the *Chicago Mercantile Exchange Yearbooks*. The d parameter estimate of FF is close to 0.5, those of DM, SF, and JY ranged from 0.2 to 0.3, and that of BP is around 0.15 according to various $n = T^{0.45}, T^{0.5}$, and $T^{0.55}$. These estimates show evidence of long memory in exchange rate changes though the result for the BP is marginal. The unit root hypotheses for the exchange rates are rejected in favor of the long memory alternatives. However, impulse response function analysis indicates that persistence in exchange rates can be difficult to detect. Furthermore, the ARFIMA model does not outperform the random walk model in out-of-sample forecast.

Several empirical studies using the post-1973 float exchange rates argue that the behavior of nominal exchange rates is well approximated

by a unit root process – a martingale. On the other hand, Baillie and Bollerslev (1989) provide some evidence supporting a hypothesis of cointegration among exchange rates. These two different empirical evidences imply two incompatible economic perspectives on exchange rate dynamics. The martingale hypothesis comes from an *efficient markets* perspective, as does the cointegration hypothesis from a *common trends* perspective. Cointegration implies that there exist one or more long-run relationships among exchange rate levels, deviations from which tend to be eliminated over time and are therefore useful in predicting future exchange rate changes, whereas nothing is useful for predicting future exchange rate changes if exchange rates evolve as a vector martingale.

Diebold, Gardeazabal, and Yilmaz (1994), hereafter DGY, compare the out-of-sample forecasting performance of VAR models with cointegration and of a martingale for the same data set used by Baillie and Bollerslev (1989). They found that the null hypothesis of no cointegration cannot be rejected for a system of nominal dollar spot exchange rates, 1980-1985, and that no improvements in forecasting performance are obtained from the use of cointegrated VARs. DGY also showed that when a drift is included into the model estimated by Baillie and Bollerslev (1989), the evidence for cointegration vanishes. They argued that evidence of cointegration arises only under the assumption that drift is known to be absent.

Baillie and Bollerslev (1994) in their reply to DGY (1994) agreed with the findings of DGY that on allowing for an intercept in the cointegrating regression, the test statistics do not reject the null hypothesis of no cointegration. On the other hand, they found that the process of the error correction term exhibits long memory characteristics with long-term cycles in their autocorrelations. In the original framework of cointegration, the error correction term is presumed to be $I(0)$ with $I(1)$ processes. The idea of cointegration, however, only requires that the error correction term be stationary. BB suggested to consider the error correction term to possibly be fractionally integrated in the sense that the deviations from the cointegrating relationship possesses long memory, according to which the effect of a shock declines at a slower rate than the usual exponential decay associated with the autocorrelation functions for the class of covariance stationary and invertible ARMA processes. Because of the long memory of the error correction term, they argued that any improvement in forecasting errors may only be apparent several years into the future.

9.5.2 Long memory in stock prices

Long memory in stock prices was discussed first in Lo and Mckinlay (1988) and Lo (1991). The paper by Lee and Robinson (1996) contains a detailed review of these and other studies on long memory in stock prices and provides a semi-parametric estimation method.

9.6 Summary and conclusions

In this chapter we presented a brief overview of fractional unit roots. Our purpose has been to emphasize the major points. Work on fractional unit roots is exploding as is evident from the long survey by Baillie (1996). This work falls in the following categories:

- (i) To investigate the sensitivity of the different unit root tests to fractional alternatives. This work is not very useful as most of the tests have been found to be sensitive. What is more important is development of new tests that can be used in the presence of fractional roots. Examples are Robinson (1994) and Gil-Alana and Robinson (1997). More work is needed in this direction.
- (ii) Estimation of ARMA models with fractional roots (ARFIMA models). Here a majority of applications have used the Geweke and Poter-Hudak (GPH) two-step method. Recent work has shown that Sowell's ML method is the preferred choice. More work is needed in this direction.
- (iii) Estimation of fractionally cointegrated models needs further study. The paper by Dueker and Startz (1995) is a beginning. Until now fractional cointegration has been confined to an examination of fractional integration of the residuals from a cointegrating regression.
- (iv) There is as yet not much discussion of the forecasting performance of fractionally integrated (and cointegrated) models.
- (v) Koop *et al.* (1997) present a Bayesian analysis of the ARFIMA model. The advantage of the Bayesian approach is that in prediction problems, the estimation uncertainty is taken into account. Also, in the presentation of impulse response functions we can plot the whole density instead of just presenting a point estimate and its standard error.

References

- Backus, D.K. and S.E. Zin (1993), "Long-Memory Inflation Uncertainty: Evidence from the Term Structure of Interest Rates," *Journal of Money, Credit, and Banking*, 25, 681–700.
- Baillie, R.T. (1996), "Long Memory Processes and Fractional Integration in Econometrics," *Journal of Econometrics*, 73, 5–59.
- Baillie, R.T. and T. Bollerslev (1989), "Common Stochastic Trends in a System of Exchange Rates," *Journal of Finance*, 44, 167–181.
- (1994), "Cointegration, Fractional Cointegration, and Exchange Rate Dynamics," *Journal of Finance*, 49, 737–745.
- Beran, J. (1994a), *Statistics for Long Memory Processes*, Chapman and Hall, New York.
- (1994b), "Maximum Likelihood Estimation of the Differencing Parameter for Invertible Short and Long Memory Autoregressive Integrated Moving Average Models," *Journal of Royal Statistical Society, B*, 57, 659–672.
- Cheung, Y. (1993), "Long Memory in Foreign Exchange Rates," *Journal of Business and Economic Statistics*, 11, 93–101.
- Cheung, Y. and K. Lai (1993), "A Fractional Cointegration Analysis of Purchasing Power Parity," *Journal of Business and Economic Statistics*, 11, 103–112.
- Cochrane, J.H. (1988), "How Big is the Random Walk in GNP," *Journal of Political Economy*, 96, 893–920.
- Crato, N. and P.J.F. de Lima (1996), "On the Power of Underdifferencing and Overdifferencing Tests Against Nearly Non-stationary Alternatives," Manuscript, Johns Hopkins University.
- Crato, N. and P. Rothman (1994), "Fractional Integration Analysis of Long-Run Behavior for US Macroeconomic Time Series," *Economics Letters*, 45, 287–291.
- Diebold, F.X. and G. Rudebusch (1989), "Long Memory and Persistence in Aggregate Output," *Journal of Monetary Economics*, 24, 189–209.
- (1991a), "Is Consumption Too Smooth? Long Memory and the Deaton Paradox," *Review of Economics and Statistics*, 74, 1–9.
- (1991b), "On the Power of Dickey–Fuller Tests Against Fractional Alternatives," *Economics Letters*, 35, 155–160.
- Diebold, F.X., J. Gardeazabal, and K. Yilmaz (1994), "On Cointegration and Exchange Rate Dynamics," *Journal of Finance*, 49, 727–735.

- Dueker, M. and R. Startz (1995), "Maximum Likelihood Estimation of Fractional Cointegration with an Application to the Short End of the Yield Curve," Working paper, October 1995, University of Washington.
- Fox, R. and M.S. Taquq (1986), "Large Sample Properties of Parameter Estimates for Strongly Dependent Stationary Gaussian Time Series," *Annals of Statistics*, 14, 517–32.
- Geweke, J. and S. Porter-Hudak (1983), "The Estimation and Application of Long Memory Time Series Models," *Journal of Time Series Analysis*, 4, 221–38.
- Gil-Alana, L.A. and P.M. Robinson (1997), "Testing of Unit Root and Other Non-stationary Hypotheses in Macroeconomic Time Series," *Journal of Econometrics*, 80, 241–268.
- Granger, C.W.J. (1980), "Long Memory Relationships and the Aggregation of Dynamic Models," *Journal of Econometrics*, 14, 227–238.
- Granger, C.W.J. and R. Joyeux (1980), "An Introduction to Long-Memory Models and Fractional Differencing," *Journal of Time Series Analysis*, 1, 15–39.
- Hassler, U. and J. Wolter (1994), "On the Power of Unit Root Tests Against Fractional Alternatives," *Economics Letters*, 45, 1–5.
- Hosking, J.R.M. (1981), "Fractional Differencing," *Biometrika*, 68, 165–176.
- Koop, G., E. Ley, J. Oseiwalski, and M.F.J. Steel (1997), "Bayesian Analysis of Long Memory and Persistence Using ARFIMA Models," *Journal of Econometrics*, 76, 149–169.
- Lee, D. and P.M. Robinson (1996), "Semiparametric Exploration of Long Memory in Stock Prices," *Journal of Statistical Planning and Inference*, 50, 155–174.
- Lee, D. and P. Schmidt (1996), "On the Power of the KPSS test of Stationarity Against Fractionally Integrated Alternatives," *Journal of Econometrics*, 73, 285–302.
- Li, W.K. and A.I. McLeod (1986), "Fractional Time Series modeling," *Biometrika*, 73, 217–221.
- Lo, A.W. (1991), "Long-Term Memory in Stock Prices," *Econometrica*, 59, 1279–1314.
- Lo, A.W. and A.C. MacKinlay (1988), "Stock Market Prices Do Not Follow Random Walks: Evidence from a Simple Specification Test," *The Review of Financial Studies*, 1, 41–66.

- Porter-Hudak, S. (1990), "An Application of the Seasonal Fractional Difference Model to the Monetary Aggregates," *Journal of American Statistical Association*, 85, 338–344.
- Robinson, P.M. (1994), "Efficient Tests of Non-stationary Hypotheses," *Journal of American Statistical Association*, 89, 1420–1437.
- Shea, G.S. (1991), "Uncertainty and Implied Variance Bounds in Long-Memory Models of the Interest Rate Term Structure," *Empirical Economics*, 16, 287–312.
- Smith, A.A., F. Sowell, and S.E. Zin (1993), "Fractional Integration with Drift: Estimation in Small Samples," Mimeo, Carnegie Mellon University.
- Smith, J. and S. Yadav (1994), "The Size and the Power of Unit Root Tests Against Fractional Alternatives: A Monte Carlo Investigation," Discussion paper No. 419, University of Warwick.
- Sowell, F. (1989), "Maximum Likelihood Estimation of Fractionally Integrated Time Series Models," unpublished manuscript, Carnegie-Mellon University.
- (1990), "The Fractional Unit Root Distribution," *Econometrica*, 58, 495–506.
- (1992a), "Maximum Likelihood Estimation of Stationary Univariate Fractionally Integrated Time Series Models," *Journal of Econometrics*, 53, 165–188.
- (1992b), "Modeling Long-Run Behavior with the Fractional ARIMA Model," *Journal of Monetary Economics*, 29, 277–302.
- Yajima, Y. (1985), "On the Estimation of Long Memory Time Series Models," *Australian Journal of Statistics*, 27, 303–320.
- (1988), "On the Estimation of a Regression Model with Long Memory Stationary Errors," *Annals of Statistics*, 16, 791–807.

10

Small sample inference: bootstrap methods

10.1 Introduction

The inferential procedures that we discussed in the previous chapters are all based on asymptotic theory. The Monte Carlo results presented in chapter 5 (section 5.7) throw light on the small sample behavior of the different estimation methods, but once an estimation method is chosen, there is still the question of appropriate inference on the parameters estimated. The relevant procedures for this are again asymptotic.

Some methods for obtaining small sample results analytically like Edgeworth expansions involve a lot of tedious algebra and are also applicable only in some special cases. The bootstrap method initiated by Efron (1979) provides a viable alternative. Another alternative is to use the Bayesian methods (discussed in chapter 8) but they are based on a different philosophy. Reviews of the bootstrap methods discussed here can be found in Jeong and Maddala (1993), Vinod (1993), Li and Maddala (1996), and Horowitz (1997).

10.2 A review of the bootstrap approach

The bootstrap method is a resampling method. Several resampling methods were in use earlier but they were disparate. Efron made the resampling method a widely applicable technique. For a history of the resampling approach going back to early papers by Barnard (1963) and Hartigan (1969), see Hall (1992).

Let $(y_1, y_2, \dots, y_n)$ be a random sample from a distribution characterized by a parameter θ . Inference about θ will be based on a statistic T . The basic bootstrap approach consists of drawing repeated samples (with replacement) of size m (which may or may not be equal

to n , although it usually is) from $(y_1, y_2, \dots, y_n)$. Call this sample $(y_1^*, y_2^*, \dots, y_m^*)$. This is the bootstrap sample. We do this B times. B is the number of bootstrap replications. For each bootstrap sample we compute the statistic T . Call this T^* . The distribution of T^* is known as the bootstrap distribution of T . We use this bootstrap distribution to make inferences about θ . Under some circumstances (to be described later) the bootstrap distribution enables us to make more accurate inferences than the asymptotic distribution of T . The bootstrap method described here is the simplest one that is valid for *iid* observations. Needless to say that when the *iid* assumption is not satisfied this method needs to be modified. These modifications in the context of time series models are described later.

Early work on the application of bootstrap methods merely consisted of using the bootstrap distribution to get standard errors. The standard error of T^* was used as a better estimate of the small sample standard error of T than that given by the asymptotic distribution of T . In some complicated models the derivation of asymptotic standard errors (*SEs*) is extremely complicated and in this case the bootstrap procedure for computation of the *SEs* were the only ones available. See Veall (1988) for this sort of argument. However, as remarked by Hartigan (1986, p. 76) in these cases one should use the bootstrap standard error to construct a pivotal statistic and then bootstrap this again for the construction of confidence intervals or testing of hypotheses. In those cases where the asymptotic *SEs* were available, the bootstrap *SEs* were justified as a finite sample approximation. However, the standard error is of interest only if the small sample distribution of T is known to be normal. In most econometric applications this is not the case. Thus, using the bootstrap distribution to get standard errors in this situation is useless. Confidence interval estimation and hypothesis testing can be based directly on the bootstrap distribution. For many early applications on the use of just bootstrap standard errors in econometrics, see Jeong and Maddala (1993, p. 580). It is not necessary to quote these misapplications here.

Even if the asymptotic and bootstrap standard errors are the same in any given example, confidence interval statements from the bootstrap distribution will be different from the confidence interval based on the asymptotic distribution if the bootstrap distribution is skewed. For the autoregressive model $y_t = \rho y_{t-1} + \varepsilon_t$ based on Wolfer's sunspot numbers for 1770–1889, Efron and Tibshirani (1986) got $\hat{\rho} = 0.815$ with asymptotic *SE* = 0.053. The bootstrap *SE* based on 1,000 replications was 0.055 agreeing nicely with the asymptotic results. However, the boot-

strap distribution was skewed to the left. We shall now list some results for confidence interval estimation, hypothesis testing, and methods of generation of bootstrap samples.

10.2.1 Confidence intervals

Many of the results on the properties of bootstrap confidence intervals, that we shall quote now, have been derived using Edgeworth expansions. The use of Edgeworth expansions to study the properties of bootstrap methods started with the papers by Singh (1981) and Bickel and Freedman (1980). They considered the distribution of the sample mean with known variance. For the case of unknown variance and the distribution of the studentized mean, see Abramovich and Singh (1985); see Hall (1992, p. 151) for a history of these methods. As Hall points out (in appendix V) there are some limitations to the results obtained from Edgeworth expansions. However, many of the conclusions that have been derived from the Edgeworth expansions have been substantiated in the Monte Carlo studies on models of interest to econometricians. Hence, we shall state the main results.

Suppose $\hat{\theta}$ is a consistent estimator for θ , and $\hat{\theta}^*$ is the bootstrap estimator of θ . We consider the following methods to construct confidence intervals:

- (i) the asymptotic distribution of $\hat{\theta}$.

The two-sided confidence interval is $\hat{\theta} \pm z_{\alpha} SE(\hat{\theta})$ where $SE(\hat{\theta})$ is the asymptotic standard error of $\hat{\theta}$ and z_{α} is the $(100 - \alpha)$ percentile from the standard normal distribution. This interval is a symmetric interval.

- (ii) Use the bootstrap distribution of $\hat{\theta}^*$.

The two-sided $(100 - 2\alpha)$ confidence interval for θ is $(\hat{\theta} - z_{1-\alpha}^*, \hat{\theta} + z_{\alpha}^*)$ where z_{α}^* is the 100α percentile of the distribution of $\hat{\theta}^* - \hat{\theta}$. This is a two-sided equal-tailed interval. It is often nonsymmetric. This is called the *percentile method*. The nominal coverage of this interval is $(100 - 2\alpha)$ and the difference between the actual coverage and nominal coverage is called the *coverage error*. To improve on the coverage error, Efron (1987) suggested two other percentile intervals (see (iii) and (iv) below).

- (iii) Bias-corrected (*BC'*) percentile interval.

- (iv) Accelerated bias-corrected (BC_a) percentile interval.
In addition there are the following other methods.
- (v) The *percentile-t* or *bootstrap-t* methods (see Hall, 1988b, 1992).
This is the percentile method based on the bootstrap distribution of the t -statistic

$$t = \frac{\sqrt{n}(\hat{\theta} - \theta)}{s}$$

where s is a $\sqrt{n}$ consistent estimator of the standard error of $\sqrt{n}(\hat{\theta} - \theta)$. This procedure is often referred to as *studentization*. t is said to be *asymptotically pivotal*. A pivotal statistic is one whose distribution is independent of the true parameter θ . Hartzigan (1986) stressed the importance of using a pivotal statistic (see also Beran, 1987, 1988).

The bootstrap- t method requires a stable estimate of the standard error while the BC_a method involves calculation of the so-called *acceleration constant* which depends on an estimate of skewness. In some cases like the Johansen procedure for cointegrated systems, the bootstrap- t has been found to perform poorly. This is because of the unstable variance of the Johansen estimator. On the other hand the bootstrap- t is more often used than the BC_a in econometric work because it is easier to compute. However, Efron suggests (in his comment on Maddala and Li, 1996) that the BC_a is not necessarily more difficult. A program for computing the BC_a interval described in the appendix of Efron and Tibshirani (1993) written in language S can be obtained by sending an e-mail to: statlib@lib.stat.cmu.edu with the message: "send bootstrap.funs from S."

These procedures for the construction of confidence intervals are all reviewed in DiCiccio and Romano (1988) and Hall (1988b, 1992) and we shall not repeat the details. The coverage errors have been shown to be $O(n^{-1/2})$ for (i), (ii), and (iii) and $O(n^{-1})$ for (iv) and (v) (where n is the sample size).

These results apply to coverage errors in each trial. Hall (1988a) shows that under quite general circumstances, the symmetric bootstrap- t confidence interval has coverage error $O(n^{-2})$. It is also proved that symmetric intervals are not necessarily any longer than equal-tailed intervals. In fact, Hall shows that in the case of the slope parameter in a simple regression, symmetric confidence intervals with coverage between 68 percent and 97 percent

tend to be shorter than their equal-tailed counterparts. (For the 99 percent confidence level, the symmetric confidence interval is longer than the two-sided interval.) These results are based on the so-called *small sample asymptotics* (that is Edgeworth expansions) and hold for skew distributions.

The important conclusion that follows from these results is that, using the simple percentile method (ii) cannot be expected to produce an improvement over the asymptotic result (i). Many econometric applications, as we shall see, are based on (ii).

Some particular cases are of interest for us. Hall (1989) shows that in a regression model $y = \alpha + \beta x + \varepsilon$, in the case of slope parameters, β , the bootstrap- t produces two-sided intervals with coverage error $O(n^{-2})$ and one-sided intervals with coverage error $O(n^{-3/2})$. For the constant term α or the conditional mean $(\alpha + \beta x_0)$, the bootstrap- t has coverage error $O(n^{-1})$ only.

(vi) Beran's pivotal method.

This was introduced in Beran (1987) and studied in Beran (1988, 1990). This is an iterative bootstrap method that involves a bootstrap within a bootstrap. Hence it is computationally intensive. This method applied to the symmetrical bootstrap- t leads to a coverage error $O(n^{-3})$ (see Hall, 1988a).

Hall (1988b) argues in favor of the bootstrap- t method in preference to Efron's BC_a . He shows that the difference between the bootstrap- t and the BC_a limits is $O(n^{-3/2})$. But more importantly, the BC_a method involves tedious analytical corrections, which the bootstrap methods are designed to avoid. In econometric work where the models are more complicated than those considered in these papers, the bootstrap- t is easier to use than the BC_a method, and hence is the preferable one. Rilstone (1993) compares the BC_a method with bootstrap- t in a Monte Carlo study and finds that the bootstrap- t performs better.

In addition to these methods there are other methods discussed in DiCiccio and Romano (1990). These are the automatic percentile method, as an alternative to the BC_a that does not require explicit calculation of the acceleration constant, nonparameteric tilting, Tibshirani's (1988) variance stabilized bootstrap- t and a couple of others. The nonparameteric tilting method was introduced by Efron (1981a). These methods, though promising, have not found widespread acceptance.

Kilian (1996) suggests a different bias corrected confidence interval.

He suggests what he calls *bootstrap after bootstrap*. This is motivated as follows: let $\hat{\theta}(x)$ be the initial estimator of θ which we use in generating bootstrap samples. Let the mean of the bootstrap estimator $\hat{\theta}(x^*)$ be denoted by $\tilde{\theta}^*$. Then the bias corrected estimator is

$$\hat{\theta}_{bc}(x) = \hat{\theta}(x) + (\hat{\theta}(x) - \tilde{\theta}^*)$$

Killian's idea is that if we bootstrap $\hat{\theta}_{bc}$ we will get better confidence intervals than if we bootstrap $\hat{\theta}$. Thus, we use the first bootstrap to get bias correction and then another bootstrap to get the confidence interval.

Note that the term bias correction in the literature on bootstrap confidence intervals refers to bias in the coverage probability and not a correction of the bootstrap estimator for bias which is what Killian's method involves. However, Killian shows that his method works very well in his application, compared with the percentile method. More detailed studies are necessary to compare it with Efron's procedure as well as the bootstrap- t .

Several objections have been raised to the bootstrap- t method. These are:

- (i) It produces bad results if the estimate of the variance is poor.
- (ii) The bootstrap- t method is not invariant to transformations.

If it produces a confidence interval $(\hat{a}, \hat{b})$ for θ , and f is a monotone increasing function of θ , then the percentile- t does not, in general, give $[f(\hat{a}), f(\hat{b})]$ as a confidence interval for $f(\theta)$. This is quite troublesome in many econometric applications where Wald tests for nonlinear hypotheses are often used and the Wald test statistic depends on how the nonlinear constraint is formulated. For instance $H_0 : \beta_1 = -\beta_2^{-1}$ and $H_0 : \beta_1\beta_2 = -1$ are equivalent formulations. By contrast, the likelihood-ratio test and Rao's score test (LM test) are invariant to the different parameterizations. Gregory and Veall (1985) show that there are substantial size distortions with the use of the Wald test statistic, and these differ considerably under different formulations. This size distortion problem and the problem of invariance with the Wald statistic have been emphasized by many others. Horowitz and Savin (1992) argue that the bootstrap-based critical values for the Wald test solve the size distortion problem and that in some cases the power of the Wald test is higher than that of the LR test. Hence, the invariance of the Wald test need not be much of a concern. Jeong and Maddala (1996) also correct the size distortions in the Wald statistic using bootstrap methods.

10.2.2 Hypothesis testing

In the statistical literature, hypothesis testing using critical values from bootstrap distributions has received less attention than construction of confidence intervals. (In the econometric literature, it is perhaps the reverse.) There are a few exceptions. Beran (1988) shows that if the asymptotic distribution of the test statistic is pivotal under the null, then under some regularity conditions the size of the bootstrap test has an error of a smaller order than the size of the asymptotic theory test. Hinkley (1988) discusses briefly bootstrap tests of significance and this is followed by a more detailed discussion in Hinkley (1989). Hall and Wilson (1991) and Hall (1992) provide two general guidelines in hypothesis testing, which we shall discuss below. These guidelines have been violated in econometric practice but with good reasons. In the econometric literature the importance of using pivotal statistics in hypothesis testing is discussed in Horowitz (1994).

The familiar duality between hypothesis testing and construction of confidence intervals is maintained under bootstrap methodology as well. Given this duality, what we said earlier regarding the importance of (asymptotically) pivotal statistics in the construction of confidence intervals and the orders of approximation for the different methods applies to hypothesis testing as well. Thus, it is important to apply significance tests using (asymptotically) pivotal statistics. Otherwise, one cannot expect much of an improvement over the asymptotic results. This is confirmed, for instance, by the conflicting results in Veall (1986) who finds that the performance of bootstrap is no better than that of asymptotic theory, and Rayner (1991) and Rilstone (1993) who come to the opposite conclusion. The former bootstraps the coefficients $\hat{\beta}$ and the latter two bootstrap $\hat{\beta}/SE(\hat{\beta})$ which is pivotal. Other examples of this will be provided in the following sections.

Two important issues concerning hypothesis testing using bootstrap methods relate to the questions about

- (i) what test statistic to bootstrap and
- (ii) how to generate the bootstrap samples.

We have said that it is important to use a pivotal (or asymptotically pivotal) statistic in hypothesis tests. But there is another issue that has been brought up by Hall and Wilson (1991). Suppose that we want to test the hypothesis $H_0 : \theta = \theta_0$ versus $H_1 : \theta \neq \theta_0$. Given an estimator $\hat{\theta}$ of θ , the usual test procedure would be based on $T = \hat{\theta} - \theta_0$ and the

significance level and p -values are obtained from the distribution of T under H_0 . A direct application of the bootstrap procedure would suggest using the bootstrap distribution of $T^* = \hat{\theta}^* - \theta_0$ instead of the distribution of T ($\hat{\theta}^*$ is the value of $\hat{\theta}$ from the bootstrap sample). However, Hall (1992, section 3.12) discusses the bad behavior of the power of this test arguing that T^* does not approximate the null hypothesis when the sample comes from a distribution with parameter θ far away from θ_0 . Hall and Wilson, therefore consider another bootstrap procedure based on the empirical distribution of $(\hat{\theta}^* - \hat{\theta})$. They compare this with the test procedure based on T and T^* .

Hall and Wilson propose two guidelines for hypothesis testing. The first suggests using the bootstrap distribution of $(\hat{\theta}^* - \hat{\theta})$ but not $(\hat{\theta}^* - \theta_0)$. The second guideline suggests using a properly studentized statistic, that is $(\hat{\theta}^* - \hat{\theta})/\hat{\sigma}^*$ but not $(\hat{\theta}^* - \hat{\theta})/\hat{\sigma}$ or $(\hat{\theta}^* - \hat{\theta})$, where $\hat{\sigma}^*$ is the estimate of $\hat{\sigma}$ from the bootstrap sample.

Van Giersbergen and Kiviet (1994) discuss these two rules in the context of hypothesis testing in regression models. To simplify the exposition, we shall discuss the case of a simple regression

$$y = \beta x + \varepsilon, \quad \varepsilon \sim iid(0, \sigma^2)$$

Let $\hat{\beta}$ and $\hat{\sigma}$ be the OLS estimators of β and σ respectively and $\hat{\varepsilon}$ the OLS residuals. Let $\hat{\varepsilon}^*$ be the bootstrap residuals obtained by resampling $\hat{\varepsilon}$. The null hypothesis to be tested is $H_0 : \beta = \beta_0$ versus $H_1 : \beta \neq \beta_0$.

Consider two sampling schemes for the generation of the bootstrap samples

$$\begin{aligned} S_1 : y^* &= \hat{\beta}x + \varepsilon^* \\ S_2 : y^* &= \beta_0x + \varepsilon^* \end{aligned}$$

Both use ε^* but they differ the way y^* is generated. For each sampling scheme consider two t -statistics

$$\begin{aligned} T_1 : T(\hat{\beta}) &= (\hat{\beta}^* - \hat{\beta})/\hat{\sigma}^* \\ T_2 : T(\beta_0) &= (\hat{\beta}^* - \beta_0)/\hat{\sigma}^* \end{aligned}$$

Thus four versions of the t -statistic can be defined. Hall and Wilson consider sampling scheme S_1 only and suggest using T_1 only. They do not consider sampling scheme S_2 which is the appropriate one for statistic T_2 . Van Giersbergen and Kiviet suggest, on the basis of a Monte Carlo study of an AR(1) model, the use of T_2 under sampling scheme S_2 in preference to T_1 under S_1 . The main conclusions of the paper are:

- (i) Inference based on T_2 under S_1 does not just have low power but in fact has size close to zero. Similarly T_1 under S_2 does not work and should not be used. The resampling scheme should mimic the null distribution of the test statistic to be bootstrapped.
- (ii) Using T_1 under S_1 and T_2 under S_2 are equivalent in nondynamic models. This equivalence extends to the multiparameter case if one bootstraps the appropriate F -statistic. However, in dynamic models this equivalence breaks down in finite samples. The Monte Carlo results suggest that it is better to use T_2 under S_2 .
- (iii) The limiting distributions of T_1 under S_1 and T_2 under S_2 are identical even with dynamic models. The conclusion that T_2 under S_2 is better is based on small sample performance.

One can also propose another resampling scheme

$$S_3 : y^* = \beta_0 x + \varepsilon_0^*$$

where ε_0^* is the bootstrap sample from $\varepsilon_0 = y - \beta_0 x$ (after centering). Note that in both S_1 and S_2 we use resampling based on the OLS residuals $\hat{\varepsilon}$. If the null $H_0 : \beta = \beta_0$ is true but the OLS estimator $\hat{\beta}$ gives a value of β far from β_0 , the empirical distribution of the residuals will suffer from a poor approximation of the distribution of the errors under the null. The intuition behind S_3 is as follows. If the null hypothesis is true, $\varepsilon_0 = y - \beta_0 x$ is exactly the true distribution of the regression errors. Hypothesis testing based on this will give (approximately) the correct test size. If the null is not true, then ε_0 is different from the true distribution of the errors. Hypothesis testing will give proper power depending on how far the null is away from the true value of β . Thus, using T_2 under S_3 is better than using T_1 under S_1 or T_2 under S_2 which are identical as shown in van Giersbergen and Kiviet. The idea of using restricted regression errors for resampling has been used in Li and Maddala (1997) and Nankervis and Savin (1996).

As we shall discuss in section 10.4, for the unit root model, test statistic T_2 under sampling scheme S_2 or S_3 are the only ones that can be justified on asymptotic grounds. It is not appropriate to use T_1 under S_1 .

10.2.3 Generation of bootstrap samples: residual-based versus direct methods

In the preceding section we discussed some issues about the generation of bootstrap samples and how this is related to the hypothesis test under consideration. The resampling methods were all residual-based methods. We shall now discuss issues related to direct methods of resampling data.

For the case of random regressors (which he calls the *correlation model* as opposed to the *regression model*) Freedman (1981) suggests resampling the pair (y, X) , which have a joint distribution with $E(y|X) = X\beta$. In this model the pairs (y_i, X_i) are assumed to be *iid*. Given a sample of size n we can compute $\hat{\beta}$, the least squares estimator of β . Denote this by $\hat{\beta}(n)$. Then $\sqrt{n}(\hat{\beta}(n) - \beta)$ is asymptotically normal with mean 0 and a certain covariance matrix which we shall denote by Σ . Let m be the size of the bootstrap sample and denote the bootstrap estimator by $\hat{\beta}^*(m)$. Freedman (1981, p. 1226) shows that under some conditions, as $m \rightarrow \infty$ and $n \rightarrow \infty$

$$\sqrt{m}(\hat{\beta}^*(m) - \hat{\beta}(n)) \rightarrow N(0, \Sigma)$$

What this result shows is that the bootstrap distribution based on direct sampling is useful as an approximation of the asymptotic distribution of $\sqrt{n}(\hat{\beta}(n) - \beta)$. In cases where the asymptotic distribution can be derived, this result is not of much use except as a consolation that using the bootstrap is alright. In this case the bootstrap method can not be expected to improve upon the asymptotic result without using a pivotal method. However, in the case of heteroskedasticity of an unknown form, this direct method even without pivoting is useful. In fact, in this case, even if the model is the usual regression model with fixed X s (as opposed to the model with random X s) the direct method of sampling (y_i, X_i) is useful because it gives an estimate of the correct covariance matrix Σ (see Jeong and Maddala, 1993, pp. 577–578). On the other hand, if one uses White's heteroskedasticity-consistent covariance matrix estimator, one can obtain asymptotically valid pivotal statistics in the presence of heteroskedasticity of an unknown form. In practice, this would be the preferred approach. An alternative method to handle heteroskedasticity of an unknown form is the *wild bootstrap* suggested by Hardle and Mammen (1990).

Efron (1981b) uses the direct method of resampling the data in a problem involving censored data. He bootstraps the data (x_i, d_i) where $d_i = 1$ if x_i is not censored and $d_i = 0$ if x_i is censored. His model

does not have covariates. Some applications of this direct method to censored regression models (that is models with covariates) are reviewed in Jeong and Maddala (1993, pp. 591-594). The problem here is one of nonlinearity of an unknown form. But the basic question is: what do you do after generating the bootstrap sample? Suppose you have the data on a latent variable y_i which is observed as a dummy variable $d_i (= 1 \text{ or } 0)$ and a set of covariates X_i . We draw the bootstrap sample (d_i^*, X_i^*) . What do we do next? If we are going to estimate a logit model, then we can also estimate the logit model with the actual data, compute the generalized residuals, and resample the generalized residuals to generate bootstrap samples. In this case the direct bootstrap method does not make use of all the information as does the method based on generalized residuals. It is only if some general semiparametric method of estimation is used that the direct bootstrap method would be useful. We cannot go into a detailed discussion of the appropriate bootstrap methods for the censored regression model here, but we will argue that the above considerations carry over to time series models as well.

The direct method of sampling the data, rather than the residuals, has not gone unnoticed in econometrics. The earliest use is in Kiviet (1984) which we will discuss later. Veall (1987, p. 205) considers the direct bootstrap approach but rejects it on grounds that it does not embody all the information used in the residual-based approach. Li and Maddala (1996) actually use the direct sampling approach combined with the moving block method (discussed in section 10.4) to data vectors (y_i, x_i) in cointegrating regression models and find its performance worse than that of the residual-based bootstrap. Here we shall review some problems with the use of the direct bootstrap approach in econometric applications once we move out of the framework of the usual regression model. These problems have not received sufficient attention.

Consider first the lagged dependent variable model. Freedman and Peters (1984) consider a q variable model and introduce the *recursive bootstrap* method. The model is

$$Y_t = Y_{t-1}B + X_tC + \varepsilon_t, \quad t = 1, 2, \dots, T$$

where Y_t and X_t are $(1 \times q)$ and $(1 \times p)$ matrices, respectively, and B and C are $(q \times q)$ and $(p \times q)$ matrices, respectively. The ε_t are assumed to be $iid(0, \Sigma)$ and Y_0 is assumed to be known. The $\hat{B}$, $\hat{C}$ are first estimated using a system method (that takes care of contemporaneous correlation among the errors ε_t), and the residuals $\hat{\varepsilon}_t = Y_t - Y_{t-1}\hat{B} - X_t\hat{C}$ are calculated. Bootstrap samples ε_t^* are generated. Since the entire vector

of residuals is resampled, this preserves the contemporaneous correlation structure of the errors. The bootstrap sample Y_t^* is generated in a recursive fashion, assuming X_t , and Y_0 are given, using the equation

$$Y_t^* = Y_{t-1}^* \hat{B} + X_t \hat{C} + \varepsilon_t^*$$

Now re-estimate the model using $(Y_t^*, X_t, \text{ and } Y_0)$. This method was used in an empirical application with $q = 10$ and $T = 18$. It is shown that the conventional asymptotic standard errors are substantially (by about 40 percent–50 percent) below the bootstrap standard errors. The empirical results, of course, merely say that the asymptotic and bootstrap standard errors are different. There is no way of telling without a Monte Carlo study how biased each is. Freedman and Peters argue that the problem lies in the estimation of the covariance matrix Σ in the conventional procedure, which can be poor with such a limited sample. It is also worth noting that Freedman and Peters concentrated on the standard errors and did not consider bootstrapping the students t -statistic, which on asymptotic grounds, is expected to give better confidence intervals, as argued earlier.

Suppose we abstract from the multiple equation issue and concentrate on the estimation of a simple dynamic equation

$$y_t = \beta_0 + \beta_1 y_{t-1} + \beta_2 x_t + \beta_3 x_{t-1} + \varepsilon_t \quad (10.1)$$

For this model Kiviet (1984) presents the results of a Monte Carlo study and finds that the bootstrap does not improve on conventional asymptotic inference. Even though this paper is out of date (and was never published) there are a few points in this paper worth noting which are useful for further work. Kiviet considered two error structures: one in which ε_t are $IN(0,1)$ and the other in which ε_t are $IL(0,1)$ where L is the Laplace distribution. Also the problem may have been caused by the use of the percentile method which is known to give no improvement over asymptotic results. The bootstrap- t bootstrap as discussed earlier, should do better. One suspects from Kiviet's Monte Carlo study that the substantial discrepancies observed by Freedman and Peters between the asymptotic and bootstrap standard errors may have been caused by the high dimensionality of their system ($q = 10$). However, one needs to investigate this by a Monte Carlo study.

It is interesting to note that Kiviet considers the recursive method suggested by Freedman and Peters (1984) but he also considers a direct resampling scheme whereby the data vector $(y_t, y_{t-1}, x_t, x_{t-1})$ is bootstrapped. This did not give any improvement over asymptotic results

either. What this suggests is that the important issue is not residual sampling versus direct sampling of data but one of percentile versus bootstrap- t . In fact this dynamic model was analyzed by van Giersbergen and Kiviet (1993) using the bootstrap- t with substantial improvement over the asymptotic results (although an *iterative* bootstrap- t that they suggest, did better), as far as the confidence interval construction is concerned.

There are, in fact, more problems with the direct method in this case. Suppose the errors ε_t in (10.1) are AR(1). Then there is a serial correlation bias in the OLS parameters in (10.1). Suppose that the information that ε_t in (10.1) are AR(1) is used in the estimation of the model from the bootstrap sample, then this information is incorporated in the residual method but is not used in the direct method of generation of the bootstrap sample by resampling the data. Thus, we feel that the direct method ignores some important information utilized in the residual-based sampling in the generation of bootstrap samples of dynamic models.

Our basic argument is that whatever information about the error term is used in the estimation of the model from the bootstrap samples, should also be used in the generation of the bootstrap samples. This is done in the residual-based bootstrap generation but not in the method of bootstrapping the data. This issue is not important in the case of models with no lagged dependent variables, where one can bootstrap the data and use White's heteroskedasticity-consistent covariance matrix to get asymptotic pivotal statistics and thus get asymptotic refinements using bootstrap methods. But in the case of lagged dependent variables with serially correlated errors, there is also a problem of consistency of the resulting estimators if the order of serial correlation is not properly taken into account.

Turning next to the application of the direct method to the cointegrated regression models, the problems are similar. Suppose that y_t and x_t are I(1) and we have the regression equation

$$y_t = \beta x_t + u_t \quad (10.2)$$

Suppose we bootstrap the data (y_t, x_t) and estimate equation (10.2) by OLS. If u_t is also I(1), then it is well known that equation (10.2) is a spurious regression. But there is no way of knowing this if we use the direct bootstrap method, without first testing whether (10.2) is indeed a cointegration relationship.

Suppose we initially apply cointegration tests to equation (10.2) and

make sure that equation (10.2) is a meaningful regression relationship. This implies that y_t and x_t are $I(1)$ and u_t is $I(0)$. (An $I(0)$ variable is stationary and an $I(1)$ variable is stationary in first differences.) But the direct bootstrap method does not use the information that u_t is $I(0)$. This point is discussed further in section 10.6. In fact, in Li and Maddala (1997) the direct method was used in several Monte Carlo studies of the comparison of the moving block procedure with asymptotic methods and it was found that the performance of the direct method was slightly worse than the one based on bootstrapping the residuals, although the direct method provided an improvement over asymptotic results. It should be noted that Li and Maddala considered bootstrapping a t -statistic in both cases and hence that is not an issue here.

We have reviewed several cases where the direct method of bootstrapping the data has been used. We have also outlined some problems in the use of the direct method as compared to bootstrapping residuals. These problems have not been appreciated in econometric work and it is often thought that the direct method, which is the simplest, is universally applicable.

10.3 The AR(1) model

The simplest AR(1) model is given by

$$y_t = \rho y_{t-1} + \varepsilon_t$$

with $y_0 = 0$, $u_t \sim iid(0, \sigma^2)$ and $-\infty < \rho < \infty$. We can divide ρ into three regions. When $|\rho| < 1$ the process $\{y_t\}$ is stationary. When $|\rho| = 1$ the process is unstable. This is known as the unit root model. When $|\rho| > 1$ the process is explosive. The unit root and explosive cases are discussed in the next section. Here we shall discuss the stationary case. The AR(1) model with intercept is

$$y_t = \alpha + \rho y_{t-1} + \varepsilon_t, \quad \varepsilon_t \sim iid(0, \sigma^2)$$

However, since the distribution of the OLS estimator $\hat{\rho}$ of ρ is invariant to α and σ^2 , we might as well set $\alpha = 0$ and $\sigma^2 = 1$ and consider the model

$$y_t = \rho y_{t-1} + \varepsilon_t, \quad \varepsilon_t \sim iid(0, 1)$$

Dufour (1990) and Andrews (1993) have developed exact inference procedures for the AR(1) parameter but these depend on the normality assumption of the errors. Hence bootstrap methods which are robust to

distributional assumptions of the errors hold promise. The procedure for the generation of the bootstrap samples is the recursive procedure described in the previous section. However, after computing the least squares residuals $\hat{u}_t$ and the bootstrap residuals u_t^* , we use ρ_0 but not $\hat{\rho}$ in generating y_t^* . This is sampling scheme S_2 discussed in section 10.2.2.

The first bootstrap application of this model is by Rayner (1990). He uses the bootstrap- t (or the bootstrap- t method) and is concerned with the problem of testing the hypothesis $H_0 : \rho = \rho_0$. He does not examine the percentile method, but examines the student statistic

$$t = (\hat{\rho} - \rho_0)/SE(\hat{\rho})$$

He considers approximating the distribution of this t -statistic by the bootstrap distribution of t^* which is defined as

$$t^* = (\rho^* - \rho_0)/SE(\rho^*)$$

where ρ^* and $SE(\rho^*)$ are the values of $\hat{\rho}$ and $SE(\hat{\rho})$ computed from the bootstrap sample. This violates the Hall and Wilson rule which says that $\hat{\rho}$ but not ρ_0 should be used in t^* . However, Rayner modifies the sampling rule by using ρ_0 but not $\hat{\rho}$ in generating y^* (see section 10.2.2), that is, he uses the t -statistic T_2 with sampling rule S_2).

Rayner finds that the use of the student- t approximation is not satisfactory, particularly for high values of ρ and that the bootstrap- t performs very well in samples of sizes five to ten, even when mixtures of normal distributions are used for the errors. He also argues that the bootstrap approximates the power well for samples as small as five to ten. Finally, for the starting value y_0 , he says, it is important to use a random starting value (from the equilibrium distribution of y_t).

Rayner's argument that the bootstrap sample generation should be consistent with the null hypothesis being tested is correct (see the discussion in section 10.2.2). So is his use of the bootstrap- t instead of the percentile method. However, his conclusion about the sample sizes (as small as five to ten) at which the performance of the bootstrap is good is very surprising. So is his conclusion regarding the robustness of the procedure to departures from normality with small sample sizes. Subsequent studies have failed to confirm these conclusions.

Van Giersbergen and Kiviet (1994) do a Monte Carlo study with the same model as Rayner. In addition to the bootstrap- t they also consider the usual percentile method and an iterative bootstrap- t . The sampling scheme used was T_1 and S_1 in the case of percentile and bootstrap- t methods and T_2 and S_2 in the case of the iterative bootstrap- t method.

As expected, the percentile method did not perform any better than the asymptotic theory. The bootstrap- t method performed much better but even for a sample size $n = 40$, the bootstrap- t gave a significance level of 0.08 compared to a nominal significance level of 0.05. In their study, the iterative bootstrap- t did better giving a significance level of around 0.05 even for samples of size $n = 20$. As far as robustness to nonnormality is concerned, they investigated a shifted χ^2 distribution (shifted to have mean zero), truncated t -distributions, and *ARCH* disturbances. The performance of all the methods deteriorated, particularly with *ARCH* disturbances, but the iterative bootstrap- t performed best. As mentioned in section 10.2.3, these authors also discuss an AR(1) model with an exogenous regressor and again find the performance of the iterative bootstrap- t the best. Due to space limitation, we shall not discuss their iterative bootstrap- t method. Details of this method can be found in van Giersbergen and Kiviet (1993).

There are a few differences in the way the bootstrap sample is generated in the van Giersbergen–Kiviet study. They condition the sample generation on y_0 , unlike the case in Rayner's study. They also adjust the least squares residuals by centering and scaling (see the description in the previous section).

Nankervis and Savin (1996) also replicate Rayner's study using Rayner's method of treating y_0 . They consider three models: (i) the AR(1) model, (ii) a linear trend model with AR(1) errors, and (iii) the unit root model.

First Nankervis and Savin use the test statistic T_2 with sampling scheme S_3 (see section 2.2) but not S_2 as done by Rayner. Thus, their results are not exactly comparable to Rayner's. They find that the bootstrap- t test has the correct level for a sample size of ten when the error distribution is normal but it suffers substantial size distortions when the errors follow a mixture of normal or lognormal distributions, although the distortions are less than with the usual t -test. However, for samples of size 50, the bootstrap- t test has the correct size even for these nonnormal distributions. Also, the empirical power of the bootstrap- t test is identical to that of the usual t -test assuming normality. Thus, there is no loss of power in using the bootstrap- t test even when the errors have a normal distribution. We shall skip more details, since this is a very restricted model.

In summary, the bootstrap- t test performs well in having the correct size and good power, for samples of size 50 (lower if errors are normal). The results presented by Rayner are not likely to be extended to non-normal errors because the sample sizes he considers are very low (five

to ten). The major differences between the three studies by Rayner, van Giersbergen–Kiviet, and Nankervis–Savin, are in the treatment of y_0 and the way that any information of the form $|\rho| < 1$ is exploited in the construction of confidence regions. Van Giersbergen and Kiviet (1996) extend these methods to bootstrapping a stable AD model.

10.4 Bootstrapping unit root tests

The literature on unit root testing is enormous and it is not possible for us to consider all the issues in bootstrapping the different unit root test procedures. The issues relate, among other things, to devising more powerful tests than the standard Dickey–Fuller test and its extensions, considering stationarity as null versus unit root as null, and using the Bhargava (1986)-type structural approach instead of the Dickey–Fuller type reduced-form approach. The conflict between the structural versus reduced form approaches has also been noted in the Bayesian approach to unit root testing and Peter Phillips argues in favor of the Bhargava approach. For an exposition of Phillips’ arguments and some recomputations, see Zivot (1994).

In spite of its defects, we shall follow the Dickey–Fuller reduced-form approach in what follows. This is because we wish to concentrate on the major issues pointed out in section 10.2 relating to the type of test statistics to use and the type of sampling scheme to be used. There is some confusion in the econometric literature that needs to be cleared up. For instance, Basawa *et al.* (1991a) prove that sampling scheme S_1 is not appropriate in the unit root case. Basawa *et al.* (1991b) use test statistic $n(\hat{\rho}^* - 1)$ with sampling scheme S_3 . Ferretti and Romo (1994) establish the result that test statistic $n(\hat{\rho}^* - 1)$ with sampling scheme S_2 can also be used in bootstrap tests of unit roots. Note that this is similar to the procedure used by Rayner (1990) for the stationary AR(1) model, except that it is not in the pivotal form. We have compared both the Basawa *et al.* (1991b) and Ferretti and Romo schemes of getting bootstrap samples and did not notice any difference, although more detailed investigation of this is under way.

Consider an AR(1) unit root model

$$y_t = \rho y_{t-1} + \varepsilon_t, \quad \rho = 1, \quad t = 1, 2, \dots, T \quad (10.3)$$

where $y_0 = 0$, $\varepsilon_t \sim iid(0, \sigma^2)$. In this case, the OLS estimator $\hat{\rho}$ of ρ is a function of the Wiener processes (for readers’ convenience, the

followings are duplicated from section 3.3.2 of chapter 3)

$$\left(\sum_{t=1}^n y_{t-1}^2 \right)^{1/2} (\hat{\rho} - 1) \Rightarrow \frac{\sigma((W(1))^2 - 1)/2}{(f(W(r))^2 dr)^{1/2}} \quad (10.4)$$

The Dickey–Fuller coefficient test (or K -test) and t -test statistics have the following limiting distributions

$$K = T(\hat{\rho} - 1) \Rightarrow \frac{((W(1))^2 - 1)/2}{(f(W(r))^2 dr)} \quad (10.5)$$

and

$$t_{\hat{\rho}} \Rightarrow \frac{((W(1))^2 - 1)/2}{(f(W(r))^2 dr)^{1/2}} \quad (10.6)$$

To construct the bootstrap test corresponding to (10.5), we start with the OLS estimation of (10.3), compute $\hat{\varepsilon}_t$ and get the bootstrap sample ε_t^* . We now generate y_t^* using ε_t^* and $\hat{\rho}$ by a recursive procedure. What Basawa *et al.* (1991a) show is that the limiting distribution of $(\sum_{t=1}^n y_{t-1}^{*2})^{1/2}(\hat{\rho}^* - \hat{\rho})$ is not (10.4). The limiting distribution of $T(\hat{\rho}^* - \hat{\rho})$ turns out to be random and does not coincide with (10.5) even if the error distribution is assumed to be normal. However, for the sampling scheme S_3 , which is resampling the restricted residuals $(y_t - y_{t-1})$ after centering, Basawa *et al.* (1991b) show that the limiting distribution of $(\sum_{t=1}^n y_{t-1}^{*2})^{1/2}(\hat{\rho}^* - 1)$ is (10.4) and the limiting distribution of $T(\hat{\rho}^* - 1)$ is (10.5).

As an alternative, consider the resampling scheme S_2 , that is the unrestricted OLS regression errors, are used to generate ε_t^* and the pseudo data y_t^* are generated using the null $H_0 : \rho = 1$. The asymptotic validity of this bootstrap method is established in Ferretti and Romo (1994). They show that the limiting distribution of $(\sum_{t=1}^n y_{t-1}^{*2})^{1/2}(\hat{\rho}^* - 1)$ is (10.4) with the resampling scheme S_2 .

It is important to note two points. First Basawa *et al.* (1991a) do not prove the invalidity of sampling scheme S_2 . Thus, the use of S_2 is not in conflict with their results. Second, both sampling schemes S_2 and S_3 are valid in the unit root case provided the test statistic is of the form T_2 , that is use a suitably normalized function of $(\hat{\rho}^* - \rho_0)$.

Turning next to the explosive case $|\rho| > 1$, Basawa *et al.* (1989) also show that if the ε_t are *iin*, conditional on $(y_1, y_2, \dots, y_n)$

$$\hat{\sigma}_n^{-1} \left(\sum_{t=1}^n y_{t-1}^{*2} \right)^{1/2} (\hat{\rho}^* - \hat{\rho}) \Rightarrow N(0, 1)$$

for all sample paths, where

$$\hat{\sigma}_n^2 = n^{-1} \sum_{t=1}^n (y_t - \hat{\rho} y_{t-1})^2$$

Note that it is $\hat{\sigma}_n$ that is used here but not $\hat{\sigma}_n^*$. We have not investigated the issue of why this is valid in the explosive case except to note that this is in conflict with the Hall and Wilson guidelines discussed in section 10.2.

Nankervis and Savin (1996) present simulation results on bootstrap tests for unit roots in the AR(1) model with linear trend

$$y_t = \alpha + \delta t + \rho y_{t-1} + \varepsilon_t$$

under different error distributions. They use sampling scheme S_3 and test statistic T_2 . In general the conclusions regarding sample sizes and empirical significance levels for the unit root case are the same as for the trend-stationary model discussed in section 10.3. The bootstrap- t test had, in general, the same power as the Dickey-Fuller test, although for some nonnormal distributions, its performance was slightly better than that of the Dickey-Fuller test.

Earlier, we stated that we did not find much difference in the form of the bootstrap- t tests under sampling schemes S_2 and S_3 . This was the case for normally distributed errors and for the three models

- (i) $y_t = \rho y_{t-1} + \varepsilon_t, \quad \rho = 1$
- (ii) $y_t = \alpha + \rho y_{t-1} + \varepsilon_t, \quad \alpha = 0, \quad \rho = 1$
- (iii) $y_t = \alpha + \delta t + \rho y_{t-1} + \varepsilon_t, \quad \delta = 0, \quad \rho = 1.$

In all cases the powers for the Dickey-Fuller test as well as two bootstrap- t tests using sampling schemes S_2 and S_3 were about the same and in all cases the powers deteriorated going from model (i) to (ii) and from (ii) to (iii). The models considered here (as well as the models considered in Nankervis and Savin, 1996) are too restrictive to be useful in practice because they are all based on the assumption that ε_t are *iid*. Extensions of the bootstrap approach to cases where there is serial correlation in ε_t (the ADF test) and other unit root tests are under investigation. Schwert (1987) analyzed many US macroeconomic time series and found that, although they appeared to be nonstationary, they contained a significant MA(1) coefficient in their ARIMA specification (see chapter 4). In view of this we do not want to make any definite recommendations on the merits of sampling schemes S_2 and S_3 .

One final point concerns the use of pivotal statistics. There are two Dickey–Fuller tests: the coefficient test (10.5) and the t -test (10.6). When it comes to the bootstrap approach there is again the question of whether to consider the coefficient test or the t -test. We have looked into this issue and found the t -test to be only marginally better, although this conclusion is highly tentative. The case for considering pivotal statistics may not be as strong for the unit root model (as in the stationary models).

The paper by Harris (1992) is the earlier one on bootstrapping unit root tests. Since the Nankervis and Savin paper is more exhaustive, we shall not go through the paper by Harris.

10.5 The moving block bootstrap and extensions

As we discussed in the previous section, application of the residual-based bootstrap methods is straightforward if the error distribution is specified to be an $\text{ARMA}(p, q)$ process with known p and q . However, if the structure of serial correlation is not tractable or is misspecified, the residual-based methods will give inconsistent estimates (if lagged dependent variables are present in the system). Other approaches which do not require fitting the data into a parametric form have been developed to deal with general dependent time series data. Carlstein (1986) first discussed the idea of bootstrapping blocks of observations rather than the individual observations. The blocks are nonoverlapping. Künsch (1989) and Liu and Singh (1992) independently introduced a more general bootstrap procedure, the *moving block bootstrap* (MBB) which is applicable to stationary time series data. In this method the blocks of observations are overlapping.

The methods of Carlstein (nonoverlapping blocks) and Künsch (overlapping blocks) both divide the data of n observations into blocks of length l and select b of these blocks (with repeats allowed) by resampling with replacement all the possible blocks. Let us for simplicity assume $n = bl$. In the Carlstein procedure, there are just b blocks. In the Künsch procedure there are $n - l + 1$ blocks. The blocks are $L_k = \{x_k, x_{k+1}, \dots, x_{k+l-1}\}$ for $k = 1, 2, \dots, (n - l + 1)$. For example with $n = 6$ and $l = 3$ suppose the data are $x_t = \{3, 6, 7, 2, 1, 5\}$. The blocks according to Carlstein are $\{(3, 6, 7), (2, 1, 5)\}$. The blocks according to Künsch are $\{(3, 6, 7), (6, 7, 2), (7, 2, 1), (2, 1, 5)\}$. Now draw a sample of two blocks with replacement in each case. Suppose, the first draw gave $(3, 6, 7)$. The probability of missing all of $(2, 1, 5)$ is $1/2$ in

Carlstein's scheme and $1/4$ in the moving block scheme. Thus there is a higher probability of missing entire blocks in the Carlstein scheme. For this reason, it is not popular, and is not often used. Our own experience with Carlstein's nonoverlapping block method is that it gave very erratic results as the block length was varied. The MBB did better.

The literature on blocking methods is mostly on the estimation of the sample mean and its variance, although Liu and Singh (1992) talk about the applicability of the results to more general statistics, and Künsch (1989, p. 1235) discusses the AR(1) and MA(1) model. In all these studies bootstrapping is done by sampling blocks of data. Li and Maddala (1997), since it was a regression model, blocks of the residuals were used. The block sampling of the data was also tried, but its performance was slightly worse (for the reasons given earlier in section 10.2). Many of the rules suggested for the optimal block length can possibly be adapted for regression models with $(n - p)$ substituted for n , where p is the number of regressors.

10.5.1 Problems with MBB

There are some important problems worth noting about the moving block bootstrap procedure.

- (i) The pseudo-time series generated by the moving block method is not stationary, even if the original series $\{x_t\}$ is stationary. For this reason, Politis and Romano (1994) suggest the *stationary bootstrap* method. This involves sampling blocks of random length, where the length of each block has a geometric distribution. They show that the pseudo-time series generated by the stationary bootstrap method is indeed stationary.

The application of moving block method to I(1) variables has more problems. Suppose that $\{x_t\}$ is I(1). Then it is not necessarily true that the pseudo-data $\{x_t^*\}$ generated by the moving block bootstrap is also I(1).

- (ii) The mean $\bar{x}_n^*$ of the moving block bootstrap is biased in the sense that $E(\bar{x}_n^* | x_1, x_2, \dots, x_n) \neq \bar{x}_n$. See the result (iii) of theorem 6 in Liu and Singh (1992, p. 241). Politis and Romano (1994) show that, in contrast, for stationary bootstrap procedures, $E(\bar{x}_n^* | x_1, x_2, \dots, x_n) = \bar{x}_n$.
- (iii) The moving block bootstrap estimator of the variance of $\sqrt{n}\bar{x}_n$ is also biased. Davison and Hall (1993) argue that this creates

problems in using the bootstrap- t method with the moving block bootstrap. They suggest that the usual estimator

$$\hat{\sigma}^2 = n^{-1} \sum_{i=1}^n (x_i - \bar{x}_n)^2$$

be modified to

$$\tilde{\sigma}^2 = n^{-1} \sum_{i=1}^n \{ (x_i - \bar{x}_n)^2 + \sum_{k=1}^{i-1} \sum_{i=1}^{n-k} (x_i - \bar{x}_n)(x_{i+k} - \bar{x}_n) \}$$

With this modification the bootstrap- t can improve substantially on the normal approximation. The reason for this bias in the estimator of the variance is that the block bootstrap method damages the dependence structure of the data. Unfortunately this formula is valid only for the variance of $\sqrt{n}\bar{x}_n$. For more complicated problems there is no such simple correction available. Hence, in the study of Li and Maddala (1997) no such corrections were applied. However, the Monte Carlo studies showed that the bootstrap- t provided considerable improvement over asymptotic results.

- (iv) In a subsequent paper, Hall and Horowitz (1993, 1996) investigate this problem in the context of tests based on GMM estimators. They argue that because the blocking methods do not replicate the dependence structure of the original data, it is necessary to develop special bootstrap versions of the test statistics and these must have the same distribution as the sample version of the test statistics through $O_p(n^{-1})$. They derive the bootstrap versions of the test statistics with Carlstein's blocking scheme (nonoverlapping blocks) but argue that Künsch's blocking scheme is more difficult to analyze owing to its use of overlapping blocks.

In the case of cointegration tests based on the moving block scheme, as studied in Li and Maddala (1997), the derivation of the appropriate bootstrap versions of the test statistics is still more complicated. Although the use of the bootstrap version of the usual test statistics cannot be theoretically justified, the Monte Carlo results unequivocally indicate considerable improvement over the asymptotic results.

10.5.2 Optimal length of blocks

There is some discussion on the optimal length of the blocks and the several rules that have been suggested are based on different criteria. However, the rules are useful as rough guides to selecting the optimal sized blocks.

First, the number of blocks should be lower under the nonoverlapping blocks rule of Carlstein than in the moving blocks rule of Künsch. In the stationary bootstrap approach, where blocks of random length are sampled, the average length of a block is $1/p$, where p is the parameter of the geometric distribution. Thus, $1/p$ should play the same role as the parameter l in the moving block bootstrap. Politis and Romano (1994) argue that the application of a stationary bootstrap is less sensitive to the choice of p than the application of a moving block bootstrap is to the choice of l .

There is some discussion of optimal block lengths in the papers by Carlstein (1986), Künsch (1989), and the more detailed discussion in Hall and Horowitz (1993). The rules are suggestive but putting some numbers in them we get a rough idea of the block sizes to consider.

Carlstein is interested in minimizing the MSE of the block bootstrap estimate of the variance of a general statistic $t(x_1, x_2, \dots, x_n)$ (e.g., a trimmed mean or a robust estimate of scale). He argues that as the block size increases the bias goes down but the variance goes up. Also, as the dependency among the x_i gets stronger, we need a longer block size. Based on these considerations, he derives the optimal block size l^* for the AR(1) model $y_t = \rho y_{t-1} + \varepsilon_t$. His rule is

$$l^* = T^{1/3} (2\rho/(1 - \rho^2))^{2/3}$$

Note that as ρ increases l^* increases. For $T = 200$ and $\rho = 0.5, 0.8$, and 0.9 we get $l^* = 7.08, 15.81$, and 26.18 respectively. In practice these numbers are rounded, e.g., for $\rho = 0.9$ we might consider eight blocks of length 25.

As for Künsch, it is commonly believed that he suggested the optimal number of blocks to be proportional to $T^{1/3}$, i.e., $l^* \propto T^{2/3}$. This implies much longer blocks. But he also suggested to use *subjective judgement* based on sample correlations. This is perhaps a better rule than the other widely quoted one.

Hall and Horowitz (1993) derive rules taking into account the MSE in the estimation of variance as in Carlstein. They argue that the rules are similar both for the moving block scheme (they call this Künsch's rule)

and the nonoverlapping block scheme (they call this Carlstein's rule). Further more they say the rule is the same for MSE of the estimate of bias. Their rule is

$$\begin{aligned} l &= (3/2)^{1/3} T^{1/3} \zeta^{-2/3} && \text{under K\"{u}nsch's rule} \\ l &= T^{1/3} \zeta^{-2/3} && \text{under Carlstein's rule} \end{aligned}$$

where

$$\zeta = \{\gamma(0) + 2 \sum_{j=1}^{\infty} \gamma(j)\} \{|\sum_{j=1}^{\infty} j\gamma(j)|\}^{-1}$$

and $\gamma(j)$ equals the covariance of y_t at lag j . For the AR(1) process

$$y_t = \rho y_{t-1} + \varepsilon_t, \quad \zeta = (1 - \rho^2)/\rho$$

For the MA(1) process

$$y_t = \varepsilon_t + \phi \varepsilon_{t-1}, \quad \zeta = (1 + \phi)^2/\theta$$

We have computed l^* for the two processes. They are for $T = 200$.

	AR(1)			MA(1)	
ϕ	K	C	θ	K	C
0.5	5.11	4.46	0.5	2.46	2.15
0.8	11.40	9.96	0.8	2.63	2.30
0.9	18.88	16.49	0.9	2.65	2.32

Note: K = K\"{u}nsch's rule
C = Carlstein's rule.

10.6 Issues in bootstrapping cointegrating regressions

To focus on the issues concerning the bootstrapping of cointegrating regressions consider the simple system consisting of two variables y_1 and y_2 which are both I(1) and the cointegrating regression

$$y_{1t} = \beta y_{2t} + u_t, \quad t = 1, 2, \dots, T \quad (10.7)$$

where u is I(0). As is well known the least squares estimator $\hat{\beta}$ of β is superconsistent, and the asymptotic distribution of $\hat{\beta}$ involves nuisance parameters arising from endogeneity and serial correlation in u . Several corrections for the problems of endogeneity and serial correlation have been proposed in the literature (see chapter 5). However, all these

methods involve estimation of equations different from equation (10.7) and enable the derivation of asymptotically pivotal statistics that are needed for the proper application of bootstrap methods. This was what was done in Li and Maddala (1997). Note that estimation of (10.7) does not lead to asymptotically pivotal statistics, except in the special case where the problem of endogeneity and serial correlation are absent.

10.6.1 Bootstrap data for cointegrated regressions

Suppose that there is no endogeneity problem nor the serial correlation problem. Then the OLS estimator $\hat{\beta}$ of β in (10.7) has an asymptotic distribution that is nuisance parameter free, and one can apply the bootstrap procedure. Even then it is better to bootstrap the pivotal *trather* than $\hat{\beta}$ itself to get confidence intervals for β . The bootstrap-*t* confidence intervals are more accurate, as discussed in section 10.2, than those based on the bootstrap distribution of $\hat{\beta}$.

The bootstrap methods applicable to regression models as in Vinod and McCullough (1995) cannot be used for cointegrated systems. Nor is bootstrapping (y_1, y_2) directly, suggested by Freedman (1981) and Efron and Gong (1983), since this method does not use the information that y_2 is $I(1)$ and (10.7) is a cointegration relationship. Note that this problem does not arise in the censored regression model considered by Efron (1981b) or the examples considered in Efron and Gong (1983). A valid procedure is the following:

- (i) get $\hat{u}_t$ by estimating (10.7) by OLS and get the set of residuals $\hat{v}_t = \Delta y_{2t}$,
- (ii) after centering these residuals, bootstrap the *pairs* $(\hat{u}_t, \hat{v}_t)$,
- (iii) construct y_{2t}^* using the recursive method and y_{1t}^* using $\hat{\beta}$, u_t^* , and y_{2t}^* in (10.7).

This method exploits the information that y_2 is $I(1)$ and that (10.7) is a cointegrating relationship.

What if there is no endogeneity but there is serial correlation in the errors u in equation (10.7)? In this case as Phillips and Park (1988) showed, the OLS and GLS estimators are asymptotically equivalent. However, to obtain the valid *t*-statistic, which we need for the purpose of bootstrap-*t* confidence intervals, we need to calculate the asymptotic variance of $\hat{\beta}$. This is described in Phillips and Park (1988) and we need not repeat it here. As for bootstrap data generation when the structure of autocorrelation is not known, we bootstrap blocks of $(\hat{u}_t, \hat{v}_t)$

as discussed in Li and Maddala (1997). If the errors are assumed to be an AR(1) process, we use recursive methods as in Li (1994). All this is not valid in the presence of endogeneity.

It is useful to discuss the bootstrap data generation for cointegrated systems under two headings:

- (i) single equation methods and
- (ii) system methods.

In both the cases an estimation method that corrects for endogeneity and serial correlation is used, e.g., FM-OLS (or any other dynamic least squares regression) in the single equation context, and the Johansen (or Box–Tiao) procedure in the case of system estimation. We can then use the estimated coefficients and residuals to generate bootstrap samples. In the case of system methods we use the residuals from the VAR we started out with. In the case of single equation methods, care should be taken to bootstrap pairs of residuals as discussed earlier.

However, if the correct lag length is not used in the VAR in the case of system methods, or the dynamic regression model in the case of single equation methods, then the residuals will be autocorrelated and it is better to use the moving block bootstrap (MBB) as done in Li and Maddala (1997). These authors illustrate how the MBB and stationary bootstrap improve asymptotic inference. The bootstrap data generation is described in their paper. These methods are not called for in their example where the structure of the model is known. However, they can be used in all circumstances. Note that, as the study of Boswijk and Franses, referred to earlier in chapter 6, pointed out, the specification of too few lags in the Johansen procedure results in substantial size distortions in the tests, and overspecification of the lag length results in a loss of power. In such circumstances it is better to specify a smaller lag and use the MBB procedure. Van Giersbergen (1996) also used the stationary bootstrap in bootstrapping the trace statistic in VAR models. He says this will take care of misspecifications.

Note that Fachin (1996) does not use the MBB. He uses the estimated coefficients from the Johansen procedure and the residuals from the VAR (filtered out for the estimated coefficients) in resampling. He finds that using the asymptotic distribution results in a size always higher than the nominal and that the bootstrap-based tests correct the size distortion. He, however, finds that the bootstrap-based tests may lack power. Li (1997), on the other hand, does find that the bootstrap-based tests have higher power. His investigation is not for the Johansen procedure. It

is for OLS and GLS testing procedures. More detailed investigation of the power of bootstrap-based tests and the role of MBB procedures is needed.

10.7 Miscellaneous other applications

Bootstrap methods have been widely used in the financial literature. Maddala and Li (1996) review the applications of bootstrap methods in finance. They classify the applications into the following categories:

- (i) to obtain small sample standard errors,
- (ii) to get significance levels of tests,
- (iii) to get significance levels for trading rule profits,
- (iv) to develop empirical approximations to population distributions,
- (v) to use trading rules on bootstrapped data as a test for model specification,
- (vi) to check the validity of long-horizon predictability,
- (vii) for impulse response analysis in nonlinear models.

They review several papers and point out the limitations of the way the bootstrap procedures were implemented.

One important application listed there (pp. 476–478) is the use of bootstrap methods for model selection using trading rules. The procedure is as follows: first, get a measure of the profits generated by a trading rule using the actual data. Next, estimate the postulated model and bootstrap the residuals to generate bootstrap samples. Generate trading rule profits for each of the bootstrap samples. Compare the bootstrap distribution of the trading rule profits with the profits from the actual data. The basic idea is to compare the time series properties of the generated series from the given model, with the actual data. Measures like R^2 and other goodness of fit measures do not capture the time series properties of the data, as trading rule profits do.

One other application worth mentioning is that by Oke and Lyhagen (1996). These authors use the bootstrap methods to study the implications of pre-testing that is so common in the literature on unit roots and cointegration. For instance, unit root tests are used as pre-tests – a prelude to cointegration analysis. Tests for cointegration rank are used as a prelude to tests on the coefficients of the cointegrating vectors. An earlier discussion of the use of bootstrap in *data mining* (which involves pre-testing) is in the papers by Freedman, Navidi, and Peters, and Dijkstra and Veldkamp in T.K. Dijkstra, ed. (1988), *On Model Uncertainty*

and *Its Statistical Implications* (Springer, Berlin). The use of bootstrap methods for the pre-testing problems in unit roots and cointegration is an important area to pursue, where the role of bootstrap methods needs to be thoroughly investigated.

10.8 Conclusions

This chapter points out how bootstrap methods can be used in making small sample inference on unit roots and cointegration. It has been found that in small samples the use of asymptotic theory results in substantial biases in the estimated coefficients and size distortions in the tests used. Bootstrap methods help solve these problems with no loss in power.

We have discussed the appropriate procedures for bootstrap data generations, methods of constructing the test statistics, and the importance of the use of pivotal methods in connection with the use of bootstrap. We have also discussed the use of moving block and stationary bootstraps.

Of some concern is the applicability of bootstrap methods to the Johansen estimator – where the bootstrap does not appear to be giving much of an improvement over asymptotic inference. More detailed analysis of this and the Box–Tiao model is needed. Another issue that needs further investigation is the pre-testing problem.

There are many instances where defective bootstrap methods have been used (see Maddala and Li, 1996). This raises the question: Is a defective bootstrap method still better than asymptotic inference? There are several examples in the literature where this is not so. One case that is relevant to our discussion is the case of bootstrapping unit root models, as shown in Basawa *et al.* (1991a). However, when no asymptotic inference is available, it is better to use a bootstrap method. Also when the correct bootstrap method is complicated and not feasible, a theoretically imperfect bootstrap method might still improve asymptotic inference, as discussed in Li and Maddala (1997). Thus, unless proven otherwise, some bootstrap may be better than no bootstrap.

References

- Abramovich, L. and K. Singh (1985), “Edgeworth Corrected Pivotal Statistics and the Bootstrap,” *Annals of Statistics*, 13, 116–132.
- Andrews, D.W.K. (1993), “Exactly Median-Unbiased Estimation of First-Order Autoregressive/Unit-Root Models,” *Econometrica*, 61, 139–165.

- Barnard, G.A. (1963), "Contribution to Discussion," *Journal of the Royal Statistical Society*, B, 25, 294.
- Basawa, I.V., A.K. Mallik, W.P. McCormick, and R.L. Taylor (1989), "Bootstrapping Explosive Autoregressive Processes," *Annals of Statistics*, 17, 1479–1486.
- Basawa, I.V. and R.L. Taylor (1991a), "Bootstrapping Unstable First Order Autoregressive Processes," *Annals of Statistics*, 19, 1098–1101.
- (1991b), "Bootstrapping Test of Significance and Sequential Bootstrap Estimation for Unstable First Order Autoregressive Processes," *Communications in Statistics: Theory and Methods*, 20, 1015–1026.
- Beran, R. (1987), "Prepivoting to Reduce Level Error of Confidence Sets," *Biometrika*, 74, 457–468.
- (1988), "Prepivoting Test Statistics: A Bootstrap View of Asymptotic Refinements," *Journal of American Statistical Association*, 83, 687–697.
- (1990), "Refining Bootstrap Simultaneous Confidence Sets," *Journal of American Statistical Association*, 85, 417–426.
- Bhargava, A. (1986), "On the Theory of Testing for Unit Roots in Observed Time Series," *Review of Economic Studies*, 53, 369–384.
- Bickel, P.J. and D.A. Freedman (1980), "On Edgeworth Expansions and the Bootstrap," Manuscript, University of California, Berkeley.
- Carlstein, E. (1986), "The Use of Subseries Values for Estimating the Variance of a General Statistic from a Stationary Sequence," *Annals of Statistics*, 14, 1171–1179.
- Davison, A.C. and P. Hall (1993), "On Studentizing and Blocking Methods for Implementing the Bootstrap with Dependent Data," *Australian Journal of Statistics*, 35, 212–224.
- DiCiccio, T. and J. Romano (1988), "A Review of Bootstrap Confidence Intervals (with discussion)," *Journal of the Royal Statistical Society*, B, 50, 338–354.
- (1990), "nonparametric Confidence Limits by Resampling Methods and Least Favorable Families," *International Statistical Review*, 58, 59–76.
- Dufour, J.M. (1990), "Exact Tests and Confidence Sets in Linear Regressions with Autocorrelated Errors," *Econometrica*, 58, 475–494.

- Efron, B., (1979), "Bootstrap Methods: Another Look at the Jackknife," *The Annals of Statistics*, 7, 1–26.
- (1981a), "nonparametric Standard Errors and Confidence Intervals," *Canadian Journal of Statistics*, 9, 139–172.
- (1981b), "Censored Data and the Bootstrap," *Journal of the American Statistical Association*, 76, 312–319.
- (1987), "Better Bootstrap Confidence Intervals," *Journal of the American Statistical Association*, 82, 171–200.
- Efron, B. and G. Gong (1983), "A Leisurely Look at the Bootstrap, the Jackknife, and Cross Validation," *American Statistician*, 37, 36–48.
- Efron, B. and R. Tibshirani (1986), "Bootstrap Methods for Standard Errors, Confidence Intervals, and Other Measures of Statistical Accuracy," *Statistical Science*, 1, 54–77.
- (1993), *An Introduction to the Bootstrap*, Chapman and Hall, New York.
- Fachin, S. (1996), "Bootstrap Tests on Cointegrating Coefficients," Manuscript, University of Rome, La Sapienza.
- Ferretti, N. and J. Romo (1994), "Unit Root Bootstrap Tests for AR(1) Models," Working paper, Division of Economics, Universidad Carlos III de Madrid.
- Freedman, D.A. (1981), "Bootstrapping Regression Models," *The Annals of Statistics*, 9, 1218–1228.
- Freedman, D.A. and S.C. Peters (1984), "Bootstrapping a Regression Equation: Some Empirical Results," *Journal of the American Statistical Association*, 79, 97–106.
- van Giersbergen, N.P.A. (1996), "Bootstrapping the Trace Statistic in VAR Models: Monte Carlo Results and Applications," *Oxford Bulletin of Economics and Statistics*, 58, 391–408.
- van Giersbergen, N.P.A. and J.F. Kiviet (1993), "A Monte Carlo Comparison of Asymptotic and Various Bootstrap Inference Procedures in First-order Dynamic Models," Discussion paper TI 93-187, Tinbergen Institute, University of Amsterdam.
- (1994), "How to Implement Bootstrap Hypothesis Testing in Regression Models," Discussion paper TI 94-130, Tinbergen Institute, University of Amsterdam.
- (1996), "Bootstrapping a Stable AD Model: Weak versus Strong Exogeneity," *Oxford Bulletin of Economics and Statistics*, 58, 631–656.

- Gregory, A.W. and M.R. Veall (1985), "Formulating Wald Tests of nonlinear Restrictions," *Econometrica*, 53, 1465–1468.
- Hall, P. (1988a), "On Symmetric Bootstrap Confidence Intervals," *Journal of the Royal Statistical Society*, B, 50, 35–45.
- (1988b), "Theoretical Comparison of Bootstrap Confidence Intervals," *Annals of Statistics*, 16, 927–953.
- (1989), "Unusual Properties of Bootstrap Confidence Intervals in Regression Problems," *Probability Theory and Related Fields*, 81, 247–273.
- (1992), *The Bootstrap and Edgeworth Expansion*, Springer Verlag, New York.
- Hall, P. and J.L. Horowitz (1993), "Corrections and Blocking Rules for the Block Bootstrap with Dependent Data," Working paper No. 93-11, Department of Economics, University of Iowa.
- (1996), "Bootstrap Critical Values for Tests Based on Generalized Method of Moments Estimators," *Econometrica*, 64, 891–916.
- Hall, P. and S.R. Wilson (1991), "Two Guidelines for Bootstrap Hypothesis Testing," *Biometrics*, 47, 757–762.
- Hardle, W. and E. Mammen (1990), "Bootstrap Methods in nonparametric Regression," CORE Discussion paper no. 9049.
- Harris, R.I.D. (1992), "Small Sample Testing for Unit Roots," *Oxford Bulletin of Economics and Statistics*, 54, 615–625.
- Hartigan, J.A. (1969), "Using Subsample Values as Typical Values," *Journal of the American Statistical Association*, 64, 1303–1317.
- (1986), "Comment on the Paper by Efron and Tibshirani," *Statistical Science*, 1, 75–76.
- Hinkley, D. V. (1988), "Bootstrap Methods (With Discussion)," *Journal of the Royal Statistical Society*, B, 50, 321–337.
- (1989), "Bootstrap Significance Tests," *Proceedings of the 47th Session of the International Statistical Institute*, 3, 65–74.
- Horowitz, J.L. (1994), "Bootstrap-based Critical Values for the Information Matrix Test," *Journal of Econometrics*, 61, 395–411.
- (1997), "Bootstrap Methods in Econometrics: Theory and Numerical Performance," in D.M. Kreps and K.F. Wallis (eds.), *Advances in Economics and Econometrics: Theory and Applications*, vol. III, Cambridge University Press, Cambridge.
- Horowitz, J.L. and N.E. Savin (1992), "Non-Invariance of the Wald Test: The Bootstrap to the Rescue," Manuscript, University of Iowa.

- Jeong, J. and G.S. Maddala (1993), "A Perspective on Application of Bootstrap Methods in Econometrics," in G.S. Maddala, C.R. Rao, and H.D. Vinod (eds.), *Handbook of Statistics*, vol. 11, North-Holland, Amsterdam, 573–610.
- (1996), "Testing the Rationality of Survey Data Using the Weighted Double-Bootstrapped Method of Moments," *Review of Economics and Statistics*, 78, 296–302.
- Kilian, L. (1996), "Small Sample Confidence Intervals for Impulse Response Functions," mimeo, University of Michigan.
- Kiviet, J.F., (1984), "Bootstrap Inference in Lagged-Dependent Variable Models," Working paper, University of Amsterdam.
- Künsch, H.R. (1989), "The Jackknife and the Bootstrap for General Stationary Observations," *The Annals of Statistics*, 17, 1217–1241.
- Li, H. (1994), "Bootstrapping Cointegrating Regressions," *Economics Letters*, 32, 229–233.
- (1997), "The Power of Bootstrap Based Tests for Parameters in Cointegrating Regressions," Manuscript, Chinese University of Hong Kong.
- Li, H. and G.S. Maddala (1996), "Bootstrapping Time Series Models," with Discussion, *Econometric Reviews*, 15, 115–195.
- (1997), "Bootstrapping Cointegrating Regressions," *Journal of Econometrics*, 80, 297–318.
- Liu, R.Y. and K. Singh (1992), "Moving Blocks Jackknife and Bootstrap Capture Weak Dependence," in R. LePage and L. Billard (eds.), *Exploring the Limits of Bootstrap*, Wiley, New York, 225–248.
- Maddala, G.S. and H. Li (1996), "Bootstrap Based Tests in Financial Models," in G.S. Maddala and C.R. Rao (eds.), *Handbook of Statistics*, vol. 14, Elsevier Science, Amsterdam, 463–488.
- Nankervis, J.C. and N.E. Savin (1996), "The Level and Power of the Bootstrap t -test in the AR(1) Model with Trend," *Journal of Business and Economic Statistics*, 14, 161–168.
- Oke, T. and J. Lyhagen (1996), "Pre-Testing for Unit Root and Bootstrapping Results in Swedish GNP," Manuscript, Dept. of Statistics, Uppsala University.
- Phillips, P.C.B. and J.Y. Park (1988), "Asymptotic Equivalence of Ordinary Least Squares and Generalized Least Squares in Regressions with Integrated Variables," *Journal of the American Statistical Association*, 83, 111–115.

- Politis, D.N. and J. Romano (1994), "The Stationary Bootstrap," *Journal of American Statistical Association*, 89, 1303–1313.
- Rayner, R.K. (1990), "Bootstrapping p Values and Power in the First-Order Autoregression: A Monte Carlo Investigation," *Journal of Business and Economic Statistics*, 8, 251–263.
- (1991), "Resampling Methods for Tests in Regression Models with Autocorrelated Errors," *Economics Letters*, 36, 281–284.
- Rilstone, P. (1993), "Some Improvements for Bootstrapping Regression Estimators under First-Order Serial Correlation," *Economics Letters*, 42, 335–339.
- Schwert, G.W. (1987), "Effects of Model Specification on Tests for Unit Roots in Macroeconomic Data," *Journal of Monetary Economics*, 20, 73–103.
- Singh, K. (1981), "On the Asymptotic Accuracy of Efron's Bootstrap," *Annals of Statistics*, 9, 1187–1195.
- Tibshirani, R. (1988), "Variance Stabilization and the Bootstrap," *Biometrika*, 75, 433–444.
- Veall, M.R. (1986), "Bootstrapping Regression Estimators under First-Order Serial Correlation," *Economics Letters*, 21, 41–44.
- (1987), "Bootstrapping the Probability Distribution of Peak Electricity Demand," *International Economic Review*, 28, 203–212.
- (1988), "Discussion of Hall's paper," *Annals of Statistics*, 16, 979–981.
- Vinod, H.D. (1993), "Bootstrap Methods: Applications in Econometrics," in G.S. Maddala, C.R. Rao, and H.D. Vinod (eds.), *Handbook of Statistics*, vol. 11, Elsevier Science, Amsterdam, 629–661.
- Vinod, H.D. and B.D. McCullough (1995), "Estimating Cointegration Parameters: An Application of the Double Bootstrap," *Journal of Statistical Planning and Inference*, 43, 147–156.
- Zivot, E. (1994), "A Bayesian Analysis of the Unit Root Hypothesis within an Unobserved Components Model," *Econometric Theory*, 10, 552–578.

11

Cointegrated systems with I(2) variables

In the Box–Jenkins methods one continues differencing a series until the correlogram of the differenced series damps out. If the series has to be differenced d times to achieve stationarity, we say the series is integrated of order d , or I(d). In the previous chapters we considered the case $d = 1$. We shall now discuss models where $d = 2$. These models are referred to as models with double unit roots. The analog of the unit root test with unit root as the null and stationarity as the alternative is a test for I(2) versus I(1). Just as we discussed (in chapter 4) tests using stationarity as null, that is tests for I(0) versus I(1), we also have to discuss tests for I(1) versus I(2). In the following sections, we shall discuss the different problems of testing for double unit roots and problems of estimation of I(2) systems.

11.1 Determination of the order of differencing

Hasza and Fuller (1979) were the first to discuss the problem of double unit roots and proposed the joint test for double unit roots. Consider the finite-order AR model

$$A(L)y_t = \varepsilon_t \quad (11.1)$$

where $A(L)$ is a lag polynomial of order p and $\varepsilon_t \sim iid(0, \sigma^2)$. By defining a new polynomial $A^*(L)$ of order $p - 1$, $A(L)$ can be rearranged such that $A(L) = A(1)L + A^*(L)(1 - L)$, and by doing the same trick on $A^*(L)$ we get

$$A(L) = A(1)L + A^*(1)(1 - L)L + A^{**}(L)(1 - L)^2 \quad (11.2)$$

If y_t has at least one unit root, then $A(1) = 0$, and if two unit roots exist, $A^*(1) = 0$ as well. Combining (11.1) and (11.2) provides a regression of

the augmented Dickey–Fuller form

$$\Delta^2 y_t = (\hat{\pi}_1 - 1)y_{t-1} + (\hat{\pi}_2 - 1)\Delta y_{t-1} + \sum_{j=1}^{p-2} \hat{\gamma}_j \Delta^2 y_{t-j} + \hat{\varepsilon}_t \quad (11.3)$$

Hasza and Fuller proposed an F -test for double unit roots testing the hypothesis $H_0 : \pi_1 = \pi_2 = 0$ in the regression (11.3). They derived the asymptotic distribution of the usual F -test under the maintained null of double unit roots. The F -statistic for testing the joint hypothesis $H_0 : \pi_1 = \pi_2 = 1$ satisfies

$$\begin{aligned} F_{\hat{\pi}} \Rightarrow \frac{\sigma^2}{2\sigma_\varepsilon^2} D^{-1} \times & \left[\left(\int_0^1 \overline{W} dW \right)^2 \int_0^1 W^2 \right. \\ & - 2 \left(\int_0^1 W dW + \lambda \right) \int_0^1 \overline{W} W \int_0^1 \overline{W} dW \\ & \left. + \left(\int_0^1 W dW + \lambda \right)^2 \int_0^1 \overline{W}^2 \right] \end{aligned} \quad (11.4)$$

where

$$D = \int_0^1 \overline{W}^2 \int_0^1 W^2 - \left(\int_0^1 \overline{W} W \right)^2$$

and

$$\lambda = \frac{\sigma^2 - \sigma_\varepsilon^2}{2\sigma^2}$$

Hasza and Fuller provide critical values, including situations in which an intercept and possibly an intercept plus a trend are included in the regression.

Dickey and Pantula (1991) investigated the effect of the standard Dickey–Fuller test and the double unit roots test suggested by Hasza and Fuller in the presence of additional unit roots. Their simulation study shows that when the process had three unit roots:

- (i) the 5 percent level Dickey–Fuller test rejected the null of one unit root in favor of stationarity 9 percent of the time,
- (ii) the 5 percent level double unit roots Hasza–Fuller test rejected the null of double unit roots in favor of one unit root and two stationary roots 9.2 percent of the time.

Intuitively, if there are more unit roots, the test for less unit roots will strongly indicate that the process needs to be differenced, and hence the

null hypothesis will be rejected less than 5 percent of the time. However, their simulation study does not support this intuition.

With this motivating study, Dickey and Pantula (1991) formalize the Box–Jenkins sequential approach and investigate its validity. Suppose one uses a sequence of unit root tests in the same order as with visual inspection of the autocorrelation function (the Box–Jenkins sequential approach). If the hypothesis of the presence of a unit root for the series levels were not rejected, one would then test the differences for the presence of a second unit root, and so on. Dickey and Pantula refer to this as the *standard* testing sequence. They pointed out that since the standard DF unit root test is based on the assumption of at most one unit root, at least the first few tests in this sequence would not be theoretically justified, if the series had more than one unit root. Sen (1995) also observed empirically that, under the hypothesis of two unit roots, the DF test rejects the null hypothesis of one unit root with probability slightly greater than α . That is, we are more likely to conclude that the process is stationary when there are really two unit roots present than when there is exactly one unit root. Thus, one should avoid testing for one unit root before testing for a higher number of unit roots. Another problem is that the single t -statistics are not invariant tests because their distributions under the null will depend on the actual number of unit roots present. For instance, the t -statistics of π_1 in (11.3) will follow the Dickey–Fuller t -distribution if $\pi_2 < 1$, but the distribution changes if $\pi_2 = 1$.

Dickey and Pantula (1991) show that an appropriate way to proceed is to reverse the order of testing by starting with the highest possible order of integration and then testing down the model. They proposed t^* - and F -tests that compare a null hypothesis of k unit roots with an alternative of $k - 1$ unit roots. Consider

$$y_t = \sum_{j=1}^p \beta_j y_{t-j} + \varepsilon_t$$

where $\{\varepsilon_t\}$ is a sequence of *iid* random variables with mean 0 and variance $\sigma^2 = 1$ and $y_{-p+1} = \dots = y_0 = 0$. They considered the AR model with $p = 3$ and three possible unit roots. Let m_1, m_2 , and m_3 denote the roots of the characteristic equation

$$m^3 - \beta_1 m^2 - \beta_2 m - \beta_3 = 0$$

The AR(3) model can be written

$$x_t = \theta_1 y_{t-1} + \theta_2 z_{t-1} + \theta_3 w_{t-1} + \varepsilon_t \quad (11.5)$$

where $z_t = y_t - y_{t-1}$, $w_t = z_t - z_{t-1}$, and $x_t = w_t - w_{t-1}$. Note that z_t , w_t , and x_t are the first, second, and third differences of the process y_t , respectively. Then the hypotheses about the different number of unit roots are:

- (i) No unit root, $H_0 : |m_1| < 1$,

or

$$H_0 : \theta_1 < 0 \text{ and some restrictions on } \theta_2 \text{ and } \theta_3.$$

For example, the restrictions on θ_2 and θ_3 are

$$-12 < \theta_2 + 2\theta_1 < 0, -2 < \theta_3 < 0$$

- (ii) One unit root, $H_1 : m_1 = 1, |m_2| < 1$,

or

$$H_1 : \theta_1 = 0, \theta_2 < 0 \text{ and some restrictions on } \theta_2 \text{ and } \theta_3.$$

The restrictions on θ_2 and θ_3 are $0 < 4 + \theta_2 + 2\theta_3, -2 < \theta_3 < 0$.

- (iii) Two unit roots, $H_2 : m_1 = m_2 = 1, |m_3| < 1$,

or

$$H_2 : \theta_1 = \theta_2 = 0, \theta_3 < 0.$$

- (iv) Three unit roots, $H_3 : m_1 = m_2 = m_3 = 1$,

or

$$H_3 : \theta_1 = \theta_2 = \theta_3 = 0.$$

Note that the subscript i in H_i denotes the number of unit roots of the corresponding null hypotheses. Dickey and Pantula showed that applying regression F -statistics for testing from the higher number of unit roots to the lower number of unit roots is valid, while the reverse order of applying F -statistics is not valid. Based on the asymptotic distributions of F -statistics derived by Pantula (1986), they suggest testing the hypotheses sequentially in the order H_3, H_2 , and H_1 :

- (i) If H_3 is rejected by F -test, then go to the next step; otherwise conclude that H_3 is true.
- (ii) If H_2 is rejected by F -test, then go to the next step; otherwise conclude that H_2 is true.

- (iii) If H_1 is rejected, then conclude that H_0 is true; otherwise conclude that H_1 is true.

The empirical percentiles of the asymptotic distributions of the F -statistics can be found in Pantula (1986).

Dickey and Pantula also proposed an alternative testing procedure based on t -test statistics. However, the t -statistics from the regression (11.5) has asymptotic distributions depending on the number of unit roots present. And thus they argue that a sequential procedure based on these statistics is not consistent. They suggest using alternative t^* -statistics which follows the standard Dickey–Fuller t -distribution:

- (i) for H_3 against H_2 , obtain the t^* -statistic from the regression x_t on w_{t-1} for testing the coefficient of w_{t-1} is 0,
- (ii) for H_2 against H_1 , use the t^* -statistic in the regression x_t on z_{t-1} and w_{t-1} for testing the coefficient of z_{t-1} is 0,
- (iii) for H_1 against H_0 , use the t^* -statistic in the regression x_t on y_{t-1} , z_{t-1} , and w_{t-1} for testing the coefficient of y_{t-1} is 0.

They show that the t^* -test will have higher power than the joint F -test against the $I(1)$ alternative, since prior imposition of a unit root in the first step will be correct, both under the null and the alternative. They also argued that some power is gained since the joint test is two-sided in nature, whereas only the one-sided alternative is addressed in their single t -test procedure. Their results of a Monte Carlo power study supports these findings and show that t^* -procedure is more powerful than F in most cases.

Haldrup (1994a) argued that the sequential testing procedure suggested by Dickey and Pantula (1991) may result in severe problems in the presence of an explosive root, which may be a likely alternative candidate to the double unit root process. He argued that prior imposition of a single unit root under the presence of an explosive root may result in the misleading conclusion that double unit roots are present because, in principle, an explosive process can be differenced an infinite number of times without being stationary. Moreover, differencing of an explosive process will produce a noninvertible error term that may cause serious problems with respect to estimation and inference.

Haldrup (1994a) developed a semiparametric testing procedure equivalent to the augmented Hasza–Fuller procedure as a straightforward generalization of the Phillips and Perron's Z -test for a unit root in the $I(1)$ model (see chapter 3 for details). He showed that the nonparametric

correction of the F -statistic in the equation (11.3) such that

$$Z_F = F_{\hat{\pi}} \frac{\hat{\sigma}_\varepsilon^2}{\hat{\sigma}^2} - \frac{1}{2} M^{-1} N$$

where

$$M = T^{-4} \sum_{t=1}^T y_t^2 \cdot T^{-2} \sum_{t=1}^T \Delta y_t \Delta y_t - \left(T^{-3} \sum_{t=1}^T y_{t-1} \Delta y_t \right)^2$$

and

$$\begin{aligned} N = & \frac{(\hat{\sigma}^2 - \hat{\sigma}_\varepsilon^2)}{\hat{\sigma}^2} \left(T^{-1} \sum_{t=1}^T \Delta y_{t-1} \Delta^2 y_t \cdot T^{-4} \sum_{t=1}^T y_t^2 \right. \\ & \left. - T^{-3} \sum_{t=1}^T y_{t-1} \Delta y_t \cdot T^{-2} \sum_{t=1}^T y_{t-1} \Delta^2 y_t \right) \\ & - \left(\frac{\hat{\sigma}^2 - \hat{\sigma}_\varepsilon^2}{2\hat{\sigma}} \right)^2 T^{-4} \sum_{t=1}^T y_t^2 \end{aligned}$$

have the same limiting distribution given in (11.4). The limiting distribution of the Z_F -statistic is invariant within a wide class of weakly dependent and potentially heterogeneously distributed errors (see chapter 3 for the wide class of errors and the Newey–West estimate of the long-run variance $\hat{\sigma}^2$). Since the distribution is identical to the situation in which $\sigma^2 = \sigma_\varepsilon^2$, the critical values tabulated by Hasza and Fuller (1979) remain valid.

The limiting distributions are affected by the inclusion of deterministic components. Since an intercept, a linear trend, and possibly a quadratic trend effectively demean and detrend the series prior to testing for unit roots, the asymptotics will carry through in a similar and straightforward fashion with the only modification being that the original Brownian motions are substituted by demeaned and detrended Brownian motions (see chapter 3).

A Monte Carlo study by Haldrup (1994a) showed that when the DGP is an explosive process, i.e., $\pi_1 > 1$ in (11.3), the Dickey and Pantula tests appear to have almost zero power, and thus the double unit root is accepted in most cases. On the other hand, the joint test procedure proposed by Hasza and Fuller and the semiparametric version by Haldrup which does not place *a priori* restriction on the data, discriminate $I(2)$ processes from explosive processes, although this property is less pronounced when the sample size is small. However, the Haldrup test

suffers from serious size distortions in case of negative MA roots, like other semiparametric tests, for example Phillips and Perron's Z -test (see chapter 4).

In practice, it is recommended to apply a collection of several tests and hope that some common evidence is obtained.

11.2 Cointegration analysis with I(2) and I(1) variables

After the determination of which variables are I(0), I(1), and I(2), the next step is the determination of any cointegrating relationships among these variables and the estimation of equations with mixtures of I(0), I(1), and I(2) variables. There are two kinds of methods that have been proposed for estimation:

- (i) Single equation methods,
- (ii) System methods.

In system methods there are two subcategories:

- (i) the nature of cointegration and the number of cointegrating vectors are known,
- (ii) these factors need to be estimated.

We will discuss these in turn.

11.2.1 Single equation methods

Haldrup (1994b) discusses, in detail, single equation models with I(0), I(1), and I(2) variables. He analyzes the conditions under which standard Gaussian inference can validly be conducted and how the literature on spurious regression for I(1) variables can be extended to the case of I(2) variables. He extends the residual-based Dickey–Fuller class of tests to models involving I(1) and I(2) variables and provides new critical values for these tests.

When dealing with I(2) variables, several possibilities exist. First, linear combinations of some I(2) variables can be I(1) or I(0). Second, some linear combinations of I(1) variables can cointegrate with differences of I(2) variables. In this situation the cointegrating vectors will be polynomials of the lag operator L . Hence the cointegration is referred to as *polynomial cointegration* (see Yoo, 1986 and Gregoir and Laroque, 1994).

Consider the model where the variable y_t is related to the m -dimensional variables x_t

$$y_t = \gamma'_0 c_t + y_t^o$$

$$x_t = (c'_t, x'_{1t}, x'_{2t})$$

$$x_{1t} = \gamma'_1 c_t + x_{1t}^o$$

$$x_{2t} = \gamma'_2 c_t + x_{2t}^o$$

where c_t is a deterministic component consisting of a constant, a linear trend, and possibly a polynomial trend, x_{1t}^o and x_{2t}^o are m_1 and m_2 dimensional $I(1)$ and $I(2)$ processes, respectively, and $m_1 + m_2 = m$. The y_t^o is $I(2)$ and is linked with x_{1t}^o and x_{2t}^o by the relation

$$y_t^o = \beta'_1 x_{1t}^o + \beta'_2 x_{2t}^o + u_t$$

If u_t is $I(2)$, then there is no cointegration. If u_t is $I(1)$, then y_t^o and x_{2t}^o are cointegrated, that is, $CI(2, 1)$, and if the error $(y_t^o - \beta'_2 x_{2t}^o)$ is cointegrated with x_{1t}^o , then we have a *full cointegrated system* and u_t is $I(0)$. We can write the whole model compactly as

$$y_t = \beta'_0 c_t + \beta'_1 x_{1t} + \beta'_2 x_{2t} + u_t$$

where $\beta'_0 = (\gamma'_0 - \beta'_1 \gamma'_1 - \beta'_2 \gamma'_2)$.

We now study the properties of the parameters of this equation. Depending upon the integration order of u_t there will be *stochastic cointegration* at different levels (for stochastic cointegration, see section 5.5.2 and also Ogaki and Park, 1989). Also, if some elements of β_0 turn out to be zero, there will be deterministic co-trending among some of the variables.

The algebra for the analysis of this model proceeds along the same lines as that discussed in the earlier chapters for $I(1)$ and $I(0)$ variables. The details can be found in Haldrup (1994b). However, the following important points regarding spurious regressions are worth repeating:

- (i) If u_t is $I(1)$ or $I(2)$, the F -test statistic of any hypothesis diverges to ∞ by the order $O_p(T)$ although the hypothesis is true.
- (ii) $R^2 \rightarrow 1$, not only when u_t is $I(0)$, but also when u_t is $I(1)$.
- (iii) DW statistic $\rightarrow 0$ at the rapid rate $O_p(T)$ regardless of whether u_t is $I(1)$ or $I(2)$.

Another important feature of the model is that the order of the different least squares regressions will differ. The orders of the coefficients, $\hat{\beta}_1$ and $\hat{\beta}_2$ corresponding to the I(1) and I(2) variables are $O_p(T^{d-1})$ and $O_p(T^{d-2})$, respectively, when u_t is I(d). Thus, $\hat{\beta}_1$ will diverge at $O_p(T)$, if u_t is I(2) and will be nondegenerate when u_t is I(1). As for $\hat{\beta}_2$ it will be nondegenerate when u_t is I(2), but is consistently estimated at the superconsistent rate $O_p(T^{-1})$ when u_t is I(1). When there is full cointegration, that is, u_t is I(0), $\hat{\beta}_2$ will tend to β_2 at the rapid rate of $O_p(T^{-2})$ and thus it is super-superconsistent whereas $\hat{\beta}_1$ is $O_p(T^{-1})$ and is just superconsistent. As for the order of $\hat{\beta}_0$, it depends on the deterministic components present in the stochastic regressors x_{1t} and x_{2t} (for details see Haldrup, 1994b, pp. 162–163).

An important question is when the inference can be conducted within the framework of the standard Gaussian distribution. In the case of I(1) variables we saw earlier that the distribution of the cointegrating vector is normal in the special case of no endogeneity and no serial correlation (see chapter 5). Haldrup shows that this result generalizes to the case of I(2) variables if there is full cointegration, that is, u_t is I(0).

Another case where asymptotic normality applies is when the stochastic trends for all the individual time series are dominated by deterministic trends and there is full cointegration as well as deterministic co-trending among the variables. For most practical situations this means that y_t and x_{2t} both contain quadratic trends and x_{1t} contains at least a linear trend. By deterministic co-trending we mean that when weighted by the cointegrating vector some or all the higher-order trends vanish. This result is a generalization of the result by West (1988) that we discussed in chapter 4. The proof also follows similar lines.

Residual-based tests for I(2) cointegration

Haldrup (1994b) considers residual-based tests for cointegration where the I(2) variables cointegrate to produce an I(1) variable. The hypotheses considered are:

H_0 : There is no cointegration among the I(2) and I(1) variables.

H_1 : The I(2) variables cointegrate into an I(1) relation, but no further cointegration is possible.

The residuals from the cointegrating regression under H_0 and H_1 are I(2) and I(1), respectively. The null of no cointegration means that the underlying levels regression is spurious. The test is analogous to the case where I(1) variables cointegrate to produce an I(0) variable. The

question is: what modifications do we need in the critical values in the residual-based tests that we use in the case of $I(1)$ variables?

Haldrup shows that the distribution of the ADF test statistic in this case is similar to that in the case of cointegration of $I(1)$ variables except that it depends on m_1 and m_2 , the number of $I(1)$ and $I(2)$ variables, respectively. Hence Haldrup tabulates the critical values for different values of m_1 and m_2 . The tables are for the sample sizes $n = (25, 50, 100, 250, 500)$, critical values $(0.01, 0.025, 0.05, 0.1)$, $m_1 = (0, 1, 2, 3, 4)$, and $m_2 = (1, 2)$. For large sample sizes ($n = 250, 500$) the critical values are the same for $m = m_1 + m_2$ and they are similar to those tabulated in Phillips and Ouliaris (1990). However, for small sample sizes, the critical values become lower as m_2 increases. Critical values where trend (linear or quadratic) is included are available from Haldrup on request. However, he says that the numerical evidence suggests that they are similar to the ones in the published tables for higher dimensions.

Haldrup illustrates the tests with a numerical example on the demand for money in UK. This will be discussed later along with other empirical illustrations of models with $I(2)$ variables.

11.2.2 System methods

In the preceding section, we discussed single equation methods. We now turn to system methods suggested by Stock and Watson (1993), Johansen (1994c, 1995), and Kitamura (1995).

Stock and Watson (1993)

Stock and Watson assume that the variables y_t can be split into y_{1t} , y_{2t} , and y_{3t} such that y_{1t} are $I(2)$, Δy_{2t} , and Δy_{1t} are cointegrated, and y_{3t} is cointegrated with levels of y_{1t} and y_{2t} and the first-differences Δy_{1t} . There is, thus, a lot of prior information used about the nature of the interrelationships. The model specifically is

$$\begin{aligned}\Delta^2 y_{1t} &= u_{1t} \\ \Delta y_{2t} &= \theta_1 \Delta y_{1t} + u_{2t} \\ y_{3t} &= \theta_2 y_{2t} + \theta_3 y_{1t} + \theta_4 \Delta y_{1t} + u_{3t}\end{aligned}$$

where $u_t = (u_{1t}, u_{2t}, u_{3t})$ is an $I(0)$ process. Stock and Watson concentrate on the estimation of the last equation. They suggest estimation of the equation augmented by leads and lags and show that the estimator

is consistent. Furthermore, the asymptotic distribution of the estimated coefficients is mixed Gaussian so that usual inference can be performed.

Kitamura (1995)

Kitamura's models are extensions for the case of I(2) variables of Phillips' ML estimation of triangular systems for I(1) variables. (Phillips and Chang (1994) use FM-OLS in a single equation context.) Thus, there is strong prior information on the nature of cointegration and the number of cointegrating vectors (as in Stock and Watson). The process y_t is an n -vector nonstationary process and u_t is an n -vector of stationary errors. It is assumed that the n -vectors y_t and u_t can be partitioned into subvectors of dimensions: n_a, n_1, n_2 , and n_3 , with $n = n_a + n_1 + n_2 + n_3$, y_{1t} and y_{2t} are I(1) processes and (y_{at}, y_{3t}) are I(2) processes.

Kitamura considers several models which differ in the type of cointegration relationships assumed as follows:

Model 1

$$y_{1t} = B_{12}y_{2t} + u_{1t} \quad (11.6)$$

$$y_{at} = B_{a3}y_{3t} + u_{at} \quad (11.7)$$

$$\Delta y_{2t} = u_{2t} \quad (11.8)$$

$$\Delta^2 y_{3t} = u_{3t} \quad (11.9)$$

Kitamura considers two estimation methods. One is estimating (11.6) and (11.8) together and (11.7) and (11.9) together using Phillips' procedures of estimation of triangular systems discussed in Phillips (1991). The other method is to estimate the four equations jointly.

Model 2

In model 1 the I(2) variables are cointegrated to produce an I(0) variable. In this model, they are cointegrated to produce an I(1) variable.

Model 3

Model 3 is an extension of model 2 to cover cases where differenced I(2) variables are cointegrated with I(1) variables.

In all cases Kitamura derives the triangular ECM representations and considers both subsystem MLE and full system MLE as well as the efficiency of the latter relative to that of the former.

Johansen

Johansen's method does not require information on the nature of cointegration and the number of cointegrating vectors. These are estimated from the model. As in the case of models with $I(1)$ variables, Johansen starts with a general vector autoregressive (VAR) model in p -dimensions. His estimation method is based on a representation of $I(2)$ systems derived in Johansen (1992a). The full maximum likelihood estimation of the system is complicated and discussed in Johansen (1994c). However, he describes a computationally feasible two-step ML method in Johansen (1994a, 1995) and this method is applied in Johansen (1992b) and Juselius (1994a, b). This is the method we will discuss here.

Consider the general VAR model with Gaussian errors in p -dimensions

$$X_t = \sum_{i=1}^p A_i X_{t-i} + \varepsilon_t, \quad t = 1, \dots, T$$

It is convenient to write this in the form

$$\Delta^2 X_t = \Gamma \Delta X_{t-1} + \Pi X_{t-2} + \sum_{i=1}^{p-2} \Gamma_i \Delta^2 X_{t-i} + \varepsilon_t \quad (11.10)$$

The relation between the parameters $(\Gamma, \Pi, \Gamma_1, \dots, \Gamma_{k-2})$ and $(A_1, A_2, \dots, A_k)$ is obtained by identifying the coefficients of the lagged values of X_t in the two different expressions.

Define the matrices α and β of dimension $p \times r$ ($r < p$) and $\alpha_\perp$ and $\beta_\perp$ of dimension $p \times (p - r)$ so that $\alpha\alpha_\perp = 0$ and $\beta\beta_\perp = 0$. Thus, $\alpha_\perp$ and $\beta_\perp$ are orthogonal complements of α and β , respectively. In the following, orthogonal complements of other matrices are denoted by a similar notation. Similarly, define matrices ϕ and η of full ranks and orders $(p - r) \times s_1$ with $s_1 < (p - r)$. Let $\alpha_1 = \alpha_\perp \phi$ and $\beta_1 = \beta_\perp \eta$. Supplement these by $\beta_2 = (\beta, \beta_1)_\perp$ and $\alpha_2 = (\alpha, \alpha_1)_\perp$. Thus $(\alpha, \alpha_1, \alpha_2)$ are mutually orthogonal and span $\mathcal{R}^p$. Similar is the case with $(\beta, \beta_1, \beta_2)$. The necessary conditions for $\{X_t\}$ to be $I(2)$ are

$$\Pi = \alpha\beta' \quad \text{is of reduced rank } r \quad (11.11)$$

and

$$\alpha'_\perp \Gamma \beta_\perp = \phi \eta' \quad \text{is of reduced rank } s_1 < (p - r) \quad (11.12)$$

Johansen (1995) shows that the space spanned by the vector X_t can be decomposed into r stationary direction β and $(p - r)$ nonstationary

direction $\beta_{\perp}$. The latter can be decomposed into directions

$$\beta_1 = \beta_{\perp} \eta \quad \text{is of dimension } p \times s_1$$

and

$$\beta_2 = (\beta, \beta_1)_{\perp} \quad \text{is of dimension } p \times s_2$$

where $s_1 + s_2 = p - r$. In the direction β , the process $\beta' X_t$ can be made stationary with suitable linear combinations of ΔX_t . In the direction β_1 , the process can be made stationary by first-order differencing and in the direction β_2 , the process can only be made stationary by second-order differencing. These properties can be described as follows

$$\begin{aligned} \beta' X_t &\sim I(1) & \beta'_1 \Delta X_t &\sim I(0) \\ \beta'_1 X_t &\sim I(1) & \beta'_2 \Delta^2 X_t &\sim I(0) \\ \beta'_2 X_t &\sim I(2) & \beta' X_t + w \beta'_2 \Delta X_t &\sim I(0) \end{aligned}$$

where $w = (\alpha' \alpha)^{-1} \alpha' \Gamma \beta_2 (\beta'_2 \beta_2)^{-1}$ and w is a $p \times s_2$ matrix of weights designed to pick out the $I(2)$ components of X_t .

Johansen's two-step ML procedure is as follows: the first step is the analysis of the $I(1)$ model, that is without the restrictions on Γ as given in (11.12). This determines r, α , and β . The second step is an analysis of the $I(2)$ model for fixed values of r, α , and β . The procedure, thus, is:

- (i) First determine r and the parameters α and β by a reduced-rank regression of $\Delta^2 X_t$ on X_{t-2} corrected for lagged differences and a constant. That is, analyze the model (11.10) with Γ unrestricted (using the Johansen procedure described in chapter 5).
- (ii) Note that the levels enter only through $\Pi X_{t-2} = \alpha \beta' X_{t-2}$. Hence, if we pre-multiply equation (11.10) by $\alpha'_{\perp}$, the levels then vanish and we are left with

$$\alpha'_{\perp} \Delta^2 X_t = \alpha'_{\perp} \Gamma \Delta X_{t-1} + \Pi X_{t-2} + \sum_{i=1}^{p-2} \alpha'_{\perp} \Gamma_i \Delta^2 X_{t-i} + \alpha'_{\perp} \varepsilon_t \quad (11.13)$$

To get to the second step in the Johansen procedure we need some more notation. Define $\bar{A} = A(A'A)^{-1}$ for any matrix A of full rank. Thus $A'\bar{A} = I$ and define the projection operation $P_A = A\bar{A}' = A(A'A)^{-1}A'$. Using this notation we note that

$$P_{\beta} + P_{\beta_{\perp}} = \beta \bar{\beta}' + \beta_{\perp} \beta'_{\perp} = I$$

Now write

$$\alpha'_{\perp} \Gamma \Delta X_{t-1} = \alpha'_{\perp} \Gamma (\beta \bar{\beta}' + \beta_{\perp} \beta'_{\perp}) \Delta X_{t-1}$$

$$= \alpha'_\perp \Gamma \beta (\bar{\beta}' \Delta X_{t-1}) + \phi \eta' (\beta'_\perp \Delta X_{t-1})$$

We have used the condition $\alpha'_\perp \Gamma \beta_\perp = \phi \eta'$ in (11.12). This suggests the reduced-rank regression to determine s, ϕ , and η for given r, α , and β such that the reduced-rank regression of $\alpha'_\perp \Delta^2 X_t$ on $\bar{\beta}'_\perp \Delta X_{t-1}$ corrected for lagged second differences $\beta' \Delta X_{t-2}$.

- (iii) It then follows that the asymptotic inference about α, β, ϕ , and η can be conducted by means of the χ^2 distribution.

This two-step estimator is not the maximum likelihood estimator and is not in general as efficient as the latter which estimates the parameters α, β, ϕ , and η jointly. This is so if the information matrix for the parameter sets (α, β) and (ϕ, η) is not diagonal. Hence, the proposed estimators of β and η are not expected to be efficient. The asymptotic properties of the two-step estimators are derived in Johansen (1994b) where it is shown that the two-step estimation of (α, β) is indeed efficient. Paruolo (1994) has shown the surprising result that the estimator of $\beta_1 = \beta_\perp \eta$ is also efficient. However, if the estimators of α and β are efficient, then it is evident that the second step estimators are also efficient. The second step estimators are not efficient only if any consistent estimator is used in the first step. So this whole issue of two-step estimators needs further investigation.

11.3 Empirical applications

There are two main areas of empirical investigation where I(2) models have been applied:

- (i) the long-run demand for money,
- (ii) the purchasing power parity (PPP).

11.3.1 Studies on the long-run demand for money

Stock and Watson (1993) apply the methods described earlier to the estimation of the long-run demand for money in the US based on annual data for the period 1900–1989. As mentioned earlier, their procedure requires at least partial knowledge of the orders of integration of the individual series and also which variables cointegrate. Defining the variables m, y , and p , respectively as logarithms of money balances, output, and price level, and r as interest rate (detailed definitions and data

sources are given in their paper), they conclude after some preliminary tests on the individual series that $(m - p)$ is $I(1)$ with drift, r is $I(1)$ with no drift, y is $I(1)$ with drift, and $(m - p)$, y , and r are cointegrated. The tests also suggest that $(r - \Delta p)$ is $I(0)$. Whether m and p are individually $I(1)$ or $I(2)$ is unclear, the inference depends on the subsample and tests considered. Their overall conclusion is that the demand for money is stable over the entire period although estimates based on the post-war data alone are unstable with variances indicating substantial sample variability. This conclusion of stability over the entire period is similar to what one observes in the case of the usual regression models. If the estimates for one of the subperiods have high sampling variances, then one would not reject the hypothesis of stability over the entire period.

The paper by Stock and Watson also reports Monte Carlo studies on several estimation methods for cointegrated relationship. These results have been reviewed in chapter 5 (section 5.7).

Haldrup (1994b) illustrates his test for $I(2)$ cointegration using a quarterly data set for the UK covering the period 1963:I-1989:II (106 observations), which is originally studied by Hendry and Ericsson (1991) and subsequently by Johansen (1992d). Four variables are considered in the analysis:

m_t : the log of nominal M1,

y_t : the log of real income measured as total final expenditures (TFE)
in 1985 prices,

p_t : the log of implicit price deflator for TFE,

R_t : an interest rate measure used by Hendry and Ericsson (1991).

By performing tests for unit roots it was found that m_t and p_t are perhaps $I(2)$ and real money $(m_t - p_t)$ and velocity $(m_t - p_t - y_t)$ are $I(1)$. Applying his residual-based test for $I(2)$ cointegration, Haldrup found no evidence of cointegration. This is in contrast to the conclusion of Johansen (1992d) who found one cointegrating vector by using the same data set. The question is not one of efficiency of single equation versus system methods. As argued by Johansen (1992c) there will be no gain in efficiency from analyzing the full and more complicated systems if the conditioning variables in a partial model are weakly exogenous, and the results in Johansen (1992d) indicate that this is indeed the case with these data. The possible explanation in this case is the absence of short-run dynamics in the single equation analysis which is taken into account in the Johansen VAR procedure. Haldrup concludes that this

example suggests that the absence of short-run dynamics in the estimation of long-run cointegrating parameters could lead to the rejection of cointegration more often compared to the evidence following from a dynamic approach.

Juselius (1994a) investigates long-run properties of aggregate money holdings of the private sector in Denmark. The data are quarterly covering the period 1974:II–1987:III. Five variables are considered: money (m_t), income (y_t), price (p_t), bond rate (i_t^b), and deposit rate (i_t^d). Allowing for two lags in the VAR model, the effective sample size is 52 which means (as Juselius admits) we have to be careful when making inference based on asymptotic distributions for the rank test statistics.

The data are almost the same as those analyzed by Johansen and Juselius (1990) earlier using the Johansen procedure for I(1) systems. The number of variables was four in Johansen and Juselius (1990) whereas it is five in Juselius (1994a). In Johansen and Juselius (1990) the difference ($i_t^d - i_t^b$) was used as the cost of holding money. Johansen and Juselius (1990) found that there was just one cointegrating vector. If there is only one cointegrating vector it is easy to interpret it as a long-run relationship. If there is more than one cointegrating vector, then all linear combinations are cointegrating vectors and one has to bring in more prior information to give economic interpretation. See Maddala (1991).

Juselius (1994a) found in the first step, when doing a reduced-rank analysis of the I(2) model by Johansen's two-step method described earlier, that $r = 2$, that is, there are two cointegrating vectors. Note that the number of variables is $p = 5$ whereas it was $p = 4$ in Johansen and Juselius (1990). Also, the starting model is different (second-differences rather than first-differences).

In the second step of the two-step procedure, Juselius found that $s_1 = p - r - s_2$ was 2. This implies $s_2 = p - r - s_1$ is 1. Thus, there is one I(2) trend in the data. If $r = 1$ is chosen, all the hypotheses $s_1 = 0, 1, 2, 3$ are rejected and thus we would not have found the I(2) relation because $s_2 = 5 - 1 - 4 = 0$. Juselius says that considering that r is not well determined is potentially a serious problem. In any case, Juselius argues that analysis with the I(2) model indicates that $r = 1$ is incorrect.

Since we have accepted that X_t is second-order stationary, $\beta'X_t$ can no longer be assumed stationary as in the I(1) model. However, $\beta'X_t$ corrected for a weighted sum of the differenced process is stationary. The $\beta'X_t$ can be decomposed into $(r - s_2) = 1$, I(0) vectors and $s_2 = 1$, I(1)

vectors. Juselius presents estimates of these vectors. The I(0) vector when normalized with respect to m_t gives

$$m \simeq 1.3y + 0.95p - 5.3i^b + 4.5i^d$$

This can be interpreted as the long-run demand for money. Johansen and Juselius (1990) arrived at roughly the same coefficients. They first tested for price homogeneity and since it was clearly accepted, the empirical analysis was for real money, real income, and some proxies for the cost of holding money. The long-run demand for money equation obtained in Johansen and Juselius (1990) was

$$m_2 \simeq 1.03y - 5.21i^b + 4.22i^d + 6.06$$

(Note that a constant was also estimated.)

Juselius (1994a) also conducts an analysis in real terms but adds another explanatory variable Δp . The empirical results are therefore not comparable to earlier results. We will skip the details since a detailed discussion can be found in the original paper. The interesting points to note are the differences in the results produced from the I(2) analysis compared with the I(1) analysis. One of the long-run relationships was the same in both the analyses.

11.3.2 Studies on PPP using I(2) variables

Johansen (1992b) and Juselius (1994b) use I(2) models to analyze purchasing power parity (PPP) theory. There have been numerous studies on PPP using unit roots, cointegration, fractional integration, and so on. The main argument of the Johansen and Juselius papers is that taking the nature of the data into consideration is important and that there is evidence to show that some of the variables are I(2) and this changes the inference on PPP.

11.4 Summary and conclusions

This chapter discussed cointegrated systems with I(2) variables. It has been observed that some variables, money supply and prices, are better characterized as I(2) rather than I(1). Thus, there are a few practical situations where I(2) variables do occur.

We first discuss the determination of the order of differencing: testing I(2) versus I(1). Next we discuss cointegration analysis with I(2) and I(1) variables: single equation methods and system methods. Finally

we illustrate applications: studies on demand for money and purchasing power parity.

Given that $I(2)$ variables occur only in a few situations we have kept our discussion brief. Omission of the topic completely leaves a hole and we wanted to avoid that. At the same time, the theoretical developments on $I(2)$ variables are far ahead of their empirical usefulness.

References

- Dickey, D.A. and S.G. Pantula (1991), "Determining the Order of Differencing in Autoregressive Processes," *Journal of Business and Economic Statistics*, 5, 455–461.
- Gregoir, S. and G. Laroque (1994), "Polynomial Cointegration: Estimation and Tests," *Journal of Econometrics*, 63 183–214.
- Haldrup, N. (1994a), "Semiparametric Tests for Double Unit Roots," *Journal of Business and Economic Statistics*, 12, 109–122.
- (1994b), "The Asymptotics of Single Equation Cointegration Regressions with $I(1)$ and $I(2)$ Variables," *Journal of Econometrics*, 63, 153–181.
- Hasza, D.P. and W.A. Fuller (1979), "Estimation of Autoregressive Processes with Unit Roots," *Annals of Statistics*, 7, 1106–1120.
- Hendry, D.F. and N.R. Ericsson (1991), "modeling the Demand for Narrow Money in the United Kingdom and the United States," *European Economic Review*, 35, 833–881.
- Johansen, S. (1992a), "A Representation of Vector Autoregressive Processes Integrated of Order 2," *Econometric Theory*, 8, 188–202.
- (1992b), "An $I(2)$ Cointegration Analysis of the Purchasing Power Parity between Australia and USA," in Colin Hargreaves (ed.), *Macroeconomic modeling of the Long-run*, Edward Elgar.
- (1992c), "Cointegration in Partial Systems and the Efficiency of Single Equation Analysis," *Journal of Econometrics*, 52, 389–402.
- (1992d), "Testing Weak Exogeneity and the Order of Cointegration in UK Money Demand Data," *Journal of Policy modeling*, 14, 313–335.
- (1994a), "The Role of the Constant and Linear Terms in Cointegration Analysis of Non-Stationary Variables," *Econometric Reviews*, 13, 205–229.
- (1994b), "Estimation of Systems of Trending Variables," *Econometric Reviews*, 13, 351–386.

- (1994c), "A Likelihood Analysis of the $I(2)$ model," *Scandinavian Journal of Statistics*.
- (1995), "A Statistical Analysis of Cointegration for $I(2)$ Variables," *Econometric Theory*, 11, 25–59.
- Johansen, S. and K. Juselius (1990), "Maximum Likelihood Estimation and Inference on Cointegration – with Applications to the Demand for Money," *Oxford Bulletin of Economics and Statistics*, 52, 169–210.
- Juselius, K. (1994a), "On the Duality between Long-Run Relations and Common Trends in the $I(1)$ versus $I(2)$ Model. An Application to Aggregate Money Holdings," *Econometric Reviews*, 13, 151–178.
- (1994b), "Do Purchasing Power Parity and Uncovered Interest Parity Hold in the Long-Run? An Example of Likelihood Inference in a Multivariate Time Series Model," Paper presented at the Econometric Society Meeting, Maastricht.
- Kitamura, Y. (1995), "Estimation of Cointegrated Systems with $I(2)$ Processes," *Econometric Theory*, 11, 1–24.
- Maddala, G.S. (1991), "On the Use of Cointegration Tests," Paper presented at the Bank of Korea, Seoul in August, 1991.
- Ogaki, M. and J.Y. Park (1989), "A Cointegration Approach to Estimating Preference Parameters," Rochester Center for Economic Research Working paper No. 209, University of Rochester.
- Pantula, S.G. (1986), "Determining the Order of Differencing in Autoregressive Processes," Department of Statistics, North Carolina State University.
- Paruolo, P. (1994), "Asymptotic Efficiency of the Two- Stage Estimator in $I(2)$ Systems," University of Bologna.
- Phillips, P.C.B. (1991), "Optimal Inference in Cointegrated Systems," *Econometrica*, 59, 283–306.
- Phillips, P.C.B. and Y. Chang (1994), "Fully Modified Least Squares in $I(2)$ Regression," *Econometric Theory*, 10, 967.
- Phillips, P.C.B. and S. Ouliaris (1990), "Asymptotic Properties of Residual Based Tests for Cointegration," *Econometrica*, 58, 165–193.
- Sen, D.L. (1995), "Robustness of Single Unit Root Test Statistics in the Presence of Multiple Unit Roots," unpublished Ph.D. Dissertation, Department of Statistics, North Carolina State University.

- Stock, J.H. and M.W. Watson (1993), "A Simple Estimator of Cointegrating Vectors in Higher Order Integrated Systems," *Econometrica*, 61, 783–820.
- West, K.D. (1988), "Asymptotic Normality When Regressors Have a Unit Root," *Econometrica*, 56, 1397–1418.
- Yoo, B.S. (1986), "Multi-Cointegrated Time Series and Generalized Error Correction Models," Discussion paper, University of California at San Diego, CA.

12

Seasonal unit roots and seasonal cointegration

As mentioned in chapter 2 and discussed in subsequent chapters, we have two types of trend: deterministic trend and stochastic trend. Similar is the case of seasonals: deterministic seasonal and stochastic seasonal. Deterministic seasonals are taken care of by seasonal dummies; stochastic seasonals are taken care of by differencing. For instance, with quarterly data we use $\Delta_4 y_t = y_t - y_{t-4}$ and with monthly data we use $\Delta_{12} y_t = y_t - y_{t-12}$ if the seasonal is a stochastic seasonal. As with the operation of first-differences to take care of trends, the Box-Jenkins approach uses seasonal differences (with quarterly data it is fourth-difference, with monthly data it is twelfth-difference, and so on). Again the tests for seasonal roots is a formal generalization of this *ad hoc* procedure.

The paper by Davidson, Hendry, Srba, and Yeo (1978), to be referred to as DHSY, which used seasonal differences also had a strong influence on subsequent research that often used seasonal differences. The main focus of the paper by DHSY was the development of an appropriate econometric methodology for modeling dynamic relationships. After examining a number of different formulations, these authors selected a model in which seasonal lags and seasonal differences played a prominent role. As with the concept of cointegration in models with stochastic trends, we have the concept of seasonal cointegration in models with stochastic seasonals. We shall discuss these concepts in what follows.

The literature on seasonality is enormous. We shall only discuss a few main findings concerning seasonal unit roots and seasonal cointegration. For more comprehensive reviews see Hylleberg (1986, 1992, 1994), Ghysels (1993), and Franses (1996). The book by Franses discusses a type of seasonality where the parameters in an AR model vary with seasons.

This leads to the concepts of periodic unit roots and periodic cointegration. We shall discuss these concepts in later sections of this chapter.

There are a variety of possible models for seasonality which may differ across series. Three classes of time series models are commonly used to model seasonality. These are:

- (i) purely deterministic seasonal processes,
- (ii) stationary seasonal processes,
- (iii) integrated seasonal processes.

A *purely deterministic seasonal process* is a process generated by seasonal dummy variables. For quarterly series, a purely deterministic seasonal process can be described as

$$x_t = \alpha_0 + \alpha_1 S_{1t} + \alpha_2 S_{2t} + \alpha_3 S_{3t} + \varepsilon_t$$

where S_{1t} , S_{2t} , and S_{3t} are seasonal dummies. This process can be perfectly forecast and will never change its shape.

A *stationary seasonal process* can be generated by a potentially infinite autoregression

$$\varphi(L)x_t = \varepsilon_t, \quad \varepsilon_t \sim iid(0, \sigma^2)$$

with all of the roots of $\varphi(L) = 0$ lying outside the unit circle but where some are complex pairs with seasonal periodicities. More precisely, the spectrum of such a process is given by

$$f(w) = \frac{\sigma^2}{|\varphi(e^{iw})|^2}$$

which is assumed to have peaks at some of the seasonal frequencies w_s . An example for quarterly data is

$$x_t = \rho x_{t-4} + \varepsilon_t$$

which has a peak at both the seasonal periodicities $\pi/2$ (one cycle per year) and π (two cycles per year) as well as at zero frequency (zero cycles per year).

A series x_t is an *integrated seasonal process* if it has a seasonal unit root in its autoregressive representation. The properties of seasonally integrated series are not immediately obvious, but are quite similar to the properties of ordinary integrated processes. They have long memory so that shocks last forever and may in fact change permanently the seasonal patterns. Section 12.2 will discuss the seasonally integrated processes in detail. Before discussing seasonal integration it is worth

questioning what the effect of seasonal adjustment (or prefiltering data) on econometric inference is.

12.1 Effect of seasonal adjustment

Most of the empirical studies in econometrics use seasonally adjusted data. The question is: what is the effect of prefiltering (seasonally adjusting) data on econometric inference and hypothesis testing? Sims (1974) and Wallis (1974) investigated the nature of the asymptotic bias due to the seasonal noise in a linear regression model. They found that linear filtering does not induce asymptotic bias as long as uniform filtering is used. Using different filters for different series would produce asymptotic biases in linear regression models estimated with OLS. However, when lagged dependent variables are used, the effect is less clear.

Ghysels (1990) showed that unit root tests applied to post-war seasonality adjusted GNP series strongly support the null hypothesis of a unit root. But the evidence with seasonally unadjusted data is far less conclusive. However, when a unit root null is rejected, it is not clear whether the rejection is due to the absence of a unit root or the presence of seasonality. This problem has been pursued in Ghysels and Perron (1993). They considered the simple dynamic regression model with lagged dependent variables and investigated the effect of prefiltering on the behavior of the OLS estimator of the sum of the autoregressive coefficients. They show the existence of a limiting upward bias when the process does not contain a unit root. When the autoregressive process has a single unit root, then the OLS estimator of the first-order AR parameter is consistent and filtering does not affect the asymptotic distribution. This result holds for two-sided symmetric filters with weights adding up to one. For example, the widely used X-11 filter of the US Bureau of the Census has this property.

There remains the question of whether unit root tests should be applied with seasonally adjusted or seasonally unadjusted data. When seasonally adjusted data are used, Ghysels and Perron show that the nonvanishing asymptotic bias can be quite substantial. This is basically due to the fact that even though seasonal adjustment eliminates correlation in the data at seasonal frequencies (albeit imperfectly) it induces a bias in the autocorrelation function at lags less than the seasonal period which does not vanish even asymptotically. Hence, the ADF and Phillips–Perron statistics for testing a unit root will be biased toward nonrejection of the unit root null if filtered data are used. From

an asymptotic perspective, it is, therefore, expected that the unit root tests will have greater power if unadjusted data are used.

Ghysels and Perron (1993) also present an extensive simulation study of finite sample behavior emphasizing both the size and power of the usual ADF and Phillips–Perron statistics (both with and without an estimated trend) and show that, in many cases, there is a considerable reduction in power. Ghysels and Perron argue that their results provide additional justification for using annual data when testing for a unit root (see Perron, 1991 and Shiller and Perron, 1985).

Diebold (1993) in his comment on the paper by Ghysels and Perron disagrees with some of their results, but argues that Ghysels–Perron’s paper is a natural extension of the paper by Sims, Stock, and Watson discussed in chapter 7. Sims *et al.* point out that, if interest centers on point forecasts or point estimates of impulse response functions, VARs can be modeled in levels with trends terms included. One can then get consistent estimates whether there is a deterministic trend, stochastic trend, or cointegration and one does not have to impose any of them. The Ghysels–Perron paper leads to an extension of this result. One can specify VAR models using seasonally unadjusted data in their levels with deterministic trend and deterministic seasonal dummies added and get consistent parameter estimates whether the data show deterministic trend, stochastic trend, cointegration, deterministic seasonality, stochastic seasonality, seasonal integration, or seasonal cointegration. The paper by Canova (1993) on forecasting makes use of their results.

12.2 Seasonal integration

The Box–Jenkins methods discussed in chapter 2 used first-difference to eliminate trend and fourth-difference to eliminate seasonality in quarterly data and twelfth-difference to eliminate seasonality in yearly data. The work on seasonal integration is an attempt to formalize and expand on this procedure.

Let us define $\Delta_k = (1 - L^k)$ so that $\Delta_1 = (1 - L)$, $\Delta_4 = (1 - L^4)$ and $\Delta_{12} = (1 - L^{12})$. If y_t is the observed quarterly time series, the Box–Jenkins procedure amounts to applying the filter $\Delta_1\Delta_4$. But we can write

$$\begin{aligned}\Delta_1\Delta_4 &= (1 - L)(1 - L^4) \\ &= (1 - L)(1 - L)(1 + L)(1 + L^2) \\ &= (1 - L)^2(1 + L)(1 - iL)(1 + iL)\end{aligned}$$

where $i = \sqrt{-1}$. Thus $\Delta_1\Delta_4$ has two unit roots, and the other roots are -1 , $+i$, and $-i$. The two unit roots correspond to nonseasonal frequency, the root -1 corresponds to $1/2$ cycle per quarter or two cycles per year and the roots $+i$ and $-i$ correspond to $1/4$ cycle per quarter or one cycle per year. Given that the filter Δ_4 also has a unit root, there is the question of the appropriateness of the double filter $\Delta_1\Delta_4$. This has been investigated by Osborn (1990) for several quarterly UK macroeconomic time series using the methods developed in Osborn *et al.* (1988). The major conclusion is that there is no evidence in favor of the double filter $\Delta_1\Delta_4$ and that in practice at most the filter Δ_4 may be useful.

The test suggested in Osborn *et al.* (1988), commonly known as the Osborn test involves estimating (in the case of quarterly data) the regression

$$\begin{aligned}\Delta_1\Delta_4y_t &= c_0 + c_1S_{1t} + c_2S_{2t} + c_3S_{3t} + \beta_1\Delta_4y_{t-1} \\ &+ \beta_2\Delta_1y_{t-4} + \sum_{i=1}^k \phi_i\Delta_1\Delta_4y_{t-i} + u_t\end{aligned}$$

where S_{it} ($i = 1, 2, 3$) are seasonal dummies. The joint hypothesis about the usefulness of the $\Delta_1\Delta_4$ operator requires the joint test of the hypothesis $c_1 = c_2 = c_3 = \beta_1 = \beta_2 = 0$. The validity of the Δ_4 operator implies $\beta_2 = 0$ with $\beta_1 < 0$ and the validity of the Δ_1 operator requires $\beta_1 = 0$ and $\beta_2 < 0$. Critical values for these three different tests are reported in Osborn (1990, table 1), but for a sample size of 136 only.

12.3 Tests for seasonal unit roots

There is next the question of discussing in greater detail the appropriateness of the Δ_4 filter. A time series is said to be seasonally integrated if the Δ_4 filter is needed to make it stationary and the time series is said to have seasonal unit roots. We shall discuss three tests:

- (i) the DHF test by Dickey, Hasza, and Fuller (1984),
- (ii) the HEGY test by Hylleberg, Engle, Granger, and Yoo (1990),
- (iii) the CH test by Canova and Hansen (1995).

Another alternative suggested by Harvey and Scott (1994) will be discussed in the next section.

The DHF test

DHF analyze

$$y_t = \phi_4 y_{t-4} + u_t$$

when u_t is some stationary invertible ARMA process and consider testing the null hypothesis $\phi_4 = 1$. The alternative is $\phi_4 < 1$. As discussed earlier, the filter Δ_4 has four roots, one of which is a unit root. The DHF test rules out the possibility that the time series y_t can be made stationary by using the Δ_1 filter to take care of the unit root and seasonality can be taken care of by seasonal dummies. The DHF test tests the hypothesis that all the four roots implied by the filter Δ_4 are unity (one nonseasonal and three seasonal unit roots) versus the alternative of no unit roots.

The HEGY test

The HEGY test is a test for seasonal and nonseasonal unit roots in a quarterly time series. An extension of this test to monthly series is given in Franses (1991b) and Beaulieu and Miron (1993). The HEGY test is a test for the appropriateness of Δ_4 versus its nested components like $(1 - L)$ or $(1 + L)$. The HEGY test is based on a result on the decomposition of polynomials (for details, see Hylleberg *et al.*, 1990, p. 222).

The test is based on the following auxiliary regression

$$\begin{aligned} \Delta_4 y_t = & \sum_{s=1}^4 \alpha_s D_{st} + \gamma T_t + \pi_1 y_{1,t-1} + \pi_2 y_{2,t-1} \\ & + \pi_3 y_{3,t-2} + \pi_4 y_{3,t-1} + \sum_{i=1}^k \phi_i \Delta_4 y_{t-i} + \varepsilon_t \end{aligned} \quad (12.1)$$

where D_{st} are seasonal dummies, T_t is the trend, and

$$\begin{aligned} y_{1t} &= (1 + L + L^2 + L^3)y_t \\ y_{2t} &= -(1 - L + L^2 - L^3)y_t \\ y_{3t} &= -(1 - L^2)y_t \end{aligned}$$

If $\pi_1 = 0$, the series contains a unit root at the zero frequency which implies that the series contains a nonseasonal stochastic trend. If $\pi_2 = 0$, this implies two cycles per year. If $\pi_3 = 0$ and $\pi_4 = 0$, the series contains the roots i and $-i$, i.e., seasonal unit roots at the annual frequencies. The appropriate filters to use are $(1 - L)$ if $\pi_1 = 0$, $(1 + L)$ if $\pi_2 = 0$ and $(1 + L^2)$ if $\pi_3 = 0$ and $\pi_4 = 0$. Simulated critical values for the t -tests for the significance of π_1 and π_2 and for an F -test for the joint

significance of π_3 and π_4 are tabulated in Hylleberg *et al.* (1990). Table 12.1 presents some of their critical values.

The results of a Monte Carlo study by Hylleberg (1995) show that the size and the power of the HEGY test is reasonable when the DGP is an AR process. However, the HEGY test performs poorly in the presence of the MA errors and cannot cope with a very weakly changing seasonal components in addition to a strong deterministic seasonal pattern.

Franses (1990) extends the HEGY (1990) procedure to monthly time series. For monthly series the differencing operator Δ_{12} will have 12 roots on the unit circle such that

$$\begin{aligned} 1 - L^{12} &= (1 - L)(1 + L)(1 - iL)(1 + iL) \\ &\times [1 + (\sqrt{3} + i)L/2][1 + (\sqrt{3} - i)L/2] \\ &\times [1 - (\sqrt{3} + i)L/2][1 - (\sqrt{3} - i)L/2] \\ &\times [1 + (\sqrt{3} + i)L/2][1 - (\sqrt{3} - i)L/2] \\ &\times [1 - (\sqrt{3} + i)L/2][1 + (\sqrt{3} - i)L/2] \end{aligned}$$

where all terms other than $(1 - L)$ correspond to seasonal unit roots. Testing for unit roots in monthly time series is equivalent to testing for the significance of the parameters in the auxiliary equation

$$\begin{aligned} \varphi^*(L)y_{8,t} &= \pi_1 y_{1,t-1} + \pi_2 y_{2,t-1} + \pi_3 y_{3,t-2} + \pi_4 y_{3,t-1} + \pi_5 y_{4,t-2} \\ &+ \pi_6 y_{4,t-1} + \pi_7 y_{5,t-2} + \pi_8 y_{5,t-1} + \pi_9 y_{6,t-2} + \pi_{10} y_{6,t-1} \\ &+ \pi_{11} y_{7,t-2} + \pi_{12} y_{7,t-1} + \mu_t + \varepsilon_t \end{aligned} \quad (12.2)$$

where $\varphi^*(L)$ is some polynomial function of L for which the usual assumption applies, μ_t is the deterministic component, and where

$$\begin{aligned} y_{1,t} &= (1 + L)(1 + L^2)(1 + L^4 + L^8)y_t \\ y_{2,t} &= -(1 - L)(1 + L^2)(1 + L^4 + L^8)y_t \\ y_{3,t} &= -(1 - L^2)(1 + L^4 + L^8)y_t \\ y_{4,t} &= -(1 - L^4)(1 - \sqrt{3}L + L^2)(1 + L^2 + L^4)y_t \\ y_{5,t} &= -(1 - L^4)(1 + \sqrt{3}L + L^2)(1 + L^2 + L^4)y_t \\ y_{6,t} &= -(1 - L^4)(1 - L^2 + L^4)(1 - L + L^2)y_t \\ y_{7,t} &= -(1 - L^4)(1 - L^2 + L^4)(1 + L + L^2)y_t \\ y_{8,t} &= (1 - L^{12})y_t \end{aligned}$$

Table 12.1. *Critical values for seasonal unit roots in quarterly data*

Auxiliary Regressions	T	$t : \pi_1 = 0$		$t : \pi_2 = 0$		$t : \pi_3 = 0$		$t : \pi_4 = 0$		$F : \pi_3 = \pi_4 = 0$	
		1%	5%	1%	5%	1%	5%	1%	5%	95%	99%
No intercept	48	-2.72	-1.95	-2.67	-1.95	-2.66	-1.93	-2.51	-2.11	3.26	5.02
No seas. dum.	100	-2.60	-1.97	-2.61	-1.92	-2.55	-1.90	-2.43	-2.01	3.12	4.89
No trend	136	-2.62	-1.93	-2.60	-1.94	-2.58	-1.92	-2.44	-1.99	3.14	4.81
	200	-2.62	-1.94	-2.60	-1.95	-2.58	-1.92	-2.43	-1.98	3.16	4.81
Intercept	48	-3.66	-2.96	-2.68	-1.95	-2.64	-1.90	-2.44	-2.06	3.04	4.78
No seas. dum.	100	-3.47	-2.88	-2.61	-1.95	-2.61	-1.90	-2.38	-1.99	3.08	4.77
No trend	136	-3.51	-2.89	-2.60	-1.91	-2.53	-1.88	-2.36	-1.98	3.00	4.73
	200	-3.48	-2.87	-2.58	-1.92	-2.57	-1.90	-2.36	-1.98	3.12	4.76
Intercept	48	-3.77	-3.08	-3.75	-3.04	-4.31	-3.61	-2.86	-2.37	6.60	9.22
Seas. dum.	100	-3.55	-2.95	-3.60	-2.94	-4.06	-3.44	-2.78	-2.32	6.57	8.74
No trend	136	-3.56	-2.94	-3.49	-2.90	-4.06	-3.44	-2.72	-2.31	6.63	8.92
	200	-3.51	-2.91	-3.50	-2.89	-4.00	-3.38	-2.74	-2.33	6.61	8.93
Intercept	48	-4.23	-3.56	-2.65	-1.91	-2.68	-1.92	-2.41	-2.05	2.95	4.64
No seas. dum.	100	-4.07	-3.47	-2.58	-1.94	-2.56	-1.89	-2.38	-1.97	2.98	4.70
Trend	136	-4.09	-3.46	-2.65	-1.96	-2.56	-1.90	-2.36	-1.97	3.04	4.57
	200	-4.05	-3.44	-2.59	-1.95	-2.58	-1.92	-2.35	-1.97	3.07	4.66
Intercept	48	-4.46	-3.71	-3.80	-3.08	-4.46	-3.66	-2.75	-2.26	6.55	9.27
Seas. dum.	100	-4.09	-3.53	-3.60	-2.94	-4.12	-3.48	-2.76	-2.32	6.60	8.79
Trend	136	-4.15	-3.52	-3.57	-2.93	-4.05	-3.44	-2.71	-2.78	6.62	8.77
	200	-4.05	-3.49	-3.52	-2.91	-4.04	-3.41	-2.65	-2.27	6.57	8.96

Note: The data generating process (DGP) is $\Delta_4 x_t = \varepsilon_t \sim iin(0, 1)$.

Source: tables 1a and 1b of Hylleberg *et al.* (1990, pp. 226–227).

Furthermore, the μ_t in equation (12.2) covers the deterministic part and might consist of a constant, seasonal dummies, or a trend. This depends on the hypothesized alternative to the null hypothesis of 12 unit roots.

Applying OLS to equation (12.2) gives estimates of the π_i s. In case there are (seasonal) unit roots, the corresponding π_i are zero. Due to the fact that pairs of complex unit roots are conjugates, it should be noted that these roots are only present when pairs of π s are equal to zero simultaneously, for example, the roots i and $-i$ are only present when π_3 and π_4 are equal to zero (see Franses, 1990, for detailed derivation). There will be no seasonal unit roots if π_2 through π_{12} are significantly different from zero. If $\pi_1 = 0$, then the presence of root 1 cannot be rejected. When $\pi_1 = 0$ and π_2 through π_{12} are not equal to zero, the seasonals can be modeled with seasonal dummies and the first-differenced series. In case all $\pi_i, i = 1, \dots, 12$, are equal to zero, it is appropriate to apply the Δ_{12} filter. Extensive tables with critical values for t -tests of the separate π s, and for F -tests of pairs of π s, as well as for a joint F -test of $\pi_3 = \dots = \pi_{12}$ can be found in Franses (1990). Table 12.2 shows part of critical values tabulated by Franses (1990).

The CH test

Canova and Hansen (1995) propose a test for the null of stationarity against seasonal unit roots which extends the KPSS test (discussed in chapter 4) to seasonal data. The asymptotic distributions are nonstandard and the critical values are provided in their paper. By a Monte Carlo study Hylleberg (1995) found that the CH test performs poorly under the presence of unit roots at other frequencies and moving-average errors.

However, as argued in Hylleberg and Pagan (1996), the HEGY test and the CH test are not directly comparable, much the same way that a test for unit root and a test for stationarity are not comparable (the null hypotheses are different and the alternatives are reversed). Thus, each of the studies can start with DGPs appropriate for their test and show that the other test does not perform well.

Hylleberg and Pagan suggest an evolving seasonal model (ESM) in which the coefficients attached to seasonal trigonometric functions follow simple autoregressive processes, and discuss its relationship to the HEGY and CH tests. Since the ESM nests the models considered by HEGY and CH, Hylleberg and Pagan are able to make comparisons between the two tests and discuss their limitations. They also suggest how

Table 12.2. Critical values for seasonal unit roots in monthly data

	Regression							
	with trend				without trend			
<i>t</i> -statistics	0.05		0.10		0.05		0.10	
π_1	-3.24		-2.92		-2.63		-2.35	
π_2	-2.65		-2.39		-2.65		-2.40	
<i>t</i> -statistics	0.0025	0.05	0.95	0.975	0.0025	0.05	0.95	0.975
π_3	-2.05	-1.71	1.72	2.10	-2.11	-1.76	1.74	2.11
π_4	-3.34	-3.12	-0.45	-0.15	-3.34	-3.12	-0.44	-0.14
π_5	-3.29	-2.99	-0.06	0.24	-3.29	-3.00	-0.05	0.25
π_6	-3.38	-3.12	-0.44	-0.11	-3.39	-3.12	-0.42	-0.09
π_7	-0.18	0.12	2.98	3.28	-0.27	0.05	3.00	3.31
π_8	-3.40	-3.15	-0.43	-0.17	-3.39	-3.14	-0.42	-0.18
π_9	-2.86	-2.54	0.81	1.12	-2.87	-2.54	0.82	1.13
π_{10}	-3.36	-3.07	-0.40	-0.09	-3.37	-3.07	-0.39	-0.07
π_{11}	-1.08	-0.73	2.55	2.80	-1.11	-0.78	2.56	2.83
π_{12}	-3.42	-3.16	-0.44	-0.17	-3.43	-3.16	-0.42	-0.14
<i>F</i> -statistics	0.90	0.95			0.90	0.95		
π_1, π_4	4.81	5.63			4.83	5.62		
π_5, π_6	4.86	5.84			4.89	5.86		
π_1, π_5	4.94	5.90			4.94	5.86		
π_9, π_{10}	4.76	5.71			4.79	5.75		
π_{11}, π_{12}	4.92	5.84			4.94	5.89		
$\pi_1, \dots, \pi_{12}$	4.00	4.45			4.00	4.46		

Notes: The DGP is $y_t = y_{12} + \varepsilon_t, \varepsilon_t \sim iin(0, 1), T = 120$.

The test for π_1 and π_2 are one-sided tests,
while the other *t*-tests are two-sided.

Source: Franses (1990, pp. 12–18)

all the unit root tests discussed in chapter 4, such as alternative tests for stationarity and efficient tests for unit roots by Elliott *et al.* can be extended to tests of seasonal unit roots.

12.4 The unobserved component model

There are two principal ways to model time series with trend, seasonal, and irregular components. The first is to fit a seasonal ARIMA model, such as *airline model* in Box and Jenkins (1970), and then decompose it by maximizing the variance of the irregular components. The second

is to specify explicitly models for the unobserved trend, seasonal, and irregular components. This is called the *unobserved component model* or *structural time series model*. The unobserved components model is widely used as a workhorse in the statistical time series literature on seasonality. The history of the unobserved components model can be found in Nerlove, Grether, and Carvalho (1979), Bell and Hillmer (1984), and Hylleberg (1986).

Consider

$$y_t = \mu_t + \gamma_t + \varepsilon_t$$

where μ_t is the trend, γ_t the seasonal, and ε_t the irregular component. ε_t is white noise, and thus stationary. The trend μ_t is only stationary in first- or second-differences and γ_t is stationary when multiplied by the seasonal summation operator

$$S(L) = 1 + L + L^2 + \cdots + L^{s-1}$$

where s is the number of seasons and $S(L)$ contains both real and complex unit roots. The seasonal γ_t is said to be seasonally integrated (see section 12.2). It ensures that the seasonal component, which is nonstationary, is not confounded with the trend and that the seasonal pattern projected into the future sums to zero over s consecutive time periods. In the context of structural time series modeling, the trend component is formulated as

$$\mu_t = \mu_{t-1} + \beta_{t-1} + \eta_t, \quad \eta_t \sim \text{nid}(0, \sigma_\eta)$$

$$\beta_t = \beta_{t-1} + \zeta_t, \quad \zeta_t \sim \text{nid}(0, \sigma_\zeta)$$

This collapses to a random walk plus drift if $\sigma_\zeta^2 = 0$ and to a deterministic linear trend if $\sigma_\eta^2 = 0$ as well. Setting σ_η^2 to zero when σ_ζ^2 is positive tends to give a trend which changes relatively smoothly. As regards seasonality, a number of different specifications exist, the simplest one is

$$S(L)\gamma_t = \sum_{j=0}^{s-1} \gamma_{t-j} = w_t, \quad w_t \sim \text{iin}(0, \sigma_w) \quad (12.3)$$

This is known as the dummy variable form of stochastic seasonality. It reduces to a standard deterministic component if σ_w^2 is zero. Another formulation, trigonometric seasonality, can be written in the following

way when s is even

$$\gamma_t = \sum_{j=1}^{s/2} \gamma_{j,t} \quad (12.4)$$

where

$$\begin{aligned} \gamma_{j,t} &= \gamma_{j,t-1} \cos \lambda_j + \gamma_{j,t-1}^* \sin \lambda_j + w_{jt} \\ \gamma_{j,t}^* &= \gamma_{j,t-1} \sin \lambda_j + \gamma_{j,t-1}^* \cos \lambda_j + w_{jt}^* \end{aligned}$$

for $j = 1, \dots, s/2 - 1$ and $\lambda_j = 2\pi j/s$, and

$$\gamma_{j,t} = -\gamma_{j,t-1} + w_{j,t}, \quad j = s/2$$

where w_{jt} and w_{jt}^* are normally distributed, zero mean, white noise processes. The advantage of (12.4) over (12.3) is that it allows for smoother changes in the seasonals, while the advantage of (12.3) is analytical simplicity and is often used in the context of developing theoretical points. All the disturbances are assumed mutually uncorrelated. The extent to which trend and seasonal components evolve over time depends on the parameters $\sigma_\eta^2, \sigma_\zeta^2, \sigma_w^2$, and σ_ε^2 which can be estimated by maximum likelihood in the time or frequency domain (see Harvey, 1989, chapter 4).

Harvey and Scott (1994) (hereafter denoted by HS) show how seasonality can be incorporated into ECMs using an unobserved components model and compared this with the earlier model studied by DHSY (1978). HS include the stochastic seasonal γ_t into the ECM such that

$$\Delta y_t = \theta + \gamma_t + \beta \Delta x_t + \lambda(y_{t-1} - \alpha x_{t-1}) + \varepsilon_t, \quad t = 2, \dots, T$$

$$\sum_{j=0}^{s-1} \gamma_{t-j} = w_t, \quad w_t \sim \text{iin}(0, \sigma_w^2)$$

Setting σ_w^2 to zero generates a deterministic seasonal pattern, while a nonzero value of σ_w^2 allows the seasonal pattern to evolve over time. The higher is σ_w^2 relative to σ_ε^2 the greater is the importance of seasonal influences in accounting for variability in Δy_t . Applying $S(L)$ yields

$$\Delta_s y_t = s\theta + \beta \Delta_s x_t + \lambda(Y_{t-1} - \alpha X_{t-1}) + u_t, \quad t = s+1, \dots, T$$

where $\Delta_s = \Delta \cdot S(L)$ and

$$u_t = S(L)\varepsilon_t + w_t$$

$$Y_t = S(L)y_t$$

$$X_t = S(L)x_t$$

Thus Y_{t-1} is the sum of the observations over the year ending in period $t - 1$ and similarly for X_{t-1} . If the disturbance u_t is assumed to be white noise, the model becomes very similar to the DHSY specification. DHSY tried to capture seasonality by means of a general lag structure and selected the following ECM

$$\Delta_s y_t = \theta + \beta \Delta_s x_t + \lambda(y_{t-s} - \alpha x_{t-s}) + v_t$$

where v_t is assumed to be white noise. The difference between the two models lies in the error correction term; the DHSY specification uses the quarterly term (if $s = 4$) $y_{t-s} - \alpha x_{t-s}$, while the HS specification uses the annual aggregate $Y_{t-1} - \alpha X_{t-1}$. Because these terms differ by an order of magnitude s we would expect the estimate of λ arising from the DHSY model to be approximately s times that from the HS model. Note that the DHSY model is a special case of the HS model in which there is no irregular component, i.e., $\sigma_\varepsilon^2 = 0$ and u_t simply equals the seasonal white noise disturbance term w_t .

HS argue that stochastic seasonality (for example, in ECM) has important implications for autoregressive distributed lag modeling. HS criticize seasonal unit root tests developed by HEGY because of (i) poor size and power properties and (ii) wrong hypothesis. Especially about (ii) HS argue that the issue is not whether seasonal movements are stationary or nonstationary, but whether they are stochastic (nonstationary) or deterministic. Thus, HS suggest that it is safer to proceed with the more general assumption that the seasonal is stochastic rather than imposing the constraint that it be deterministic. They show that letting a seasonal component in a regression model be stochastic has little impact on efficiency. On the other hand, the use of deterministic seasonals when seasonality is stochastic is likely to lead to spurious dynamics.

HS estimate the above unobserved components model with the trigonometric specification of seasonals for quarterly UK nondurable consumption and disposable income from 1957:III to 1992:II. For nondurable consumption they found that the fitted model may be written as

$$c_t = \mu_t + \gamma_t$$

where μ_t is a random walk plus drift. The drift parameter was estimated as 0.0069 with a standard error of 0.0008 and $\hat{\sigma}_\varepsilon^2 = 0$. The nonzero estimate of σ_w^2 implies that the seasonal component γ_t changes over

time. Taking seasonal differences yields

$$\Delta_s c_t = S(L)\eta_t + s\beta + \Delta S(L)\gamma_t$$

where $\Delta_s = 1 - L^s = \Delta \cdot S(L)$. When $S(L)\gamma_t$ is at most an $\text{MA}(s-2)$ process, $\Delta_s c_t$ is at most an $\text{MA}(s-1)$ process and so

$$E_t \Delta_s c_{t+s} = s\beta$$

indicating that the expected value of consumption this time next year is what it is now plus trend growth. Hence this model is a generalization of the random walk theory of consumption, which takes account of the fact that consumption will be different in different seasons and which implies a degree of predictability when forecasting less than one year ahead. The authors use a computer package called STAMP to perform the ML estimation. (Details of this package can be obtained from the authors.)

HS also applied the ECM with the stochastic seasonals for UK consumption and income. They found that the DHSY model provided a good fit over the sample period used, 1958:II–1970:IV, but it breaks down when the sample is extended to 1958:II–1992:II. The general HS model which encompassed the DHSY model as a special case shows that it is possible to arrive at a stable ECM between consumption and income over the extended period 1958:II–1992:II.

12.5 Seasonal cointegration

The HEGY procedure allows one to separately test for the four roots: 1, -1 , $\pm i$ implied by the Δ_4 filter. Corresponding to this Engle, Granger, Hylleberg, and Lee (1993) (to be referred to as EGHL) suggest different levels of seasonal cointegration. Suppose that two series y_t and z_t are seasonally cointegrated so that $\Delta_4 y_t$ and $\Delta_4 z_t$ are stationary. When there is cointegration at the zero frequency, i.e., at the root 1, we have

$$u_t = (1 + L + L^2 + L^3)y_t - \alpha_1(1 + L + L^2 + L^3)z_t \quad (12.5)$$

is stationary. When there is cointegration at frequency $1/2$, i.e., at the root (-1) , we have

$$v_t = (1 - L + L^2 - L^3)y_t - \alpha_2(1 - L + L^2 - L^3)z_t \quad (12.6)$$

is stationary. Finally when there is cointegration at frequency $1/4$, i.e., at the root $\pm i$ or corresponding to the filter $(1 - L^2)$, we have

$$w_t = (1 - L^2)y_t - \alpha_3(1 - L^2)z_t - \alpha_4(1 - L^2)y_{t-1} - \alpha_5(1 - L^2)z_{t-1} \quad (12.7)$$

is stationary. We have to consider cointegration with one period lagged because the filter $(1 - L^2)$ covers only the half-year period.

In the case where all the three variables u_t , v_t , and w_t are stationary a simple version of the seasonal cointegration model is

$$\begin{aligned}\Delta_4 y_t &= \beta_{11}u_{t-1} + \beta_{21}v_{t-1} + \beta_{31}w_{t-2} + \beta_{41}w_{t-3} + e_{1t} \\ \Delta_4 z_t &= \beta_{21}u_{t-1} + \beta_{22}v_{t-1} + \beta_{32}w_{t-2} + \beta_{42}w_{t-3} + e_{2t} \quad (12.8)\end{aligned}$$

where the β s are the error correction terms. One can include a constant, seasonal dummies, and trend in these equations.

EGHL propose tests for seasonal cointegration similar to the Engle-Granger two-step test for cointegration, i.e., estimate equations (12.5)–(12.7) by OLS and apply unit root tests (ADF tests) to $\hat{u}_t$, $\hat{v}_t$, and $\hat{w}_t$. The critical values for the tests applied to $\hat{u}_t$ and $\hat{v}_t$ are the same as those tabulated in Engle and Granger (1987), but for the test applied to $\hat{w}_t$ they are different. EGHL derive the asymptotic distributions in this case and tabulate the critical values.

Of course, in principle, any of the cointegration tests discussed in chapter 6 can be applied here, but the appropriate asymptotic distributions have not been studied yet. Also, ECM tests applied to (12.8) need further study. However, some more fundamental problems with the seasonal cointegration model analyzed by EGHL need to be addressed before all these extensions can be carried out. In her comment on the EGHL paper, Osborn argues that the seasonal cointegration model (12.8) implies that the equilibrium relations between y_t and z_t vary with the lag, e.g., the long-run relation at time $(t - 1)$ is different from that at time $(t - 2)$. She argues that a more reasonable model is one where coefficients vary with seasons. This leads to the periodic cointegration model, which we shall discuss in section 12.9 later.

12.6 Estimation of seasonally cointegrated systems

In the previous section we discussed the EGHL extension of the Engle-Granger two-step method to the case of seasonal cointegration. Lee (1992) extends the Johansen procedure to the case of seasonal cointegration. The procedure consists of the same type of decomposition as adopted earlier with the HEGY test as in equation (12.1).

Starting with a vector Y_t of $I(1)$ variables and a VAR, the equation we end up with corresponding to (12.1) is (for quarterly data)

$$\Delta_4 Y_t = AD_t + \Pi_1 Y_{1,t-1} + \Pi_2 Y_{2,t-1} + \Pi_3 Y_{3,t-2}$$

$$+ \Pi_4 Y_{3,t-1} + \sum_{i=1}^{p-4} \Gamma_i \Delta_4 Y_{t-i} + \varepsilon_t \quad (12.9)$$

where D_t denotes the matrix of the deterministic variables (constant, trend and seasonal dummies) and

$$\begin{aligned} Y_{1t} &= (1 + L + L^2 + L^3)Y_t \\ Y_{2t} &= -(1 - L + L^2 - L^3)Y_t \\ Y_{3t} &= -(1 - L^2)Y_t \end{aligned}$$

The ranks of the matrices Π_1, Π_2 , and Π_3 determine the number of cointegrating vectors at zero, $1/2$, and $1/4$ frequencies. Lee derives trace statistics as done in the Johansen procedure and derives the asymptotic distributions. The details are too lengthy to be reproduced here. They can be found in Lee's paper. Lee points out that the distribution of the ML cointegration test statistics may be quite different from their asymptotic distributions. In fact, we tend to reject the true null hypothesis more often than what is suggested by the asymptotic distribution. This bias, however, decreases in very large samples. Hence Lee and Siklos (1995) present finite sample critical values for the Johansen seasonal cointegration test statistics. In particular, they consider a DGP which contains unit roots at the seasonal and zero frequencies, as well as a time trend. They present these for sample sizes 50, 100, 150, 200, and ∞ and also present a comparison of finite versus asymptotic critical values (see tables 1, 2, 3 of their papers. We shall not reproduce those tables here.)

Kunst (1992) applies Lee's procedure to analyze seasonal cointegration in the Austrian economy using quarterly data for 1964–1990 on five variables. Although he finds evidence of seasonal cointegration, this does not help in obtaining better forecasts. Kunst (1993) contains more applications of Lee's procedure.

The regularity conditions for using the Johansen procedure require that the DGP does not have unit roots other than the zero frequency. However, Ghysels *et al.* (1994) show that seasonal unit roots do not pose a problem in the use of the Johansen procedure, which can still be used to test the hypothesis of cointegration at the zero frequency.

There are other extensions of cointegration methods (besides the Johansen procedure) to seasonal cointegration. For instance, Joyeux (1992) suggests an extension of the frequency domain cointegration tests of Phillips and Ouliaris (see chapter 6) to the case of seasonal cointegra-

tion. In principle all the methods discussed in chapters 5 and 6 can be extended to seasonal cointegration

12.7 Empirical evidence

EGHL (1993) applied the testing procedure discussed in the section 12.3 (except the test of seasonal cointegration at frequency $1/4$ with estimated cointegration vector) to consumption ($c = \log$ of consumption expenditures on durables) and income ($y = \log$ of personal disposable income) in UK from 1955:I to 1984:IV. They investigate the permanent income hypothesis where the log of income and the log of consumption may be thought to be cointegrated with the cointegrating vector $(1, -1)$. Their results show strong evidence for the presence of a unit root at zero frequency in all c, y , and $c - y$ which implies that there is no cointegration between c and y at the long-run frequency, at least not for the cointegrating vector $(1, -1)$. The hypotheses that c, y , and $c - y$ are $I_{1/2}(1)$ cannot be rejected and thus c and y are not cointegrated at the bi-annual cycle either. The hypothesis that y is $I_{1/4}(0)$ is rejected, but that for c is not, and thus c and y cannot possibly be cointegrated at this frequency. They conclude that the unit-elasticity error-correction model is not valid at any frequency for the UK consumption and income data.

EGHL (1993) applied the above testing procedure (including the case with estimated cointegration vectors) to total consumption and disposable income in Japan from 1961:I to 1987:IV. Their results indicate that the log of the income series is integrated of order 1 at both the long-run frequency ($\theta = 0$) and at the seasonal frequencies. This implies that income is nonstationary and that the seasonal pattern has significant variation over the period. A similar result is obtained for consumption, although the seasonal pattern is more regular and it is in fact a question whether a deterministic annual seasonal pattern is to be preferred. The results of the cointegration analysis show that the log of the two series is not cointegrated at any frequency with the cointegrating vector $(1, -1)$. It is also shown that the seasonally adjusted data, which are adjusted for seasonal unit roots by summing over four consecutive quarters, are not cointegrated at the long-run frequency, neither with cointegrating vector $(1, -1)$ nor with an estimated cointegrating vector. The test for seasonal cointegration indicates no cointegration at the bi-annual frequency, while there may be signs of cointegration at the annual frequency. This indicates that

the annual seasonal components of consumption and income are similar.

In her comment on EGHL, Osborn (1993) argues that as long as there is no good economic rationalization for seasonal unit roots, we should not worry about them (see the criticism of EGHL in section 12.5).

Sarantis and Stewart (1993) apply the HEGY test to nine exchange rate series and relative prices and conclude that none of the exchange rates and relative prices contain seasonal unit roots, but all have an autoregressive unit root. On the other hand, Ghysels, Lee, and Siklos (1993) and Lee and Siklos (1991) examine respectively, for the US and Canada, several macroeconomic series and find unit roots at some seasonal (but not all) frequencies. Further evidence on seasonal unit roots is also presented in Franses (1996) for 11 quarterly series from the US, Canada, Germany, and the UK, and Granger and Siklos (1995) for international data consisting of 16 quarterly and eight monthly time series. The purpose of this latter study is to investigate the effect of skip sampling and time aggregation. Overall these latter studies show more evidence of seasonal unit roots at least at some seasonal frequencies.

To investigate the robustness of the results from the seasonal unit root tests, Franses (1996) also investigates recursive unit root tests. The results suggest that the observed significance of the seasonal unit root test statistics, might be due to seasonal mean shifts. The evidence in favor of seasonal unit roots may become less pronounced, or may even disappear, when one allows for seasonal mean shifts. This observation is similar to the one about the evidence in favor of unit roots when one allows for broken trends, as noted by Perron (1989).

Also, as discussed in section 12.4, Harvey and Scott point out the limitations of the seasonal unit root tests suggested by Hylleberg if there are moving-average errors. Thus, there are several parallels in the limitations between the unit root tests discussed in chapters 3 and 4 and the seasonal unit root tests discussed here.

12.8 Periodic autoregression and periodic integration

Here we shall discuss briefly the concepts of *periodic autoregression* and *periodic integration*. More detailed discussion can be found in Osborn (1990, 1991), Lütkepohl (1991), and the comprehensive book by Franses (1996).

Periodic autoregression

A periodic autoregression (PAR) model is one where the autoregressive parameters vary with the seasons, e.g., if we consider an AR(1) model, the PAR(1) model is (for quarterly data):

$$y_t = \phi_s y_{t-1} + u_t, \quad s = 1, 2, 3, 4$$

When t falls in season s , y_t is seasonal because of seasonal variation in the autoregressive parameter ϕ .

One important point to note about periodic models is that although using annual data are thought to remove seasonality, the effects of periodic autoregressive coefficients are not necessarily removed by taking annual sums. One common complaint against PAR models is that they increase the number of parameters estimated, but this is the case only if one uses a long autoregression and say monthly data. However, it is shown in Osborn (1991) that the orders of time invariant models can be higher than that of the PAR model. Neglecting periodicity may necessitate the inclusion of several (seasonal) lags in a nonperiodic model.

Periodic integration

Define the operator

$$\delta_s y_t = y_t - \phi_s y_{t-1}, \quad s = 1, 2, 3, 4$$

Then we say that y_t is periodically integrated of order 1 or PI(1) if y_t is nonstationary and $\delta_s y_t$ is stationary. This concept is related to time varying parameter integration TVP(1) discussed by Granger (1986, p. 224).

It is shown in Osborn (1991) as well as Franses (1996, p. 99) that y_t is stationary if $\phi_1 \phi_2 \phi_3 \phi_4 < 1$. If $\phi_1 \phi_2 \phi_3 \phi_4 = 1$ the PAR(1) process has a unit root. A special case of interest is $\phi_i = \phi$ for all i . In this case the condition $\phi_1 \phi_2 \phi_3 \phi_4 = 1$ can hold if $\phi = 1$ and also if $\phi = -1$. The first case corresponds to the case where y_t has a nonseasonal unit root and the second case corresponds to a seasonal unit root (half-year cycle). Thus, both seasonal and nonseasonal unit roots are nested in the PAR(1) model. Franses suggests first testing whether $\phi_1 \phi_2 \phi_3 \phi_4 = 1$ and second testing whether $\phi_i = 1$ or $\phi_i = -1$ for all i .

As for estimation of PAR processes, one can use OLS methods which give maximum likelihood estimators under the assumption of normality. The next step is to test the hypothesis of the constancy of the autoregressive parameters over the different seasons. For a PAR(p) process,

one tests the hypothesis (for quarterly data)

$$H_0 : \phi_{is} = \phi_i \quad \text{for } s = 1, 2, 3, 4 \text{ and } i = 1, 2, \dots, p$$

It has been shown in Boswijk and Franses (1996) that the likelihood ratio test for this null hypothesis has an asymptotic $\chi^2(3p)$ distribution irrespective of whether the y_t series has nonseasonal or seasonal unit roots.

The concept of periodic autoregression can be extended to VAR models as well (although this results in too many parameters to be estimated). This extension is discussed in Lütkepohl (1991) and Franses (1996, p. 124).

Tests for periodic unit roots

As discussed earlier, the PAR(1) process has a unit root if $\phi_1\phi_2\phi_3\phi_4 = 1$. Tests for periodic unit roots are slightly more complicated than tests for seasonal unit roots. Since a discussion of these tests involves more algebra, we shall skip the details and give the appropriate references: Boswijk and Franses (1995a, 1996) and Franses (1996, pp. 127–149).

12.9 Periodic cointegration and seasonal cointegration

Periodic cointegration refers to the case of seasonally varying long-run relationships. A detailed discussion of this concept involves repeating a lot of algebra from Franses (1996). Hence we refer to the sources: Boswijk and Franses (1995b) and Franses (1996, chapter 9).

Periodic models involve many more parameters than seasonal models. If one has a lot of data, they are worth considering because they make more sense. But with few observations, of course, one cannot afford the luxury of estimating so many parameters.

It has been well documented that seasonal patterns are not constant over time. One alternative to the periodic model is the evolving seasonal model discussed in Hylleberg and Pagan (1996). This involves estimation of fewer parameters than the periodic model.

12.10 Time aggregation and systematic sampling

Most macroeconomic data are obtained from time aggregation or systematic sampling. Granger and Siklos (1995) investigate the effects of these procedures on tests for seasonal and nonseasonal unit roots and on tests for cointegration. Systematic sampling occurs when observations

are at fixed intervals, e.g., end of March, June, September, and December for quarterly data. This is the case for variables like interest rates and prices. Temporal aggregation occurs when flows are aggregated over some fixed periods of time. This is the case with GNP, consumption, money supply, and so on.

Granger and Siklos also investigate the effects of official seasonal adjustment procedures on tests for seasonal unit roots (see section 12.1 for earlier discussion of effects on tests for unit roots). They show that:

- (i) A time series which is integrated of order 1 at the zero frequency, denoted by $I_0(1)$ can arise because of systematic sampling and/or seasonal adjustment. This suggests that practicioners should resort to unadjusted data.
- (ii) Two time series can appear cointegrated at the zero frequency but this can be due to a unit root at zero frequency in one series and a seasonal unit root in another.
- (iii) We can miss seasonal unit roots when using skip sampling. For instance the existence of seasonal unit roots in, say, monthly data can lead to missing seasonal unit roots with quarterly data. This problem stems from skip sampling while temporal aggregation produces no such results. They find that M1 has a unit root at the bi-annual frequency when monthly data are averaged, but not when the third monthly observation is used to generate quarterly data.
- (iv) Cointegration relationships are affected by the type of seasonal adjustments used. It is important to consider cointegration by seasonal frequency.

We shall not go into the detailed derivations here. Granger and Siklos present empirical evidence on these propositions using an international data set on US, UK, Canada, France, Japan, Germany, and Italy with monthly and quarterly observations.

12.11 Conclusion

The present chapter discusses the statistical issues relating to seasonality in time series. The problems discussed are the effects of seasonal adjustment procedures on tests for unit roots, sesonal integration and seasonal cointegration, tests for sesonal unit roots, periodic integration and cointegration, and effects of time aggregation and skip sampling. There is a large amount of literature on the economics of seasonality, which we

have not discussed here. See the references in Hylleberg (1994) and the papers in Hylleberg (1992). As mentioned in Hylleberg (1994, pp. 171–172), some things about seasonality are widely accepted. These are:

- (i) It is the best to use seasonally unadjusted data, rather than *officially* seasonally adjusted data.
- (ii) Seasonal variation often accounts for a major part of variation in time series.
- (iii) The seasonal and nonseasonal components are dependent and economic considerations are important in explaining both the components.

There is some controversy over the relative merits of modeling seasonality through periodic models.

References

- Beaulieu, J.J. and J.A. Miron (1993), "Seasonal Unit Roots in Aggregate US Data", *Journal of Econometrics*, 55, 305–328.
- Bell, W.R. and S.C. Hillmer (1984), "Issues Involved with the Seasonal Adjustment of Economic Time Series," *Journal of Business and Economic Statistics*, 2, 526–534.
- Boswijk, H.P. and P.H. Franses (1995a), "Testing for Periodic Integration," *Economic Letters*, 48, 241–248.
- (1995b), "Periodic Cointegration : Representation and Inference," *Review of Economics and Statistics*, 77, 436–454.
- (1996), "Unit Roots in Periodic Autoregression," *Journal of Time Series Analysis*, 17, 221–245.
- Box, G.E.P. and G.M. Jenkins (1970), *Time Series Analysis Forecasting and Control*, Holden-Day, San Francisco.
- Canova, F. (1993), "Forecasting Time Series with Common Seasonal Patterns," *Journal of Econometrics*, 55, 173–200.
- Canova, F. and B.E. Hansen (1995), "Are Seasonal Patterns Constant over Time: A Test for Seasonal Stability," *Journal of Business and Economic Statistics*, 13, 237–252.
- Davidson, J.E.H., D.F. Hendry, F. Srba, and S. Yeo (1978), "Econometric Modeling of the Aggregate Time Series Relationship between Consumers' Expenditure and Income in the UK," *The Economic Journal*, 88, 661–692.

- Dickey, D.A., H.P. Hasza, and W.A. Fuller (1984), "Testing for Unit Roots in Seasonal Time Series," *Journal of the American Statistical Association*, 79, 355–367.
- Diebold, F.X. (1993), "Discussion : The Effect of Seasonal Adjustment Filters on Tests for a Unit Root," *Journal of Econometrics*, 55, 99–103.
- Engle, R.F. and C.W.J. Granger (1987), "Co-integration and Error Correction: Representations, Estimation and Testing," *Econometrica*, 55, 252–276.
- Engle, R.F., C.W.J. Granger, S. Hylleberg, and H.S. Lee (1993), "Seasonal Cointegration: The Japanese Consumption Function," *Journal of Econometrics*, 55, 275–298.
- Franses, P.H. (1990), "Testing for Seasonal Unit Roots in Monthly Data," Econometric Institute Report No. 9032A, Erasmus university, Rotterdam.
- (1991a), *Model Selection and Seasonality in Time Series*, Timbergen Institute: Thesis Publisher.
- (1991b), "Seasonality, Nonstationarity and the Forecasting of Monthly Time Series," *International Journal of Forecasting*, 7, 199–208.
- (1992), "Testing for Seasonality," *Economics Letters*, 38, 259–262.
- (1996), *Periodicity and Stochastic Trends in Economic Time Series*, Oxford university press, Oxford.
- Franses, P.H. and T.J. Vogelsang (1995), "Testing for Seasonal Unit Roots in the Presence of Changing Seasonal Means," Report no. 9532A, Econometric Institute, Erasmus University, Rotterdam.
- Ghysels, E. (1990), "Unit Root Tests and the Statistics Pitfalls of Seasonal Adjustment: The Case of US Postwar Real Gross National Product," *Journal of Business and Economic Statistics*, 8, 145–152.
- (1993), "On the Economics and Econometrics of Seasonality," in C. Sims (ed.), *Advanced in Econometrics*, Cambridge University Press, Cambridge.
- (1994), "On the Periodic Structure of the Business Cycle", *Journal of Business and Economic Statistics*, 12, 289–298.
- Ghysels, E. and P. Perron (1993), "Effect of Seasonal Adjustment Filters on Tests for a Unit Root," *Journal of Econometrics*, 55, 57–98.

- Ghysels, E., H.S. Lee, and P.L. Siklos (1993), "On the Misspecification of Seasonality: An Empirical Investigation with US Data," *Empirical Economics*, 18, 747–760.
- Ghysels, E., H.S. Lee, and J. Noh (1994), "Testing for Unit Roots in Seasonal Time Series: Some Theoretical Extensions and a Monte Carlo Investigation," *Journal of Econometrics*, 62, 415–442.
- Granger, C.W.J. (1986), "Developments in the Study of Cointegrated Economic Variables," *Oxford Bulletin of Economics and Statistics*, 48, 213–228.
- Granger, C.W.J. and P.L. Siklos (1995), "Systematic Sampling, Temporal Aggregation, Seasonal Adjustment and Cointegration: Theory and Evidence," *Journal of Econometrics*, 66, 357–369.
- Harvey, A.C. (1989), *Forecasting Structural Time Series Models and the Kalman Filter*, Cambridge University Press, Cambridge.
- Harvey, A.C. and A. Scott (1994), "Seasonality in Dynamic Regression Models," *Economic Journal*, 104, 1324–1345.
- Hylleberg, S. (1986), *Seasonality in Regression*, Academic Press, New York.
- (1994), "Modeling Seasonal Variation," in C.P. Hargreaves (ed.), *Nonstationary Time Series Analysis and Cointegration*, Oxford University Press, Oxford.
- (1995), "Tests for Seasonal Unit Roots: General to Specific or Specific to General?" *Journal of Econometrics*, 69, 5–25.
- Hylleberg, S. (ed.), (1992), *Modeling Seasonality*, Oxford University Press, Oxford.
- Hylleberg, S. and A.R. Pagan (1996), "Seasonal Integration and the Evolving Seasonal Model," manuscript, Australian National University.
- Hylleberg, S., R.F. Engle, C.W.J. Granger, and B.S. Yoo (1990), "Seasonal Integration and Cointegration," *Journal of Econometrics*, 44, 215–238.
- Joyeux, R. (1992), "Testing for Seasonal Cointegration Using Principal Components," *Journal of Time Series Analysis*, 13, 109–118.
- Kunst, R.M. (1992), "Seasonality in the Austrian Economy : Common Seasonality and Forecasting," Research memorandum, No. 305, Institute for Advance Studies, Vienna, Austria.
- (1993), "Seasonal Cointegration in Macroeconomic Systems: Case Studies for Small and Large European Countries," *Review of Economics and Statistics*, 75, 325–330.

- Lee, H.S. (1992), "Maximum Likelihood Inference and Cointegration," *Journal of Econometrics*, 54, 1–47.
- Lee, H.S. and P.L. Siklos (1991), "Unit Roots and Seasonal Unit Roots in Macroeconomic Time Series: Canadian Evidence," *Economics Letters*, 35, 273–277.
- (1995), "A Note on the Critical Values for the Maximum Likelihood (Seasonal) Cointegration Tests," *Economics Letters*, 49, 137–145.
- Lütkepohl, H. (1991), *Introduction to Multiple Time Series Analysis*, Springer Verlag, Berlin.
- Nerlove, M., D.D. Grether, and J.L. Carvalho (1979), *Analysis of Economic Time Series – A Synthesis*, Academic Press, New York.
- Osborn, D.R. (1990), "A Survey of Seasonality in UK Macroeconomic Variables," *International Journal of Forecasting*, 6, 327–336.
- (1991), "The Implication of Periodically Varying Coefficients for Seasonal Time Series Processes," *Journal of Econometrics*, 48, 373–84.
- (1993), "Discussion: Seasonal Cointegration: The Japanese Consumption Function," *Journal of Econometrics*, 55, 275–298.
- Osborn, D.R. and J.P. Smith (1989), "The Performance of Periodic Autoregressive Models in Forecasting Seasonal UK Consumption," *Journal of Business and Economic Statistics*, 7, 117–27.
- Osborn, D.R., A.P.L. Chiu, J.P. Smith, and C.R. Birchenhall (1988), "Seasonality and the Order of Integration for Consumption," *Oxford Bulletin of Economics and Statistics*, 50, 361–377.
- Perron, P. (1989), "The Great Crash, the Oil Price Shock and the Unit Root Hypothesis," *Econometrica*, 57, 1361–1401.
- (1991), "Test Consistency with Varying Sampling Frequency," *Econometric Theory*, 7, 341–368.
- Sarantis, N. and C. Stewart (1993), "Seasonality, Cointegration, and the Long-Run Purchasing Power Parity: Evidence for Sterling Exchange Rates," *Applied Economics*, 25, 243–250.
- Sims, C.A. (1974), "Seasonality in Regression," *Journal of American Statistical Association*, 69, 618–626.
- Shiller, R. and P. Perron (1985), "Testing the Random Walk Hypothesis: Power versus Frequency of Observation," *Economics Letters*, 18, 381–386.
- Wallis, K.F. (1974), "Seasonal Adjustment and Relations Between Variables," *Journal of American Statistical Association*, 69, 18–31.

Part IV

Structural change

Structural change is an important problem in time series and affects all the inferential procedures discussed in the previous chapters. The area is so vast that it warrants a separate book. However, omitting it completely would leave an important gap in this book. We have therefore decided to cover it briefly.

This part has 3 chapters.

Chapter 13 on structural change, unit roots, and cointegration starts with a discussion of tests for structural change with known and unknown break points. We next discuss tests for unit roots under structural change. Both the classical and Bayesian approaches are covered. Next we discuss tests for structural change in cointegrated relationships.

Chapter 14 is on outliers and unit roots. We first discuss the different kinds of outliers and their effects on unit root tests. Given that outliers have serious effects on unit root tests, some outlier-robust unit root testing procedures as well as outlier-robust estimation procedures for CI systems, are discussed.

Chapter 15 is on regime switching models. These have been popularized by Hamilton who suggested using the Markov switching regression (MSR) models. This chapter discusses Hamilton's basic MSR model and several extensions of this basic model. The models in this chapter complement the outlier models discussed in the previous chapter.

One problem with the MSR model is that it models only sudden changes. In practice structural change is gradual. We, therefore, discuss model with gradual change. An important class of models with gradual change is the structural time series model introduced by Harvey. This model is estimated by using the Kalman filter. We discuss this model briefly and give the relevant references because this is a vast area to cover here at length.

13

Structural change, unit roots, and cointegration

One major drawback of unit root tests is that, in all of them, the implicit assumption is that the deterministic trend is correctly specified. Perron (1989) argued that if there is a break in the deterministic trend, then unit root tests will lead to a misleading conclusion that there is a unit root, when in fact there is not. Perron's paper started a controversy about the effect of trend breaks on unit root tests, and a criticism that Perron assumed the break point to be known. It was argued that if the break point is treated as endogenous, then Perron's conclusions are reversed. Subsequent literature partially reversed this conclusion and also extended the problem of trend breaks to other areas such as cointegration and seasonality.

The purpose of this chapter is to present a selective survey of the literature on trend breaks in unit roots, cointegration, and seasonal integration and comment on their implications for further research. Because of space limitations technical details will be omitted.¹ Since the literature is vast and it is easy to get lost in the details, the present chapter should help readers to get a perspective on this area. In this chapter we shall first review the literature on tests for structural change, and tests for unit roots under structural changes with a known break point and an unknown break point. We then move on to the Bayesian approach and multiple structural changes. Next we discuss the problems of the effect of structural change on cointegration tests.

¹ A technical account of this area from the classical point of view is given in a lengthy paper by Stock(1994). This chapter covers the Bayesian aspects as well, and also discusses cointegration tests and problems of seasonality.

13.1 Tests for structural change

Because of events like the great depression, oil price shocks, abrupt policy changes, and so on, models with constant coefficients have been found to perform poorly, either for forecasting purposes or for the purpose of analyzing the effect of policy changes. The solutions to this problem have been

- (i) Models with continuous parameter changes: these are estimated using some recursive algorithms like the Kalman filter. The problem with these models is that they do not capture sudden shifts.
- (ii) Outlier models: these models argue that sudden shocks produce outliers (with temporary or permanent level shifts). These are discussed in the next chapter.
- (iii) Switching regression models, with abrupt switches and gradual switches: one popular model during recent years in this category has been the Markov switching regression (MSR) model. These are discussed in chapter 15.

Before adopting any of these approaches, it is customary to test for structural changes. There is now an enormous number of statistical tests for testing structural change. The tests can conveniently be classified under the categories:

- (i) known break points versus unknown break points.
- (ii) single break versus multiple breaks.
- (iii) univariate versus multivariate relationships.
- (iv) stationary versus nonstationary variables.

We shall discuss the different tests within this classification scheme.

13.2 Tests with known break points

The earliest tests for structural breaks in the economic literature are the tests in Chow (1960) which are for stationary variables and a single break. The model is a linear regression model with k variables and two regimes with observations n_1 and n_2 respectively. Chow derives two tests: (a) for $n_1 > k$, $n_2 > k$ and (b) for $n_1 > k$, $n_2 < k$. The first is known as an analysis of variance test, and has a long history. It was discussed in Rao (1952) and Kullback and Rosenblatt (1957) among others. The second test is known as a predictive test because it does not test for stability of the coefficients but in fact tests for unbiasedness

of predictions for the n_2 observations, from a regression estimated with n_1 observations. This test is properly called the Chow test. Rao (1965) also considers this deficient rank case under the heading: "The Third Fundamental Theorem on Least Squares."

Dufour (1982) extends these tests to the case of multiple regimes and covers the cases where subsamples size are less than k . He also provides a single derivation for both the cases of full rank and deficient rank (tests (a) and (b)). Consider the case of J groups with n_j observations in the j th group ($j = 1, 2, \dots, J$) (some but not all of n_j can be less than k). Let RSS_0 be the residual sum of squares from a regression with the pooled data, and RSS_T be the total residual sum of squares from the sets for which $n_j > k$. Let there be r such sets ($r \leq J$). Let $n = \sum n_j$ and $n_1 = \sum_r \sum n_j$ over r sets for which $n_j > k$. Then $RSS_0/\sigma^2 \sim \chi^2$ with d.f. $(n - k)$ and $RSS_T/\sigma^2 \sim \chi^2$ with d.f. $(n_1 - rk)$ and

$$\frac{(RSS_0 - RSS_T)/(n - k - n_1 + rk)}{RSS_T/(n_1 - rk)}$$

has an F -distribution with degree of freedom $(n - k - n_1 + rk)$, $(n_1 - rk)$. Further discussion of Dufour's test can be found in Cantrell *et al.* (1991).

There are several extensions of the Chow tests (though one has to bear in mind which of these are extensions of the analysis of the variance test, and which are extensions of the predictive test). Lo and Newey (1985) and Park (1991) extend these tests to simultaneous equations. Andrews and Fair (1988) provide extensions of the analysis of the variance test to general nonlinear models. Dufour, Glyssels, and Hall (1994) provide extensions of the predictive test to general nonlinear models.

13.3 Tests with unknown break points

The traditional Chow (1960) test is developed to test the null hypothesis of parameter constancy against the alternative of a known break point *a priori* under the assumption of constant variances. The paper by Quandt (1960) discusses testing the null hypothesis of constant coefficients against a more general alternative, where a structural change has occurred at some unknown time and the error variance is also allowed to change. Quandt considered a switching regression such that the observations are thought, for theoretical reasons, to have been generated by two distinct regression regimes. Thus, for some subset of the

observations

$$y_i = X_i\beta_i + u_{1i}, \quad (i \in I)$$

and for the complementary subset

$$y_j = X_j\beta_j + u_{2j}, \quad (j \in J)$$

The essence of this simple formulation is that all the observations up to the unknown time m come from one regime and all the observations after that point come from the other. If we wish to test the hypothesis that there is no switch in regimes against the alternative of one switch, the appropriate likelihood ratio is

$$\lambda = \frac{1}{2}m \log \hat{\sigma}_1^2 + \frac{1}{2}(T - m) \log \hat{\sigma}_2^2 - \frac{1}{2}T \log \hat{\sigma}^2$$

The estimate of the point at which the switch from one relationship to another has occurred is then the value of m at which λ attains its minimum. However, implementation of this procedure has been hindered by the lack of a distribution theory. Quandt noted on the basis of a Monte Carlo experiment that a proposed chi-squared approximation to the significance level of the likelihood ratio test is very poor. It was shown empirically in Quandt that the χ^2 distribution is a poor approximation to that of $-2 \log \lambda$.

Kim and Siegmund (1989) examined likelihood ratio tests to detect a structural change in a simple linear regression when the alternative, H_1 , specifies that only the intercept changes and when the alternative, H_2 , permits the intercept and the slope to change. They show the asymptotic distributions of the maximum value of likelihood ratio, i.e., Quandt's LR statistic for the null of no structural change against two different alternatives, which are functions of Wiener processes. They tabulated the critical values by Monte Carlo methods.

Brown, Durbin, and Evans (1975) suggest the CUSUM test based on recursive residuals. This test is for the general alternatives, including the case of a single break. The CUSUM test involves considering the plot of the quantity

$$W_m = \frac{1}{\hat{\sigma}} \sum_{t=k+1}^m w_t, \quad m = k+1, \dots, T$$

where w_t is the recursive residual. Under H_0 , probabilistic bounds for the path of W_m can be determined and H_0 is rejected if W_m crosses the boundary (associated with the level of the test) for some m . This

test is aimed mainly at detecting systematic movements of coefficients. Against haphazard rather than systematic types of movements, Brown *et al.* proposed the CUSUM of squares test, which uses the squared recursive residuals and is based on a plot of the quantities

$$S_m = \frac{\sum_{t=i}^m w_t^2}{S^2}, \quad S^2 = \sum_{t=k+1}^T w_t^2, \quad m = k+1, \dots, T$$

The H_0 is rejected if the path of S_m crossed a boundary determined by the level of the test. These tests are of the goodness-of-fit type in the sense that they seem applicable against a wide variety of alternatives.

Extensions of this test have been made by Ploberger, Kramer, and Alt (1989) and Kramer, Ploberger, and Alt (1988) for models with lagged dependent variables, and Kao and Ross (1992) for models with serially correlated disturbances. Ploberger and Kramer (1990) extend the CUSUM test to OLS residuals and argue that it could as well be applied with OLS residuals and not just recursive residuals. A drawback of the CUSUM tests is that they have asymptotically low power against instability in the intercept but not the entire coefficient vector. Kramer, Ploberger, and Alt (1988) show this using asymptotic local power and Garbade (1977) shows this using simulations.

For the power problem of the CUSUM test, Ploberger, Kramer, and Kontrus (1989) proposed the fluctuation test based on successive parameter estimates rather than on recursive residuals. A similar procedure for single regression model has been suggested by Sen (1980) and the fluctuation test was first suggested by Ploberger (1983).

Ploberger, Kramer, and Kontrus (1989) considered the varying parameter models and proposed the fluctuation test which is based on rejecting the null hypothesis of parameter constancy whenever these estimates fluctuate too much. Their test statistic is

$$S^{(T)} = \max_{t=k, \dots, T} \frac{t}{\hat{\sigma}T} \|(X^{(T)'} X(T))^{1/2} (\hat{\beta}_i^{(t)} - \hat{\beta}_i^{(T)})\|_{\infty}$$

where

$$\hat{\sigma} = [\sum (y_t - x_t' \hat{\beta}^{(T)})^2 / (T - K)]^{1/2}$$

and $\|\cdot\|_{\infty}$ denotes the maximum norm. They derived the limiting distribution of the test statistic and tabulated the critical values by Monte Carlo methods. They also show that the fluctuation test has nontrivial local power irrespective of the particular type of structural change. Kramer, Ploberger, and Kontrus (1989) compare the fluctuation test

with the CUSUM test and find that the former does better for many alternatives. However, Kontrus and Ploberger's (1984) Monte Carlo results show that neither the CUSUM test nor the fluctuation test dominates the other in small samples.

Another class of tests is for continuous parameter variation as the alternative. The tests in this category are those by Leybourne and McCabe (1989), Nabeya and Tanaka (1988), and Nyblom (1989) who consider tests for parameter constancy against nonstationary alternatives (e.g., that the parameters follow a random walk or a martingale). Nyblom (1989) developed the locally most powerful test against a parameter variation in the form of a martingale. The martingale specification has an advantage of covering several types of departure from constancy: for example, a single jump at an unknown time point (change point model) or slow random variation (typically random walk specification of parameters).

The problem of deriving the appropriate asymptotic distribution of the test statistic in these problems involves hypothesis testing when the nuisance parameter is present only under the alternative. This problem was dealt with in Davies (1977, 1987) and more recently in Hansen (1996).

Quandt's LR test and Kim and Siegmund's (1989) procedures can be considered in the line of Max Chow test, since their procedure is to find out the significance of the maximum value of the likelihood ratio statistic from recursive switching models. Andrews (1993) derives the asymptotic distribution of the LR-like test for one-time structural change with an unknown change point (the Quandt test) as well as analogous Wald (W) and Lagrange Multiplier (LM) tests. They are for models with no deterministic or stochastic trends. Whereas, the CUSUM test and the fluctuation test apply only to linear models, the Andrews' tests apply to nonlinear models estimated by ML or GMM methods. Andrews shows that his Sup F test (or Max Chow) has better power properties than the CUSUM test and the fluctuation test (in the context of a linear model). He provides asymptotic critical values for 1 percent, 2.5 percent, 5 percent, and 10 percent significance levels, for his tests. (These are the critical values for the Sup F or Max Chow tests.)

In the case of a simple AR(1) model Andrews' three tests are defined as follows: consider

$$y_t = \rho y_{t-1} + \varepsilon_t, \quad \varepsilon_t \sim IN(0, 1), \quad t = 1, 2, \dots, T$$

Under the null of no structural change we estimate this equation by OLS and let $\hat{\varepsilon}_t$ be the residuals.

The model with a one time break at m is

$$\text{subsample 1: } y_t = \rho_1 y_{t-1} + \varepsilon_{1t}, \quad t = 1, 2, \dots, m$$

$$\text{subsample 2: } y_t = \rho_2 y_{t-1} + \varepsilon_{2t}, \quad t = m + 1, \dots, T$$

Let $\hat{\varepsilon}_{1t}$ and $\hat{\varepsilon}_{2t}$ be the residuals from the OLS estimation of these equations. Define $S = \hat{\varepsilon}'\hat{\varepsilon}$, $S_1 = \hat{\varepsilon}'_1\hat{\varepsilon}_1$, and $S_2 = \hat{\varepsilon}'_2\hat{\varepsilon}_2$. Then “Sup” statistics suggested by Andrews are:

$$\text{Sup } W = \max_{\pi} T \left[\frac{S - S_1 - S_2}{S_1 + S_2} \right]$$

$$\text{Sup } LM = \max_{\pi} T \left[\frac{S - S_1 - S_2}{S} \right]$$

$$\text{Sup } LR = \max_{\pi} T \left[\frac{S}{S_1 + S_2} \right]$$

where $\pi = m/T$. It is customary to take $\pi \in [0.15, 0.85]$, so that breaks toward the ends are ruled out.

Diebold and Chen (1996) argue that the use of the asymptotic critical values suggested by Andrews leads to size distortions and suggest a better procedure generating small sample critical values using bootstrap methods. This is similar to the conclusions in several other models (see chapter 10) where critical values generated by bootstrap methods were found to be better than those derived from asymptotic approximations.

Andrews and Ploberger (1994) develop tests with stronger optimality properties than those in Andrews (1993). We shall define these test statistics for the Wald test. The definitions for the LM and LR statistics are similar and the asymptotic distributions are the same for the W, LM, and LR statistics.

Let $T_1 < m < T_2$ be the range within which the structural break occurs. Let W^* be the Wald statistic for a break at $t = m$. Then

$$\text{Sup } W = \max_{T_1 < m < T_2} W^*$$

The test statistics suggested by Andrews and Ploberger are

$$\text{Exp } W = \ln \left[\frac{1}{T_2 - T_1 + 1} \sum_{m=T_1}^{T_2} \exp(W^*/2) \right]$$

$$\text{Ave } W = \frac{1}{T_2 - T_1 + 1} \sum_{m=T_1}^{T_2} W^*$$

Under a wide set of regularity conditions, Andrews (1993) and Andrews and Ploberger (1994) show that the asymptotic distributions of the test statistics are given by functions of the Wiener processes (these are the same for the LM and LR test statistics)

$$\begin{aligned}\text{Sup } W &\Rightarrow \max_{\pi_1 \leq \tau \leq \pi_2} Q(\tau) \\ \text{Exp } W &\Rightarrow \ln \left[\frac{1}{\pi_2 - \pi_1} \int_{\pi_1}^{\pi_2} \exp(Q(\tau)/2) d\tau \right] \\ \text{Ave } W &= \frac{1}{\pi_2 - \pi_1} \int_{\pi_1}^{\pi_2} Q(\tau) d\tau\end{aligned}$$

where $\pi_1 = T_1/T$, $\pi_2 = T_2/T$, and

$$Q(\tau) = \frac{(W(\tau) - \tau W(1))'(W(\tau) - \tau W(1))}{\tau(1 - \tau)}$$

where W is a vector of k independent Wiener process (see section 3.2). When τ is known, this is simply $\chi^2(k)$.

Hansen (1995) provides numerical approximations to these asymptotic distributions and provides detailed tables whereby one can compute p -values, rather than use significance tests. (Note that many regression programs now present p -values.)

Andrews, Lee, and Ploberger (1996) present a detailed Monte Carlo study comparing all these tests: the three Sup tests, three Exp tests, and the three Ave tests.

Instead of the Max Chow test of Andrews, Hansen (1990) suggests the Mean Chow test. Note that this is the Ave test discussed by Andrews and Ploberger. The Mean Chow statistic is the average of the likelihood ratios from the separate regressions over the subsamples $1, \dots, m$ and $m + 1, \dots, T$. As mentioned above, test statistics of Quandt (1960) and Kim and Siegmund (1989) are the Max Chow type test – the maximum of the Chow sequence that was designed to detect a single abrupt structural change over sample periods. Hansen's Mean Chow statistic is testing for the null hypothesis of parameter constancy against more general alternatives such as those in the fluctuation test and the locally most powerful test. He suggested two statistics, one for stationary regressors and the other for nonstationary regressors.

On the basis of a limited Monte Carlo experiment, Hansen claims that the Mean Chow test performs better in regression models with both trending and nontrending regressors in terms of size accuracy and power.

Note that the Monte Carlo experiment reported in Andrews *et al.* (1996) is more extensive. Since this diagnostic test is based on the average value rather than maximum value, it does not provide any guidance for the identification of breakpoints. Hansen (1992) suggests the Max F , Mean F , and L_c statistics. The last one is a test statistic for testing constancy of the parameter against the random walk alternative. The Mean F and Max F statistics are for the alternative of sudden breaks. The L_c statistic is for the alternative of a gradual change. Hansen (1992) gives tables for critical values of the Max F , Mean F , and L_c statistics.

Other papers that consider unknown breaks with nonstationary regressors are: Banerjee, Lumsdaine, and Stock (1992), Zivot and Andrews (1992), and Chu and White (1992). Banerjee *et al.* consider recursive, rolling, and sequential tests for unit roots and/or changing coefficients. The recursive and rolling tests are based on changing subsamples of the data. The sequential statistics are computed using the full sample partitioned into two groups sequentially. Chu and White consider a test for a broken trend at an unknown point. One interesting result in their paper is that the critical values for the asymptotic distribution for testing structural change in nonstationary time series differ from those in a stationary time series by a factor of $\sqrt{3}$. (Note that the test statistics are different.) Of all these tests, the tests by Hansen (1992) are the easiest to apply.

In the Bayesian approach, the case of a single unknown break has been considered by Holbert (1982) and Salazar (1982). The case of multiple unknown breaks has been discussed by Kim and Maddala (1991). All these procedures are applicable for stationary regressors and also nonstationary regressors.

A commonly used approach for multiple breaks is the Markov switching regression (MSR) model – also sometimes called the Hamilton model (though the Hamilton model is a particular MSR model. This is discussed in chapter 15.) There is the question of how many regimes to consider, though most empirical work is on the two-state model. For example the date of the second switch in the paper by Garcia and Perron (1995) depended on whether the MSR model was considered as a two-state model or a three-state model. The tests for three-state versus two-state, two-state versus one-state (no structural break) are nonstandard tests. Their distribution is discussed in Garcia and Perron, as well as Garcia (1997).

13.4 A summary assessment

As the previous survey indicates there is a large number of tests: tests of no break versus single break, tests of no break versus multiple breaks. One can raise the question of testing for one break versus two breaks and so on. But the problem is not one of testing for one versus two breaks, but one of estimation of the number of breaks. This is a model selection problem as noted in Kim and Maddala (1991) for which they use the BIC criterion. This problem has been analyzed further by Peter Phillips using the PIC criterion (see chapter 8). The analysis of this problem of multiple breaks from the classical point of view and the asymptotic distributions are discussed in Bai and Perron (1995). In tests for no break versus single break, note that rejecting the null does not necessarily imply one-time structural change.

We not only need to know that breaks exist, but also the location of the breaks. Several tests for breakpoints merely test whether a break exists or not, but do not identify the location of the breaks. The CUSUM tests, the Bayesian procedure used by Kim and Maddala (1991), the MSR models, all identify the location of the break points. In the MSR models, as found by Garcia and Perron (1996), the location of the switch points depends on the number of states assumed.

There is a lot of work on testing with unknown switch points. In practice, there is a lot of prior information and there is no reason why we should not use it. For instance, suppose there is a drastic policy change or some major event (for example, oil price shock) that occurred at time t_0 . It does make sense to ask the question of whether there was a structural change around that period. It is not very meaningful to search for a break over the entire period ignoring this prior information. Thus criticism of Perron (1989) and the subsequent proliferation of the literature on *endogenizing* the breakpoint are not totally justified. If a search is conducted, it should be around the events.

Finally, there is the question of the relative merits of tests based on residuals versus tests based on scores. Tests based on residuals like simple CUSUM type tests have been found to be not appropriate for tests concerning the stability of the entire coefficient vector. Hansen (1992) shows that tests using the full score $x_i\hat{u}_i$ rather than just the residual $\hat{u}_i$ generate test statistics with better power.

13.5 Tests for unit roots under structural change

The exact definition of structural changes has not been given in the literature. Usually it is interpreted as changes of regression parameters. We shall use this conventional definition. Thus our focus is on finding the effect of changes in regression parameters on the tests for unit roots. Recent developments in unit root tests are concerned with the effect of changes in the coefficients of the deterministic trends.

13.5.1 Single known break

Given a known structural break which is assumed to be given exogenously, Perron (1989) has proposed a modified DF test for a unit root in the noise function with three different types of deterministic trend function. The time of a structural change is referred to as T_B . First, the *crash* model, model (A), allows for a one-time change in the intercept of the trend function

$$DT_t = \mu_0 + \mu_1 DU_t + \delta t$$

where

$$DU_t = \begin{cases} 1 & \text{if } t > T_B \\ 0 & \text{otherwise} \end{cases}$$

The *changing growth* model, model (B), allows for a change in the slope of the trend function without any sudden change in the level at the time of the break

$$DT_t = \mu + \delta_0 t + \delta_1 DT_t$$

$$DT_t = \begin{cases} t - T_B & \text{if } t > T_B \\ 0 & \text{otherwise} \end{cases}$$

And both effects are allowed in the model (C)

$$DT_t = \mu_0 + \mu_1 DU_t + \delta_0 t + \delta_1 DT_t^*$$

$$DT_t^* = \begin{cases} t & \text{if } t > T_B \\ 0 & \text{otherwise} \end{cases}$$

The null hypothesis of a unit root is different, since the deterministic trend function includes dummy variables (DU_t, DT_t, DT_t^*). The alternative hypothesis is a *broken-trend* stationary system which also incorporates the same dummy variables.

To assess the effects of the presence of a shift in the level or a shift in the slope (at a single point of time) on tests for the presence of a unit root, Perron (1989) first performed a Monte Carlo experiment. The Monte Carlo results show that if the magnitude of the shift is significant, one could hardly reject the unit root hypothesis, even if the series is stationary with a broken trend and *iid* disturbances (thus, there is no unit root in the noise term). Perron extended the Dickey–Fuller testing strategy to ensure a consistent testing procedure against shifting trend functions, which is first to detrend the series and then analyze the behavior of the estimated residuals. Let $\{\tilde{y}_t^i\}$ be the detrended series according to either model (A), (B), or (C), where $i = \text{model (A), (B), (C)}$. Then $\{\tilde{y}_t^i\}$ are the residuals from a regression of y_t corresponding to the models (A), (B), and (C).

Perron derived the limiting distributions of the normalized least squares estimators $\hat{\rho}$ and their t -statistics from the following regression

$$\tilde{y}_t^i = \rho^i \tilde{y}_{t-1}^i + e_t$$

Those distributions are functions of functionals of Wiener processes and the parameter $\lambda = T_B/T$, the ratio of the pre-break sample size to total sample size. For the exact forms of the distributions, see theorem 2 of Perron (1989). He tabulated the percentage points of the limiting distributions for given values of λ . When λ is either 0 or 1, the limiting distributions are identical over all models and the critical values are identical to those of Dickey and Fuller. When λ is not equal to either 0 or 1 (thus, there is a structural break over the sample period), the critical values under the various models are noticeably smaller (greater in absolute value) than the standard Dickey–Fuller critical values. Perron applied the modified Dickey–Fuller test for the same US macroeconomic series used by Nelson and Plosser (1982) and found the quite strikingly different result that the unit root hypothesis can be rejected for all but three series: consumer price, velocity, and interest rate.

Perron's procedure is a conditional test given a known break point. This assumption of a known break date (treated as an exogenous event) raised the problem of pre-testing and data-mining regarding the choice of the break date. After Perron (1989), several methods have been developed for endogenizing the choice of a break point into testing procedures. These procedures incorporate the estimation of a break point and use recursive methods (using subsamples) or sequential methods (using full sample with dummies).

13.5.2 Single unknown break

Perron's study was criticized on the grounds that he treated the date of the break as known. Subsequent work used a variety of recursive and sequential tests endogenizing the break point. The recursive test statistics are computed over subsamples $t = 1, \dots, m$ for $m = m_0, \dots, T$ where m_0 is the start-up value and T is the size of the full sample. The sequential test statistics are computed using the full sample, sequentially incrementing the date of the hypothetical break (thus using different dummy variables).

Zivot and Andrews (1992) use a sequential test, derive the asymptotic distribution of the test statistic, $\min_{\lambda} t(\lambda)$, and tabulate the critical values. They fail to reject the unit root hypothesis for four of ten series for which Perron rejected the unit root null (real per capita GNP, GNP deflator, money stock, and real wages). With finite sample critical values obtained by bootstrap methods (see chapter 10 for bootstrap methods) they fail to reject the unit root null for three more series (employment, nominal wages, and common stock prices).

Banerjee, Lumsdaine, and Stock (1992), to be referred to as BLS, apply a variety of recursive and sequential tests endogenizing the break point to international data. For instance, in the recursive tests they consider the maximum DF statistic, the minimum DF statistic, and the difference between the two. They derive the asymptotic distributions of the recursive and sequential test statistics and tabulate the critical values. Further results endogenizing the break points are also presented in Perron (1994).

The summary picture one gets from these studies is that endogenizing the break point reverses the conclusions arrived at by Perron (1989). However, these early results have been partially reversed by recent empirical work. One is the study by Ben-David and Papell (1995) covering a long time series which we shall discuss later (see section 13.8). The other is the argument by Lumsdaine and Papell (1997) (to be referred to as LP) and Ben-David, Lumsdaine, and Papell (1997) (to be referred to as BLP) that the previous literature considered endogenizing only one break point but do not consider the possibility of two break points.

LP follows the earlier study by BLS (1992) and consider the sequential tests for two breaks with unknown break points. Following the methods in BLS, they derive the asymptotic distributions of the test statistics. Like Zivot and Andrews, they also use bootstrap methods to get finite sample critical values. (The tables of critical values are available from

the authors on request.) Applying these tests to the Nelson–Plosser data LP find more evidence against the unit root hypothesis than Zivot and Andrews but less than Perron. There is of course the question: what about three breaks instead of two? Ultimately, determination of the number of breaks is best viewed as a model selection problem.

BLP apply the procedure developed in LP to international data for 16 countries. They find evidence of two breaks in three quarters of the countries and reject the unit root hypothesis in 50 percent more cases than in models that consider only one break.

13.6 The Bayesian approach

Turning to the Bayesian work on structural breaks, there are two broad avenues to view. One starts with Perron’s work along classical lines, the other can be viewed along the lines of Bayesian analysis of switching regressions with unknown break points. We shall discuss these in turn.

13.6.1 Bayesian analysis of classical models

Zivot and Phillips (1994) start with Perron’s model. They allow for a one-time change in the level (constant) and a one-time change in the slope of time. These occur at an unknown point t_0 and they assume a uniform prior for t_0 . However, for the autoregressive root ρ , they assume Jeffreys’ prior. To counteract the objection to this prior, that it places increasing weight on extreme and unlikely values of ρ as far as economic time series are concerned, they limit the effective range of ρ and hence the degree of nonstationarity allowed *a priori*. Finally, to evaluate the evidence for the unit root model, when allowance is made for structural change, they let $q \in [0, 1]$ denote the prior probability for the model with no structural change (NSC). Zivot and Phillips find that the trend structural change model with $t_0 = 1929$ is quite likely for nominal GNP and nominal wages if the prior probability of NSC is not too large. Thus, the Bayesian analysis allowing for an endogenous break supports some of Perron’s conclusions. As shown by Banerjee *et al.* (1992) and Zivot and Andrews (1992) the asymptotic theory for testing for unit roots with endogenous breaks is quite complicated and also as shown by Zivot and Andrews (1992) and Perron (1990), the finite sample distributions of the unit root test statistics can be very different from their asymptotic counter parts. This factor might also account for the reversal under

Bayesian methods of the conclusions arrived at earlier on the effect of endogenous breaks on unit root tests.

DeJong (1996) also starts with Perron's model. However, he considers a third-order autoregression

$$y_t = \alpha_0 + \alpha_1 t + \beta_1 D_{1t} + \beta_2 D_{2t} + \sum_{i=1}^3 \rho_i y_{t-i} + e_t$$

where D_{1t} and D_{2t} are dummies for the breaks in the intercept and slope coefficients. He assumes flat priors for the location of these breaks as well as for ρ_1, ρ_2, ρ_3 . The parameter of interest is the dominant root (note that a flat prior on ρ_i does not imply a flat prior on the dominant root). The marginal distribution of this dominant root λ_{max} is obtained through integration by Monte Carlo. DeJong calculates the odds in favor a unit root: $\lambda_{max} \geq 0.98$ versus $\lambda_{max} < 0.98$. This avoids the problem of testing a point null hypothesis which is tricky under Bayesian inference (see chapter 8). DeJong concludes that, allowing for uncertainty about the location of the break point, the data support trend stationarity.

13.6.2 Bayesian analysis of switching regression models

The switching regression approach for detecting a structural break is as follows: consider the two-phase regression model with an unknown break point m ,

$$\begin{aligned} y &= X_1 \theta_1 + \varepsilon_1, & t &= 1, \dots, m \\ y &= X_2 \theta_2 + \varepsilon_2, & t &= m+1, \dots, T \end{aligned}$$

We can rewrite the model in the standard form

$$y = W\theta + \varepsilon$$

where

$$W = \begin{bmatrix} X_1 & 0 \\ 0 & X_2 \end{bmatrix}, \quad \theta = \begin{bmatrix} \theta_1 \\ \theta_2 \end{bmatrix}, \quad \varepsilon = \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \end{bmatrix}$$

ε_1 and ε_2 are assumed to be independently and normally distributed with a zero mean vector and unknown variance σ^2 . Then the likelihood function for m, θ , and σ^2 is given by

$$L(\theta_1, \theta_2, \sigma^2, m | y, W) \propto (\sigma^2)^{-T/2} \exp \left[\frac{-1}{2\sigma^2} (y - W\theta)' (y - W\theta) \right]$$

We assume that the break point m is independent of θ and σ^2 , and m is equally likely at any observation point between τ_0 and $T - \tau_0$ where $\tau_0 > k$ is the end point. These assumptions result in a uniform prior over m . We further assume a flat prior on θ and σ^2 , and thus the joint prior is given by

$$\pi(m, \theta, \sigma^2) \propto 1/\sigma^2$$

Combining the prior with the likelihood function, we have the joint posterior density function

$$p(\theta, \sigma^2, m|y, W) \propto (\sigma^2)^{-(T/2)+1} \exp \left[-\frac{1}{2\sigma^2} \{S + (\theta - \hat{\theta})'W'W(\theta - \hat{\theta})\} \right]$$

Integrating out θ and σ^2 yields the marginal posterior distribution function of a break point

$$p(m|y, W) \propto |W'W|^{-1/2} |S|^{-(T-2k)/2}$$

where $|W'W| = |X_1'X_1||X_2'X_2|$ and $S = S_1 + S_2$ with

$$S_i = (y_i - X_i\hat{\theta}_i)'(y_i - X_i\hat{\theta}_i) \quad \text{and} \quad \hat{\theta}_i = (X_iX_i)^{-1}X_i'y_i$$

The marginal posterior density function of the regression parameters is obtained by

$$p(\theta|y, W) \propto \sum_{m=k}^{T-k} P_1(\theta|m, y, W)P_1(m|y, W)$$

where $p(\theta|m, y, W)$ is the conditional marginal posterior density function of θ given m and it is given by a matrix t density

$$p(\theta|m, y, W) \propto [I + (\theta - \hat{\theta})'H(\theta - \hat{\theta})]^{-(\nu+2k-1)/2}$$

where $\nu = T - 2k + 1$ and $H = W'W/S$.

The procedure outlined above has been used by Holbert (1982), Salazar (1982), and Kim (1991). The procedure has also been checked by performing Monte Carlo experiments. The Monte Carlo results reported in chapter 3 of Kim (1991) showed that a break point can be identified as the peak of the marginal posterior distribution of within the sample period, if $0.1T \leq m \leq 0.9T$, regardless of the stationarity of regressors. The problem arises only if the break is at the beginning or the end.

13.6.3 Multiple unknown breaks

We shall now discuss the case of multiple structural breaks. Suppose that there are q multiple breaks, then

$$\begin{aligned} y &= X_1\theta_1 + \varepsilon_1, & t = 1, \dots, m_1 \\ &= X_2\theta_2 + \varepsilon_2, & t = m_1 + 1, \dots, m_2 \\ &\vdots \\ &= X_{q+1}\theta_{q+1} + \varepsilon_{q+1}, & t = m_q + 1, \dots, T \end{aligned}$$

Define the matrix W as

$$W = \begin{bmatrix} X_1 & 0 & \dots & 0 \\ 0 & X_2 & \dots & 0 \\ & & \ddots & \\ 0 & \dots & 0 & X_{q+1} \end{bmatrix}$$

and assume that $\varepsilon_i, i = 1, \dots, q+1$ are independently and normally distributed with a zero mean vector and unknown variance σ^2 . Then the expressions of posterior distributions derived in the previous section remain the same. The only difference is that the posterior distribution of m should be the joint discrete distribution of $m_1, \dots, m_q$ which will be given by

$$p(m_1, \dots, m_q | y, W) \propto |W'W|^{-1/2} |S|^{-(T-2k)/2}$$

where $|W'W| = \prod_{j=1}^{q+1} |X'_j X_j|$ and $S = \sum_{j=1}^{q+1} S_j$ with

$$S_i = (y_i - X_i \hat{\theta}_i)' (y_i - X_i \hat{\theta}_i) \quad \text{and} \quad \hat{\theta}_i = (X_i X_i)^{-1} X_i' y_i$$

To obtain the marginal posterior distribution of m_j , we need to integrate out $m_1, \dots, m_{j-1}, m_{j+1}, \dots, m_q$. Analytical derivation is not possible, so we have to use numerical methods. One possible way of numerical calculation is to compute the probabilities in each grid of $m_1, \dots, m_q$ and sum it corresponding to m_j . This requires $(T - 2\tau_0)^q$ iterations. Thus when T and q are large, it is impossible to use this method for integration. Kim and Maddala (1991) use Gibbs sampling to get the marginal distributions out of the joint distribution of break points. Gibbs sampling is a Markovian updating scheme based on the cycle of conditional distributions. It has been found to be very useful in computing marginal densities in Bayesian inference. For an exposition of Gibbs sampling, see Gelfand and Smith (1990).

The problem of choosing between difference-stationary and trend-stationary models has been often approached as a hypothesis testing problem (that of testing for unit roots), but it should properly be approached as a model selection problem. This is done in McCulloch and Tsay (1994) and Phillips (1995) as discussed earlier in chapter 8.

The model selection approach can also be used to select the number of breaks. Kim and Maddala (1991) use the BIC criterion for selecting the best model among the models with different numbers of breaks. Kashiwagi (1991) analyzed the same problem of determining the number of breaks by the predictive likelihood approach. The method used by Kim and Maddala (1991) is described in Maddala and Kim (1996a, 1996b) and consists of the following steps:

- (i) Given the maximum number of breaks, calculate the marginal posterior distributions of $m_1, \dots, m_q$.
- (ii) Find the peaks of the marginal posterior distributions of $m_1, \dots, m_q$, say $\hat{m}_1, \dots, \hat{m}_q$.
- (iii) Calculate the BIC corresponding to $\hat{m}_1, \dots, \hat{m}_q$

$$BIC = -2 \log(ML) + (\log T)(\text{number of parameters})$$

where the number of parameters is that $(q+1) \times k$ (number of regressors) since the q breaks correspond to the $(q+1)$ subsamples. Compute these values in the case of no break ($q=0$), BIC^0 , one break ($q=1$), BIC^1 , ..., and so on, up to q , BIC^q .

- (iv) Find the $\min\{BIC^1, BIC^2, \dots, BIC^q\}$. The number of breaks can be estimated as

$$\hat{q} = \arg \min\{BIC^1, BIC^2, \dots, BIC^q\}$$

This procedure estimates the true number of breaks and location of break points $(m_1, \dots, m_q)$.

Kashiwagi (1991) proposes a method to evaluate the posterior distribution of the number of breaks and the posterior probability of each break point by using the predictive log likelihood. Kashiwagi's approach is not completely Bayesian. In his approach the probabilities of break points and the number of breaks are computed using Bayes theorem, but the likelihood function is evaluated using predictive likelihood, which can be described as follows. Consider the predictive density of a future observation y_f given data y_d and parameter θ . In practice θ is not known. There are two solutions to this problem. One is to specify a prior density for θ and integrate it out to get the posterior density of y_f . The other

is to replace θ by a sufficient statistic. This is the predictive likelihood approach.

For choosing the best model among alternatives under the different assumptions of the number of breaks, the procedures of Kashiwagi (1991) and Kim and Maddala (1991) are the same in that they choose the model which has the highest penalized likelihood. Also both procedures adopt numerical methods for computing the posterior distributions of the break points – Kashiwagi uses an approximation method and Kim and Maddala use Gibbs sampling. The main difference between the two approaches is the way they penalize the likelihood function, Kashiwagi penalized the likelihood function with the expected bias

$$\frac{T(t + m_1 + m_2)}{\sum_{i \in I_2} k_i - 2m_2 - 2} - T$$

where m_1 is the number of outliers and $m_2 = T - m_1$ and k_i is the number of regression parameters. On the other hand, Kim and Maddala (1991) penalized the likelihood function with

$$(\log T) \times (\text{number of parameters})$$

The two methods can be viewed in the context of comparing the (classical) predictive likelihood approach and the Bayesian approach for selecting the best model among alternatives with different number of breaks.

Bianchi (1995) extends Kashiwagi's approach to the case of autoregressive errors. See Bianchi (1995, p. 45). This extension is useful because the assumption of independent errors is not likely to be valid in practice. Bianchi shows that the results are sensitive to this assumption.

13.7 A summary assessment of the empirical work

The empirical results in the preceding sections suggest that in many areas allowing for structural change has changed conclusions about the inference on unit roots. Here we shall review some more empirical results. Some Monte Carlo evidence on this issue is provided in Hendry and Neale (1991) who argue that (i) regime shifts can mimic unit roots in stationary autoregressive time series and (ii) such shifts may be very hard to detect using conventional parameter constancy tests.

As mentioned earlier, Perron (1989) argued that, allowing for structural breaks, most macroeconomic variables are trend-stationary. Rapoport and Reichlin (1989) present similar evidence and Reichlin (1989)

analyzes the systematic bias in tests for $I(d)$ models with constant parameters against $I(d - 1)$ models with time-varying parameters, and shows that if the true process has deterministic structural change, we will accept the unit root hypothesis against a model corresponding to the true one but without the structural change; moreover, if the true process is a random walk, a goodness of fit criterion will lead one to accept the hypothesis of structural change against that of no change.

Perron's method of testing for unit roots allowing for trend breaks at known break points has also been used by Corbae and Ouliaris (1991) to test for long-run PPP for the Australian dollar. The conclusions are similar, that allowing for structural breaks reverses conclusions about unit roots. Perron (1990), with correction in Perron and Vogelsang (1992), considers testing for a unit root in a time series characterized by a structural change in the mean level (rather than in the trend). Again the analysis is for a known break point, but it is shown that allowing for a break reverses previous conclusions that the real interest rate for the US is characterized by a unit root. The data Perron considers are the US *ex-post* real interest rate over the period 1961:I–1986:III. Rose (1988) analyzed related series and concluded that the real interest rate was characterized by a unit root. Perron (1990) observes a marked discontinuity around 1980. Before 1980:III the average was close to zero and it was close to 6 percent during 1980:IV–1986:III.

Garcia and Perron (1995) analyze this same problem of structural breaks in real interest rates using the Markov switching model (discussed in chapter 15). The break points are thus estimated, and not known *a priori*. The sample used is monthly 1961:1–1986:12, though they put more emphasis on using the quarterly version of this data set. The results suggest that the *ex-post* real interest rate is essentially random around a mean that is different for the periods 1961–1973, 1973–1980, and 1980–1986. The variance is also different, being higher in both the 1973–1980 and 1980–1986 subperiods. Though the first break can be identified as due to the oil price shock, the dating of the second break is ambiguous. The two-regime model locates it at the end of 1979, suggesting that the change in monetary policy is its origin, and the three-regime model places it in the middle of 1981, a date more in line with the federal budget deficit explanation. In any case, the unit root hypothesis is rejected when allowance is made for structural breaks.

As mentioned earlier, the effect of endogenizing the break point in Perron's (1989) study has been shown by Banerjee, Lumsdaine, and Stock (1992), who use a variety of recursive and sequential tests, and

Zivot and Andrews (1992), who use sequential Dickey–Fuller tests to reverse Perron’s conclusions. However, Ben-David and Papell (1995) replicate the sequential Dickey–Fuller tests used by Zivot and Andrews (1992) and Banerjee *et al.* (1992) using a larger time span (130 and 120 years) and including more countries (sixteen) than is common in unit root studies. The model used is an endogenous trend break model. After determination of each country’s break date, they find that they can reject the unit root null in 20 of the 32 cases at a 10 percent significance level (16 of these at the 1 percent level) compared with only two rejections at the 5 percent level in the absence of a break (these are for the US aggregate and per capita GDP). They also find that on average post-break growth of aggregate GDP for the countries with significant trend breaks was twice the pre-break growth. They conclude that with longer time spans, incorporating trend breaks provides significant evidence for stationarity. Two issues, emphasized by Campbell and Perron (1991) guided the choice of their data: first, the power of unit root tests is better with long spans of data. Second, lengthening the span of the data increases the possibility of major structural changes. The papers by Lumsdaine and Papell (1997) and Ben-David, Lumsdaine, and Papell (1997), as discussed earlier, argue that allowing for the possibility of two endogenous break points, shows more evidence against the unit root hypothesis than Zivot and Andrews (although less than Perron).

In summary, the work on endogenizing the break point in Perron’s study (1989) seems to suggest that his conclusions are only partially reversed, if at all. In the area of exchange rates, we have referred to the study by Corbae and Ouliaris earlier. Perron and Vogelsang (1992) show how allowing for level shifts affects conclusions about unit roots in real exchange rates and thus conclusions about the purchasing power parity. Dropsy (1996) performs several structural stability tests to five foreign exchange rates relative to the dollar and five foreign exchange rates relative to the Deutsche Mark using quarterly data over 20 years (starting in the first quarter of 1974). He thus misses the important structural break of the switch to flexible exchange rates in March 1973. He identifies several series with structural breaks thus changing conclusions about unit roots and PPP. However, the dates for structural breaks in his study vary a lot and need further study, whether they can be identified with major events or policy changes. In summary, structural breaks in foreign exchange rates need further study. Several structural breaks also were identified earlier in the study

by Kim and Maddala (1991) but they have not been investigated further.

In conclusion, the evidence on unit roots in real exchange rates and real interest rates has been found to be weaker than thought earlier, once allowance is made for structural breaks.

13.8 Effect of structural change on cointegration tests

As noted earlier, the studies by Rappoport and Reichelin (1989), Hendry and Neale (1991), and Perron (1989) among others, showed that inference on unit roots is affected by structural change (the unit root tests tend to underreject the null of a unit root). The same is the case with tests for cointegration. However, when considering cointegrated relationships one has to distinguish between breaks in the relationships and breaks in the individual variables. In the latter case there is the problem that the dates of the breaks in the different variables may not coincide (see Hendry, 1996). We shall not, however, go into the latter issue because it is much more complicated than we wish to discuss here. We shall review two studies investigating the effect of structural breaks on tests for cointegration relationships in the single equation context. In the next section we shall consider the case of cointegrated systems.

The paper by Gregory, Nason, and Watt (1996) studies the sensitivity of the ADF test for cointegration in the presence of a single break. Their Monte Carlo results show that the rejection frequency of the ADF test decreases substantially. That is, in the presence of a break, you tend to underreject the null of no cointegration. (The underrejection is similar to the underrejection of the unit root null in the case of unit root tests.) However, in this case, the underrejection of the null indicate correctly that the constant parameter cointegration relation is not appropriate.

Campos, Ericsson, and Hendry (1996) investigate the properties of several cointegration tests when the marginal process of one of the variables is stationary with a structural break. They find that the break has little effect on the test size (compared with the underrejection noted by Gregory *et al.*). However, tests based on ECM (see chapter 6, section 6.3) are more powerful than the Engle–Granger two-step procedure employing the DF unit root test. These authors again emphasize the importance of common factor restrictions discussed in section 6.3 of chapter 6.

13.9 Tests for structural change in cointegrated relationships

There have been several studies deriving tests for structural change in cointegration relationships. These fall into two categories. Those in single equations and those in cointegrated systems (following the Johansen procedure). Before we discuss these we shall consider a simple diagnostic test for structural change.

13.9.1 A simple diagnostic test

Wright (1993) extends the CUSUM test (see section 13.3) to nonstationary trended variables and to integrated variables. Hoa and Inder (1996) extend the OLS-based CUSUM test discussed by Ploberger and Kramer (1992) for nonstationary regressors and since the test does not explicitly specify the nature of the alternative, they suggest its use as a diagnostic test for structural change. Instead of considering OLS residuals, they consider the FM-OLS residuals and the estimated error variance is replaced by the long-run variance estimate. Hoa and Inder derive the asymptotic distribution of the FM-OLS-based CUSUM test statistic, tabulate the critical values, and show that the test has nontrivial local power irrespective of the particular type of structural change. They tabulate the asymptotic critical values for two models

$$\begin{array}{ll} \text{Model 1 } (M_1) & y_t = \alpha_t + \beta_t x_t + u_t \\ & x_t = x_{t-1} + v_t, \quad t = 1, 2, \dots, T \\ \text{Model 2 } (M_2) & y_t = \alpha_t + \beta_t x_t + u_t \\ & x_t = \mu + x_{t-1} + v_t, \quad t = 1, 2, \dots, T \end{array}$$

In M_1 , x_t is I(1) without drift. In M_2 , x_t is I(1) with drift. The asymptotic critical values are shown in table 13.1. The bootstrap-based small-sample critical values can be computed and compared with these asymptotic critical values.

13.9.2 Single equation based tests

We shall discuss three papers related to structural change in cointegrated regressions. Quintos and Phillips (1993) develop tests for parameter constancy in cointegrated regressions. The alternative hypothesis is that the coefficients follow a random walk. These types of tests were considered earlier by Nabeya and Tanaka (1988), Leybourne and McCabe (1989), and extended to tests for cointegration by Hansen (1992).

Table 13.1. *Asymptotic critical values for the diagnostic test*

No. of Regressors	Significance Levels					
	10%		5%		1%	
	M_1	M_2	M_1	M_2	M_1	M_2
1	1.048	0.768	1.168	0.833	1.425	0.969
2	0.934	0.721	1.041	0.786	1.256	0.916
3	0.838	0.680	0.934	0.739	1.178	0.876
4	0.769	0.642	0.847	0.695	1.038	0.827
5	0.712	0.617	0.789	0.668	0.968	0.779

Source: Hoa and Inder (1996, tables 1 and 2).

Quintos and Phillips generalize these to tests for breaks in subvectors of the cointegrated regression. They derive an LM test for the hypothesis of cointegration versus the alternative of no cointegration (unlike the usual tests for which the null hypothesis is of no cointegration). They derive the (nonstandard) asymptotic distributions of the test statistics, tabulate the critical values, and present an empirical illustration.

The paper by Gregory and Hansen (1996) can be viewed as complementary to that of Hansen (1992) and Quintos and Phillips (1993). They propose ADF, Z_α , and Z_t -type tests designed to test the null of no cointegration against the alternative of cointegration in the presence of a possible regime shift. The shift considered is a single break in the intercept and/or slope coefficients in the cointegration relationship at an unknown break point. They derive the asymptotic distributions and tabulate the critical values using simulation methods. They also provide an empirical example of the money demand function in the US using annual data for the period 1901–1985 and quarterly data covering the period 1960:I–1990:IV. Apart from some detail, for the annual data, on balance both the tests, ignoring the breaks and allowing for the breaks, reject the null of no cointegration. For the quarterly data, however, the conventional ADF test fails to reject the null of no cointegration but allowing for shifts in both the intercept and slope coefficients results in rejecting the null of no cointegration.

Gregory, Nason, and Watt (1996) investigate in detail tests for cointegration following the testing procedures in Hansen (1992), in the context of a quadratic adjustment cost model. The detailed conclusions are too long to be discussed here.

13.9.3 System tests

When it comes to analysis of structural change in CI systems, there are two types of structural change to consider:

- (i) A change in the number of CI vectors or the rank of CI space,
- (ii) A change in the coefficients of a CI vector. In this case, since only the CI space is determined by the Johansen procedure, one needs to impose some normalization condition before one can devise tests for structural change.

Tests for a change in the rank of the CI space

These tests have been considered by Hansen and Johansen (1993) and Quintos (1995). Both start from the equation

$$R_{0t} = \alpha\beta'R_{1t} + u_t$$

defined in section 5.5.1 of chapter 5. This can be written as

$$R_{0t} = \Pi R_{1t} + u_t$$

Quintos separates the sample into different periods assuming the break dates known. For instance, let there be one known break date and let H and (H_1, H_2) be the parameters for the whole sample and the split samples. Her hypothesis is

$$H_0 : \text{Rank}(\Pi_1) = \text{Rank}(\Pi_2) = \text{Rank}(\Pi)$$

She uses the trace test statistic and derives the LR test for H_0 and its distribution. (Details of the test statistic and its distribution can be found in her papers.) Applying this to quarterly data on US government revenues and expenditures for the period 1947–1992 and using the fourth quarter of 1980 as the break date, she found the variables to be cointegrated before the break but not after the break.

Hansen and Johansen (1993) do not assume the break point to be known. They, instead, use recursive methods (which will enable the break point to be located). It would be difficult to plot the recursive coefficients. They, therefore, suggest looking at the larger roots λ_i (in the trace statistic) and derive tests for the significance of the differences in the successive estimates of the λ_i .

When the break points are known, as in the case that Quintos considers, one can always estimate the model in the separate regimes and find the number of CI vectors in the different regimes. What she has is a test for the significance of the observed differences.

When the break point is not known, the recursive estimation method in Hansen and Johansen would be useful in locating the break point. This is in the spirit of the CUSUM tests.

Tests for changes in the coefficients of a given CI vector

The tests we considered in the previous section by Hansen (1992), Quintos and Phillips (1993), etc. are for changes in the coefficients of a CI vector. However, these are based on the FM-OLS (or a single equation) estimation method. Tests for changes in the coefficient vector estimated by the Johansen procedure (or a system method) have been discussed by Seo (1995). He derives the tests and shows that the distributions are nonstandard, but different from those found by Andrews and Ploberger (1994). He also illustrates the methods with some empirical examples on money demand functions and term structure of interest rates.

13.10 Miscellaneous other issues

13.10.1 Seasonal models

Smith and Otero (1995) examine the effects of structural breaks on the HEGY test for seasonal integration (see chapter 12, section 12.3 for the HEGY test). They consider an exogenous change in the seasonal pattern and find that the distribution of the test statistic associated with the HEGY test is more skewed in the presence of the structural break. In their application to Columbian money supply and GDP as well as UK transportation, the results show that inference about a unit and seasonal roots is overturned. This is similar to the finding of Perron (1989) about the effects of structural breaks on unit root tests.

Ghysels and Perron (1996) show, both theoretically and via Monte Carlo studies that for linear models with a one-time structural break, seasonal adjustment using the X-11 seasonal adjustment procedure results in smoothed data and disguises structural instability and decreases the probability of detecting such a break. This result reinforces the conclusions of several studies (discussed in chapter 12) that analysis (including tests for structural breaks) should be conducted with seasonally unadjusted data.

13.10.2 Exact tests

Dufour and Kiviet (1996) discuss several finite sample tests of parameter constancy against the presence of structural change in first-order

dynamic models. They suggest a method for the derivation of an exact confidence set for the autoregressive parameter using an expanded regression and derive CUSUM, CUSUM of squares, and predictive tests (see section 13.3). The details are too long to be presented here. The methods depend on Monte Carlo simulations and their relationships with the bootstrap methods discussed in chapter 10 needs to be explored. They also use a union-intersection principle for combining several tests.

13.10.3 Spurious breaks

As we have seen in section 13.7, for the detection of multiple structural breaks, Maddala and Kim (1996a, 1996b) suggested Schwartz's Bayesian information criterion (BIC) to estimate the number of breaks based on the posterior distributions of break points. On the other hand, Nunes, Kuan, and Newbold (1995) and Nunes, Newbold, and Kuan (1996) (hereafter NNK) showed that the BIC failed to detect the true number of breaks, but estimated the maximum permitted number of breaks when the data generating process (DGP) is a random walk.

Maddala and Kim (1996a, 1996b) and NNK suggested using the BIC for detecting the number of structural breaks as follows (see section 3.7.1)

$$\hat{q} = \operatorname{argmin}_{q \leq q_{max}} \{BIC^q\}$$

$$BIC^q = \log(\hat{\sigma}_q^2) + [k + q(k + 1)] \frac{\log T}{T}$$

where k is the number of regressors and q_{max} is a given maximum number of breaks. For $\hat{\sigma}_q^2$, Maddala and Kim estimated them by using the posterior distributions of parameters, while NNK used the maximum likelihood estimators. Yao (1988) and Maddala and Kim (1996a) show that $\hat{q}$ is a consistent estimator of the true number of breaks, q_0 , provided $q_0 \leq q_{max}$ and errors are normally distributed.

To evaluate the performance of the BIC for detecting the number of breaks, NNK considered two cases for the DGP: (i) $\Delta y_t = e_t$ and (ii) $y_t = e_t$ where $e_t \sim IIN(0, 1)$. For the estimating models they used only deterministic regressors as follows: (i) $x_t = \{1\}$ (only constant) and (ii) $x_t = \{1, t\}$ (constant and trend). NNK's Monte Carlo results show that the BIC estimates q_0 consistently when the DGP is a white noise. When the DGP is a random walk, however, the BIC selects q_{max} in an overwhelming majority of occasions.

The estimating models of NNK, where only constant and/or trend

were used as regressors, belong to the trend-stationary processes (TSP) class, while one of the true DGPs was a random walk which belongs to the difference-stationary processes (DSP) class. Since it is well known that the inference based on the model of the TSP class leads to spurious conclusions when the true DGP belongs to the DSP class, Kim (1997) argues that the correct estimating model should include y_{t-1} as a regressor.

In contrast to the results reported by NNK (1996), Kim (1997) shows that the BIC for detecting the number of structural breaks estimates the true number of breaks accurately provided that appropriate regressors are used. When the DGP is a random walk with no break, however, additional deterministic regressors in the estimating model significantly lowers the probability of detecting the true number of breaks by the BIC. These results are similar to the fact that the power of a test of the unit root hypothesis against stationary alternatives decreases as additional deterministic regressors are included (Campbell and Perron, 1991, Rule 5). On the other hand, when the DGP is a white noise with no break, the BIC estimates the true number of breaks on most occasions regardless of the maximum number of breaks and additional deterministic regressors. These results highlight the well-known fact that when a true DGP belongs to the DSP class, inference based on the models of the TSP class misleads to spurious conclusions.

13.11 Practical conclusions

We have presented an overview of:

- (i) tests for structural change,
- (ii) effects of structural change on tests for unit roots and on cointegration tests,
- (iii) tests for unit roots under structural change,
- (iv) tests for structural change in cointegration relationships, and
- (v) structural change in seasonal models.

As far as possible, both the classical and Bayesian viewpoints have been presented.

The predominant conclusion that emerges from this voluminous work is that structural change does affect inference on unit roots, on cointegration, and seasonal integration, and it is important to allow for its possibility at the estimation stage itself. Some pieces of advice can be gleaned from all this work. These are:

- (i) Tests for structural change that merely test whether or not there is structural change are not very useful. What is of importance is the determination of the number and location of break points. One can argue that the detection of the number and location of break points can proceed after testing whether a break exists or not. However, this line of investigation results in an unknown pre-testing bias.
- (ii) There is a problem with the consistent estimation of the break point (which we have not discussed). The max statistics we discussed do not necessarily give a consistent estimate of the break point. They merely test whether or not there is a break. The problem of consistent estimation of the break point is dealt with in the paper by Bai and Perron (1995).
- (iii) There is no point in searching for a break over the entire time period of observations. Some prior information always exists about the dates of major shocks (real or financial) and this suggests the approximate location of the breaks. Searching for breaks around these dates is a valid procedure. The Bayesian procedure is more flexible in taking account of this limited prior information rather than the classical procedure.
- (iv) The Bayesian approach to unit roots and structural breaks is simpler and more straightforward than the classical approach. But the choice of the prior is an important issue and is more intricate in time series models (especially nonstationary models) than in the usual regression models. Much work has been done in this area but there is no consensus yet. Hence, some sensitivity analysis with alternative priors should be presented. Given the progress made in numerical methods, many more priors can now be handled than was the case earlier.
- (v) Since the switch from one regime to another is never sudden, models involving gradual structural change should receive more attention.
- (vi) Given the importance of structural breaks on inference about unit roots, cointegration, and seasonal integration, it is best to start with models allowing for structural breaks and then simplify them rather than test for structural breaks first and then complicate them. This is consistent with the recent philosophy of dynamic model building of going from general to specific models rather than the reverse. The model selection approach seems to be especially promising in this respect.

- (vii) More work needs to be done for a comparison of the different methods of determining the number of structural breaks.

References

- Andrews, D.W.K. (1993), "Tests for Parameter Instability and Structural Change with Unknown Change Point," *Econometrica*, 61, 821–856.
- Andrews, D.W.K. and Fair, R.C. (1988), "Inference in Nonlinear Econometric Models with Structural Change," *Review of Economic Studies*, 55, 615–640.
- Andrews, D.W.K. and W. Ploberger (1994), "Optimal Tests When a Nuisance Parameter is Present Only under the Alternative," *Econometrica*, 62, 1383–1414.
- Andrews, D.W.K., I. Lee, and W. Ploberger (1996), "Optimal Change-point Tests for Normal Linear Regression," *Journal of Econometrics*, 70, 9–38.
- Bai, J. and P. Perron (1995), "Estimating and Testing Linear Models with Multiple Structural Changes," Manuscript, Dept. of Economics, MIT.
- Banerjee, A., R.L. Lumsdaine, and J.H. Stock (1992), "Recursive and Sequential Tests of the Unit Root and Trend Break Hypotheses: Theory and International Evidence," *Journal of Business and Economic Statistics*, 10, 271–287.
- Ben-David, D. and D.H. Papell (1995), "The Great Wars, the Great Crash, and the Unit Root Hypothesis," *Journal of Monetary Economics*, 36, 453–475.
- Ben-David, D., R.L. Lumsdaine, and D.H. Papell (1997), "The Unit Root Hypothesis in Long-Term Output: Evidence from Two Structural Breaks in 16 Countries," Manuscript, University of Houston.
- Bianchi, M. (1995), "Time Series modeling in the Presence of Structure Change," Ph.D. dissertation, Universite Catholique de Louvain.
- Brown, R.L., J. Durbin, and J.M. Evans (1975), "Techniques for Testing the Constancy of Regression Relationships over Time," *Journal of the Royal Statistical Society, Series B*, 37, 149–192.
- Campbell, J.Y. and P. Perron (1991), "Pitfalls and Opportunities: What Macroeconomists Should Know about Unit Roots," in O.J. Blanchard and S. Fisher (eds.), *NBER Macroeconomics Annual*, The MIT Press, Cambridge, MA, 141–201.

- Campos, J., N.R. Ericsson, and D.F. Hendry (1996), "Cointegration Tests in the Presence of Structural Breaks," *Journal of Econometrics*, 70, 187–220.
- Cantrell, R.S., P.M. Burrows, and Q.H. Vuong (1991), "Interpretation and Use of Generalized Chow Tests," *International Economic Review*, 32, 725–741.
- Chow, G.C. (1960), "Tests of Equality between Sets of Coefficients in Two Linear Regressions," *Econometrica*, 28, 591–605.
- Christiano, L.J. (1992), "Searching for a Break in GNP," *Journal of Business and Economic Statistics*, 10, 237–250.
- Chu, C.S. and H. White (1992), "A Direct Test for a Changing Trend," *Journal of Business and Economic Statistics*, 10, 289–299.
- Corbae, D. and S. Ouliaris (1991), "A Test of Long-run Purchasing Power Parity Allowing for Structural Breaks," *The Economic Record*, 67, 26–33.
- Davies, R.B. (1977), "Hypothesis Testing When a Nuisance Parameter is Present Only Under the Alternative," *Biometrika*, 64, 247–254.
- (1987), "Hypothesis Testing When a Nuisance Parameter is Present Only Under the Alternative," *Biometrika*, 74, 33–43.
- DeJong, D. N. (1996), "A Bayesian Search for Structural Breaks in US GNP," in T. Fomby (ed.), *Advances in Econometrics: Bayesian Methods Applied to Time Series Analysis*, vol. 11, Part B, JAI Press, Greenwich, Conn.
- Diebold, F.X. and C. Chen (1996), "Testing Structural Stability with Endogenous Breakpoint: A Size Comparison of Analytic and Bootstrap Procedures," *Journal of Econometrics*, 70, 221–241.
- Dropsy, V. (1996), "Real Exchange Rates and Structural Breaks," *Applied Economics*, 28, 209–219.
- Dufour, J.M. (1982), "Recursive Stability Analysis of Linear Regression Relationships," *Journal of Econometrics*, 19, 31–76.
- Dufour, J.M. and J.F. Kiviet (1996), "Exact Tests for Structural Change in First Order Dynamic Models," *Journal of Econometrics*, 70, 39–68.
- Dufour, J.M., E. Ghysels, and A. Hall (1994), "Generalized Predictive Tests and Structured Change Analysis in Econometrics," *International Economic Review*, 35, 199–229.
- Garbade, K. (1977), "Two Methods for Examining the Stability of Regression Coefficients," *Journal of the American Statistical Association*, 72, 54–63.

- Garcia, R. (1997), "Asymptotic Null Distribution of the Likelihood Ratio Test in Markov Switching Models," *International Economic Review*, Forthcoming.
- Garcia, R. and P. Perron (1995), "An Analysis of Real Interest Rate under Regime Shifts," *Review of Economics and Statistics*, 78, 111–25.
- Gelfand, A.E. and A.F.M. Smith (1990), "Sampling-based Approaches to Calculating Marginal Densities," *Journal of the American Statistical Association*, 85, 398–409.
- Ghysels, E. and P. Perron (1996), "The Effects of Linear Filters on Dynamic Time Series with Structural Change," *Journal of Econometrics*, 70, 69–98.
- Gregory, A.W. and B.E. Hansen (1996), "Residual Based Tests for Cointegration in Models with Regime Shifts," *Journal of Econometrics*, 70, 99–126.
- Gregory, A.W., J.M. Nason, and D. Watt (1996), "Testing for Structural Breaks in Cointegrated Relationships," *Journal of Econometrics*, 71, 321–341.
- Hansen, B.E. (1990), "Lagrange Multiplier Tests for Parameter Instability in nonlinear Models," Paper presented at the Sixth World Congress of the Econometric Society, Barcelona, August 1990.
- (1992), "Tests for Parameter Instability in Regressions with I(1) Processes," *Journal of Business and Economic Statistics*, 10, 321–335.
- (1995), "Approximate Asymptotic p -values for Structural Change Tests," Manuscript, Boston College.
- (1996), "Inference When a Nuisance Parameter is Not Identified Under the Null Hypothesis," *Econometrica*, 64, 413–430.
- Hansen, H. and S. Johansen (1993), "Recursive Estimation in Cointegrated VAR Models," Institute of Mathematical Statistics, Working paper No. 1, University of Copenhagen.
- Hendry, D.F. and A.J. Neale (1991), "A Monte Carlo Study of the Effects of Structural Breaks on Tests for Unit Roots," in P. Hackl and A. Westlung (eds.), *Economic Structural Change*, Springer Verlag, New York, 95–119.
- Hoa and Inder (1996) "Diagnostic Test for Structural Change in Cointegrated Regression Models," *Economics Letters*, 50, 179–187.
- Holbert, D. (1982), "A Bayesian Analysis of a Switching Linear Model," *Journal of Econometrics*, 19, 77–87.

- Kao, C. and S.L. Ross (1992), "A CUSUM Test in the Linear Regression Model with Serially Correlated Disturbances," Syracuse University, May 1992.
- Kashiwagi, N. (1991), "Bayesian Detection of Structural Changes," *Annals of the Institute of Statistical Mathematics*, 43, 77–93.
- Kim, H.J. and D. Siegmund (1989), "The Likelihood Ratio Test for a Change-point in Simple Linear Regression," *Biometrika*, 76, 409–423.
- Kim, I.M. (1991), "Structural Change and Unit Roots," unpublished Ph. D. dissertation, University of Florida.
- (1997), "Detecting the Number of Structural Breaks," *Economics Letters*, 57, 145–148.
- Kim, I.M. and G.S. Maddala (1991), "Multiple Structural Breaks and Unit Roots in Exchange Rates," Paper presented at the Econometric Society Meeting at New Orleans, Dec. 1991.
- Kontrus, K. and W. Ploberger (1984), "A Simulation Study of Some Tests for Parameter Constancy in Linear Models," Paper given at the Econometric Society European meeting, Madrid.
- Kramer, W., W. Ploberger, and R. Alt (1988), "Testing for Structural Change in Dynamic Models," *Econometrica*, 56, 1355–1369.
- Kramer, W., W. Ploberger, and K. Kontrus (1989), "A Comparison of Two Significance Tests for Structural Stability in the Linear Regression Model," in B. Raj (ed.), *Advances in Econometrics and Modeling*.
- Kullback, S. and H.M. Rosenblatt (1957), "On the Analysis of Multiple Regression in k Categories," *Biometrika*, 67–83.
- Leybourne, S.J. and B.P.M. McCabe (1989), "On the Distribution of Some Test Statistics for Coefficient Constancy," *Biometrika*, 76, 169–177.
- Lo, A.W. and W.K. Newey (1985), "A Large Sample Chow Test for the Linear Simultaneous Equation," *Economics Letters*, 18, 351–353.
- Lumsdaine, R.L. and D.H. Papell (1997), "Multiple Trend Breaks and the Unit Root Hypothesis," *The Review of Economics and Statistics*, 79, 212–218.
- Maddala, G.S. and I.M. Kim (1996a), "Structural Change and Unit Roots," *Journal of Statistical Planning and Inference*, 49, 73–103.
- (1996b), "Bayesian Detection of Structural Changes," in D.A. Berry, K.M. Chaloner, and J.K. Geweke, Maddala, G.S. and

- I.M. Kim (eds.), *Bayesian Analysis in Statistics and Econometrics*, Wiley, New York.
- McCulloch, R.E. and R.S. Tsay (1994), "Bayesian Inference of Trend and Difference Stationary," *Econometric Theory*, 10, 596–608.
- Mizrach, B. (1992), "Mean Reversion in EMS Exchange Rates," Manuscript, University of Pennsylvania. paper presented at the Econometric Society meetings, Seattle, June 1992.
- Nabeya, S. and K. Tanaka (1988), "Asymptotic Theory of a Test for the Constancy of Regression Coefficients against the Random Walk Alternative," *The Annals of Statistics*, 16, 218–235.
- Nelson, C.R. and C.I. Plosser (1982), "Trends and Random Walks in Macroeconomic Time Series," *Journal of Monetary Economics*, 10, 139–162.
- Nunes, L.C., C.M. Kuan, and P. Newbold (1995), "Spurious break," *Econometric Theory*, 11, 736–749.
- Nunes, L.C., P. Newbold, and C.M. Kuan (1996), "Spurious Number of Breaks," *Economics Letters*, 50, 175–178.
- Nyblom, J. (1989), "Testing for the Constancy of Parameters over Time," *Journal of the American Statistical Association*, 84, 223–230.
- Park, S.B. (1991), "The Wald and LM Tests for Structural Change in a Linear Simultaneous Equations Model," Working paper, Carleton University.
- Perron, P. (1989), "The Great Crash, the Oil Price Shock and the Unit Root Hypothesis," *Econometrica*, 57, 1361–1401.
- (1990), "Testing for a Unit Root in a Time Series with a Changing Mean," *Journal of Business and Economic Statistics*, 8, 153–162.
- (1994), "Further Evidence on Breaking Trend Functions in Macroeconomic Variables," Manuscript, Université de Montreal.
- Perron, P. and T.J. Vogelsang (1992), "Testing for a Unit Root in a Time Series with a Changing Mean: Corrections and Extensions," *Journal of Business and Economic Statistics*, 10, 467–470.
- Phillips, P.C.B. (1995), "Bayesian Model Selection and Prediction with Empirical Applications," with comments by F.C. Palm and J.F. Richard and Reply by P.C.B. Phillips, *Journal of Econometrics*, 69, 289–365.
- Ploberger, W. (1983), "Testing the Constancy of Parameters in Linear

- Models," Paper given at the Econometric European meeting, Pisa.
- Ploberger, W. and W. Kramer (1990), "The Local Power of the CUSUM and CUSUM of Squares Tests," *Econometric Theory*, 6, 335–347.
- (1992), "The CUSUM Test with OLS residuals," *Econometrica*, 60, 271–285.
- Ploberger, W., W. Kramer, and R. Alt (1989), "A Modification of the CUSUM test in the Linear Regression Model with Lagged Dependent Variables," *Empirical Economics*, 14, 65–75.
- Ploberger, W., W. Kramer, and K. Kontrus (1989), "A New Test for Structural Stability in the Linear Regression Model," *Journal of Econometrics*, 40, 307–318.
- Quandt, R. (1960), "Tests of the Hypothesis that a Linear Regression System Obeys Two Separate Regimes," *Journal of the American Statistical Association*, 55, 324–330.
- Quintos, C.E. (1995), "Sustainability of the Deficit Process with Structural Shifts," *Journal of Business and Economic Statistics*, 13, 409–417.
- Quintos, C.E. and P.C.B. Phillips (1993), "Parameter Constancy in Cointegrating Regressions," *Empirical Economics*, 18, 675–706.
- Rao, C.R. (1952), *Advanced Statistical Methods in Biometric Research*, John Wiley, New York.
- (1965), *Linear Statistical Inference and Its Applications*, 2nd edn, 1973, John Wiley, New York.
- Rappoport, P. and L. Reichlin (1989), "Segmented Trends and Nonstationary Time Series," *Economic Journal*, 99 (supplement), 168–177.
- Reichlin, L. (1989), "Structural Change and Unit Root Econometrics," *Economics Letters*, 31, 231–233.
- Rose, A.K. (1988), "Is the Real Interest Rate Stable?" *Journal of Finance*, 43, 1095–1112.
- Salazar, D. (1982), "Structural Changes in Time Series Models," *Journal of Econometrics*, 19, 147–163.
- Sen, P.K. (1980), "Asymptotic Theory of Some Tests for a Possible Change in the Regression Slope Occurring at an Unknown Time Point," *Zeitschrift für Wahrscheinlichkeitstheorie und Verwandte Gebiete*, 52, 203–218.
- Seo, B.S. (1995), "Tests for Structural Change in Cointegrated Systems," Manuscript, University of Rochester.

- Smith, J. and J. Otero (1995), "Structural Breaks and Seasonal Integration," Warwick economic research papers No. 435, University of Warwick.
- Stock, J.H. (1994), "Unit Roots, Structural Breaks and Trends," in R.F. Engle and D.L. McFadden (eds.), *Handbook of Econometrics*, vol. IV, Elsevier, Amsterdam, 2740–2841.
- Wright, J.H. (1993), "The CUSUM Test Based on Least Squares Residuals in Regressions with Integrated Variables," *Economics Letters*, 41, 353–8.
- Yao, Y.C. (1988), "Estimating the Number of Change-Points via Schwartz's Criterion," *Statistics and Probability Letters*, 6, 181–189.
- Zivot, E. and D.W.K. Andrews (1992), "Further Evidence on the Great Crash, the Oil-Price Shock and the Unit Root Hypothesis," *Journal of Business and Economic Statistics*, 10, 251–287.
- Zivot, E. and P.C.B. Phillips (1994), "A Bayesian Analysis of Trend Determination in Economic Time Series," *Econometric Reviews*, 13, 291–336.

14

Outliers and unit roots

14.1 Introduction

In the preceding chapter we discussed the effects of structural breaks on unit root tests and cointegration tests. The breaks considered were breaks in trends, mostly one break and in a few cases multiple breaks. In this chapter we shall discuss another related problem, that of outliers. The structural breaks considered in the previous chapter correspond to one type of outliers.

There is a large amount of literature on outliers in time series, but this has had an impact on unit roots only in recent years. As Balke and Fomby (1991a, 1994) noted, there is considerable evidence of different types of outliers in macroeconomic time series. In this chapter we shall consider the different types of outliers and also outlier-robust unit root tests. Outlier-robust estimation methods for nonstationary time series will also be discussed briefly. Much of this material is covered in Maddala and Yin (1996) and Yin and Maddala (1996, 1997).

14.2 Different types of outliers in time series models

Outliers are aberrant observations that are away from the rest of the data. They can be caused by recording errors or unusual events such as changes in economic policies, wars, disasters, and so on, as for instance those considered by Perron (1989). They are also likely to occur if the errors have fat-tailed distributions (as is the case with financial time series). Outliers can also be caused by misspecification of the estimated relationships, e.g., linear instead of nonlinear relationships, omitted variables, and so on.

There is no question that outliers can cause problems with inference.

The only question is what we should do with them. Legendre, in the first paper on least squares in 1805, suggested throwing these observations out. So did Edgeworth in 1887. However, if outliers are caused by misspecification of the relationships estimated, then a proper course is to change the specification and if they are caused by fat-tailed error distributions, a proper course is to use robust methods. Thus, there are three courses of action one can take (see Donald and Maddala, 1993, p. 680):

- (i) Throw the rascals out. (It will be argued in a later section that although often suggested, this is not always a desirable course of action in time series models.)
- (ii) Leave them in, but under control (robust methods).
- (iii) Change the model.

For (i) we need to identify the outliers. For the linear regression model, there are many diagnostics to identify outliers. Most of these have been built into the several standard computer programs. However, many of them are not applicable for time series data.

In time series problems, because successive observations are correlated, outliers can cause more problems for detection. Fox (1972) first addressed outlier problems in time series by classifying outliers as *additive outliers* (AO) and *innovation outliers* (IO). An additive outlier occurs when a particular observation is bad but subsequent observations in time series are not affected. An innovation outlier occurs when the effect of a large innovation is perpetrated through the dynamics of the model. Tsay (1988) extended this classification to allow for structural changes as well and considers *transient changes* (TC), *level changes* (LC), and *variance changes* (VC). This classification will be followed here.

Although the discussion of these outliers can be conducted in terms of more general time series models, we shall define them in terms of first-order autoregressive models. The definition of the different types of outliers according to Tsay's classification is as follows: let y_t be the time series without any disturbances and $f(t)$ denote the disturbance so that the observed time series is $z_t = y_t + f(t)$. We shall consider three models:

$$\begin{array}{ll}
 \text{(Model A)} & \text{No drift case: } y_t = x_t \\
 \text{(Model B)} & \text{Drift case: } y_t = \alpha + x_t \\
 \text{(Model C)} & \text{Trend case: } y_t = \alpha + \delta t + x_t
 \end{array} \tag{14.1}$$

where

$$x_t = \rho x_{t-1} + e_t, \quad e_t \sim IN(0, \sigma^2)$$

These are standard models for studying the properties of unit root tests. The different outlier models in this simplified setup are defined as follows:

$$\begin{aligned} \text{AO model: } f(t) &= w_{AO} \zeta_t^m \\ \text{IO model: } f(t) &= \frac{1}{1-\rho L} w_{IO} \zeta_t^m \\ \text{LC model: } f(t) &= \frac{1}{1-L} w_{LC} \zeta_t^m \\ \text{VC model: } f(t) &= \frac{1}{1-\rho L} w_{VC} e_t^m \end{aligned} \quad (14.2)$$

where L is the lag operator, $w_i, i = AO, IO, LC, VC$, denote the magnitudes of the respective disturbances, m denotes the location of the outlier, and

$$\zeta_t^m = \begin{cases} 1 & \text{if } t = m \\ 0 & \text{otherwise} \end{cases}$$

$$e_t^m = \begin{cases} 0 & t < m \\ e_t & t \geq m \end{cases}$$

The following important points are worth noting regarding these models:

- (i) The behavior of the series z_t in the IO and VC models depends on the AR structure of y_t which, for simplicity, we have assumed to be an AR(1) model. On the other hand, the behavior of z_t under the AO and LC models does not depend on the structure of y_t .
- (ii) IO affects the m th observation by w_{IO} and affects the subsequent observations by $\rho^{t-m} w_{IO}$. Thus, its impact on the subsequent observations decays at the rate of ρ . In the case of $\rho = 1$, the impact is constant and permanent over time and thus the pattern of the impact will coincide with that of LC. For the VC model we have

$$z_t = \rho z_{t-1} + w_{VC} e_t^m + e_t = \rho z_{t-1} + e'_t$$

where

$$e'_t = \begin{cases} e_t & t < m \\ (1 + w_{VC}) e_t & t \geq m \end{cases}$$

In this case, the variance of the innovation driving up the series z_t changes at time m from σ^2 to $(1 + w_{VC})^2 \sigma^2$. Thus the subsequent z_t s are also affected.

- (iii) The sources of these disturbances can be different. AO can come from recording errors, natural disasters, a bizarre day on the stock market, and so on. IO normally represents the onset of an external cause. Several factors can cause LC which represents a changing mean of the series. This falls into the category of structural change discussed in the previous chapter. VC is another example of structural change, but this time it is the variance of innovations that changes. One example of this is the exchange rate when it switches from the fixed exchange rate regime to the flexible exchange rate regime.
- (iv) The implications of these models on unit root testing are different. Note that in the presence of these disturbances, both the null and the alternative hypotheses should encompass these disturbances which makes them different from the ordinary unit root hypotheses. For example, to test the unit root hypotheses in the model with drift against the alternative of trend stationarity in the presence of LC, the null hypotheses now become a one-time change in the drift of the random walk model while the alternative becomes a changing mean in the trend-stationary model.
- (v) Finally in following the literature on outliers and structural breaks, it is important to bear in mind the distinction between outliers in levels and in first-differences of the time series. In Perron (1989) the terms AO and IO refers to breaks in the trend function (in his models B and C) because he considers outliers (AO and IO) in first-differenced series. In Perron (1990) the outliers refer to breaks in means (levels). In Balke and Fomby (1991a) also, outliers corresponds to changes in trend whereas in Hendry and Neale (1990), Franses and Haldrup (1994), Lucas (1995a,b), and Yin and Maddala (1997) outliers affect the levels of the series. When discussing the properties of unit root tests, this distinction is important.

14.3 Effects of outliers on unit root tests

It has been noted in several papers (which will be reviewed in the following sections) that outliers lead to size distortions (overrejection or underrejection of the null) in unit root tests. The type of distortion depends on the nature of the outlier (AO, IO, etc.). Also, as mentioned earlier, the studies fall into two categories: outliers in first-differences

and outliers in levels. Outliers in first-differences refer to breaks in trends as in Perron (1989, models B and C in his paper) whereas outliers in levels refer to breaks in the mean as in Perron (1989, his model A) and Perron (1990).

14.3.1 Outliers in first-differences

These are discussed in Perron (1989) and Balke and Fomby (1991a). Perron (1989) first pointed out that standard unit root tests of the unit root hypothesis against trend-stationary alternatives are biased toward the nonrejection of the unit root hypothesis if the data generation process is that of stationary fluctuations around a trend with a one-time break. He suggested a two-step procedure. First, he detrended the series. Then if $\tilde{y}_t$ are the detrended series, he considered the limiting distributions of the normalized $\tilde{\rho}$ and the associated t -statistics for the three models he considered, based on the regression

$$\tilde{y}_t = \rho \tilde{y}_{t-1} + e_t$$

(see the discussion of Perron's paper in chapter 13).

Balke and Fomby (1991a) is another example of outliers in first-differences. They proposed a random-level, random-slope model which encompasses deterministic trends, shifting and segmented trends, and stochastic trends. While the primary purpose of this model is to embody both the hypotheses of small, frequent permanent shocks and large, infrequent permanent shocks, this model also links the outliers in the first-difference of the series to the different types of trend models. In their model, the trend component X_t is given by

$$X_t = \beta + X_{t-1} + G_{t-1} + I_{1t}S_{1t}$$

and

$$G_t = \theta + G_{t-1} + I_{2t}S_{2t}$$

where G_t is the stochastic growth component, S_{1t} and S_{2t} are independent with $S_{it} \sim N(\mu_i, \sigma_i^2)$, $i = 1, 2$. I_{1t} and I_{2t} are independent random variables taking on values of 0 or 1 and are assumed to have an independent Bernoulli distribution with $\Pr(I_{it} = 1) = p_i$, $i = 1, 2$ and $0 \leq p_i \leq 1$. In their model the outliers or disturbances are modeled as random because the original purpose is to model random shocks.

There are several special cases of this model that are of interest. Assume, for simplicity that $\theta = 0$ and $G_{-1} = 0$.

- Case 1** $p_1 = 0$ and $p_2 = 0$. In this case we have $X_t = X_0 + \beta t$ which is a deterministic trend.
- Case 2** $p_1 = 1$ and $p_2 = 0$. In this case, we have $X_t = X_0 + \beta t + \sum_{i=1}^t S_{1i}$. Thus X_t is a random walk with drift.
- Case 3** $p_1 = 0$ and $p_2 = 1$. In this case, we get $X_t = \beta + \sum_{i=1}^t S_{2,i-1}$ which means ΔX_t is a random walk. Note that we get this only by assuming $\theta = 0$ and $G_{-1} = 0$. If either one is nonzero, ΔX_t will be a random walk with a drift.
- Case 4** $p_1 > 0$ and $p_2 > 0$. In general, this will make X_t similar to a shifting or segmented trend where the breaks in trend are determined by Bernoulli draws.

Balke and Fomby (1991b) argue that the usual unit root tests will break down in this framework. They showed that the Dickey–Fuller test statistics will converge to one in probability and their asymptotic distributions depend on $p_i \sigma_i^2$ and thus the Dickey–Fuller tests are incapable of distinguishing between small frequent and large infrequent shocks. This finding is similar to those in Perron (1989, 1990), Rappoport and Reichlin (1989), and Hendry and Neale (1990) who also showed that standard Dickey–Fuller tests have problems in distinguishing between a shifting or segmenting deterministic trend and a stochastic trend.

14.3.2 Outliers in levels

One of the models considered in Perron (1989) is that of a break in the mean. This model has been analyzed in Perron (1990). These results by Perron for a break in mean have been extended in several directions and detailed critical values have been tabulated for the different cases. (We shall not go into the details of these tables here.) The extensions are as follows

- (i) Perron and Vogelsang (1992b) make some corrections to the distributions of unit root tests for the AO model in Perron (1990). Note that this model referred to nontrending data.
- (ii) Perron and Vogelsang (1992a) extend the paper of Perron (1990) to the case of the unknown break point, and present detailed critical values.
- (iii) Perron and Vogelsang (1993) make corrections to the distributions for the AO model in Perron (1989) with trending data. They also present extensions to the case of the unknown break

point and provide critical values for several models that could be useful in applications.

- (iv) Vogelsang and Perron (1994) provide further extensions to the unknown break point case. Unlike the previous studies, where the break is considered only under the alternative of trend stationarity, they derive the limiting distributions of statistics in both the AO and IO frameworks where a shift is permitted under the null hypothesis of a unit root. They show that the limiting distributions are invariant to a mean shift but not to a change in slope. The paper provides several tables of critical values for the different tests.
- (v) Finally, Vogelsang (1994) examines different procedures for dealing with additive outliers in unit root tests.

One important issue in all this work is how to actually apply unit root tests in the presence of outliers. There are two possible approaches in the case of additive outliers. The first is to formally test for the presence and location of outliers and then to perform a modified DF unit root test. There are two ways of modifying the DF regression. One is to use dummy variables to the DF regression to account for outliers. This is the procedure followed by Franses and Haldrup (1994). Vogelsang (1994) argues that this is the preferred approach. The other method involves first removing the outliers from the data, treating the corresponding observations as missing, and then estimating the DF regressions. Vogelsang (1994) shows that this leads to misleading inferences and this procedure should not be used.

The second approach to dealing with additive outliers is to use unit root tests that are robust to MA errors with a root close to unity (which is a consequence of additive outliers). These modified tests suggested by Perron and Ng (1996) have been discussed in chapter 4 (section 4.3.4).

Franses and Haldrup (1994) addressed the problem of unit root tests in the presence of AO by using a probabilistic setup. They showed that there is *overrejection* of the null hypothesis using standard DF tests in the presence of additive outliers. Yin and Maddala (1997) adopted a similar approach and studied the effect of different types of disturbances on unit root tests. They considered not only additive outliers but also innovation outliers, level changes, and variance changes. Lucas (1995a) also looked at the problems of standard DF tests from the point of view of robust estimation and also showed that standard DF tests in the

presence of AO and IO suffer from size distortions. In all these papers, outliers are defined in levels of the series.

14.3.3 The impact of disturbances

To focus on the impact of the presence of disturbances, we assume that the error term is *iid*. Consider the DF regression

$$z_t = D_t + \rho z_{t-1} + u_t \quad (14.3)$$

where D_t is the deterministic trend and we shall consider three different cases:

- (i) Model A: $D_t = 0$,
- (ii) Model B: $D_t = \alpha$,
- (iii) Model C: $D_t = \alpha + \delta t$.

AO model

Because the impact of finite AOs will go to zero as sample size goes to infinity, it is necessary to use a probabilistic model to study the AO model. Franses and Haldrup (1994) used the following probabilistic setup to study the large-sample theory for model A

$$z_t = y_t + \theta d_t$$

where d_t is a Bernoulli variable taking the value of 1 or -1 with a probability $\pi/2$ and θ is the magnitude of AO. This is similar to the measurement error models except that a Bernoulli variable instead of a normal variable controls the contamination process.

Using this setup, Franses and Haldrup showed that the asymptotic distributions of both the Dickey–Fuller K -test statistics and t -test statistics in model A shift to the left in the presence of AO. Thus by using the usual critical values, one tends to overreject the null hypothesis of unit root in the presence of AO. Their results can be readily extended for models B and C as well. Yin and Maddala (1997) adopt their setup but use a different method to derive the asymptotic distributions. We shall present their results here (they will be referred to as YM).

Let y_t be generated by model A to model C as defined in (14.1) and let the observed z_t be generated by

$$z_t = y_t + w_{AO} d_t$$

where d_t is *iid* Bernoulli with

$$d_t = \begin{cases} 1 & \text{with probability } \pi/2 \\ 0 & \text{with probability } 1 - \pi \\ -1 & \text{with probability } \pi/2 \end{cases}$$

Note that in this case the u_t in the DF regression (14.3) are no longer *iid*. They follow an MA(1) process

$$u_t = e_t + w_{AO}d_t - w_{AO}d_{t-1} = v_t + \theta v_{t-1}$$

with $\gamma_0 = \text{var}(u_t) = \sigma^2 + 2\pi w_{AO}^2$ and $\lambda^2 = \sigma_v^2(1 + \theta)^2 = \sigma^2$. It is well known that the asymptotic distribution of the DF-test statistics has an extra term in the numerator

$$\frac{1}{2} \left(1 - \frac{\gamma_0}{\lambda^2} \right) = -\pi \left(\frac{w_{AO}}{\sigma} \right)^2$$

Thus the asymptotic distributions of the DF-test statistics are

$$T(\hat{\rho} - 1) = \frac{\int W^*(r)dW(r) - \pi(w_{AO}/\sigma)^2}{\int W^*(r)^2 dr}$$

$$t_{\hat{\rho}} = \left(1 + \pi \left(\frac{w_{AO}}{\sigma} \right)^2 \right)^{-1/2} \frac{\int W^*(r)dW(r) - \pi(w_{AO}/\sigma)^2}{\int W^*(r)^2 dr}$$

where $W(r)$ is the standard Wiener process and

$$W^*(r) = \begin{cases} W(r) & \text{for model A} \\ \overline{W}(r) & \text{for model B} \\ \widehat{W}(r) & \text{for model C} \end{cases} \quad (14.4)$$

$$\overline{W}(r) = W(r) - \int W(r)dr$$

$$\widehat{W}(r) = W(r) - 4 \left[\int W(r)dr - \frac{3}{2} \int rW(r)dr \right] \\ + 6r \left[\int W(r)dr - 2 \int rW(r)dr \right]$$

The extra term in the numerator of the asymptotic distributions are due to the MA structure of the error terms, and they are the same for all three models A, B, and C. These terms are positive and cause the asymptotic distributions to shift to the left, and this results in overrejection if the standard critical values are used. The extra terms depend on the probability of occurrence of AO, π , the magnitude of AO, w_{AO}^2 , and

the variance of the innovation e_t, σ^2 . One thing to note is that only the product of π and $(w_{AO}/\sigma)^2$ matters. In other words, large shocks with small chance of occurrence have the same effect as that of small shocks with high probability of occurrence. Balke and Fomby (1991b) pointed out this fact and argue that the DF tests cannot distinguish infrequent large shocks from frequent small shocks. Note that a positive AO and a negative AO with the same magnitude have the same effect on the limiting distributions which depend on $(w_{AO}/\sigma)^2$, not (w_{AO}/σ) .

From the derivation we know that the problem is caused by the MA structure of the error terms in the DF regression, a certain departure from the *iid* case where the DF tests are appropriate. This might suggest the use of more general tests like the ADF test and/or PP tests to handle the serial correlation in errors. However, the tests discussed in Perron and Ng (1996) would be more appropriate, as mentioned earlier.

The IO and LC models

Under the null hypothesis, the impact of IO will be the same as that of LC. Thus, in testing for a unit root, the test statistics derived from these two models have the same limiting distributions under the null hypothesis. This is why they are discussed together in the same place. However, the test statistics have different power functions. We shall discuss models B and C only, since model A is meaningless under this circumstance.

Perron (1989, 1990) and Perron and Vogelsang (1992a) have studied these cases in detail. But there are some differences in their models and those considered by YM. Perron and Vogelsang consider outliers in the first-difference while YM, following the literature, consider the traditional definitions of IO and LC. Although the results are similar, one should notice that the models are different.

First, YM get the limiting distribution of the DF tests. Let y_t be generated by equation (14.1) and (14.2) for models B and C, respectively, and let the observed z_t be generated by

$$z_t = \begin{cases} y_t & \text{for } t < m \\ y_t + w & \text{for } t \geq m \end{cases}$$

where m is the date of break. It is relatively easy to show that both the limiting distributions of the DF, K -, and t -statistics will not change in this case under the null hypothesis. The key feature is that the partial sum process has a higher convergence rate than the extra constant term

w after time m . Because the proof is quite lengthy and tedious, we leave it out here.

A question arises as to why one should worry about the existence of IO and LC in the model if they do not cause any trouble in the sense that the limiting distributions do not change even if they exist in the model. There are two reasons. The first reason is the problem of small sample size distortions. Since the DF tests suffer from small sample size distortions, in empirical work researchers normally use those critical values adjusted for small sample biases. But those critical values are normally derived via simulations based on the uncontaminated models. Apparently in small samples the test statistics are definitely different from the usual ones. So, even if researchers use those adjusted critical values, small sample size distortions are still expected to exist in the presence of IO and LC.

Secondly, Perron (1989) has shown that under the alternative hypothesis he specified, $\hat{\rho}$ would be inflated toward 1. This is also true for the LC model. It can be shown that for the LC model we have

$$\hat{\rho} \rightarrow \frac{\rho\gamma_0 + w^2\lambda(1 - \lambda)}{\gamma_0 + w^2\lambda(1 - \lambda)}$$

for both models B and C under the alternative, where γ_0 is the variance of z_t , w is the magnitude of LC, and λ is a fixed constant defined as $m = \lambda T$. YM use a fixed λ to make sure that both before-break sample and after-break sample grow at the same rate as the full sample. Thus under the alternative $\hat{\rho}$ is not a consistent estimator and it will overestimate the true parameter. This will make the low power problem of the unit root tests worse.

Surprisingly, this result does not hold for the IO model. Using the same procedure, it can be shown that $\hat{\rho}$ is still consistent under the alternative for the IO model. So there will be no corresponding power loss for the IO model.

These results on the behavior of the DF test statistics in the presence of IO and LC can be summarized as follows: both statistics converge to the usual DF distributions under the null, while they normally have extra small sample biases caused by the contamination. These biases persist even if one uses the small sample critical values generated via simulations of the uncontaminated models. However, under the alternative, the test statistics in IO and LC models have different properties. In the LC model the estimate is biased toward 1 and goes to 1 when the magnitude of the level change becomes extremely large relative to the

variance of the innovations. Thus there is a tendency to underreject the unit root hypothesis for the LC model. On the other hand, for the IO model, the estimate of the autoregressive parameter converges to its true value asymptotically. Nevertheless in small samples, it may behave quite differently from that in the uncontaminated model. YM confirm these results using some simulations to show the small sample size distortions and power functions.

The VC model

In the presence of VC, the model reduces to

$$z_t = y_t + \frac{1}{1 - \rho L} w_{VC} e_t^m \quad (14.5)$$

with e_t^m defined after equation (14.2). YM investigate the behavior of the DF-test statistics in the presence of VC. Let y_t be generated by model A to model C in (14.1) and let the observed z_t be generated by (14.5). They then derived the following

$$T(\hat{\rho} - 1) \Rightarrow \frac{\int W_i^*(r) dW^*(r)}{\int W_i^*(r)^2 dr}$$

$$t_{\hat{\rho}} \Rightarrow \frac{\int W_i^*(r) dW^*(r)}{(\int W_i^*(r)^2 dr)^{1/2} [\lambda \sigma^2 + (1 - \lambda) \sigma_1^2]^{1/2}}$$

where

$$W_i^*(r) = \begin{cases} W^*(r) & i = \text{model A} \\ W^*(r) - \int W^*(r) dr & i = \text{model B} \\ W^*(r) - 4 \left[\int W^*(r) dr - \frac{3}{2} \int r W^*(r) dr \right] \\ \quad + 6r \left[\int W^*(r) dr - 2 \int r W^*(r) dr \right] & i = \text{model C} \end{cases} \quad (14.6)$$

and

$$W^*(r) = \begin{cases} \sigma W(r) & r < \lambda \\ \sigma_1 W(r) & r \geq \lambda \end{cases}$$

$$\sigma_1^2 = \text{var}((1 + w_{VC})e_t) = (1 + w_{VC})^2 \sigma^2$$

Note that the difference between the above distributions and the standard DF distributions is that instead of $W(r)$ we have $W^*(r)$ defined in (14.6). This is simply because the variance of the innovation driving up the series changes at time m . By modifying those derivations using the functional central limit theorem, these will give an integral from 0 to λ with the initial variance σ^2 and an integral from λ to 1 with the

changed variance σ_1^2 . Thus by defining $W^*(r)$ one can easily derive the above limiting distributions.

The implications of these limiting distributions are not obvious to observe. One cannot tell immediately whether the limiting distributions shift to the left or to the right. One obvious thing is that these limiting distributions are not the same as the standard DF distributions unless $\sigma^2 = \sigma_1^2$, i.e., there is no variance change. Another observation is that the date when the variance changes matters here. Different values of λ result in different $W^*(r)$ and thus different limiting distributions. Although the shift of the distributions are not clear, the shift depends on λ and $k = \sigma_1^2/\sigma^2$.

One can tackle this problem by computing the distributions using some numerical methods for given λ . Yin and Maddala (1997) investigate this through a Monte Carlo simulation.

The presence of the VC results in the most complicated situation. In the presence of the AO, there is a tendency for overrejection of the unit root null, while there is a tendency of underrejection in the presence of the LC. But for the VC case, whether there is a tendency for underrejection or overrejection of unit root hypothesis depends on the values of λ and k .

14.4 Outlier detection

In the preceding section we noted that outliers produce misleading inferences in the application of unit root tests and that the nature of the distributions depends on the type of the outlier.

There are two solutions often suggested for handling outliers. The first is to detect the outliers and discard them (throw the rascals out). The other is to use robust methods. We shall discuss these two solutions in turn. For the use of the robust methods, detection of the outliers is not necessary although, as will be noted later, robust estimation methods can also help the detection of outliers.

In the regression context, the procedures used for the detection of outliers fall into the categories of detection methods and likelihood displacement methods. See Cook (1987) and, for a critique, Donald and Maddala (1993).

Fox (1972) first proposed distinguishing between two types of outliers in time series: AO and IO. He also proposed two parametric models for these categories and suggested likelihood ratio tests for outliers when the location and type of the outlier is known. Abraham and Yatawara

(1988) propose a score test for identifying the location of an outlier and a procedure for distinguishing between AO and IO, assuming that the location of the outlier is known. On the other hand, Abraham and Chung (1989) and Bruce and Martin (1989) discuss the use of deletion diagnostics to detect outliers in time series. The deletion diagnostics can refer to the effect on the estimates of the regression coefficients or the error variance. Bruce and Martin found that in the case of time series models, deletion diagnostics based on innovation variance σ^2 give a clearer indication of an outlier than diagnostics based on estimates of the regression parameters. Ledolter (1990) discusses outlier diagnostics based on the likelihood displacement method in time series models. Ljung (1993) contains a detailed discussion of a comparison between likelihood-based methods and deletion methods, although the discussion is in the context of only additive outliers in ARMA models.

Atkinson, Koopman, and Shephard (1997) also use deletion methods to detect outliers and breaks in time series. Their model is based on measuring the influence of deletion measures. In the usual regression model, an outlier can be detected by the effect of its deletion on the residual sum of squares. An equivalent procedure is to add an extra variable to the mean of the deleted observation (dummy variable method), also known as intervention analysis. Atkinson *et al.* argue that such intervention analysis can also be conducted in unobserved components and structural time series models. However, these models contain more than one variance, so the effect of intervention has to be measured by the change in the individual variances. The procedures for doing this are rather complicated to be reviewed here, but Atkinson *et al.* demonstrate that this can be done by the use of score statistics and recent developments in smoothing and filtering. They also provide four examples that illustrate the fragility of inferences about outliers and level breaks.

As for methods suggested for distinguishing between different types of outliers, there are many studies. But many of these are valid under very restrictive assumptions. For instance the procedure in Abraham and Yatawara (1988) is valid for only one outlier with known location. Fiebig and Maasoumi (1990) suggest nesting both AO and IO in a comprehensive model

$$y_t = \alpha y_{t-1} + \beta' x_t + \phi_1 z_t^{(m)} + \phi_2 z_{t-1}^{(m)} + e_t$$

where $z_t^{(m)} = 1$ if $t = m$ and $z_t^{(m)} = 0$ otherwise (m denotes the location of the outlier). For the AO model $\phi_2 = -\alpha\phi_1$, but the IO model requires $\phi_2 = 0$.

The main drawback of the Fox method is that it is designed for the case of a single outlier. This procedure was extended for several outliers and studied later in full length by several researchers. Tsay (1988) combined some previous work on this issue and proposed a unified model to deal with different types of outliers and structural breaks. He considered not only AO and IO but also TC, LC, and VC, and put them into a unified model. This procedure was later modified by Chen and Liu (1993) to take care of bogus detected outliers, and by Balke (1993) to better distinguish IOs and LCs. The procedure described by Tsay (1988) consists of the following steps: specifying and estimating an initial ARMA model, detecting outliers based on some prespecified criteria, discarding the outliers, and then re-specifying and re-estimating the ARMA model. Balke (1993) argues that Tsay's method does not always perform satisfactorily if level shifts are present. He suggests a simple modification of Tsay's procedure which consists of two steps: (i) estimating an initial ARMA (0,0) or white noise model in addition to the ARMA model as Tsay suggests and (ii) combining the results from the two outlier searches to form a single intervention model.

There is an important problem with the procedure by Tsay noted by Balke and Fomby (1991a, p. 73) and Balke (1993). The estimation of the parameter from models based on the data with outliers left in, may produce a bias in the parameter estimates and thus affect the efficiency of outlier detection. The paper by Chen and Liu (1993) addresses this problem.

In addition to these classical approaches there are Bayesian approaches as well. Some references in this area are West (1986), West and Harrison (1986), and Durbin and Cordero (1994).

Another avenue is to use a robust estimation procedure at the first stage in outlier detection. The idea is, as suggested in Franses and Lucas (1995), that a useful by-product of robust estimators is that weights are obtained for the individual observations and these weights can be used to detect outliers and structural breaks. A low weight indicates that the observation does not correspond to the rest of the data. However, the main idea of outlier detection is to throw them out. The main idea of robust estimation is to leave them in. But the procedures of starting outlier detection on the basis of models estimated with outliers left in as in Tsay and Balke produces biased estimates and can lead to spurious outliers. This is precisely the reason for all the deletion diagnostics and methods suggested in regression models (Cook and Weisberg, 1982, Donald and Maddala, 1993).

14.5 Robust unit root tests

In the previous section we presented a brief review of outlier detection methods. The purpose of these methods is to throw the outliers out. Another alternative is to use robust methods. In the area of robust estimation, Huber (1964, 1973) introduced the class of M estimators. For the linear regression model

$$y_t = x_t' \beta + u_t$$

An M estimator is the minimand of

$$\min_{\beta} \sum \rho \left(\frac{y_t - x_t' \beta}{\sigma} \right)$$

where $\rho(\cdot)$ is a function defined on $\Re$ and σ^2 is the variance of u_t . For $\rho(u) = u^2$ one obtains the OLS estimator. The first-order condition for the minimization problem is

$$\sum \psi \left(\frac{y_t - x_t' \beta}{\sigma} \right) x_t' = 0$$

where $\psi(u) = d\rho(u)/du$. Two commonly used functions for $\psi(\cdot)$ are the Huber ψ function

$$\psi(u) = \min(c, \max(-c, u))$$

and the bi-square ψ function

$$\psi(u) = u(c^2 - u^2)^2 1_{[-c, c]}$$

where c is a tuning constant and 1_A is an indicator function of the set A .

One would expect that the M estimators would be applicable in the case of time series to take care of outliers as well. But Martin and Yohai (1986) pointed out that the performance of M estimators in the presence of outliers, especially AO, is not satisfactory. This is because in ARMA models, contaminated y_t s are also included in explanatory variables. Therefore Martin and Yohai (1986) proposed to use *generalized* M (GM) estimators. In an AR(1) model

$$y_t = \rho y_{t-1} + u_t$$

a GM estimator is a solution of

$$\sum \eta \left(\frac{y_t - \rho y_{t-1}}{\sigma}, y_{t-1} \right) y_{t-1} = 0$$

where $\eta(\cdot, \cdot)$ is some robustifying function. A typical example of $\eta(\cdot, \cdot)$ as suggested by Martin and Yohai is

$$\eta(u, v) = \psi(u)\psi(v)$$

with ψ being the Huber ψ function or the bisquare ψ function. Martin and Yohai showed that the GM estimator can handle both AO and IO quite successfully in the AR(1) model. The problem with the GM estimator is its performance when the order of the AR structure increases.

In the area of robust estimations, *high breakdown point* (HBP) estimators have been widely suggested over the last decade. These estimators are concerned with the concept of the breakdown point of an estimator. The breakdown point of an estimator measures the largest possible proportion of outliers in the data set an estimator can tolerate before it collapses to some nonsensical value. Because the performance of the M estimator is not always satisfactory, HBP estimators have been introduced into time series analysis as well. For a brief discussion see Lucas (1995a), who presents the class of S estimators introduced by Rousseeuw and Yohai (1984) and the MM estimator introduced by Yohai (1987). The S estimator can be calculated as follows: suppose that for a given β , a scalar estimate $\sigma_S(\beta) > 0$ is the solution of

$$n^{-1} \sum \theta_1 \left(\frac{y_t - x'_t \beta}{\sigma} \right) = \frac{1}{2} \sup(\theta_1(u) | u \in \mathbb{R})$$

where $\theta_1(\cdot)$ is some bounded function. The S estimator of β is the value which minimizes $\sigma_S(\beta)$. It has been shown that the S estimator is a HBP estimator compared with Huber's M estimator. The MM estimator is defined as follows: let σ_{MM} be $\sigma_S(\beta)$ evaluated at the S estimator of β . The MM estimator of β is the one which minimizes

$$\sum \theta_2 \left(\frac{y_t - x'_t \beta}{\sigma_{MM}} \right)$$

with respect to β . The function $\theta_1(\cdot)$ and $\theta_2(\cdot)$ have to satisfy certain conditions stated in Yohai (1987). In Lucas (1995a), for his purpose of estimating the AR(1) model, he used the following functions

$$\theta_i(u) = \begin{cases} (3c_i^4 u^2 - 3c_i^2 u^4 + u^6)/6 & \text{for } |u| \leq c_i \\ c_i^6/6 & \text{for } |u| > c_i \end{cases} \quad i = 1, 2$$

with $c_1 = 1.547$ and $c_2 = 4.685$. Note that the derivative of $\theta_i(u)$ gives the bisquare ψ function with different tuning constant.

The preceding discussion is concerned with estimation problems. The

effects of outliers on forecasts in time series models have been discussed in Ledolter (1991). Turning to the robust unit root tests, the tests that have been proposed are:

- (i) Lucas (1995a): based on the MM estimator proposed by Yohai (1987),
- (ii) Lucas (1995b): based on M estimators,
- (iii) Herce (1994, 1996): based on LAD estimators,
- (iv) Rothenberg and Stock (1997): based on nonnormal likelihoods,
- (v) Hoek *et al.* (1995): based on student t -likelihood rather than Gaussian likelihood (MLT estimation),
- (vi) Hasan and Koenker (1997): based on ranks.

We shall discuss these in turn.

Lucas (1995a)

In this paper Lucas develops outlier-robust unit root tests, using the MM estimator proposed by Yohai (1987). Lucas first established the asymptotic distribution for the t -test statistic, t_ρ based on the M estimator of ρ in the AR(1) model

$$y_t = \rho y_{t-1} + e_t$$

In the case of the AR(p) model

$$y_t = \rho_1 y_{t-1} + \rho_2 \Delta y_{t-2} + \cdots + \rho_p \Delta y_{t-p+1} + e_t$$

an M estimator $\hat{\rho}$ of ρ is the solution of

$$\sum \psi \left(\frac{y_t - x_t' \hat{\rho}}{\sigma} \right) x_t = 0$$

where $\rho = (\rho_1, \dots, \rho_p)'$ and $x_t = (y_{t-1}, \Delta y_{t-1}, \dots, \Delta y_{t-p+1})'$. The covariance matrix of $(\hat{\rho} - \rho)$ can be estimated by $\sigma^2 A^{-1} C A^{-1}$ with

$$A = \sum \psi' \left(\frac{y_t - x_t' \hat{\rho}}{\sigma} \right) x_t x_t'$$

$$C = \sum \psi^2 \left(\frac{y_t - x_t' \hat{\rho}}{\sigma} \right) x_t x_t'$$

and in this case the DF t -test statistic based on M estimator is

$$t_{\rho_1} = \frac{\hat{\rho}_1 - 1}{S_\rho}$$

where S_ρ is the square root of the (1,1) element of $\sigma^2 A^{-1} C A^{-1}$. Lucas proved that both t_ρ and t_{ρ_1} have the same limiting distribution given by

$$\frac{\int B dS}{[\int B^2 dt]^{1/2}} \quad (14.7)$$

where B and S are generated by the following partial sums

$$B_T(s) = T^{-1/2} \sum_{t=1}^{[sT]} e_t$$

$$S_T(s) = T^{-1/2} \sum_{t=1}^{[sT]} \psi\left(\frac{e_t}{\sigma}\right)$$

He showed that t_ρ and t_{ρ_1} based on the MM estimator instead of the M estimator have the limiting distribution (14.7). The critical values of (14.7) are computed via simulations. In his table 1, Lucas provides critical values for t -tests based on MM estimates for samples of 50, 100, and 200. He compares the performance of the DF t -test based on the OLS estimator with the one based on the MM estimator and finds that the DF t -tests based on the MM estimator are much more robust to the outliers he considered.

Lucas (1995b)

Lucas (1995b) considers unit root tests based on M estimators and develops the asymptotic theory for these estimators. The behavior of the M estimator in nearly stationary models is treated in Cox and Liatsas (1991). Lucas uses these results to derive the asymptotic distributions for his M estimators. Lucas (1995a) develops unit root tests based on the high breakpoint estimators suitable for a large number of outliers. The M estimators, by contrast have a low break down point, but Lucas (1995b) considers them because they are much easier to compute and are useful for fat-tail error distributions and provide some protection against outliers.

To apply the unit root tests based on the M estimators, we need the variance matrix of the M estimators. Lucas uses the standard one in Hampel *et al.* (1986, p. 316) which is a heteroskedasticity-consistent covariance matrix. For comparison he uses this type of standard error for both the M estimator and the OLS estimator. Thus the unit root tests based on the OLS estimator are different from the PP tests. But

Lucas shows that the difference between the two vanishes in the limit, although it is important in finite samples.

From simulation experiments Lucas provides some evidence that unit root tests based on M estimators are more powerful than those based on OLS estimators, if the errors come from fat-tailed distributions, although they are less powerful if the errors are normally distributed.

Herce (1994, 1996)

Herce (1996) derives the asymptotic distributions of the *least absolute deviation* (LAD) estimator of the autoregressive parameter under the unit root null, when errors have finite variance, and suggests LAD-based unit root tests. He shows that the asymptotic distributions of the LAD estimators depend on nuisance parameters and suggests corrections to the test statistics similar to those used in the PP tests for OLS-based unit root tests. He also provides extensive simulation evidence to show that the LAD-based tests are more powerful than the PP tests in the case of heavy tailed distributions of errors (which is the case, for instance, of many financial time series) but their performance is poor if the errors are normally distributed.

The results of Herce (1996) are for the case of finite error variances. Herce (1994) suggests a unified unit root testing framework, allowing for both finite and infinite variance errors.

Rothenberg and Stock (1997)

This is an extension of the likelihood-based tests in Elliott *et al.* (1996) to the case of nonnormal errors. They assume that the log density possesses the first- and second-derivatives. Rothenberg and Stock develop the asymptotic theory of the test statistics following the lines of Cox and Liatas (1991) and Lucas (1995a, 1995b). Like the tests by Lucas (1995b) based on M estimators and by Herce (1996) based on LAD estimators, these tests are designed for fat-tailed distributions commonly encountered in financial time series. A comparison of these three tests would, therefore, be useful.

Hoek *et al.* (1995)

Hoek *et al.* show that in both the classical and Bayesian frameworks, the presence of additive outliers leads to a biased inference toward stationarity (the overrejection of the unit root null). They suggest using an independent student *t*-distribution instead of a Gaussian distribution for an inference on unit roots. This yields maximum like-

likelihood and posterior results that are less sensitive to the presence of outliers. The use of the t -distribution for robust inference has been suggested earlier by Lange *et al.* (1989). The t -distribution is also amenable to Bayesian analysis. They argue that the MM estimator can cope with a large number of outliers, but the MLT (ML estimator based on the t -distribution) is robust only to a few outlying observations.

Hasan and Koenker (1997)

Rank-based methods play an important role in nonparametric inference but they have not been used much in econometrics. Hasan and Koenker (1997) apply rank-based methods to the problem of unit root testing. The paper also gives extensive references to the literature on rank-based methods in statistics. Hasan and Koenker develop rank-based unit root tests exploiting the formal duality between rank statistics and quantile regression. We shall omit the details, which are complicated and can be found in their paper, but note that they found that the rank tests they develop have lower than nominal sizes. They, therefore, examine size-corrected power, which is comparable to that of the ADF test in the case of normal innovations, and substantially better than that of the ADF test for t - and Cauchy errors. The rank tests have been suggested to tackle nonnormal distributions of the errors and thus are comparable to the other tests discussed here. There is as yet no comparative study of the different tests here to give advice to practitioners.

14.6 Robust estimation of cointegrating regressions

In the preceding section we discussed some robust unit root tests. We shall discuss briefly robust estimation methods for single cointegrating regression. We shall discuss the FM-LAD and FM-M estimation methods due to Phillips (1996) and the extension by Hoek *et al.* (1995), using t -distributed errors, and the cointegrating regressions by Franses and Lucas (1995).

In the case of nonstationary time series, Phillips (1996) suggests two robust methods of estimation: the fully modified least absolute deviation (FM-LAD) estimation and fully modified M (FM-M) estimation. These are extensions of the FM-OLS estimator of Phillips and Hansen suitable for heavy tailed error distributions.

Consider the cointegrated system

$$\begin{aligned}y_t &= \beta' x_t + u_{1t} \\ \Delta x_t &= u_{2t}\end{aligned}$$

where $u_t = (u_{1t}, u_{2t})$ is a vector of stationary errors. x_t and y_t are $I(1)$ and cointegrated since u_{1t} is $I(0)$. The least squares estimator of β is superconsistent but its asymptotic distribution depends on nuisance parameters arising from the endogeneity and serial correlation in the errors u_t . The FM-OLS estimator starts with $\hat{\beta}_{OLS}$ and applies semi-parametric corrections to it, to take care of the two problems of endogeneity and serial correlation.

The FM-LAD estimator starts with $\hat{\beta}_{LAD}$ and then applies semi-parametric corrections to it in the same fashion as with FM-OLS estimator. Similarly, the FM-M estimator starts with $\hat{\beta}_M$. The semi-parametric corrections are for the problems of endogeneity and serial correlation mentioned above. Thus, the FM-LAD and FM-M estimators are designed to combine the features of estimators for nonstationary regression like FM-OLS with the characteristics of LAD and M estimators that are robust to the presence of outliers.

Phillips develops the asymptotic theory necessary for these estimators which we shall not discuss here. He also presents some simulation results and outlines the possible extensions to multivariate regressions or subsystem cointegration. Finally, an empirical illustration of the practical usefulness of the FM-LAD procedure is presented. The example refers to the efficiency of the Australian foreign exchange market (Australian dollar in terms of the US dollar), using the daily exchange rate data over the period January 1984–April 1991 and a forward contract of three months (a total of 1,830 observations). The equation estimated is

$$s_{t+k} = \alpha + \beta f_{t,k} + u_{t+k}$$

where s_{t+k} is the log of the spot exchange rate and $f_{t,k}$ is the log of the forward exchange rate for a k -period ahead contract delivery. The forward rate unbiasedness hypothesis states that $\beta = 1$. The FM-OLS and FM-LAD estimators of β (with standard errors in parentheses) are

$$\text{FM-LAD : } 0.700(0.040)$$

$$\text{FM-OLS : } 0.883(0.092)$$

The nonrobust estimates are biased in favor of the unbiasedness hypothesis, whereas the robust estimates do not support it.

Phillips, McFarland, and McMahon (1996) also investigate the unbiasedness hypothesis based on daily data on spot and one-month forward exchange rates over the period May 1922–May 1925. The currencies considered are the Belgian Franc, French Franc, Italian Lira, and the US dollar, all measured in terms of pounds sterling. The estimates of β (with standard errors in parentheses) are:

Belgium	FM-LAD :	0.880(0.040)
	FM-OLS :	0.952(0.03)
France	FM-LAD :	0.863(0.046)
	FM-OLS :	0.942(0.032)
Italy	FM-LAD :	0.863(0.055)
	FM-OLS :	0.950(0.043)

The results for US did not change the inference and hence are not reported here. But for the other currencies, the robust method does not support the unbiasedness hypothesis, whereas the OLS method is biased in favor of the hypothesis. More details and tests of market efficiency can be found in Phillips *et al.* The major conclusion of these studies is that the use of robust methods makes an important difference to the conclusions. The papers also provide evidence to show that the distributions are fat tailed.

Franses and Lucas (1995) discuss another method for robust estimation of cointegrating regressions. Their method which follows the suggestion of Lange *et al.* (1989) consists of using t -distribution errors rather than Gaussian errors and using the Johansen maximum likelihood procedure. They argue, based on the results presented in Hoek *et al.* (1995), that maximum likelihood estimation of the model with student t -distributed errors (MLT) guards against the presence of outliers, if there are not too many of them. They present an empirical example involving interest rate spreads in the Netherlands where the maximum likelihood method based on Gaussian likelihood and student t -likelihood led to different conclusions. Franses and Lucas also argue in favor of using the weights generated by the use of robust methods to identify outliers and structural breaks.

In the preceding discussion, it was assumed that nonstationarity was caused by an autoregressive unit root in the process. Nonstationarity can also arise from fractional roots (see chapter 9). Beran (1994, chapter 7) discusses robust estimation methods. Following Martin and Yohai (1986), he develops bounded influence estimation for long-memory models.

14.7 Outliers and seasonal unit roots

Franses and Vogelsang (1995) extend the HEGY test for the seasonal unit root test (see chapter 12 for the HEGY test) to the case of AO and IO models. They consider the problems of testing for seasonal unit roots in the presence of a single break in each season in one particular year. They consider both the cases where the date of break is known and unknown. They consider the additive outlier (AO) model, which treats the shifts as immediate and the innovative outlier (IO) model which treats the shifts as gradual. They derive the limiting distributions of test statistics for seasonal unit roots, present asymptotic critical values, and also investigate how well the asymptotic distributions approximate finite sample distributions for a sample size of 20 quarterly observations. The proofs of asymptotic distributions are presented in appendix A and the table of critical values in appendix B of their paper.

Franses and Vogelsang also analyze quarterly industrial production for the US based on data for 1960:I–1991:IV (128 quarterly observations). When they do not incorporate mean shifts, they find evidence of a seasonal unit root at the bi-annual frequency. However, when they test for seasonal unit roots in the AO or IO model, the evidence for a seasonal unit root disappears.

14.8 Conclusions

The chapter considers several problems relating to outliers in time series models, and robust estimation methods in time series models.

First the different types of outliers are defined: additive outliers (AO), innovation outliers (IO), level changes (LC), and variance changes (VC). Next the effects of these different types of outliers on unit root tests have been discussed. For this an important distinction has to be made between outliers in levels versus outliers in differences.

Given that the unit root tests are sensitive to outliers, different outlier-robust unit root tests have been discussed. Some of these robust tests have been suggested as robust to distributional assumptions. There is as yet no comparative study of these different robust unit root tests. Finally, two robust cointegration methods: FM-LAD and FM-M have been discussed. The area of robust unit root tests and robust cointegration methods is new and will no doubt be an active area of research in the future.

What are the new insights obtained by these outlier models? The

answer depends on whether the outliers detected agree or disagree with the breakpoints identified by extraneous information and other models incorporating structural changes, and, if they disagree, what new information is provided by the outliers. There is some evidence on this in the paper by Balke and Fomby (1991a) but the evidence is not conclusive. A major problem with outlier analysis that has yet not been satisfactorily solved is the *masking problem*, that the initial model estimated with the outliers might mask the true outliers. In the usual regression models this is analyzed by the deletion methods. An extension of these methods to time series models within the context of a *structural time series model* can be found in Atkinson *et al.* (1997) who apply it to the analysis of gas sales data with quarterly observation for 1960–1986.

Assuming that the masking problem is not serious or is approximately solved, the outlier analysis can give us answers to questions whether some shocks have temporary or permanent effects (innovative outliers and level changes). The outlier analysis can be a useful complement to the regime switching models discussed in the next chapter.

We have discussed removing outliers as a solution to the outlier problem. This has some consequence on the power of unit root tests. Large shocks (which may or may not be outliers) contain a great deal of information as to whether or not there is a unit root. If the series mean reverts quickly, a conclusion of stationarity is likely. If not, then a unit root will not be rejected. Removing this observation as an outlier thus has important power implications.

How does one use the outlier models for prediction purposes. There is not much discussion or empirical evidence on this. McCulloch and Tsay (1993) suggest tagging on a probit model to the outlier model. Though their analysis is in the Bayesian framework using Gibbs sampling, the idea can be equally implemented (if the necessary data are available) in the usual framework. Suppose that you identify a number of outliers. If we have a set of explanatory variables relating to their occurrence, then we can use this set of variables to predict when some future outliers would occur. In principle, this can work (and McCulloch and Tsay show through a Monte Carlo study that it does) but the only question is how to implement it in practice.

References

- Abraham, B. and A. Chung (1989), "Outlier Detection and Time Series Modeling," *Technometrics*, 31, 241–248.

- Abraham, B. and N. Yatawara (1988), "A Score Test for Detection of Time Series Outliers," *Journal of Time Series Analysis*, 9, 109–119.
- Atkinson, A.C., S.J. Koopman and N. Shephard (1997), "Detecting Shocks : Outliers and Breaks in Time Series," *Journal of Econometrics*, 80, 387–422.
- Balke, N.S.(1993), "Detecting Level Shifts in Time Series," *Journal of Business and Economic Statistics*, 11, 81–92.
- Balke, N.S. and T.B. Fomby (1991a), "Shifting Trends, Segmented Trends, and Infrequent Permanent Shocks," *Journal of Monetary Economics*, 28, 61–85.
- (1991b), "Infrequent Permanent Shocks and the Finite-Sample Performance of Unit Root Tests," *Economics Letters*, 36, 269–273.
- (1994), "Large Shocks, Small Shocks, and Economic Fluctuations: Outliers in Macroeconomic Time Series," *Journal of Applied Econometrics*, 9, 181–200.
- Beran J. (1994), *Statistics for Long Memory Processes*, Chapman and Hall, New York.
- Bruce, A.G. and R.D. Martin (1989), "Leave K-Out Diagnostics for Time Series (with Discussion)," *Journal of Royal Statistical Society, Series B*, 51, 363–424.
- Chen, C. and L.M. Liu(1993), "Joint Estimation of Model Parameters and Outlier Effects in Time Series," *Journal of the American Statistical Association*, 88, 284–297.
- Cook, R.D. (1987), "Influence Assessment," *Journal of Applied Statistics*, 14, 117–131.
- Cook, R.D. and S. Weisberg (1982), *Residuals and Influence in Regression*, Chapman and Hall, New york.
- Cox, D.D. and I. Liatas (1991), "Maximum Likelihood Type Estimation for Nearly Nonstationary Autoregressive Time Series," *Annals of Statistics*, 19, 1109–1128.
- Donald, S.G. and G.S. Maddala (1993), "Identifying Outliers and Influential Observations in Econometric Models," in G. S. Maddala, C. R. Rao, and H. D. Vinod (eds.), *Handbook of Statistics*, vol. 11, North-Holland, Amsterdam, 663–701.
- Durbin, J. and M. Cordero (1994), "Handling Structural Shifts, Outliers and Heavy-Tailed Distributions in State Space Time Series Models," Manuscript, London School of Economics.
- Elliott, G., T.J. Rothenberg, and J.H. Stock (1996), "Efficient Tests for an Autoregressive Unit Root," *Econometrica*, 64, 813–836.

- Fiebig, D.G. and E. Maasoumi (1990), "Specification Analysis in Dynamic Models," *Journal of Business and Economic Statistics*, 8, 443–451.
- Fox, A.J. (1972), "Outliers in Time Series," *Journal of the Royal Statistical Society, Series B*, 34, 350–363.
- Franses, P.H. and N. Haldrup (1994), "The Effects of Additive Outliers on Tests for Unit Roots and Cointegration," *Journal of Business and Economic Statistics*, 12, 471–478.
- Franses, P.H. and A. Lucas (1995), "Outliers Robust Cointegration Analysis," Report 9529A, Erasmus University, Rotterdam.
- Franses, P.H. and T.J. Vogelsang (1995), "Testing for Seasonal Unit Roots in the Presence of Changing Seasonal Means," Report no. 9532A, Econometric Institute, Erasmus University, Rotterdam.
- Hampel, H.R., E.M. Ronchetti, P.J. Rousseeuw, and W.A. Stahel (1986), *Robust Statistics: The Approach Based on Influence Functions*, Wiley, New York.
- Hasan, M.N. and R.W. Koenker (1997), "Robust Rank Tests of the Unit Root Hypothesis," *Econometrica*, 65, 133–161.
- Hendry, D. F. and A.J. Neale (1990), "The Impact of Structural Breaks on Unit Root Tests," in P. Hackl and A. Westlung (ed.), *Economic Structural Change: Analysis and Forecasting*, Springer Verlag, Berlin.
- Herce, M.A. (1994), "A Unified Approach to Unit Root Testing Based on LAD Estimation," Manuscript, University of North Carolina at Chapel Hill.
- (1996), "Asymptotic Theory of LAD Estimation in a Unit Root Process with Finite Variance Errors," *Econometric Theory*, 12, 129–153.
- Hoek, H., A. Lucas and H.K. van Dijk (1995), "Classical and Bayesian Aspects of Robust Unit Root Inference," *Journal of Econometrics*, 69, 27–59.
- Huber, P.J. (1964), "Robust Estimation of a Local Parameter," *Annals of Mathematical Statistics*, 35, 73–101.
- (1973), "Robust Regression: Asymptotics, Conjectures, and Monte Carlo," *Annals of Statistics*, 14, 781–818.
- Lange, K.L., R.J.A. Little, and J.M.G. Taylor (1989), "Robust Statistical Modeling Using the *t*-Distribution," *Journal of American Statistical Association*, 84, 881–896.
- Ledolter, J. (1990), "Outlier Diagnostics in Time Series Analysis," *Journal of Time Series Analysis*, 11, 317–324.

- (1991), "Outliers in Time-Series Analysis: Some Comments on Their Impact and Their Detection," in W. Stahel and S. Weisberg (eds.), *Directions in Robust Statistics and Diagnostics*, vol. I, Springer Verlag, New York, 159–165.
- Ljung, G.M. (1993), "On Outlier Detection in Time Series," *Journal of Royal Statistical Society, Series B*, 55, 559–567.
- Lucas, A. (1995a), "An Outlier Robust Unit Root Test with An Application to the Extended Nelson–Plosser Data," *Journal of Econometrics*, 66, 153–173.
- (1995b), "Unit Root Tests Based on M-Estimators," *Econometric Theory*, 11, 331–346.
- Maddala, G.S. and Y. Yin (1996), "Outliers, Unit Roots and Robust Estimation of Non-Stationary Time Series," in G. S. Maddala and C. R. Rao (eds.), *Handbook of Statistics*, vol. 15, Elsevier, Amsterdam, 237–266.
- Martin, R. D. and V. J. Yohai (1986), "Influence Functionals for Time Series," *Annals of Statistics*, 14, 781–818.
- McCulloch, R.E. and R.S. Tsay (1993), "Bayesian Inference and Prediction for Mean and Variance Shifts in Autoregressive Time Series," *Journal of the American Statistical Association*, 88, 968–978.
- Perron, P. (1989), "The Great Crash, the Oil Price Shock and the Unit Root Hypothesis," *Econometrica*, 57, 1361–1401.
- (1990), "Testing for a Unit Root in a Time Series with a Changing Mean," *Journal of Business and Economic Statistics*, 8, 153–162.
- Perron, P. and S. Ng (1996), "Useful Modifications to Some Unit Root Tests with Dependent Errors and Their Local Asymptotic Properties," *Review of Economic Studies*, 63, 435–465.
- Perron, P. and T.J. Vogelsang (1992a), "Nonstationarity and Level Shifts with an Application to Purchasing Power Parity," *Journal of Business and Economic Statistics*, 10, 301–320.
- (1992b), "Testing for a Unit Root in a Time Series with a Changing Mean: Corrections and Extensions," *Journal of Business and Economic Statistics*, 10, 467–470.
- (1993), "A Note on the Asymptotic Distributions of Unit Root Tests in the Additive Outlier Model with Breaks," *Revista De Econometric*, 8, 181–202.
- Phillips, P.C.B. (1996), "Robust Non-Stationary Regression," *Econometric Theory*, 11, 912–951.

- Phillips, P.C.B., J.W. McFarland, and P.C. McMahon (1996), "Robust Tests of Forward Market Efficiency with Empirical Evidence from the 1920s," *Journal of Applied Econometrics*, 11, 1-22.
- Rappoport, P. and L. Reichlin (1989), "Segmented Trends and Non-Stationary Time Series," *Economic Journal*, 99, 168,177.
- Rothenberg, T.J. and J.H. Stock (1997), "Inference in a Nearly Integrated Autoregressive Model with nonnormal Innovations," *Journal of Econometrics*, 80, 269-286.
- Rousseeuw, P.J. and V.J. Yohai (1984), "Robust Regression by means of S-Estimators," in J. Franke, W. Hardle, and R. D. Martin (eds.), *Robust and Nonlinear Time Series*, Springer Verlag, New York, 256-272.
- Tsay, R.S. (1988), "Outliers, Level Shifts, and Variance Changes in Time Series," *Journal of Forecasting*, 7, 1-20.
- Vogelsang, T.J. (1994), "On Testing for a Unit Root in the Presence of Additive Outliers," CAE Working paper No. 94-30, Cornell University.
- Vogelsang, T.J. and P. Perron (1994), "Additional Tests for a Unit Root Allowing for a Break in the Trend Function at an Unknown Time," CAE Working paper No. 94-13, Cornell University.
- West, M. (1986), "Bayesian Model Monitoring," *Journal of the Royal Statistical Society, Series B*, 48, 70-78.
- West, M. and P. J. Harrison (1986), "Monitoring and Adaptation in Bayesian Forecasting Models," *Journal of the American Statistical Association*, 81, 741-750.
- Yin, Y. and G.S. Maddala (1996), "Effects of Variance Change on Unit Root Tests," Working Paper, Department of Economics, The Ohio State University.
- (1997), "The Effects of Different Types of Outliers on Unit Root Tests," in T. Fomby and R. C. Hill (eds.), *Advance in Econometrics*, vol. 13, JAI Press, Greenwich, Conn.
- Yohai, V.J. (1987), "High Breakdown-Point and High Efficiency Robust Estimates for Regression," *Annals of Statistics*, 15, 642-656.

15

Regime switching models and structural time series models

In the previous two chapters we considered some methods of an analysis of structural change. In this chapter we shall discuss regime switching models of which the widely used Markov switching model is a special case. We also discuss several extensions of the Markov switching model. This model implies sudden switches. More reasonable models are gradual switching models – Harvey’s structural time series model being a special case. These models are also known as *state-space models*.

We shall discuss the main ideas and the problems these models try to solve. We omit the detailed mathematics, which is quite involved, because we do not see any point in reproducing it from the papers cited. Those interested in the details can refer to the particular paper relevant for their work, on which they can decide from our review. We feel it is important to understand the merits and limitations of the different procedures before getting bogged down in the mathematical detail.

The estimation problems with the Markov switching models are not as complex as those of testing. An illustration of the problems of testing is the paper by Garcia (1997) which has been around for many years in its several unpublished versions.

15.1 The switching regression model

A commonly used model for the analysis of structural change is the switching regression model. This model was first introduced into the econometric literature by Quandt (1958), though it has a long history (see Lindgren, 1978). The simplest two-regime model is given by

$$\begin{aligned} Y_t &= X_{1t}\beta_1 + u_{1t} && \text{in regime 1} \\ Y_t &= X_{2t}\beta_1 + u_{2t} && \text{in regime 2} \end{aligned} \tag{15.1}$$

X_{1t} and X_{2t} are explanatory variables. The variable Y_t is generated from regime 1 or regime 2, but not both. Suppose we define the indicator variable I_t by

$$\begin{aligned} I_t &= 1 && \text{if } Y_t \text{ comes from regime 1} \\ I_t &= 0 && \text{if } Y_t \text{ comes from regime 2} \end{aligned}$$

and I_t is observed. In this case, equations (15.1) can be estimated by OLS using the observations in the respective regimes. If the indicator I_t is not observed but its determinates X_{3t} are, we can define another equation relating a latent variable Z_t to X_{3t}

$$Z_t = X_{3t}\beta_3 + u_{3t} \quad (15.2)$$

and define

$$I_t = \begin{cases} 1 & \text{if } Z_t > 0 \\ 0 & \text{if } Z_t \leq 0 \end{cases}$$

In this case we have probabilistic sample separation. Let the covariance matrix u_1, u_2 , and u_3 be denoted by

$$\Sigma = \begin{bmatrix} \sigma_{11} & \sigma_{12} & \sigma_{13} \\ \sigma_{12} & \sigma_{22} & \sigma_{23} \\ \sigma_{13} & \sigma_{23} & 1 \end{bmatrix}$$

Since Z_t is observed only as a dichotomous variable, we assume $\text{var}(Z_t) = 1$. If $\sigma_{13} = \sigma_{23} = 0$, we have a switching regression model with exogenous switching. Otherwise we have endogenous switching (see Maddala, 1983, p. 284). With endogenous switching, even if the sample separation is known, equation (15.1) cannot be estimated by OLS.

The model given by equations (15.1) and (15.2) results in a mixture distribution. For these problems, the maximum likelihood method fails because the likelihood function explodes for some values of the error variances. This problem has been noted by Quandt (see Maddala, 1983, p. 299). For this reason, Goldfeld and Quandt (1973) suggested a Markov switching regression (MSR) model.

15.2 The Markov switching regression model

In the Markov switching regression (MSR) model, the probabilities of switching from one regime to the other in the next period are assumed to be constant. These are given by the following transition matrix

		Regime at time t	
		$I_t = 1$	$I_t = 0$
Regime at time $t - 1$	$I_{t-1} = 1$	p_{11}	p_{10}
	$I_{t-1} = 0$	p_{01}	p_{00}

Another modification of the switching regression model is that by Lee and Porter (1984) who consider the availability of a dichotomous indicator W_t that provides sample separation but this indicator is imperfect. They postulate a transition probability matrix

		$W_t = 1$	$W_t = 0$
I_t	$I_t = 1$	p_{11}	p_{10}
	$I_t = 0$	p_{01}	p_{00}

If $p_{11} = p_{01}$, the indicator does not provide any information on sample separation. If $p_{11} = 1$ and $p_{01} = 0$, the indicator W_t provides perfect sample separation. In the case $p_{11} \neq 1$ and $p_{01} \neq p_{11} (> p_{01})$ we have imperfect sample separation. Lee and Porter examine the transportation prices charged by the Joint Executive Committee railroad cartel from 1880–1886 using the price equation

$$\log P_t = \beta_0 + X_t\beta_1 + \beta_2 I_t + u_t \quad (15.3)$$

where $I_t = 0$ if there is a price war and $I_t = 1$ otherwise. Thus, there is a regime switch. They used the indicator w_t obtained from the *Railroad Review* which reported in each period whether a price war was going or not. One reason Lee and Porter suspected measurement error in W_t is that it conflicted with an index of cartel adherence constructed by MacAvoy.

Cosslett and Lee (1985) extend this model by Lee and Porter to the case where the error term u_t in (15.3) is serially correlated. In addition, they assume that the true indicator variable I_t follows a Markov chain with transition probability matrix

$$\Lambda = [\lambda_{ij}] \quad \text{and} \quad \lambda_{ij} = \text{Prob}(I_t = j | I_{t-1} = i), \quad i, j = 0, 1$$

Cosslett and Lee derive an algorithm for computing the likelihood and its derivatives for this model by means of a recurrence relation. In the appendix we review how this algorithm can also be used for the ML estimation of the MSR model. They also argue that the original method proposed by Goldfeld and Quandt (1973) yields consistent but inefficient estimates. This is because it does not take into account an

essential feature of a Markov chain, namely that the state probabilities at time t depend on the realized state at time $(t - 1)$. Cosslett and Lee present the correct log likelihood function for this model.

15.3 The Hamilton model

The MSR model considered by Hamilton (1989, 1990, 1992) is an extension of the idea of Quandt (1958), Goldfeld and Quandt (1973), and Cosslett and Lee (1985) to an autoregression. Related models have been discussed by Sclove (1983), though he did not discuss maximum likelihood (ML) estimation of the parameters.

Suppose that the time series $\tilde{y}_t$ is composed of two parts

$$\tilde{y}_t = n_t + \tilde{z}_t \quad (15.4)$$

where

$$n_t = n_{t-1} + \alpha_0 + \alpha_1 s_t$$

and

$$\tilde{z}_t - \tilde{z}_{t-1} = \phi(\tilde{z}_{t-1} - \tilde{z}_{t-2}) + \varepsilon_t$$

with $\varepsilon_t \sim iin(0, \sigma^2)$ and $|\phi| < 1$. Suppose that there exists a finite set of K states (regimes) and that each $\tilde{y}_t$ is associated with an unobserved state s_t . We also assume that the process $\{s_t\}$ is a Markov chain with the stationary transition probability matrix $P = (p_{ij})$ where

$$p_{ij} = \text{Prob}(s_t = j | s_{t-1} = i), \quad i = 0, 1, \quad j = 0, 1$$

with $p_{10} = 1 - p_{11}$ and $p_{00} = 1 - p_{01}$. This model is referred to as the MSR model or Hamilton's model.

The MSR model implies that the level of the series, $\tilde{y}_t$, is an ARIMA (1,1,1) and its representation conditional on s_t is an ARIMA(1,1,0) with shifts in the mean. Define $y_t = \Delta \tilde{y}_t$ and $z_t = \Delta \tilde{z}_t$. Differencing (15.4) gives

$$\begin{aligned} y_t &= \alpha_0 + \alpha_1 s_t + z_t \\ &= \mu_{s_t} + z_t \end{aligned}$$

where z_t follows a Gaussian AR(p) process. Suppose that z_t is AR(1), then the MSR can be written as

$$\begin{aligned} y_t - \mu_{s_t} &= z_t = \phi z_{t-1} + \varepsilon_t \\ &= \phi(y_{t-1} - \mu_{s_{t-1}}) + \varepsilon_t \end{aligned}$$

The dynamic structure of MSR in the framework of AR models implies that the effect of a change in regime is immediate, and is not tied down by the dynamic consequences of ε_t .

In the two-state MSR, there are 2^n state configurations for a sample of size n . To overcome this problem, Cosslett and Lee (1985) suggest a recursive algorithm for the ML estimation of the model. Hamilton (1989) uses an adaptation of probability smoothers and the EM algorithm. Earlier, Baum *et al.* (1970) derived the EM algorithm for ML estimation of probabilistic functions of Markov chains. Suppose that to each $\lambda \in \Lambda$, we have a smooth assignment $\lambda \rightarrow [p(\lambda), f(\lambda)]$. For each fixed y_t , $L(\lambda)$ is a smooth function of λ . The likelihood for the MSR model can be rewritten as

$$\begin{aligned} L(\lambda) &= \sum_{s \in S} p(s, \lambda) \\ &= \sum_{s \in S} [p(s, \lambda) \prod_{t=1}^T p(s, \lambda) f(\lambda, y_t)] \end{aligned}$$

The EM algorithm is based on the fact that the derivatives of the log-likelihood have a representation as the difference of an unconditional and a conditional expectation of the sufficient statistics. The expectation step is to estimate the complete-data sufficient statistics (in our case observed y and unobserved s) by computing the auxiliary function

$$\begin{aligned} Q(\lambda, \lambda') &= E(\log p(s, \lambda) | y, \lambda) \\ &= \int_s p(s, \lambda) \log p(s, \lambda') d\mu(s) \end{aligned}$$

The maximization step is to find λ' which maximizes $Q(\lambda, \lambda')$. With a given normal density function $f_i(m_i, \sigma_i)$, $L(\lambda)$ can be maximized by maximizing

$$\begin{aligned} Q(m, \sigma, m', \sigma') &= \sum_s \nu_s \sum_i \log f_i(m'_i, \sigma'_i) \\ &= \sum_{t=1}^T \gamma_t \log f_t \end{aligned}$$

Hamilton(1989) shows how to derive $\gamma_t = p(s_t | y_t, \hat{\theta})$ in the MSR model and named it as *smooth inferences* or *smooth probabilities*. The expectation step of the EM algorithm requires the estimation of the components of $\gamma_t = p(s_t | y_t, \hat{\theta})$ given the observed y_t and the current fitted parameters $\hat{\theta}$. The maximization step is then equivalent to the complete-data

maximization with weights given by the K -estimated components of s_i . For the MSR model

$$y_t = X_t' \beta_i + e_t, \quad i = 1, \dots, k$$

Hamilton shows that the M-step is an OLS regression of $y_t \sqrt{\gamma_t}$ on $X_t \sqrt{\gamma_t}$ with smoothed probabilities. To obtain the ML estimates, with an initial guess on parameters, the E- and M-step needs to be iterated until the change in parameter values between iterations is less than some target convergence criterion, say 10^{-8} .

Kim (1993) suggests an alternative two-stage estimation method based on the Maximum A Posteriori (MAP) decision rule. The MAP decision rule originated from the decision-directed learning procedure which is, with a quasi-Bayes learning procedure (Makov and Smith, 1977), one of the approximations to the formal Bayes learning procedures (Titterton, Smith, and Makov, 1985). He shows that the sequence of $\{\hat{s}_t\}$ can be obtained by imposing the MAP decision rule and it is a solution path of Bellman's dynamic programming algorithm. Then the sequence of $\{\hat{s}_t\}$ can be used in place of the state (dummy) variables s_t in MSR models. A Monte Carlo experiment performed by Kim shows that the MAP procedure outperforms Hamilton's method in terms of the probabilities of error events (misclassification of states). And the bound of the probabilities of error events is tighter than in Hamilton's method regardless of sample size. When he applied the MAP procedure for the same data set used by Hamilton (1989), Kim found that the dating of the US business cycles by the MAP procedure is closer to the dating by NBER than that given by Hamilton's method.

The Markov switching autoregressive model has been applied with success to a variety of economic and financial data. Hamilton (1988, 1989) uses the model to study the term structure of interest rates and to date the timing of recessions and booms with the US GNP data. Engel and Hamilton (1990) use it to explain the long swings in exchange rates. Pagan and Schwert (1990) use it to model conditional stock volatility in stock returns. Turner, Startz, and Nelson (1989) also use it to study risk-return relationships in the stock market. Cecchetti, Lam, and Mark (1990) use it to study regime switches in century-long data on consumption, output, and dividends. McCulloch and Tsay (1994b) even use it to make inference about trend and difference stationarity. They treat trend stationarity and difference stationarity as two competing models and allow each observation to switch from one model to the other, with the transition being governed by a Markov process. This method, they

argue, is superior to unit root tests based on the entire data, because it can identify periods during which trend stationarity is more appropriate than difference stationarity.

15.4 On the usefulness of the MSR model

The MSR model is a model that allows us to take into account multiple structural breaks in a given time series. It is also a model that allows us to explain nonlinearities in the data. One major limitation of the model is that it restricts us to the two-regime case. However, there have been extensions of the model to the case of several regimes, which we shall discuss in the next section.

The MSR model has been used for a variety of purposes and it has yielded some new insights in some cases, and not in some others. For instance, the model has been used by Hamilton (1989) to date the turning points in the US business cycle using the quarterly data for 1951:II–1984:IV. Hamilton shows that the turning points identified by his model agree closely with those identified by the NBER. No new insights are obtained except to confirm the NBER method of timing business cycles or to demonstrate the usefulness of the MSR model. Rabault (1992) studies the ability of the Hamilton model to reproduce the business cycles of six industrialized countries (UK, France, Germany, US, Canada, and Japan) and the OECD as a whole. He argues that only three countries (US, Germany, and Japan) show a satisfactory coincidence between the Markovian transition and cycle chronology. He also complains about the fragility of the ML estimation method because multiple local maxima were often found. Kim (1993) shows that the dates of the US business cycle identified using the same MSR model used by Hamilton can be estimated even more closely with those identified by the NBER, if the MAP estimation procedure is applied.

Engel and Hamilton (1990) found that the simple two-state MSR model in which the log of the exchange rate is a random walk with drift in each of the two states provided a good description of exchange rate behavior. However, Engel (1994) investigated the forecasting behavior of the MSR model. He fits the model for 18 bilateral exchange rates at quarterly and monthly frequencies. The 18 exchange rates were: US/Canada, US/France, US/Italy, US/Japan, US/Switzerland, US/UK, US/Germany, Japan/Canada, Japan/France, Japan/Italy, Japan/Switzerland, Japan/UK, Japan/Germany, UK/Canada, UK/France, UK/Italy, UK/Switzerland, and UK/Germany. The estimation period was 1973:3–1986:1 and the post-sample forecast period

was 1986:2–1991:1. The model was also estimated for the period 1973–1988 and forecasts were constructed for the remaining period. Engel's general conclusion was that the MSR model fits well with the sample, but it does not generate superior forecasts in terms of mean square errors (MSE) to a random walk or the forward rate. Engel, however, finds some evidence that the MSR model is superior to the others in forecasting the direction of change of the exchange rate. He speculates that with the Louvre accord of March 1987, perhaps the foreign exchange rates entered a third state, and the MSR model might perform better in a three-state model.

Lam (1990) extends the Hamilton model of GNP growth to a model with a general autoregressive component, in which the Markov component, n_t , is integrated but the Gaussian component, $\tilde{z}_t$, is not. He specified that $\tilde{z}_t$ follows the Gaussian AR(p) process, while $z_t = \Delta\tilde{z}_t$ is the Gaussian AR(p) process in Hamilton's model. Lam's framework, thus, could be viewed as the generalization of Perron's (1989) specification of a stationary process around an occasionally shifting linear trend. Define $\tilde{s}_t$ to represent the cumulative number of times that the event $s_\tau = 1$ has occurred during dates $\tau = 1, 2, \dots, t$; thus $\tilde{s}_t$ takes on an integer value in $(0, 1, \dots, t)$. Then, in the framework of an AR(1) model, Lam's model can be written as

$$\tilde{y}_t - w_t = \phi(\tilde{y}_{t-1} - w_{t-1}) + \varepsilon_t$$

where $w_t = n_0 + \mu_1\tilde{s}_t + (t - \tilde{s}_t)\mu_2$. Unlike the Hamilton model, Lam's model does not restrict one of the roots of the AR process to unity. His algorithm is consequently more complicated than that of Hamilton. For the same data used by Hamilton, Lam compares the forecasting performance of his model with that of the Hamilton model and the traditional ARIMA model by considering within-sample forecasts over different horizons of 1, 4, 5, 20, and 40 quarters. None of the models dominated the others over all horizons. Out-of-sample forecasting was not considered. Also, unlike the conclusion of Hamilton (1989), the dichotomous shift in states identified by Lam do not resemble the NBER dating of US business cycles very closely. Both Hamilton's and Lam's papers, however, argue that there were shifts in the trend of US GNP other than that caused by the major oil price shock of 1974–1975.

Albert and Chib (1993) and McCulloch and Tsay (1994b) present a Bayesian analysis of the MSR model using the Gibbs sampling approach. The Gibbs sampling method enables us to obtain marginal distributions of random variables numerically from the full cycle of conditional

distributions (see Gelfand and Smith, 1990 and Casella and George, 1992, and chapter 16 for its adaptation to the analysis of multiple structural changes by Kim and Maddala, 1992). The method is useful when the joint distribution is complicated but the conditional distributions are well defined, which is the case with the MSR model.

Albert and Chib (1993) claim that Hamilton's smooth inferences $\gamma_t = p(s_t|y_t, \hat{\theta})$ are based on the observed y_t and the current fitted parameters $\hat{\theta}$ and thus, the uncertainty about $\hat{\theta}$ is not incorporated in the inferences about the states. In the Bayesian approach this uncertainty is taken into account because it is the marginal distribution $p(s_t)$ that is considered. Albert and Chib present a Bayesian analysis of the Hamilton model and implement it using Gibbs sampling. However, when applied to the US GNP data, they found that the timing of booms and recessions is not different from those obtained by Hamilton.

McCulloch and Tsay (1994a) also use a Bayesian analysis with the Gibbs sampler, but they allow the dynamic pattern of US GNP growth to be different during expansions and contractions. They first estimate the Hamilton-type MSR model, where the dynamic pattern of US GNP growth is the same in both states, and get estimated probabilities close to those of Hamilton, but the posterior probabilities are in general smaller and smoother than those obtained by the Hamilton method. This is because the uncertainty in $\hat{\theta}$ is taken into account. They then expand the model to allow for different dynamics in the two states

$$\begin{aligned} y_t &= a_{01}y_{t-1} + a_{02}y_{t-2} + a_{03}y_{t-3} + a_{04}y_{t-4} + C_0 + e_{0t}, \quad \text{if } s_t = 0 \\ &= a_{11}y_{t-1} + a_{12}y_{t-2} + a_{13}y_{t-3} + a_{14}y_{t-4} + C_1 + e_{1t}, \quad \text{if } s_t = 1 \end{aligned}$$

They find that the classification of expansion and contraction becomes somewhat less clear when the dynamic pattern of GNP growth is made to be state dependent. There is, however, no evidence in either of these two Bayesian papers on the post-sample predictive ability of the MSR model relative to say the ARIMA model. Albert and Chib (1993) also apply the Gibbs sampler to the Markov switching model. However, they consider a model where only the level parameter and innovational variances differ from state to state. The autoregressive coefficients are the same in both the states. Thus, it is a special case of the model considered by McCulloch and Tsay (1994a).

McCulloch and Tsay (1994b) treat trend stationarity and difference stationarity as two competing models and cast them into the framework of a Markov switching model. Considering the monthly series of

industrial production index for the US from 1947:1 to 1992:1 (541 observations), they conclude that there might be a structural change around $t = 300$ (December 1971) and that the series was difference-stationary before that, but moved closer to trend stationarity after December 1971. It would be interesting to examine this same time series using the gradual switching models and see the switch date and the length of the adjustment period.

Garcia and Perron (1996) consider the time series behavior of US real interest rates from 1961 to 1986. They consider a three-state MSR model in levels rather than in first-differences as in the Hamilton model. They find that the real interest rate is essentially random around a mean that is different for the periods 1961–1973, 1973–1980, and 1980–1986. The variance of the random process is also different in the three periods, being higher in both the last two subperiods. Garcia and Perron compared the within-sample forecasting ability of the three-state MSR model to a simple random walk model and find the former better (though no out-of-sample forecasts were made). They argue that the earlier investigations that found the real interest rate to be a random walk arrived at the (wrong) conclusion because of ignoring structural breaks. Garcia and Perron also discuss three tests: Davies-test, Gallant-test, and the J -test to determine the number of states (1, 2, or 3) and find that the tests generally favored a three-state model. These tests are discussed further in Garcia (1997). One important difference in the economic interpretation of the structural break point is that the two-state model picks the second break at the end of 1979, suggesting that a change in monetary policy is its origin, whereas, the three-state model puts the second break in the middle of 1981, a date more in line with the federal deficit explanation.

Overall, apart from the paper by Engel (1994) there is not much evidence on out-of-sample forecasting ability of the MSR model. There is evidence on the ability of the MSR model in forecasting turning points and identifying trend breaks, though there is some skepticism on this issue in the papers by McCulloch and Tsay (1994a) and Rabault (1992). The paper by Garcia and Perron (1996) demonstrates the importance of tests for the number of states.

15.5 Extensions of the MSR model

In the preceding section we considered an example of a three-state MSR model by Garcia and Perron (1996), who concluded that the appar-

ent unit root in the real interest rate is an artifact due to structural changes. Another example of multiple states is the paper by Raymond and Rich (1992) who consider a four-state model and analyze the quarterly growth rates of the GNP deflator over the period 1948:I–1987:IV. They find that the time series is characterized by a low mean high variance regime over 1949–1954, a low mean, low variance regime over 1955–1967, and recurrent shifts between medium mean and high mean regimes over 1968–1983, before returning to the low mean, low variance regime in 1984. They also found that the apparent evidence of a unit root in the conditional mean and ARCH structure in the conditional variance are artifacts due to shifts in the mean and shifts in variance.

One major drawback of the models considered by Hamilton (1989) and Lam (1990) is that they are pure broken trend models with no explanatory variables. There are now several papers that include explanatory variables in the MSR model. Rich and Raymond (1992) generalize the Hamilton model of real GNP growth to include lagged growth rates of the real price of oil and analyze the importance of real oil prices on the GNP growth rate within the context of a two-state MSR model. They find that real GNP growth rates are significantly negatively correlated with real oil price increases but have no significant correlation with real oil price decreases. But the fluctuations in GNP cannot be fully accounted for by fluctuations in real oil prices. Thus, other variables need to be included in the model.

One other extension of the Hamilton model is by Durland and McCurdy (1992). They argue that extending the MSR model to multiple states increases the computations polynomially and that unrestricted higher-order Markov processes result in substantial loss in degrees of freedom. They, therefore, suggest a semi-Markov process (which could also be interpreted as a restricted higher-order Markov process). In a semi-Markov process the successive states are determined not only by the transition probabilities but also by the holding and waiting periods in each state. Durland and McCurdy consider a restricted version of the semi-Markov process, suggest an algorithm to estimate the model, and present evidence with the US GNP data. They find evidence of significant asymmetry between recessions and expansions and some duration dependence for the former but not the latter. They also suggest that Hamilton's first-order Markov models can be rejected when compared with the AR(4) model, whereas the semi-Markov model cannot be rejected against the AR(4) model. These preliminary results suggest that duration dependence is an important factor in MSR models.

Bekaert and Hodrick (1993) consider endogenous regime shifts in the MSR model. In their model the conditional means in each state are allowed to depend autoregressively on lagged values of the rate of depreciation and the forward premium. Kaminsky (1993) considers another generalization of the MSR model by Hamilton. She assumes that rational investors not only recognize the possibility of changes in the regime but also incorporate the information provided by the Fed announcements into their forecasts for the future. However, the Fed announcements are a noisy indicator. Thus the model is in the spirit of the papers by Lee and Porter (1984) and Cosslett and Lee (1985). Kaminsky also extends this model by allowing serial correlation in the measurement errors. During the period from March 1976 to December 1987 there were 274 announcements by the Fed classified as easy money policy (indicator = 1) and tight money policy (indicator = 0). Kaminsky reports results from out of sample forecasts with the MSR models compared with the random walk model and finds the latter superior. However, she argues that even if the MSR models does not give better out of sample forecasts, it still explains most of the variation in the exchange rate and gives new insights into the behavior of rational investor's response to changing fundamentals.

Lahiri and Wang (1994) applied the two-state MSR model to evaluate the Commerce Department's Composite Index of Leading Indicators (CLI) as a predictor of business cycle turning points. They found that the predictive performance of CLI is quite good. However, they found that imposing any degree of autoregression in the errors on the simple regime-shift model caused the MSR model to signal turning points inappropriately. Evans (1993) extended Hamilton's estimation methods to the MSR model with serial correlation in errors. He applied the two-state MSR model for quarterly US inflation and could not reject the presence of serial correlation of errors. Diebold *et al.* (1994) proposed a class of MSR models in which the transition probabilities vary with underlying (economic) fundamentals. They develop an EM algorithm for estimation of the model and illustrate it with a simulation example.

Lee (1991) and Filado (1994) include explanatory variables with transition probability depending on economic fundamentals. Lee (1991) uses real interest rate difference as the fundamental factor. Kim and Lee (1996) employ the magnitude of the deviation of the exchange rate from a monetary equilibrium value as the economic fundamental with which the transition probabilities vary. They find that the predictions from the TVTP (time-varying transition probability) Markov switching model, using this fundamental variable as an explanatory variable, fore-

casts exchange rates much better than FTP (fixed transition probability) models as used in Engel (1994). The TVTP Markov model can identify both the appreciation state and depreciation state better than the FTP Markov model. Finally, the TVTP model is also better at predicting the direction of the exchange rate than the FTP model.

Kim and Lee (1996) note that in the Markov switching model, the lack of observations on states can be regarded as an incomplete data problem. Hence they suggest and use the EM algorithm to estimate the TVTP model. Rabault (1992) also uses the EM algorithm. Kim and Lee, also derive the exchange rate from the monetary equilibrium value, the latter being determined by a time-varying coefficient cointegration regression as described in Kim and Yu (1996).

Finally Hamilton and Susmel (1994) extend the Markov switching model to ARCH models. The resulting SWARCH models have been used by Kim (1994) and found to be better than ARCH models in explaining risk premia.

Franses and Rapp (1996) investigate a different issue: do the usual seasonal adjustment procedures (like the census X-11) affect inference about the probabilities in the MSR model? They answer this question in the affirmative. The commonly used seasonal adjustment procedures (since they smooth the data) result in increasing the probabilities in the MSR model of staying in the same regime. Hence they suggest estimating MSR models using seasonally unadjusted data. (See chapter 12 for arguments in several other situations, where it is suggested that using seasonally adjusted data should be avoided.) They illustrate this problem through Monte Carlo analysis as well as two empirical examples concerning German unemployment and US industrial production. Several studies (e.g., Golodwin, 1993) study business cycles with an MSR model using seasonally adjusted data. Franses and Rapp argue that this can affect business cycle turning points.

15.6 Gradual regime switching models

In the previous sections we discussed regime switching models where the switch takes place at a certain point of time. Suppose there is a policy change at time t . In actual practice, the switch to the new regime need not take place at time t , nor need the switch be sudden. There are several reasons for this. Some economic agents may anticipate the policy change and adjust before it takes place. Some economic agents may not consider the policy change as credible and may take time to

learn. Also, there will be some costs of adjustment that accounts for the delays. Because of these factors, some gradual regime switching models have been suggested.

The earliest example is that of Bacon and Watts (1971), followed by Tsurumi (1983), Ohtani and Katayama (1985), and Ohtani *et al.* (1990). In Bacon and Watts and Tsurumi, the transition path from one regime to the other is specified by a hyperbolic tangent of time. In this method the start point of the transition from one regime to the other is estimated but the end point is obtained only indirectly through the adjustment coefficient. Tsurumi (1988) used this method to examine the stability of the US demand for money function using quarterly data from 1959:II to 1979:IV. As an estimate of the switch point he obtains 1973:I for the abrupt switching model and 1972:IV for the gradual switching model. However, the estimate of the speed of adjustment he gets implies that it takes roughly 26 quarters or 6.5 years to reach the second regime, that is 1972:IV to almost the end of the sample period 1979:IV as an adjustment period.

Ohtani and Katayama (1985) suggested a gradual switching regression model where both the start point and end point are estimated but the transition path is assumed to be a linear function of time. The transition path is thus very restrictive as compared to the one used by Bacon and Watts, and Tsurumi. However, in Ohtani *et al.* (1990) this procedure is extended by assuming the transition path to be a polynomial of time, the degree of the polynomial being decided by a model selection criterion (such as Akaike's AIC or Schwarz's BIC).

The model considered by Ohtani *et al.* is

$$y_t = x'_t(\beta + \lambda_t\delta) + e_t, \quad t = 1, 2, \dots, T$$

where

$$\begin{aligned} \lambda_t &= 0, & t \leq t_1 \\ &= \alpha_0 + \alpha_1 t + \alpha_2 t^2 + \dots + \alpha_n t^n, & t_1 < t < t_2 \\ &= 1, & t \geq t_2 \end{aligned}$$

Since $\lambda_t = 0$ for $t = t_1$ and $\lambda_t = 1$ for $t = t_2$, any two of the α s can be eliminated. (This is similar to the end-point restrictions in the Almon lag.) We can eliminate α_0 and α_1 . The unknown parameters are thus $\beta, \delta, \sigma^2, t_1, t_2, n$ and $\alpha_2, \alpha_3, \dots, \alpha_n$ with two of the α s eliminated, and the number of unknown parameters is thus $2k + n + 3$.

Ohtani *et al.* illustrate this with an estimation of structural change

in the import demand function for Japan following the second oil crisis. The data used were quarterly 1975:I–1986:IV. They estimated the start date of the structural change as 1978:III (slightly earlier than the second oil crisis) and that the adjustment to the new regime took about five years. We have given only two empirical examples of gradual switching models but both of them have yielded long lags in adjustment. Further experimentation with alternative functional forms for adjustment is needed to see whether these long lags are a statistical artifact. The Bayesian analysis of the model in Ohtani *et al.* is, however, rather complicated. The Bayesian analysis of the gradual switching model of Tsurumi is discussed in detail in Broemling and Tsurumi (1987).

In addition to these methods there is the D method suggested by Goldfeld and Quandt (1972) who consider the two-regime switching regression model

$$\begin{aligned} y_t &= \alpha_1 + \beta_1 x_t + u_{1t}, & 1 \leq t \leq n_0 \\ y_t &= \alpha_2 + \beta_2 x_t + u_{2t}, & n_0 < t \leq n \end{aligned}$$

where $u_i \sim iin(0, \sigma_i^2)$ for $i = 1, 2$. Define an indicator variable D_t such that if $D_t = 0$, then observations come from regime 1, and if $D_t = 1$, the observations come from regime 2. The Quandt method assumes that the indicator function D_t is a step function at $t = n_0$. Goldfeld and Quandt assume the switch to be gradual and that this is given by the indicator D_t to be cumulative normal

$$D_t = \Phi\left(\frac{t - \mu}{\sigma}\right)$$

where $\Phi(c) = Prob(x \leq c)$ with $x \sim N(0, 1)$, μ gives the mean point of the switch, and the lower σ is, the more sudden the switch is. The advantage of this method is that the distribution theory for μ is simpler because one can use standard asymptotic results. We combine the two equations as

$$y_t = \alpha_1(1 - D_t) + \alpha_2 D_t + [\beta_1(1 - D_t) + \beta_2 D_t]x_t + u_{1t}(1 - D_t) + u_{2t} D_t$$

and maximize the likelihood function for this model after substituting the expression for D_t . A recent application of this method is in Varoufakis and Sapsford (1991), where the methods of sudden switching produced anomalous results but the use of the D -method suggested that the problem was the assumption of a sudden switch, rather than a smooth transition to a second regime. The analysis with sudden shifts identified

two switch points, 1967 and 1973, but the D -method suggested that there was one switch which was gradual during 1967–1973.

For a Bayesian analysis of the gradual switching regression models, the D -method appears to be a very fruitful model to consider. There is, as yet, no Bayesian analysis of this method. The other models that incorporate gradual switching are the models with continuous parameter variation. A Bayesian analysis of these models is discussed extensively in West and Harrison (1989) and hence this area will not be pursued here.

15.7 A model with parameters following a random walk

Another class of gradual switching models is the one where the parameters follow a random walk. These models are also known as *state-space models*.

Consider the model

$$\begin{aligned} y_t &= X_t \beta_t + u_t \\ \beta_t &= \beta_{t-1} + v_t, \quad t = 1, \dots, n \end{aligned} \quad (15.5)$$

where

$$\begin{pmatrix} u_t \\ v_t \end{pmatrix} \sim IN \left[\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} \sigma_u^2 & 0 \\ 0 & \sigma_v^2 \end{pmatrix} \right]$$

Substituting the second equation into the first yields

$$\begin{aligned} y_t = X_t(\beta_{t-1} + v_t) + u_t &= X_t \hat{\beta}_{t-1} + X_t(\beta_{t-1} - \hat{\beta}_{t-1}) + X_t v_t + u_t \\ &= X_t \hat{\beta}_{t-1} + w_t \end{aligned}$$

where $\hat{\beta}_t$ is the estimator of β_t based on observations up to t and w_t is the one-step ahead prediction error. If $\text{var}(\hat{\beta}_{t-1}) = \lambda_{t-1}$, then $\text{var}(w_t) = f_t$ where $f_t = X_t^2 \lambda_{t-1} + X_t^2 \sigma_v^2 + \sigma_u^2$. Then maximizing the likelihood is equivalent to maximizing

$$-\frac{1}{2} \sum_t \left[\log f_t + \frac{w_t^2}{f_t} \right]$$

Note that this model can be written as an expanded regression as

follows

$$\begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \\ 0 \\ 0 \\ \vdots \\ 0 \end{bmatrix} = \begin{bmatrix} 0 & X_1 & 0 & \cdots & 0 \\ 0 & 0 & X_2 & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \cdots & X_n \\ 1 & -1 & 0 & \cdots & 0 \\ 0 & 1 & -1 & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & 1 & -1 \end{bmatrix} \begin{bmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_n \end{bmatrix} + \begin{bmatrix} u_1 \\ u_2 \\ \vdots \\ u_n \\ v_1 \\ v_2 \\ \vdots \\ v_n \end{bmatrix}$$

We can first estimate this equation by OLS, get $\hat{\sigma}_u^2$ and $\hat{\sigma}_v^2$ and then apply GLS. This expanded regression approach can be cumbersome. However, some have argued (see Snyder, 1990) that this does not pose a problem. Snyder argues that although the expanded regression is large in size, Page and Saunders (1977) have shown that the least squares problem can be solved efficiently with orthogonalization methods, such as Givens transformation (see Gentleman, 1973), which exploit the sparsity of the associated design matrix.

The ML estimates of σ_u^2 and σ_v^2 should be identical to those obtained from a representation of (15.5) in terms of the expanded regression model (written in the dummy variable form). Also, the likelihood function is conditional on β_0 but it should be possible to estimate β_0 as well, as can be seen from the expanded regression model. The estimate of β_t based on data up to t are obtained by estimating this expanded regression using data up to and including t only. This can be done for all t . On the other hand, if we use a regression with all the n observations, we get the *smoothed* estimates (see section 15.9).

15.8 A general state-space model

We shall now consider a general state-space model, from which we can derive many other models as special cases. The general model is

$$\text{Measurement equation: } y_t = X_t \beta_t + u_t$$

$$\text{Transition equation: } \beta_t = T_t \beta_{t-1} + v_t, \quad t = 1, \dots, n$$

where y_t is $n \times 1$ vector, X_t is $n \times m$ matrix, β_t is $m \times 1$ vector of coefficients, and

$$\begin{pmatrix} u_t \\ v_t \end{pmatrix} \sim N \left[\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} R_t & 0 \\ 0 & Q_t \end{pmatrix} \right]$$

In the usual applications of these models, we assume that T_t , Q_t , and R_t are known.

Define b_{t-1} as the GLS estimator of β_{t-1} using data at time $t-1$ and $b_{t|t-1}$ as the predictor of β_t using information up to $t-1$. Let V_{t-1} denote the variance of b_{t-1} . Note the following

$$b_{t|t-1} = T_t b_{t-1}$$

is the prediction of β_t based on the information up to time $t-1$. Since

$$\beta_t = T_t b_{t-1} + w_t$$

where $w_t = T_t(\beta_{t-1} - b_{t-1}) + v_t$, we have the variance of the prediction error

$$\text{var}(w_t) = T_t V_{t-1} T_t' + Q_t$$

Denote this by $V_{t|t-1}$. At time t , using the new observations, we have two estimators for β_t :

- (i) $(X_t' R_t^{-1} X_t)^{-1} X_t' R_t^{-1} y_t$ with variance $(X_t' R_t^{-1} X_t)^{-1}$
- (ii) $b_{t|t-1}$ with variance $V_{t|t-1}$

We get the efficient GLS estimator as the matrix weighted average of these estimators. The estimator b_t and its variance are given by

$$b_t = V_t [V_{t|t-1}^{-1} b_{t|t-1} + X_t' R_t^{-1} y_t] \quad (15.6)$$

$$V_t = [V_{t|t-1}^{-1} + X_t' R_t^{-1} X_t]^{-1} \quad (15.7)$$

Suppose we follow a different route and use the recursive prediction error method. Then we get a different set of equations. The one-step ahead prediction error in y_t is

$$\hat{u}_t = y_t - X_t b_{t|t-1}$$

Since we can write

$$y_t = X_t b_{t|t-1} + X_t(\beta_t - b_{t|t-1}) + u_t$$

the variance of the prediction error is

$$D_t = X_t V_{t|t-1} X_t' + R_t$$

Kalman's recurrence relations are

$$b_t = b_{t|t-1} + V_{t|t-1} X_t' D_t^{-1} (y_t - X_t b_{t|t-1}) \quad (15.8)$$

This can also be written as

$$b_t = b_{t|t-1} + G_t \hat{u}_t$$

where $G_t = V_{t|t-1} X_t' D_t^{-1}$ is called the *Kalman gain*. It is the correction term applied to the one-step ahead prediction error $\hat{u}_t$. Also

$$V_t = V_{t|t-1} - V_{t|t-1} X_t' D_t^{-1} X_t V_{t|t-1} \quad (15.9)$$

Note that the GLS equations (15.6) and (15.7) are different from the Kalman recursive equations in (15.8) and (15.9). However, we can show their equivalence using the following two lemmas.

Lemma 15.1 *If $V = (M^{-1} + X'R^{-1}X)^{-1}$, then $V = M - MX'D^{-1}XM$ where $D = R + XMX'$.*

The result can be checked by showing that $V^{-1}V = I$. This shows that the equations (15.7) and (15.9) are equivalent.

Lemma 15.2 *If $b_1 = V[M^{-1}b_0 + X'R^{-1}y]$, then $b_1 = b_0 + MX'D^{-1}(y - Xb_0)$.*

Note that $b_1 = [M - MX'D^{-1}XM][M^{-1}b_0 + X'R^{-1}y]$ using lemma 15.1. This shows that the equations (15.6) and (15.8) are equivalent. This proves the equivalence of Kalman's recurrence relation and GLS. This elegant proof is due to Duncan and Horn (1972).

15.9 Derivation of the Kalman filter

To derive the Kalman filter given by equation (15.8) and (15.9) we shall use the following fact based on normal regression theory. Suppose

$$\begin{pmatrix} y_1 \\ y_2 \end{pmatrix} \sim N \left[\begin{pmatrix} \mu_1 \\ \mu_2 \end{pmatrix}, \begin{pmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{pmatrix} \right]$$

Then the conditional distribution of y_1 given y_2 is normal with mean $\mu_1 - \Sigma_{12}\Sigma_{22}^{-1}(y_2 - \mu_2)$ and variance $\Sigma_{11} - \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21}$.

We shall use this fact to derive the distribution of β_t conditional on y_t . First note that

$$\beta_t = b_{t|t-1} + \beta_t - b_{t|t-1}$$

$$y_t = X_t b_{t|t-1} + X_t(\beta_t - b_{t|t-1}) + u_t$$

Hence we have a multivariate normal distribution as follow

$$\begin{pmatrix} \beta_t \\ y_t \end{pmatrix} \sim N \left[\begin{pmatrix} b_{t|t-1} \\ X_t b_{t|t-1} \end{pmatrix}, \begin{pmatrix} V_{t|t-1} & V_{t|t-1} X_t' \\ X_t V_{t|t-1} & D_t \end{pmatrix} \right]$$

Taking an estimator of b_t as the conditional mean $E(\beta_t|y_t)$ and using the above result we get the Kalman filter equation (15.8) and (15.9).

The above analysis is based on the assumption that $E(\beta_t|y_t)$ is the best estimator of β given the observations up to and including time t . Also, the Kalman filter estimator of β_t depends on the normality assumption. If the disturbances in the state-space model are not normally distributed, it is no longer true, in general, that the Kalman filter yields the conditional mean of the state vector. The computation of the conditional mean when the disturbances are nonnormal is described in section 3.7.3 of Harvey (1989). However, even if the disturbances are not normal the estimator b_t given by the Kalman filter is the minimum mean-squared-error linear estimator of β_t based on the observations up to and including time t . Proof of this result can be found in Duncan and Horn (1972) or Anderson and Moore (1979).

Smoothing

Consider the estimation of β_t based on data up to and including y_s . The terminology used in state-space models is as follows:

$s < t$: prediction

$s = t$: filtering

$s > t$: smoothing

Another term used is *signal extraction*. This refers to estimating the signal $X_t\beta_t$. In many models the dimension of y_t is less than that of β_t . In this case there is a more efficient way of estimating the signal. In the case of prediction several steps ahead, the prediction equations are used repeatedly but the updating equations (15.8) and (15.9) are not used. As for smoothing one particular case of interest is $s = t$. In this case we use all the sample data to derive estimates of β_t . The expanded regression model discussed in section 15.7 produces estimates of all the β_t using the entire set of n observations. When rational expectation models are estimated in the state-space framework, the estimates of β_t have to be obtained by the filtering method because the assumption underlying these models is that economic agents use all the information currently available. In these models, getting the smoothed estimates of β_t does not make sense.

To get the smoothed estimates of β_t we use the backward recursions starting with the final period estimates b_n and V_n , which are obtained by using the forward recursions given by equations (15.8) and (15.9).

There are two backward recursion methods for smoothing. The first, given in Harvey (1989, section 3.6) and derived in Anderson and Moore (1979, chapter 7) is as follows

$$b_{t|n} = b_t + V_t^*(b_{t+1|n} - T_{t+1}b_t)$$

$$V_{t|n} = V_t + V_t^*(V_{t+1|n} - V_{t+1|t})V_t^*$$

where

$$V_t^* = V_t T'_{t+1} V_{t+1|t}^{-1}, \quad t = n-1, \dots, 1$$

with $b_{n|n} = b_n$ and $V_{n|n} = V_n$. The problem with this smoother is that it involves inversion of $(n-1)$ matrices $V_{t+1|t}$ for $t = 1, 2, \dots, n-1$. An alternative smoother given by DeJong (1989) involves less computation. It involves no other inversion than that of D_t , which anyhow is inverted in the forward recursion of the Kalman filter.

Let us define

$$K_t = T_{t+1}G_t$$

where G_t is the Kalman Gain defined earlier as $G_t = V_{t|t-1}X'_t D_t^{-1}$. Define

$$L_t = T_{t+1} - K_t X_t$$

Then the DeJong smoothing equations are

$$b_{t|n} = b_{t|t-1} + V_{t|t-1}r_{t-1}$$

where r_t satisfies the backward recursion

$$r_{t-1} = X'_t D_t^{-1} \hat{u}_t + L'_t r_t, \quad t = n, n-1, \dots, 1$$

with $r_n = 0$. The $\hat{u}_t$ is the one-step ahead forecast error in u_t . This smoother is more elegant than the earlier one.

Note that in the expanded regression approach discussed in section 15.7 we estimate the β_t from all the observations $1, \dots, n$. Thus, the estimates we have are the smoothed estimates. We can get the filtered estimate of β_s by just estimating the expanded regression model using data $t = 1, \dots, s$. Thus, to get the filtered estimates we have to run many regressions. For the smoothed estimates, we have only one regression equation to estimate. Thus, in the expanded regression approach smoothing is easier than filtering.

Why use the Kalman filter?

There have been some papers that trace the Kalman filter to much earlier work. Sorenson (1970) suggests that the Kalman filter is not entirely due to Kalman. Lauritzen (1981) traces the recursive least squares estimation and prediction method to papers by a Danish statistician T.N. Thiele in 1880 (long before Kalman was born). Snyder (1990) questions the usefulness of the Kalman filter. He argues that when properly implemented, the regression approach not only involves lower computational loads than numerically stable versions of the Kalman filter, but it is inherently simpler and more transparent. Steyn (1989) argues that for some simple commonly used state-space models the asymptotic least squares (ALS) method can be recommended in preference to the Kalman filter.

15.10 Harvey's structural time series model (1989)

Consider

$$y_t = z\beta_t + d + u_t, \quad t = 1, 2, \dots, n$$

where z is $1 \times m$ vector of coefficients assumed to be time-invariant, β_t is $m \times 1$ vector, d is a time invariant scalar. The evolution of β_t is given by

$$\beta_t = T\beta_{t-1} + v_t, \quad t = 1, 2, \dots, n$$

where $E(v_t) = 0$, $E(v_t v_t') = Q$, $E(u_t v_t) = 0$, $E(u_t) = 0$, and $E(u_t^2) = \sigma_u^2$.

Let b_{t-1} be the estimator of β_{t-1} obtained from observations up to time $(t-1)$ and let its variance be denoted by V_{t-1} . Let us denote by $b_{t|t-1}$ the predictor of β_t obtained from b_{t-1} . Then we have

$$b_{t|t-1} = Tb_{t-1}$$

Also if we denote the variance of the prediction error $T(\beta_{t-1} - b_{t-1}) + v_t$ by $V_{t|t-1}$, then we have

$$V_{t|t-1} = TV_{t-1}T' + Q_t$$

Now consider the one-step ahead prediction error in y_t . If we denote this by $\hat{u}_t$, we have

$$\hat{u}_t = y_t - zb_{t|t-1} - d$$

and its variance, which we shall denote by D_t is given by

$$D_t = \sigma_u^2 + zV_{t|t-1}z'$$

Note that D_t is a scalar. The ML estimates of the parameters in the model are obtained by maximizing

$$L = -\frac{1}{2} \sum_1^n \left[\log |D_t| + \frac{\hat{u}_t^2}{D_t} \right]$$

To carry on the iterations, we also need the recursive relations for the estimator b_t of β_t and its variance V_t . These are given by

$$b_t = b_{t|t-1} + \frac{V_{t|t-1}z'\hat{u}_t}{D_t}$$

and

$$V_t = V_{t|t-1} - \frac{V_{t|t-1}z'zV_{t|t-1}}{D_t}$$

To simplify matters, consider the model

$$\begin{aligned} y_t &= \mu_t + u_t \\ \mu_t &= \mu_{t-1} + \gamma_{t-1} + \eta_t \\ \gamma_t &= \gamma_{t-1} + \zeta_t \end{aligned}$$

This is a slowly evolving trend model.

Case I $\sigma_\eta^2 = \sigma_\zeta^2 = 0, \gamma_0 = 0$. Then we get

$$y_t = \mu_0 + u_t$$

Case II $\sigma_\eta^2 = \sigma_\zeta^2 = 0, \gamma_0 \neq 0$. Then we get

$$y_t = \mu_0 + \gamma_0 t + u_t$$

which is a deterministic trend model.

Case III $\sigma_\eta^2 \neq 0$ or $\sigma_\zeta^2 \neq 0$. Then the time series is difference-stationary.

The case $\sigma_\eta^2 \neq 0$ allows for a parallel shift in the linear time trend, and the case $\sigma_\zeta^2 \neq 0$ allows for a change in the slope of the time trend.

The correspondence between this model and the previous one is given by the following equations

$$\beta_t = [\mu_t \ \gamma_t]'$$

$$v_t = [\eta_t \ \zeta_t]'$$

$$d = 0, \quad z = [1 \ 0]$$

$$T = \begin{bmatrix} 1 & 1 \\ 0 & 1 \end{bmatrix}, \quad Q = \begin{bmatrix} \sigma_\eta^2 & 0 \\ 0 & \sigma_\zeta^2 \end{bmatrix}$$

15.11 Further comments on structural time series models

The estimation of structural time series models is discussed in Harvey and Peters (1990). The method is maximum likelihood (ML) using the Kalman filter. A computer program to do this (that does not need any knowledge of the Kalman filter) is the STAMP program of Koopman *et al.* (1995).

There is one problem with the ML estimation of structural time series models discussed in Shephard (1993). Consider the model

$$\begin{aligned} y_t &= \beta' x_t + u_t \\ u_t &= \mu_t + \varepsilon_t, \quad \varepsilon_t \sim IN(0, \sigma^2) \\ \mu_t &= \mu_{t-1} + \eta_t, \quad \eta_t \sim IN(0, q\sigma^2) \end{aligned}$$

Interest centers on q . If $q = 0$ we have a deterministic trend. If $q > 0$, then the trend is a random walk.

Shephard (1993) argues that it is often observed that the ML estimate of q is zero even when the true $q > 0$. This probability of getting an estimate of $q = 0$ (implying a deterministic trend) is very sensitive to the type of the likelihood function used as the basis of inference. He considers two types of likelihood: *profile likelihood* (which, in econometric terminology is *concentrated likelihood*) and *marginal likelihood*. For profile likelihood, we maximize the likelihood function with respect to the other parameters, and substitute the ML estimates to get the concentrated likelihood function $L(q)$. In the marginal likelihood method, we integrate out the other parameters and get the marginal likelihood $L^*(q)$. In both cases we get an estimate of q by maximizing $L(q)$ or $L^*(q)$. Shephard argues that we get an estimate $q = 0$ more often with profile likelihood than with marginal likelihood. Thus, the estimation of the structural time series models should be done with marginal likelihood.

We shall not go into the details of the different problems with Kalman filtering. These are dealt with in the book by Harvey and several papers by him. He has presented several arguments for the use of the structural time series model. But the most scathing attack on traditional

time series modeling (with its emphasis on unit roots, VAR models, and cointegration) is in Harvey (1997). We feel that this has a sobering effect on current time series modeling (the unit root revolution, cointegration revolution, and so on) discussed in the previous chapters. Here are some quotes from Harvey (1997):

Since a deterministic trend is too restrictive, the obvious thing to do is to make it more flexible by letting the level and slope parameters change over time. (p. 192)

The problem with ARIMA class is that there are many models and parameter values which have no sensible interpretation and give forecast functions which may have undesirable properties. (p. 194)

Testing for unit roots has become almost mandatory in applied economics. This is despite the fact that, much of the time, it is either unnecessary or misleading or both. (p. 196)

Why worry about testing for unit roots in the first place? ... within the structural framework deciding on the degree of integration is not crucial ... very little is lost by starting off with a stochastic trend model which has the deterministic slope and level as a special case. If there really is a need to test whether a component can be treated as deterministic, it is best to use a test that is not based on an autoregressive approximation. (pp. 196–197)

Trying to draw conclusion about whether the world is Keynesian or Neoclassical on the basis of the presence or otherwise of unit root is complete nonsense. (p. 197)

To many econometricians, VAR stands for *Very Awful Regression*. (I am indebted to Arnold Zellner for introducing me to this delightful terminology.) (p. 199)

There are a number of reasons why my enthusiasm for VAR based cointegration methods is somewhat muted. The fundamental objection is that autoregressive approximation can be very poor, even if a large number of parameters is useful. (p. 199)

... what have economists learnt from fitting such models? The answer is very little. I cannot think of one article which has come up with a cointegrating relationship which we did not know about already from economic theory. (p. 199)

The solution is to combine the flexibility of a time series model with the interpretation of regression ... The recent emphasis on unit roots, vector autoregression and cointegration has focused too much attention on tackling uninteresting problems by flawed methods. (p. 200)

15.12 Summary and conclusions

The present chapter reviews the literature on regime switching models that have been used for identifying break points in time series. It also reviews the work on another alternative: outlier analysis. A major concern of these studies has been the consequence of structural breaks on inference about unit roots in macroeconomic time series. Another objective has been the testing of some economic theories within the context of a flexible nonlinear framework. As far as prediction goes, most of the papers have concentrated on within-sample predictions and compared them with simple random walk models (or in some cases ARIMA models). The out-of-sample predictions have not been investigated in detail, but the meager evidence available suggests that it is not encouraging. A major defect of the regime switching models in use, particularly the Markov switching regression (MSR) models (often termed the Hamilton model, though the terminology is unfortunate) is that they are simple trend change models with no explanatory variables. There have been some recent attempts to include explanatory variables in the MSR models, but more work needs to be done. The outlier models, on the other hand, start with an ARIMA model and then identify the outliers through residual analysis. These models again do not have any explanatory variables. Also the outlier models are less flexible than the regime switching models in this respect.

The switching regression models are based on sudden switches. In practice, the switch is gradual rather than sudden. Hence some gradual switching models are also reviewed in this chapter.

Another class of models that have been found useful is the set of models with continuous parameter variation rather than sudden or gradual switches. These models are also called *state-space models*. An example of this is Harvey's *structural time series model* which is estimated using the Kalman filter. In this chapter we present a brief review of this model and explain the Kalman filter and its limitations. We also review Harvey's criticism of the current emphasis on unit roots, vector autoregressions, and cointegration in time series modeling.

References

- Albert, J. and S. Chib (1993), "Bayesian Inference via Gibbs Sampling of Autoregressive Time Series Subject to Markov Mean

- and Variance Shifts," *Journal of Business and Economic Statistics*, 11, 1–15.
- Anderson, B.D.O. and J.B. Moore (1979), *Optimal Filtering*, Prentice Hall, New York.
- Bacon, D.W. and D.G. Watts (1971), "Estimating the Transition between Two Intersecting Lines," *Biometrika*, 58, 525–534.
- Baum, L. E., T. Petrie, G. Soules, and N. Weiss (1970), "A Maximization Technique Occurring in the Statistical Analysis of Probabilistic Functions of Markov Chains," *Annals of Mathematical Statistics*, 41, 164–171.
- Bekaert, G. and R.J. Hodrick (1993), "On Biases in the Measurement of Foreign Exchange Risk Premiums," *Journal of International Money and Finance*, 12, 115–138.
- Broemling, L.D. and H. Tsurumi (1987), *Econometrics and Structural Change*, Marcel Dekker, New York.
- Casella, G. and E. I. George (1992), "Explaining the Gibbs Sampler," *The American Statistician*, 46, 167–174.
- Cecchetti, S.G., P.S. Lam, and N.C. Mark (1990), "Mean Reversion in Equilibrium Asset Prices," *American Economic Review*, 80, 398–418.
- Cosslett, S. R. and L. Lee (1985), "Serial Correlation in Discrete Variable Models," *Journal of Econometrics*, 27, 79–97.
- DeJong, P. (1989), "Smoothing and Interpolation with the State-Space Model," *Journal of the American Statistical Association*, 84, 1085–1088.
- Diebold, F.X., J.H. Lee, and G.C. Weinbach (1994), "Regime Switching with Time-Varying Transition Probabilities," in C. Hargreaves (ed.), *Non-Stationary Time Series Analysis and Cointegration*, Oxford University Press, Oxford.
- Duncan, D.B. and S.D. Horn (1972), "Linear Dynamic Recursive Estimation from the Viewpoint of Regression Analysis," *Journal of the American Statistical Association*, 67, 815–821.
- Durland, J.M. and T.H. McCurdy (1992), "modeling Duration Dependence in Cyclical Data Using a Restricted Semi-Markov Process," Manuscript, Queens University.
- Engel, C. (1994), "Can the Markov Switching Model Forecast Exchange Rates?" *Journal of International Economics*, 36, 151–165.
- Engel, C. and J.D. Hamilton (1990), "Long Swings in the Dollar: Are They in the Data and Do Market Know It?" *American Economic Review*, 80, 689–713.

- Evans, M.D.D. (1993), "Estimating General Markov Switching Models," Manuscript, Stern School of Business, New York University.
- Filado, A.J. (1994), "Business Cycle Phases and Their Traditional Dynamics," *Journal of Business and Economic Statistics*, 12, 299–308.
- Franses H. and R. Rapp (1996), "Does Seasonal Adjustment Change Inference from Markov Switching Models?" Report 9615/A, Econometric Institute, Erasmus University, Rotterdam.
- Garcia, R. (1997), "Asymptotic Null Distribution of the Likelihood Ratio Test in Markov Switching Models," *International Economic Review*, Forthcoming.
- Garcia, R. and P. Perron (1996), "An Analysis of Real Interest Rate under Regime Shifts," *Review of Economics and Statistics*, 78, 111–125.
- Gelfand, A.E. and A.F.M. Smith (1990), "Sampling-Based Approaches to Calculating Marginal Densities," *Journal of the American Statistical Association*, 85, 398–409.
- Gentleman, W.M. (1973), "Least Squares Computations by Givens Transformations with Square Roots," *Institute of Mathematical Applications*, 12, 329–336.
- Goldfeld, S.M. and R. Quandt (1972), *nonlinear Methods in Econometrics*, North-Holland Publishing Co., Amsterdam.
- (1973), "A Markov Model for Switching Regressions," *Journal of Econometrics*, 1, 3–16.
- Golodwin, T.H. (1993), "Business Cycle Analysis with a Markov Switching Model," *Journal of Business and Economic Statistics*, 11, 331–339.
- Gurland, J.M. and T.H. McCurdy (1992), "modeling Duration Dependence in Cyclical Data Using a Restricted Semi-Markov Process," Manuscript, Queen's University, Kingston, Ontario, Canada.
- Hamilton, J.D. (1988), "Rational Expectations Econometric Analysis of Changes in Regime," *Journal of Economic Dynamics and Control*, 12, 385–423.
- (1989), "A New Approach to the Economic Analysis of Nonstationary Time Series and the Business Cycle," *Econometrica*, 57, 357–384.
- (1990), "Analysis of Time-Series Subject to Changes in Regime," *Journal of Econometrics*, 45, 39–70.

- (1992), "Estimation, Inference, and Forecasting of Time Series Subject to Changes in Regime," in G.S. Maddala, C.R. Rao, and H.D. Vinod (eds.), *Handbook of Statistics*, vol. 11, North-Holland, Amsterdam, chapter 9.
- Hamilton, J.D. and R. Susmel (1994), "Autoregressive Conditional Heteroskedasticity and Changes in Regime," *Journal of Econometrics*, 64, 307–333.
- Harvey, A.C. (1989), *Forecasting, Structural Time Series Models and the Kalman Filter*, Cambridge University Press.
- (1997), "Trends, Cycles, and Autoregressions," *The Economic Journal*, 107, 192–201.
- Harvey, A.C. and S.C. Peters (1990), "Estimation Procedures for Structural Time Series Models," *Journal of Forecasting*, 9, 89–108.
- Kaminsky, G. (1993), "Is there a Peso Problem? Evidence from Dollar/Pound Exchange Rate," *American Economic Review*, 83, 450–472.
- Kim, B.H. (1994), "A Study of Risk Premiums in the Foreign Exchange Market," unpublished Ph.D. dissertation, The Ohio State university.
- Kim, B.H. and J.H. Lee (1996), "Can the Markov Switching Model with Time-Varying Probabilities Forecast Exchange Rates?" Manuscript, Inje University, Korea.
- Kim, B.H. and B.C. Yu (1996), "On Time-Varying Coefficients Improve the Empirical Performance of Monetary Exchange Rate Determination Models?" Manuscript, Inje university, Korea.
- Kim, I.M. (1993), "A Dynamic Programming Approach to the Estimation of Markov Switching Regression Models," *Journal of Statistical Computation and Simulation*, 45, 61–76.
- Kim, I.M. and G.S. Maddala (1992), "Multiple Structural Breaks and Unit Roots in Exchange Rates," Paper presented at the Econometric Society Meeting at New Orleans.
- Koopman, S.J., A.C. Harvey, J.D. Doornik, and N. Shephard (1995), *STAMP 5.0 Structural Time Series Analysis, modeler and Predictor*, Chapman and Hall, London.
- Lahiri, K. and J.G. Wang (1994), "Predicting Cyclical Turning Points with Leading Index in a Markov Switching Model," *Journal of Forecasting*, 13, 243–263.
- Lam, P.S. (1990), "The Hamilton Model with a General Autoregressive Component: Estimation and Comparison with Other Models of

- Economic Time Series," *Journal of Monetary Economics*, 26, 409–432.
- Lauritzen, S.L. (1981), "Time Series Analysis in 1880: A Discussion of Contributions Made by T.N. Thiele," *International Statistical Review*, 49, 319–331.
- Lee, J.H. (1991), "Non-Stationary Markov Switching Models of Exchange Rates: The Pound-Dollar Exchange Rate," Ph.D. dissertation, University of Pennsylvania.
- Lee, L.F. and R.H. Porter (1984), "Switching Regression Models with Imperfect Sample Separation Information," *Econometrica*, 52, 391–418.
- Lindgren, G. (1978), "Markov Regime Models for Mixed Distributions and Switching Regressions," *Scandinavian Journal of Statistics*, 5, 81–91.
- Maddala, G.S. (1983), *Limited-Dependent and Qualitative Variables in Econometrics*, Cambridge University Press.
- Makov, U.E. and A.F.M. Smith (1977), "A Quasi-Bayes Unsupervised Learning Procedure for Priors," *IEEE Transactions on Information Theory*, IT-23, 761–764.
- McCulloch, R.E. and R.S. Tsay (1994a), "Statistical Analysis of Economic Time Series via Markov Switching Models," *Journal of Time Series Analysis*, 15, 523–539.
- (1994b), "Bayesian Inference of Trend- and Difference-Stationarity," *Econometric Theory*, 10, 596–608.
- Ohtani, K. and S. Katayama (1985), "An Alternative Gradual Switching Regression Model and Its Applications," *Economic Studies Quarterly*, 36, 148–153.
- Ohtani, K., S. Kakimoto, and K. Abe (1990), "A Gradual Switching Model with a Flexible Transition Path," *Economics Letters*, 32, 43–48.
- Pagan A.R. and G.W. Schwert (1990), "Alternative Models for Conditional Stock Volatility," *Journal of Econometrics*, 45, 267–290.
- Page, C.C. and M.A. Saunders (1977), "Linear Least Squares Estimation of Discrete Linear Dynamical Systems Using Orthogonal Transformations," *SIAM Journal of Numerical Analysis*, 15, 180–193.
- Perron, P. (1989) "The Great Crash, the Oil Price Shock and the Unit Root Hypothesis," *Econometrica*, 57, 1361–1401.
- Quandt, R. (1958), "The Estimation of the Parameters of a Linear

- Regression System Obeying Two Separate Regimes," *Journal of American Statistical Association*, 53, 873-880.
- Rabault, G. (1992), "Une Application du Modèle de Hamilton a L'Estimation des Cycles Economiques" (in French), Discussion paper No. 9210, INSEE, Paris, France.
- Raymond, J.E. and R.W. Rich (1992), "Changes in Regime and the Behavior of Inflation," Manuscript, Vanderbilt University.
- Rich, R.W. and J.E. Raymond (1992), "Oil and the Macro-economy: A Markov State Switching Approach," Manuscript, Vanderbilt University.
- Sclove, S. L. (1983), "Time-Series Segmentation: a Model and a Method," *Information Sciences*, 29, 7-25.
- Shephard, N. (1993), "Maximum Likelihood Estimation of Regression Models with Stochastic Trend Components," *Journal of American Statistical Association*, 88, 590-601.
- Snyder, R.D. (1990), "Why Kalman Filter?" Working paper No.11/90, Monash University.
- Sorenson, H.W. (1970), "Least Squares Estimation from Gauss to Kalman," *IEEE Spectrum*, 7, 63-68.
- Steyn, I.J. (1989), "ALS Estimation of Parameters in a State Space Model," Institute of Actuarial Science and Econometrics Report AE 2/89, University of Amsterdam.
- Titterton, D.M., A.F.M. Smith, and U.E. Makov (1985), *Statistical Analysis of Finite Mixture Distributions*, John Wiley and Sons, New York.
- Tsurumi, H. (1983), "A Bayesian and Maximum Likelihood Analysis of a Gradual Switching Regression Model with Sampling Experiments," *Economic Studies Quarterly*, 34, 237-248.
- (1988), "A Survey of Bayesian and Non-Bayesian Testing of Model Stability in Econometrics," in J.C. Spall (ed.), *Bayesian Analysis of Time Series and Dynamic Models*, Marcel Dekker, New York.
- Turner, C.M., R. Startz, and C.R. Nelson (1989), "A Markov Model of Heteroskedasticity, Risk and Learning in the Stock Market," *Journal of Financial Economics*, 25, 3-22.
- Tyssedal, J.S. and D. Tjøstheim (1988), "An Autoregressive Model with Suddenly Changing Parameters and an Application to Stock Market Prices," *Applied Statistics*, 37, 353-369.
- Van Norden, S. (1994), "Regime Switching as a Test for Exchange Rate Bubbles," Manuscript, Bank of Canada.

- Varoufakis, Y. and D. Sapsford (1991), "Discrete and Smooth Switching Regressions for Australian Labor Productivity Growth," *Applied Economics*, 23, 1299–1304.
- West, M. and J. Harrison (1989), *Bayesian Forecasting and Dynamic Models*, Springer Verlag, New York.

16

Future directions

A chapter on future directions will have several subjective elements. Here we list topics that should be pursued in the future and also mention those topics that should not be pursued in the future.

Nonlinear models

Throughout the book we considered linear models only. Obviously all the topics discussed earlier need to be extended to nonlinear models. Thus, nonlinear error correction, nonlinear cointegration, and nonlinear structural change are possible topics. Some references here are: Granger (1995), Granger and Swanson (1996, 1997), Granger and Terasvirta (1993), and Granger, Inoue, and Norin (1997).

Bootstrap methods

There is currently too much asymptotic theory. Significance levels obtained from asymptotic distributions have been found to be misleading unless samples are very large. As reviewed in chapter 10, bootstrap methods are promising in making inferences with moderate-sized samples. More work needs to be done in this direction.

Robust methods

In chapter 14, we discussed robust unit root tests (section 14.5) and robust estimation methods for cointegrating regressions (section 14.6). Given that many error distributions are nonnormal, more developments of robust methods will be practically useful. Lucas (1995) develops tests for cointegration using pseudo-likelihood methods.

Pre-testing problems

The problems with unit roots and cointegration analysis within the con-

text of linear models are far from solved. Thus, future developments in nonlinear models (as suggested by Granger and Swanson 1996, 1997) maybe premature. One important question that has not even been asked is the appropriate significance levels at which the unit root tests should be applied, when they are pre-tests – prelude to cointegration analysis. A similar pre-testing problem occurs when a test is used to determine the number of cointegrating vectors in the Johansen procedure and then subsequently the coefficients of the estimated cointegrating vectors are tested. More work needs to be done in this important area.

Bayesian methods

The Bayesian methods (reviewed in chapter 8) argue that the commonly used procedure of applying a unit root test and behaving as if the series is a unit root process leads to misleading inferences. There is a nonzero probability that the process is a stationary process. In this respect the Bayesian methods are on a sounder footing than the classical methods because the posterior probabilities of the process being a unit root process and a stationary process are taken into account. There are still many open issues particularly regarding the appropriate priors for VAR models and cointegration analysis. The priors used for multivariate regression models are not appropriate here because the VAR model is different from a multivariate regression model with exogenous explanatory variables. Also, the Bayesian approach to the pre-testing problem in cointegration analysis needs to be investigated.

Structural time series models

Harvey (1997) has vehemently attacked the whole approach to unit root testing and offers structural time series models as an alternative. Harvey and Koopman (1996) discuss multivariate structural time series models.

Structural time series models offer a promising alternative to the *unit root route*. One often gets the feeling that the unit root mania has gone too far. The concepts of unit roots and cointegration have acquired a life of their own because the mathematics is fascinating and many just get carried away without any thinking about why we are doing what we are doing. It is time to ask the questions: why use unit root tests and why do cointegration analysis?

An important problem that has received attention recently and on which more work needs to be done is that of analyzing structural breaks within the framework of state-space models. Some references here are: McCulloch (1997) and Iyer and Andrews (1997).

Identification problems

Without feeding in some economic information, the cointegrating vectors do not have any economic interpretation. The starting point is a VAR (which is a-theoretical) and cointegration is a purely statistical concept. As we discussed in chapter 4, there is the issue of at what stage we feed in economic information, at the beginning by starting with a structural system rather than by VAR, or after determining the cointegration space as done by Johansen. The issues in identification need more detailed study.

The frequent procedure of doing cointegration analysis without discussing the identification problems has prompted Wickens (1995, p. 1645) to say that

Far from uncovering long-run relations, it can be shown that cointegration analysis is more likely to obscure them.

As we discussed earlier in chapter 6 not all cointegrating relationships can be interpreted as long-run relationships and any linear combinations of cointegrating vectors is also a cointegrating relationship.

Miscellaneous other problems

There have been other avenues in which unit roots and cointegration have been extended, e.g., stochastic unit roots by Granger and Swanson (1997) and multicointegration by Granger and Lee (1990) which has been applied to study present value relationships by Engsted *et al.* (1995). More such extensions will be forthcoming in the future. It remains to be seen what more comes out of all these extensions.

What is not needed

What we do not need is more unit root tests (each of which uses the Nelson–Plosser data as a guinea pig).

References

- Engsted, T., J. Gonzalo, and N. Haldrup (1995), “Multicointegration and Present Value Models,” Manuscript, Aarhus University, Denmark.
- Granger, C.W.J. (1995), “Modeling nonlinear relationships Between Extended Memory Variables,” *Econometrica*, 63, 265–279.
- Granger, C.W.J. and T.H. Lee (1990), “Multicointegration,” in G.F.

- Rhodes and T.B. Fomby (eds.), *Advances in Econometrics*, 8, JAI Press, Greenwich, Conn., 71–84.
- Granger, C.W.J. and N.R. Swanson (1996), “Future Developments in the Study of Cointegrated Variables,” *Oxford Bulletin of Economics and Statistics*, 58, 537–553.
- (1997), “An Introduction to Stochastic Unit Root Processes,” *Journal of Econometrics*, 80, 35–62.
- Granger, C.W.J. and T. Terasvirta (1993), *Modeling nonlinear Economic Time Series*, Oxford University Press, Oxford.
- Granger, C.W.J., T. Inoue, and N. Norin (1997), “Nonlinear Stochastic Trends,” *Journal of Econometrics*, Forthcoming.
- Harvey, A.C. (1997), “Trends, Cycles, and Autoregressions,” *The Economic Journal*, 107, 192–201.
- Harvey, A.C. and S.J. Koopman (1996), “Multivariate Structural Time Series Models,” in C. Heij *et al.* (eds.), *Systematic Dynamics in Economic and Financial Models*, John Wiley, New York.
- Iyer, S. and R. L. Andrews (1997), “Forecasting with Multi-Regime Structural Time Series Models: An Application to Nominal Interest Rates,” Working paper 97-07, Department of Economics, University of Delaware.
- Lucas, A., (1995), “Cointegration Testing Using Pseudo-Likelihood Ratio Tests,” Discussion paper TI 95-202, Tinbergen Institute, Rotterdam.
- McCulloch (1997), “Structural Shifts within the Context of State-Space Models,” Manuscript, Department of Economics, The Ohio State University.
- Wickens, M.R. (1995), “Real Business Cycle Analysis: A Needed Revolution in Macroeconometrics,” *The Economic Journal*, 105, 1637–1648.

Appendix 1

A brief guide to asymptotic theory

As we mentioned in the introduction we omitted completely asymptotic theory in our discussion saying that it can be found in the respective papers cited. However, the following references give the basics of asymptotic theory needed in most cases. For more specialized results, the respective papers have to be consulted.

The papers by Chan and Wei (1987), Phillips (1987a), and Chan (1988) are on local to unity asymptotics.

Andrews, D.W.K. (1993), "An Introduction to Econometric Applications of Empirical Process Theory for Dependent Random Variables," *Econometric Reviews*, 12, 183–216.

Billingsley, P. (1968), *Convergence of Probability Measures*, Wiley and Sons, New York.

Chan, N.H. (1988), "On the Parameter Inference for Nearly Nonstationary Time Series," *Journal of the American Statistical Association*, 83, 857–862.

(1989), "Asymptotic Inference for Unstable Autoregressive Time Series with Drifts," *Journal of Statistical Planning and Inference*, 23, 301–312.

Chan, N.H. and C.Z. Wei (1987), "Asymptotic Inference for Nearly Nonstationary AR(1) Processes," *The Annals of Statistics*, 15, 1050–1063.

(1988), "Limiting Distributions of Least Squares Estimates of Unstable Autoregressive Processes," *The Annals of Statistics*, 16, 367–401.

Johansen, S. (1995), *Likelihood Based Inference in Cointegrated Vector Autoregressive Models*, Oxford University Press, Oxford.

- Park, J.Y. and P.C.B. Phillips (1988), "Statistical Inference in Regressions with Integrated Processes I," *Econometric Theory*, 4, 468–497.
- (1989), "Statistical Inference in Regressions with Integrated Processes II," *Econometric Theory*, 5, 95–131.
- Phillips, P.C.B. (1986), "Understanding Spurious Regressions in Econometrics," *Journal of Econometrics*, 33, 311–340.
- (1987a), "toward a Unified Asymptotic Theory for Autoregression," *Biometrika*, 74, 535–547.
- (1987b), "Time Series Regression with a Unit Root," *Econometrica*, 55, 277–301.
- Phillips, P.C.B. and V. Solo (1992), "Asymptotics for Linear Processes," *The Annals of Statistics*, 20, 971–1001.
- Tanaka, K. (1996), *Time Series Analysis: Non-Stationary and Non-Invertible Distribution Theory*, John Wiley, New York.

Author index

- Abadir, K. M., 140, 144–146
 Abe, K., 483
 Abraham, B., 437, 438, 449, 450
 Abramovich, L., 311, 336
 Abuaf, N., 241, 242
 Agiakoglou, C., 77, 93, 101, 146
 Ahn, S. K., 168, 187, 191, 192, 196, 225
 Albert, J., 461, 462, 480
 Algoskoufis, G., 35, 42, 192
 Alt, R., 393, 421, 423
 Amano, R. A., 125, 126, 147
 Anderson, B. D. O., 473, 474, 480
 Anderson, R. L., 61, 93
 Anderson, T. W., 3, 6, 87, 93, 116, 147, 168, 192
 Andrews, D. W. K., 64, 97, 141–144, 147, 277, 292, 322, 336, 391, 394–397, 401, 402, 409, 414, 418, 424, 490
 Andrews, R. L., 487, 489
 Ansley, C. F., 19, 42, 93
 Arellano, C., 120, 125, 147
 Atkinson, A. C., 438, 449, 450

 Backus, D. K., 303, 306
 Bacon, D. W., 467, 480
 Bai, J., 398, 417, 418
 Baillie, R. T., 234, 235, 242, 300, 302–306
 Balke, N. S., 190, 192, 425, 428–430, 434, 439, 449, 450
 Banerjee, A., 37, 42, 157, 162, 176–178, 192, 203–205, 242, 251, 252, 258, 397, 401, 402, 408, 409, 418
 Bardsen, G., 36, 42
 Barnard, G. A., 309, 337
 Basawa, I. V., 325, 326, 336, 337
 Baum, L. E., 458, 480
 Bauwens, L., 288–290, 292

 Beaulieu, J. J., 367, 383
 Bekaert, G., 465, 480
 Bell, W. R., 91, 94, 372, 383
 Beltratti, A., 237, 244
 Ben-David, D., 401, 409, 418
 Beran, J., 300, 306, 447, 450
 Beran, R., 312, 313, 315, 337
 Berger, J. O., 277, 278, 284, 292
 Bernardo, J. M., 284, 292, 293
 Beveridge, S., 32, 42, 101, 147
 Bewley, R., 36, 37, 42, 164, 169–171, 181, 182, 192, 211, 222–225, 242, 243, 248, 254
 Bhargava, A., 38–40, 42, 70, 82–86, 88, 92, 93, 97, 105, 115, 116, 118, 152, 159, 196, 325, 337
 Bianchi, M., 407, 418
 Bickel, P. J., 311, 337
 Bierens, H. J., 125, 140, 147
 Billingsley, P., 54, 93, 490
 Birchenhall, C. R., 386
 Blanchard, O. J., 93, 193
 Blough, S. R., 102, 103, 147
 Bollerslev, T., 234, 235, 242, 302, 304, 306
 Bossaerts, P., 169, 181, 192, 222, 223, 243
 Boswijk, H. P., 175, 192, 205, 215, 219, 243, 334, 381, 383
 Box, G. E. P., 3, 4, 6, 8, 17–19, 42, 43, 47, 93, 140, 165, 169, 173, 181, 182, 184, 191, 192, 211, 222, 230, 334, 336, 342, 344, 362, 365, 371, 383
 Breitung, J., 119, 120, 134, 147
 Breusch, T. S., 37, 44, 197, 254
 Broemling, L. D., 463, 480
 Brown, R. L., 392, 393, 418
 Bruce, A. G., 438, 450
 Burke, S. P., 127, 128, 147, 206

- Burridge, P., 140, 147
 Burrows, P. M., 419
- Campbell, J. Y., 73, 90, 91, 93, 168,
 169, 175, 192, 193, 196, 202,
 212, 217, 233, 243, 247, 409,
 416, 418
- Campos, J., 193, 410, 419
- Canova, F., 365, 366, 370, 383
- Cantrell, R. S., 391, 419
- Cappuccio, N., 182, 193
- Carlstein, E., 328–332, 337
- Carvalho, J. L., 372, 386
- Casella, G., 277, 292, 462, 480
- Cavanagh, C. L., 255–258
- Cecchetti, S. G., 87, 93, 459, 480
- Chan, N. H., 66, 93, 490
- Chang, Y., 352, 360
- Chen, C., 395, 419, 439, 450
- Chen, D., 142, 144, 147
- Cheung, Y., 134, 147, 216, 225, 243,
 302, 303, 306
- Chib, S., 461, 462, 480
- Chiu, A. P. L., 386
- Choi, I., 106, 107, 109, 110, 117, 118,
 120, 125–127, 129–131, 148,
 206, 207, 247
- Chow, G. C., 390, 391, 394, 396, 419
- Christiano, L. J., 419
- Chu, C. S., 397, 419
- Chung, A., 438, 449
- Chung, B., 130, 148
- Clements, M. P., 186, 193
- Cochrane, J. H., 73, 86–88, 91, 93, 101,
 102, 127, 139, 146, 148, 298,
 299, 306
- Cook, R. D., 437, 439, 450
- Copeland, L. S., 241, 246
- Corbae, D., 408, 409, 419
- Cordero, M., 439, 450
- Cosslett, S. R., 456–458, 465, 480
- Cox, D. D., 443, 444, 450
- Cramer, H., 131, 148
- Crato, N., 300, 303, 306
- Cuthbertson, K., 237, 243
- Davidson, J. E. H., 35, 43, 94, 362, 383
- Davies, R. B., 394, 419
- Davison, A. C., 329, 337
- Dawid, A. P., 292
- de Lima, P. J. F., 300, 306
- DeJong, D. N., 79, 82, 94, 100, 148,
 268–270, 278, 292, 294, 403,
 419
- DeJong, P., 474, 480
- Delampady, M., 278, 292
- DiCiccio, T., 312, 313, 337
- Dickey, D. A., 39, 43, 47, 61–63, 65–67,
 71–77, 79, 82, 90–92, 94, 97,
 100, 101, 104–107, 109, 112,
 113, 119, 127, 131, 133, 148,
 150, 152, 203, 223, 231, 298,
 306, 325–328, 343–348, 359,
 366, 384, 400, 409, 430, 432
- Diebold, F. X., 129, 148, 298, 302–304,
 306, 365, 384, 395, 419, 465,
 480
- Dolado, J. J., 42, 176, 192, 246, 258
- Domowitz, I., 80, 97
- Donald, J. J., 242
- Donald, S. G., 426, 437, 439, 450
- Donsker, M. D., 54, 55, 61, 94
- Doornik, J. D., 482
- Dropsy, V., 409, 419
- Dueker, M., 302, 303, 305, 307
- Dufour, J. M., 322, 337, 391, 414, 419
- Duncan, D. B., 472, 473, 480
- Durbin, J., 392, 418, 439, 450
- Durland, J. M., 464, 480
- Eddy, W. F., 278, 292
- Edison, H. J., 228, 243
- Efron, B., 309–314, 318, 333, 338, 339,
 341
- Eitrheim, O., 182, 193, 215, 243
- Elliott, G., 99, 113, 134, 148, 193, 219,
 221, 222, 230, 231, 233, 243,
 244, 254, 255, 258, 259, 371,
 444, 450
- Engel, C., 459–461, 463, 466, 481
- Engle, R. F., 40, 41, 43, 97, 153, 157,
 162, 169, 176, 178, 181, 185,
 186, 190, 193, 197–199, 202,
 205, 215, 220, 221, 224, 234,
 235, 241, 244, 246, 366, 375,
 376, 384, 385, 410
- Engsted, T., 488
- Entorf, H., 29, 33, 43, 180, 181, 193
- Ericsson, N. R., 163, 193, 246, 356, 359,
 410, 419
- Escribano, A., 193
- Evans, G. B. A., 61, 64, 94
- Evans, J. M., 392, 418
- Evans, M. D. D., 465, 481
- Fachin, S., 334, 338
- Fair, R. C., 391, 418
- Ferretti, N., 325, 326, 338
- Fiebig, D. G., 438, 451
- Filado, A. J., 465, 481
- Fisher, L. A., 192
- Fisher, R. A., 134, 137, 149, 230
- Fisher, S., 93, 193
- Fomby, T. B., 152, 190, 192, 194, 196,

- 247, 292, 419, 425, 428–430,
434, 439, 449, 450, 453
Fox, A. J., 426, 437, 439, 451
Fox, R., 300–302, 307
Franses, P. H., 205, 215, 219, 241, 243,
244, 334, 362, 367, 368, 370,
371, 379–381, 383, 384, 428,
431, 432, 439, 445, 447, 448,
451, 466, 481
Freedman, D. A., 311, 318–320, 333,
335, 337, 338
Friedman, M., 63, 94
Fuller, W. A., 3, 7, 39, 43, 47, 61–63,
65–67, 71–77, 79, 81, 82, 92,
94, 99–101, 104–107, 109,
112, 113, 127, 148–152, 203,
223, 231, 244, 298, 306,
325–328, 342–344, 346–348,
359, 366, 384, 400, 409, 430,
432
Gagnon, J. E., 243
Galbraith, J. W., 42, 192, 258
Garbade, K., 393, 419
Garcia, R., 397, 398, 408, 420, 454, 463,
481
Gardeazabal, J., 304, 306
Geisser, S., 278, 292
Gelfand, A. E., 405, 420, 462, 481
Gentleman, W. M., 470, 481
George, E. I., 462, 480
Geweke, J., 91, 95, 267, 270, 288, 289,
293, 300, 305, 307, 421
Ghysels, E., 362, 364, 365, 377, 379,
384, 385, 414, 419, 420
Gil-Alana, L. A., 300, 305, 307
Goldfeld, S. M., 455–457, 468, 481
Gong, G., 333, 338
Gonzalez-Farias, G., 151
Gonzalo, J., 101, 149, 170, 171, 173,
181, 182, 191, 193, 216, 220,
221, 244, 488
Goodwin, T. H., 466, 481
Gould, J. P., 63, 94
Granger, C. W. J., 3, 7, 26, 28, 29, 35,
40, 41, 43, 157, 162, 169, 176,
178, 181, 185, 186, 188–190,
193, 194, 198, 202, 203, 205,
215, 220, 221, 224, 234, 235,
241, 244, 246, 296, 307, 366,
375, 376, 379–382, 384, 385,
410, 486–489
Gregoir, S., 348, 359
Gregory, A. W., 216, 218, 219, 227, 244,
314, 339, 410, 412, 420
Grether, D. D., 372, 386
Grilli, V., 237, 244
Guerre, E., 140, 147
Guo, S., 125, 147
Gurland, J. M., 481
Hakkio, C. S., 229, 234, 236, 244
Haldrup, N., 252, 259, 346–351, 356,
359, 428, 431, 432, 451, 488
Hall, A., 78, 79, 94, 104–106, 149, 151,
164, 391, 419
Hall, P., 54, 94, 309, 311–313, 315, 316,
323, 327, 329–331, 337, 339
Hamilton, J. D., 48, 94, 397, 457–466,
479, 481–484
Hampel, H. R., 443, 451
Hansen, B. E., 161, 162, 177–179, 184,
194, 196, 202, 203, 231, 232,
241, 244, 246, 256, 366, 370,
383, 394, 396–398, 411, 412,
414, 420, 445
Hansen, H., 413, 414, 420
Hardle, W., 318, 339
Hargreaves, C., 96, 183, 194, 244, 385
Harris, D., 171, 172, 182, 194, 206, 209,
210, 226, 244
Harris, R., 194
Harris, R. I. D., 328, 339
Harrison, J., 439, 453, 469, 485
Hartigan, J. A., 309, 310, 312, 339
Harvey, A. C., 117, 122, 149, 366, 373,
379, 385, 454, 473, 474,
477–479, 482, 487, 489
Hasan, M. N., 442, 445, 451
Hassett, K., 106, 149
Hassler, U., 298, 307
Hasza, D. P., 112, 148, 342, 343, 346,
347, 359, 366, 384
Haug, A. A., 218, 219, 244
Hendry, D. F., 3, 7, 35, 42–44, 157, 162,
163, 165, 176–180, 186,
192–194, 258, 356, 359, 362,
383, 407, 410, 419, 420, 428,
430, 451
Henricsson, R., 126, 149
Herce, M. A., 442, 444, 451
Herrndorf, N., 54, 95
Heyde, C. C., 54, 94
Hillmer, S. C., 372, 383
Hinkley, D. V., 315, 339
Ho, M. S., 220, 245
Hoa, K., 411, 412, 420
Hodges, J., 277, 293
Hodrick, R. J., 465, 480
Hoek, H., 442, 444, 445, 447, 451
Holbert, D., 397, 404, 420
Hooker, M. A., 228, 229, 245
Horn, S. D., 472, 473, 480

- Horowitz, J. L., 309, 314, 315, 330, 331, 339
- Horvath, M. T. K., 227, 228, 245
- Hosking, J. R. M., 296, 297, 307
- Hsiao, C., 174, 194
- Hu, W., 138, 150, 229, 245
- Huber, P. J., 440, 441, 451
- Hurwicz, L., 141, 149
- Hwang, J., 114–116, 149
- Hylleberg, S., 4, 7, 362, 366–368, 370, 372, 375, 379, 381, 383–385
- Im, K. S., 136, 137, 149
- Imhof, J. P., 64, 82, 95, 142, 149
- Inder, B., 162, 164, 179, 180, 191, 194, 206, 209, 210, 244, 411, 412, 420
- Inoue, T., 486, 489
- Iyer, S., 487, 489
- Jeffreys, H., 263, 265, 266, 271–276, 284, 286, 289, 291, 293, 295
- Jenkins, G. M., 3, 4, 6, 8, 17, 18, 42, 43, 47, 93, 140, 173, 342, 344, 362, 365, 371, 383
- Jeong, J., 309, 310, 314, 318, 319, 340
- Johansen, S., 165, 166, 168–175, 181–184, 187, 188, 191, 192, 194–196, 202, 203, 211, 212, 214–221, 223–227, 229, 230, 234, 235, 237–242, 245–247, 249, 312, 334, 336, 351, 353–360, 376, 377, 411, 413, 414, 420, 447, 490
- Jorion, P., 241, 242
- Joutz, F., 187, 195
- Joyeux, R., 296, 307, 377, 385
- Juselius, K., 173, 174, 195, 211, 218, 237, 240, 245, 353, 357, 358, 360
- Künsch, H. R., 328–332, 340
- Kakimoto, S., 483
- Kaminsky, G., 465, 482
- Kang, H., 34, 44
- Kang, K. M., 116, 149
- Kao, C., 393, 421
- Kashiwagi, N., 406, 407, 421
- Kass, R. E., 284, 293
- Katanaka, M., 259
- Katayama, S., 467, 483
- Kendall, M. G., 141, 149
- Kilian, L., 313, 314, 340
- Kim, B. H., 465, 466, 482
- Kim, H. J., 392, 394, 396, 421
- Kim, I. M., 230, 245, 271, 293, 397, 398, 404–407, 409, 415, 416, 421, 459, 460, 462, 482
- King, M. L., 113, 149
- Kitamura, Y., 256, 257, 259, 351, 352, 360
- Kiviet, J. F., 316, 317, 319–321, 323–325, 338, 340, 414, 419
- Kleibergen, F., 188, 195, 227, 228, 245, 288, 289, 293
- Kloek, T., 270, 293
- Koedijk, K., 241, 246
- Koenker, R. W., 442, 445, 451
- Kofman, P., 244
- Kolmogorov, A., 8, 43
- Kontrus, K., 393, 394, 421, 423
- Koop, G., 283, 287, 288, 293, 305, 307
- Koop, K., 291, 293
- Koopman, S. J., 438, 450, 477, 482
- Koopmans, T. C., 96
- Koopmen, S. J., 487, 489
- Kramer, W., 393, 411, 421, 423
- Kremers, J. J., 203–205, 232, 246
- Kuan, C. M., 415, 422
- Kullback, S., 390, 421
- Kunst, R. M., 377, 385
- Kwiatkowski, D., 120, 149
- Lütkepohl, H., 379, 381, 386
- Lahiri, K., 139, 150, 229, 246, 465, 483
- Lai, K., 134, 147, 216, 225, 243, 302, 306
- Lam, P. S., 87, 93, 459, 461, 464, 480, 483
- Lange, K. L., 445, 447, 451
- Laroque, G., 348, 359
- Larson, H. K., 235, 247
- Lauritzen, S. L., 475, 483
- Layton, A. P., 235, 246
- Le Breton, A., 141, 150
- Leamer, E., 239, 246, 271, 277, 280, 281, 293
- Ledolter, J., 438, 442, 451, 452
- Lee, D., 299, 305, 307
- Lee, H. S., 375–377, 379, 384–386
- Lee, I., 396, 418
- Lee, J., 105, 150
- Lee, J. H., 465, 466, 480, 482, 483
- Lee, L. F., 456–458, 465, 480, 483
- Lee, T. H., 101, 149, 173, 220, 221, 244, 488
- LeSage, J. P., 187, 195
- Levin, A., 135, 136, 138, 145, 150
- Leybourne, S. J., 99, 109–112, 117, 120, 122–124, 126, 127, 131, 150, 205–207, 246, 394, 411, 421
- Li, H., 105, 150, 184, 195, 309, 312, 317, 319, 322, 329, 330, 333–336, 340

- Li, W. K., 300, 301, 307
 Liatas, I., 443, 444, 450
 Lin, C-F., 135, 136, 138, 145, 150
 Lindgren, G., 454, 483
 Lindley, D. V., 266, 293
 Little, R. J. A., 451
 Liu, L. M., 439, 450
 Liu, P. C., 101, 137, 150, 240, 246
 Liu, R. Y., 328, 329
 Ljung, G. M., 438, 452
 Lo, A. W., 87, 95, 298, 299, 305, 307,
 391, 421
 Loretan, M., 161-165, 177, 179-181, 196
 Lovell, M. C., 4, 7
 Lubian, D., 182, 193
 Lubrano, M., 275, 276, 288-290, 292,
 294
 Lucas, A., 428, 431, 439, 441-445, 447,
 451, 452, 486, 489
 Lumsdaine, R. L., 397, 401, 408, 409,
 418, 421
 Lundbäck, E., 126, 149
 Luukkonen, R., 152
 Lyhagen, J., 335, 340

 Maasoumi, E., 438, 451
 MacDonald, R., 234, 235, 246
 MacKinlay, A. C., 87, 95, 298, 299, 307
 MacKinnon, J. G., 199, 201, 202, 216,
 246
 Maddala, G. S., 20, 43, 95, 137, 138,
 150, 151, 184, 187, 195, 215,
 233, 246, 253, 259, 271, 293,
 309, 310, 312, 314, 317-319,
 322, 329, 330, 333-336, 340,
 357, 360, 397, 398, 405-407,
 409, 415, 421, 425, 426, 428,
 431, 432, 437, 439, 450, 452,
 453, 455, 462, 482, 483
 Maekawa, K., 253, 254, 259
 Makov, U. E., 459, 483, 484
 Mallik, A. K., 337
 Mamingi, N., 229, 246
 Mammen, E., 318, 339
 Mann, H. B., 61, 95
 Mark, N. C., 459, 480
 Marriott, F. H. C., 141, 151
 Martin, R. D., 438, 440, 441, 447, 450,
 452
 McCabe, B. P. M., 117, 120, 122-124,
 126, 127, 131, 150, 151,
 205-207, 246, 394, 411, 421
 McCallum, B. T., 34, 44, 90, 95
 McCormick, W. P., 337
 McCulloch, J. H., 487, 489
 McCulloch, R. E., 406, 422, 449, 452,
 459, 461-463, 483
 McCullough, B. D., 341
 McCurdy, T. H., 464, 480, 481
 McDonald, R., 138, 151
 McFarland, J. W., 446, 453
 McLeish, D. L., 54, 95
 McLeod, A. I., 300, 301, 307
 McMahon, P. C., 446, 453
 Meese, R., 91, 95
 Melick, W. R., 243
 Mestre, R., 242
 Meyer, W., 134, 147
 Miller, R. B., 91, 94
 Min, C., 285, 295
 Miron, J. A., 367, 383
 Mizrach, B., 422
 Moore, J. B., 473, 474, 480
 Moore, M. J., 241, 246
 Morgan, M. S., 3, 7
 Moser, J., 244

 Nabeya, S., 65, 95, 120, 121, 151, 394,
 411, 422
 Nankervis, J. C., 64, 94, 95, 148, 317,
 324, 325, 327, 328, 340
 Nason, J. M., 410, 412, 420
 Neale, A. J., 407, 410, 420, 428, 430, 451
 Nelson, C. R., 3, 7, 32, 34, 42, 44, 47,
 48, 63, 73, 87-89, 95, 101,
 111, 127, 143, 144, 146, 147,
 151, 232, 268-270, 276, 294,
 400, 402, 422, 459, 484
 Nerlove, M., 246, 372, 386
 Newbold, P., 3, 7, 28, 29, 43, 77, 93,
 101, 146, 415, 422
 Newey, W. K., 80, 83, 95, 125, 162, 207,
 347, 391, 421
 Ng, S., 77, 78, 82, 83, 95, 96, 99, 108,
 109, 130, 151, 152, 157, 158,
 162, 165, 195, 431, 434, 452
 Nicholls, D. F., 131, 151
 Noh, J., 385
 Norin, N., 486, 489
 Nunes, L. C., 415, 422
 Nyblom, J., 394, 422

 Ogaki, M., 168, 196, 349, 360
 Oh, K. Y., 134, 136, 138, 151
 Ohanian, L. E., 187, 196
 Ohtani, K., 467, 468, 483
 Oke, T., 335, 340
 Orcutt, G. H., 141, 151
 Orden, D., 181, 192, 211, 222, 223, 242
 Osborn, D. R., 366, 376, 379, 380, 386
 Oseiwalski, J., 293, 307
 Osterwald-Lenum, M., 212-214, 225,
 247
 Otero, J., 414, 424

- Ouliaris, S., 199, 200, 203, 226, 247,
351, 360, 377, 408, 409, 419
- Pagan, A. R., 35, 43, 189, 196, 251, 259,
370, 381, 385, 459, 483
- Page, C. C., 470, 483
- Pan, J. J., 64, 95
- Pantula, S. G., 104–106, 113, 119, 120,
125, 131, 133, 147, 148, 151,
343–347, 359, 360
- Papell, D. H., 130, 136, 151, 401, 409,
418, 421
- Park, H. J., 112, 152
- Park, J. Y., 120, 124, 125, 127, 152,
161, 168, 196, 206–209, 247,
249, 253, 259, 333, 340, 349,
360, 491
- Park, S. B., 391, 422
- Paruolo, P., 355, 360
- Paul, M., 139, 150
- Pedroni, P., 241, 247
- Pena, D., 193
- Perron, P., 47, 48, 58, 65, 73, 75, 77–80,
82, 83, 90, 91, 93, 95, 96, 99,
100, 108, 109, 118, 121, 122,
129, 130, 145, 152, 153, 157,
158, 162, 165, 168, 169, 192,
195, 196, 199, 202, 207, 212,
217, 218, 233, 243, 247, 298,
346, 348, 364, 365, 379, 384,
386, 389, 397–403, 407–410,
414, 416–418, 420, 422, 425,
428–431, 434, 435, 452, 453,
461, 463, 481, 484
- Pesaran, M. H., 136, 137, 145, 149, 194
- Peters, S. C., 319, 320, 335, 338, 477,
482
- Petrie, T., 480
- Pham, D. T., 141, 150
- Phillips, P. C. B., 28, 29, 39, 44, 47, 48,
58, 59, 61, 62, 64, 75, 78–84,
86, 90, 91, 96, 97, 99, 100,
105, 115, 118, 120–122, 130,
145, 149, 152, 153, 156,
161–165, 173, 177–184, 194,
196, 197, 199, 200, 203, 207,
209, 218–220, 226, 241,
246–249, 253, 255–257, 259,
263, 266, 268, 270–273,
276–278, 281–283, 286, 288,
291, 294, 295, 298, 325, 333,
340, 346, 348, 351, 352, 360,
364, 365, 377, 398, 402, 406,
411, 412, 414, 422–424,
445–447, 452, 453, 490, 491
- Pierce, D. A., 19, 43, 93
- Ploberger, W., 255, 278, 281–283, 291,
294, 393–396, 411, 414, 418,
421, 423
- Plosser, C. I., 3, 7, 34, 44, 47, 48, 73,
87–89, 95, 111, 116, 127, 143,
144, 146, 151, 152, 232,
268–270, 276, 294, 400, 402,
422
- Podivinsky, J. M., 215, 219, 239, 247
- Politis, D. N., 329, 331, 341
- Pope, J. A., 141, 151
- Porter-Hudak, S., 300, 303, 307, 308
- Praschnik, J., 101, 150
- Quah, D., 135, 152
- Quandt, R., 391, 392, 394, 396, 423,
454–457, 468, 481, 484
- Quenouille, M., 18, 44, 96
- Quinn, B. G., 131, 151
- Quintos, C. E., 411–414, 423
- Rabault, G., 460, 463, 466, 484
- Raiffa, H., 265, 267, 294
- Rao, A. S., 253, 259
- Rao, C. R., 107, 152, 314, 390, 391, 423
- Rao, M. M., 3, 7, 61, 96
- Rapp, R., 466, 481
- Rappoport, P., 407, 410, 423, 430, 453
- Raymond, J. E., 464, 484
- Rayner, R. K., 315, 323–325, 341
- Reichelin, L., 410
- Reichlin, L., 407, 423, 430, 453
- Reimers, H. E., 216, 225, 247
- Reinsel, G. C., 103, 104, 154, 168, 187,
191, 192, 196, 225
- Rich, R. W., 464, 484
- Rilstone, P., 313, 315, 341
- Robinson, P. M., 300, 305, 307, 308
- Romano, J., 312, 313, 329, 331, 337, 341
- Romo, J., 325, 326, 338
- Ronchetti, E. M., 451
- Rose, A. K., 408, 423
- Rosenblatt, H. M., 390, 421
- Ross, S. L., 393, 421
- Rothenberg, T. J., 99, 113, 148, 219,
442, 444, 450, 453
- Rothman, P., 303, 306
- Rousseeuw, P. J., 441, 451, 453
- Rubin, H., 61, 96
- Rudebusch, G., 128, 141, 148, 152, 298,
302, 303, 306
- Rush, M., 229, 234, 236, 244
- Sørensen, B. E., 220, 245
- Said, S. E., 75, 76, 82, 90, 97, 100, 152
- Saikkonen, P., 120, 152, 161–163, 165,
196, 227, 247
- Salazar, D., 397, 404, 423

- Sampson, M., 91, 92, 97
 Sapsford, D., 468, 485
 Sarantis, N., 379, 386
 Sargan, J. D., 35, 43, 44, 82, 83, 92, 97,
 116, 118, 152, 159, 196
 Satchell, S. E., 64, 97
 Saunders, M. A., 470, 483
 Savin, N. E., 61, 64, 94, 95, 148, 314,
 317, 324, 325, 327, 328, 339,
 340
 Schlaifer, R., 265, 267, 294
 Schmidt, P., 39, 44, 84, 86, 97, 105,
 114–116, 120, 149, 150, 153,
 299, 307
 Schotman, P., 143, 144, 153, 241, 246
 Schotman, P. C., 267, 272, 275, 276,
 278, 294
 Schwartz, A. J., 63, 94
 Schwert, G. W., 77, 78, 81, 82, 97, 100,
 103, 116, 123, 124, 127, 152,
 153, 327, 341, 459, 483
 Schwert, W. G., 34, 44
 Sclove, S. L., 457, 484
 Scott, A., 366, 373, 379, 385
 Sellke, T., 277, 292
 Sen, D. L., 344, 360
 Sen, P. K., 393, 423
 Seo, B. S., 414, 424
 Sephton, P. S., 235, 247
 Shea, G. S., 303, 308
 Shephard, N., 117, 153, 438, 450, 477,
 482, 484
 Shiller, R., 129, 130, 153, 175, 193, 365,
 386
 Shin, Y., 120, 136, 137, 145, 149, 206,
 208, 209, 248
 Siegmund, D., 392, 394, 396, 421
 Siklos, P. L., 377, 379, 381, 382, 384–386
 Sims, C. A., 34, 44, 249, 259, 263,
 266–268, 271, 277, 278, 280,
 286, 291, 294, 364, 365, 384,
 386
 Singh, K., 311, 328, 329, 336, 340, 341
 Slutsky, E., 13, 44
 Smith, A. A., 303, 308
 Smith, A. F. M., 292, 405, 420, 459,
 462, 481, 483, 484
 Smith, G. W., 176, 192
 Smith, J., 298, 299, 308, 386, 414, 424
 Smith, R., 35, 42, 192
 Snyder, R. D., 470, 475, 484
 Solo, V., 491
 Sorenson, H. W., 475, 484
 Soules, G., 480
 Sowell, F., 270, 294, 298, 300–303, 305,
 308
 Spanos, A., 10, 44
 Srba, F., 43
 Stahel, W. A., 451
 Startz, R., 302, 303, 305, 307, 459, 484
 Steel, M., 293, 307
 Stewart, C., 379, 386
 Steyn, I. J., 475, 484
 Stock, J. H., 79, 83, 97, 99, 108, 109,
 113, 117, 139, 145, 148, 153,
 157, 161–163, 171, 172, 193,
 196, 197, 203, 218, 219, 221,
 222, 226, 244, 248, 254, 255,
 258, 259, 351, 352, 355, 356,
 361, 365, 389, 397, 401, 408,
 418, 424, 442, 444, 450, 453
 Susmel, R., 466, 482
 Swanson, N. R., 486–489
 Takanehi, Y., 259
 Takemura, A., 116, 147
 Tanaka, K., 65, 66, 95, 97, 119–121,
 151, 153, 394, 411, 422, 491
 Taqqu, M. S., 300–302, 307
 Taylor, J. M. G., 451
 Taylor, M. P., 234, 235, 237, 243, 246
 Taylor, R. L., 337
 Terasvirta, T., 486, 489
 Thornber, H., 267, 295
 Tiao, G. C., 165, 169, 181, 182, 184,
 191, 192, 211, 222, 230, 334,
 336
 Tibshirani, R., 310, 312, 313, 338, 339,
 341
 Titterton, D. M., 459, 484
 Tjøstheim, D., 485
 Toda, H. Y., 217, 220, 248, 258, 259
 Tran, H. A., 193
 Tremayne, A. R., 131, 150, 151
 Trost, R. P., 187, 195
 Tsay, R. S., 406, 422, 426, 439, 449, 452,
 453, 459, 461–463, 483
 Tsurumi, H., 467, 468, 480, 484
 Turner, C. M., 459, 484
 Tyssedal, J. S., 485
 Uhlig, H., 267, 268, 275, 285, 286,
 289–291, 294, 295
 Urbain, J. D., 188, 197, 245
 van Dijk, H. K., 294
 van Dijk, D., 143, 144, 153, 227, 245,
 451
 van Dijk, H. K., 267, 270, 272, 275, 276,
 278, 288, 289, 293
 van Giersbergen, N. P. A., 316, 317,
 321, 323–325, 338
 van Norden, S., 125, 126, 147, 485
 Varoufakis, Y., 468, 485

- Veall, M. R., 310, 314, 315, 319, 339, 341
 Vinod, H. D., 309, 333, 341
 Vogelsang, T. J., 384, 408, 409, 422, 430, 431, 434, 448, 451–453
 Vuong, Q. H., 419

 Wald, A., 61, 95
 Wallis, K. F., 44, 364, 386
 Wang, J. G., 465, 483
 Wasserman, L., 284, 293
 Watson, M. W., 139, 153, 161–163, 171, 172, 196, 197, 203, 218, 226–228, 245, 248, 259, 351, 352, 355, 356, 361, 365
 Watt, D., 410, 412, 420
 Watts, D. G., 467, 480
 Wei, C. Z., 66, 93, 280, 282, 295, 490
 Weinbach, G. C., 480
 Weisberg, S., 439, 450
 Weiss, A. A., 35, 43
 Weiss, N., 480
 West, K. D., 67, 80, 83, 95, 97, 125, 162, 207, 347, 350, 361
 West, M., 439, 453, 469, 485
 White, H., 80, 97, 318, 321, 397, 419
 White, J. S., 3, 7, 61, 97
 Whiteman, C. H., 94, 148, 268–270, 278, 292, 294
 Wickens, M. R., 37, 44, 175, 197, 251, 254, 259, 488, 489
 Wilson, S. R., 315, 316, 323, 327, 339
 Winokur, H. S., Jr., 141, 151
 Wolter, J., 298, 307

 Wooldridge, J. M., 251, 260
 Wright, J. H., 411, 424
 Wu, S., 137, 151
 Wu, Y., 134, 136, 138, 153

 Xiao, Z., 197, 219, 220, 248

 Yajima, Y., 302, 308
 Yamamoto, T., 258, 259
 Yang, C. W., 173, 194
 Yang, M., 169–171, 192, 223–225, 243, 248
 Yang, R., 284, 292
 Yao, Y. C., 415, 424
 Yap, S. F., 103, 104, 154
 Yatawara, N., 437, 438, 450
 Yeo, S., 43
 Yilmaz, K., 304, 306
 Yin, Y., 425, 428, 431, 432, 437, 452, 453
 Yohai, V. J., 440–442, 447, 452, 453
 Yoo, B. S., 185, 186, 193, 199, 244, 348, 361, 366, 385
 Yu, B. C., 117, 118, 125, 127, 131, 148, 466, 482
 Yule, G. U., 3, 7, 14, 15, 28, 44

 Zellner, A., 266, 267, 275, 285, 286, 288, 295
 Zin, S. E., 303, 306, 308
 Zivot, E., 64, 97, 203, 205, 248, 275, 276, 295, 325, 341, 397, 401, 402, 409, 424

Subject index

- Abadir test, 141
- aggregation problem
 - cointegration, 182, 190–191, 229
 - unit root test, 129, 130, 145
- Akaike Information Criterion (AIC), 19, 78, 127, 191, 217, 467
- ARCH disturbances, 324
- asymptotic convergence, 21
- asymptotic independence, 9
- augmented Dickey–Fuller (ADF)
 - choice of lag-length, 77–79
 - comparison with the PP test, 81–82
- augmented Dickey–Fuller (ADF) distribution, 75
- augmented Dickey–Fuller (ADF) regression, 75
- augmented Dickey–Fuller (ADF) test, 45–47, 98–100, 103, 122, 124, 127, 129, 130, 133, 135, 137, 140, 144, 145, 161, 162, 164, 165, 199, 202, 205, 215, 218, 228, 229, 234, 252, 276, 298, 299, 327, 351, 364, 365, 376, 410, 412, 434, 445
 - size distortions and low power, 100–103
- autocorrelation function (acf), 11
- autocovariance function (acvf), 11
- autoregressive (AR) processes, 13–15
- autoregressive distributed lag (ADL)
 - model, 35–37
- autoregressive fractionally integrated moving average (ARFIMA)
 - model, 296–298
- frequency-domain method, 301–302
- Geweke and Porter-Hudak procedure, 300–301
- ML estimation, 301
 - unit root tests, *see* unit root tests
 - against ARFIMA alternatives
- autoregressive moving-average (ARMA) processes, 15–17
- autoregressive unit root, 49
- Bardsen transformation, 36
- Bayesian inference
 - introduction, 264–266
 - Markov switching regression, 461–463
 - on Nelson–Plosser data, 268–271
 - structural change, 402–407
- Bayesian Information Criterion (BIC), 19, 78, 79, 191, 217, 280, 282, 283, 398, 406, 415, 416, 467
- Bayesian likelihood ratio (BLR), 283, 284, 291
- Bayesian methods
 - future direction, 487
- Bayesian vector autoregressive (BVAR)
 - model, 185, 187
- Beran’s pivotal method, 313
- Berger–Bernardo method, 284
- Beveridge–Nelson decomposition, 101
- Beveridge–Nelson decomposition, 31
- Bhagava test, 82–84
- Bhargava–Schmidt–Phillips (BSP) test, 86
- Bieren–Guo test, 125
- Bierens test, 140
- bootstrap
 - application in finance, 335
 - AR(1) model, 322–325
 - cointegrating regressions, 332–335
 - confidence intervals, 311–314
 - future direction, 486
 - hypothesis testing, 315–317
 - moving block, *see* moving block bootstrap (MBB)

- percentile- t method, *see* bootstrap- t method
- pre-testing problem, 335–336
- review, 309–311
- samples, 318–322
- unit root tests, 325–328
- bootstrap- t method, 311–313
- Box–Jenkins methods, 17–20
- Box–Tiao method, 46, 169–171, 334
- Breitung–Meyer test, 134–135
- Brownian motion, *see* Wiener processes
 - geometric, 51
- Burrige–Guerre test, 140
- Canova–Hansen (CH) test, 370–371
- central limit theorem (CLT), 52
- Chow test, 390–391
- cointegrated system
 - general, 155
 - triangular system, 160
 - two variable model, 156
- cointegrating regression Durbin–Watson (CRDW) test, 202
- cointegration, 26–28
- cointegration and Granger causality, 188–189
- cointegration test
 - cointegration as null, 205–210
 - effect of time disaggregation, 228–229
 - pre-testing problem, 229–231
- cointegration with exogenous variables, 187–188
- cointegration with $I(2)$ and $I(1)$ variables
 - Johansen method, 353–355
 - Kitamura model, 352
 - residual-based tests, 350–351
 - single equation methods, 348–350
 - Stock and Watson method, 351–352
- confirmatory analysis, 126–128
- continuous mapping theorem (CMT), 54
- CUSUM test, 392–393
 - OLS residual, *see* OLS-based CUSUM test
 - power, 393
 - structural change in cointegration, *see* OLS-based CUSUM test
- deterministic trend, 29–33
- detrending methods, 33
- Dickey–Fuller (DF) distribution, 62–66, 84, 104, 113, 232, 435–437
 - AO model, 433
 - IO and LC models, 434
 - Dickey–Fuller (DF) forward regression, 109
 - Dickey–Fuller (DF) regression, 107, 431–434
 - Dickey–Fuller (DF) reverse regression, 110
 - Dickey–Fuller (DF) test, 45, 47, 63, 82, 85, 86, 107, 113, 115, 126, 200, 254, 255, 300, 343, 344, 399–401, 409, 410, 431, 434, 435, 442, 443
 - additive outliers, 431
 - AO model, 432, 434
 - innovation outliers, 431
 - MM estimator, 443
 - outliers in first-differences, 430
 - VC model, 436
 - Dickey–Fuller GLS (DF–GLS) test, 113–115
 - Dickey–Pantula test, 343–346
 - Dickey–Hasza–Fuller (DHF) test, 367
 - difference-stationary process (DSP), 25, 26, 33, 34, 37, 42, 73, 74, 88, 89, 143, 144, 416
 - double unit root, 131, *see* Hasza–Fuller test, Dickey–Pantula test, Halrup test
 - Dufour test, 391
 - Durbin–Hausman test, 106–107
 - Durbin–Watson (DW) statistic, 28
 - Durbin–Watson statistic, 82, 202, 215, 252
 - dynamic generalized least squares (DGLS) method, 163–164
 - dynamic linear regression (DLR) model, 36
 - dynamic OLS (DOLS) method, 163
 - ECM tests, 203–205
 - economic identification, 174
 - Elliot–Rothenberg–Stock procedure, *see* Dickey–Fuller GLS (DF–GLS) test
 - EM algorithm, 458–459
 - empirical identification, 174
 - endogeneity of regressors, 158
 - Engle–Granger two-step method, 156–159, 410
 - Engle–Granger–Hylleberg–Lee (EGHL) tests, 375–376
 - error correction model (ECM), 35–37, 40–42, 46, 157, 162, 163, 177–180, 182, 185–187, 189, 195, 202–205, 212–215, 218, 220, 233, 242, 352, 373–376, 410
 - exchange rates dynamics

- using ARFIMA model, 303–304
 - using Markov switching model, 460–461
- Fisher test, 137
- fluctuation test, 393–394, 396
- forecast
 - ARIMA model, 20
 - cointegrated system, 184–187
 - interval forecasts, 91–92
 - TS versus DS, 90–91
- fractional unit roots, *see* ARFIMA model
- fractionally cointegrated models, 302–303
- Franses test, 368–370
- frequency of observations, 129–131
- fully modified least absolute deviation (FM-LAD) estimation, 445
- fully modified M (FM-M) estimation, 445
- fully modified OLS (FM-OLS)
 - estimation, 46, 161–162, 165, 172, 178–180, 182, 183, 256, 334, 352, 411, 414, 445, 446
- fully parametric least squares (FPLS)
 - method, 164
- functional central limit theorem (FCLT), 52
- Gaussian process, 9
- general to specific rule, 78
- generic identification, 174
- Geweke and Porter-Hudak procedure, 300–301
- gradual regime switching models, 467–469
 - Bayesian analysis, 469
- Granger causality, 188–189
- Granger representation theorem, 41
- Gregory-Hansen tests, 412
- Haldrup test, 346–348
- Hamilton model, 457
- Hansen-Johansen tests, 413–414
- Harris-Inder test, 209–210
- Hasza-Fuller test, 342–343
- Horvath-Watson test, 228
- Hylleberg-Engle-Granger-Yoo (HEGY) test, 367–368
- identification problem, 173–175
 - future direction, 488
- Im-Pesaran-Shin test, 136–137
- integrated variables, $I(k)$, 25
- invertibility condition, 12
- Ito process, 51
- IV test, 104–106
- Johansen procedure, 165–168
 - limitation, 173
 - with trends, 168–169
- Johansen tests, 211–222
 - critical values, 212–214
 - finite sample evidence, 214–220
 - maximum eigenvalue test, 212
 - nearly unit root variables, 221–222
 - spurious cointegration, 220–221
 - trace test, 211–212
- Kalman filter, 472–473
- Kleibergen-van Dijk tests, 227–228
- KPSS test, 120–122
- LAD-based tests, 444
- lag length, 77–79, *see* selection of lag length
- lag operator, 12
- lagged dependent variables, 252–254
- Lam model, 461
- levels canonical correlation analysis (LCAA) tests, 222–226
- Levin-Lin test, 135–136
- Leybourne-McCabe test, 122–124, 206–207
- local level model, 117
- locally most powerful test, *see* tests for varying parameter, 396
- long memory in stock price, 305
- long-run demand for money, 236–239
 - cointegration with $I(2)$ and $I(1)$ variables, 355–358
- LR test in ARMA models, 103–104
- MA unit roots, 49, 116
 - Choi-Yu test, 117–119
 - ML estimator, 116–117
 - spectral density test, 119
 - Tanaka test, 119–120
 - VD test, 119
- MacKinnon critical values, 199
- MAP procedure, 459
- market efficiency hypothesis (MEH), 234–235
- Markov switching regression model, 455–457
 - ARCH models, 466
 - Bayesian analysis, 461–463
 - effects of X-11, 466
 - endogenous regime shifts, 464
 - MAP procedure, 459
 - ML estimation, 458–459
 - multiple state models, 463, 464
 - three-state model, 463

- time-varying transition
 - probability, 466
 - with explanatory variables, 464
- martingale, 304, 394
- martingale difference sequences, 74, 75
- max Chow type tests, 394–396
- mean Chow test, 396–397
- mean-reverting, 23, 190, 298
- median-unbiased estimation, 141–144
- minimum MSE estimator, 144–145
- moving block bootstrap (MBB),
 - 328–329
 - optimal length of blocks, 331–332
 - problems, 329–330
- moving-average (MA) processes, 12–13
- moving-average (MA) unit root
 - Arellano–Pantula test, 125
 - Choi test, 125
- multiple structural breaks, 405–407
- Nabeya–Tanaka LM test, 121
- near unit root, 66, 298
- nearly stationary, 157
- Newey–West estimator, 83, 162, 207, 347
- Newey–West estimator of σ^2 , 80
- nonlinear least squares (NLS) method, 163
- nonlinear models
 - future direction, 486
- nonstationary, 20
- normalization, 46, 60, 107, 156, 157, 162, 165, 168, 189, 211, 226, 227, 250, 289, 413
- OLS-based CUSUM test, 393
 - nonstationary variables, 411
- Ornstein–Uhlenbeck process, 65
- outliers
 - additive outliers (AO), 426–428, 430–434, 437–441, 448
 - detection, 437–439
 - different types, 425–428
 - effects on unit root tests, 428–437
 - innovation outliers (AO), 426–428, 431, 432, 434–439, 441, 448
 - level changes (LC), 426–428, 434–437, 439, 448
 - transient changes (TC), 426, 439
 - variance changes (VC), 426–428, 436, 437, 439, 448
- overdifferencing, 49, 89
- overdifferencing, 116
- panel data unit root, 133–134
 - power of tests, 137
 - the hypothesis of interest, 138–139
- Park *J* test, 124–125, 207–208
- periodic
 - autoregression, 379–380
 - cointegration, 381
 - integration, 380–381
 - models, 381
 - unit roots tests, 381
- Phillips–Perron (PP) test, 45, 46, 79–81, 83, 86, 92, 98–100, 103, 107–109, 122, 124, 126, 129, 130, 140, 145, 161, 162, 364, 434, 443, 444
 - comparison with the ADF test, 81–82
 - modifications, 107–109
 - size distortions and low power, 100–103
- Phillips–Perron test, 83
- PICF, 283, 291
- posterior
 - distribution, 264–270, 279–281, 283, 291
 - odds, 276–281, 283, 284, 291
- posterior information criterion (PIC), 282, 283, 291
- pre-testing problem, 46, 100, 139–140, 229–231, 242, 255, 286, 290, 335, 336, 417, 487
 - future direction, 486
- predictive
 - approach, 278
 - density, distribution, 280–283, 287
 - odds, 281, 283, 291
 - sample reuse (PSR), 281
- principal components and cointegration, 171–172
- principal components method, 226
- prior
 - Berger–Bernardo, 284
 - conjugate, 264, 265
 - diffuse, 264, 266, 270, 282, 289, 292
 - distribution, 264, 265, 279
 - flat, 263, 265–269, 271–273, 278, 288, 291
 - ignorance, 263, 266, 271, 272
 - improper, 265
 - Jeffreys', 263, 266, 271–276, 284, 286, 289, 291, 295
 - maximal data information priors (MDIP), 285
 - Normal–Wishart, 286, 289
 - Wishart, 289
- purchasing power parity (PPP), 126, 134, 138, 156, 232, 233, 236, 237, 239–241, 355, 358, 408
- purely random process, 11

- Q test, 19–20
- Quandt test, 391–392
- Quintos tests, 413–414
- Quintos-Phillips tests, 411–412
- random walk, 20, 24, 27, 28, 33, 48, 50, 52, 54, 57, 67, 82, 83, 87, 102, 123, 160, 190, 214, 303, 408, 415, 416, 430, 461, 463, 465, 479
 - continuous, 51
 - parameters, 120, 394, 397, 411, 469
 - weak, 157
 - with a change in the drift, 428
 - with drift, 29, 31, 33, 38, 67, 70, 101, 372, 374, 430
 - with two states (Markov switching), 460
- reduced rank regression, 168
- regression with $I(1)$ and $I(0)$ variables
 - $I(1)$ regressors cointegrated, 250–251
 - $I(1)$ regressors not cointegrated, 249–250
- residual-based tests for cointegration, 198–203
 - Cochrane-Orcutt version, 202–203
- robust estimation of cointegrating regressions, 445–447
- robust methods
 - future direction, 486
- robust unit root tests, 440–445
- Saikkonen tests, 227
- Sargan-Bhargava test, 82
- Sargan-Bhargava test, 82
- scaling factors, 59–60
- Schmidt-Phillips LM test, 84–86
- seasonal adjustment, 364–365
- seasonal cointegration, 375–376
 - estimation, 376–378
- seasonal integration, 365–366
- seasonal processes
 - integrated, 363
 - purely deterministic, 363
 - stationary, 363
- seasonal unit roots, 366–371, 448
- selection of lag length, 164–165
- Seo tests, 414
- Shin test, 208–209
- signal-noise ratio, 178, 184
- single equation ECM (SEECM)
 - method, 162–163
- skip sampling problem
 - cointegration, 182, 229
 - seasonal unit root, 382
 - unit root, 129–131
- Slutsky effect, 13
- smooth estimates, 470
- smooth inferences, 458
- smooth probabilities, 458
- smoothing, 473–474
- spurious breaks, 415–416
- spurious periodicity, 34
- spurious regression, 28–29
- STAMP program, 477
- state-space models, 469
 - general form, 470–471
 - GLS and Kalman's recurrence relation, 471–472
- stationarity, 10–11
 - covariance, 24
- stochastic process, 8
- stochastic trend, 29–33
- Stock and Watson method, 172
- structural change
 - Bayesian approach, 402–407
 - effects on cointegration tests, 410
 - empirical works, 407–410
 - exact tests, 414–415
 - in cointegrated relationships, 411–414
 - seasonal models, 414
 - tests with known break, 390–391
 - tests with unknown break, 391–397
 - unit root tests, *see* unit root tests under structural change
- structural time series model, 475–477
 - future direction, 487
 - ML estimation, 477
 - program for estimation, *see* STAMP program
- superconsistency, 157
- switching regression model, 454–455
- systematic sampling, 381–382
- tests for TSP versus DSP, 87–90
- tests for varying parameter, 394
- threshold cointegration, 190
- time aggregation, 381–382
- trend-stationary process (TSP), 25, 26, 33, 34, 37, 42, 73, 74, 88, 89, 143, 144, 416
- unbalanced equations, 251–252
- uncertain unit roots, 139–140, 254–256
- uncertain unit roots and cointegration, 256–258
- uncovered interest parity (UIP), 240
- underdifferencing, 33, 89
- unit root tests
 - against ARFIMA alternatives, 298–300

- bootstrapping, *see* bootstrap,
 - unit root tests
- LAD estimator, 444
- low power, 100–103
- M estimator, 444
- rank-based method, 445
- size distortion, 100–103
- stationary as null, 120–126
- under structural change, 399–402
- using cointegration, 231–233
- unobserved components (UC) model,
 - 117, 371–375
- variance ratio (VR) test, 86–87
- vector autoregressive (VAR) model, 34,
 - 35, 42, 155, 156, 164, 165,
 - 172, 174, 175, 182, 184–187,
 - 195, 212, 215–217, 220, 227,
 - 232, 233, 238, 248, 257, 258,
 - 285, 287, 288, 290, 292, 304,
 - 334, 338, 353, 356, 357, 365,
 - 376, 381, 478, 487, 488
- weighted symmetric (WS) estimator,
 - 111–113
- white noise, 12
- wide class of errors, 74
- Wiener process, 50–51
 - generalized, 51
 - relationships with normal distributions, 56–59
 - some basic results, 54–56
- X-11 filter, *see* seasonal adjustment
- Yule–Walker equations, 14, 15