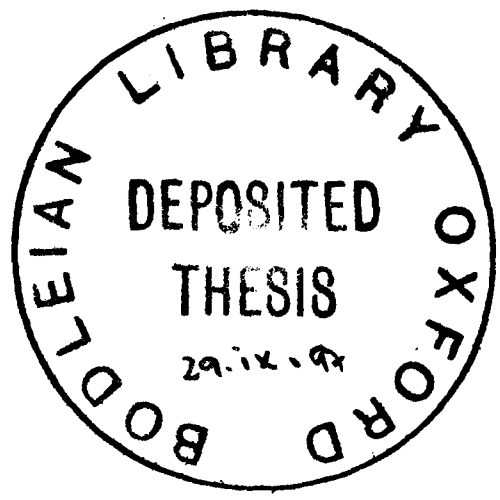
Adjusted Profile Likelihood Applied to Estimation and Testing of Unit Roots

Pekka Pere

St. Antony's College
Oxford

Hilary Term 1997



Submitted for the Degree of Doctor of Philosophy to the Faculty of Social Studies, University of Oxford

To my family

Acknowledgements

I have benefited tremendously from the help of a number of people to whom I wish to express my sincere gratitude. It is a great pleasure to thank:

Professor **David Hendry** for pointing out to me the potential of adjusted profile likelihood for unit-root econometrics and for insightful supervision. His excitement in econometrics, in discovering new things, and in sharing ideas with students have left a permanent impact on me.

Sir **David Cox** for giving me his manuscript with Nancy Reid before the publication of it 1993. Much of the theory in this thesis derives from it. I am grateful to him for useful discussions on the subject as well.

Professor **Clive Granger**, Professor **Søren Johansen**, Dr. **Ilkka Mellin**, Dr. **Neil Shephard**, and especially Dr. **Pentti Saikkonen** for helpful comments on the manuscript at its various stages. I have been fortunate to benefit from their advice. I also appreciate Dr. Saikkonen for always finding the time to discuss econometrics in general. His insight has aided me greatly throughout my post-graduate studies.

Dr. **Markku Rahiala** for comments and for the exercise he made for me pointing out how a large starting value can distort inference on unit roots.

Professor **Seppo Mustonen** for supplying his versatile statistical software SURVO and quickly writing new code for analysis of extremely large data sets when I needed it.

Dr. **Jürgen Doornik** for sharing his helpful TEX styles.

Professor **Timo Teräsvirta** for supervising my Master's thesis and thereby smoothing my path to the D.Phil. degree.

Dr. **Lars-Erik Öller** for drawing my attention to unit-root econometrics and time-series analysis in general. I appreciate his openheartedness and shall try to follow his example.

Participants of the Econometrics seminar at Nuffield College, University of Oxford, Econometric Theory seminar at the Department of Economics, University of Exeter, and **Winter Symposium of the Econometric Society, Lisbon** for comments and encouragement.

The Department of Economics at the University of Helsinki for first equipping me with excellent facilities and then accepting me as a colleague in the department to complete my thesis.

The Finnish Cultural Foundation, the Finnish Academy, and particularly **the Yrjö Jahnsson Foundation** for financing my postgraduate studies.

My parents **Kirsti** and **Aimo Pere** for affectionately supporting me throughout my life and providing me with such a good background for studies.

My dear wife **Anneli** and our lovable children **Sampo**, **Tuulia**, and **Ilpo**, all born during this research, for their love and bringing depth to life. I am fortunate to have a wife who is not only excited to go abroad to share adventures but who can also cope with a sometimes absent minded husband. We have been blessed with a memorable time in beautiful Oxford which we shall treasure for ever.

I am solely responsible for the hopefully few errors in the thesis.

Adjusted Profile Likelihood Applied to Estimation and Testing of Unit Roots

Pekka Pere

St. Antony's College
Oxford

D.Phil. in Economics
Hilary Term 1997

Abstract

A short review of unit-root econometrics is given from the point of view of testing. The adjusted likelihoods of Cox and Reid (1987, 1993) are presented and applied to the usual AR(1) with constant, an AR(1) process suggested by Bhargava (1986), and an AR(2) process. Biases of the associated maximum-likelihood estimates (MLEs) are pondered briefly. A Wald statistic based on adjusted profile likelihood is proposed.

The Cox–Reid adjusted estimate (AE) for the autoregressive coefficient of the unit-root AR(1) model with zero constant is even asymptotically more accurate, in terms of mean-square error (MSE), than the MLE. The derived tests are more powerful than the corresponding Dickey–Fuller tests if the starting value of the process deviates sufficiently from the unconditional mean. An iteratively adjusted estimate is introduced which can also be more accurate than the MLE. We obtain also an estimate and a Wald statistic which are asymptotically distributed compactly and symmetrically around zero under a unit root but the estimate is not consistent in general.

The MLE and the AE are consistent not only as the sample size tends to infinity but also when the (absolute value of the) deviation of the starting value from the unconditional mean of the time series is tuned towards infinity. The finding exposes why Wald-kind of tests are more powerful than tests based on standardised coefficients when the starting value lies far from the unconditional mean.

The AE and the corresponding Wald statistic are derived for the Bhargava AR(1) model. We obtain the asymptotic distributions of them and simulate the previously unknown finite sample distributions of the MLE and the usual Wald statistic under a unit root. Again the AE is the more accurate estimate. Distortion towards a unit root is pointed out.

The adjusted estimate and the Wald statistic follow their asymptotic distributions better than the unadjusted when the process is a unit-root AR(1) with drift or the Bhargava AR(1). Accuracy is gained also under the unit-root AR(2) model.

A practical advice is to apply a unit-root test based on the Bhargava model when the process can be assumed to have started from the unconditional mean under the alternative and otherwise a test based on the ordinary AR(1) with constant model.

The adjustment often decreases the bias at the cost of variance but it can yield a reduction in both, too, which happens under the Bhargava model and 'typically' under the unit-root AR(2) model.

The two most distinctive findings are perhaps that the AE can be asymptotically more accurate than the corresponding MLE or in finite samples when the AE is calculated from an embedding model.

Contents

1 Introduction **1**

2 The Framework of Unit-Root Econometrics **4**
2.1 Economic Time Series and the Process Generating them 4
2.2 A Sketch of Unit-Root Econometrics and its Historical Background . 6

3 Testing for Unit Roots **9**
3.1 Two Classes of Processes 9
3.2 Dickey–Fuller Tests 14
3.3 Augmented Dickey–Fuller Tests 19
3.4 The Generalization of Said and Dickey 20
3.5 Problems with Testing for Unit Roots 22

4 Likelihood Theory **29**
4.1 Likelihood and Profile Likelihood 29
4.2 Adjustments to Profile Likelihood 34
4.3 Examples of Adjusted Profile Likelihood 41

5 On Biases in the Estimation of Autoregressive Models **52**
5.1 Introductory Remarks 52
5.2 Incitements for Bias Corrections 52
5.3 Biases in the Estimation of Stationary Autoregressive Models 54
5.4 Biases in the Estimation of Nonstationary Autoregressive Models . . 61
5.5 Side Effects of Bias Corrections 63

6 Application of Adjusted Profile Likelihood to Some AR Models **66**
6.1 Introductory Remarks 66
6.2 Adjusted Profile Likelihood for an AR(1) Model 67
6.3 Adjusted Profile Likelihood for an AR(1) Model with Constant . . . 72
6.4 Adjusted Profile Likelihood for a Unit-Root AR(1) Model with Drift 105
6.5 Adjusted Profile Likelihood for the Bhargava AR(1) Model with Constant 111
6.6 Adjusted Profile Likelihood for a Unit-Root AR(2) Model 124

7 Finite-Sample Properties **141**
7.1 Introductory Remarks 141
7.2 Finite-Sample Distributions for the Unit-Root AR(1) Model with Constant 142
7.3 Power Comparisons for the AR(1) Model with Constant 151
7.4 Finite-Sample Distributions for the Unit-Root AR(1) Model with Drift 165
7.5 Fractiles for the Unit-Root Bhargava AR(1) Model with Constant . . 174

7.6 Power Comparisons for the Bhargava AR(1) Model with Constant . . 180
7.7 Finite-Sample Results for the Unit-Root AR(2) Model 187
7.8 Comparison to the Optimally Multiplicatively Corrected Estimates . 195

8 Discussion **200**
8.1 General Comments 200
8.2 Discussion on Estimation and Related Distributional Results 204
8.3 Discussion on Testing 211
8.4 Suggestions for Theoretical Research 215
8.5 Suggestions for Empirical Research 218

A1 Proofs for the AR(1) Model with Constant **221**

A2 Proofs for the Unit-Root AR(1) Model with Drift **230**

A3 Proofs for the Bhargava AR(1) Model with Constant **236**

A4 Proofs for the AR(2) Model **244**

A5 Starting Value Asymptotics **252**

A6 Simulation Techniques **262**
A6.1 General Notes 262
A6.2 Generation of the Random Numbers 262
A6.3 Calculation of the Confidence Intervals 263
A6.4 A Simulation Programme for the AR(1) Model with Constant 264
A6.5 A Simulation Programme for the Bhargava AR(1) Model with Constant 270
A6.6 A Simulation Programme for the Unit-Root AR(2) Model 279

References **286**

List of Figures

5.1 The stationarity triangle, the zero-bias line, and the unit-root line for ρ_1 and ρ_2 of the AR(2) model 58

6.1 Empirical asymptotic distribution of $T(\hat{\psi}_{ml} - 1)$ for $\psi = 1$ (AR(1)) . . . 70
6.2 Empirical asymptotic distribution of the t-value for $\psi = 1$ (AR(1)) . . 71
6.3 A realisation from the random-walk process 72
6.4 Empirical asymptotic distribution of $T(\hat{\psi}_{ml} - 1)$ for $\psi = 1$ ($\text{AR}_\mu(1)$) . . 75
6.5 Empirical asymptotic distribution of the t-value for $\psi = 1$ ($\text{AR}_\mu(1)$) . . 76
6.6 Empirical asymptotic distribution of $T(\hat{\psi}_{ad2,ap} - 1)$ for $\psi = 1$ ($\text{AR}_\mu(1)$) 89
6.7 Empirical asymptotic distribution of the data-based $\hat{c}$ for $\psi = 1$ ($\text{AR}_\mu(1)$) 91
6.8 Empirical asymptotic distribution of $T(\hat{\psi}_{ad2} - 1)$ for $\psi = 1$ ($\text{AR}_\mu(1)$) . . 92
6.9 Empirical asymptotic joint distribution of $T(\hat{\psi}_{ml} - 1)$ and $T(\hat{\psi}_{ad2} - 1)$ for $\psi = 1$ ($\text{AR}_\mu(1)$) 93
6.10 Empirical asymptotic distribution of $T(\hat{\psi}_{ad2,i} - 1)$ for $\psi = 1$ ($\text{AR}_\mu(1)$) . 97
6.11 Empirical asymptotic distribution of $\sqrt{W_{ad2,ap}}$ for $\psi = 1$ ($\text{AR}_\mu(1)$) . . . 101
6.12 Empirical asymptotic distribution of $\sqrt{W_{ad2}}$ for $\psi = 1$ ($\text{AR}_\mu(1)$) 102

7.1 Empirical distribution of $\hat{\psi}_{ml}$ for $\psi = 1$ and $T = 100$ ($\text{AR}_\mu(1)$) 148
7.2 Empirical distribution of $\hat{\psi}_{ad2}$ for $\psi = 1$ and $T = 100$ ($\text{AR}_\mu(1)$) 148
7.3 Empirical distribution of $\hat{\psi}_{ad2,i}$ for $\psi = 1$ and $T = 100$ ($\text{AR}_\mu(1)$) 149
7.4 Empirical distribution of the data based $\hat{c}$ for $\psi = 1$ and $T = 100$ ($\text{AR}_\mu(1)$) 150
7.5 Empirical joint distribution of $T(\hat{\psi}_{ml} - 1)$ and $T(\hat{\psi}_{ad2} - 1)$ for $\psi = 1$ and $T = 100$ ($\text{AR}_\mu(1)$) 150
7.6 Empirical distribution of $\hat{\psi}_{ml}$ for $\psi = 1$ and $T = 100$ ($\text{AR}^B_\mu(1)$) 179
7.7 Empirical distribution of $\hat{\psi}_{ad}$ for $\psi = 1$ and $T = 100$ ($\text{AR}^B_\mu(1)$) 179
7.8 Empirical distribution of $\hat{\psi}_{ml}$ for $\psi = 1$, $\phi_1 = 0.95$, and $T = 100$ (AR(2)) 189
7.9 Empirical distribution of $\hat{\psi}_{ml}$ for $\psi = 1$, $\phi_1 = -0.95$, and $T = 100$ (AR(2)) 189
7.10 Empirical distribution of $\hat{\psi}_{ml}$ for $\psi = 1$, $\phi_1 = 0.5$, and $T = 100$ (AR(2)) 196
7.11 Empirical distribution of $\hat{\psi}_{ad}$ for $\psi = 1$, $\phi_1 = 0.5$, and $T = 100$ (AR(2)) 196

A5.1 Variance of the standardised MLE as a function of ψ for $|y_0|$ large and $T = 25$ ($\text{AR}_\mu(1)$) 256
A5.2 Residual variances with two minima for $|x_0|$ large, $T = 100$, and $\psi_0 = -0.3$ or $\psi_0 = 0.3$ ($\text{AR}^B_\mu(1)$) 260

A5.3 Residual variance with two minima for $|x_0|$ large, $T = 100$, and $\psi_0 = 0$ ($\mathrm{AR}_\mu^B(1)$ with x_0 incorporated in the model) 260

A6.1 An example of a bimodal profile log-likelihood ($\mathrm{AR}_\mu^B(1)$) 273
A6.2 An example of adjusted profile log-likelihood ($\mathrm{AR}_\mu^B(1)$) 273

List of Tables

4.1 Suitable classical tests under multimodal profile likelihoods with a mode at $H_0 : \psi = \psi_0$ 40

7.1 Empirical fractiles of $T(\hat{\psi}_{ad2} - 1)$ for $\psi = 1$ ($\mathrm{AR}_\mu(1)$) 144

7.2 Differences between the empirical fractiles of $T(\hat{\psi}_{ad2} - 1)$ and $T(\hat{\psi}_{ml} - 1)$ for $\psi = 1$ ($\mathrm{AR}_\mu(1)$) 144

7.3 Empirical fractiles of $T(\hat{\psi}_{ad2,i} - 1)$ for $\psi = 1$ ($\mathrm{AR}_\mu(1)$) 146

7.4 Differences between the empirical fractiles of $T(\hat{\psi}_{ad2,i} - 1)$ and $T(\hat{\psi}_{ml} - 1)$ for $\psi = 1$ ($\mathrm{AR}_\mu(1)$) 146

7.5 Empirical fractiles of $T(\hat{\psi}_{ad2,ap} - 1)$ for $\psi = 1$ ($\mathrm{AR}_\mu(1)$) 146

7.6 Empirical fractiles of $\sqrt{W_{ad2}}$ for $\psi = 1$ ($\mathrm{AR}_\mu(1)$) 152

7.7 Empirical fractiles of $\sqrt{W_{ad2,ap}}$ for $\psi = 1$ ($\mathrm{AR}_\mu(1)$) 152

7.8 Empirical power ($\times 100$) of the ψ_μ, $\psi_{\mu,ad}$, and $\psi_{\mu,ad,i}$ tests when size is 5 % 160

7.9 Empirical power ($\times 100$) of the τ_μ and $\tau_{\mu,ad}$ tests when size is 5 % . . . 161

7.10 Empirical power ($\times 100$) of the $\psi_{\mu,ad,ap}$ and $\tau_{\mu,ad,ap}$ tests when size is 5 % 162

7.11 Suggested test for different values of $|y_0 - \mu|/\sigma$ for $|y_0 - \mu|/\sigma$ in the range [0,250] and T in the range [25,250] ($\mathrm{AR}_\mu(1)$) 164

7.12 Suggested test for different values of $|y_0 - \mu|/T\sigma$ for $|y_0 - \mu|/\sigma$ in the range [0,250] and T in the range [25,250] ($\mathrm{AR}_\mu(1)$) 164

7.13 Deviations of the standardised empirical fractiles of the SCs from the Standard Normal (unit-root $\mathrm{AR}_\mu(1)$ with drift) 170

7.14 Deviations of the standardised empirical fractiles of the SCs from the Standard Normal (unit-root $\mathrm{AR}_\mu(1)$ with drift) 171

7.15 Empirical size of the SC tests when nominal size is 5 % (unit-root $\mathrm{AR}_\mu(1)$ with drift) 172

7.16 Empirical size of the Wald tests when nominal size is 5 % (unit-root $\mathrm{AR}_\mu(1)$ with drift) 172

7.17 Empirical fractiles of $T(\hat{\psi}_{ml} - 1)$ for $\psi = 1$ ($\mathrm{AR}^B_\mu(1)$) 176

7.18 Differences between the empirical fractiles of the MLEs under the AR(1) and $\mathrm{AR}^B_\mu(1)$ models for $\psi = 1$ 176

7.19 Differences between the empirical fractiles of the MLEs under the $\mathrm{AR}_\mu(1)$ and $\mathrm{AR}^B_\mu(1)$ models for $\psi = 1$ 176

7.20 Empirical fractiles of $T(\hat{\psi}_{ad} - 1)$ for $\psi = 1$ ($\mathrm{AR}^B_\mu(1)$) 177

7.21 Differences between the empirical fractiles of $T(\hat{\psi}_{ml} - 1)$ and $T(\hat{\psi}_{ad} - 1)$ for $\psi = 1$ ($\mathrm{AR}^B_\mu(1)$) 177

7.22 Empirical fractiles of $\sqrt{W}$ for $\psi = 1$ ($\mathrm{AR}^B_\mu(1)$) 178

7.23 Empirical fractiles of $\sqrt{W_{ad}}$ for $\psi = 1$ ($\mathrm{AR}^B_\mu(1)$) 178

7.24 Empirical power (×100) of the ψ_{μ}^{B} and $\psi_{\mu,ad}^{B}$ tests when size is 5 % . . . 184
7.25 Empirical power (×100) of the τ_{μ}^{B} and $\tau_{\mu,ad}^{B}$ tests when size is 5 % . . . 185
7.26 Suggested test for different values of $|x_0|/T\sigma$ for $|x_0|/\sigma$ in the range [0,250] and T in the range [25,250] ($AR_{\mu}^{B}(1)$) 186
7.27 Suggested test for different values of $|x_0|/T\sigma = [y_0 - \alpha/(1-\psi)]/T\sigma$ for $|x_0|/\sigma$ in the range [0,250] and T in the range [25,250] ($AR_{\mu}(1)$ and $AR_{\mu}^{B}(1)$) 187
7.28 Empirical size of the Wald statistics when nominal size is 5 % (AR(2)) 188
7.29 Empirical bias (×10) of the MLE, the AE, and the SAE (AR(2)) 193
7.30 Empirical SD (×10) of the MLE, the AE, and the SAE (AR(2)) 194
7.31 Empirical MSE (×1000) of the MLE, the AE, and the SAE (AR(2)) . . 194
7.32 The empirical reduction in MSE (per cent) of the AE relative to the MLE (AR(2)) 194
7.33 The range of the biases, the SDs, and the MSEs (× 100) of the AEs for $T = 100$ under the AR(2) model (with $\phi_1 \in [-0.95, 0.95]$) and the other models (when the process is a random walk) 195
7.34 The bias, the SD, and the MSE (× 100) of the AE and the OMCE for a random walk and $T = 100$ 199

8.1 The order of magnitude of the Cox–Reid adjustment to the standardised MLE 209

A6.1 The seeds used in the simulation of the fractiles in Tables 7.1, 7.3, 7.5, 7.17, 7.20, 7.33 and 7.34 (excluding the results for the AR(2) model) . . 263
A6.2 The seeds used in the simulation of powers in Tables 7.8, 7.9, 7.10, 7.24, and 7.25 263
A6.3 The seeds used in the simulation of the entries in Tables 7.13 and 7.14 . 263
A6.4 The seeds used in the simulation of the entries in Tables 7.28–7.34 for the AR(2) model 264
A6.5 Per cents of the cases where the residual variance took a smaller value under the AE than the MLE when simulating the unit-root distributions ($AR_{\mu}^{B}(1)$) 271
A6.6 Per cents of the cases where the residual variance took a larger value under the AE than the MLE when simulating powers of the tests ($AR_{\mu}^{B}(1)$) . 272
A6.7 Per cent of the draws in which the AE or the SAE were substituted a value of minus unity (AR(2)) 279
A6.8 Per cent of the draws in which the AE or the SAE were substituted a value of unity (AR(2)) 279

Notation

Abbreviations

AD	A quantity based on **Ad**justed profile likelihood[1]
AE	**A**djusted **E**stimate
AE_{ap}	**A**djusted **E**stimate which makes use of the *a priori* information $c = (T-1)/2$
AE_i	**I**teratively calculated **A**djusted **E**stimate
AR	**A**uto**r**egressive
ARIMA	**A**uto**r**egressive **I**ntegrated **M**oving **A**verage
ARMA	**A**uto**r**egressive **M**oving **A**verage
DF	**D**ickey–**F**uller
$I(d)$	**I**ntegrated of order **d**
IID	**I**dentically **I**ndependently **D**istributed
LM	**L**agrange **M**ultiplier
LR	**L**ikelihood **R**atio
MA	**M**oving **A**verage
Med	**Med**ian
ML	**M**aximum **L**ikelihood
MLE	**M**aximum-**L**ikelihood **E**stimate
MSE	**M**ean-**S**quare **E**rror
N	**N**ormally distributed
NID	**N**ormally **I**ndependently **D**istributed
OLS	**O**rdinary **L**east **S**quares
OMCE	**O**ptimally **M**ultiplicatively **C**orrected **E**stimate
SAE	**S**hortcut **A**djusted **E**stimate
SC	**S**tandardised **C**oefficient
SD	**S**tandard **D**eviation
Skew	**Skew**ness

Notational Conventions

Observations of univariate random variables are written with lower-case letters. The random variables themselves are denoted by upper-case letters. Italic type font is used for both cases. Bolded lower-case letters stand for vectors in general. However, bolded upper-case letters are used to emphasize that the vector is a random variable. Matrices are denoted by bolded upper-case letters. Heads of sections or tables, say, may be printed in bold so these rules do not necessarily apply there.

[1] E.g. AD tests stand for unit-root tests based on adjusted profile likelihood.

Symbols

Parameters are denoted by Greek letters. The ones having a very general meaning are:

$\boldsymbol{\lambda}$	Vector of nuisance parameters which has been orthogonalized with respect to ψ
$\boldsymbol{\phi}$	Vector of nuisance parameters (not necessarily orthogonal with respect to ψ)
ψ	Scalar parameter of interest
$\boldsymbol{\omega}$	Vector encompassing the parameter of interest and the nuisance parameters

Estimates of parameters are expressed as follows. The list should make the practice of notation clear but it does not include all variations of the theme.[2]

$\hat{\boldsymbol{\lambda}}_{\psi}$	Maximum-likelihood estimate of $\boldsymbol{\lambda}$ assuming a known value of ψ
$\hat{\boldsymbol{\phi}}_{\psi}$	Maximum-likelihood estimate of $\boldsymbol{\phi}$ assuming a known value of ψ
$\hat{\boldsymbol{\phi}}_{ml}$	Maximum-likelihood estimate of $\boldsymbol{\phi}$
$\hat{\psi}_{ad}$	Adjusted maximum-likelihood estimate of ψ (assumes orthogonality of ψ with respect to the nuisance parameter)
$\hat{\psi}_{ad2}$	Adjusted maximum-likelihood estimate of ψ (does not assume orthogonality of ψ with respect to the nuisance parameter)
$\hat{\psi}_{ad2,ap}$	As $\hat{\psi}_{ad2}$ but the *a priori* information $\hat{c}= (T-1)/2$ is employed in the calculation of the estimate
$\hat{\psi}_{ad2,i}$	Iteratively calculated $\hat{\psi}_{ad2}$
$\hat{\psi}_{ols}$	Ordinary least-squares estimate of ψ
$\hat{\rho}_{b}$	Bias-corrected estimate of ρ
$\hat{\boldsymbol{\omega}}_{ml}$	Maximum-likelihood estimate of $\boldsymbol{\omega}$

[2]We make an exception to the practise by allowing for a 'hat' to occur above two entities which are not a parameters by following the notation of Dickey and Fuller (1979). They denote by $\hat{\tau}_{\mu}$ and $\hat{\tau}_{\tau}$ the t statistic for the null of a unit root when the model is AR(1) with constant or with constant and time trend, respectively.

Tests and test statistics are denoted as follows:

ψ_μ	Unit-root test based on the statistic $T(\hat{\psi}_{ml} - 1)$ under the AR(1) model with constant
$\psi_{\mu,ad}$	Unit-root test based on the statistic $T(\hat{\psi}_{ad2} - 1)$ under the AR(1) model with constant
$\psi_{\mu,ad,i}$	Unit-root test based on the statistic $T(\hat{\psi}_{ad2,i} - 1)$ under the AR(1) model with constant
$\psi_{\mu,ad,ap}$	Unit-root test based on the statistic $T(\hat{\psi}_{ad2,ap} - 1)$ under the AR(1) model with constant
τ_μ	Unit-root test based on the statistic $\sqrt{W}$ or $\hat{\tau}_\mu$ under the AR(1) model with constant (i.e. the statistics are essentially the same so a common symbol for the tests is used)
$\tau_{\mu,ad}$	Unit-root test based on the statistic $\sqrt{W_{ad2}}$ under the AR(1) model with constant
$\tau_{\mu,ad,ap}$	Unit-root test based on the statistic $\sqrt{W_{ad2,ap}}$ under the AR(1) model with constant
ψ_μ^B	Unit-root test based on the statistic $T(\hat{\psi}_{ml} - 1)$ under the Bhargava AR(1) model with constant
$\psi_{\mu,ad}^B$	Unit-root test based on the statistic $T(\hat{\psi}_{ad} - 1)$ under the Bhargava AR(1) model with constant
τ_μ^B	Unit-root test based on the statistic $\sqrt{W}$ under the Bhargava AR(1) model with constant
$\tau_{\mu,ad}^B$	Unit-root test based on the statistic $\sqrt{W_{ad}}$ under the Bhargava AR(1) model with constant
$\hat{\tau}_\mu$	t statistic for the null of a unit root under the AR(1) model with constant
W	Wald statistic
W_{ad}	Wald statistic calculated from adjusted profile log-likelihood
$W_{ad,s}$	Wald statistic calculated from adjusted profile log-likelihood in a shortcut way
$W_{ad,ap}$	Wald statistic calculated from adjusted profile log-likelihood employing *a priori* information of a unit root

Other symbols include:

AD_μ	Asymptotic distribution of $T(\hat{\psi}_{ad2} - 1)$ when the model is AR(1) with constant (as defined in Section 6.3) and $\psi = 1$
$AD_{\mu,ap}$	As AD_μ but *a priori* information $\psi = 1$ is used in the construction of $\hat{\psi}_{ad2}$
$AD_{\mu,i}$	As AD_μ but for the iterative estimate $\hat{\psi}_{ad2,i}$
$\mathrm{AR}_\mu(1)$	AR(1) model with constant (as defined in Section 6.3)
$\mathrm{AR}_\mu^B(1)$	AR(1) model with constant due to Bhargava (as defined in Section 6.5)
c	Coefficient c of the Cox–Reid (1993) formula (4.12)
$\hat{c}$	Coefficient c evaluated at $\hat{\omega}_{ml}$
$\hat{c}_i$	Coefficient c evaluated at $\hat{\psi}_{ad2}$ (the iteratively estimated c)
d	$\sum_{t=1}^T y_t \Delta y_{t-1} / \sum_{t=1}^T (\Delta y_{t-1})^2$
d_{-1}	$\sum_{t=1}^T y_{t-1} \Delta y_{t-1} / \sum_{t=1}^T (\Delta y_{t-1})^2$
DF_μ	Asymptotic distribution of $T(\hat{\psi}_{ml} - 1)$ when the model is AR(1) with constant (as defined in Section 6.3) and $\psi = 1$
$l_{ad2}(\psi)$	The adjusted profile log-likelihood of Cox and Reid (1993)
$l_p(\psi)$	Profile log-likelihood
$l_p(\psi, \hat{\phi}_\psi)$	Same as $l_p(\psi)$ but the dependence on the data is made more explicit
$O_p(\cdot)$	Order in probability
$plim$	Limit in probability
r_i	Autocorrelation at lag i (after start-up effects have faded)
$\hat{r}_i$	Sample autocorrelation at lag i
$r_{1,T}$	Remainder term of a first order Taylor-series expansion
$r_T,\ r_T^*$	Remainder terms of approximations
s^2	Degrees-of-freedom corrected estimate of residual variance
t	Time index
T	Sample size or last observation in a time series
$tr\,[\cdot]$	Trace of matrix $[\cdot]$
$\bar{y}$	$T^{-1} \sum_{t=1}^T y_t$
$\bar{y}_{-1}$	$T^{-1} \sum_{t=1}^T y_{t-1}$
$W(r)$	Wiener or Brownian motion process with variance r.
$W_*(r)$	$W(r) - \int_0^1 W(r)dr$ (demeaned Wiener or Brownian motion process)
$\hat{l}$	Quantity (context dependent) which converges in probability to 1 as the sample size increases and the process is nonstationary
$\sqrt{\cdot}$	$\sqrt{(T-3)^2/4\ \hat{c}^2 + \hat{\psi}_{ml}^2} - \hat{l}$
$\xrightarrow[p]{}$	Tends in probability
$x_T \xrightarrow[p]{} -\infty$	Diverges in probability to minus infinity or $\exists N \geq T$ so that $P(e^{x_T} > \delta) < \epsilon\ \forall \epsilon, \delta > 0$
$\Rightarrow$	Weak convergence

$\propto$	Equals up to irrelevant additive or proportional terms for maximisation or minimisation of a function with respect to the quantities of relevance.[3]

[3]In the present context $f(\mathbf{y}, \psi) \propto g(\mathbf{y}, \psi)$ if $f(\mathbf{y}, \psi) = a(\mathbf{y}) + b(\mathbf{y})g(\mathbf{y}, \psi)$ where $f(\mathbf{y}, \psi)$ is a function to be maximised or minimised with respect to parameter ψ, $\mathbf{y}$ stands for data, and $g(\mathbf{y}, \psi)$ is a function which shares the arguments with $f(\mathbf{y}, \psi)$.

Chapter 1
Introduction

Our principal aim is to improve the accuracy of the estimates of the unit-root parameter and the power and size properties of the corresponding unit-root tests via the adjusted profile likelihood functions of Cox and Reid (1987, 1993).

Adjusted profile likelihood is designed to decrease the effects of nuisance parameters in likelihood analysis. It has been able to decrease the bias of the MLE or improve the fit of the small-sample distribution of the likelihood ratio test with the asymptotic χ^2 distribution. It is well known that nuisance parameters cause bias in, and increase the variance of, the MLE of the sum of the autoregressive parameters in the ARMA(p, q) model (Sections 3.2 and 3.5). The outcome is a decrease in power or size distortion of unit-root tests (Section 3.5). There are thus good reasons to expect that adjusted profile likelihood will prove useful in this context.

Alternative theoretical devices for analysis of a parameter of interest have been proposed, such as the theories laid out in Barndorff-Nielsen (1983) and Christensen and Kiefer (1993). These theories are not without interest, and indeed Christensen and Kiefer consider the cointegration model important in the unit-root literature. Application of the Cox–Reid theory appears yet the most straightforward, e.g. no *ancillary* statistic needs to be found.

Adjusted profile likelihood has been employed only with stationary processes so the second aim is to learn about the performance of the adjustment when the process is nonstationary. Supplementary, the new estimates for the autoregressive coefficients of the AR(1) models can be applied also in the stationary case. They should be more accurate than the MLEs in such a conventional circumstance — for most values in the

parameter space at least — and hence useful as such (Chapter 5).

We start by giving a short reevaluation of unit-root econometrics from the point of view of testing and methods of interest to the thesis.[1] Apart from explaining the background of the thesis, a purpose is to establish a case for amending inference distorted by nuisance parameters. A review of likelihood theory follows in Chapter 4. The emphasis is here on Cox and Reid's adjusted profile likelihoods. The subsequent chapter addresses biases arising in the estimation of autoregressive models. Impetus for new adjustments is also sought. These results are used in the following chapter to consider the extent to which the Cox–Reid adjustment operates to diminish bias.

Adjusted profile likelihoods for an AR(1) process with constant, an alternative AR(1) process with constant suggested by Bhargava (1986), and the simple AR(2) model are derived in the leading Chapter 6. The two first models are chosen for analysis not only because of their simplicity or by their great empirical relevance but also because of the serious distortion which estimation of the constant can induce. The last mentioned model is empirically important, too, but another motivation for the analysis of it is to learn about the prospects of generalising the adjustment to longer autoregressions. It will turn out that grave distortions can arise under the AR(2) model as well.

Small-sample properties of the statistics are uncovered in Chapter 7 which is based on simulation experiments. The small-sample fractiles of the derived test statistics under a unit root with a zero constant or nonzero drift, the associated powers when the model is AR(1) with zero constant, and the fit of the small-sample distributions with the asymptotic Normal when the drift is nonzero, are reported. Next, the fractiles and corresponding powers of the relevant statistics for the Bhargava AR(1) model with constant are exposed. Finally the match of the finite sample distributions of the Wald statistics with the asymptotic unit-root distribution and the accuracy of the estimates are pondered when the model is AR(2).

[1] Banerjee and Hendry (1992) and Campbell and Perron (1991) are very readable surveys on unit-root econometrics whereas Stock (1994) and Cribari-Neto (1996) are the most up-to-date. They discuss also other related issues such as tests for the null of stationarity against the alternative of nonstationarity. The text books by Banerjee *et al.* (1993) and Hamilton (1994) offer more thorough expositions.

The findings are summarised and discussed, and suggestions for future research are outlined in the last chapter. Some laborious or supplementary derivations, complementary details, and computer programme codes are presented in the appendices.

Chapter 2

The Framework of Unit-Root Econometrics

2.1 Economic Time Series and the Process Generating them

The behaviour of many economic time series can be interpreted in the context of a formal statistical model. For example, observations of an economic series may be autocorrelated in a way corresponding to a specific model.

Most economic time series are also nonstationary: their means and variances are not constant over time. Since the beginning of the 1970s, nonstationary time series have frequently been modeled with ARIMA (autoregressive integrated moving average) models propagated by Box and Jenkins (1970). This class of processes is quite flexible because time series with very different autocorrelation structures can be handled with it. ARIMA models have been useful descriptions of time series in many cases, even though they are based on only the history of a specific time series. Indeed, ARIMA forecasts have sometimes been more accurate than forecasts of elaborate econometric models or professional economic forecasts (cf. Granger and Newbold (1975), Longbottom and Holly (1985), Naylor *et al.* (1972), the résumé by Granger and Newbold (1986, section 9.4) and Leitch and Tanner (1995) for a critique of this literature).

Debate on the nature of the nonstationarity of economic time series has been dominating time-series econometrics since the seminal article by Nelson and Plosser (1982). According to these authors, most macroeconomic time series in the United States, such as gross national product, consumer prices, wages, and money stock, are ARIMA processes. Other suggested models for (univariate) economic time series include *trend-*

stationary processes (cf. Section 3.1), *segmented trend models*[1], *fractionally integrated processes*[2], *periodically integrated processes*[3], *unobserved components models*[4] (as a special case of ARIMA models), and *nonlinear models*[5]. An empirical problem is that for short realizations, sample autocorrelation functions of these processes can closely resemble each other. Nevertheless, the time series and their properties are quite different in the long run. It is the theoretical long-run properties which make the processes interesting and important for both economists and statisticians.

A primary differentiating feature of the above mentioned processes is the long term effect on the series that an innovation imposes on the process. ARIMA models belong to the general category of *long-memory processes.*[6] Shocks driving the series have a lasting effect in this class.

The focus of this thesis is on testing the null of an ARIMA process against, loosely speaking, the alternative of a stationary process. Most of the theoretical and empirical work fits within this framework. Before going into detail, a remark on the history of the issues of the thesis is made.

---

[1]Cf. Balke and Fomby (1991), Banerjee et al. (1992), Christiano (1988), Dolado and Maravall (1989), Gallant and Fuller (1973), Linden (1995a, chapter II), Perron (1989, 1990), Perron and Vogelsang (1991) and Rappoport and Reichlin (1989).

[2]References include Cheung (1993), Cheung and Lai (1993), Crato and Rothman (1994), Crato and de Lima (1994), Diebold (1989), Diebold and Rudebusch (1989, 1991a), Geweke and Porter-Hudak (1983), Granger and Joyex (1980), Haubrich and Lo (1989), Hosking (1981), Li and McLeod (1986), Newbold and Agiakloglou (1993), Peiris and Perera (1988), Porter-Hudak (1990) and Sowell (1990, 1992a,b).

[3]Cf. Boswijk and Franses (1996), Franses and Boswijk (1991), Osborne et al. (1988), Osborne and Smith (1989), Tiao and Grube (1980), Vecchia (1985) and Vechia and Ballerini (1991), say.

[4]Cf. Harvey (1990) and the references there in, and Mocan (1994), say.

[5]Examples of this line of work are Beaudry and Koop (1993), Granger and Teräsvirta (1993), Lin and Teräsvirta (1994) and Teräsvirta and Anderson (1993). Balke and Fomby (1991a,b), Chen and Tiao (1990), Engel and Hamilton (1990), Hamilton (1989, 1994, Ch. 22), Jorion (1989), Lam (1990) and Layton (1994) consider specifically models with randomly-switching regimes. Leyborne *et al.* (1996) study a model with a stochastic autoregressive coefficient (a randomized unit-root process).

[6]The specific meaning of a long-memory process varies in the literature, cf. Mills (1990, p. 225), Granger and Joyeux (1980), and Parzen (1982), say. We allude here to the sense in Granger and Joyeux (*op. cit.*).

2.2 A Sketch of Unit-Root Econometrics and its Historical Background

The type of econometrics described in the previous section is called unit-root econometrics. The name derives from the fact that the processes studied have characteristic equations with roots equal to unity (cf. equation (3.3)). Economists, econometricians, and statisticians interested in this field are concerned, among other things, with the following themes:

- properties of unit-root processes
- economic implications of unit roots
- testing of unit roots
- the effect of unit roots on statistical inference in models encompassing unit-root processes
- empirical relevance.

These issues are of great significance whether one is theoretically or empirically oriented. Indeed, this path of research attracts researches with a wide spectrum of backgrounds. An applied econometrician may consider persistence of shocks in unit-root processes a feature of vital importance. The arising extraordinary statistical theory — usual optimality results do not hold and asymptotic distributions are nonstandard, say — points to a fascinating research area for theoreticians. (Chapter 3 provides some detail on these issues.)

According to Phillips (1995), no subject in econometrics has excited more interest than nonstationary time series since the development of simultaneous equations theory. The thrill has undoubtedly contributed to the speed at which unit-root econometrics has transformed empirical time-series econometrics (the study of Nelson and Plosser (*op. cit.*) may be considered to be the turning point).

Unit-root econometrics is a modern methodology, but unit-root processes, especially the famous random-walk model (equation (3.5)), have been used for a long time in

economics. There is a short history of using the random-walk model in describing stock market behaviour in the book by Granger and Morgenstern (1970, pp. 76-7) and in the survey by LeRoy (1989). Bachelier (1900) first utilized the random-walk model according to these authors. Orcutt (1948) was one of the first time-series analysts, if not the first, to suggest that most economic time series are unit-root processes. The following citation of Working (1934, p. 11) suggests[7] that the idea can be even older:

> 'The fact that series commonly used as indexes of business activity closely resemble series obtainable by cumulating random numbers has given support to the theory that so-called business cycles result in large degree from cumulative effects of independent random influences—. Economic theory has fallen far short of recognizing the full implications of the resemblance of many economic time series to random-difference [random-walk] series; and methods of statistical analysis in general use have given these implications virtually no recognition.'

Working went on to propose visual comparison of actual series with the artificial random-walk series he had created to judge whether a series was a random-walk process. He posed (1949) a further visual procedure apparently starting the literature of so called *variance-ratio tests* (cf. Diebold (1988) or Faust (1992), say). Unit-root econometrics has developed quickly during the last two decades, but the wait had been for a long time!

Modern economics implies random-walk behaviour for varying phenomena. There is a martingale epidemic going on in economics according to Blanchard (1981, p. 151).[8] In this respect, there is an intrinsic demand for unit-root econometrics by economists. Having said this, it should be remembered that economic models do not in general enforce strong conditions for the processes describing economic data.

---

[7]Confusingly, Working refers to a paper by Slutsky (published later or 1937 in Econometrica) where the famous Slutsky effect of moving averages causing spurious cycles was discovered (see e.g. Harvey (1981)).

[8]Lo and MacKinley (1989, footnote 1) and Perron (1988, p. 297) list articles where economic reasoning produces unit-root behaviour.

Other seminal articles include White (1958), Granger and Newbold (1974) and Phillips (1986, 1987) which revealed the exceptional asymptotic theory of unit-root processes. The idea of *spurious* or *nonsense correlation* due to (stochastic) trends in variables can be traced back as far as Yule (1926).[9]

[9]Yule's analysis was, of course, based on random-walk analysis. Pearson (1897) introduced the concept of spurious correlation. He used it to describe correlation between ratios of independent random variables. Hooker (1901) is an early study focusing on the importance of taking trends into account in analysis.

Chapter 3
Testing for Unit Roots

3.1 Two Classes of Processes

We analyse discrete time series measured at equidistant intervals generated by a well-defined statistical process. Our interest is focused on two classes of processes. The following variant of an *autoregressive moving-average process of order* (p, q) (denoted ARMA(p, q)) essentially encompasses the two classes and is helpful in distinguishing them:

$$\begin{aligned} y_t &= \mu_t + x_t \\ \rho(B)x_t &= \theta(B)\epsilon_t. \end{aligned} \tag{3.1}$$

Here $\rho(B) = 1 - \rho_1 B - \cdots - \rho_p B^p$, $\theta(B) = 1 - \theta_1 B - \cdots - \theta_q B^q$, B is the backward shift operator ($B^k y_t = y_{t-k}$), the characteristic equations $\rho(B) = 0$ and $\theta(B) = 0$ have no roots in common, μ_t is a deterministic function of time, ϵ_t is a series of identically and independently distributed random innovations so that $E(\epsilon_t) = 0$ and $E(\epsilon_t^2) = \sigma^2$ or $\epsilon_t \sim \text{IID}(0, \sigma^2)$, $\sigma^2 > 0$, $\epsilon_{-q}, \ldots, \epsilon_0 = 0$, the starting values $x_{-p}, \ldots, x_0$ are fixed and $t = 1, \ldots, T$. If $q = 0$ then the process is said to be an AR(p) process and if $p = 0$ then it is said to be an MA(q) process. An additional assumption of normality of the innovations ϵ_t will be introduced in Section 3.2.

A closely related and probably a more common variant of an ARMA(p, q) model is

$$\rho(B)(y_t - \mu_t) = \theta(B)\epsilon_t \tag{3.2}$$

where the starting values $y_{-p}, \ldots, y_0$ are fixed and the model is otherwise the same. The

models differ in the treatment of the starting values. The models and their dissimilarities are analysed in detail in sections 6.3 and 6.6 when $p = 1$ and $q = 0$ or when the process is AR(1).

Model (3.1) facilitates intuition in the present context. Assume first that the roots of the characteristic equations lie outside the unit circle drawn around the origin on the complex plane. Then μ_t can be interpreted as the unconditional mean of y_t and the process generating $y_t - \mu_t$ is (asymptotically) stationary, invertible and has unconditional mean zero.[1] μ_t can be a polynomial of time (t), say, of order k or $\mu_t = \gamma_0 + \gamma_1 t + \cdots + \gamma_k t^k$. In practice, the most interesting circumstances are $k = 0$ and $k = 1$ or $\mu_t = \gamma_0$ and $\mu_t = \gamma_0 + \gamma_1 t$, respectively. y_t is *stationary* if $k = 0$ and *trend stationary* if $k \geq 1$, a terminology originating from Nelson and Plosser (1982).[2] In the latter case, y_t deviates randomly from the deterministic time trend but follows it on average in the long run. In other words, shocks ϵ_t have only a temporary effect. Model (3.1) may aid intuition in this respect. A formal argument would involve finding the infinitely long MA-representation of x_t in which the coefficients of the innovations would eventually damp exponentially (along the lines of equation (3.19) below). Trend stationarity with $\mu_t = \gamma_0 + \gamma_1 t$ is by visual inspection a possible characterization of many important economic time series over some periods.[3]

The behaviour of the process is distinctly different if the characteristic equation $\rho(B) = 0$ has d roots equal to unity, or d *unit roots*. In this case the process generating y_t can be written as

---

[1] It should be understood that a process can be stationary only asymptotically if there are start-up effects. For simplicity, we shall ignore them and call a process stationary if the characteristic equation fulfills the stationarity condition stated in the text. The unconditional mean interpretation of the deterministic term ignores start-up effects as well.

[2] Actually they give a more general definition of trend stationarity but the one given here is adequate for our purposes.

[3] See e.g. the graphs of some Finnish, UK, and US time series in Linden (1995a), Duck (1992) and Perron (1989), respectively, and the graphs of per capita GDPs of the so called G7 countries in Cheung (1994) and Serletis (1992) for the same, but longer, series for Canada.

$$\begin{aligned} y_t &= \mu_t + x_t \\ \rho^*(B)\Delta^d x_t &= \theta(B)\epsilon_t \end{aligned} \tag{3.3}$$

or

$$\rho^*(B)\Delta^d(y_t - \mu_t) = \theta(B)\epsilon_t \tag{3.4}$$

depending on which model is used. Here $\Delta = (1-B)$ is the backward difference operator and $\rho^*(B) = 1 - \rho_1^* B - \cdots - \rho_{p-d}^* B^{p-d} = (1-B)^{-d}\rho(B)$. If the roots of the equations $\rho^*(B) = 0$ and $\theta(B) = 0$ lie outside the unit circle, then $\Delta^d(y_t - \mu_t)$ is (unconditionally) a zero-mean stationary and invertible time series. Box and Jenkins called this model an *autoregressive integrated moving-average process of order* $(p-d, d, q)$, $p \geq d \geq 0$ (with acronym ARIMA$(p-d, d, q)$). Equivalently, $y_t - \mu_t$ can be described as being integrated of order d and denoted $(y_t - \mu_t) \sim I(d)$ as in Engle and Granger (1987) ($I(0)$ corresponding to stationarity). A consequence of the(se) unit root(s) is that y_t is a sum of random terms (as the adjective integrated suggests) with no tendency to stay close to μ_t since shocks have a permanent effect. This may be more apparent from the form (3.3) than from the usual expression (3.4). This is in sharp contrast to the trend-stationary model in which the impact of shocks fades eventually. A further difference is that in the unit-root case, $(1-B)^d\mu_t$ is the unconditional mean of $(1-B)^d y_t$ whereas in the stationary case μ_t and y_t were associated similarly.

We remark that μ_t could again be a polynomial in time. The order of the polynomial is empirically likely to be at most d or $\mu_t = \gamma_0 + \gamma_1 t + \cdots + \gamma_d t^d$.[4] This being the case, in general, only the parameter γ_d is identifiable as $(1-B)^d(\gamma_0 + \gamma_1 t + \cdots + \gamma_d t^d) = \gamma_d d!$ where $d! = d(d-1)\cdots 1$. Likewise, if μ_t is a polynomial of order $k < d$ then $(1-B)^d\mu_t = 0$ and $\mu_0, \ldots, \mu_k$ are redundant parameters in the model (3.3) except for start-up effects.

Example. The popular *random-walk* model follows from setting $\rho^*(B) = \theta(B) = d = 1$

[4] Agbeyegbe (1993) makes an exception by claiming that some mineral prices would feature a quadratic trend and $d = 1$.

and $\mu_t = \gamma_0$ in equation (3.3). The result is the simplest example of an $I(1)$ series:

$$y_t = y_{t-1} + \epsilon_t \tag{3.5}$$

where $\epsilon_t \sim \text{IID}(0, \sigma^2)$. γ_0 is now not identifiable unless the starting value x_0 is known.□

Only the cases $d = 1$, $d = 0$, $k = 1$ and $k = 0$ will be considered in what follows. Conditions $d = 1$ and $k = 1$ both imply that the process is nonstationary. This is empirically the most appealing situation though $d = 2$ may be a possible characterization of some nominal series (see e.g. Caporale and Pittis (1993), Clements and Mizon (1991), Engsted (1993), Dickey and Pantula (1987), Hall (1986), Hylleberg and Mizon (1989), Juselius (1991) and Taylor (1991)).

Introductory economic background for unit-root processes was given in Chapter 2. The consumption theory of Hall (1978), hysteresis in unemployment (Blanchard and Summers (1986)), and real business cycle models (Nelson and Plosser (1982)) provide subsequent examples of economics yielding unit-root processes. Further intuitive reasons have been given for integratedness. For example, Deaton (1986, p. 22, 1992, p. 112) considers unit-root processes more likely and claims (1986, p. 22) that wearing out of the capital stock is more likely to produce a permanent change to the level of production in the fashion of a unit-root model than only a temporary drop as with the trend-stationary model. Hendry presents the following interesting argument (personal communication) based on a remark of Granger (1986, p. 217). Consider the process

$$\begin{aligned} \Delta y_t &= \kappa(y_{t-1} - \lambda x_{t-1}) + \epsilon_t \\ \Delta x_t &= \kappa^*(y_{t-1} - \lambda x_{t-1}) + \epsilon_t^*. \end{aligned}$$

Here κ and κ^* are less than one in absolute value and at least one or the other is nonzero, $\epsilon_t \sim \text{IID}(0, \sigma^2)$ and $\epsilon_t^* \sim \text{IID}(0, \sigma^{*2})$ (uncorrelated), $\sigma^2, \sigma^{*2} > 0$, and $y_{t-1} - \lambda x_{t-1}$ is $I(0)$ or the variables y_t and x_t are *cointegrated* (Granger (1981) and Engle and Granger (1987)). This is a simple *error-correction model* (originating from Phillips (1954, 1957), Sargan

(1964), and Davidson *et al.* (1978)). Bivariateness and IIDness are assumed only for simplicity. The linear combination $y_t - \lambda x_t$ may be called an equilibrium relation between the variables as they tend towards the line $y_t = \lambda x_t$ in the long run. Both variables, y_t and x_t, are obviously $I(1)$. The point is that if there are to be error-correction mechanisms of the above kind in the economy then the associated variables have to be $I(1)$. This is a consequence of the *Granger Representation Theorem* (Engle and Granger (*op. cit.*)). The remark has pragmatic relevance as error-correction models have gained empirical success.

Though economic theory leads often to a random-walk process instead of a more general unit-root process it does not necessarily conflict with the empirical time series being autocorrelated in differences. Measurement errors or temporal aggregation of the true underlining series create autocorrelated observed series (e.g. Harvey (1981, pp. 43–4)).

Few present day econometricians have argued for the likeliness of the trend-stationary model as such. A major reason appears to be that in the case of a (known) trend-stationary process one would be able to forecast the distant level of the variable with negligible forecast error (e.g. Perron (1989, p. 1387)).[5] Relatedly, a deterministically determined level of a time series, even in the very distant future, does not appear an economically reasonable assumption. Instead, many analysts prefer the segmented trend model where the process follows a trend-stationary process for a period of time with the time trend changing abruptly. These authors assert that persistent shocks to the economy take place, but they are seldom and large when they occur (e.g. Balke and Fomby (1991, 1994), Perron (*op. cit.*) and Rappoport and Reichlin (1989)). The advocates of this approach refer to wars, natural disasters, oil-price shocks, financial panics, changes in policy regimes, and technological breakthroughs as prime reasons for persistence of shocks in the sense of breaking trends.

[5]On the other hand, Sampson (1991) argues that the forecast variance increases with the square of the forecast horizon if the parameters are estimated whether the model has a unit root or is trend stationary.

Trend-stationary and integrated processes are two extreme occurrences of segmented trend models (Rappoport and Reichlin (*op. cit.*)). A single segment corresponds to a trend-stationary model, but as the number of trend breaks becomes equal to the number of observations an integrated process results.[6]

The dichotomy appears of great relevance to economic theory, cf. Deaton (1992, Chapter 4), say, for an example from the theory of consumption. A statistical — and important — aspect is that asymptotic distribution theory depends crucially on the existence or nonexistence of a unit-root in the process generating the data (as explained in the following sections).[7]

Which model characterizes economic variables better remains an unsolved question despite remarkable theoretical innovations to the field and numerous empirical studies. Though the trend-stationary model is incredible on the long run it is a feasible alternative, particularly as an approximation rather than a structure, for shorter time periods.[8]

To point out the empirical plausibility of unit-root processes, we refer to the study of Koop and Poirier (1995). The null hypothesis of a unit root could not be rejected even at the 10 per cent level for the GDP of 86 of the 91 countries they studied (they employed so called ADF tests, cf. Section 3.3).

3.2 Dickey–Fuller Tests

Dickey (1976), Fuller (1976) and Dickey and Fuller (1979) developed the first, and by far the most utilized, tests for the null hypothesis of an integrated process against the alternative hypothesis of a (trend) stationary process. They assumed that the process is purely autoregressive i.e. $\theta(B) = 0$ in model (3.1). Basically, they proposed six tests denoted ρ, τ, ρ_μ, τ_μ, ρ_τ and τ_τ which we shall refer to as the DF tests. (We will later

[6] A problem with segmented trends is the specification of the segments but we shall pass that issue (cf. Christiano (1992) and Dolado and Maravall (1989), say).

[7] Though sometimes one may be able to base inference on the traditional asymptotic distribution theory if one can formulate the problem in a suitable way, see Campbell and Perron (1991, Section 3.4).

[8] The tests surveyed and derived below do not specify an explicit alternative. Still in practice, the conclusion from a rejection of a null of a unit root may well be a trend stationary model, at least segmentwise.

substitute the ρ_μ symbol by ψ_μ.)

We start by considering equation (3.2) with $\rho(B) = 1 - \rho B$, $|\rho| \leq 1$, $\theta(B) = 1$ and $\mu_t = \gamma_0 + \gamma_1 t$:

$$y_t - \gamma_0 - \gamma_1 t - \rho(y_{t-1} - \gamma_0 - \gamma_1(t-1)) = \epsilon_t \tag{3.6}$$

which is equivalent to

$$y_t = \alpha + \beta t + \rho y_{t-1} + \epsilon_t \tag{3.7}$$

and

$$\Delta y_t = \alpha + \beta t + \rho' y_{t-1} + \epsilon_t. \tag{3.8}$$

Here $\alpha = (1-\rho)\gamma_0 + \rho\gamma_1$, $\beta = (1-\rho)\gamma_1$ and $\rho' = (\rho - 1)$. This parameterisation was established by Bhargava (1986) and helps to interpret the tests (see Section 6.6 for a detailed analysis of the model with a constant). From now on we shall assume that $\epsilon_t \sim$NID$(0, \sigma^2)$ (normally independently distributed).

The first two tests of Dickey and Fuller assume that $\gamma_0 = \gamma_1 = 0$ in which case equation (3.7) shrinks to

$$y_t = \rho y_{t-1} + \epsilon_t \tag{3.9}$$

or equivalently

$$\Delta y_t = \rho' y_{t-1} + \epsilon_t. \tag{3.10}$$

The null hypothesis is $\rho = 1$ or $\rho' = 0$ the alternative being $|\rho| < 1$.[9] Intuitively, an obvious idea to test the null is to estimate ρ and study if $\hat{\rho} - 1$ (where $\hat{}$ denotes an esti-

[9]This is the interpretation of the tests in practice. None of the tests of Dickey and Fuller (1979) state explicitly an alternative hypothesis.

mate) differs statistically significantly from zero. The obstacle is that y_t is nonstationary under the null and as a consequence the standard distribution theory of Mann and Wald (1943) does not apply. Nevertheless, Dickey and Fuller proposed two tests based on this idea. The first test statistic is $T(\hat{\rho}_{ml} - 1)$ where T is the sample size and $\hat{\rho}_{ml}$ is the MLE conditional on the initial value (assumed to be zero) from equation (3.9), which is numerically equivalent to the corresponding ordinary least-squares (OLS) estimate.[10] The second statistic is the ordinary 't-value' of $\hat{\rho}_{ml} - 1$ which is labeled $\hat{\tau}$ in this context. It can be considered to be a Wald test as it features the same asymptotic distribution as a Wald statistic constructed along the lines set out in Section 4.1.

It was as early as 1958 when White concluded that $\hat{\rho}_{ml}$ converges at a rate of $O_p(T^{-1})$ when $\rho = 1$ instead of $O_p(T^{-1/2})$ as under $|\rho| < 1$ which was then the standard result (and notation) of Mann and Wald (*op. cit.*). This is the *super consistency* property of Stock (1987). (Rubin (1950) had showed even earlier that $\hat{\rho}_{ml}$ is weakly consistent when $|\rho| \geq 1$). White found[11] that the asymptotic distribution of $T(\hat{\rho}_{ml} - 1)$ is a functional of standard Brownian motion or Wiener process ($W(r)$):

$$T(\hat{\rho}_{ml} - 1) \Rightarrow \frac{\frac{1}{2}\{[W(1)]^2 - 1\}}{\int_0^1 [W(r)]^2 dr} = \frac{\int_0^1 W(r) dW(r)}{\int_0^1 [W(r)]^2 dr} \tag{3.11}$$

where $\Rightarrow$ signifies weak convergence (see Banerjee and Hendry (1992) or Hamilton (1994) for a detailed and heuristic derivation of this formula).[12] But it was Dickey (1976) who tabulated the small-sample distribution of $T(\hat{\rho}_{ml} - 1)$ and $\hat{\tau}$ by simulation, devised the generalizations below, and made testing of the unit root practical. Fuller (1976) has published the corresponding tables.[13] The distribution (3.11), or equivalently the

[10] This applies to all autoregressive models discussed in this chapter.

[11] His formula is slightly in error, see Phillips (1987, p. 282).

[12] The asymptotic theory holds also under much less stringent assumptions than those made here, cf. Phillips and Perron (1988). Interesting results can emerge if the innovations follow a (non-Normal) distribution with an infinite variance. *M-estimates* (maximum-likelihood type estimates) can converge then even faster than $O_p(T^{-1})$ (Knight (1989), say).

[13] The fractiles in Fuller (*op. cit.*) are based on a smaller simulation study, though, than those in Dickey (1976). The fractiles for the ρ_t and τ_τ test in Guilkey and Schmidt (1989) should be more accurate than those published by Fuller. Nabeya and Tanaka (1990a) have numerically calculated accurate fractiles for the asymptotic distributions of the different variants of the ρ tests.

asymptotic distribution of the standardised $\hat{\rho}_{ml}$ (if $\rho = 1$) is depicted in Figure 6.1 in Section 6.2. The distribution is skewed to the left and the associated mean is less than zero (references include Abadir (1993), Evans and Savin (1981), Fuller (1976) and Shenton and Johnson (1965)).[14] A related example is given at the end of the section.

The asymptotic distribution of $\hat{\tau}$ is

$$\hat{\tau} \Rightarrow \frac{\int_0^1 W(r)dW(r)}{\sqrt{\int_0^1 [W(r)]^2 dr}} \tag{3.12}$$

(an up-to-date reference is Phillips (1987)). It is illustrated in Figure 6.2 in Section 6.2. Thus $\hat{\tau}$ does not follow the t-distribution even asymptotically, the critical values of which would lead to notable over-rejection of the null hypothesis. For example, the critical values of $T(\hat{\rho}_{ml} - 1)$ and $\hat{\tau}$ are -7.9 and -1.95 when the size of the tests is 0.05 and $T = 100$ (the equivalent critical value from the t-distribution being around -1.64). The critical value of $T(\hat{\rho}_{ml} - 1)$ corresponds to $\hat{\rho}_{ml} \simeq 0.92$ in this case. Dickey and Fuller named the tests as ρ and τ tests, respectively.

A convenient way to calculate the tests is to estimate equation (3.10). The first statistic is then $T\,\hat{\rho}'_{ml}$ and the second its 't-value' as reported automatically by most regression packages.

The implicit assumption of the ρ and τ tests is that the process has zero mean under the alternative. The ρ_μ and τ_μ tests of Dickey and Fuller allow a nonzero mean (γ_0) under the alternative as one often wants to do in applied work. These tests still assume that $\gamma_1 = 0$ in equation (3.7) which then becomes equal to:

$$y_t = \alpha + \rho y_{t-1} + \epsilon_t \tag{3.13}$$

[14]The distribution is skewed to the left also in the stationary case for $\rho > 0$, especially for $0.8 < \rho < 1$, say, cf. Section. 5.3, Fuller (1976, p. 339), Davidson and MacKinnon (1993, Chapter 21), Hendry *et al.* (1990), Nankervis and Savin (1988), Phillips (1977, 1978), Tuan (1992), and White (1961), say.

We note in passing that the Yule-Walker estimate of the autoregressive coefficient suffers from a yet more severe bias (Shaman and Stine (1988) and Tjøstheim and Paulsen (1983)). An intuitive reason is that the Yule-Walker estimate is guaranteed to lie in the stationary region.

or

$$\Delta y_t = \alpha + \rho' y_{t-1} + \epsilon_t. \tag{3.14}$$

The test statistics are now $T(\hat{\rho}_{\mu,ml} - 1)$ and $\hat{\tau}_\mu$ where $\hat{\rho}_{\mu,ml}$ is the MLE (conditional on initial values) of ρ in equation (3.13) and $\hat{\tau}_\mu$ is the 't-value' of $\hat{\rho}_{\mu,ml} - 1$. $\alpha = (1-\rho)\gamma_0$ becomes zero under the null of the unit root but is different from zero (in general) under the alternative and has to be estimated.[15] Estimation of α also makes the tests invariant with respect to the starting values of the process (Section 6.3). $T(\hat{\rho}_{\mu,ml} - 1)$ and $\hat{\tau}_\mu$ have critical values -13.7 and -2.89 at risk level 0.05 when $T = 100$. The former value implies that the null will be rejected if $\hat{\rho}_{\mu,ml} < 0.86$. Both values are larger in absolute value than the corresponding critical values of ρ and τ tests. Again, it is a bit easier to calculate the test statistics from the regression of model (3.14) just as in the zero-mean case.

Finally, it may be appropriate to allow a drift under the null or let the process be trend stationary under the alternative. This means that no restrictions are imposed on equation (3.6) under the alternative. One should then estimate equation (3.7) or (3.8) by maximum likelihood (ML). The resulting test statistics are now denoted by $T(\hat{\rho}_{\tau,ml} - 1)$ and $\hat{\tau}_\tau$ and the tests are called ρ_τ and τ_τ tests, respectively. The corresponding critical values are larger in absolute value than the critical values of the previous tests. For example, the critical value of $T(\hat{\rho}_{\tau,ml} - 1)$ is -20.7 which means that the null is rejected if $\hat{\rho}_{\tau,ml} < 0.79$ when the size of the test is 0.05 and $T = 100$. The critical value of $\hat{\tau}_\tau$ is -3.45 in the same situation. We note that the coefficient of the time trend β equals zero and the constant α becomes a drift equal to γ_1 under the null.[16] Invariance of the tests with respect to α is achieved via estimation of the additional parameter β.

The asymptotic distributions of $T(\hat{\rho}_{\mu,ml} - 1)$, $\hat{\tau}_\mu$, $T(\hat{\rho}_{\tau,ml} - 1)$ and $\hat{\tau}_\tau$ are also functionals of Brownian motion but of somewhat more complicated form than formulae (3.11) and (3.12), cf. formulae (6.8) and (6.9) in Section 6.3 and Phillips and Perron (1988). The

[15] We are arguing heuristically.

[16] It is possible to test the unit root property with a likelihood ratio test which makes use of the fact that also the coefficient of the time trend is zero under the null (cf. Dickey and Fuller (1981)). However, the likelihood ratio test is really not more powerful than the other tests, so there is no special reason to use it.

associated asymptotic p-values can be inferred with the help of the tables in MacKinnon (1994) which are based on intensive simulation experiments. (Adda and Gonzalo (1996) give similar results which have been derived analytically.) Estimation of the nuisance parameters α and β increases the downward bias and the variance of estimates of ρ, cf. the figures in MacKinnon (*op. cit.*), say. Heuristically, detrending forces the time series to appear more stationary than the original series was.[17]

Example (Nelson and Plosser (1982)). If $\rho = 1$, $\beta = 0$ and $T = 100$ in model (3.7) then $E(\hat{\rho}_{\tau,ml}) = 0.9$ and $E(\hat{\tau}_\tau) = -2.22$.□

3.3 Augmented Dickey–Fuller Tests

The three τ tests can be extended to testing of AR(p) processes with $p \geq 2$. Then it is assumed that y_t is generated from

$$\rho(B)(y_t - \gamma_0 - \gamma_1 t) = \epsilon_t \tag{3.15}$$

where $\rho(B) = 1 - \rho_1 B - \cdots - \rho_p B^p$. This is a generalization of model (3.6) (and a special case of (3.2) with $\theta(B) = 1$ and a Normal error). From the point of view of testing for unit roots, a more useful parameterisation is

$$y_t = \alpha + \beta t + \psi y_{t-1} + \textstyle\sum_{i=1}^{p-1} \phi_i \Delta y_{t-i} + \epsilon_t \tag{3.16}$$

or

$$\Delta y_t = \alpha + \beta t + \psi' y_{t-1} + \textstyle\sum_{i=1}^{p-1} \phi_i \Delta y_{t-i} + \epsilon_t. \tag{3.17}$$

Here $\alpha = \gamma_0(1 - \sum_{i=1}^{p} \rho_i) + \gamma_1 \sum_{i=1}^{p} i\rho_i$, $\beta = \gamma_1(1 - \sum_{i=1}^{p} \rho_i)$, $\psi = \sum_{i=1}^{p} \rho_i$ (or $\psi' = \sum_{i=1}^{p} \rho_i - 1$) and $\phi_k = -\sum_{j=k+1}^{p} \rho_j$, $k = 1, \ldots, p-1$. The trick is that under the null of a unit root $\psi = 1$ or $\psi' = 0$ (and $\beta = 0$) and thus the existence of a unit root depends again on one single parameter (ψ or ψ'). Dickey (1976) and Fuller (1976) showed that

[17] Of course, this is the reason why detrending has so often been done.

the asymptotic distributions of the $\hat{\tau}$ statistic and adaptations of it stay the same even if the equations (3.9), (3.13) and (3.7) are augmented with extra $p-1$ lags. Thus we can, say, apply the *augmented* τ_μ *test* or $\tau_\mu(p)$ *test* by calculating the following regression:

$$y_t = \alpha + \psi y_{t-1} + \sum_{i=1}^{p-1} \phi_i \Delta y_{t-i} + \epsilon_t$$

(assuming $\gamma_1 = 0$). The nulls and alternatives of the augmented tests correspond to those of the original τ tests.

Dickey and Fuller did not generalize the ρ tests to the AR(p) case as the corresponding asymptotic distributions turned out to depend on the nuisance parameters ρ_i, $i = 1,\ldots,p$. It is namely $Tf(\hat{\psi} - 1)$ which follows the corresponding Dickey–Fuller distribution asymptotically. Here $f = [\phi(1)]^{-1}$, $\phi(B) = 1 - \phi_1 B - \cdots - \phi_{p-1} B^{p-1}$ and $\hat{\psi}$ is estimated with or without a constant or a time trend according to the above-stated nulls and alternatives (Dickey (1976, Chapter 6) and Fuller (1976, p. 374)).

If p is not known *a priori* then it can be chosen with a usual t test (by eliminating apparently unnecessary lagged values from an AR(k) model with $k \geq p$) or by an information criterion, say. Limited simulation evidence suggests that the empirical size of the test remains close to the nominal size despite the data-based determination of lag order (Ng and Perron (1995)).[18]

3.4 The Generalization of Said and Dickey

Said and Dickey (1984) were able to generalize the τ tests further for testing ARMA(p,q) processes with $q \geq 1$ or with an MA-part. Their idea is to approximate an ARMA(p,q) process with a long autoregression and test for the unit root from the approximating autoregression. They used the parameterisation

[18]Nominal size is the size which applies under the hypothetical situation in which one has an infinite number of observations. Distributions of test statistics and the associated critical values have often been derived only under this assumption. Researches approximate the (unknown) exact small-sample critical values with the asymptotic critical values when working with empirical (finite) data. Empirical size is the size that results from this practise.

$$\begin{aligned} y_t &= \psi y_{t-1} + x_t \\ \rho(B)x_t &= \theta(B)\epsilon_t \end{aligned} \tag{3.18}$$

for an ARMA$(p+1,q)$ process. Here $\mid \psi \mid \leq 1$, $\rho(B)x_t = \theta(B)\epsilon_t$ is a stationary and invertible ARMA(p,q) process and $\epsilon_t \sim$IID$(0,\sigma^2)$. If $\psi = 1$ then y_t is an ARIMA$(p,1,q)$ process. If the process has been in operation indefinitely then x_t can be expressed as

$$x_t = \sum_{i=1}^{\infty} \pi_i x_{t-i} + \epsilon_t \tag{3.19}$$

where π_k, $k = 1,2,\ldots$, are coefficients of B^k in the polynomial $\pi(B) = \theta^{-1}(B)\rho(B)$. The series π_i converges to zero exponentially from the coefficient π_m onwards with $m = \max(p, q+1)$ (e.g. Harvey (1981, p. 38)). Using the fact that under the null of a unit root $\Delta y_t = x_t$ we get an approximate result

$$y_t = \psi y_{t-1} + \sum_{i=1}^{m} \pi_i \Delta y_{t-i} + \epsilon_t. \tag{3.20}$$

The by now familiar trick is that under the null $\psi = 1$ and thus the existence of a unit root depends again on a sole parameter (ψ). Said and Dickey proved that the 't-value' ($\hat{\tau}$) of the OLS estimate $\hat{\psi}_{ols}$, follows asymptotically the same distribution as in the AR(1) case if the length of the autoregression m is made to increase at a rate of $o(T^{1/3})$.[19] Thus the τ test can be applied also in this context with the help of asymptotic theory. Further, a constant can be added to the equation (3.20). Then the $\hat{\tau}_\mu$ (with obvious notation) statistic can be applied just as in the case of the AR(1) with an asymptotic interpretation. The exponential decay of the π_i coefficients and the increase of the length of the autoregression provide the intuition of the result.

A general comment concludes the introduction of tests. Numerous other tests such as the ones in Andrews (1993), Bhargava (1986), Choi and Phillips (1993), Elliot *et al.*

[19] They employed also a condition on a lower bound on m but Ng and Perron (1995) showed that this condition is unnecessary for the result to hold.

(1996), Pantula *et al.* (1994), Pantula and Hall (1991), and Phillips and Perron (1988) have been advanced. A pragmatic reason for their exclusion from the chapter is that our aim is to improve unit-root tests via improvements to the ML method, on which these excluded tests do not rely. Another reason is that the empirical size of the asymptotics based excluded tests may differ notably from nominal size if the model includes (non-zero) MA- or AR-terms (over the first lag). Schwert (1987, 1989) recommends use of the ADF- or Said–Dickey tests in general, Phillips and Perron (*op. cit.*) and Pantula (1991) if $\theta > 0$ in equation (3.21) — the most common case empirically — and DeJong *et al.* (1992) and Pantula and Hall (*op. cit.*) if a constant and a time trend are included in the regressions.[20] The difficulty with the size of the asymptotics-based tests is rationalized next.

3.5 Problems with Testing for Unit Roots

There are both economic and statistical reasons motivating devices to test if a process is integrated or (trend) stationary. However, unit-root tests have been severely censured. Christiano and Eichenbaum (1990), Cochrane (1991a,b), Miron (1991), and Sims (1988) have an economic viewpoint on top of a statistical one while Blough (1992) with Faust (1993, 1994) focus on statistical aspects. Sims (*op. cit.*) and Sims and Uhlig (1991) among others express Bayesian criticism which Phillips (1991a,b) disputes (see Phillips (1995), though). West (1988a) opposes the tendency to draw conclusions from unit-root tests about the kind of shocks ('real' or 'nominal') which an economy has faced. We shall highlight the statistical restraints.

To start with, the tests lack power. Power of the ρ and τ tests is modest and gets much worse if the nuisance parameters of the ρ_μ, τ_μ, ρ_τ, and τ_τ tests are estimated. As already explained, detrending has a seemingly stationarising impact on the series.

Example. Power of the τ_τ test is merely 0.08 if $\rho = 0.95$ in equation (3.7), size of the test

[20]The tests proposed by Elliot *et al.* (*op. cit.*) were not included in the comparisons. Haug (1993) reports related results favourable to ADF type tests.

is 0.05, and $T = 100$ (Dickey *et al.* (1986)).□

Further, the power seems to decrease monotonically as the tests are augmented or more lags are introduced to the equation (3.15) (as long as empirical size agrees with nominal size, see Agiakloglou and Newbold (1992), Blough (1992), DeJong *et al.* (1992), Ng and Perron (1995), Phillips and Perron (1988), Stock (1994), and the related results by Boswijk and Franses (1992)). A striking example is given below.[21]

An issue largely neglected in the literature is deterioration of power due to a starting value which deviates from the unconditional mean of the process (μ_t in equation (3.2)). (Schmidt and Phillips (1992) are an exception, see also Abadir (1993) for a related example.) If the autoregressive parameters are such that the process is autocorrelated in the first place, then the starting value may induce trend-like behaviour, amplify the estimated autocorrelations, and weaken the tests. The ρ_μ test is scrutinized in Chapter 7 where it is shown that the test can be even biased if the starting value is large enough.

Next, the tests suffer from size distortions due to nuisance parameters. As argued by Schwert (1989) one should often use the Said–Dickey $\tau(p)$ test or adaptations of it because processes generating economic time series are likely to include MA-terms due to measurement errors or temporal aggregation, say (e.g. Harvey (1981, pp. 43–4)). As explained in the previous section, these test statistics follow only asymptotically the Dickey–Fuller distribution. As a consequence, actual size of the tests does not necessarily coincide with nominal size. This phenomena has been documented by Agiakloglou and Newbold (1992), Elliot *et al.* (1992), Ng and Perron (1995), Pantula (1991), Pantula and Hall (1991), Phillips and Perron (1988) and Schwert (1987, 1989); cf. also Molinas (1986). They analyzed the ARIMA$(0,1,1)$ model

$$(1 - \rho B)y_t = (1 - \theta B)\epsilon_t \tag{3.21}$$

where $\rho = 1$ (or a modification of this formula with a constant). The main point is

[21] Another, though subtler, related matter is that observations are lost in the sense of some of them becoming starting values when long autoregressions are estimated.

that as the nuisance parameter θ approaches 1 we have y_t approximately equal to ϵ_t, i.e. y_t starts to resemble an IID series. The consequence, in small-samples, is severe size distortion of almost all unit-root tests in use. Simulation evidence (cf. the aforementioned articles) suggests that only the $\tau(p)$ type of tests perform adequately.[22] According to the simulations of Schwert (1989) the nominal size is reasonably close to empirical size for both $\tau_\mu(m_{12})$ and $\tau_\tau(m_{12})$ tests when $-0.8 \leq \theta \leq 0.8$ and $25 \leq T \leq 1000$ and $m_{12} = int[12(T/100)^{1/4}]$. The results of Agiakloglou and Newbold confirmed (for $T = 100$) that autoregressions of about this length usually capture the autocorrelation structure of the process (3.21) well enough for the Said–Dickey tests to have empirical size close to nominal size for aforementioned values of θ.[23]

However, the $\tau(m_{12})$ family of tests (when $25 \leq T \leq 1000$) is anything but flawless. First, a minor point, which is that we do not yet know how the tests perform if the process belongs to the ARIMA$(p, 1, q)$ class where $q \geq 2$. (There was no problem in a small experiment in Stock (1990).) Secondly, and more importantly, Agiakloglou and Newbold found disappointingly, though not surprisingly, that the empirical size of the $\tau_\mu(11)$ test (the longest lag and the only $\tau(p)$-type of test they studied) does not anymore coincide with the nominal size if $\theta = 0.9$ and $T = 100$ (the sole sample size they experimented). (In this case, empirical size was 0.14 while asymptotic theory suggested 0.05.) The test breaks down, of course, as θ is too close to 1. Thirdly, we do know that the power of the $\tau(p)$ tests can be extremely poor.

Example. Power of the $\tau_\mu(11)$ test with size 0.05 is 0.52 when $\theta = \rho = 1$ in equation (3.21) and $T = 100$ (Agiakloglou and Newbold, *op. cit.*). A unit-root test which has difficulties to distinguish an IID series from a nonstationary process does not sound useful![24] □

---

[22] The tests in Elliot *et al.* (*op. cit.*) have much potential in terms of achieving a reasonable size and an increase in power, see also Stock (1994).

[23] Results of Phillips and Perron (1988, Table 1a) seem to be in error on this matter as their results contradict both those of Agiakloglou and Newbold (1992) and Schwert (1989). Agiakloglou and Newbold point out that a possible source of disparities in simulation evidence in general may be the different estimation algorithms used.

[24] Power deteriorates to an absurd 0.20 (about) if a time trend is added to the regression (Blough (1992)).

The poor power in the above example is due to the heavy over parameterization of the model. Ng and Perron (1995) and Stock (1994) argue that sample information should be made use of to increase the power. However, a trade-off between the size and the power is unavoidable, and is exemplified by their own simulations. They showed that a t test based procedure for lag length selection (favored by them) leaves more room for size distortions than the deterministic m_{12} rule, say (see also Pantula *et al.* (1994)). In this context, employment of sample information effectively means that the idea of significance testing is replaced by a subjective weighting of the gain in power and loss from allowing for a variable and unknown test size.

Blough (1992) gives a theoretical framework for the conflict between size and power. We modify slightly the process he considers to

$$\begin{aligned} y_t &= a y_{1t} + (1-a) y_{2t}, \quad 0 \leq a \leq 1, \\ y_{1t} &= y_{1t-1} + \epsilon_{1t}, \\ y_{2t} &= \frac{\theta(B)}{\rho(B)} \epsilon_{2t}, \end{aligned} \tag{3.22}$$

where $\epsilon_{1t} \sim \text{IID}(0, \sigma_1^2)$, $\epsilon_{2t} \sim \text{IID}(0, \sigma_2^2)$ (not necessarily uncorrelated), $\sigma_1^2, \sigma_2^2 > 0$, $\rho(B)$ and $\theta(B)$ are as stated under equation (3.1) with the additional assumption that the roots of the characteristic equations lie outside the unit circle, and $t = 1, 2, \ldots, T$. The time series y_{2t} is a stationary and invertible ARMA$(1,1)$ process while y_{1t} is a random walk. If $a = 1$ then y_t is a random walk, too, but if $a = 0$ then y_t becomes the stationary ARMA(p,q) process y_{2t}. The intermediate cases $0 < a < 1$ yield an ARIMA$(p, 1, q+1)$ process with the root of the characteristic equation for the MA-polynomial tending to -1 when a tends to 0. While not completely general, the formulation makes it clear that there are integrated and stationary processes whose realizations must be very much alike (cases with a close to 0). We shall return to this issue shortly. Moreover, the above structure enables Blough to prove the following proposition: a unit-root test which holds its size for every model in the null with a finite-sample size T cannot have power of no more than the size against *any* model in the alternative. Intuitively, if a unit-root test

has a specified size when $a \approx 0$ then one really cannot expect it to display much more power if a is exactly zero regardless of the other parameters specifying the model. The result explains the very poor power performance of the Said–Dickey procedure in the above example. The $\tau_\mu(11)$ test holds its size well for a wide range of models in the alternative but at the cost of severe power loss for any model in the alternative including white noise.

Faust (1993, 1994) proves associated results claiming additional problems in the asymptotic properties of generic unit-root tests. We shall pass the detail and remark that the above finite-sample results of Blough and the asymptotic results of Faust are not unimportant but do not carry over to the AR(p) class of models (with finite p) in which we are interested in Chapter 6. (It is fairly easy to see that model (3.22) cannot accommodate an integrated AR(p) process in general.)

Relatedly, Saikkonen (1993) and Luukkonen and Saikkonen (1996) point out that the unit-root tests of Solo (1984) and Ahn (1993) based on the LM principle can be inconsistent when MA terms are present.

Another line of criticism was initiated by Cochrane (1991a) (see also Blough (1992)). Let us suppose that a unit-root test really could discriminate a unit-root process from a (trend) stationary process. Cochrane argued that this might not be useful after all. If an $I(1)$ series is stationary for practical purposes though $I(1)$ theoretically (as when $\theta = 0.99$ in model (3.21)) then distribution theory based on an assumption of stationarity may be more accurate than a unit-root distribution. (A small simulation study in Stock (1990) supports this view.) Similarly, if a process is theoretically stationary but appears nonstationary in small-samples (as when $\rho = 0.99$ in (3.9)) then unit-root theory may be more appropriate. A correct answer to the question 'Is there a unit root?' may thus be misleading when one should decide which distribution theory to apply. Cochrane's reasoning certainly has pertinence but it still remains hypothetical. Present tests are not capable of differentiating such processes and an applied econometrician would usually

choose a suitable distribution theory just as Cochrane asserts he should.[25]

Finally, a few notes that are slightly out of the focus of the thesis but are still worth mentioning. Seasonal adjustment decreases power of the tests further (Ghysels, (1990), Ghysels and Perron (1993), Jager and Kunst (1990) and Olekalns (1994)). Structural breaks or innovative outliers can ruin the tests in the sense of making the tests extremely conservative (cf. Balke and Fomby (1990a,b), Hendry and Neale (1990), Perron (1989, 1990), Reichlin (1989), Ghysels (1992), the survey by Stock (1994), and Lucas (1996, p. 103) on outliers). DF tests feature low power against many fractionally integrated processes, too (Diebold and Rudebusch (1991b) and Hassler and Wolters (1994)). Of course, other types of misspecification can weaken the tests as well (Pippenger and Goering (1993)). On the other hand, additive outliers can lead to a high frequency of type I errors (Franses and Haldrup (1994) and Lucas (1996, p. 78)). De Long and Lang (1992) argue that publication bias ('surprising is interesting') distorts reported inference against the (well-established) unit-root null.

Altogether, these results throw doubt on the usefulness of unit-root tests in judging the degree of integratedness of a time series, and on their value in general especially when MA terms are present. The problem in the very heart of unit-root econometrics is that the properties of the two classes of processes, integrated and trend stationary, are not as prominently different in finite samples as they are asymptotically. All the above mentioned problems have ample empirical relevance. For example, many time series of economic interest have nonzero mean and they feature drift-like behaviour introducing the nuisance parameters α and β into the regressions of Section 3.2. Also MA-components seem to exist in many empirical time-series processes (Campbell and Mankiw (1987, 1989) and Schwert (1987), say) implying long approximative autoregressions with many nuisance parameters π_k as in Section 3.4. Of course, long pure autoregressions are possible as well.[26] A conclusion is that there is much scope for improving the present unit-root

[25]This and a related point concerning forecasting are debated further in Campbell and Perron (1991) and Cochrane (1991b); see also Blough (1992) .

[26]Harris (1992) had only 32 observations but needed an AR(8) model. Maravall and Mathis (1994)

testing tool kit in terms of empirical effectiveness. Having concluded this, we remark a caveat extracted from Blough (1992, p. 307) whom we let summarize the above discussion:

> '...unit-root tests should not be judged by the conventional criteria of power and level, but rather by their ability to make the appropriate diagnosis in a particular application.'

Stock (1994, pp. 2828–31) differentiates the pros and cons of unit-root tests for different purposes.

---

empoyed the AR(8) model, too, in the context of quarterly macroeconomic variables. Osborne and Smith (1989) estimated an AR(16) model (with some zero restrictions) for annual nondurable consumption expenditure. Copeland (1991) used 50 lags when testing a unit root from daily exchange rates.

Chapter 4
Likelihood Theory

4.1 Likelihood and Profile Likelihood

This section gives a compact review of likelihood theory of relevance to the thesis.[1] Let us suppose that we have a statistical model $p_{Y_i}(y_i) = f(y_i;\boldsymbol{\omega})$ where Y stands for a random variable, y is an outcome or observation of it, the index i is a running number, and $\boldsymbol{\omega}$ is a $(k+1)\times 1$ vector of parameters. If there are T independent observations we have $p_{\mathbf{Y}}(\mathbf{y}) = f(\mathbf{y};\boldsymbol{\omega}) = \prod_{i=1}^{T} f(y_i;\boldsymbol{\omega})$ in an obvious notation. It is further assumed that $\boldsymbol{\omega}$ can be decomposed and interpreted as $\boldsymbol{\omega}' = [\psi \quad \boldsymbol{\phi}'] = [\psi \quad \phi_1 \ldots \phi_k]$ with ψ a scalar parameter of interest and $\boldsymbol{\phi}$ a subvector of nuisance parameters. It will become clear that it is not a coincidence that the parameters ψ and $\boldsymbol{\phi}$ coincide with those already used in the former chapter, but for the moment one should think in general terms, with no explicit connection to any specific model.

Probably the most common method of estimating $\boldsymbol{\omega}$ is the *method of maximum likelihood.* That is, a *likelihood function*

$$L(\boldsymbol{\omega};\mathbf{y}) \;=\; a(\mathbf{y})f(\mathbf{y};\boldsymbol{\omega}) \;\propto\; f(\mathbf{y};\boldsymbol{\omega}),$$

is constructed and maximised with respect to $\boldsymbol{\omega}$. $a(\mathbf{y})$ is any constant which can (but does not have to) depend on $\mathbf{y}$.[2] Usually it is theoretically and computationally more

[1] Edwards (1974) gives a fascinating history of the concept of likelihood as invented by Ronald Fisher.

[2] Observations can be treated as given or fixed for the purpose of maximizing the likelihood function with respect to the parameters.

convenient to maximise the *log-likelihood function*

$$l(\psi,\ \boldsymbol{\phi}) \ \equiv \ l(\boldsymbol{\omega};\mathbf{y}) \ = \ \log L(\boldsymbol{\omega};\mathbf{y}) \ \propto \ \log f(\mathbf{y};\boldsymbol{\omega}) \tag{4.1}$$

where the dependence on y is suppressed in the left-most form. Maximisation can be done in two steps. First, $l(\boldsymbol{\omega};\mathbf{y})$ is maximised with respect to parameters $\boldsymbol{\phi}$ regarding ψ fixed, yielding estimates $\hat{\boldsymbol{\phi}}_{\psi}\equiv\hat{\boldsymbol{\phi}}\ (\psi)$ which in general depend on ψ. Evaluating the log-likelihood function at these estimates yields the *profile log-likelihood function*[3]

$$l_p(\psi) \equiv l_p(\psi,\hat{\boldsymbol{\phi}}_{\psi}) \equiv l(\psi,\hat{\boldsymbol{\phi}}_{\psi}) = \sup_{\boldsymbol{\phi}} l(\psi\ \boldsymbol{\phi}). \tag{4.2}$$

The notation $l_p(\psi,\hat{\boldsymbol{\phi}}_{\psi})$ will be used when it is helpful to indicate the parameterisation or the nuisance parameters to which the log-likelihood corresponds. Next, the profile log-likelihood is maximised with respect to ψ. The value of ψ corresponding to the maximum is the MLE or $\hat{\psi}_{ml}$ of ψ. If needed, $\hat{\boldsymbol{\phi}}_{ml}$, the MLE of $\boldsymbol{\phi}$, is found by substituting $\hat{\psi}_{ml}$ into the formula for $\hat{\boldsymbol{\phi}}_{\psi}$, i.e. $\hat{\boldsymbol{\phi}}_{ml}=\hat{\boldsymbol{\phi}}_{\hat{\psi}_{ml}}$. (Seber and Wild (1989, Section 2.2.3), say, prove this).

If some regularity conditions are fulfilled then the MLEs $\hat{\boldsymbol{\omega}}_{ml}= [\hat{\psi}_{ml}\ \hat{\boldsymbol{\phi}}_{ml}']'$ enjoy attractive asymptotic properties which include consistency, efficiency in the sense of achieving the minimum of (generalized) variance of the estimates, and normality (cf. Cramer (1946, Chapter 33) or Stuart and Ord (1991), say). MLEs need not be efficient in finite samples, though (e.g. Judge and Bock (1983)). The usual asymptotic results are not useful from the perspective of this thesis as we shall also be interested in small-sample properties and more importantly, nonstationary processes do not fulfill the regularity conditions. Indeed, we have already seen that the asymptotic distribution of the MLE is not normal when the process has a unit root (equation (3.11)).

Profile likelihood functions share some properties with genuine likelihood functions

[3]Concentrated log-likelihood function has been a common nomenclature in econometrics after the pioneering work of the Cowles Commission, see e.g. Koopmans and Hood (1953, pp. 156-7) who give also a helpful graphical illustration of stepwise maximisation.

(cf. Barndorff-Nielsen (1991, p. 242–3)). Actually, an example of this was just given, namely that maximisation of (4.2) yields the same estimate as maximisation of (4.1). This implies that (4.2) is invariant with respect to reparameterisation like an ordinary likelihood.

Another analogy is estimation of the variance of the MLE for ψ. A common estimate is the upper left-hand corner element of the inverse of the *observed information matrix*, or of the negative of the matrix of second derivatives of the log-likelihood, evaluated at the MLE:

$$\begin{aligned}\left\{\left[-\frac{\partial^2 l(\omega;\mathbf{y})}{\partial\omega\partial\omega'}\Big|_{\omega=\hat{\omega}_{ml}}\right]^{-1}\right\}_{1,1} &\equiv \left\{\left[\mathbf{J}_{\omega\omega}(\omega)|_{\omega=\hat{\omega}_{ml}}\right]^{-1}\right\}_{1,1} \\ &= \left\{\left[j_{\psi\psi}(\omega)-\mathbf{j}_{\psi\phi}(\omega)\mathbf{J}_{\phi\phi}^{-1}(\omega)\mathbf{j}_{\phi\psi}(\omega)\right]\Big|_{\omega=\hat{\omega}_{ml}}\right\}^{-1}.\end{aligned} \tag{4.3}$$

Here the index $(1,1)$ extracts the corner element, the last equality follows from a well-known rule of a partioned inverse, and the dependence of the observed information measures on $\mathbf{y}$ is suppressed in the last two forms. The observed information matrix $\mathbf{J}_{\omega\omega}(\omega)$ is partitioned as follows:

$$\mathbf{J}_{\omega\omega}(\omega) = \begin{bmatrix} j_{\psi\psi}(\omega) & \mathbf{j}_{\psi\phi}(\omega) \\ \mathbf{j}_{\phi\psi}(\omega) & \mathbf{J}_{\phi\phi}(\omega) \end{bmatrix}$$

so, for example, $\mathbf{J}_{\phi\phi}(\omega)$ is the $k \times k$ submatrix for the nuisance parameters ϕ or

$$-\frac{\partial^2 l(\psi,\phi)}{\partial\phi\partial\phi'},$$

(with $(i,j)^{th}$ element $j_{\phi_i\phi_j}$). By analogy, if the profile log-likelihood function were treated as a genuine log-likelihood for a scalar parameter then the estimate of the variance should be found as the inverse of the minus of the second derivative of the profile log-likelihood with respect to ψ. Indeed, it can be shown to be the case so:

$$\left[-\frac{\partial^2 l_p(\psi)}{\partial^2 \psi}\bigg|_{\psi=\hat{\psi}_{ml}}\right]^{-1} = \left\{\left[j_{\psi\psi}(\omega) - \mathbf{j}_{\psi\phi}(\omega)\mathbf{J}^{-1}_{\phi\phi}(\omega)\mathbf{j}_{\phi\psi}(\omega)\right]\Big|_{\omega=\hat{\omega}_{ml}}\right\}^{-1}$$

(Seber and Wild (*op. cit.*, pp. 39–42) give a proof).

Consequently, the classical likelihood based tests or *Wald, Lagrange multiplier* (LM), and the *likelihood ratio* (LR) tests (for a parameter of interest) can be calculated conveniently via the profile log-likelihood function. For example, the Wald test statistic for the null hypothesis $\psi = \psi_0$ can be expressed either in the form

$$W \equiv (\hat{\psi}_{ml} - \psi_0)^2 \left\{\left[j_{\psi\psi}(\omega) - \mathbf{j}_{\psi\phi}(\omega)\mathbf{J}^{-1}_{\phi\phi}(\omega)\mathbf{j}_{\phi\psi}(\omega)\right]\Big|_{\omega=\hat{\omega}_{ml}}\right\},$$

or in the form

$$W = (\hat{\psi}_{ml} - \psi_0)^2 \left[-\frac{\partial^2 l_p(\psi)}{\partial^2 \psi}\bigg|_{\psi=\hat{\psi}_{ml}}\right]. \tag{4.4}$$

Having pointed out these parallelisms, it should be stated that function (4.2) is not a proper log-likelihood function. For example, it is not the log of a density function for an observable random variable. Relatedly,

$$E\left[\frac{\partial l_p(\psi)}{\partial \psi}\right] = O(1) \tag{4.5}$$

or the expectation of the derivative of $l_p(\psi, \hat{\boldsymbol{\phi}}_\psi)$ with respect to ψ is not zero in general as if the derivative were a score of a *bona fide* likelihood function. This is because the expectation of the score depends here on the random variable $\mathbf{Y}$ also through the term $\hat{\boldsymbol{\phi}}_\psi$ (cf. McCullagh and Tibshirani (1990)). The intuition is that we have replaced $\boldsymbol{\phi}$ by its estimate $\hat{\boldsymbol{\phi}}_\psi$ in the profile log-likelihood (4.2).

Profile likelihood function is an example of a *pseudo-likelihood function.* These are likelihood functions which regard the inference frame as if it was composed of only a

scalar parameter (ψ) though that was not necessarily the case.[4] However, they may differ from an ordinary likelihood function in some important respects as pointed out above (Davidson and MacKinnon (1993, p. 269), McCullagh and Tibshirani (1990, p. 330), and Seber and Wild (*op. cit.*, p. 229) point out further differences). Examples of other pseudo-likelihoods are listed in Barndorff-Nielsen (*op. cit.*, p. 256).

Profile likelihood functions are useful also in proving various properties of estimates like invariance of estimates (pp. 103 and 122) and consistency or inconsistency (Appendices A3 and A5). Profile likelihoods prove particularly beneficial when the nonlinear model of Section 6.6 is studied. Proofs for consistency and asymptotic distribution of the MLE for this model are easily derived with the help of profile likelihoods. This is the case despite recent results indicating additional complexities in such proofs in general for nonlinear models with integrated regressors (Saikkonen (1995)).

A rule of thumb is that the greater is the dimension of the nuisance parameter $\boldsymbol{\phi}$ compared to the sample size T the greater is the distortion caused by using $\hat{\boldsymbol{\phi}}_{\psi}$ instead of $\boldsymbol{\phi}$ in (4.2). As an extreme case, the MLEs of ψ may become inconsistent.

Example on estimation of the variance in the presence of incidental parameters (Neyman and Scott (1948)). Let us assume that T pairs of data $(y_{i1}, y_{i2}) \sim N(\boldsymbol{\phi}_i, \boldsymbol{\Sigma})$ are observed where $\boldsymbol{\phi}_i' = [\phi_i \ \phi_i]$ and $\boldsymbol{\Sigma} = diag[\psi \ \psi]$, $\psi > 0$. Consequently y_{i1} and y_{i2} are independent but share the same mean ϕ_i and variance ψ, taken to be the parameter of interest. The profile log-likelihood for ψ is easily found to be

$$l_p(\psi) \quad = \quad -T \log \psi - s^2/2\psi \tag{4.6}$$

where $s^2 = \sum_{i=1}^{T}(y_{i1} - y_{i2})^2/2$. Maximisation of $l_p(\psi)$ yields $\hat{\psi}_{ml} = s^2/2T$. $\hat{\psi}_{ml}$ is a harshly biased estimate as $E[\hat{\psi}_{ml}] = \psi/2$ and, even worse, it is inconsistent. This can be seen from the fact that T is not an argument in the formula for the expected value. The

[4] A likelihood function which is applied in a situation when it is not strictly valid has been called a pseudo-likelihood, too. One aims for robustness by such a device, cf. White (1982) and Gouriéroux *et al.* (1984) for the concept and Lucas (1996) for unit-root applications.

incidental parameters ϕ_i, $i = 1,\ldots,T$, the number of which increases one-to-one with the sample size, explain inconsistency of the MLE in this context.□[5]

Neyman and Scott give also a related example where the MLE is an asymptotically inefficient estimate. Cruddas *et al.* (1989) have an example of the inconsistency of the MLE in a time-series context. Davidson and MacKinnon (1993, pp. 143–4) and Hendry (1995, p. 720 ff.) point out another instance, and a fresh calculation on it is given in Section 4.3. An instance of asymptotic inefficiency of the MLE, with a close association with unit-root models, arises with the *nearly nonstationary model* (model (3.9) with ρ replaced by $\rho_T = 1 - \beta/T$, $\beta > 0$) which is analysed by Cox and Llatas (1991). Lucas (1996, Chapter 7) gives similar results in multivariate (cointegration) framework. The time-series model advocated by Bhargava (1986) induces a constant, the MLE of which is inconsistent under a unit root. The model is considered in Sections 4.3 and 6.6.

4.2 Adjustments to Profile Likelihood

The example in the previous section demonstrates that cases exist where nuisance parameters cause a deterioration in the ML method. Instead, some other procedure is needed in that sort of situation. We shall present three adjustments to profile likelihood proposed recently. All three aim to decrease the effect of using estimated values in place of true values of the nuisance parameter ϕ in the profile log-likelihood (4.2). The adjustments may be advantageous even though the applications of interest are not as drastic as the example of the previous section.

Barndorff-Nielsen (1983, 1988) presented the formula

$$l_{mod}(\psi) \;=\; l_p(\psi, \hat{\phi}_\psi) - \tfrac{1}{2}\log\left|\mathbf{J}_{\phi\phi}(\psi, \hat{\phi}_\psi)\right| + \log\left|\frac{\partial\hat{\phi}_{ml}}{\partial\hat{\phi}_\psi}\right| \qquad (4.7)$$

and coined it *modified profile log-likelihood.* $\mathbf{J}_{\phi\phi}(\psi, \hat{\phi}_\psi)$ is the observed information matrix

[5]A familiar occurrence of incidental parameters for time-series analysts is the periodogram. The periodogram is an inconsistent (though unbiased) estimate of the spectral density.

for ϕ evaluated at $\hat{\phi}_\psi$. As before, $\hat{\phi}_{ml}$ equals $\hat{\phi}_\psi$ evaluated at $\hat{\psi}_{ml}$. Intuitively, the last term echoes the effect of ψ in estimation. The second, and also the last, term can be interpreted more easily when discussing equations (4.9) and (4.12).

Modified profile likelihood is parameterisation invariant like profile likelihood. The likelihoods are also asymptotically equal to first-order. The formula (4.7) is sometimes exact in the sense of giving the conditional distribution of the MLE conditionally on an *ancillary statistic*. In other cases it can be interpreted as a higher-order approximation to a *marginal* or *conditional likelihood* for ψ.

A problem with the formula (4.7) is that the term $|\partial \hat{\phi}_{ml} / \partial \hat{\phi}_\psi|$ may be difficult to calculate. Cox and Reid (1987) pointed out that the term is zero to first-order if ψ and ϕ are (globally) orthogonal, i.e. if the Fisher information vector $\mathbf{i}_{\phi\psi}$ between ψ and ϕ is zero:

$$\mathbf{i}_{\phi\psi} = E\left[-\frac{\partial^2 l(\psi,\phi)}{\partial\phi\partial\psi}\right] = \mathbf{0},$$

(with $(i,1)^{th}$ element $i_{\phi_i\psi}$) for all ψ and ϕ in the parameter space. Consequently, the term could then be ignored with little effect. (The term may be approximately zero in other circumstances as well, see Barndorff-Nielsen and McCullagh (1993).) In general, models do not feature orthogonal parameters (with notable exceptions). Cox and Reid presented a procedure to reparameterise the model so that the new parameters were orthogonal and thus simplification of the formula (4.7) would be justified. The argument starts by expressing the original log-likelihood in terms of new parameters $\boldsymbol{\lambda}' = [\lambda_1 \ldots \lambda_k]'$ (of the same dimension as ϕ):

$$l(\psi,\phi) = l(\psi,\phi(\psi,\boldsymbol{\lambda})) = l^*(\psi,\boldsymbol{\lambda}).$$

It follows that

$$\frac{\partial l^*(\psi,\boldsymbol{\lambda})}{\partial\psi} = \frac{\partial l(\psi,\phi)}{\partial\psi} + \sum_{i=1}^{k} \frac{\partial l(\psi,\phi)}{\partial\phi_i}\frac{\partial\phi_i}{\partial\psi}$$

and

$$\frac{\partial l^*(\psi,\boldsymbol{\lambda})}{\partial\psi\partial\lambda_r} = \sum_{j=1}^{k}\frac{\partial^2 l(\psi,\phi)}{\partial\psi\partial\phi_j}\frac{\partial\phi_j}{\partial\lambda_r} + \sum_{i=1}^{k}\sum_{j=1}^{k}\frac{\partial^2 l(\psi,\phi)}{\partial\phi_i\partial\phi_j}\frac{\partial\phi_j}{\partial\lambda_r}\frac{\partial\phi_i}{\partial\psi} + \sum_{i=1}^{k}\frac{\partial l(\psi,\phi)}{\partial\phi_i}\frac{\partial^2\phi_i}{\partial\psi\partial\lambda_r}.$$

The last term of the latter equation vanishes after taking expectations. We get

$$\sum_{j=1}^{k}\frac{\partial\phi_j}{\partial\lambda_r}\left(i_{\psi\phi_j} + \sum_{i=1}^{k} i_{\phi_i\phi_j}\frac{\partial\phi_i}{\partial\psi}\right) = 0,$$

$r = 1,...,k$, where $i_{\phi_i\phi_j} = E[-\partial^2 l(\psi,\phi)/\partial\phi_i\partial\phi_j]$. The term in parentheses must be zero for the transformation from (ψ,ϕ) to (ψ,λ) to have nonzero Jacobian (otherwise the same linear combination of two different columns of the Jacobian matrix could equal zero). This implies that

$$\sum_{i=1}^{k} i_{\phi_i\phi_j}\frac{\partial\phi_i}{\partial\psi} = -i_{\psi\phi_j},$$

$j = 1,\ldots,k$. More compact matrix notation is

$$\frac{\partial\boldsymbol{\phi}}{\partial\psi} = -\mathbf{I}_{\boldsymbol{\phi}\boldsymbol{\phi}}^{-1}\mathbf{i}_{\boldsymbol{\phi}\psi} \tag{4.8}$$

where $\mathbf{I}_{\boldsymbol{\phi}\boldsymbol{\phi}}$ is the information matrix with $(i,\ j)^{th}$ component $i_{\phi_i\phi_j}$ and $\mathbf{i}_{\boldsymbol{\phi}\psi}$ is the information vector with components $i_{\phi_i\psi}$. Thus, if these differential equations are solved then the model can be expressed using orthogonal parameters. Equation (4.8) has a continuum of solutions, so a choice of an appropriate solution has to be made (cf. Cox and Reid (1989)).

After reparameterisation equation (4.7) becomes

$$l_{ad}(\psi) = l_p(\psi,\hat{\boldsymbol{\lambda}}_\psi) - \tfrac{1}{2}\log\mid\mathbf{J}_{\boldsymbol{\lambda}\boldsymbol{\lambda}}(\psi,\hat{\boldsymbol{\lambda}}_\psi)\mid \tag{4.9}$$

and is labeled *adjusted profile log-likelihood.* The second term on the right-hand side has now an obvious interpretation as a penalty term for values of ψ which produce a lot of information on $\boldsymbol{\lambda}$, the parameter (vector) we are not interested in. Of course, the

log-likelihood is shaped accordingly. The cost of the simplicity of formula (4.9) over (4.7) is forfeit of invariance to order $O_p(T^{-1})$ (in conventional circumstances).[6]

Also, adjusted profile likelihood may be hard to derive because of the complexity of the differential equation (4.8). Cox and Reid (1993) proposed an approximation to formula (4.9) which does not require orthogonalisation of parameters. They make use of the relation

$$\log \mid \mathbf{J}_{\boldsymbol{\lambda}\boldsymbol{\lambda}}(\psi, \hat{\boldsymbol{\lambda}}_\psi) \mid \;=\; \log \mid \mathbf{J}_{\boldsymbol{\phi}\boldsymbol{\phi}}(\psi, \hat{\boldsymbol{\phi}}_\psi) \mid + \left(2 \log \left|\frac{\partial \boldsymbol{\phi}'}{\partial \boldsymbol{\lambda}}\right|\right) \mid_{(\psi, \hat{\boldsymbol{\lambda}}_\psi)},$$

which is established through similar calculations to those which led to (4.8). A first-order expansion of the last term around an arbitrary point $(\psi_0, \boldsymbol{\lambda}_0)$ (or a corresponding point $(\psi_0, \boldsymbol{\phi}_0)$) substituted into formula (4.9) yields

$$\begin{aligned} l_{ad}(\psi) \;&\approx\; l_p(\psi, \hat{\boldsymbol{\phi}}_\psi) - \tfrac{1}{2} \log \mid \mathbf{J}_{\boldsymbol{\phi}\boldsymbol{\phi}}(\psi, \hat{\boldsymbol{\phi}}_\psi) \mid - \log \left(\left|\frac{\partial \boldsymbol{\phi}'}{\partial \boldsymbol{\lambda}}\right|\right) \\ &\quad -(\psi - \psi_0) \left(\frac{\partial}{\partial \psi} \log \left|\frac{\partial \boldsymbol{\phi}'}{\partial \boldsymbol{\lambda}}\right|\right) \mid_{(\psi_0, \boldsymbol{\lambda}_0)} \\ &\propto\; l_p(\psi, \hat{\boldsymbol{\phi}}_\psi) - \tfrac{1}{2} \log \mid \mathbf{J}_{\boldsymbol{\phi}\boldsymbol{\phi}}(\psi, \hat{\boldsymbol{\phi}}_\psi) \mid -(\psi - \psi_0) tr \left[\left(\frac{\partial \boldsymbol{\phi}'}{\partial \boldsymbol{\lambda}}\right)^{-1} \frac{\partial^2 \boldsymbol{\phi}'}{\partial \boldsymbol{\lambda} \partial \psi}\right] \mid_{(\psi_0, \boldsymbol{\lambda}_0)} . \end{aligned} \tag{4.10}$$

Differential and trace manipulation shows that the negative of the trace term is equal to

$$- tr \left[\frac{\partial}{\partial \boldsymbol{\phi}} \left(\frac{\partial \boldsymbol{\phi}'}{\partial \psi}\right)\right] \mid_{(\psi_0, \boldsymbol{\phi}_0)} \;=\; tr \left(\frac{\partial}{\partial \boldsymbol{\phi}} (\mathbf{I}^{-1}_{\boldsymbol{\phi}\boldsymbol{\phi}} \mathbf{i}_{\boldsymbol{\phi}\psi})'\right) \mid_{(\psi_0, \boldsymbol{\phi}_0)} \equiv \; c \tag{4.11}$$

where the equality arose from (4.8). Substituting this into the former equation and taking the point of expansion to be $[\hat{\psi}_{ml} \;\; \hat{\boldsymbol{\phi}}'_{ml}]'$, i.e. the overall MLE, we get

$$l_{ad2}(\psi) \;=\; l_p(\psi, \hat{\boldsymbol{\phi}}_\psi) - \tfrac{1}{2} \log \mid \mathbf{J}_{\boldsymbol{\phi}\boldsymbol{\phi}}(\psi, \hat{\boldsymbol{\phi}}_\psi) \mid + \hat{c}\,(\psi - \hat{\psi}_{ml}). \tag{4.12}$$

---

[6]Invariance is prevailed when $\hat{\boldsymbol{\phi}}_\psi = \hat{\boldsymbol{\phi}}_{ml}$ in which case the term $\mid \partial \hat{\boldsymbol{\phi}}_{ml} / \partial \hat{\boldsymbol{\phi}}_\psi \mid$ of formula (4.7) collapses to zero or is unnecessary. This can happen for example when the generating process for y_t belongs to the exponential family (cf. Reid (1992, pp. 429–30)).

where $\hat{c}= c\,|_{(\hat{\psi}_{ml},\hat{\phi}_{ml})}$. This is the version of adjusted likelihood derived by Cox and Reid (1993). The second term does not have as straightforward an interpretation as a penalty term in (4.12) as the corresponding term in (4.9) because the initial parameterisation is not orthogonal (by assumption). This being the case, more information on ϕ may imply more information on ψ, say. The last term compensates for this feature. If the parameters are orthogonal then $\hat{c}= 0$, as can be seen from the definition of c. Accordingly, $\hat{c}$ might be expected to be small in general when the parameterisation is close to orthogonal. The last term in (4.7) can also be interpreted in this fashion. A small cost of the apparent simplicity of (4.12) is that evaluation of it requires explicit use of expected as well as of observed information measures. However, the expected information measures need to be derived to check orthogonality even if formula (4.9) is to be used.

An intuitive benefit of calculating either of the the adjusted profile likelihoods, the latter version say, is that

$$E\left[\frac{\partial l_{ad2}(\psi)}{\partial \psi}\right] = O(T^{-1}) \tag{4.13}$$

where the convergence speed $O(T^{-1})$ holds in standard cases (Ferguson *et al.* (1991) and Cox and Reid (1993)). That is, the expectation of the derivative of $l_{ad2}(\psi, \hat{\phi}_{\psi})$ with respect to ψ tends to zero with the sample size bridging the gap between a profile likelihood function and a genuine likelihood function. Another interpretation is that the adjusted likelihood equation is more nearly an unbiased estimating equation for ψ than the unadjusted.

Though intuition suggests that $\hat{c}$ is nonzero when the parameters are nonorthogonal and small whenever the parameters are close to orthogonal, our calculations show that this is not necessarily the case. The coefficient equals zero even though the parameters are not orthogonal in an example at the end of Section 4.3, under a model considered in Section 6.6, and another one which is referred to in Section 8.1. Relatedly, the magnitude of the coefficient may behave even mirror wise to the orthogonality measure over a range of the parameter space as shown in Section 6.6.

It is worth remarking that formula (4.9) is not applicable if $\mathbf{i}_{\phi\psi} = \mathbf{0}$ only at one value

of the nuisance parameter ϕ (called local orthogonality). There is then no guarantee that the derivative $\frac{\partial}{\partial\phi}(\mathbf{I}_{\phi\phi}^{-1}\mathbf{i}_{\phi\psi})'$, and thus c (or $\hat{c}$) in equation (4.12), match zero universally. However, in principle, formulae (4.9) and (4.12) can both be employed under the assumption of a fixed value of ψ (in orthogonal and nonorthogonal circumstances, respectively). We shall study this possibility in Sections 6.3 and 6.6.

Another fact worth pointing out is that in the application we are interested in, adjusting the profile likelihood in the context of a nonstationary time-series process, standardised information measures evaluated at the MLE, and hence the Cox–Reid adjustment, do not necessarily converge to constants. Under stationarity, convergence would take place. Relatedly, it turns out in Section 6.3 that $\hat{c}$ (as such or after a standardisation to $O_p(1)$) can be asymptotically stochastic when the model is nonstationary (equation (6.28)). A further eccentric property of an expected information measure under nonstationarity is presented in the introduction to Appendix A5.

It will turn out in Chapters 6 and 7 that in the applications we are interested in, adjusted profile likelihood can possess multiple modes even in the relevant range of the parameter. In such cases, a mode occurs consistently at unity even though it were not the true value (Figures A6.1 in Appendix A6 and Figures A5.2 and A5.3 in Appendix A5). We consider at a heuristical level the suitability of the three classical tests in such a circumstance in Table 4.1.[7] Four cases are considered: *i)* H_0 is true and the MLE is consistent *ii)* H_0 is true but the MLE is inconsistent *iii)* H_0 is not true and the MLE is consistent *iv)* H_0 is not true and the MLE is inconsistent. The following discussion is based on asymptotics although this fact is not emphasized. To fix ideas, we shall assume that the null hypothesis H_0 states that $\psi = \psi_0$. We assume that the MLE converges stochastically to a constant as Fisher information on ψ grows towards infinity but the flavour of the conclusions would not change if the MLE did not converge at all.

All the tests are valid (in the sense that they do not reject) in circumstance *i)*. The LM

[7] 'True value' in the table refers to the true value of the parameter of interest ψ the value of which is hypothesized by the null hypothesis. 'Tends to' alludes to stochastic convergence as Fisher information on ψ grows towards infinity.

test is alone fit in case *ii)* or when the MLE is inconsistent. The LM test is based on the slope of the likelihood function at ψ_0 and takes value zero or does not reject H_0. Instead, the LR and the Wald tests reject erroneously and are invalid. The former measures the 'vertical' distance between the modes and the latter the 'horizontal' distance of the MLE and ψ_0 which distances are both positive. (See e.g. Buse (1982) for these interpretations of the classical tests.) The LM test is not among the suitable tests when H_0 does not hold or in situations *iii)* and *iv)* because it lacks power completely or takes value zero due to the artificial mode at ψ_0. Under situation *iii)* the power of the Wald test is unaffected by the artificial mode because the test focuses on the horizontal distance. However, the artificial mode does probably deteriorate the power of the LR test based on the 'vertical' distance measure (hence the parentheses around LR in zone *iii)*) but asymptotically the test should be adequate. In zone *iv)*, the LR and the Wald tests possess power despite inconsistency of the MLE unless the MLE converges to ψ_0 (hence the question marks in the zone).

Two conclusions can be drawn. First, the LM test displays consistently zero power so it is not useful despite the reasonable performance in situation *ii)*. Second, the LR and the Wald tests are valid in the same circumstances though the LR test might be weaker in situation *iii)*.

	The highest mode takes place at a value which	
	tends to the true value	does not tend to the true value
H_0 true	LM, LR, W	LM
H_0 false	W (LR)	LR?, W?

Table 4.1 Suitable classical tests under multimodal profile likelihoods with a mode at $H_0 : \psi = \psi_0$.

We will consider an adjusted Wald statistics in the thesis. The statistic

$$W_{ad2} \equiv (\hat{\psi}_{ad2} - \psi_0)^2 \left[-\frac{\partial^2 l_{ad2}(\psi)}{\partial^2 \psi} \bigg|_{\psi=\hat{\psi}_{ad2}} \right] \tag{4.14}$$

is constructed analogously to the usual Wald statistic (4.4) but is founded on the adjusted profile likelihood (here on the 1993 version but the 1987 version could be utilised under

orthogonality). Remark that the term in square brackets is evaluated at the adjusted estimate. The performance of (the square root of) W_{ad2} is explored in Chapters 6 and 7 in the context of AR models.

Cruddas *et al.* (1989) and An and Bloomfield (1993) have successfully employed adjusted profile likelihood in time-series contexts.

We will sometimes refer to a profile or adjusted profile likelihood function as a likelihood or adjusted likelihood function, respectively, when confusion is unlikely. Similarly, a 'likelihood equation' or an 'adjusted likelihood equation' will often be based on a profile or an adjusted profile likelihood function, respectively, instead of a likelihood function.

4.3 Examples of Adjusted Profile Likelihood

The first four examples involve models with orthogonal parameters and the last one with nonorthogonal parameters. This model is quite interesting as it appears exceptional in its class because the coefficient c shrinks to zero under this model despite the nonorthogonality. Similar results are reported at the end of Section 6.6 and in Section 8.1 where a result on a modification of the second last example is referred to. Probably more representative nonorthogonal models are considered in depth in Sections 6.3 and 6.6.

Example of Section 4.1 (on estimation of the variance in the presence of incidental parameters) revisited (Cox and Reid (1993)). It is easily seen that ψ and the incidental parameters ϕ_i, $i = 1,\ldots,T$, are orthogonal. It is also found that the observed information matrix of the nuisance parameters $\mathbf{J}_{\phi\phi}(\psi, \hat{\phi}_\psi) = diag[2/\psi \ldots 2/\psi]$ and thus $\mid \mathbf{J}_{\phi\phi}(\psi, \hat{\phi}_\psi) \mid = 2^T\psi^{-T}$. It follows that

$$\begin{aligned} l_{ad}(\psi) &= -T\log\psi - s^2/2\psi - \tfrac{1}{2}\log(2^T\psi^{-T}) \\ &\propto -\tfrac{T}{2}\log\psi - s^2/2\psi. \end{aligned} \tag{4.15}$$

The calculation of the extra terms in (4.15) compared to (4.6) yields a good return. The adjusted estimate (AE), in the sense of Cox and Reid (1987), is

$$\hat{\psi}_{ad} = \frac{s^2}{T}$$

which is unbiased and consistent. Accordingly, the adjustment of Cox and Reid (1987) to the profile likelihood is a very effective treatment in this case.□

Example on estimation of the variance parameter of the AR(1) model. Consider the simple AR(1) process

$$y_t = \phi y_{t-1} + \epsilon_t$$

where $\phi \in (-1, 1]$, $\epsilon_t \sim$NID$(0, \psi)$, $\psi > 0$, $t = 1, \ldots, T$ and y_0 is a fixed constant. This model will be studied in more detail in Section 6.2 (with different notation) but it is used already here for illustrative purposes. As the notation suggests, the variance ψ is the parameter of interest and the autoregressive coefficient ϕ is the nuisance parameter. The profile log-likelihood is

$$l_p(\psi) = -\tfrac{T}{2}\log\psi - (2\psi)^{-1}\textstyle\sum_{t=1}^{T}(y_t - \hat{\phi}_\psi\, y_{t-1})^2$$

where $\hat{\phi}_\psi = \hat{\phi}_{ml} = (\sum_{t=2}^{T} y_{t-1}^2)^{-1}\sum_{t=2}^{T} y_t y_{t-1}$, which does not depend on ψ at all. Thus it is not surprising that the parameters are orthogonal (cf. Section 6.2).[8] It is easily shown that $|j_{\phi\phi}(\psi, \hat{\phi}_\psi)| = \psi^{-1}\sum_{t=2}^{T} y_{t-1}^2$. We are ready to construct the adjusted profile log-likelihood of Cox and Reid (1987):

$$\begin{aligned} l_{ad}(\psi) &= -\tfrac{T}{2}\log\psi - (2\psi)^{-1}\textstyle\sum_{t=1}^{T}(y_t - \hat{\phi}_{ml}\, y_{t-1})^2 + \tfrac{1}{2}\log\psi - \tfrac{1}{2}\log\textstyle\sum_{t=2}^{T} y_{t-1}^2 \\ &\propto -\tfrac{(T-1)}{2}\log\psi - (2\psi)^{-1}\textstyle\sum_{t=1}^{T}(y_t - \hat{\phi}_{ml}\, y_{t-1})^2. \end{aligned}$$

It is now apparent that

[8] Orthogonality of parameters implies that the MLEs of the parameters are asymptotically independent.

$$\hat{\psi}_{ad} = \frac{\sum_{t=1}^{T}(y_t - \hat{\phi}_{ml}\, y_{t-1})^2}{T-1}.$$

Thus the adjustment produces a degrees-of-freedom correction in this application (as $\hat{\psi}_{ml} = T^{-1}(T-1)\,\hat{\psi}_{ad}$). If y_t is not autocorrelated, or $\phi = 0$, then $\hat{\psi}_{ad}$ is an unbiased estimate of the variance. As is well known, $\hat{\psi}_{ad}$ features then a larger variance and MSE than $\hat{\psi}_{ml}$. In general, the exact distribution of either $\hat{\psi}_{ml}$ or $\hat{\psi}_{ad}$ has not been derived as far as we know.□

Example on estimation of the variance parameter of the AR(1) model with constant. The previous model is typically extended to possess a constant (α):

$$y_t = \alpha + \phi y_{t-1} + \epsilon_t,$$

where y_0 is any fixed constant, but otherwise the above restrictions on the model are assumed to be in force. Also this model will be perused later (Section 6.3). The variance ψ is again taken as the parameter of interest; the nuisance parameters are denoted by $\phi' = [\alpha\ \phi]$. The profile log-likelihood is now:

$$l_p(\psi) = -\tfrac{T}{2}\log\psi - (2\psi)^{-1}\sum_{t=1}^{T}(y_t - \hat{\alpha}_\psi - \hat{\phi}_\psi\, y_{t-1})^2.$$

Here

$$\hat{\alpha}_\psi = \hat{\alpha}_{ml} = T^{-1}\sum_{t=1}^{T} y_t - \hat{\phi}_{ml}\, T^{-1}\sum_{t=1}^{T} y_{t-1} \equiv \bar{y} - \hat{\phi}_{ml}\bar{y}_{-1}$$

and

$$\hat{\phi}_\psi = \hat{\phi}_{ml} = \frac{\sum_{t=1}^{T}(y_t - \bar{y})(y_{t-1} - \bar{y}_{-1})}{\sum_{t=1}^{T}(y_{t-1} - \bar{y}_{-1})^2}$$

neither of which depend on ψ. It is easy to show that the nuisance parameters are again orthogonal to ψ (cf. p. 75 where ψ stands for the autoregressive parameter and σ^2 for

the variance). Uncomplicated calculations give

$$\begin{aligned}\left|\mathbf{J}_{\phi\phi}(\psi,\hat{\phi}_\psi)\right| &= \begin{vmatrix} \psi^{-1}T & \psi^{-1}\sum_{t=1}^{T} y_{t-1} \\ \psi^{-1}\sum_{t=1}^{T} y_{t-1} & \psi^{-1}\sum_{t=1}^{T} y_{t-1}^2 \end{vmatrix} \\ &= \psi^{-2}\left[T\sum_{t=1}^{T} y_{t-1}^2 - (\sum_{t=1}^{T} y_{t-1})^2\right],\end{aligned}$$

so

$$-\tfrac{1}{2}\log \mid \mathbf{J}_{\phi\phi}(\psi,\hat{\phi}_\psi) \mid \;=\; \log\psi - \tfrac{1}{2}\left[T\sum_{t=1}^{T} y_{t-1}^2 - (\sum_{t=1}^{T} y_{t-1})^2\right].$$

The adjusted profile log-likelihood turns out to be

$$\begin{aligned} l_{ad}(\psi) &= -\tfrac{T}{2}\log\psi - (2\psi)^{-1}\textstyle\sum_{t=1}^{T}(y_t - \hat{\alpha}_{ml} - \hat{\phi}_{ml}\, y_{t-1})^2 + \log\psi \\ &\quad -\tfrac{1}{2}\left[T\textstyle\sum_{t=1}^{T} y_{t-1}^2 - (\sum_{t=1}^{T} y_{t-1})^2\right] \\ &\propto -\tfrac{(T-2)}{2}\log\psi - (2\psi)^{-1}\textstyle\sum_{t=1}^{T}(y_t - \hat{\alpha}_{ml} - \hat{\phi}_{ml}\, y_{t-1})^2. \end{aligned}$$

The maximum of this log-likelihood lies above

$$\hat{\psi}_{ad} \;=\; \frac{\sum_{t=1}^{T}(y_t - \hat{\alpha}_{ml} - \hat{\phi}_{ml}\, y_{t-1})^2}{T-2}.$$

The outcome is a new degrees-of-freedom correction (since $\hat{\psi}_{ad} = (T-2)^{-1}T\,\hat{\psi}_{ml}$).

It is quite intuitive that introduction of a constant — which may have a visually stationarising impact on the series or equal zero though it is estimated, say — may lead to an overly good fit and downward bias in the MLE of ψ motivating an adjustment to the MLE. Indeed, the simulations by Orcutt and Winokur (1969) suggest that this is the case and that the above degrees-of-freedom correction produces an approximately unbiased estimate for a wide range of values of ϕ. However, the correction does not seem to be able to master the bias for values of ϕ close to one (cf. the *op. cit.*).[9]□

[9]The simulation results of Orcutt and Winokur (*op. cit.*) do not apply quite straightforwardly here though the complications may be unimportant. They assumed that already the starting value follows the stationary distribution (when $\mid \phi \mid < 1$), as opposed to being a fixed constant as in the present model. (If

Example on the estimation of the constant of the AR(1) model with constant due to Bhargava (1986). We will give only a sketch of the derivations to avoid repetition with Section 6.6 and Appendix A3 where the model is studied in more depth (with a different notation).

The model is

$$y_t = \begin{cases} \psi + \phi x_0 + \epsilon_1 & \text{for } t = 1 \\ \psi(1-\phi) + \phi y_{t-1} + \epsilon_t & \text{for } t = 2, \ldots, T, \end{cases}$$

where $\phi \in (-1, 1]$, $\epsilon_t \sim \text{NID}(0, \sigma^2)$, $\sigma^2 > 0$, and $x_0 = 0$. The constant ψ is the parameter of interest. The vector of nuisance parameters (the vector of the autoregressive coefficient ϕ and the variance parameter σ^2) is denoted by $\boldsymbol{\phi}' = [\phi \ \sigma^2]$.

The profile log-likelihood is

$$l_p(\psi) \propto -\frac{T}{2} \log \hat{\sigma}^2_\psi \tag{4.16}$$

where

$$\begin{aligned} \hat{\sigma}^2_\psi &= T^{-1} \left\{ (y_1 - \psi)^2 + \textstyle\sum_{t=2}^{T} \left[y_t - \psi - \hat{\phi}_\psi \, (y_{t-1} - \psi) \right]^2 \right\} \\ &= T^{-1} \left\{ \textstyle\sum_{t=1}^{T} (y_t - \psi)^2 - \dfrac{\left[\sum_{t=2}^{T} (y_t - \psi)(y_{t-1} - \psi) \right]^2}{\sum_{t=2}^{T} (y_{t-1} - \psi)^2} \right\} \end{aligned} \tag{4.17}$$

$\psi = 1$ then the specification of the starting value is immaterial as $\hat{\phi}_{ols}$ is invariant with respect to it, cf. p. 103 accounting for the different notation there.) For example, Evans and Savin (1981) demonstrate that the distribution of $\hat{\phi}_{ols}$ is affected by a fixed or a stationary starting value of the process (the effect on the mean value is likely to be moderate for a reasonable T). Moreover, as remarked by the authors themselves, the stationarity of the starting value in the simulations is in doubt for experiments with ϕ close to ± 1 as only the first five observations were discarded of each generated series (with starting value zero).

Another matter to clarify is Orcutt's and Winokur's statement (*op. cit.*, p. 3) that they can assume without loss of generality that $\alpha = 0$ in their simulation experiments. Actually, the value of the constant α can affect the distribution profoundly if $\phi = 1$ — a value included in their simulation experiments. With the benefit of hindsight we know that $\hat{\phi}_{ml}$ converges at a rate of $O_p(T^{-3/2})$ if $\alpha \neq 0$ instead of $O_p(T^{-1})$ which applies if $\alpha = 0$ (Dickey (1976), Evans and Savin (1984) and West (1988b)). So asymptotics suggests that the bias in the estimation of ϕ, and hence in ψ, disappears faster if $\alpha \neq 0$ than if $\alpha = 0$. A counter remark is that α may have little effect in small samples unless it is very large in absolute value relative to σ (Hylleberg and Mizon (1989)).

and

$$\hat{\phi}_{\psi} = \frac{\sum_{t=2}^{T}(y_t - \psi)(y_{t-1} - \psi)}{\sum_{t=2}^{T}(y_{t-1} - \psi)^2}.$$

The likelihood equation for ψ

$$\frac{\partial l_p(\psi)}{\partial \psi} = \frac{-T}{2\,\hat{\sigma}^2_{\psi}} \frac{\partial\, \hat{\sigma}^2_{\psi}}{\partial \psi} = 0 \tag{4.18}$$

presents an equation equivalent to a polynomial of degree five for which analytical solutions are not available in general. Hence the MLE of ψ has to be found by numerical maximisation of likelihood function (4.16). However, it is easy to derive that the estimate of the autoregressive coefficient as a function of ϕ is

$$\hat{\psi}_{\phi} = \frac{y_1 + (1-\phi)\sum_{t=2}^{T}(y_t - \phi y_{t-1})}{1 + (T-1)(1-\phi)^2}. \tag{4.19}$$

The formula reveals that $\hat{\psi}_{\phi}$ equals y_1 when evaluated at $\phi = 1$. The constant becomes then a kind of dummy and the model has some resemblance to the Neyman–Scott model of Section 4.1 in the sense that a sole observation (the first) carries all of the information on a parameter. Indeed, we show in Appendix A3 that $\hat{\psi}_{ml}$ equals y_1 asymptotically. This implies, after noting that $y_1 = \psi + \epsilon_1$, that the MLE of ψ is asymptotically unbiased but inconsistent under a unit root. In fact, ψ is not identified under a unit root. No adjustment can be expected to remove such an identification problem but we will check anyway how the Cox–Reid adjustment operates in such a circumstance.

Under stationarity, usual asymptotic theory should apply and imply consistency and asymptotic unbiasedness. In finite samples the MLE is likely to be biased unless $\phi = 0$ because this is usually the case with MLEs for parameters of time-series models.

All the relevant information measures are recorded in Section 6.6. The constant is found there orthogonal to the nuisance parameters so the theory of Cox and Reid (1987) applies. Some algebra shows that the observed information measure $j_{\phi\sigma^2}$ collapses to zero

when evaluated at $\hat{\phi}_\psi$, and we find the expression

$$-\tfrac{1}{2}\log \mid \mathbf{J}_{\phi\phi}(\psi, \hat{\phi}_\psi) \mid \;=\; \tfrac{3}{2}\log \hat{\sigma}^2_\psi - \tfrac{1}{2}\textstyle\sum_{t=2}^{T}(y_{t-1}-\psi)^2 + \tfrac{1}{2}\log 2 - \tfrac{1}{2}\log T$$

for the determinant term of the adjusted profile log-likelihood. An adjustment to the profile log-likelihood emerges:

$$\begin{aligned} l_{ad}(\psi) &= -\tfrac{T}{2}\log \hat{\sigma}^2_\psi + \tfrac{3}{2}\log \hat{\sigma}^2_\psi - \tfrac{1}{2}\textstyle\sum_{t=2}^{T}(y_{t-1}-\psi)^2 + \tfrac{1}{2}\log 2 - \tfrac{1}{2}\log T \\ &\propto -\frac{(T-3)}{2}\log \hat{\sigma}^2_\psi - \frac{1}{2}\log \textstyle\sum_{t=2}^{T}(y_{t-1}-\psi)^2. \end{aligned} \tag{4.20}$$

A completely new additive term is included relative to the profile log-likelihood (4.16). Also the multiplier of the term $\log \hat{\sigma}^2_\psi$ is modified apparently according to the number of parameters in the model — which is one more than is 'estimated' so far. Like the MLE, also the AE has to be found by numerical methods because an analytic solution does not seem to exist to the corresponding adjusted likelihood equation (cf. equation (4.18)).

The first logarithmic quantity is multiplied by an $O(T)$ term so the additive adjustment term fades in importance as the number of observations tends to infinity. This holds whether the model is stationary or nonstationary. Intuitively, an adjustment emerges in finite samples when the MLE is biased but does not asymptotically when no bias exists.□

Example on the estimation of the coefficient of the inverse time trend model. Let the model be

$$y_t \;=\; \alpha + \psi t^{-1} + \epsilon_t$$

where $\epsilon_t \sim \text{NID}(0, \sigma^2)$, $\sigma^2 > 0$ and $t = 1,\ldots,T$. (Davidson and MacKinnon (1993, pp. 143–4) and Hendry (1995, p. 720 ff.) study properties of OLS estimates under a related model.) As before, the parameter of interest is ψ and the nuisance parameters

are collected into the vector $\phi' = [\alpha\ \sigma^2]$. As in the previous example (under a unit root), the parameter of interest is not identified asymptotically.

The OLS estimate and the MLE

$$\hat{\psi}_{ml} = \frac{T\sum_{t=1}^{T} y_t t^{-1} - \sum_{t=1}^{T} y_t \sum_{t=1}^{T} t^{-1}}{T\sum_{t=1}^{T} t^{-2} - (\sum_{t=1}^{T} t^{-1})^2}$$

are the same for this model. The regressor is deterministic so standard least-squares theory applies to the model and it can be inferred that the MLE of ψ is unbiased. The formula for the variance of $\hat{\psi}_{ml}$ is

$$var(\hat{\psi}_{ml}) = \frac{T\sigma^2}{T\sum_{t=1}^{T} t^{-2} - (\sum_{t=1}^{T} t^{-1})^2} \underset{T\to\infty}{\longrightarrow} \frac{6\sigma^2}{\pi^2}$$

where $\underset{T\to\infty}{\longrightarrow}$ stands for convergence as T tends to infinity. The variance of the MLE tends hence to a non-zero constant and the MLE is inconsistent. The information on ψ carried by the observations converges, too, which explains why the non-standardised MLE features a positive variance even asymptotically.

One might wonder whether the Cox–Reid adjustment would be beneficial in this context. The answer turns out to be no. The ingredients for, and the derivation of, the AE are recorded below. The profile log-likelihood is

$$l_p(\psi) \propto -\frac{T}{2}\log \hat{\sigma}^2_{\psi}$$

where

$$\hat{\sigma}^2_{\psi} = T^{-1}\sum_{t=1}^{T}(y_t - \hat{\alpha}_{\psi} - \psi t^{-1})^2$$

and

$$\hat{\alpha}_{\psi} = T^{-1}\sum_{t=1}^{T} y_t - \psi T^{-1}\sum_{t=1}^{T} t^{-1}.$$

The nuisance parameter σ^2 or variance is orthogonal to both the parameter of interest ψ

and the nuisance parameter α or constant whereas the latter two are not orthogonal to each other:

$$i_{\psi\sigma^2} = E\left[\sigma^{-4}\textstyle\sum_{t=1}^{T}(y_t - \alpha - \psi t^{-1})t^{-1}\right] = 0,$$

$$i_{\alpha\sigma^2} = E\left[\sigma^{-4}\textstyle\sum_{t=1}^{T}(y_t - \alpha - \psi t^{-1})\right] = 0,$$

and

$$i_{\psi\alpha} = E\left[\sigma^{-2}\textstyle\sum_{t=1}^{T} t^{-1}\right] = \sigma^{-2}\textstyle\sum_{t=1}^{T} t^{-1},$$

respectively. The theory of Cox-and Reid (1993) which allows for nonorthogonality should hence be applied. The coefficient c of the adjusted likelihood (cf. equations (4.11) and (4.12)) is here:

$$\hat{c} = tr\left(\tfrac{\partial}{\partial\phi}(\mathbf{I}_{\phi\phi}^{-1}\mathbf{i}_{\phi\psi})'\right)\Big|_{(\hat{\psi}_{ml},\hat{\phi}_{ml})} = tr\left\{\frac{\partial}{\partial\phi}\left(\begin{bmatrix}\sigma^{-2}T & 0\\ 0 & T/2\sigma^4\end{bmatrix}^{-1}\begin{bmatrix}T\sigma^{-2}\sum_{t=1}^{T} t^{-1}\\ 0\end{bmatrix}\right)'\right\}$$

$$= tr\left(\begin{bmatrix}0 & 0\\ 0 & 0\end{bmatrix}\right) = 0.$$

The coefficient c happens to equal zero even though the vector of nuisance parameters is not orthogonal to the parameter of interest.

The term based on the observed information matrix is

$$
\begin{aligned}
-\tfrac{1}{2}\log \mid \mathbf{J}_{\phi\phi}(\psi,\hat{\phi}_\psi) \mid &= -\tfrac{1}{2}\log \begin{vmatrix} \hat{\sigma}_\psi^{-2}\, T & \hat{\sigma}_\psi^{-4} \sum_{t=1}^{T}(y_t - \hat{\alpha}_\psi - \psi t^{-1}) \\ \hat{\sigma}_\psi^{-4} \sum_{t=1}^{T}(y_t - \hat{\alpha}_\psi - \psi t^{-1}) & T/2\, \hat{\sigma}_\psi^{2} \end{vmatrix} \\
&= -\tfrac{1}{2}\log \begin{vmatrix} \hat{\sigma}_\psi^{-2}\, T & 0 \\ 0 & T/2\, \hat{\sigma}_\psi^{2} \end{vmatrix} \\
&= \tfrac{3}{2}\log \hat{\sigma}_\psi^{2} - \tfrac{1}{2}\log(T^2/2).
\end{aligned}
$$

The adjusted profile likelihood of Cox and Reid (1993) collapses to the simpler version (1987) because $c = 0$ as demonstrated above. The adjusted likelihood function is thus

$$
l_{ad2}(\psi) = l_{ad}(\psi) \propto -\tfrac{T}{2}\log \hat{\sigma}_\psi^2 + \tfrac{3}{2}\log \hat{\sigma}_\psi^2 = -\frac{(T-3)}{2}\log \hat{\sigma}_\psi^2 .
$$

The maximum of this function arises at the same value of ψ as the maximum of the usual profile log-likelihood so $\hat{\psi}_{ml}$=$\hat{\psi}_{ad2}$. That is to say, there is no adjustment. A possible reason for the impotency of the Cox–Reid theory for this model is that the MLE is unbiased in the first place.□

The cases we have investigated can be classified by the existence of bias and an adjustment as follows:

i) A bias and an adjustment exist in both finite and large samples in the

- estimation of the variance parameter in the presence of incidental mean parameters

ii) A bias and an adjustment exist only in finite samples in the

- estimation of the variance parameter of the AR(1) model
- estimation of the variance parameter of the AR(1) model with constant
- estimation of the constant of the AR(1) model with constant due to Bhargava

iii) No bias nor an adjustment arise at all in the

- estimation of the coefficient of the inverse of the time trend.

The adjustment appears to relate to the existence of the bias of the MLE according to these examples. However, the last two models asymptotically face exceptional identification problems of the parameter of interest and are thus far from being prototype models. Furthermore, analysis of the autoregressive time-series models in Chapter 6, digested in Section 8.1, will indicate that existence of bias does not alone explain the adjustment.

The adjustment is not necessarily able to, and sometimes cannot be expected to, account for inconsistency. The adjustment can be very efficacious but, perhaps not surprisingly, it is not a panacea to all problems associated with the method of maximum likelihood. Indeed, Cox and Reid (1992) give an example in which the adjustment yields a decrease in bias but does not achieve consistency.

Chapter 5

On Biases in the Estimation of Autoregressive Models

5.1 Introductory Remarks

The Cox–Reid (1987, 1993) adjustment often appears to correct for the bias (recall the examples on pp. 41–43). Impetus for bias removal in the contexts of time-series and unit-root analyses, insight into the biases arising in these contexts, and contemplation of related theory is presented in this chapter. The exposition focuses on the aspects deemed germane to the theme of the thesis. Hence many aspects, including the detailed results of Evans and Savin (1981, 1984) on unit-root AR(1) models and, to a large extent, the more general results of MacKinnon and Smith (1996) are passed over. The adjustments of Andrews (1993) and Andrews and Chen (1994) for median-unbiasedness are not contemplated either.

The theoretical insight gained aids us later in assessing the extent to which the Cox–Reid adjustment focuses on the bias and other features of the MLEs under study (cf. the discussion on p. 80 on the adjustment for the AR(1) model with constant and p. 132 for the case of the AR(2) model). The corrections of the present chapter also supply benchmarks which enable comparisons to adjustments to be introduced in Chapter 6.

5.2 Incitements for Bias Corrections

There are many reasons to bias correct in time-series contexts on top of the simple intuitive idea of hitting the true value in estimation on average. These include achieving an adequate agreement with the asymptotic distributions and the (usually unknown) small-

sample distributions of the estimates and related test statistics, evaluation of prediction accuracy[1], autoregressive spectral estimation, and measuring the permanent effect of the innovations of the Wold (1938) decomposition on the level of the time series or persistence in the unit-root literature (see p. 132). (Shaman and Stine (1988) supply further references and examples).

As regards unit-root analysis, there are often many nuisance parameters which have to be estimated as made clear in Chapter 3. Further, some nuisance parameters, like γ_0 and γ_1 in model (3.6) induce a notable bias to the MLEs and increase their variance (Fuller (1976) and Section 3.2). Others, like θ in model (3.21) cause size distortions to the tests. The fact that nuisance parameters are a major reason for the biases makes the Cox–Reid adjustments especially attractive in this setting.

Abadir (1995) claims an additional motivation for adjusting the MLE in the context of unit-root testing. He minimizes the MSE of estimates of the form

$$a_T \hat{\rho}_{ml} \tag{5.1}$$

with respect to the multiplying constant a_T. Here $\hat{\rho}_{ml}$ is the usual MLE, and a_T depends on the sample size (T). It is assumed that the model is the simple AR(1) model with a unit root ($\rho = 1$) (equation (5.2)). (Chapter 6 provides analytical expressions for the MLEs for the AR(1) and AR(2) models.) The MSE minimizing a_T turns out to produce an approximately unbiased estimate, and to be in general larger than unity for samples up to two hundred observations, at least. Hence Abadir concludes that bias-reducing techniques seem to be needed to improve present unit-root tests (based on biased estimates).

It may be remarked that eliminating the bias does not *necessarily* improve unit-root tests. The bias of the MLE could be abolished under the unit-root hypothesis by

[1]The MSE of the prediction error depends greatly on the process, so knowledge of the latter is informative on the former. (For example the size of the autoregressive parameter influences the MSE in the AR(1) model.) However, unbiased estimates of the parameters do not necessarily lead to better *predictions* in terms of MSE (Orcutt and Winokur (1969) provide an example).

multiplying the MLE as in $a_T \hat{\rho}_{ml}$ above, or by adding a sample size dependent constant (a_T^*) to the MLE as in $\hat{\rho}_{ml} + a_T^*$. A combination of the procedures is possible too and has been realized in the more universally corrected estimate $\hat{\rho}_b$ of equation (5.6) underneath. These techniques rely on the sample information solely through the MLE and hence induce the same partition of the sample space. The consequence is that no increase in power can take place if these statistics are used as a sole basis for a unit-root test. Other methods are required. Besides, the above kind of a modification could easily lead to an increase in the bias for smaller values of ψ if the modification produced an unbiased estimate under a unit root.

5.3 Biases in the Estimation of Stationary Autoregressive Models

It is only recently that simple analytical formula for the biases of the MLEs for stationary AR(p) models have been discovered (Shaman and Stine (1988), cf. also Stine and Shaman (1989)).[2] An uncomplicated expression has long been known for the AR(1) model alone (Marriot and Pope (1954), Kendall (1954)). (Tjøstheim and Paulsen (1983), among others, had examined more complex models like the AR(2) model but they did not come up with as simple formulae as Shaman and Stine (1988) did.) Our interest lies especially in the bias formulae for the AR(1) model (5.4) with constant and the AR(2) model (5.7) for which we shall derive the Cox–Reid adjusted estimates in Chapter 6.

We shall consider first, for illustrative purposes, the simple AR(1) model:

$$y_t = \rho y_{t-1} + \epsilon_t \tag{5.2}$$

where $\rho \in (-1, 1)$, $\epsilon_t \sim \text{IID}(0, \sigma^2)$, $\sigma^2 > 0$, $t = 1, \ldots, T$ and the starting value y_0 is a fixed constant. Marriot and Pope (*op. cit.*) (and Shaman and Stine (1988)) proved that the

[2]The Mathematica (Wolfram (1991)) package of Stine (1992) gives, among other things, the (approximate) biases of the MLEs of the coefficients of the AR(p) model.

expected value of the MLE of the autoregressive coefficient ρ is

$$E(\hat{\rho}_{ml}) = \rho - \frac{2\rho}{T} + O(T^{-2}). \tag{5.3}$$

It can be seen that $\hat{\rho}_{ml}$ is unbiased (to order $O_p(T^{-2})$) only if $\rho = 0$ and is biased towards zero otherwise, that the absolute value of the bias increases with the absolute value of ρ, and that the bias decreases at a rate of $O(T^{-1})$. The relative bias remains constant at $(\pm)2/T$ percent for all ρ different from zero which is not too bad in general. Evans and Savin (1981, Table III) report numerically calculated exact values of the bias for specific values of ρ, T, and y_0/σ.[3] We shall comment on the impact of the starting value at the end of this section.

Let us import a constant (α) to the model (which remains otherwise unchanged):

$$y_t = \alpha + \rho y_{t-1} + \epsilon_t. \tag{5.4}$$

The expected value of the MLE is then:

$$E(\hat{\rho}_{ml}) = \rho - \frac{1+3\rho}{T} + O(T^{-2}) \tag{5.5}$$

(Marriot and Pope (*op. cit.*), Kendall (*op. cit.*), and Shaman and Stine (1988)). Estimation of the constant has engendered a more severe bias for all nonnegative ρ. The symmetry of bias around zero is lost with the introduction of a constant. Positive values of ρ imply a greater bias (in absolute value) than the corresponding negative values. Indeed, the bias vanishes now for $\rho = -1/3$ illustrating that if ρ is negative then the bias can be smaller with the present model than with the simple AR(1) model. The bias decreases in relative terms as ρ tends from zero to one, but even for large values of ρ, at least twice the sample size seems to be needed to constrain the bias to its pre-constant level. The magnitude of the constant does not affect the bias. Nankervis and Savin

[3]Perron (1991, p. 228) points out two small errors in the Table.

(1988) tabulate some numerically calculated exact values for the bias.

Copas (1966) proposed the estimator

$$\hat{\rho}_b = \hat{\rho}_{ml} + \frac{1 + 3\,\hat{\rho}_b}{T} \quad \text{or} \quad \hat{\rho}_b = \frac{T\,\hat{\rho}_{ml} + 1}{T - 3}, \tag{5.6}$$

which is, of course, based on the above expression for the bias of $\hat{\rho}_{ml}$ (the subscript b stands for a bias-corrected estimator). It is straightforward to show that $\hat{\rho}_b$ is unbiased to order $O(T^{-2})$.

Let us consider next the AR(2) model

$$y_t = \rho_1 y_{t-1} + \rho_2 y_{t-2} + \epsilon_t \tag{5.7}$$

where $\epsilon_t \sim \text{IID}(0, \sigma^2)$, $\sigma^2 > 0$, $t = 1, ..., T$, and the starting values y_0 and y_{-1} are fixed constants. The (asymptotic) stationarity conditions

$$\begin{aligned} \rho_1 + \rho_2 &< 1, \\ \rho_2 - \rho_1 &< 1, \quad \text{and} \\ -1 < \rho_2 &< 1 \end{aligned} \tag{5.8}$$

are also assumed to be in force (Box and Jenkins (1976, p. 58)). The conditions define a triangle which is graphed in Figure 5.1. (The line $\rho_2 = -(1 + \rho_1)/3$ in the figure does not relate to the above conditions. The unit-root line $\rho_2 = 1 - \rho_1$ is bolded to point out a parameter region which we shall study in Section 6.6.)

The MLEs of the autoregressive coefficients are again biased, with expectations given as:

$$\begin{aligned} E(\hat{\rho}_{1_{ml}}) &= \rho_1 - \frac{\rho_1}{T} + O(T^{-2}), \\ E(\hat{\rho}_{2_{ml}}) &= \rho_2 - \frac{1 + 3\rho_2}{T} + O(T^{-2}), \end{aligned} \tag{5.9}$$

and hence

$$E(\hat{\rho}_{1_{ml}} + \hat{\rho}_{2_{ml}}) = \rho_1 + \rho_2 - \frac{1 + \rho_1 + 3\rho_2}{T} + O(T^{-2}) \tag{5.10}$$

(Shaman and Stine (1988)). The bias of the MLE of the ρ_1 coefficient lies in $(-2/T, 2/T)$ while the bias of the MLE of the ρ_2 coefficient lies in $(-4/T, 2/T)$ when the parameters fulfill the stationarity conditions (formulae (5.8)). Consequently, the MLE of the parameter ρ_2 can introduce more bias to the MLE of the sum than the MLE of the parameter ρ_1.

The stationarity conditions imply that the sum of the autoregressive coefficients $\rho_1+\rho_2$ must be smaller than unity with the line $\rho_1 + \rho_2 = 1$ corresponding to a unit root. We can see the possibly unexpected outcome that $E(\hat{\rho}_{1_{ml}} + \hat{\rho}_{2_{ml}})$ may be smaller or *larger* than $\rho_1 + \rho_2$ even when the sum lies relatively close to one. In other words, the biases have a joint effect which may push the estimates in either a stationary or a nonstationary direction.[4] The bias in the estimation of the quantity $\rho_1 + \rho_2$ appears zero along the line $\rho_2 = -(1 + \rho_1)/3$ and positive, or towards one, if $\rho_2 < -(1 + \rho_1)/3$. If $\rho_1 > 0$ then the condition implies that the roots of the characteristic equation have to be complex for the process to be stationary. The unit-root line $\rho_2 = 1 - \rho_1$ and the zero-bias line $\rho_2 = -(1 + \rho_1)/3$ are indicated in Figure 5.1.

The bias seems to be negative and hence towards stationarity in the immediate neighbourhood of the unit-root line $\rho_1+\rho_2 = 1$. Substituting the unit-root condition $\rho_2 = 1-\rho_1$ into formula (5.10) suggests that

$$E(\hat{\rho}_{1_{ml}} + \hat{\rho}_{2_{ml}})|_{\rho_1+\rho_2\approx 1} \approx 1 - \frac{4 - 2\rho_1}{T} \leq 1, \quad 0 < \rho_1 < 2. \tag{5.11}$$

The bias, along or close to the unit-root line, appears to culminate for ρ_1 close to zero (with $\rho_2 = 1$) and fade to zero around $\rho_1 = 2$ (with $\rho_2 = -1$) where the $I(2)$ process $\Delta^2 y_t = \epsilon_t$ emerges. Also formula (5.10) suggests that the bias would fade for values of ρ_1 tending to two and values of ρ_2 tending to minus one.

[4] A bias towards nonstationarity is possible with higher-order autoregressions as well.

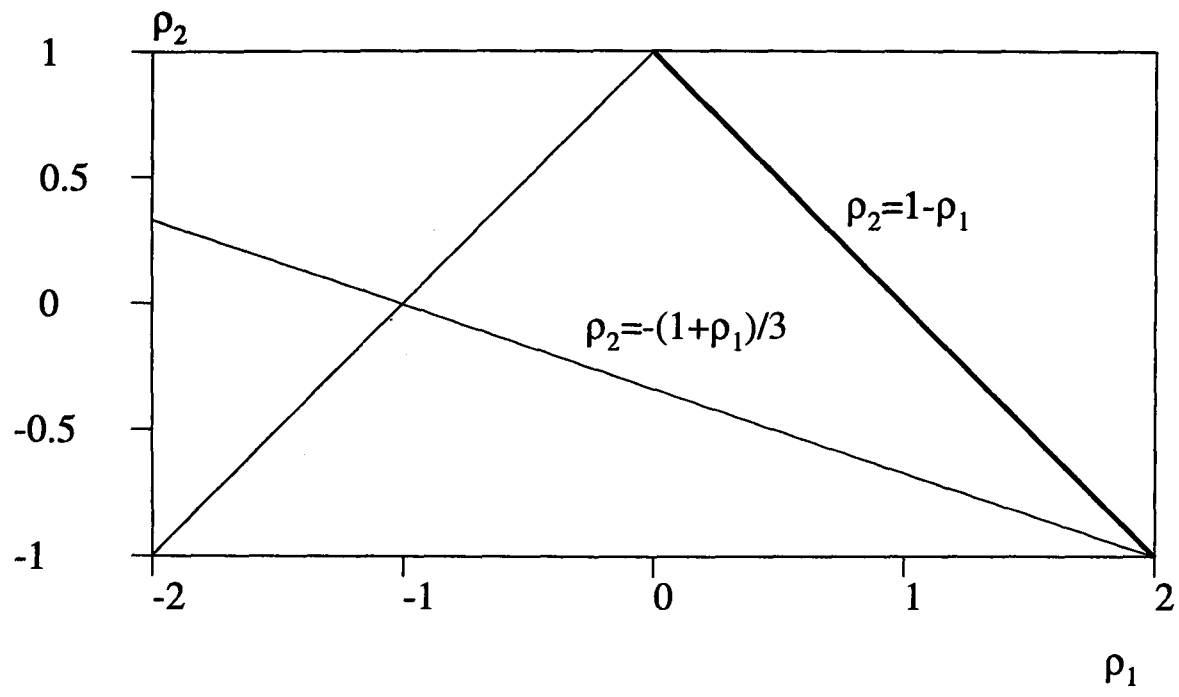


Figure 5.1 The stationarity triangle, the zero-bias line, and the unit-root line for ρ_1 and ρ_2 of the AR(2) model.

Formulae (5.10) and (5.11) imply the following pattern for the bias in the triangle defining the stationarity region (a further glance at Figure 5.1 may aid comprehension). The bias is positive and peaks at $(\rho_1, \rho_2) = (-2, -1)$ starting at about $4/T$ from which it collapses to zero as the ρ_1 parameter is tuned from -2 to 2 along the line $\rho_2 = -1$. The bias becomes negative as the parameters are varied along the line $\rho_1 + \rho_2 = 1$ and the bias deteriorates to about $-4/T$ around $(\rho_1, \rho_2) = (0, 1)$.[5] The line from $(0, 1)$ to $(-2, -1)$ closes the triangle. The plane defining the bias is steepest along this line: the bias rises from $-4/T$ gradually to $4/T$ as the parameters are revolved from $(0, 1)$ to $(-2, -1)$. The following examples may be conducive to understanding.

Example. The parameter pair $\rho_1 = 1.8$ and $\rho_2 = -0.95$ lies close to the unit-root line $\rho_1 + \rho_2 = 1$ but fulfills the stationarity conditions. Formula (5.10) suggest that $E(\hat{\rho}_{1_{ml}} + \hat{\rho}_{2_{ml}}) \approx 0.85 + 0.05/T$. The bias does not appear remarkable but the message is clear: the MLE for the sum of the autoregressive parameters is not necessarily biased towards stationarity with the AR(2) model.□

Example. The bias in the estimation of the sum $\rho_1 + \rho_2$ can be large with the AR(2)

[5]The fact that the autocorrelation of y_t at lag 1 is zero if $\rho_1 = 0$ may give intuition for the bias (this follows from Box and Jenkins (1976, p. 60)).

model. If we let $\rho_1 = 0$ then the bias is of the same form, to order $O(T^{-2})$, as with the AR(1) model with constant. A case in point is $\rho_1 = 0$ and $\rho_2 = 0.9$ establishing a stationary process and implying $E(\hat{\rho}_{1_{ml}} + \hat{\rho}_{2_{ml}}) \approx 0.9 - 3.7/T$ by formula (5.10). The bias is here towards stationarity.□

Relatedly, it can be inferred from Table 1 of Shaman and Stine (1988) that if $\rho_i = 0$ (with obvious notation), $i \geq 2$, but an AR(p), $p = 2, ..., 6$, model is estimated then

$$\begin{aligned}
E(\textstyle\sum_{i=1}^{2} \hat{\rho}_{i_{ml}}) &= \rho_1 - \frac{1+\rho_1}{T} + O(T^{-2}), \\
E(\textstyle\sum_{i=1}^{3} \hat{\rho}_{i_{ml}}) &= \rho_1 - \frac{1+\rho_1}{T} + O(T^{-2}), \\
E(\textstyle\sum_{i=1}^{4} \hat{\rho}_{i_{ml}}) &= \rho_1 - \frac{2}{T} + O(T^{-2}), \qquad (5.12) \\
E(\textstyle\sum_{i=1}^{5} \hat{\rho}_{i_{ml}}) &= \rho_1 - \frac{2}{T} + O(T^{-2}), \qquad \text{and} \\
E(\textstyle\sum_{i=1}^{6} \hat{\rho}_{i_{ml}}) &= \rho_1 - \frac{3-\rho_1}{T} + O(T^{-2})
\end{aligned}$$

(IID$(0, \sigma^2)$ innovations are assumed). The number of terms in the sums indicates the order of the estimated model. Estimation of a sole redundant parameter appears to add little to the bias in the AR(2) model if ρ_1 is close to unity (as can be seen by comparing the expression for $E(\sum_{i=1}^{2} \hat{\rho}_{i_{ml}})$ with formula (5.3)). The extra bias tends to grow in absolute value with the number of redundant autoregressive terms unless ρ_1 is close to unity leading to approximate constancy of the above formulae. In this case, even five excess autoregressive parameters do not import as much bias as a single constant. For ρ_1 smaller than 0.5 the extra autoregressive parameters induce a more severe bias culminating for ρ_1 close to minus one. (Cf. the expression for $E(\sum_{i=1}^{6} \hat{\rho}_{i_{ml}})$ with formula (5.5)). The distortion is then of the same magnitude as with the AR(1) model with constant when the autoregressive coefficient is close to one. Curiously, the bias remains negative in the above formulae regardless of the value of ρ_1 (in the range $(-1, 1)$) and the number of redundant autoregressive parameters so the bias tends to tune the MLE towards -1 or nonstationarity when ρ_1 is negative. The tendency gains strength as the number of estimated excessive parameters increases.

The formulae of this section hold for values well within the stationarity regions but the approximations are likely to run into trouble if the process is close to a nonstationary boundary.[6] The related corrections may nevertheless be able to lessen the bias considerably, though not satisfactorily, even in such a case (Orcutt and Winokur (1969) provide an example). There is ample small-sample simulation evidence that distributions of the MLEs should convert smoothly when the parameters are tuned from the stationary region to the nonstationary region. The above approximations should hence throw some light on the behaviour of the biases with unit-root models in finite samples, too. The example below on the bias with the unit-root AR(1) model illustrates the case. Of course, conclusions derived this way are tentative, and it would be unwise to take them literally.

Starting values do not play a role in the above asymptotic expressions as those effects can be relegated to the $O(T^{-2})$ terms. They can influence the exact small-sample biases substantially, though, as exemplified by the aforementioned numerical calculations of Evans and Savin (1981), the related calculations of Abadir and Hadri (1996), and the approximations of Perron (1991) for the model (5.2) or the simple AR(1). The impact of the starting value tends to be more prone the closer the process is to the nonstationary boundary but Abadir and Hadri (*op. cit.*) yet show that the bias does not increase monotonically in ρ and, moreover, that the bias can *increase* with T when ρ and y_0 are fairly large.[7] White (1961) had proved for the same model that the bias of the MLE vanishes, for a given T, if the starting value is large enough. Perron reasoned from the approximation he derived that, furthermore, the MLE converges to the true value as y_0/σ tends to infinity. (Evans and Savin had also noted that the distribution of the MLE grows denser when y_0/σ increases.) We prove in Appendix A5 consistency, and hence fading of the bias, for large y_0 for the simple AR(1) model and also for the AR(1) model

[6]Tuan (1992) has found an exact formula for the bias of the MLE when the model is (5.2) with $y_0 = 0$. He recognised that formula (5.3) gives slight overestimates for the bias. Orcutt and Winokur (1969) provide simulation evidence confirming adequate performance of $\hat{\rho}_b$.

[7]Relatedly, Schmidt and Phillips (1992) employ Monte Carlo experiments and *response-surface methodology* (cf. Hendry (1984) or Davidson and MacKinnon (1993)) to find a functional form for the power of some unit-root tests as a function of the starting value.

with constant. To our knowledge, the result has not been proved for the latter model previously.

5.4 Biases in the Estimation of Nonstationary Autoregressive Models

Analytical results on the biases of the MLEs of the autoregressive parameters are scarce when the process is nonstationary. Evans and Savin (1981) found exact numerical values for the bias in the simple unit-root AR(1) model, but they had to employ complicated numerical procedures. Recently, Abadir (1993, 1995) uncovered a simple yet accurate approximation for the bias $b_T \equiv E[(\hat{\rho}_{ml} - 1)]$ when the model is (5.2) with $\rho = 1$, $y_0 = 0$ and the innovations follow NID$(0, \sigma^2)$. The formula is

$$b_T \approx (b^*_\infty / T) \exp(-2.6138/T) \tag{5.13}$$

where $b^*_\infty = -1.78143$ or the limiting value of the standardised bias $E[T(\hat{\rho}_{ml} - 1)]$ (inferred from Table 3 of Evans and Savin (1981)). The formula gives an accurate representation of the bias up to five decimal places.

Example. If $T = 100$ and $\psi = 1$ in the simple AR(1) model (5.2) then the bias $E(\hat{\rho}_{ml} - 1) \approx -0.017$ by formula (5.13). The (unreliable) prediction from formula (5.3) is -0.02 which coincides with the previous figure to two decimal places.□

The above referred MSE minimizing multiplying correction (5.1) leads to the bias

$$b^m_T = \frac{-v_T}{v_T + (1 + b_T)^2} \tag{5.14}$$

where the bias of the MLE $b_T = E(\hat{\rho}_{ml} - 1)$, and the corresponding variance $v_T = E(\hat{\rho}_{ml} - 1)^2$ depend, of course, on the sample size. If the model is the simple AR(1) then b_T and v_T, needed for evaluation of b^m_T, can be found from Table 3 of Evans and Savin (1981) or for any sample size by the approximations given by Abadir (1995).

Abadir's calculations deal explicitly with the AR(1) model, but the expression for b_T^m applies more generally. The moments have not been reported for more general nonstationary models so they have to be found by simulation for b_T^m to be assessed. We shall follow this route when comparing the bias of the Cox–Reid adjusted estimates in Chapter 6.

Next, let the model be (5.4) or the AR(1) with constant with $\rho = 1$ and let the innovations follow NID$(0, \sigma^2)$. Rothenberg (1995) found, by a simple innovation permutation argument, that the statistic

$$T\left[\frac{\sum_{t=1}^{T}(y_t - \bar{y})(y_{t-1} - \bar{y}_{-1}) + T\sigma^2/2}{\sum_{t=1}^{T}(y_{t-1} - \bar{y}_{-1})^2} - 1\right] \tag{5.15}$$

follows asymptotically a symmetric distribution with mean zero under $\rho = 1$. That is, an asymptotically unbiased estimate for ρ under $\rho = 1$ is achieved if T times one half of the variance of the innovations is added to the numerator of the MLE of ρ (see formula (6.6) which defines the MLE). Rothenberg argues that approximate unbiasedness should apply for small samples too because of the relatively rapid convergence of the standardised MLE to its asymptotic distribution. A consequence of the result is that the bias of the MLE, under $\rho = 1$, is approximately

$$E(\hat{\rho}_{ml} - 1) \approx -E\left[2T^{-2}\sigma^{-2}\sum_{t=1}^{T}(y_{t-1} - \bar{y}_{-1})^2\right]^{-1}.$$

More generally, the results of Rothenberg (*op. cit.*) hold also when ϵ_t is autocorrelated but Normal and stationary or *reversible* (cf. Lawrance (1991)).

Starting values can have a great impact on the bias with nonstationary models, too. The impact appears more sturdy than under stationarity when the model is (5.2) or the simple AR(1), see Evans and Savin (1981, p. 764)[8], Perron (*op. cit.*), and Phillips (1987). Abadir (1993) and Abadir and Hadri (1996) point out that the bias can increase

[8]They show also that the starting value affects the asymptotic distribution when $\psi > 1$ in process (5.2).

with the sample size with this model because of the starting value. On the other hand, the MLE and hence the bias are invariant with respect to the starting value under a unit root when the model is (5.4) or the AR(1) with a constant, see p. 103 or Appendix A5. Biases in unit-root models will be elucidated further, though not concentrated on, through simulation experiments in Chapters 6 and 7.

5.5 Side Effects of Bias Corrections

Tuning likelihood based statistics derives at least from Bartlett (1937, 1954) who considered the likelihood ratio. The Bartlett correction, as the technique is nowadays called, is often motivated as a bias correction for the mean in the first place (e.g. Stuart and Ord (1991, p. 873)) but actually it can be shown to amend the higher order cumulants, too (Lawley (1956), cf. Cribari-Neto and Cordeiro (1997) for a survey). The outcome is better agreement with the small-sample distribution and the asymptotic (χ^2) distribution of the statistic. Hence while a correction may explicitly not tackle the higher order cumulants, it may nevertheless have a desired impact on them as well.

Unfortunately, the overall effect of a bias correction is not always as rewarding. Orcutt and Winokur (1969) studied by simulation the small-sample distribution of the bias-corrected estimator $\hat{\rho}_b$ of the autoregressive parameter ρ of the AR(1) model with constant. They found that $\hat{\rho}_b$ is essentially unbiased unless ρ is very close to one (only nonnegative values of ρ were scrutinized). It featured a smaller MSE than the MLE for large values of ρ and *vice versa* for small values of ρ (the sample sizes studied were less than or equal to forty). However, it can be inferred from Table V in the *op. cit.* that the variance of the corrected estimator is larger for all values of ρ. The variance seems to be in general between one- or twofold bigger for most values of ρ when the sample size is forty, say.[9] In conclusion, the bias-corrected estimator achieves a smaller bias and often also a smaller MSE than the MLE but at the cost of an increase in the variance.

[9]The two simulation result sets of Orcutt and Winokur (*op. cit.*) appear ambiguous for $\psi = 0$ and $\psi = 1$. Copas (*op. cit.*) reports related results.

The multiplying correction (5.1) studied by Abadir (1995) induces an increase in the variance too, for typical sample sizes, as the multiplying constant tends to be larger than one. It is straightforward to infer from the results of Abadir (*op. cit.*) that the minimum achievable MSE (MSE_T^m) and the associated standard deviation (SD_T^m) are:

$$\begin{aligned} \mathrm{SD}_T^m &= a_T^m \sqrt{v_T} \equiv \frac{(1+b_T)}{v_T + (1+b_T)^2}\sqrt{v_T} \\ \mathrm{MSE}_T^m &= \frac{v_T}{v_T + (1+b_T)^2} = -b_T^m. \end{aligned} \tag{5.16}$$

Here a_T^m is the multiplier producing the minimum MSE. The last equality follows from formula (5.14). As remarked above, the formulae capturing the impact of the multiplication pertain in general; not only in the simple AR(1) case. The formulae can hence provide a point of comparison for the Cox–Reid adjusted estimates for more complex models as well. (Of course, the bias b_T and the variance v_T in the above equations refer then to the characteristics of the corresponding MLE.)

The *jackknifing* technique could be utilised in this setting as well (Quenouille (1949)). The simulations of Orcutt and Winokur revealed that the method can produce an essentially unbiased estimate, too, but the variance is consistently higher than with the above parametrically-corrected estimate $\hat{\rho}_b$ and hence cannot be recommended. Moreover, the technique runs into extra trouble if the model is nonstationary, see Abadir (1995) for a discussion.

Bias (or size correction) by means of Monte Carlo or response surface methodology might be employed. A problem is that estimation of such surfaces becomes increasingly difficult as the dimension of the parameter space increases. For example, Cheung and Lai (1995) were forced to a severely limited response surface analysis due to this obstacle.

MacKinnon and Smith (1996) argue that eliminating the bias altogether increases the variance of the estimate (though not necessarily the MSE) or even the MSE of it when the slope b'_T of the *bias function* $b_T = b(\psi, T) = E(\hat{\psi}_{ml}) - \psi$ (assumed approximately linear) is negative. A perhaps unexpected increase in MSE should occur if the following

inequality holds:

$$\frac{1}{(1+b'_T)^2}V_T \;>\; b_T^2 + V_T \;=\; MSE_T$$

where V_T and MSE_T are the variance and the MSE of the MLE, respectively, at sample size T, and the left–hand side of the inequality gives the (approximative) variance of the bias-corrected estimate. The bias correction increases the MSE, under a negative b'_T, when the bias is not large enough compared to the variance. The authors argue that reducing the bias only partially may be a better solution in terms of MSE.

The slope is negative for the AR(1) models we have looked at according to formulae (5.3) and (5.5) but they may not hold in the neighbourhood of ± 1 (cf. Figure 1 in *op. cit.*). Indeed, the bias correction of MacKinnon and Smith (*op. cit.*) yielded an increase in MSE in general except at around the extremes when the model is AR(1) with constant.[10] Formula (5.11) for the bias of the sum of the AR(2) coefficients under a unit root would suggest that the slope is positive and hence that deletion of bias could only have a desirable effect on both the variance and the MSE. This is a conjecture only because the approximation accuracy of the formula has not been grasped.

These considerations motivate the study of alternative ways such as Cox–Reid adjusting to correct for the bias. The effect of the Cox–Reid adjustments on the distribution, or its first cumulants, is not known in the present context but is scrutinized in Chapters 6 and 7.

[10] Associatively, the multiplying correction (5.1) of Abadir (1995) essentially expunges the bias while it simultaneously increases the variance yet reduces the MSE when the model is unit-root AR(1).

Chapter 6

Application of Adjusted Profile Likelihood to Some AR Models

6.1 Introductory Remarks

The Cox–Reid theory is applied to autoregressive processes of orders 1 and 2 in this chapter. First the case of a simple AR(1) model without constant is briefly analysed. The next, and longer, section is devoted to the AR(1) model with constant. Then follows a section devoted to an analysis of the previous model when a unit root exists simultaneously with a non-zero constant. A non-zero constant or drift reshapes the asymptotic distributions radically which is the rationale for a separate exposition. The analysis of the AR(1) dynamics is closed with an analysis of a model due to Bhargava (1986) who introduced an alternative way of including a constant in the model. The dynamically richer AR(2) model without a constant is the final scrutinised model.

Two families of statistics, or standardised autoregressive coefficients and Wald test statistics, are explored. The analysis is in the spirit of Dickey and Fuller (1979) but the statistics are constructed from the adjusted profile likelihood in the present study. The original Dickey–Fuller statistics provide points of comparison. As pointed out at the end of Chapter 4 the LM and perhaps also the LR principles may be less fit for construction of tests than the Wald principle when the adjusted profile likelihood is employed.

In general, the asymptotic distributions under both nonstationarity and stationarity are analysed. However, the information measures related to the AR(2) model have been derived under the assumption of a unit root. This facilitates the calculations which have already become pretty complicated. The emphasis on testing in the unit-root literature

motivates to an extent this simplification. The stationary AR(2) model provides an interesting point of comparison but it is commented on solely from the point of view of asymptotics.

Only an outline of the sometimes extensive derivations is presented. The ML and the OLS estimates of ψ are equal for the models of Sections 6.2, 6.3, 6.4, and 6.6 but the notation $\hat{\psi}_{ml}$ is used to emphasize the fact that the adjustments spring from likelihood considerations.

We make the assumption of the normality of the innovations when deriving the Cox–Reid adjusted estimate. The asymptotic unit-root distributions which we shall obtain apply also under IIDness of the innovations or even under certain conditions which allow for some heterogeneity in the innovation process (but not autocorrelatedness, cf. Phillips and Perron (1988)). IIDness of the innovations with some standard technical conditions is also enough to guarantee convergence to the stated asymptotic distributions when the process is stationary.[1]

The focus is on asymptotic distributions; Chapter 7 is devoted to small-sample characteristics.

6.2 Adjusted Profile Likelihood for an AR(1) Model

The simple AR(1) or Markov model

$$y_t = \psi y_{t-1} + \epsilon_t, \tag{6.1}$$

[1]The calculations related to the examples and all of the graphs in the thesis were done with the statistical software SURVO (Mustonen (1992)). The STATL command for analysis of very large data sets and to be included in future versions of SURVO was additionally provided for our use by Professor Mustonen. SURVO was also used to check numerically derivations and properties of the obtained quantities. The Mathematica (Wolfram (1991)) programme was used for analytical checks and finding the maxima of the likelihood equations of the example in Section 6.6. All of the simulations were implemented with programmes written by the author using the GAUSS language (1992). The experiments referred to in this chapter are based on 100 000 replications unless pointed out otherwise. The scales of the graphs do not always cover the whole range of values of the variable in question. Appendix A6 supplies other details of the simulations.

where $\psi \in (-1,1]$, $\epsilon_t \sim$NID$(0,\sigma^2)$, $\sigma^2 > 0$, $t = 1,\ldots,T$, and $y_0 = 0$ is scrutinized in this section. The autoregressive parameter ψ is the parameter of interest while the variance parameter σ^2 is the nuisance parameter.

As already remarked, the model has a long history in economics. Indeed, if economic theory can predict time series properties of a variable of interest then the outcome is typically this model with $\psi = 1$. The variable in question is usually of financial or expectational character. (Cf. Sections 2.2 and 3.1 and the references there in).

The log-likelihood generated by the model is

$$l(\psi,\sigma^2;\mathbf{y}) \;=\; -\tfrac{T}{2}\log\sigma^2 - \tfrac{1}{2\sigma^2}\textstyle\sum_{t=1}^{T}(y_t - \psi y_{t-1})^2.$$

The profile log-likelihood is

$$l_p(\psi) \;\propto\; -\tfrac{T}{2}\log\hat{\sigma}^2_{\psi}$$

where

$$\hat{\sigma}^2_{\psi} \;=\; T^{-1}\textstyle\sum_{t=1}^{T}(y_t - \psi y_{t-1})^2.$$

The expected and observed information measures of relevance are

$$i_{\psi\sigma^2} \;=\; E\left[\sigma^{-4}\textstyle\sum_{t=1}^{T}(y_t - \psi y_{t-1})y_{t-1}\right] \;=\; 0$$

and

$$j_{\sigma^2\sigma^2} \;=\; -\frac{T}{2\sigma^4} + \frac{1}{\sigma^6}\textstyle\sum_{t=1}^{T}(y_t - \psi y_{t-1})^2,$$

respectively. Hence the parameters are orthogonal which inflicts $\hat{c}= 0$ in equation (4.12), and the adjusted profile likelihood (4.9) is applicable. We remark that

$$\begin{aligned} -\tfrac{1}{2}\log \mid j_{\sigma^2\sigma^2}(\psi,\hat{\sigma}^2_{\psi}) \mid \;&=\; -\tfrac{1}{2}\log\left|-\tfrac{T}{2}\,\hat{\sigma}^{-4}_{\psi} + \hat{\sigma}^{-6}_{\psi}\textstyle\sum_{t=1}^{T}(y_t - \psi y_{t-1})^2\right| \\ &=\; \log\hat{\sigma}^2_{\psi} - \tfrac{1}{2}\log(\tfrac{T}{2}). \end{aligned}$$

It follows that the adjusted profile log-likelihood is

$$l_{ad}(\psi) \;\propto\; -\tfrac{T}{2}\log\hat{\sigma}^2_\psi + \log\hat{\sigma}_\psi \;=\; -\frac{(T-2)}{2}\log\hat{\sigma}^2_\psi\,. \tag{6.2}$$

Obviously, maximization of this log-likelihood will produce the same estimate as maximization of the ordinary profile log-likelihood or

$$\hat{\psi}_{ad} \;=\; \hat{\psi}_{ml} \;=\; \frac{\sum_{t=1}^{T} y_t y_{t-1}}{\sum_{t=1}^{T} y_{t-1}^2}. \tag{6.3}$$

Intuitive explanation for the equivalence of the estimates is that σ^2 is a scale parameter: it is not really a nuisance from the point of view of estimation. This is so regardless of the value of ψ which would be one in the case of a unit root. Contrastingly, an adjustment arose in the example of p. 42 where the variance was regarded as the parameter of interest. In that case the autoregressive coefficient is a nuisance even from the point of view of estimation. As pointed out in Section 5.3, the bias is not too severe for this model in general which might account for the result, too.

It is interesting that the innovation permutation argument of Rothenberg (1995), referred to in Section 5.4, likewise fails to modify the estimate under this model (personal communication). A bias correction (for median-unbiasedness) for the AR(1) model did not seem feasible in Andrews (1993) either.

The familiar asymptotic distribution of $T^{1/2}(\hat{\psi}_{ml} - \psi)$ is $N(0, 1-\psi^2)$. The asymptotic unit-root distribution of $T(\hat{\psi}_{ml} - 1)$ was introduced already in formula (3.11). The histogram of Figure 6.1, calculated from simulated data, illustrates the latter distribution. The Monte Carlo median (Med), skewness (Skew), mean, standard deviation (SD), and MSE are documented in the figure as well.[2] The distribution is skewed to the left and

[2]The Monte-Carlo mean may be compared with the previously reported exact value -1.7814 (p. 61). The other exact moments may be checked from the numerical calculations of Evans and Savin (1981, p. 768).

Nabeya and Tanaka (1990a) have calculated accurate asymptotic fractiles and plotted the corresponding density function for this model and for the other variants with a constant or a constant and a time trend. Literally, their model differs from the present slightly, neither do they analyse explicitly the MLE but the resulting asymptotic distributions are the same as those in which we are interested.

has most of its probability mass on negative values (or values of $\hat{\psi}_{ml}$ less than 1).

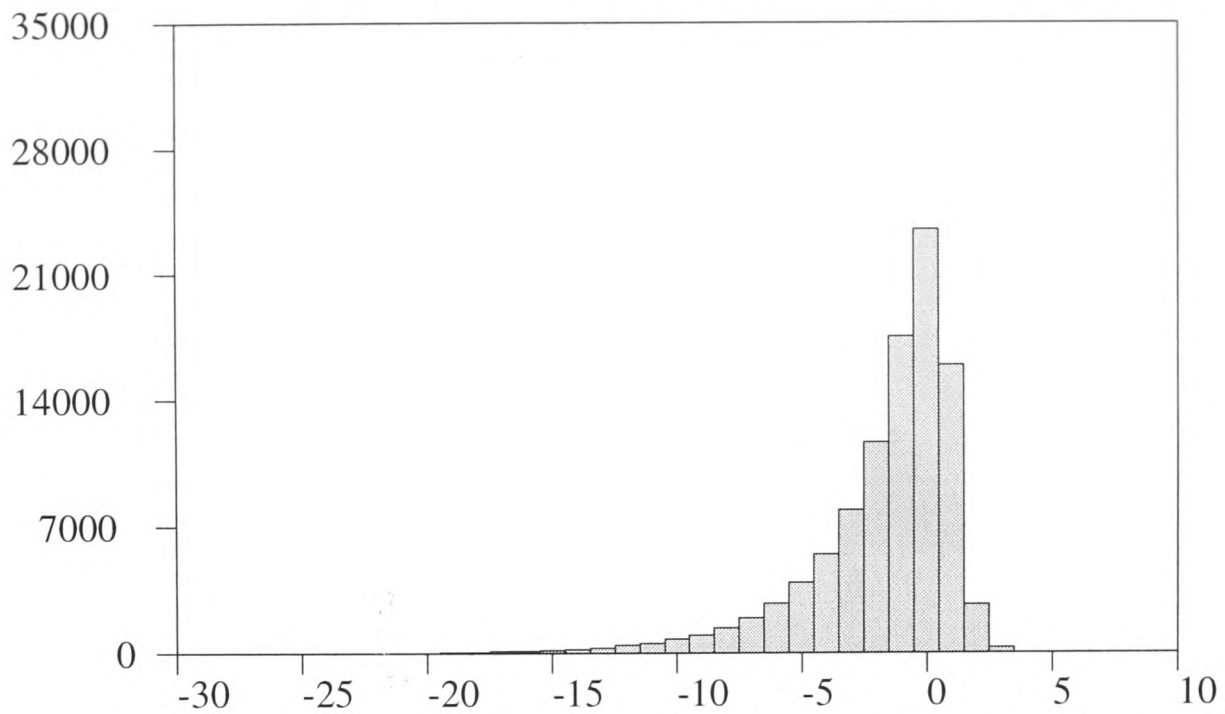


Figure 6.1 Empirical asymptotic distribution of $T(\hat{\psi}_{ml} - 1)$ for $\psi = 1$ (AR(1)).

The t-statistic for the null hypothesis of $\psi = 1$ is

$$\frac{(\hat{\psi}_{ml} - 1)}{s/\sqrt{\sum_{t=1}^{T} y_{t-1}^2}}$$

where s is the standard degrees-of-freedom corrected estimate of σ. This is the statistic which Dickey (1976) employed for testing for unit roots, and the asymptotic unit-root distribution of it was recorded in formula (3.12). The same asymptotic distribution is generated by the (square root of the) Wald statistic

$$\sqrt{W} = \frac{(\hat{\psi}_{ml} - 1)}{\hat{\sigma}_{ml}/\sqrt{\sum_{t=1}^{T} y_{t-1}^2}}$$

which complies with definition (4.4). The distribution is sketched in Figure 6.2. It can be seen that the distribution possesses a negative mean and, perhaps a bit surprisingly, that it is somewhat skewed to the right which contrasts with the asymptotic distribution of the MLE. The asymptotic distribution would be Standard Normal under stationarity.

It was pointed out in Section 4.1 that the statistic

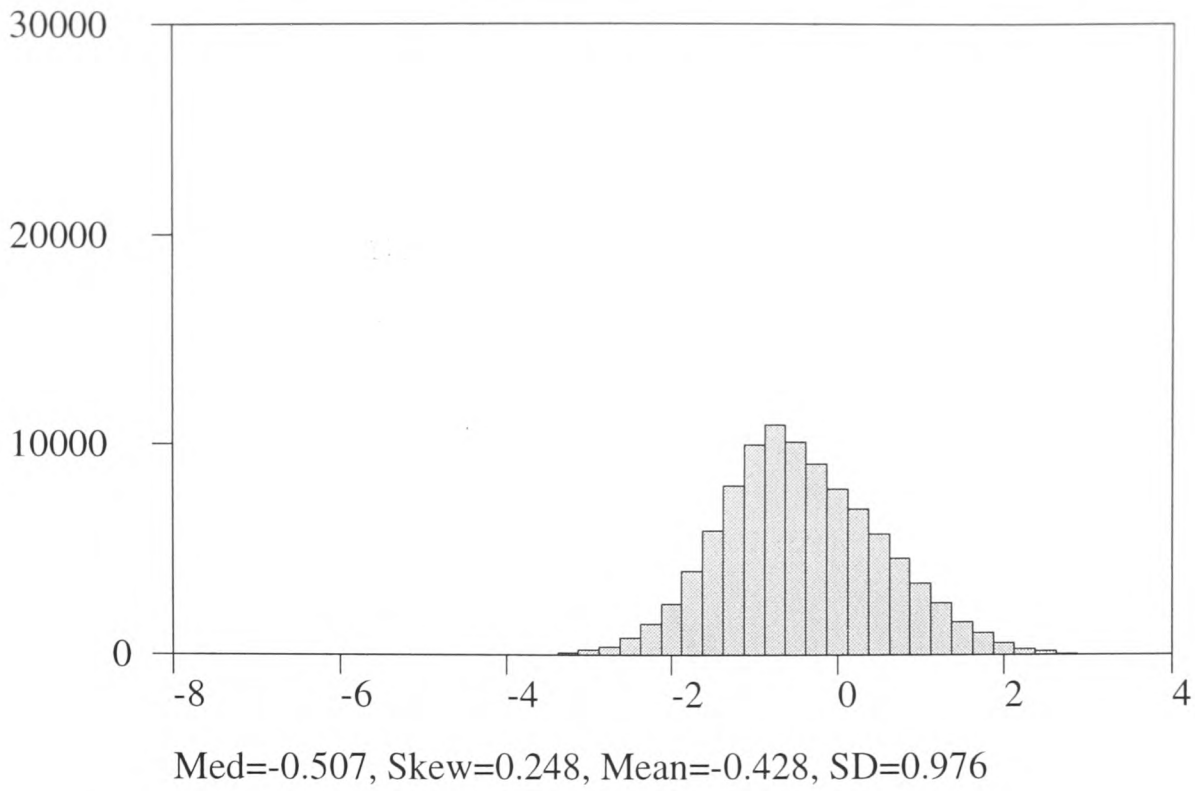


Figure 6.2 Empirical asymptotic distribution of the t-value for $\psi = 1$ (AR(1)) .

$$\sqrt{W_{ad}} \equiv \sqrt{(\hat{\psi}_{ad} - 1)^2 \left[-\frac{\partial^2 l_{ad}(\psi)}{\partial^2 \psi} \Big|_{\psi=\hat{\psi}_{ad}} \right]}$$

is based on adjusted profile likelihood but is analogous to the square root of the ordinary Wald statistic. With the above results in hand, it is a simple task to show that

$$\sqrt{W_{ad}} = \frac{(\hat{\psi}_{ml} - 1)}{\hat{\sigma}_{ml} / \sqrt{\sum_{t=1}^{T} y_{t-1}^2}} \frac{T-2}{T}$$

for the maintained model. The adjusted Wald statistic differs from the ordinary one only by a multiplying constant (tending to unity) so the adjustment is not helpful when a unit root is tested for. When the null value of ψ (ψ_0) lies in the region $(-1, 1)$ (and '1' above is replaced by ψ_0) then a better concordance with the Normal approximation to the small-sample distribution is possible, but we shall not dwell on this side-track.

Example. 100 observations were generated from the random-walk process $y_t = \psi y_{t-1} + \epsilon_t$ where $\psi = 1$, $\epsilon_t \sim$NID$(0, 1)$ and $y_0 = 0$. A graph of the time series is plotted in Figure 6.3. If only the 25 first observations are analyzed then $\hat{\psi}_{ml}$=$\hat{\psi}_{ad}$= 0.959. The estimates equal 0.994 when all the observations are considered. These estimates will serve as a benchmark against which estimates to be introduced in the following sections can be

compared.□

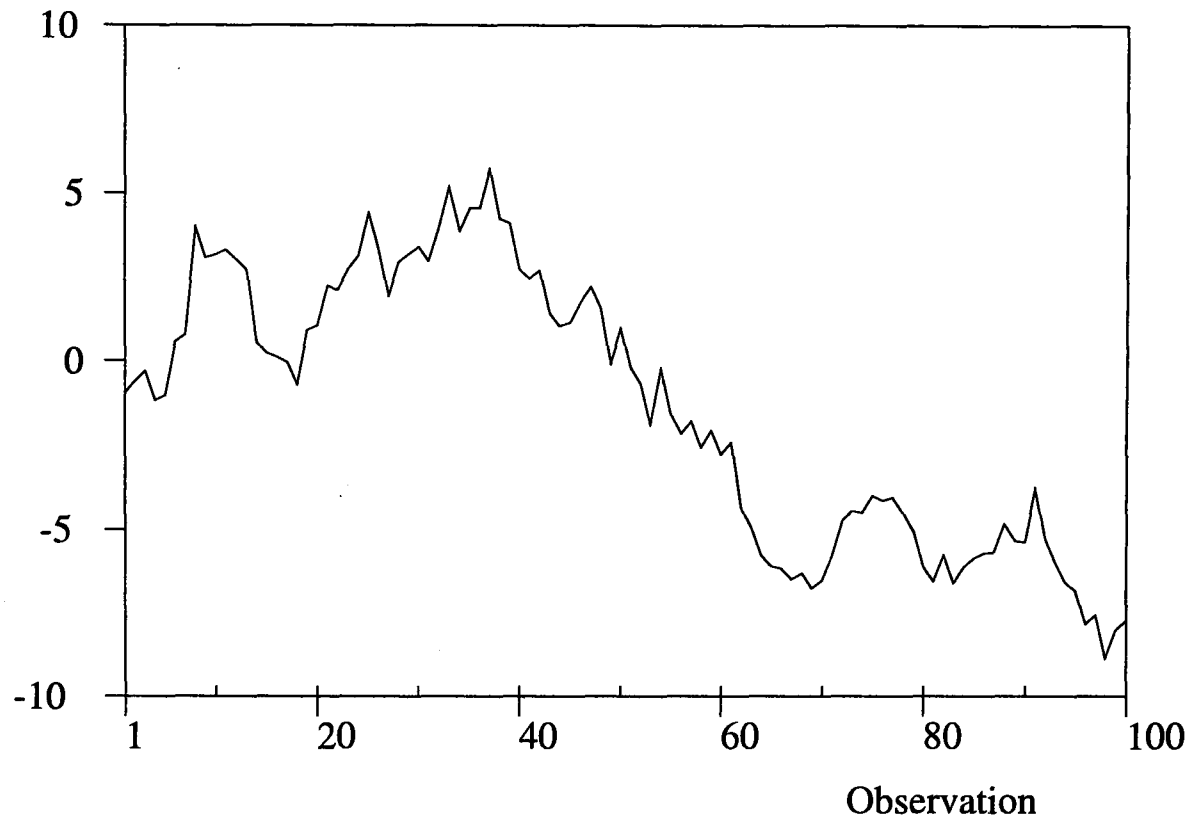


Figure 6.3 A realisation from the random-walk process.

6.3 Adjusted Profile Likelihood for an AR(1) Model with Constant

A typical way of introducing a constant to the AR(1) model is

$$y_t = \alpha + \psi y_{t-1} + \epsilon_t \tag{6.4}$$

where $\psi \in (-1, 1]$, $\epsilon_t \sim$NID$(0, \sigma^2)$, $\sigma^2 > 0$, $t = 1, \ldots, T$, and $y_0 = 0$. If $\psi = 1$ then it is assumed that $\alpha = 0$. The autoregressive parameter ψ is the parameter of interest while the constant α and the variance parameter σ^2 are the nuisance parameters. In tune with our previous notation denote the nuisance parameters by $\phi' = [\alpha\ \sigma^2]$ and the whole parameter vector by $\omega' = [\psi\ \phi']$. The model will sometimes be referred to as AR$_\mu$(1). It is one of the most important workhorses of applied econometrics. As already pointed out, inclusion of the constant substantially magnifies the bias and spread in the estimation of

ψ, for most values of ψ.

The assumption $\alpha = 0$ under $\psi = 1$ will not be used in the derivation of the information measures or the adjusted profile log-likelihood in any way. (Indeed, α will appear in many formulae below even when $\psi = 1$.) The assumption is needed only when the asymptotic (and small-sample) distributions are derived (and simulated). The case $\alpha \neq 0$ and $\psi = 1$ is examined in the next section.[3]

The log-likelihood is

$$l(\psi, \alpha, \sigma^2; \mathbf{y}) = -\tfrac{T}{2} \log \sigma^2 - \tfrac{1}{2\sigma^2} \textstyle\sum_{t=1}^{T} (y_t - \alpha - \psi y_{t-1})^2$$

and the profile log-likelihood is

$$l_p(\psi) = -\tfrac{T}{2} \log \hat{\sigma}^2_{\psi} \,. \tag{6.5}$$

Here

$$\begin{aligned} \hat{\sigma}^2_{\psi} &= T^{-1} \textstyle\sum_{t=1}^{T} (y_t - \hat{\alpha}_{\psi} - \psi y_{t-1})^2 \\ &= T^{-1} \left[\textstyle\sum_{t=1}^{T} (y_t - \bar{y})^2 - 2\psi \sum_{t=1}^{T} (y_t - \bar{y})(y_{t-1} - \bar{y}_{-1}) + \psi^2 \sum_{t=1}^{T} (y_{t-1} - \bar{y}_{-1})^2 \right] \end{aligned}$$

and

$$\begin{aligned} \hat{\alpha}_{\psi} &= T^{-1} \textstyle\sum_{t=1}^{T} y_t - \psi T^{-1} \sum_{t=1}^{T} y_{t-1} \\ &\equiv \bar{y} - \psi \, \bar{y}_{-1} \,. \end{aligned}$$

The maximum of $l_p(\psi)$ occurs at ψ equal to

$$\hat{\psi}_{ml} = \frac{\sum_{t=1}^{T} (y_t - \bar{y})(y_{t-1} - \bar{y}_{-1})}{\sum_{t=1}^{T} (y_{t-1} - \bar{y}_{-1})^2}. \tag{6.6}$$

[3]The time series would feature a drift α if $\alpha \neq 0$ and $\psi = 1$ (α/σ might be called a scale-free drift). A drift would alter the shape of the asymptotic distributions radically as noted before.

The standardised estimate, $T^{1/2}(\hat{\psi}_{ml} - \psi)$, shares the asymptotic distribution $N(0, 1-\psi^2)$ with the MLE of ψ of the previous section under stationarity. The well-known asymptotic unit-root distribution of $T(\hat{\psi}_{ml} - 1)$ is

$$T(\hat{\psi}_{ml} - 1) \;\Rightarrow\; \frac{\frac{1}{2}\left\{[W(1)]^2 - 1\right\} - W(1)\int_0^1 W(r)dr}{\int_0^1 [W(r)]^2 dr - \left[\int_0^1 W(r)dr\right]^2} \tag{6.7}$$

or more compactly

$$T(\hat{\psi}_{ml} - 1) \;\Rightarrow\; \frac{\int_0^1 W_*(r)dW(r)}{\int_0^1 [W_*(r)]^2 dr}. \tag{6.8}$$

Here $W_*(r) = W(r) - \int_0^1 W(r)dr$ (so called demeaned Brownian motion). (Hamilton (1994) provides a detailed derivation of formula (6.7). Phillips and Perron (1988) established the latter form). The distribution will be referred to as DF_μ after Dickey and Fuller (1979). The histogram of Figure 6.4 portrays the distribution. The scale of the graph is kept the same as in Figure 6.1 to elucidate the increase in bias and spread of the distribution compared to the MLE of the previous section. Estimation of the constant increases the bias relatively even more than the SD. The consequent inflation of MSE can be read from the attached statistics. Perhaps the sole positive finding is that the MLE with constant features less skewness than the MLE without a constant.[4]

The t-value, or $\hat{\tau}_\mu$ in the notation of Section 3.2, for the null hypothesis $\psi = 1$ and the asymptotic distribution of it are:

$$\hat{\tau}_\mu = \left.\frac{\hat{\psi}_{ml} - \psi}{s/\sqrt{\sum_{t=1}^{T}(y_{t-1} - \bar{y}_{-1})^2}}\right|_{\psi=1} \;\Rightarrow\; \frac{\int_0^1 W_*(r)dW(r)}{\left\{\int_0^1 [W_*(r)]^2 dr\right\}^{1/2}} \tag{6.9}$$

$$\text{or to} \quad DF_\mu \left\{\int_0^1 [W_*(r)]^2 dr\right\}^{1/2}.$$

[4]It is not clear if a symmetric distribution is desired when estimating an autoregressive parameter near (but not larger) than unity. For one thing, estimates larger than unity may lead to very inaccurate forecasts.

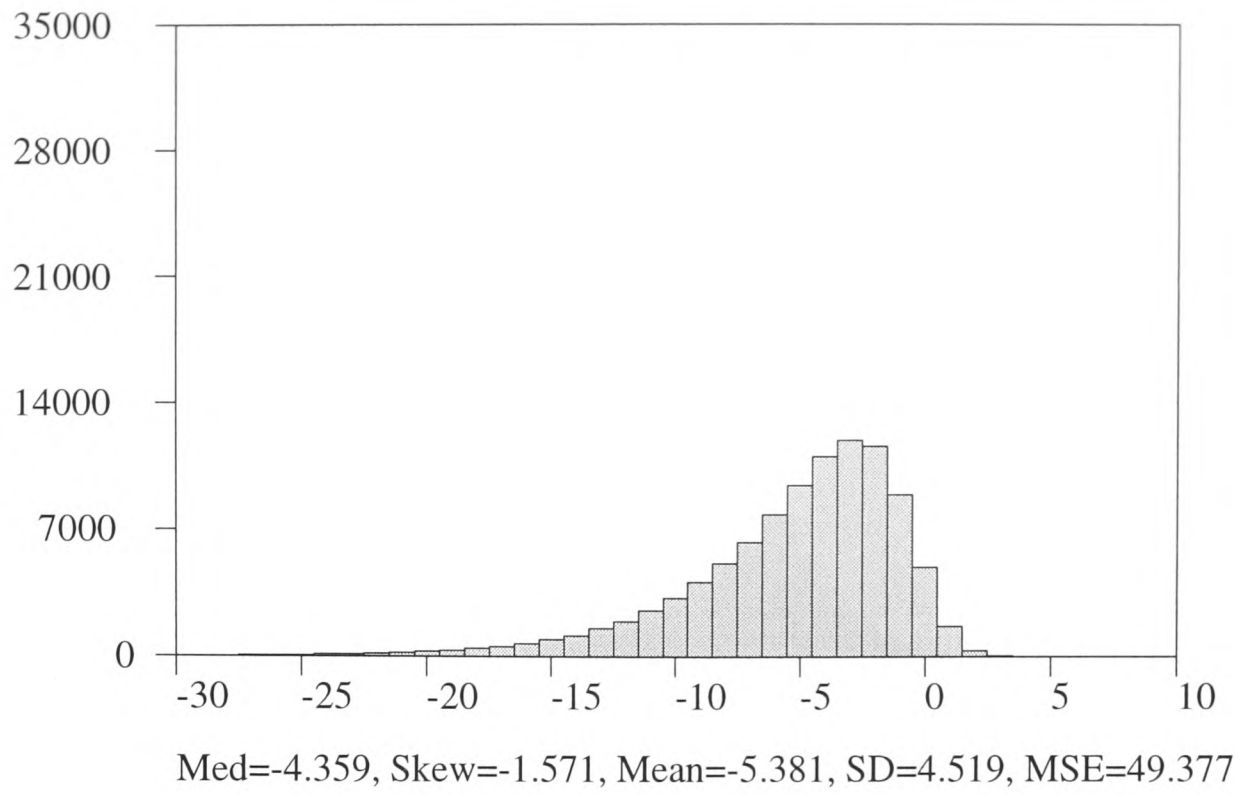


Figure 6.4 Empirical asymptotic distribution of $T(\hat{\psi}_{ml} - 1)$ for $\psi = 1$ ($\mathrm{AR}_\mu(1)$).

Lower case s^2 is the degrees-of-freedom corrected estimate of variance or:

$$s^2 = (T-2)^{-1} \sum_{t=1}^{T} (y_t - \hat{\alpha}_{ml} - \hat{\psi}_{ml}\, y_{t-1})^2.$$

The same asymptotic distribution is generated by (the square root of) the Wald statistic ($\sqrt{W}$) which is calculated otherwise similarly but the MLE of σ is substituted in place of s. The distribution will carry the tag $DF{-}\tau_\mu$ in what follows. It is drafted in Figure 6.5. The distribution features a negative mean but the observations are scattered fairly symmetrically as can be read from the small difference between the mean and the median. Somewhat surprisingly, the standard deviation and skewness coefficient are smaller than when a constant was not estimated. Of course, a corresponding t-value for a null within the stationary region would follow the Standard Normal asymptotically under stationarity.

We begin the derivation of adjusted profile log-likelihood from an easy starting point:

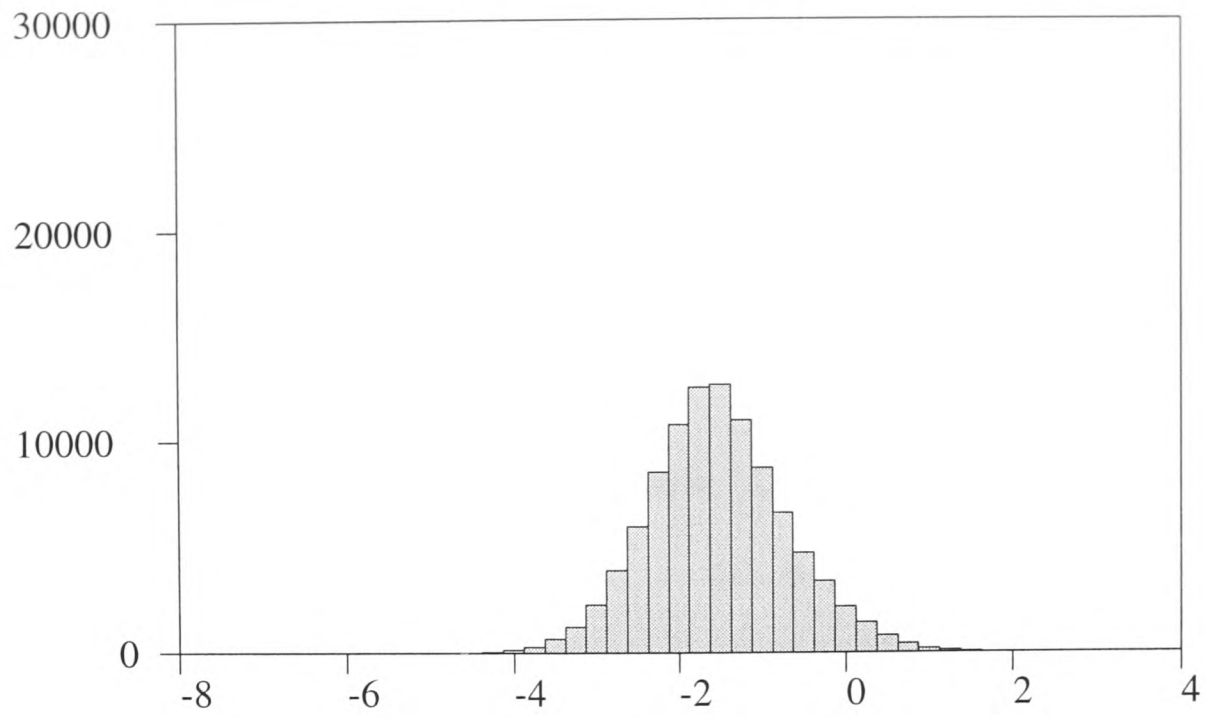


Figure 6.5 Empirical asymptotic distribution of the t-value for $\psi = 1$ ($\mathrm{AR}_\mu(1)$).

$$i_{\psi\sigma^2} = E(j_{\psi\sigma^2}) = E\left[\sigma^{-4}\sum_{t=1}^{T}(y_t - \alpha - \psi y_{t-1})y_{t-1}\right] = 0$$

$$i_{\alpha\sigma^2} = E(j_{\alpha\sigma^2}) = E\left[\sigma^{-4}\sum_{t=1}^{T}(y_t - \alpha - \psi y_{t-1})\right] = 0.$$

Thus σ^2 is orthogonal with respect to ψ and α. Instead, ψ and α are in general nonorthogonal:

$$\begin{aligned} i_{\psi\alpha} &= E(\sigma^{-2}\textstyle\sum_{t=1}^{T} y_{t-1}) \\ &= \begin{cases} \frac{\alpha}{\sigma^2(1-\psi)}\sum_{t=1}^{T-1}(1-\psi^t) & \text{if } \mid\psi\mid < 1 \\ \frac{\alpha}{\sigma^2}\sum_{t=1}^{T-1} t & \text{if } \psi = 1 \end{cases} \qquad (6.10) \\ &= \begin{cases} \frac{\alpha}{\sigma^2(1-\psi)}(T - \frac{1-\psi^T}{1-\psi}) & \text{if } \mid\psi\mid < 1 \\ \frac{\alpha}{2\sigma^2}T(T-1) & \text{if } \psi = 1. \end{cases} \end{aligned}$$

The second equality follows from the formula

$$y_t = \begin{cases} \psi^t y_0 + \alpha\dfrac{1-\psi^t}{1-\psi} + \sum_{i=1}^{t}\psi^{t-i}\epsilon_i & \text{if } \mid\psi\mid < 1 \\ y_0 + \alpha t + \sum_{i=1}^{t}\epsilon_i & \text{if } \psi = 1 \end{cases} \qquad (6.11)$$

and the assumption $y_0 = 0$.[5]

The parameters are orthogonal only if $\alpha = 0$.[6] This kind of local orthogonality is not useful from our perspective as pointed out in Section 4.2. The formula for the unit-root case arises, say, as a limit of the formula for the stationary case by applying sequentially l'Hospital's rule (e.g. Chiang (1984, p. 429)).

We shall apply the Cox–Reid (1993) formula (4.12) which does not require orthogonality. Orthogonality of σ^2 with ψ and α is still beneficial because it simplifies the formula of c in that equation:

$$\begin{aligned} c &= tr\left\{ \frac{\partial}{\partial(\alpha,\sigma^2)} \left(\begin{bmatrix} i_{\alpha\alpha} & i_{\alpha\sigma^2} \\ i_{\sigma^2\alpha} & i_{\sigma^2\sigma^2} \end{bmatrix}^{-1} \begin{bmatrix} i_{\psi\alpha} \\ i_{\psi\sigma^2} \end{bmatrix} \right)' \right\} \\ &= tr\left\{ \frac{\partial}{\partial(\alpha,\sigma^2)} \left(\begin{bmatrix} i_{\alpha\alpha} & 0 \\ 0 & i_{\sigma^2\sigma^2} \end{bmatrix}^{-1} \begin{bmatrix} i_{\psi\alpha} \\ 0 \end{bmatrix} \right)' \right\} \\ &= \frac{\partial}{\partial\alpha}(i_{\alpha\alpha}^{-1} i_{\psi\alpha}). \end{aligned} \tag{6.12}$$

A single additional information measure is needed to construct c but we also note the observed measure for σ^2 for use below:

$$\begin{aligned} i_{\alpha\alpha} &= E(j_{\alpha\alpha}) = E(\sigma^{-2}T) = \sigma^{-2}T \\ j_{\sigma^2\sigma^2} &= \frac{-T}{2\sigma^4} + \sigma^{-6}\sum_{t=1}^{T}(y_t - \alpha - \psi y_{t-1})^2. \end{aligned} \tag{6.13}$$

Coefficient c is here:

[5] The expression $\sum_{i=1}^{t} \psi^{t-i}\epsilon_i$ should be understood to stand for $\sum_{i=1}^{t-1} \psi^{t-i}\epsilon_i + \epsilon_t$ when $\psi = 0$.

[6] Cruddas *et al.* (1989) deviced the orthogonal parameterisation $\boldsymbol{\omega}^{*\prime} = [\psi \;\; \alpha \;\; \log\{\sigma(1-\psi^2)^{-1/2}\}]$ for the fully stationary AR(1) model with constant where the first observation already follows the stationary distribution. Such a parameterisation is not functional for our purposes as we want to allow for the circumstance $\psi = 1$.

$$
\begin{aligned}
c &= \begin{cases} \frac{\partial}{\partial\alpha}\left\{\alpha\left[\frac{1}{1-\psi} - \frac{1-\psi^T}{T(1-\psi)^2}\right]\right\} & \text{if } \mid\psi\mid < 1 \\ \frac{\partial}{\partial\alpha}\left(\alpha\frac{T-1}{2}\right) & \text{if } \psi = 1 \end{cases} \\
&= \begin{cases} \dfrac{1}{1-\psi} - \dfrac{1-\psi^T}{T(1-\psi)^2} & \text{if } \mid\psi\mid < 1 \\ \dfrac{T-1}{2} & \text{if } \psi = 1. \end{cases}
\end{aligned} \tag{6.14}
$$

Again, the formula for the nonstationary model arises as a limit from the first formula. In what follows it is assumed that c or the estimated c, $\hat{c} = c\mid_{\psi=\hat{\psi}_{ml}}$, is calculated according to the first formula whether $\mid\psi\mid < 1$ or $\psi = 1$ unless it is pointed out otherwise.

Of course, c is always positive according to the formula for $\psi = 1$ but it is also easy to see that $c > 0$ for all $T \geq 2$ if $\mid\psi\mid < 1$. Actually, c can then be negative only if ψ were less than, say, -1 — indeed an unlikely circumstance with economic time series. The estimated c converges in probability to $(1-\psi)^{-1} > 0$ under $\mid\psi\mid < 1$. If $\psi = 1$ then the term $\left(1-\psi^T\right)/T(1-\psi)^2$ does not fade as T increases and $\hat{c}$ is $O_p(T)$. Numerical analysis reveals that c increases monotonically with ψ for $\psi \in (-1,1)$ and $T \geq 3$ (as can easily be seen from the asymptotic expression for c but the property appears to hold as well for T finite).[7] The adjustment to the profile likelihood thus tends to magnify with ψ. Though theoretically equal at $\psi = 1$, the formulae are likely to differ substantially numerically if the first is evaluated at $\hat{\omega}_{ml}$ because of the downward bias of $\hat{\psi}_{ml}$ and random variation. The coefficient increases rapidly if it is evaluated at a ψ larger than one.

We note that c does not depend on the magnitude of α, the reason for the existence of c in the first place. This harmonizes with formula (5.5) according to which the bias does not depend on the magnitude of the constant if the model is stationary. It is somewhat unanticipated that c is positive for $\psi \in (-1,-1/3]$ about which more is said below. The fact that α does not appear in the formula for c fits our assumption of $\alpha = 0$ if $\psi = 1$

[7]The coefficient c grows monotonically for ψ larger than one as well, because then $\partial c/\partial\psi = T(1-\psi)\left(1-\psi^{T-1}\right) - 2\left(1-\psi^T\right)$ is positive for all T. In general, the coefficient is not monotonic for values of ψ less than -1.

well. The intuition is less clear when $\alpha \neq 0$ and $\psi = 1$; the case is commented on in Section 6.4

The final element needed is

$$-\tfrac{1}{2}\log \mid \mathbf{J}_{\phi\phi}(\psi, \hat{\phi}_\psi) \mid \;=\; \tfrac{3}{2}\log \hat{\sigma}^2_\psi - \tfrac{1}{2}\log\left(\tfrac{T^2}{2}\right)$$

by the above results. We have now all the ingredients to evaluate the adjusted profile log-likelihood of Cox and Reid (1993):

$$\begin{aligned} l_{ad2}(\psi) &= -\tfrac{T}{2}\log \hat{\sigma}^2_\psi + \tfrac{3}{2}\log \hat{\sigma}^2_\psi - \tfrac{1}{2}\log\left(\tfrac{T^2}{2}\right) + \hat{c}\,(\psi - \hat{\psi}_{ml}) \\ &\propto -\frac{(T-3)}{2}\log \hat{\sigma}^2_\psi + \hat{c}\,(\psi - \hat{\psi}_{ml}). \end{aligned} \tag{6.15}$$

The adjustment has turned out to be a kind of degrees-of-freedom correction amplified by the $\hat{c}\,(\psi - \hat{\psi}_{ml})$ term. The impact of the former correction is not apparent yet but will be considered later (p. 84) after defining $\hat{\psi}_{ad2}$ explicitly. It is easy to see that the local maximum of $l_{ad2}(\psi)$ lies to the right of $\hat{\psi}_{ml}$, or $\hat{\psi}_{ad2} > \hat{\psi}_{ml}$, if $\hat{c} > 0$. It is highly likely that the condition is satisfied, as remarked above, and the modification serves to increase the AE over that of $\hat{\psi}_{ml}$. As remarked above, the magnitude of $\hat{c}$ depends greatly on the magnitude of $\hat{\psi}_{ml}$. The likelihood (6.15) makes it clear that the strength of the adjustment depends straightforwardly on the size of $\hat{c}$: the greater $\hat{\psi}_{ml}$, the greater $\hat{c}$, and the greater the difference between $\hat{\psi}_{ad2}$ and $\hat{\psi}_{ml}$. Relatedly, the adjustment appears more substantial in the nonstationary than in the stationary case: the adjustment term $(\psi - \hat{\psi}_{ml})$ is multiplied by $\hat{c}$ which is $O_p(T)$ or $O_p(1)$ if the process is nonstationary or stationary, respectively.[8] The disparity in the orders of magnitude of the adjustment to the MLE will be formalized later.

---

[8]The magnitude of the term $-[(T-3)/2]\log \hat{\sigma}^2_\psi$ can increase as well if a unit root is allowed for: $-[(T-3)/2]\log \hat{\sigma}^2_\psi$ is then $O_p(T)$ when evaluated at $\psi = 1$, as in the stationary case, but is of larger magnitude otherwise as $\hat{\sigma}^2_\psi$ is $O_p(T)$ when evaluated at $\psi \neq 1$ under $\psi = 1$. However, the magnitude does not increase by as much as the magnitude of the coefficient $\hat{c}$ because of the logarithmic transformation of $\hat{\sigma}^2_\psi$.

It was hinted above that the intuition of c for ψ in the range $(-1, -1/3]$ is not evident. The MLE is then unbiased or biased towards zero (Section 5.3) so a shift towards -1 would be needed for such values of ψ from the point of view of correcting for bias. The theoretical c is positive for the values of ψ under discussion so $\hat{c}$ should be positive, too with the consequence that the AE would be shifted towards 0 instead of -1. Coefficient c is moderate though when $\psi \in (-1, -1/3]$ which suggests a modest shift. The lesson seems to be that the Cox–Reid adjustment does not operate solely to diminish the bias or that it is not able to do it, in the present application, everywhere in the parameter space.

Differentiating equation (6.15) with respect to ψ and setting it equal to zero produces the adjusted likelihood equation[9]:

$$\begin{aligned} \frac{\partial l_{ad2}(\psi)}{\partial \psi} &= -\frac{(T-3)}{2\,\hat{\sigma}^2_{\psi}} \frac{\partial\, \hat{\sigma}^2_{\psi}}{\partial \psi} + \hat{c} \\ &= \frac{(T-3)(\psi_{ml} - \psi)}{\hat{l} - 2\psi\, \hat{\psi}_{ml} + \psi^2} + \hat{c} = 0 \end{aligned} \tag{6.16}$$

where

$$\frac{\partial\, \hat{\sigma}^2_{\psi}}{\partial \psi} = 2T^{-1}\left[\psi \sum_{t=1}^{T}(y_{t-1} - \bar{y}_{-1})^2 - \sum_{t=1}^{T}(y_t - \bar{y})(y_{t-1} - \bar{y}_{-1})\right]$$

and

$$\hat{l} = \frac{\sum_{t=1}^{T}(y_t - \bar{y})^2}{\sum_{t=1}^{T}(y_{t-1} - \bar{y}_{-1})^2} \geq 0. \tag{6.17}$$

Equation (6.16) is equivalent to:

$$\hat{\psi}_{ml} + \frac{\hat{c}}{T-3}\,\hat{l} - \left(1 + \frac{2\,\hat{c}}{T-3}\,\hat{\psi}_{ml}\right)\psi + \frac{\hat{c}}{T-3}\psi^2 = 0$$

---

[9]The notation $\hat{l}$ is designed to point out that the ratio converges stochastically to 1 as T tends to infinity whether $\psi = 1$ or $|\,\psi\,| < 1$. However, $\hat{l}$ tends to ψ^2 when the starting value relative to the unconditional mean tends to infinity, see Appendix A5.

which is a quadratic in ψ. The roots of the equation are:

$$\hat{\psi}_{ml} + \frac{T-3}{2\,\hat{c}} \pm \sqrt{\frac{(T-3)^2}{4\,\hat{c}^2} + \hat{\psi}_{ml}^2 - \hat{l}},$$

assuming $\hat{c} \neq 0$ (which is the case with probability one). It is useful to distinguish the cases:

i) $\hat{c} < 0$ and the roots are complex (with nonzero imaginary parts)

ii) $\hat{c} < 0$ and the roots are real (and unequal)

iii) $\hat{c} = 0$ (in which case the adjusted likelihood equation implies a polynomial of order one)

iv) $\hat{c} > 0$ and the roots are real (and unequal)

v) $\hat{c} > 0$ and the roots are complex (with nonzero imaginary parts).

Mathematically, two further circumstances exist: the roots are real but equal and $\hat{c} < 0$ or $\hat{c} > 0$. However, these events can happen only with probability zero. If the roots are real then $l_{ad2}(\psi)$ features a local maximum and a local minimum but increases monotonically if the roots are complex. The case $\hat{c} = 0$ occurs with probability zero too and would imply that the adjusted likelihood equation would shrink (essentially) to the usual likelihood equation.

Assuming real roots, the maximum of $l_{ad2}(\psi)$ occurs at the root $\hat{\psi}_{ml} + (T-3)/2\,\hat{c} - \sqrt{\cdot}$ if $\hat{c} > 0$ or at the root $\hat{\psi}_{ml} + (T-3)/2\,\hat{c} + \sqrt{\cdot}$ if $\hat{c} < 0$ where $\sqrt{\cdot} = \sqrt{(T-3)^2/4\,\hat{c}^2 + \hat{\psi}_{ml}^2 - \hat{l}}$ (Appendix A1). If the roots are complex then the term $\hat{c}\,(\psi - \hat{\psi}_{ml})$ must dominate $l_{ad2}(\psi)$ forcing it to decrease or increase monotonically with ψ if $\hat{c} > 0$ or $\hat{c} < 0$, respectively. As the parameter space is restricted to $(-1, 1]$ it seems inherent to interpret $l_{ad2}(\psi)$ as supporting an estimate equal to -1 or 1, respectively, in these situations. The incident $\hat{c} = 0$ would imply a maximum at $\hat{\psi}_{ml}$. Based on this reasoning we define

$$\hat{\psi}_{ad2} = \begin{cases} -1 & \text{if } i) \text{ applies} \\ \hat{\psi}_{ml} + \frac{T-3}{2\hat{c}} + \sqrt{\frac{(T-3)^2}{4\hat{c}^2} + \hat{\psi}_{ml}^2 - \hat{l}} & \text{if } ii) \text{ applies} \\ \hat{\psi}_{ml} & \text{if } iii) \text{ applies} \\ \hat{\psi}_{ml} + \frac{T-3}{2\hat{c}} - \sqrt{\frac{(T-3)^2}{4\hat{c}^2} + \hat{\psi}_{ml}^2 - \hat{l}} & \text{if } iv) \text{ applies} \\ 1 & \text{if } v) \text{ applies.} \end{cases} \tag{6.18}$$

The definition is devised so that $\hat{\psi}_{ad2}$ tends to increase relative to $\hat{\psi}_{ml}$ as the cases are scanned downwards. The ranking holds always for the circumstances from *ii)* to *iv)*.[10]

The terms of the adjusted profile log-likelihood are clearly of different magnitude if $|\psi| < 1$: $-[(T-3)/2]\log\hat{\sigma}_{\psi}^2$ remains $O_p(T)$ regardless of the point of evaluation, while the term $\psi - \hat{\psi}_{ml}$ becomes multiplied by $\hat{c}$ which is $O_p(1)$. If $\psi = 1$ then $\hat{c}$ is $O_p(T)$ but the first term dominates still as it must be of greater order than $O_p(T)$ in general ($\hat{\sigma}_{\psi}^2$, evaluated at a value of ψ different from unity, is $O_p(T)$ under $\psi = 1$). So $l_{ad2}(\psi)$ should feature curvature, a local maximum and real roots regardless of ψ according to asymptotics. Further, it is shown in Appendix A1 that case *iv)* applies asymptotically in general.[11] The simulations reported in Section 7.2 confirm that it is the standard even for very small samples when $\psi = 1$.

Complex roots relate to events in which the variation of the term $-[(T-3)/2]\log\hat{\sigma}_{\psi}^2$ with ψ is overshadowed by the variation of the term $\hat{c}\,(\psi - \hat{\psi}_{ml})$ in the adjusted profile log-likelihood. The simulations in Section 7.2 will reveal that this can happen only if the

---

[10]The ranking would break, say, if $\hat{\psi}_{ml}$ were larger than one but the roots were complex. Another possibility would be to define $\hat{\psi}_{ad2}$ to equal $\hat{\psi}_{ml}$ also in cases *i)* and *v)*. The AE could not then be closer to zero than the MLE. A further alternative would be to define $\hat{\psi}_{ad2}$ equal to the modulus of the root or the real part of the root. The complex cases could also be ignored, i.e. dropped from the data sets in the simulation exercises of Chapter 7 but this would be the least acceptable solution in our opinion. It is not a settlement for practitioners, and it might bias Monte Carlo statistics, too.

[11]Heuristically, this can be seen as follows. The radical in *iv)* is $\sqrt{\frac{(T-3)^2}{4c^2} + \psi^2 - 1}$ ignoring sampling errors. The $\psi^2 - 1$ term becomes asymptotically negligible under $|\psi| < 1$ and equal to zero under $\psi = 1$, so $\hat{\psi}_{ml} + \frac{T-3}{2c} - \sqrt{\cdot} \approx \hat{\psi}_{ml}$ $(\underset{p}{\rightarrow} \psi)$ for large T. Relatedly, it is shown in Appendix A1 that asymptotically $\partial^2 l_{adj2}(\psi)/\partial^2\psi < 0$ (indicating a local maximum) at the smaller root, and that the other root lies in the region $(1, \infty)$ or outside the parameter space of ψ.

sample size is extremely small or $\hat{c}$ is exceedingly large due to a MLE larger than one.

It is proved in Appendix A1 that $\hat{\psi}_{ad2}$ can be expressed as

$$\hat{\psi}_{ad2} = \hat{\psi}_{ml} - \frac{\hat{c}\,(\hat{\psi}^2_{ml} - \hat{l})}{T-3} + r_T, \qquad r_T = \begin{cases} O_p(T^{-3}) & \text{if } |\psi| < 1 \\ O_p(T^{-2}) & \text{if } \psi = 1. \end{cases} \tag{6.19}$$

The approximation is apt for deriving the asymptotic distributions but it provides insight, too.

It is also shown in Appendix A1 that $\hat{\psi}^2_{ml} - \hat{l} \leq 0$ with the equality applying only asymptotically.[12] The term $\hat{c}\ (\hat{\psi}^2_{ml} - \hat{l})$ tends thus to be negative as $\hat{c}$ is positive for large T implying then that $\hat{\psi}_{ad2} > \hat{\psi}_{ml}$ in harmony with our previous remarks. A closer scrutinisation of the expression affirms our previous comments on the effect of the magnitude of $\hat{c}$ on the difference between $\hat{\psi}_{ad2}$ and $\hat{\psi}_{ml}$ (a large $\hat{\psi}_{ad2}$ relative to $\hat{\psi}_{ml}$ is related to a large $\hat{c}$).[13]

The strength of the adjustment depends also on the informativeness of the sample. Substituting the MLE into equation (A1.4) in Appendix A1 reveals that the second derivative of the adjusted profile log-likelihood evaluated at the MLE is:

---

[12]It is easy to see that $\hat{\psi}^2_{ml} / \hat{l} = (\hat{r}_1)^2$ or the square of the sample autocorrelation coefficient between y_t and y_{t-1}. The coefficient lies between zero and one so $\hat{\psi}^2_{ml}$ must be smaller than or equal to $\hat{l}$. The proof in Appendix A1 makes use of the Cauchy–Schwarz inequality explicitly. The equality implies that $\hat{l} = \hat{\psi}^2_{ml} / (\hat{r}_1)^2$. This makes it also clear that $\hat{l}$ tends in probability to 1: the terms on the right-hand side both tend in probability to ψ^2 and hence the ratio tends in probability to 1. Furthermore, the relation $\hat{\psi}^2_{ml} = \hat{l}\ (\hat{r}_1)^2$ provides an alternative expression for the approximation: $\hat{\psi}_{ad2} \approx \hat{\psi}_{ml} - \hat{c}\ [(\hat{r}_1)^2 - 1]\ \hat{l}\ /(T-3)$.

[13]The terms in the multiplication $\hat{c}\ (\hat{\psi}^2_{ml} - \hat{l})$ are not independent so it is not quite legitimate to focus solely on the size of $\hat{c}$ on the adjustment. Actually, a large $\hat{c}$ is related to a large (and positive) $\hat{\psi}^2_{ml}$ and hence small $(\hat{\psi}^2_{ml} - \hat{l})$. However, asymptotically at least, this is more than compensated by a larger $\hat{c}$ which increases more with ψ than $\hat{\psi}^2_{ml} - \hat{l}$ decreases with ψ. If ψ is negative then $\hat{\psi}^2_{ml} - \hat{l}$ should increase in absolute value with ψ and hence contribute to a larger adjustment together with $\hat{c}$.

$$\left[\frac{\partial^2 l_{ad2}(\psi)}{\partial^2 \psi}\right]\Bigg|_{\psi=\hat{\psi}_{ml}} = \frac{T-3}{\hat{\psi}_{ml}^2 - \hat{l}}.$$

A similar derivation for the profile log-likelihood yields:

$$\left[\frac{\partial^2 l_p(\psi)}{\partial^2 \psi}\right]\Bigg|_{\psi=\hat{\psi}_{ml}} = \frac{T}{\hat{\psi}_{ml}^2 - \hat{l}}. \tag{6.20}$$

The two likelihoods feature quite similar curvature at the MLE. Comparison to approximation (6.19) reveals that the additive adjustment term there depends on the curvature of the adjusted profile log-likelihood. The smaller $(T-3)/(\hat{\psi}_{ml}^2 - \hat{l})$ is in absolute value or the less curved the (adjusted) profile log-likelihood is, the larger the additive adjustment term in equation (6.19) is (given the MLE and hence $\hat{c}$). The adjustment gains strength when the sample emerges uninformative.

The effect of multiplying $\log \hat{\sigma}_{\psi}^2$ by $-(T-3)/2$ in the adjusted profile log-likelihood instead of $-T/2$ (as in the ordinary profile log-likelihood (6.5)) can be considered now, too. It can be seen from approximation (6.19) that substituting T in place of $T-3$ in approximation (6.19) would shrink the AE towards the MLE. Consequently, the $T-3$ modification of the adjusted log-likelihood (6.5) reinforces the tendency of the term $\hat{c}$ $(\hat{\psi}_{ml} - \psi)$ to increase the AE over that of the MLE. In conclusion, the degrees-of-freedom correction tends to make the AE depart from the MLE, or increase it (assuming a positive $\hat{c}$), in parallel with the standard case of such a correction when estimating residual variance (cf. the examples on pp. 42–44). Asymptotically it does not matter whether $\log \hat{\sigma}_{\psi}^2$ is multiplied by $-(T-3)/2$ or $-T/2$.

It can be seen that:

$$T^{1/2}(\hat{\psi}_{ad2} - \psi) = T^{1/2}(\hat{\psi}_{ml} - \psi) - T^{1/2}\frac{\hat{c}\,(\hat{\psi}_{ml}^2 - \hat{l})}{T-3} + r_T^*,$$

$$r_T^* = O_p(T^{-5/2}), \quad |\psi| < 1$$

when the model is stationary. The second term on the right-hand side is $O_p(T^{-1/2})$ because $\hat{c}$ is $O_p(1)$ or tends stochastically to $(1-\psi)^{-1}$ and $(\hat{\psi}^2_{ml} - \hat{l})$ is $O_p(1)$ or tends stochastically to $\psi^2 - 1$ if $|\psi| < 1$ (Appendix A1). It follows that the asymptotic distribution of the standardised AE or $T^{1/2}(\hat{\psi}_{ad2} - \psi)$ is the same as that of the standardised ordinary MLE if the process is stationary. In other words, there is no adjustment asymptotically if $|\psi| < 1$.

The nonstationary case is different. The appropriate standardising factor is T and we find

$$T(\hat{\psi}_{ad2} - 1) = T(\hat{\psi}_{ml} - 1) - T\frac{\hat{c}\,(\hat{\psi}^2_{ml} - \hat{l})}{T-3} + r^*_T, \tag{6.21}$$

$$r^*_T = O_p(T^{-1}), \quad \psi = 1.$$

The second term on the right-hand side is now $O_p(1)$ as $\hat{c}$ is $O_p(T)$ and $(\hat{\psi}^2_{ml} - \hat{l})$ is $O_p(T^{-1})$ when $\psi = 1$ (Appendix A1). The adjustment persists even asymptotically, and the disparity in the order of magnitude of the adjustment in the stationary and nonstationary case — noted when discussing the adjusted profile log-likelihood — is verified.

A further way to interpret the strength of the adjustment is to notice that

$$-T\left(\hat{\psi}^2_{ml} - \hat{l}\right) = \frac{\hat{\sigma}^2_{ml}}{T^{-2}\sum_{t=1}^{T}(y_{t-1} - \bar{y}_{-1})^2} \geq 0.$$

The numerator is $O_p(1)$ whether the model is stationary or not but the denominator or the sample variance multiplied by T^{-1} is $O_p(T^{-1})$ under stationarity or $O_p(1)$ under a unit root. The ratio and the adjustment hinge on the magnitude of the sample variance of y_t which is larger under nonstationarity.

We can write the right-hand side of approximation (6.21) to form:

$$\begin{aligned} & T(\hat{\psi}_{ml} - 1) - \tfrac{\hat{c}}{T-3}(\hat{\psi}_{ml} + 1)T(\hat{\psi}_{ml} - 1) + \tfrac{\hat{c}}{T-3}T(\hat{l} - 1) + r_T^* \\ & = T(\hat{\psi}_{ml} - 1)\left[1 - \tfrac{\hat{c}}{T-3}(\hat{\psi}_{ml} + 1)\right] + \tfrac{\hat{c}}{T-3}T(\hat{l} - 1) + r_T^*. \end{aligned} \tag{6.22}$$

The distribution of this expression depends on the method of evaluating $\hat{c}$, namely if it is specified as $\hat{c}= (1- \hat{\psi}_{ml})^{-1} - (1- \hat{\psi}_{ml}^{T})/T(1- \hat{\psi}_{ml})^2$, or $\hat{c}= (T-1)/2$ which employs the *a priori* information $\psi = 1$ (formula (6.14)). We shall first consider the latter case which is simpler.

The term in square brackets, and hence the first term as a whole in the second line of expression (6.22), tends stochastically to zero because

$$\lim_{T\to\infty}\left[\hat{c}\,/(T-3)\right] = 1/2 \tag{6.23}$$

when $\hat{c}= (T-1)/2$,

$$plim\left[1 - \tfrac{\hat{c}}{T-3}(\hat{\psi}_{ml} + 1)\right] \;=\; 0,$$

and because $T(\hat{\psi}_{ml} - 1)$ is $O_p(1)$. The $T(\hat{l} - 1)$ part in expression (6.22) becomes for large T (cf. Appendix A1)

$$\begin{aligned} T(\hat{l} - 1) \;&=\; \frac{T^{-1}y_T(y_T - 2\,\bar{y}_{-1})}{T^{-2}\sum_{t=1}^{T}(y_{t-1} - \bar{y}_{-1})^2} + O_p(T^{-1/2}). \\ &\Rightarrow\; \frac{W(1)\left[W(1) - 2\int_0^1 W(r)dr\right]}{\int_0^1 [W_*(r)]^2\,dr} \end{aligned} \tag{6.24}$$

where the weak convergence is due to well-known theorems (Hamilton (1994, p. 486), say). It follows that

$$T(\hat{\psi}_{ad2,ap} - 1) \;\Rightarrow\; \frac{\frac{1}{2}W(1)\left[W(1) - 2\int_0^1 W(r)dr\right]}{\int_0^1 [W_*(r)]^2\,dr} \tag{6.25}$$

where $\hat{\psi}_{ad2,ap}$ (*ap* for *a priori*) is the adjusted estimate employing the *a priori* information

$\hat{c}= (T-1)/2$. (We shall sometimes refer to this sort of estimate as AE_{ap}). The distribution will be alluded to as $AD_{\mu,ap}$ (AD for adjusted). Of course, this is a different distribution from the Dickey–Fuller distribution to which $T(\hat{\psi}_{ml} -1)$ tends.

The following alternative way to derive the asymptotic distribution provides extra insight. It is shown in Appendix A1 that

$$T(\hat{\psi}^2_{ml} - \hat{l}) \Rightarrow -\left\{\int_0^1 [W_*(r)]^2\, dr\right\}^{-1} . \tag{6.26}$$

Combining this information with formulae (6.7), (6.21), and (6.23) yields

$$T(\hat{\psi}_{ad2,ap} -1) \;\Rightarrow\; \frac{\frac{1}{2}\left\{[W(1)]^2 - 1\right\} - W(1)\int_0^1 W(r)dr}{\int_0^1 [W_*(r)]^2\, dr} - \frac{-1/2}{\int_0^1 [W_*(r)]^2\, dr}$$

or

$$T(\hat{\psi}_{ad2,ap} -1) \;\Rightarrow\; \frac{\frac{1}{2}W(1)\left[W(1) - 2\int_0^1 W(r)dr\right]}{\int_0^1 [W_*(r)]^2\, dr}.$$

The Cox–Reid adjustment term $-\hat{c}\, T(\hat{\psi}^2_{ml} - \hat{l})/(T-3)$ in equation (6.21) has the impact of adding a '1/2' into the numerator of the ratio which defines the asymptotic distribution of the standardised MLE. The '1/2' and the '$-1/2$' cancel in the numerators. Comparison with formula (5.15) (or the asymptotic distribution of the statistic defined by it) reveals that this is the asymptotic correction which Rothenberg (1995) has independently devised and proved to yield an asymptotically unbiased symmetrically distributed estimate under $\psi = 1$.[14,15] Another remark on this derivation is that it embodies

[14]We will not prove this result but the fact that $W(1)$ in the numerator is a Standard Normal variate hints that this might be the case. The square-bracket term in the numerator has expectation zero, too: $\int_0^1 W(r)dr$ is a Normal variate with mean zero and variance 1/3 (Banerjee *et al.* (1993, pp. 43-5), say). More specifically, it is shown in Appendix A1 that $T^{-1/2}(y_T - 2\,\bar{y}_{-1}) \Rightarrow N(0, \sigma^2/3)$. Thus the numerator tends in distribution to $N(0,\sigma^2/3)\cdot N(0,\sigma^2/3)$ (where the two Normal variates are not independent).

We had independently derived the asymptotic distribution (6.25) in essence by May 1994 and in detail by October 1994. The discussion relating the present results to those of Rothenberg (1995) was added in March 1995.

[15]The canceling of the one half in the numerator of the asymptotic distribution of the MLE might

our comments on formula (6.19): the Cox–Reid adjustment shifts the distribution of the MLE asymptotically to the right as the adjustment term $\frac{1}{2}\left\{\int_0^1 [W_*(r)]^2\,dr\right\}^{-1}$ is positive (with probability one).

Figure 6.6 illuminates distribution (6.25). Monte Carlo statistics are again provided with the figure. The distribution features zero mean (within simulation error) and symmetry, and is far less scattered than the distribution of the corresponding MLE (Figure 6.4). The adjustment performs extremely well in asymptotic terms as it removes the bias altogether and produces a relatively tight distribution as measured by SD or MSE. Furthermore, the distribution is much more compact than the asymptotic distribution of the MLE without constant (6.4). The present SD is only about a third of the SD of the corresponding MLE and about a half of the SD of the MLE without a constant. The analogous ratios for the MSE are about one eighteenth and about a fifth, respectively. The distribution in Figure 6.6 is not Normal as the Monte Carlo kurtosis is 2.045 (the standardised measure takes value zero under normality, not reported in the figure).[16]

Unfortunately, use of *a priori* information of $\psi = 1$ or using the formula $\hat{c} = (T-1)/2$ in the calculation of the AE is not sensible even when a unit root is considered the null. Inspection of definition (6.18), case *iv*), reveals that

$$\begin{aligned}\hat{\psi}_{ad2,ap} &= \hat{\psi}_{ml} + \frac{T-3}{2\hat{c}} - \sqrt{\frac{(T-3)^2}{4\hat{c}^2} + \hat{\psi}_{ml}^2 - \hat{l}} \\ &\xrightarrow[p]{} \psi + \tfrac{1}{2}\cdot 2 - \sqrt{\tfrac{1}{4}\cdot 2^2 + \psi^2 - 1} \quad = \quad 1\end{aligned}$$

regardless of the true value of ψ. In other words, $\hat{\psi}_{ad2,ap}$ would be inconsistent for

---

suggest that a similar adjustment would yield an unbiased estimate under a unit root for the simple AR(1) model of Section (6.2). This does not happen. The asymptotic distribution of the standardised MLE of ψ under this model is $\frac{1}{2}\left\{[W(1)]^2 - 1\right\} / \int_0^1 [W(r)]^2$ (formula (3.11)). Hence there is a 'one half' which would cancel asymptotically if $T\sigma^2/2$ were added to the numerator of the standardised MLE as in equation (5.15) for the $AR_\mu(1)$ model. However, such a modification cannot yield a zero-mean symmetric distribution because $\frac{1}{2}[W(1)]^2 / \int_0^1 [W(r)]^2$ is positive with probability one. Interestingly, neither Rothenberg's argument (personal communication) or the Cox–Reid procedure yield this nor any adjustment for the simple AR(1) model.

[16]Some caution is in place when interpreting the sample kurtosis of a very large sample. The number of decimal points included in the original data can cause vast discrepances in the sample kurtosis, at least with the algorithm we have used.

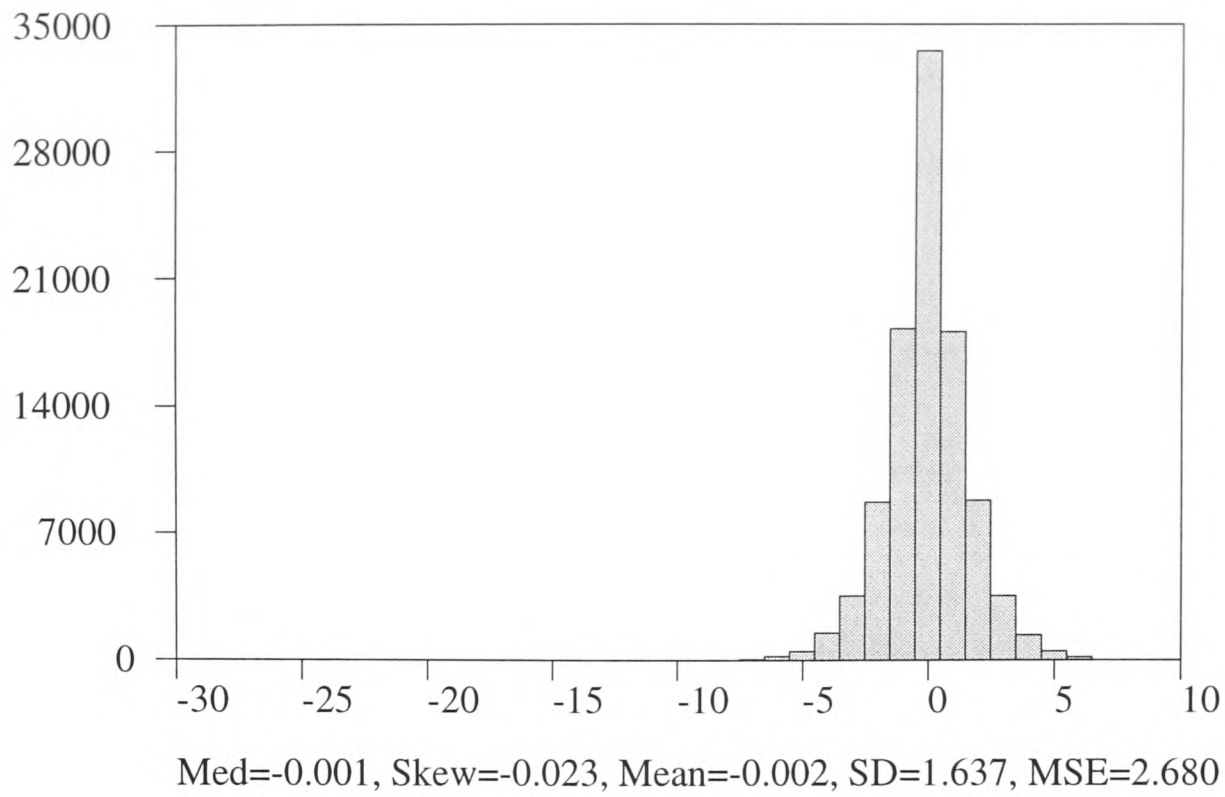


Figure 6.6 Empirical asymptotic distribution of $T(\hat{\psi}_{ad2,ap} - 1)$ for $\psi = 1$ ($AR_{\mu}(1)$).

$| \psi | < 1$.[17] The result is discouraging from the point of view of testing for unit roots but it is noted that a test statistic like $T(\hat{\psi}_{ad2,ap} - 1)$ can display power if the convergence to unity under stationarity takes place at rate $O_p(T^{-(1-\delta)})$ where $0 < \delta < 1$. In Section 7.3 we point out an alternative sense in which $\hat{\psi}_{ad2,ap}$ and a test based on it can be consistent and useful.

The standardised coefficient $\hat{c}\ /T$ converges stochastically to a distribution instead of a constant if the *a priori* information $\psi = 1$ is not made use of in the estimation. The sole tricky part in finding the distribution is attaining an asymptotic expression for $\hat{\psi}^{T}_{ml}$.

The mean-value theorem states that $f(\psi) = f(\psi_0) + [\partial f(\psi)/\partial\psi]\ |_{\psi=\psi^*}\ (\psi - \psi_0)$ where ψ^* lies between ψ and ψ_0. It follows that $\log \hat{\psi}^{T}_{ml} = T(\hat{\psi}_{ml} - 1)/\psi^*$ by letting $f(\hat{\psi}_{ml}) = \log \hat{\psi}^{T}_{ml}$ and $\psi_0 = 1$.[18] Noting that $|\psi^* - 1| \leq \left|\hat{\psi}_{ml} - 1\right|$ and that the latter tends in probability to zero, so the former must tend in probability to zero, too, or ψ^*

[17]Note that approximations like (6.22) or (6.19) are not serviceable for analysing consistency in the case under study because the remainder terms are of larger order of magnitude than stated in the text if $c = (T-1)/2$ simultaneously with $| \psi | < 1$.

[18]We are assuming here that $\hat{\psi}_{ml}$ does not take negative values so that $\log(\hat{\psi}_{ml})$ is defined which should hold well for a large enough sample size. A technically more sound proof would redefine the MLE so that it cannot take negative values and reason the asymptotic distribution from the asymptotic distribution of the redefined MLE.

to one. It follows that $\log \hat{\psi}_{ml}^{T}$ converges weakly to the same distribution as $T(\hat{\psi}_{ml} - 1)$ which is the Dickey–Fuller distribution (6.8) or DF_{μ}. We find that

$$\hat{\psi}_{ml}^{T} \Rightarrow \exp(DF_{\mu}) \tag{6.27}$$

after taking anti-logarithms. Now it is straightforward to infer that

$$\begin{aligned} \hat{c}\,/T &= \frac{-1}{T(\hat{\psi}_{ml} - 1)} - \frac{1 - \hat{\psi}_{ml}^{T}}{[T(\hat{\psi}_{ml} - 1)]^2} \\ &\Rightarrow -\frac{1}{DF_{\mu}} - \frac{1 - \exp(DF_{\mu})}{(DF_{\mu})^2} \end{aligned}$$

or

$$\hat{c}\,/T \Rightarrow \frac{\exp(DF_{\mu}) - (1 + DF_{\mu})}{(DF_{\mu})^2} \in (0, \infty) \tag{6.28}$$

where the '$\in$' sign follows from the properties of the exponential function and applies with probability one. Though extreme values of $\hat{c}\,/T$ are possible in principle it should lie most likely within the span $(0, 1)$: DF_{μ} should take a value larger than (approximately) 1.8 for $\hat{c}\,/T$ to exceed one and the probability of such an event is quite small (about 0.003). Figure 6.7 sketches the distribution. Most of the probability mass lies between zero and one with emphasis on values close to zero, and the right tail is long. In contrast, coefficient c evaluated at $\psi = 1$ tends to exactly one half as T tends to infinity.

Coefficient c tends here to a distribution instead of a constant because appropriately standardised information measures evaluated at the MLE do not necessarily converge to a constant under a unit root in the present context. In the stationary case this does not happen and $\hat{c}$ converges stochastically to $(1 - \psi)^{-1}$.

The asymptotic distribution of $T(\hat{\psi}_{ad2} - 1)$ can now be readily derived from expression (6.22):

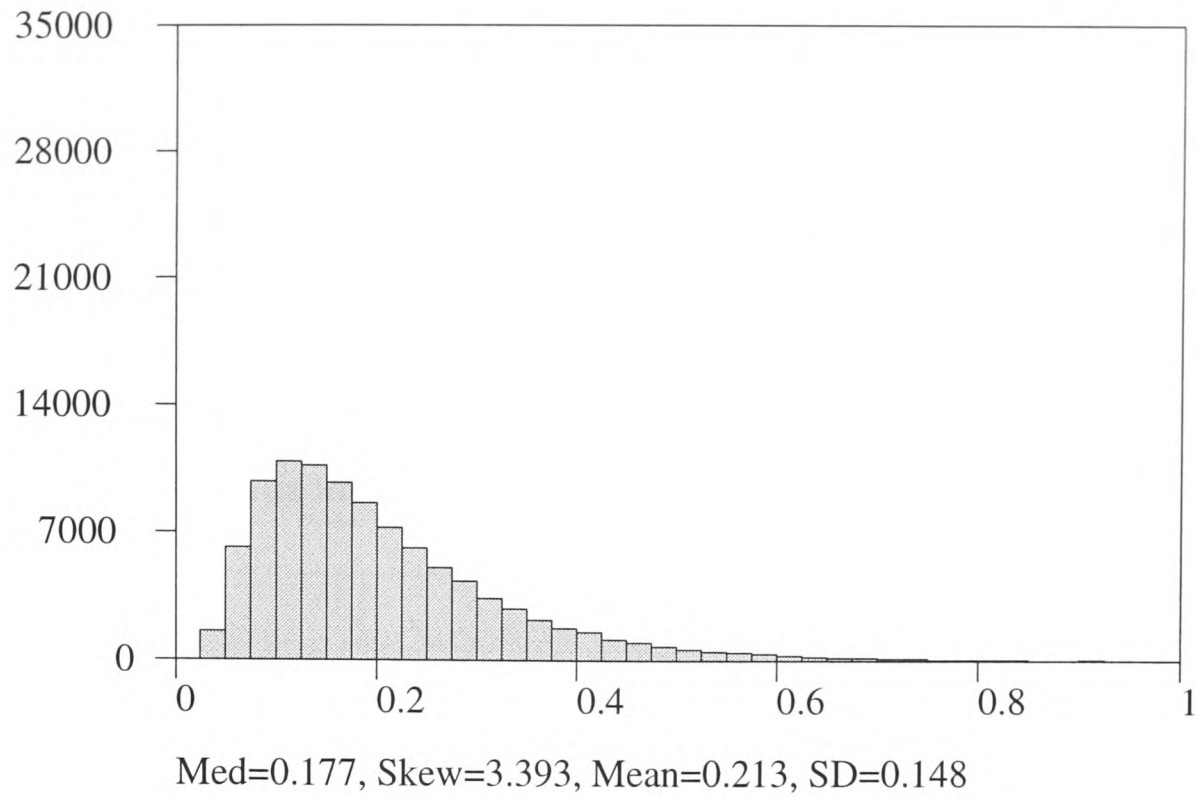


Figure 6.7 Empirical asymptotic distribution of the data-based $\hat{c}$ for $\psi = 1$ ($\mathrm{AR}_\mu(1)$).

$$T(\hat{\psi}_{ad2} - 1) \;\Rightarrow\; DF_\mu - 2\left[\frac{\exp(DF_\mu) - (1 + DF_\mu)}{(DF_\mu)^2}\right](DF_\mu - AD_{\mu,ap})\,. \tag{6.29}$$

or

$$T(\hat{\psi}_{ad2} - 1) \;\Rightarrow\; DF_\mu + \left[\frac{\exp(DF_\mu) - (1 + DF_\mu)}{(DF_\mu)^2}\right]\left\{\int_0^1 [W_*(r)]^2\, dr\right\}^{-1}. \tag{6.30}$$

We shall label the distribution as AD_μ.

It is interesting to remark that if the term $\exp(DF_\mu)$ were approximated by $1 + DF_\mu + \frac{1}{2}(DF_\mu)^2$ then the asymptotic distribution $AD_{\mu,ap}$ or distribution (6.25) of the *a priori* adjusted MLE would result. However, the approximation is not good as can be seen by comparing Figures 6.6 and 6.8 which depict $AD_{\mu,ap}$ and AD_μ, respectively. (The fractiles of AD_μ are reported in Table 7.1 of Section 7.2.) The distributions are quite different.

AD_μ resembles the Dickey–Fuller distribution DF_μ (Figure 6.4) quite a bit, especially at the left tail. Visually, apparent differences include the right-ward shift and the stretched right tail of AD_μ compared to DF_μ. The Monte Carlo statistics, especially

the medians and the means, accompanying Figures 6.4 and 6.8, confirm the shift. The slightly larger SD of AD_μ over that of DF_μ is probably due to the longer right tail of the former. The same phenomena is reflected in the reduction of skewness, too. Most interestingly, we have been able to beat the MSE of the MLE by about 30 per cent. However, the differences between the distributions are not too drastic in general. (The finite-sample distributions are more clearly distinct, cf. Section 7.2.)

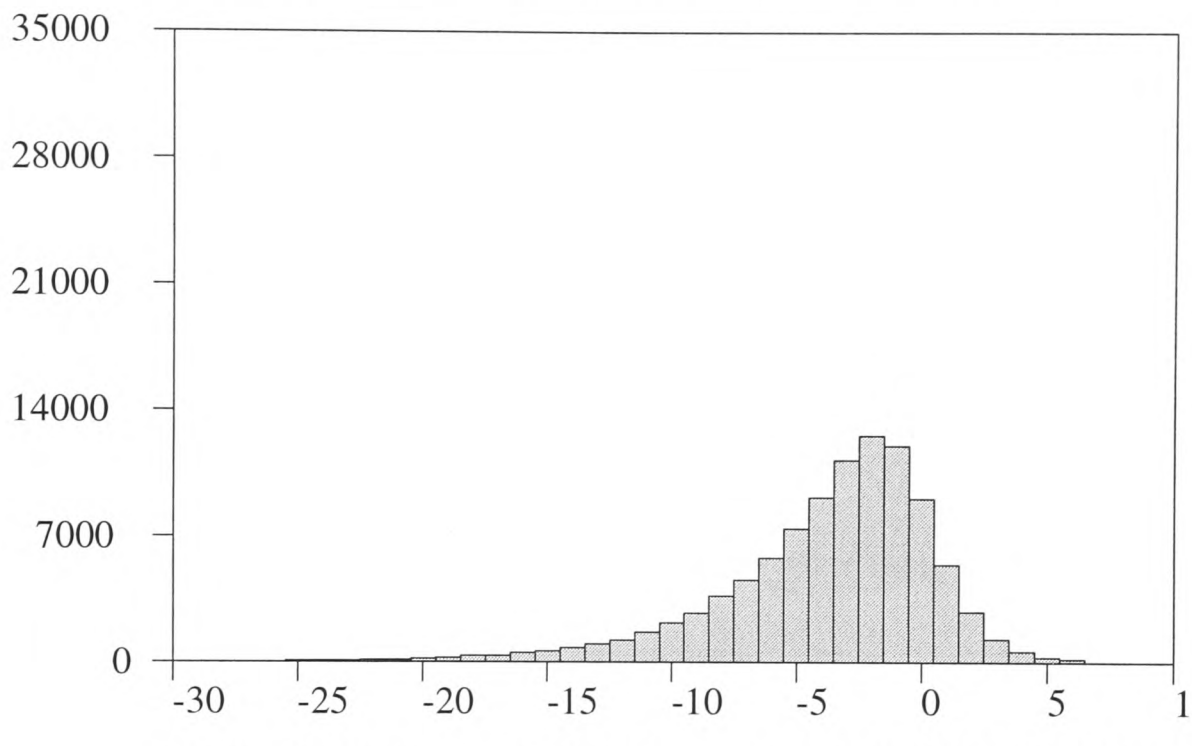


Figure 6.8 Empirical asymptotic distribution of $T(\hat{\psi}_{ad2} - 1)$ for $\psi = 1$ ($\mathrm{AR}_\mu(1)$).

Figure 6.9 presents the asymptotic joint distribution of $T(\hat{\psi}_{ml} - 1)$ and $T(\hat{\psi}_{ad2} - 1)$ based on 100 000 draws from DF_μ and AD_μ (each point corresponds to the same realization of an approximation to Brownian motion). The observations from AD_μ lie consistently above those from DF_μ. (The line in the figure indicates the points were equality would occur.) The figure reinforces our previous comments on the disparity of $\hat{\psi}_{ml}$ and $\hat{\psi}_{ad2}$: if $\hat{\psi}_{ml}$ is very large — over one say — then the difference between the estimates tends to widen. This explains why the right tail of AD_μ is stretched compared to DF_μ. The Monte Carlo correlation coefficient (r) reported in the figure is very high. (The coefficient is a measure of linear correlation so it does not completely capture the strength of the relationship between the variables.) An explanation for the similarity of

the distributions (excluding the right tail) seems to be the strong correlation.

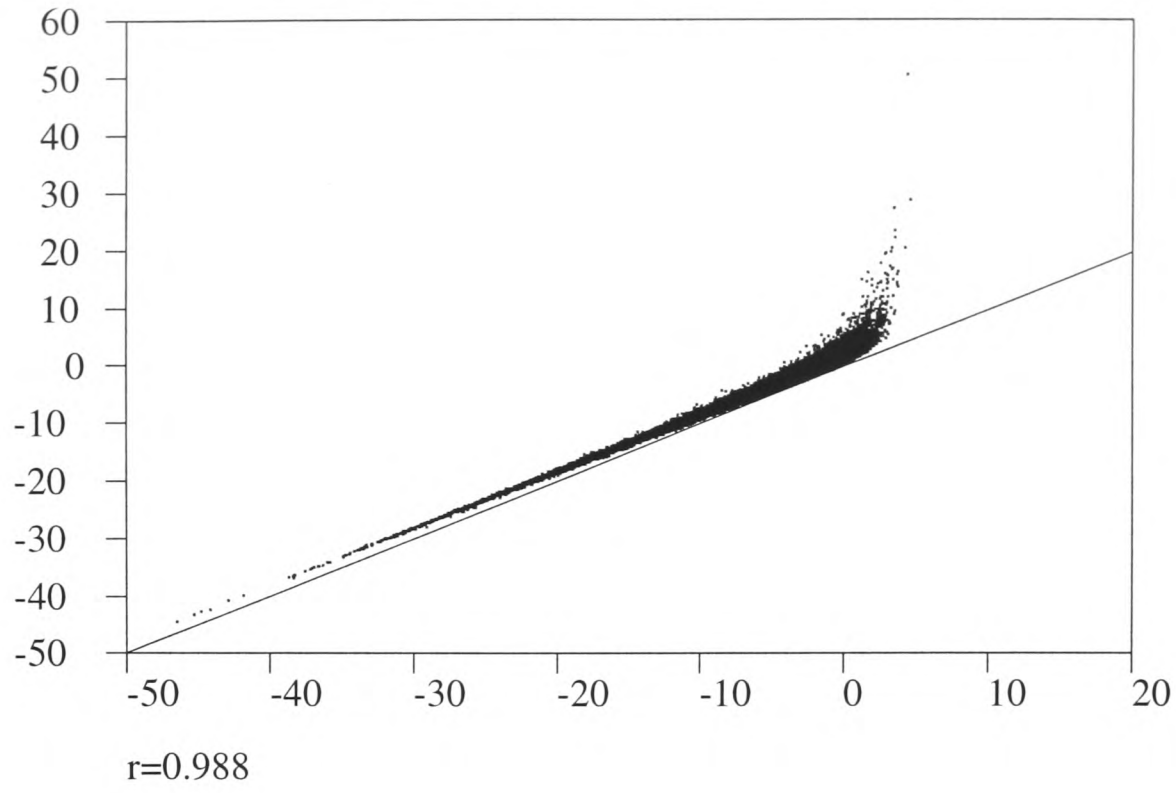


Figure 6.9 Empirical asymptotic joint distribution of $T(\hat{\psi}_{ml} - 1)$ and $T(\hat{\psi}_{ad2} - 1)$ for $\psi = 1$ ($\mathrm{AR}_\mu(1)$).

It is not often when one would be willing to adjust the MLE asymptotically. However, the standard regularity conditions behind the usual optimality results are not valid in the present context as already pointed out in Section 4.1. Indeed, we have found that the MSE of the AE is asymptotically notably smaller than the MSE of the corresponding MLE.

Inaccuracy in the estimation of coefficient $\hat{c}$ deteriorates the (asymptotic) distribution of the AE compared to the circumstance when the *a priori* information is employed. Presumably the coefficient could be estimated more accurately after finding $\hat{\psi}_{ad2}$ and evaluating the coefficient at it. The new estimate of c could be substituted into the adjusted profile log-likelihood, and the ψ that would maximize the revised likelihood might be a more accurate estimate.[19] Repeated iteration does not seem sensible though, as it would probably (if not definitely) lead to an ever-increasing estimate of ψ and c

[19]One might wonder whether the AE should be substituted in place of the MLE in the term $(\psi - \hat{\psi}_{ml})$ as well. This would change only the value of the likelihood at the maximum but not the ψ at which the maximum takes place.

because a larger (positive) c implies a larger adjusted estimate *ceteris paribus.*[20] Alternatively, the expression $\hat{\psi}_{ml} + \frac{T-3}{2\hat{c}} - \sqrt{\frac{(T-3)^2}{4\hat{c}^2} + \hat{\psi}_{ml}^2 - \hat{l}}$ might be evaluated iteratively once or until convergence takes place with $\hat{\psi}_{ml}$ replaced by the AE in the first iteration and by the resulting revised estimate thereafter with $\hat{c}$ evaluated similarly at the revised estimate from the previous iteration.[21] A third possibility would be to evaluate $\hat{\psi}_{ml} + \frac{T-3}{2\hat{c}} - \sqrt{\frac{(T-3)^2}{4\hat{c}^2} + \hat{\psi}_{ml}^2 - \hat{l}}$ again once[22] or repeatedly until convergence by replacing $\hat{\psi}_{ml}$ by the revised estimate iteratively but with the original $\hat{c}$. The two latter procedures would be quite exploratory as they do not share the intuition of the first procedure of maximizing the revised likelihood. The above iteration methods will be referred to as procedure 1) (a one-step iteration is assumed with this technique), procedure 2), and procedure 3), respectively. We shall study only procedure 1) in detail, but the other two procedures are experimented within the example below, too. For clarity, we define the iteratively adjusted estimate $\hat{\psi}_{ad2,i}$ of procedure 1) formally:

$$\hat{\psi}_{ad2,i} = \begin{cases} -1 & \text{if } i) \text{ or } i^*) \text{ applies} \\ \hat{\psi}_{ml} + \frac{T-3}{2\hat{c}_i} + \sqrt{\frac{(T-3)^2}{4\hat{c}_i^2} + \hat{\psi}_{ml}^2 - \hat{l}} & \text{if } ii) \text{ applies} \\ \hat{\psi}_{ml} & \text{if } iii) \text{ applies} \\ \hat{\psi}_{ml} + \frac{T-3}{2\hat{c}_i} - \sqrt{\frac{(T-3)^2}{4\hat{c}_i^2} + \hat{\psi}_{ml}^2 - \hat{l}} & \text{if } iv) \text{ applies} \\ 1 & \text{if } v) \text{ or } v^*) \text{ applies} \end{cases} \tag{6.31}$$

where $\hat{c}_i$ and cases $i)$ to $v^*)$ are defined below.

---

[20]Relatedly, the root of the new likelihood equation would become complex if $\hat{c}$ were allowed to increase without limit: The first term $(T-3)^2/4\,\hat{c}^2$ in the radical of the root would tend to zero with an increasing $\hat{c}$ while the second term $\hat{\psi}_{ml}^2 - \hat{l}$ would remain negative.

[21]Zigzag convergence would be possible with this procedure as there is no guarantee that the revised estimate squared minus $\hat{l}$ would be nonpositive as with the MLE.

[22]The values from the one-time iterations appear better than the ones from the repeated iteration for the data of the example below. The results for the one-time iterations are not reported there, though.

i) $\hat{c}_i < 0$ and the roots are complex (with nonzero imaginary parts)

*i**) $\hat{c} < 0$ and the roots of the original adjusted likelihood equation are complex (with nonzero imaginary parts)

ii) $\hat{c}_i < 0$ and the roots are real (and unequal)

ii) $\hat{c}_i = 0$ (in which case the adjusted likelihood equation implies a polynomial of order one)

iv) $\hat{c}_i > 0$ and the roots are real (and unequal)

v) $\hat{c}_i > 0$ and the roots are complex (with nonzero imaginary parts).

*v**) $\hat{c} > 0$ and the roots of the original adjusted likelihood equation are complex (with nonzero imaginary parts)

Cases without an asterisk read the same as on p. 81 apart from the fact that coefficient $\hat{c}$ has been replaced by $\hat{c}_i$, and that the roots refer here to the roots of the revised adjusted likelihood equation

$$\frac{\partial l_{ad2,i}(\psi)}{\partial \psi} = -\frac{(T-3)}{2\,\hat{\sigma}^2_{\psi}} \frac{\partial\, \hat{\sigma}^2_{\psi}}{\partial \psi} + \hat{c}_i = 0$$

where $l_{ad2,i}(\psi)$ is defined by

$$l_{ad2,i}(\psi) = -\frac{(T-3)}{2} \log \hat{\sigma}^2_{\psi} + \hat{c}_i\,(\psi - \hat{\psi}_{ml}).$$

Cases *i**) and *v**) allow for the possibility that the roots of the original adjusted likelihood equation are complex. In other words, we check first that the roots of the original adjusted likelihood equation are real and then apply the same logic in the definition of the iterated estimate as when defining the original adjusted estimate. If the original adjusted likelihood equation had complex roots then we define $\hat{\psi}_{ad2,i} = \hat{\psi}_{ad2}$ (and $\hat{c}_i = \hat{c}$).[23]

[23]It seems sensible to define $\hat{\psi}_{ad2,i} = \hat{\psi}_{ad2}$ and $\hat{c}_i = \hat{c}$ or in a sense not to enter the iteration if the roots of the original adjusted likelihood equation were complex as the way $\hat{\psi}_{ad2}$ is defined in the complex cases is not the sole meaningful definition. (For example, one could argue for defining $\hat{\psi}_{ad2} = \hat{\psi}_{ml}$ in such cases.) Under *v**) another possibility would be to define $\hat{c}_i$ equal to its theoretical value $(T-1)/2$ at

The asymptotic distribution of $\hat{\psi}_{ad2,i}$, after standardisation, is easily derived. The sole difference to the distribution AD_μ arises from the fact that c is evaluated at $\hat{\psi}_{ad2}$ instead of $\hat{\psi}_{ml}$. Hence

$$\begin{aligned} \hat{c}_i\,/T &= \frac{-1}{T(\hat{\psi}_{ad2}-1)} - \frac{1-\hat{\psi}_{ad2}^{T}}{[T(\hat{\psi}_{ad2}-1)]^2} \\ &\Rightarrow -\frac{1}{AD_\mu} - \frac{1-\exp(AD_\mu)}{(AD_\mu)^2} \end{aligned} \tag{6.32}$$

or

$$\hat{c}_i\,/T \;\Rightarrow\; \frac{\exp(AD_\mu)-(1+AD_\mu)}{(AD_\mu)^2} \in (0,\infty) \tag{6.33}$$

where $\hat{c}_i$ stands for c evaluated at $\hat{\psi}_{ad2}$, and

$$T(\hat{\psi}_{ad2,i}-1) \;\Rightarrow\; DF_\mu - 2\left[\frac{\exp(AD_\mu)-(1+AD_\mu)}{(AD_\mu)^2}\right](DF_\mu - AD_{\mu,ap})\,. \tag{6.34}$$

The distribution and the corresponding random variable will be referred to as $AD_{\mu,i}$.[24]

The random variable in square brackets or the asymptotic distribution of $\hat{c}_i\,/T$ is more prone to take large values than its noniterated counterpart above (no figure is presented). The probability mass has shifted to the right, and the right tail has been elongated compared to the asymptotic distribution of $\hat{c}\,/T$. The probability of a value within the range $(0,1)$ is still over 0.95, though. Truly extreme values remain unlikely, but unfortunately they emerge, too.[25] The consequence is that $AD_{\mu,i}$ may take excessive values, as well. The distribution $AD_{\mu,i}$ is presented in Figure 6.10. The stretched right

---

$\psi = 1$ — as $\hat{\psi}_{ad2}$ was set to 1 in the first round of the iteration — and define $\hat{\psi}_{ad2,i}$=$\hat{\psi}_{ad2,ap}$. However, the last estimate is consistent only for $\psi = 1$ so the procedure would not be too different from the one adopted here.

[24]If the iterations were done twice then the asymptotic distribution would be found by replacing AD_μ by $AD_{\mu,i}$.

[25]The Monte Carlo asymptotic median of $\hat{c}_i\,/T$ is 0.231. The distribution appears not to have finite moments because extreme values seem to have nonneglible probability.

tail, and consequent transfer towards symmetry, are apparent. The distribution does not seem to feature moments due to some outlying values, so only the Monte Carlo median is reported in the Figure.[26] (The fractiles of $AD_{\mu,i}$ are tabulated in Table 7.3 of Section 7.2.)

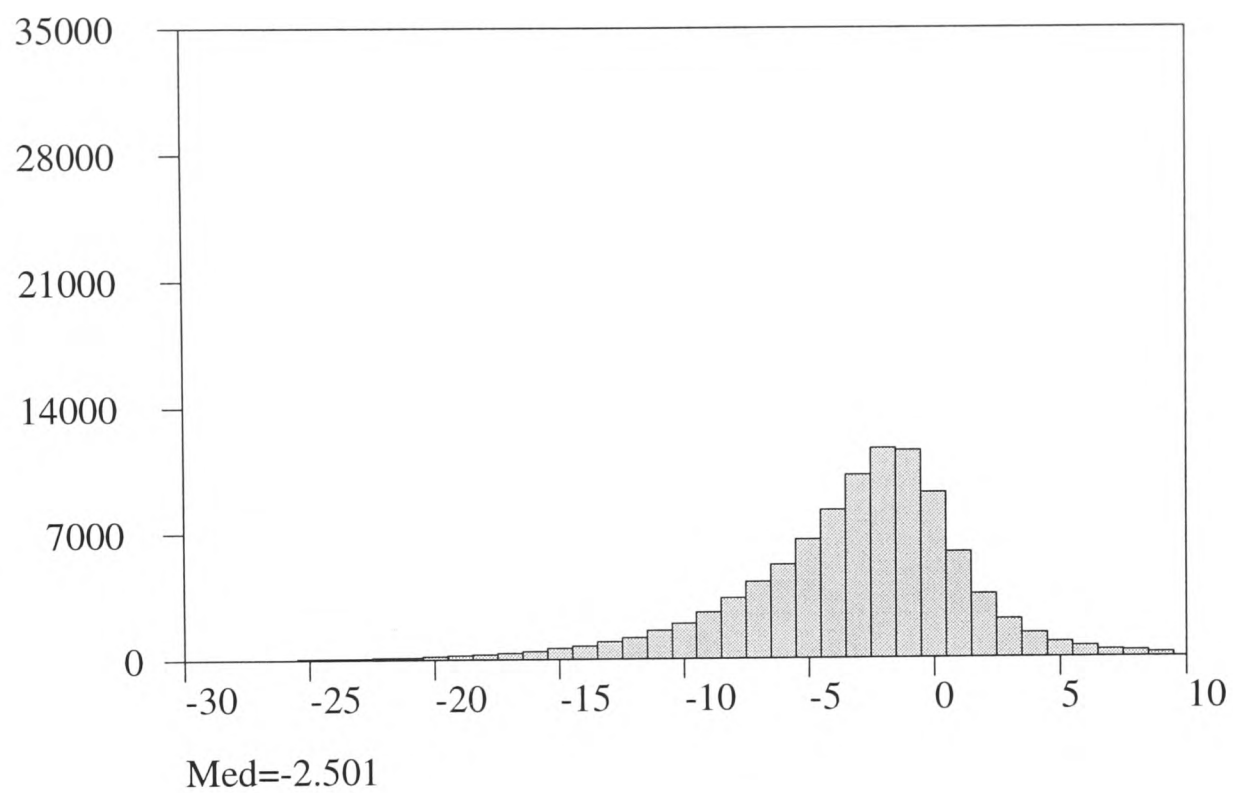


Figure 6.10 Empirical asymptotic distribution of $T(\psi_{ad2,i} - 1)$ for $\psi = 1$ ($\mathrm{AR}_\mu(1)$).

The lack of moments suggests that it might be appropriate to confine the estimate of c to lie in the theoretical range $(0, (T-1)/2]$.[27] However, exaggerated values of $\hat{c}_i$ are likely to arise at the right tail of the distribution of the MLE so they should be inconsequential for unit-root tests where the focus is at the left tail.

Iteration makes sense only, at least in asymptotic terms, if the model is nonstationary and the adjustment does not fade with T. The new asymptotic distribution of the iterated AE or AE_i arises here from the fact that the asymptotic distribution of $T(\hat{\psi}_{ad2} - 1)$ and $\hat{\psi}^T_{ad2}$ (cf. equation (6.32)) is different from the asymptotic distribution of $T(\hat{\psi}_{ml} - 1)$ and $\hat{\psi}^T_{ml}$, respectively. These distributions would agree if the model were stationary and the standardised AE_i would face the same asymptotic distribution as the standardised MLE.

[26] The simulations of Section 7.2 indicate that finite-sample distributions for AE_i possess the first moments. This is interesting because the more typical case in econometrics is that moments do not exist in finite samples but do asymptotically (cf. Maddala (1977, p. 149)).

[27] We have not implemented this idea in the simulation. The resulting estimate could not then exceed the AE_{ap}.

The adjusted Wald statistic

$$(\hat{\psi}_{ad2} - \psi_0)^2 \left[-\frac{\partial^2 l_{ad2}(\psi)}{\partial^2 \psi} \bigg|_{\psi=\hat{\psi}_{ad2}} \right]$$

(cf. equation (4.14)) for the present model is easily found with the above results. The second derivative in the preceding formula can be found from equation (6.16) and substituted above to give (cf. p. 227)

$$W_{ad2} = (\hat{\psi}_{ad2} - \psi_0)^2 \left[\frac{-(T-3)(\psi^2 - \hat{l} - 2\psi\,\hat{\psi}_{ml} + 2\,\hat{\psi}_{ml}^2)}{\left(\hat{l} - 2\psi\,\hat{\psi}_{ml} + \psi^2\right)^2} \bigg|_{\psi=\hat{\psi}_{ad2}} \right]. \tag{6.35}$$

This is the formula to calculate the statistic in practice. It is proved in Appendix A1 that the second derivative of the adjusted profile log-likelihood, evaluated at the AE (as defined in case *iv*) of definition (6.18)), is

$$\frac{\partial^2 l_{ad2}(\psi)}{\partial^2 \psi} \bigg|_{\psi=\hat{\psi}_{ad2}} = \frac{(T-3)\left(\hat{\psi}_{ml}^2 - \hat{l} + r_T^*\right)}{\left[-(\hat{\psi}_{ml}^2 - \hat{l}) + r_T^*\right]^2}, \qquad r_T^* = O_p(T^{-2}) \text{ for } |\psi| < 1 \text{ and } \psi = 1.$$

so

$$\sqrt{W_{ad2}} = (\hat{\psi}_{ad2} - \psi_0) \frac{\sqrt{-(T-3)\left(\hat{\psi}_{ml}^2 - \hat{l} + r_T^*\right)}}{-(\hat{\psi}_{ml}^2 - \hat{l}) + r_T^*}. \tag{6.36}$$

We note that $\hat{\psi}_{ml}^2 - \hat{l}$ is nonpositive (see inequality (A1.1) in Appendix A1) so the square-root term is asymptotically a real number and the ratio positive (with probability one). In principle, the argument of the square-root term may be nonnegative for finite

T, though, and the statistic not calculable (in real terms).[28] We analyse the square root of the Wald statistic or $\sqrt{W_{ad2}}$ because the sign of the deviation $\hat{\psi}_{ad2} - \psi_0$ can now be taken account of and also because the $\hat{\tau}_\mu$ statistic (6.9) studied by Dickey and Fuller (1979) corresponds to $\sqrt{W}$ where W is the ordinary Wald statistic.[29] The asymptotic distribution of $\sqrt{W_{ad2}}$ depends again on the method of evaluating coefficient c and the magnitude of ψ. The stationary circumstance is contemplated again first.

It is proved in Appendix A1 that $\hat{\psi}^2_{ml} - \hat{l}$ is $O_p(1)$ when $\mid \psi \mid < 1$. The second thing to register is that

$$\frac{\sqrt{-T\left(\hat{\psi}^2_{ml} - \hat{l}\right)}}{-\left(\hat{\psi}^2_{ml} - \hat{l}\right)} = \sqrt{\frac{-T}{\hat{\psi}^2_{ml} - \hat{l}}} = \sqrt{\frac{\sum_{t=1}^{T}(y_{t-1} - \bar{y}_{-1})^2}{\hat{\sigma}^2_{ml}}} \geq 0, \tag{6.37}$$

a result already touched on and easily revealed by some algebra. The remainder terms in the formula for $\sqrt{W_{ad2}}$ may be ignored for asymptotic purposes so we find that

$$\sqrt{W_{ad2}} \approx (\hat{\psi}_{ad2} - \psi_0)\sqrt{\frac{\sum_{t=1}^{T}(y_{t-1} - \bar{y}_{-1})^2}{\hat{\sigma}^2_{ml}}}.$$

This is essentially the t-value (6.9) so the asymptotic distribution is Standard Normal. It is noted that formulae (4.4), (6.20), and (6.37) imply that the ordinary Wald statistic can be expressed as a function of $\hat{\psi}_{ml}$ and $\hat{l}$ alone.

The most distinctive result originates again from the assumptions $\psi = \psi_0 = 1$ in combination with exploitation of the *a priori* formula $\hat{c} = (T-1)/2$. Formula (6.36) for

---

[28] We have not come across such an incident in our simulations when the data based formula for $\hat{c}$ is used and a unit root is assumed. In practise the term can be negative under a unit root only if the *a priori* form for $\hat{c}$ is utilized and the sample size is small (25, say). For comparison, the ordinary (square root of the) Wald statistic is $(\hat{\psi}_{ml} - \psi_0)\sqrt{\hat{\sigma}^{-2}_{ml}\sum_{t=1}^{T}(y_{t-1} - \bar{y}_{-1})^2}$. The square-root term of this quantity is always nonnegative.

[29] Literally, the square root of the usual Wald statistic according to formula (4.4) is $(\hat{\psi}_{ml} - \psi_0)\,\hat{\sigma}^{-1}_{ml}\sqrt{\sum_{t=1}^{T}(y_{t-1} - \bar{y}_{-1})^2}$ whereas $\hat{\tau}_\mu$ equals $(\hat{\psi}_{ml} - \psi_0)s^{-1}\sqrt{\sum_{t=1}^{T}(y_{t-1} - \bar{y}_{-1})^2}$.

$\sqrt{W_{ad2}}$ holds even though we multiplied $(\hat{\psi}_{ad2} - \psi_0)$ (for $\psi_0 = 1$) and the denominator by T. We can then reason from convergences (6.25) and (6.26) that

$$\sqrt{W_{ad2,ap}} \Rightarrow AD_{\mu,ap} \cdot \frac{\left\{\int_0^1 [W_*(r)]^2\, dr\right\}^{-1/2}}{\left\{\int_0^1 [W_*(r)]^2\, dr\right\}^{-1}}$$

$$\text{or to} \qquad AD_{\mu,ap} \left\{\int_0^1 [W_*(r)]^2\, dr\right\}^{1/2}$$

or that

$$\sqrt{W_{ad2,ap}} \Rightarrow \frac{\frac{1}{2} W(1) \left[W(1) - 2\int_0^1 W(r)dr\right]}{\sqrt{\int_0^1 [W_*(r)]^2\, dr}} \tag{6.38}$$

where we have indicated utilization of *a priori* information with the subscript *ap*. Recalling the alternative form for the numerator from formula (6.7) and comparison to the DF–τ_μ distribution (6.9) exposes that

$$\sqrt{W_{ad2,ap}} \Rightarrow \frac{\frac{1}{2}\left\{[W(1)]^2 - 1\right\} - W(1)\int_0^1 W(r)dr}{\sqrt{\int_0^1 [W_*(r)]^2\, dr}} - \frac{-1/2}{\sqrt{\int_0^1 [W_*(r)]^2\, dr}}$$

or that the adjustment afresh has the impact of adding a one half to the numerator of the asymptotic distribution of the original statistic. The outcome is depicted in Figure 6.11. The benefits of the adjustment include a shift to the right to zero, symmetrisation, and a notable decrease in the standard deviation compared to the asymptotic distribution of the ordinary Wald statistic (Figure (6.5)). The SD is less than one and is the smallest SD in the thesis. The distribution is platycurtic (sample kurtosis is 1.529, not reported in the figure) and hence is not Normal. On the negative side, the potential serviceability of this statistic suffers from the inconsistency of $\hat{\psi}_{ad2,ap}$ under $|\psi| < 1$. It was suggested on p. 88 that the statistic $T(\hat{\psi}_{ad2,ap} - 1)$ may still exhibit power and we can conclude by a similar argument here that so may $\sqrt{W_{ad2,ap}}$, too. We address the question of power in more detail in Chapter 7. Rothenberg (*op. cit.*) has applied the permutation argument to this statistic, too, and revealed asymptotic unbiasedness and symmetry.[30]

[30] We derived the result independently of Rothenberg but he found the results first. He did not

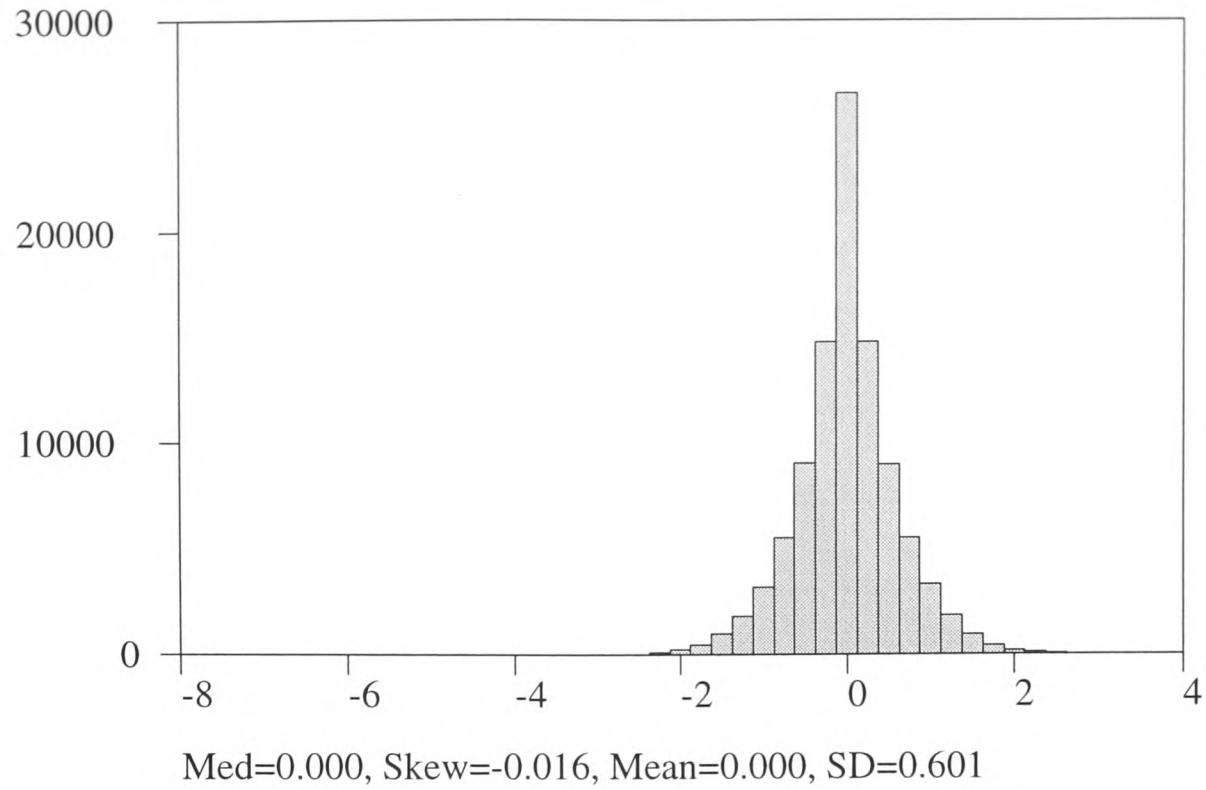


Figure 6.11 Empirical asymptotic distribution of $\sqrt{W_{ad2,ap}}$ for $\psi = 1$ ($\mathrm{AR}_\mu(1)$).

If the data-based formula for estimating c is utilized under $\psi = 1$ then the asymptotic distribution is found by now easily from convergences (6.30) and (6.26):

$$\sqrt{W_{ad2}} \;\Rightarrow\; AD_\mu \left\{ \int_0^1 [W_*(r)]^2 \, dr \right\}^{1/2}$$

or

$$\sqrt{W_{ad2}} \;\Rightarrow\; \left\{ DF_\mu + \left[\frac{\exp(DF_\mu) - (1 + DF_\mu)}{(DF_\mu)^2 \int_0^1 [W_*(r)]^2 \, dr} \right] \right\} \left\{ \int_0^1 [W_*(r)]^2 \, dr \right\}^{1/2}. \qquad (6.39)$$

The histogram of Figure 6.12 illustrates the distribution. This time the advantage of the adjustment is less clear. The mean and the median of the distribution lie closer to zero than for the original distribution (6.9). This might be considered an improvement. However, the right tail of the distribution has elongated which contributes to increased skewness and standard deviation of the variate.

Example. The observations generated for the example of Section 6.2 can be interpreted to have arisen from the process $y_t = \alpha + \psi y_{t-1} + \epsilon_t$ where $\alpha = 0$, $\psi = 1$, $\epsilon_t \sim \mathrm{NID}(0, 1)$ and

---

comment on the other moments, though.

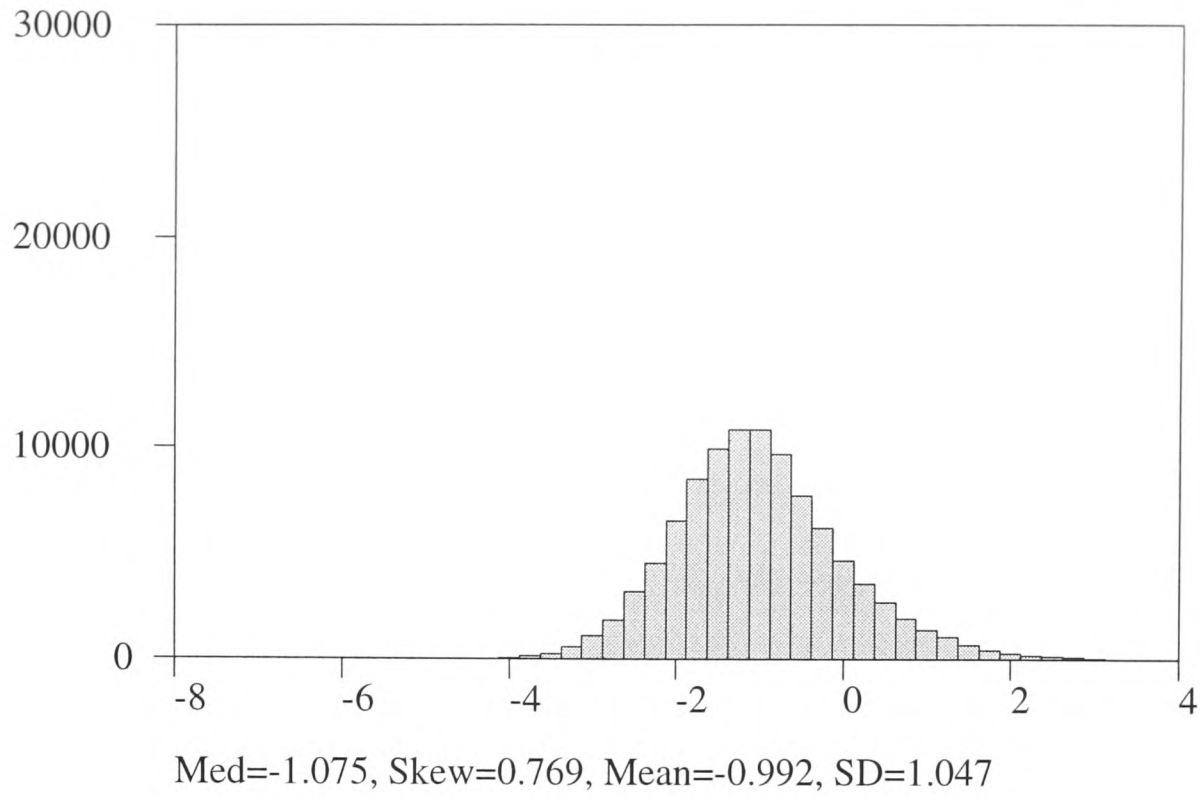


Figure 6.12 Empirical asymptotic distribution of $\sqrt{W_{ad2}}$ for $\psi = 1$ $(\mathrm{AR}_\mu(1))$.

$y_0 = 0$. Estimation of the model is thus legitimate. The estimates are $\hat{\psi}_{ml}$= 0.856 and $\hat{\psi}_{ad2}$= 0.950 for the 25 first observations. The estimates are $\hat{\psi}_{ml}$= 0.987 and $\hat{\psi}_{ad2}$= 1.005 when all the 100 observations are examined. The *a priori* adjusted estimate $\hat{\psi}_{ad2,ap}$ equals 1.109 or 1.013 for the two sample sizes, respectively.

The minima of the adjusted profile likelihoods, relating to the other roots of the adjusted likelihood equation, occur at 5.117 and 3.839, respectively, for the samples. They lie outside the parameter space of ψ or $(-1, 1]$ in accordance with asymptotic theory (Appendix A1).

Coefficient c evaluated at the MLEs equals 5.052 or 33.803 for the 25 and 100 observations, respectively. The estimated cs are smaller than the corresponding theoretical or *a priori* values 12 and 49.5. The revised estimates of c or $\hat{c}_i$s are 8.431 or 58.221, respectively, for the two data spans. The estimate for the smaller sample size is markedly better than the original. The estimate has jumped over the theoretical value with the larger sample size which suggests, as remarked above, that it could be useful to restrict the estimate to lie in the theoretical range $(0, (T-1)/2]$.

Procedure 1) gives the estimates (one-step iteration) 1.020 and 1.018 for the two

sample sizes, respectively. Procedure 2) yields[31] corresponding estimates 1.063 and 1.012 while procedure 3) returns 1.064 and 1.012, respectively. If $\hat{c}$ were confined to the theoretical range, or $\hat{c}$ were defined to equal 49.5 after observing the over-estimate with the larger sample size, then procedure 1) would return $\hat{\psi}_{ad2,ap}$ or 1.013. Procedure 2) would then coincidentally give the same number (one-step iteration).

The MLE was the worst estimate for the shorter span of data though not so bad for the longer span.[32] The AEs were quite good, for example it was the most accurate estimate for the longer span of data. The AE_{ap} did not perform too well especially with the shorter span of data.[33] The results on the iterated estimates are not evident apart from the fact they were overestimates of ψ. For example, procedure 1) gave the best estimate for the shorter span of data but the worst for the longer span — because of the exaggerated $\hat{c}_i$. The MLEs from the simple AR(1) model (cf. Section 6.2) were far better than the present MLEs with constant, but some of the modified estimates drew as, or more, accurate values.□

Quadratic adjusted score equations will appear in the following sections as well. One arose in the time-series application of the Cox–Reid (1987) formula (4.9) in Cruddas *et al.* (1989).[34]

Finally, we note two invariance properties of $\hat{\psi}_{ml}$ and $\hat{\psi}_{ad2}$.[35] First, neither changes if a constant is added to the observations including y_0 (due to a recording practice, say).

---

[31] Actually, we did not run procedure 2) until convergence because that would have taken a lot of time (we did the iterations 'by hand'). The iterative estimates using the 25 data points zigzaged around 1.063 but they seemed to converge towards it.

[32] The measurement criterion used is the absolute or squared deviation from the true value.

[33] The estimate is below the 90^{th} fractile, though, according to unreported fractiles of $\hat{\psi}_{ad2,i}$ for $T = 100$ and $\psi = 1$.

[34] Multiple roots are not rare in time-series analysis. For example, ordinary likelihood equations are cubic or quartic for the autoregressive parameters of stationary AR(1) and AR(2) models, respectively, when the first observations already follow the stationary distribution (White (1961) and Minozzo and Azzalini (1993), respectively).

Relatedly, Zivot (1994) applies Bayesian analysis to the present model. He employs the Jeffreys prior $|I(\omega^*)|^{1/2}$ where $I(\omega^*)$ is the information matrix for $\omega^{*\prime} = [\psi \ \alpha \ \sigma]$ and reports (*op. cit.*, p. 558) that the posterior distribution for ψ may feature two modes.

[35] Kiviet and Phillips (1992) prove in a more general context that invariance (under the null) may often be achieved by extending the regression to include an otherwise redundant regressor. This is exactly what is done here as the constant equals zero under the null of a unit root but a constant is estimated.

The residual variance is

$$\hat{\sigma}^2_{\psi} = T^{-1}\sum_{t=1}^{T}\left[y_t - \bar{y} - \psi(y_{t-1} - \bar{y}_{-1})\right]^2$$

before and

$$\hat{\sigma}^2_{\psi} = T^{-1}\sum_{t=1}^{T}\left\{y_t + a - (\bar{y} + a) - \psi\left[y_{t-1} + a - (\bar{y}_{-1} + a)\right]\right\}^2$$

after adding a constant a to the observations, respectively. Obviously, the as cancel in the latter formula, so $\hat{\sigma}^2_{\psi}$ does not change. It follows that the profile log-likelihood, $\hat{\psi}_{ml}$ and the determinant term of the adjusted profile log-likelihood do not change either. Coefficient $\hat{c}$ depends solely on $\hat{\psi}_{ml}$ with this model, so it remains unmodified as well. Consequently, all the terms of the adjusted likelihood $l_{ad2}(\psi)$, and thus $\hat{\psi}_{ad2}$, persist untouched.

Second, if $\psi = 1$ then a starting value $y_0 \neq 0$ has an analogous effect to the process of adding a constant of equal value after the generation of the time series. The above argument implies that the distribution of $\hat{\psi}_{ml}$ is invariant with respect to the starting value of y_0 if $\psi = 1$. So the assumption $y_0 = 0$ can be replaced by y_0 is a constant when a unit-root test based on $\hat{\psi}_{ml}$ is used.

Invariance of $\hat{\psi}_{ad2}$ with respect to the starting value y_0 under $\psi = 1$ is not apparent as the starting values have some impact on the information measures to which the adjustments hinge. However, it can be shown that of the relevant information measures only $i_{\psi\alpha}$ is altered[36]:

$$i_{\psi\alpha} = \begin{cases} i_{\psi\alpha}\mid_{y_0=0} + \dfrac{1-\psi^T}{\sigma^2(1-\psi)}y_0 & \text{if } \mid\psi\mid < 1 \\ i_{\psi\alpha}\mid_{y_0=0} + \sigma^{-2}Ty_0 & \text{if } \psi = 1 \end{cases}$$

where $i_{\psi\alpha}\mid_{y_0=0}$ stands for formula (6.10) which assumed $y_0 = 0$.[37] Coefficient c is

[36]The (here irrelevant) measure for ψ under a unit root is documented in the introduction to Appendix A5.

[37]The expression for $\sum_{t=0}^{T-1} E(y_{t-1})$ in the stationary case in the Appendix of Guilkey and Schmidt

$$c = \frac{\partial}{\partial \alpha}\left\{ i_{\alpha\alpha}^{-1}\left[i_{\psi\alpha} \mid_{y_0=0} + \frac{1-\psi^T}{\sigma^2(1-\psi)} y_0 \right] \right\} = \left(\frac{\partial}{\partial \alpha} i_{\alpha\alpha}^{-1} i_{\psi\alpha} \right) \mid_{y_0=0}$$

by equation (6.12), the above formula for $i_{\psi\alpha}$ and the fact that $i_{\alpha\alpha}$ does not depend on α (equation (6.13)). The outcome is that coefficient c needs no modification. It follows that $\hat{\psi}_{ad2}$ is nailed to an invariant value regardless of the value the process started from if $\psi = 1$ i.e. the assumption $y_0 = 0$ can be relaxed to y_0 is any constant when using $\hat{\psi}_{ad2}$ as a core of a unit-root test. Invariance of the Wald statistics $\sqrt{W_{ad2}}$, $\sqrt{W_{ad2}}$, and $\sqrt{W_{ad2,ap}}$ follows from the above statements.

6.4 Adjusted Profile Likelihood for a Unit-Root AR(1) Model with Drift

The model we shall study next is a unit-root AR(1) model with drift:

$$y_t = \alpha + \psi y_{t-1} + \epsilon_t \tag{6.40}$$

where $\alpha \neq 0$, $\psi = 1$, $\epsilon_t \sim$NID$(0, \sigma^2)$, $\sigma^2 > 0$, and $t = 1,\ldots,T$. The difference to the model of the previous section is that only the unit-root case is allowed for and that the constant is assumed to be different from zero even though $\psi = 1$. As before, the parameter of interest is ψ while the nuisance parameters are α and σ^2. The model will be referred to as the unit-root AR$_\mu$(1) with drift.

The variability of y_t is asymptotically dominated by the drift. The model appears to be a reasonable description of many economic time series (cf. Hylleberg and Mizon (1989), say).

The asymptotic properties of the MLE of ψ, and of coefficient c evaluated at the MLE, are drastically different from the case $\alpha = 0$ and $\psi = 1$ considered in Section 6.3. This is the reason to devote a section of its own for the process. The derivations and the

(1991) agrees with the above expression confirming our derivation.

quantities of interest for the model — apart from the asymptotic expressions — agree exactly with those presented in Section 6.3 so they are not repeated. We shall only give a few additional comments on the formulae. Asymptotics will be focused, instead.

Dickey (1976, pp. 58–61) was a forerunner in proving that

$$T^{3/2}(\hat{\psi}_{ml} - 1) \;\Rightarrow\; N(0, 12\sigma^2/\alpha^2) \tag{6.41}$$

under the present assumptions (Dickey and Fuller (1979, p. 429) refer to the fact, too). The result was generalised and publicised by West (1988), see also Evans and Savin (1984). The deterministic drift outweighs the fluctuation due to the unit root in y_t which brings the asymptotic distribution back to the Normal family. The drift induces also the faster rate of convergence compared to the driftless case. The variance of $T^{3/2}(\hat{\psi}_{ml} - 1)$ explodes as α tends to zero which is in line with the fact that T is the appropriate multiplying factor when $\alpha = 0$. Dickey (*op. cit.*) proved also that

$$\hat{\tau}_{\mu} \;=\; \frac{\hat{\psi}_{ml} - 1}{s/\sqrt{\sum_{t=1}^{T}(y_{t-1} - \bar{y}_{-1})^2}} \;\Rightarrow\; N(0,1),$$

or that the t-statistic for the null of a unit root follows asymptotically the Standard Normal distribution.

Improved accuracy in the estimation of ψ implies more accurate estimates of c, too. Precision in the estimation of c is a reason which intrigues us to study the Cox–Reid adjustment under the present model. Inaccuracy in the estimation of c made the adjustment less effective in the $\alpha = 0$, $\psi = 1$ situation (cf. the previous section).

The counterpart to formula (6.19) is here

$$\hat{\psi}_{ad2} \;=\; \hat{\psi}_{ml} - \frac{\hat{c}\,(\hat{\psi}_{ml}^2 - \hat{l})}{T - 3} + r_T, \qquad r_T = O_p(T^{-3}), \tag{6.42}$$

so the accuracy of the approximation ameliorates compared to the driftless case (Appendix A2 supplies the proofs for the asymptotics). It follows that

$$T^{3/2}(\hat{\psi}_{ad2} - 1) = T^{3/2}(\hat{\psi}_{ml} - 1) - T^{3/2}\frac{\hat{c}\,(\hat{\psi}_{ml}^2 - \hat{l})}{T-3} + r_T^*, \qquad r_T^* = O_p(T^{-3/2}) \tag{6.43}$$

where the order of magnitude of the term $T^{3/2}\,\hat{c}\,(\hat{\psi}_{ml}^2 - \hat{l})/(T-3)$ determines whether the MLE is adjusted asymptotically or not.

It is shown in Appendix A2 that

$$\hat{c}\,/T \underset{p}{\longrightarrow} \frac{1}{2}, \tag{6.44}$$

or coefficient c evaluated at $\hat{\psi}_{ml}$ and divided by T tends stochastically to the theoretical limit of c/T. The result is distinct from the case in Section 6.3 where $\hat{c}\,/T$ tended to a random variable instead of a constant under a unit root. Our expectations of improved accuracy in the estimation of c are hence fulfilled. On another intuitive point of view, the convergence to the constant $1/2$ can be attributed to the fact that information measures evaluated at the MLE, after standardisation, tend here stochastically to constants. This is because the drift, as opposed to the nonstationarity due to the unit-root, dominates the asymptotic behaviour of the process. A consequence is that the asymptotics are the same under the present model whether $\hat{c}$ is estimated or defined to equal the *a priori* theoretical value $(T-1)/2$ (which utilizes the information $\psi = 1$).

It is proved in Appendix A2 that the order of magnitude of $\hat{\psi}_{ml}^2 - \hat{l}$ is $O_p(T^{-2})$. It follows from formulae (6.43) and (6.44) that the adjustment to the standardised MLE is $O_p(T^{-1/2})$ or that the adjustment fades asymptotically. In other words, the standardised AE converges weakly to the same Normal distribution as the standardised MLE:

$$T^{3/2}(\hat{\psi}_{ad2} - 1) \Rightarrow N(0, 12\sigma^2/\alpha^2). \tag{6.45}$$

Precision in the estimation of c is hence not beneficial in asymptotic terms, to first order, in the present context as the asymptotic distributions of the MLE and the AE

agree. There is no asymptotic bias to remove but neither does the Cox–Reid adjustment operate on any other property of the asymptotic distribution. The adjustment can still be advantageous in small samples by reducing the bias or the gap between the small-sample distributions and the asymptotic distribution among other things. We shall return to this issue below and in Section 7.4.

We note a further asymptotic result which is of interest of its own. Paralleling the analysis in Section 6.3, the right-hand side of equation (6.43) can be expressed as

$$\begin{aligned} & T^{3/2}(\hat{\psi}_{ml} - 1) - \frac{\hat{c}}{T-3}(\hat{\psi}_{ml} + 1)T^{3/2}(\hat{\psi}_{ml} - 1) + \frac{\hat{c}}{T-3}T^{3/2}(\hat{l} - 1) + r_T^* \\ & = T^{3/2}(\hat{\psi}_{ml} - 1)\left[1 - \frac{\hat{c}}{T-3}(\hat{\psi}_{ml} + 1)\right] + \frac{\hat{c}}{T-3}T^{3/2}(\hat{l} - 1) + r_T^*. \end{aligned}$$

The square-bracket term tends stochastically to $1 - (1/2) \cdot 2 = 0$ by convergence (6.44) and convergence of $\hat{\psi}_{ml}$ to 1. This implies that

$$\frac{\hat{c}}{T-3}T^{3/2}(\hat{l} - 1) \Rightarrow N(0, 12\sigma^2/\alpha^2),$$

i.e. $\left[\hat{c}/(T-3)\right] \cdot T^{3/2}(\hat{l} - 1)$ follows asymptotically the same distribution to which $T^{3/2}(\hat{\psi}_{ml} - 1)$ tends. We prove explicitly in Appendix A2 that $T^{3/2}(\hat{l} - 1) \Rightarrow N(0, 48\sigma^2/\alpha^2)$, from which the result follows after recalling convergence (6.44). Furthermore, we note that $T^{-3/2}$ times the numerator of $(\hat{l} - 1)$ converges weakly to $N(0, \alpha^2\sigma^2/3)$ which can be expressed as $\sigma\alpha\left[W(1) - 2\int_0^1 W(r)dr\right]$ (Appendix A2). The square bracket term has appeared previously in the numerator of the asymptotic distribution of $T(\hat{l} - 1)$ under $\alpha = 0$ and $\psi = 1$, cf. formula (6.24). Here the term is multiplied by $\sigma\alpha$ whereas in Section 6.3 the term was multiplied by $\sigma W(1)$ (σ does not appear in formula (6.24) as it cancels with σ in the denominator).

It was remarked in Section 6.3 that coefficient c does not depend explicitly on the constant α. It might be argued that this appears somewhat counterintuitive if $\psi = 1$. The asymptotic distribution of $T(\hat{\psi}_{ml} - 1)$ depends crucially on the size of α under

a unit root: the DF_μ distribution (6.8) with a negative mean results if $\alpha = 0$ but a Normal distribution with no bias involved arises if $\alpha \neq 0$, as can be seen from equation (6.41). Furthermore, a sizable α squeezes also the small-sample distribution towards the Normal distribution (Evans and Savin (1984) and Hylleberg and Mizon (1989)). The Cox–Reid adjustment does not take these features explicitly into account. (It may be worth repeating that the assumption $\alpha = 0$ if $\psi = 1$ of Section 6.3 was not made use of when deriving c.) However, the asymptotic behaviour of the Cox–Reid adjustment makes sense: the bias fades asymptotically for all values of α apart from zero (formula (6.41)) which is the sole circumstance for which the adjustment persists.

Similar derivations as in Section 6.3 prove that the Wald statistic calculated from the adjusted profile log-likelihood may be expressed as:

$$\sqrt{W_{ad2}} = (\hat{\psi}_{ad2} - 1)\frac{\sqrt{-(T-3)\left(\hat{\psi}_{ml}^2 - \hat{l} + r_T^*\right)}}{-(\hat{\psi}_{ml}^2 - \hat{l}) + r_T^*}, \tag{6.46}$$

$$= T^{3/2}(\hat{\psi}_{ad2} - 1)\frac{\sqrt{-T(T-3)\left(\hat{\psi}_{ml}^2 - \hat{l} + r_T^*\right)}}{T^2\left[-(\hat{\psi}_{ml}^2 - \hat{l}) + r_T^*\right]}, \qquad r_T^* = O_p(T^{-3}).$$

(The approximation accuracy can be reasoned with the help of Appendix A2.) It is shown in Appendix A2 that $T^2\left(\hat{\psi}_{ml}^2 - \hat{l}\right) \underset{p}{\longrightarrow} -12\sigma^2/\alpha^2$. The standardised coefficient (SC), which is asymptotically a $N(0, 12\sigma^2/\alpha^2)$ variate, becomes asymptotically multiplied by a factor converging to $(12\sigma^2/\alpha^2)^{-1/2}$ so the limiting distribution of $\sqrt{W_{ad2}}$ is obtained to be:

$$\sqrt{W_{ad2}} \Rightarrow N(0,1).$$

The limiting distribution of the adjusted Wald statistic is the same, Standard Normal, as

that of the conventional Wald statistic (or of the closely related t-statistic). Employing the *a priori* information of a unit root would not change the asymptotic distribution.

The previous results are of theoretical interest but the practical implications are not evident. We are not suggesting that one should conduct inference on unit roots assuming asymptotic normality even though the null hypothesis would match with model (6.40) and such a practice has been recommended.[38] Apart from the complication that the asymptotic distribution of $T(\hat{\psi}_{ml} - 1)$ and $T(\hat{\psi}_{ad2} - 1)$ depends on the nuisance parameter α, there are two more fundamental hindrances:

i) Normality does not apply if the drift α or the sample size is not fairly large. Hylleberg and Mizon (1989) and Banerjee *et al.* (1993, pp. 170–2) present simulation evidence on the normality of the 't-value' of $\hat{\psi}_{ml} - 1$. In both sources it is suggested that α/σ should be about unity for normality to apply in samples of typical length in economics. Banerjee *et al.* (cf. *op. cit.*, p. 171) argue that this may be a realistic value but Hylleberg and Mizon (*op. cit.*) found that the ratio is far smaller for the economic series they studied so the conditions for normality to hold do not apply in general. The extent to which the Cox–Reid adjustment speeds convergence to the asymptotic distribution is scrutinized in Section 7.4.

ii) Even though the conditions did hold, the result may be of little value from the point of view of testing for unit roots. The natural (implicit) alternative to the null of a drifted unit-root process is a trend-stationary process. Perron (1988) and West (1987) have shown that the ρ_μ (based on the statistic $T(\hat{\psi}_{ml} - 1)$) and τ_μ tests (Section 3.2) are inconsistent if the process is (6.40) or the process features a drift (see also Table 2, Experiment 3 in Schmidt and Phillips (1992) and Table VI in Evans and Savin (1984)). A fully stationary alternative is not a proper (implicit) alternative either as it can hardly

[38] We refer here to Dolado *et al.* (1990). Literally, the authors advise (p. 255) to base inference on the 't-value' of $\hat{\psi}_{ml} - 1$. The asymptotic distribution of the t-value is Standard Normal and does not hence depend on the drift α. The small-sample distributions of the t-value should be more symmetric than that of $T(\hat{\psi}_{ml} - 1)$ as the finite sample distributions of the former are much more skewed than those of the latter under $\alpha = 0$ and $\psi = 1$, cf. Banerjee *et al.* (1993, Tables 4.1. and 4.2), Figures 6.5 and 6.4, and Section 7.4.

be regarded as a substitute of a drifted unit-root process as a description for an empirical time series.

We can imagine one special case when the result would be useful in testing for unit roots. It is when a sensible alternative is an asymptotically stationary process which has started from a starting value very far away from the unconditional mean but a trend-stationary process is out of question. Such a process, *gradually stationary process* from now on, can trace a graph with typically nonlinear trend and could be sometimes considered a reasonable alternative to a drifted random walk. We shall argue in Chapter 7 that the test is consistent against such an alternative as the starting value diverges to infinity from the unconditional mean.

6.5 Adjusted Profile Likelihood for the Bhargava AR(1) Model with Constant

The model favored by Bhargava (1986) exploits an alternative manner of adding a constant to the AR(1) model:

$$\begin{aligned} y_t &= \gamma_0 + x_t \\ x_t &= \psi x_{t-1} + \epsilon_t \end{aligned}$$

or

$$y_t = \begin{cases} \gamma_0 + \psi x_0 + \epsilon_1 & \text{for } t = 1 \\ \gamma_0(1-\psi) + \psi y_{t-1} + \epsilon_t & \text{for } t = 2, \ldots, T. \end{cases} \tag{6.47}$$

Here $\psi \in (-1, 1]$, $\epsilon_t \sim$NID$(0, \sigma^2)$, $\sigma^2 > 0$, and $x_0 = 0$. The nuisance parameters will be denoted by $\boldsymbol{\phi}' = [\gamma_0 \ \sigma^2]$. We shall often refer to (6.47) as the Bhargava AR$_\mu$(1) or the AR$_\mu^B$(1) model. The model or close relatives of it have gained popularity recently (cf. Ahn (1993), Andrews (1993), Elliot *et al.* (1992), Pantula *et al.* (1994), Schmidt and Phillips (1992), and Zivot (1994)).

Handling of the first observation differentiates the model from that of the previous sections. The outcome is that the present time series is composed as follows:

$$y_t = \begin{cases} \psi^t x_0 + \gamma_0 + \sum_{i=1}^{t} \psi^{t-i} \epsilon_i & \text{if } \mid \psi \mid < 1 \\ x_0 + \gamma_0 + \sum_{i=1}^{t} \epsilon_i & \text{if } \psi = 1 \end{cases} \tag{6.48}$$

which can be compared with the corresponding formula (6.11) for the $\mathrm{AR}_\mu(1)$ model of the previous section. Here the process is assumed to start from the neighbourhood of its unconditional mean (γ_0, assuming that σ^2 is not large relative to γ_0 and $x_0 = 0$). The assumption is evident if the process is nonstationary ($\psi = 1$ in which case the process shrinks to a random walk which has started from γ_0). It is quite often made also when the inspected time series has been in operation for some time and the process is known to be stable (i.e. $\mid \psi \mid < 1$). The variance of y_t is affected by start-up effects in this case. The model is especially reasonable when the time series truely originates at $t = 1$ as opposed to that being the moment to begin recording.

The y_t processes models (6.47) and (6.4) generate agree in two circumstances (for $t = 1, 2, \ldots$):

- Under a unit root and zero constant (α) for the $\mathrm{AR}_\mu(1)$ model. Then γ_0 should be understood to account for y_0 of the $\mathrm{AR}_\mu(1)$ model.
- Under stationarity when the autoregressive coefficients are equal, the processes share a common unconditional mean, and the time series generated by the $\mathrm{AR}_\mu(1)$ model starts from the unconditional mean or $\gamma_0 = \alpha/(1-\psi) = y_0$.

If x_0 different from zero were allowed for then the processes would harmonise in the following two cases (for $t = 1, 2, \ldots$):

- Under a unit root and zero constant (α) for the $\mathrm{AR}_\mu(1)$ model. Then $x_0 + \gamma_0$ should be understood to account for y_0 of the $\mathrm{AR}_\mu(1)$ model.
- Under stationarity when the autoregressive coefficients are equal and the processes share a common unconditional mean $\gamma_0 = \alpha/(1-\psi)$ to which the starting values relate so that $x_0 = y_0 - \alpha/(1-\psi)$. A special case emerges when the unconditional

means equal zero and the starting values each other or $\gamma_0 = \alpha/(1-\psi) = 0$ and $x_0 = y_0$.

We shall carry on with the assumption $x_0 = 0$. Empirically x_0 is likely to be unobserved and estimation of it difficult (as reasoned by Schmidt and Phillips (*op. cit.*)).

The model lies then in a sense between the $AR_\mu(1)$ model and the fully stationary AR(1) with constant model where the first observation already follows the stationary distribution. A common feature with the former is the nonconstancy of the variance of y_t (due to start-up effects) and the constancy of the mean with the latter.[39] If $\gamma_0 = \alpha/(1-\psi)$ and $|\psi| < 1$ then the differences between the processes fade for large T. As a whole the model can be comprehended to be an AR(1) process with constant $\gamma_0(1-\psi)$ with start-up effects altering the variance. Furthermore, the model shrinks to the simple AR(1) (equation (6.1)) if the constant γ_0 equals zero. As will be shown below, the present and the last mentioned model have much in common in terms of asymptotics, too.

An important advantage of this model over that of the previous section is that if $\psi = 1$ then γ_0 is nothing but the starting value of the process. This means that asymptotically estimation of ψ is not distorted by estimation of a (in a sense redundant) constant as explained below.

The first observation receives a separate treatment in the log-likelihood function:

$$l(\psi, \gamma_0, \sigma^2; \mathbf{y}) = -\frac{T}{2}\log\sigma^2 - \frac{1}{2\sigma^2}\left\{(y_1 - \gamma_0)^2 + \sum_{t=2}^{T}\left[y_t - \gamma_0(1-\psi) - \psi y_{t-1}\right]^2\right\}$$

The profile log-likelihood has the standard appearance at the outset:

[39] One can argue further that the Bhargava $AR_\mu(1)$ model shares more characteristics with the $AR_\mu(1)$ model than with the stationary model: The stationary model cannot accommodate a ψ equal to one (the reason it is not serviceable here). The variance of y_t is the same for the $AR_\mu(1)$ model and the present model. (The variance is $\sigma^2(1-\psi^{2t})/(1-\psi^2)$ or $\sigma^2 t$ if $\psi = 1$). The model shrinks to the simple AR(1) (equation (6.1)), which is a special case of the $AR_\mu(1)$ model, if $\gamma_0 = 0$. If $x_0 \neq 0$ is allowed for in model (6.47) then the mean of y_t is affected by start-up effects and the common feature of the present and of the stationary model vanishes. The present model would be just a reparameterisation of the $AR_\mu(1)$ model if it were taken as defined in (6.47) for observations $t = 2, 3, ..., T$ and y_1 were regarded as a given constant (cf. Fuller (1976, Section 8.1)).

$$l_p(\psi) = -\frac{T}{2}\log\hat{\sigma}^2_\psi \tag{6.49}$$

where

$$\begin{aligned}\hat{\sigma}^2_\psi &= T^{-1}\left\{(y_1-\hat{\gamma}_{0\psi})^2+\textstyle\sum_{t=2}^{T}\left[y_t-\hat{\gamma}_{0\psi}\,(1-\psi)-\psi y_{t-1}\right]^2\right\}\\ &= T^{-1}\left\{\ \textstyle\sum_{t=1}^{T}y_t^2-2\psi\sum_{t=2}^{T}y_ty_{t-1}+\psi^2\sum_{t=2}^{T}y_{t-1}^2\right.\\ &\quad \left.-\frac{\left[y_1+(1-\psi)\left(\sum_{t=2}^{T}y_t-\psi\sum_{t=2}^{T}y_{t-1}\right)\right]^2}{1+(T-1)(1-\psi)^2}\ \right\}\end{aligned} \tag{6.50}$$

and

$$\hat{\gamma}_{0\psi} = \frac{y_1+(1-\psi)\sum_{t=2}^{T}(y_t-\psi y_{t-1})}{1+(T-1)(1-\psi)^2}. \tag{6.51}$$

The formula for $\hat{\sigma}^2_\psi$ reveals the particular handling of the first observation. In the unit-root case, it alone carries information on γ_0. Indeed $\hat{\gamma}_{0\psi}$ equals y_1 when evaluated at $\psi = 1$. The constant becomes then a kind of dummy and the model has some resemblance to the Neyman–Scott model of Section 4.1 in the sense that a sole observation (the first) carries all of the information on a nuisance parameter.[40,41]

The likelihood equation for ψ

$$\frac{\partial l_p(\psi)}{\partial\psi} = \frac{-T}{2\,\hat{\sigma}^2_\psi}\frac{\partial\,\hat{\sigma}^2_\psi}{\partial\psi} = 0 \tag{6.52}$$

turns out to be equivalent to a fifth order polynomial in ψ which means that analytical explicit solutions for $\hat{\psi}_{ml}$ do not necessarily exist. (The precise formula for $\partial\,\hat{\sigma}^2_\psi\,/\partial\psi$ is given in Appendix A3.) The MLE can still be found by numerical maximisation of likelihood (6.49) or by solving the likelihood equation numerically, say (cf. equation

[40]The number 1 in formula (6.55) arises from the first observation while the term $(T-1)(1-\psi)^2$ ensues from the rest of the observations. Hence the first observation remains the most informative on γ_0 as long as $\psi \neq 0$. If $\psi = 0$ then all of the observations contribute equally to the information.

[41]Formula 6.51 is the usual one for the MLE of γ_0 conditional on a fixed value of y_1 and given ψ (by the general properties of MLEs and section (6.3) or Fuller (1976, section 8.1)).

(6.57)). The likelihood equation has always at least one real root as the order of the polynomial is odd.

The MLE for a known γ_0 is

$$\hat{\psi}_{\gamma_0} = \frac{\sum_{t=2}^{T}(y_t - \gamma_0)(y_{t-1} - \gamma_0)}{\sum_{t=2}^{T}(y_{t-1} - \gamma_0)^2}.$$

It is easy to show that $\hat{\gamma}_{0\psi} \approx (\bar{y} - \psi\ \bar{y}_{-1})/(1-\psi) \approx \bar{y}$ for $\psi \neq 1$ and large T. It follows that in the stationary case, $\hat{\psi}_{ml}$ will asymptotically agree numerically and share the distribution $N(0, 1-\psi^2)$ with the MLE of ψ of the previous section (cf. formula (6.6)). We prove in Appendix A3 that $\hat{\gamma}_{0_{ml}}$ tends in probability to y_1 or is $O_p(1)$ when $\psi = 1$. The $O_p(1)$ness of $\hat{\gamma}_{0_{ml}}$ and the well-known orders of magnitude of expressions like $\sum_{t=2}^{T} y_{t-1}$ suffice for proving that:

$$\begin{aligned}
T(\hat{\psi}_{ml} - 1) &= T\left[\frac{\sum_{t=2}^{T}(y_t - \hat{\gamma}_{0_{ml}})(y_{t-1} - \hat{\gamma}_{0_{ml}})}{\sum_{t=2}^{T}(y_{t-1} - \hat{\gamma}_{0_{ml}})^2} - 1\right] \\
&= T\left[\frac{\sum_{t=2}^{T} y_t y_{t-1} - \sum_{t=2}^{T} y_{t-1}^2 - \hat{\gamma}_{0_{ml}}(y_T - y_1)}{\sum_{t=2}^{T} y_{t-1}^2 - 2\,\hat{\gamma}_{0_{ml}} \sum_{t=2}^{T} y_{t-1} + (T-1)\,\hat{\gamma}_{0_{ml}}^2}\right] \qquad (6.53) \\
&= \frac{T^{-1}\sum_{t=2}^{T} y_{t-1}\epsilon_t}{T^{-2}\sum_{t=2}^{T} y_{t-1}^2 - O_p(T^{-1/2}) + O_p(T^{-1})} - O_p(T^{-1/2}) \\
&\Rightarrow \frac{\int_0^1 W(r)dW(r)}{\int_0^1 [W(r)]^2\,dr}.
\end{aligned}$$

Most interestingly, the Bhargava model accompanied with the assumption $\psi = 1$ yields a MLE with the same asymptotic distribution as the MLE of ψ for the simple AR(1) model (equation (3.11)). Heuristically, the MLE picks $\psi = 1$ and $\gamma_0(1-\psi) = 0$ or the random-walk model as T increases. This is in sharp contrast with the model of the previous section where the estimated constant remained a nuisance even asymptotically. Consequently, the present $\hat{\psi}_{ml}$ follows asymptotically the tighter Dickey–Fuller distribution (3.11) than the MLE of ψ for the model of Section 6.3.[42] A conjecture is that the bias and variance of

[42] In other words, the asymptotic distribution of $\hat{\psi}_{ml}$ is free of the nuisance parameter γ_0 under $\psi = 1$.

the present $\hat{\psi}_{ml}$ remain smaller for a finite but reasonable T, too. Another special feature of the model is that the asymptotic distributions of the appropriately standardised MLEs of ψ arise similar to those of the simple AR(1) model of Section 6.2 whether the model is stationary or not. It does not yet follow that the MLEs from this model and the simple AR(1) model could not take very different values (cf. p. 178).

Parallel results can be derived for the (square root of the) Wald statistic

$$\sqrt{W} = (\hat{\psi}_{ml} - \psi)\sqrt{-\frac{\partial^2 l_p(\psi)}{\partial^2 \psi}\bigg|_{\psi=\hat{\psi}_{ml}}}$$

(by equation (4.4)). We cannot present an explicit analytic form for the Wald statistic as the explicit form of the MLE is unknown. However, the numerical value of the statistic, for a specific sample, can be straightforwardly exposed after substituting the numerical value of the MLE to the formulae defining the statistic, see Appendix A3 for details.

The asymptotic distribution of the statistic depends, of course, on whether the process is stationary or not. In the stationary circumstance, the appropriate standardisation is

$$T^{1/2}(\hat{\psi}_{ml} - \psi)\sqrt{-T^{-1}\frac{\partial^2 l_p(\psi)}{\partial^2 \psi}\bigg|_{\psi=\hat{\psi}_{ml}}}$$

because the MLE is $O_p(T^{-1/2})$. The essentials of the derivations for the stationary case are laid out in Appendix A3 because the model is not as familiar as the other contemplated models. It is shown there that

$$\sqrt{W} \;\Rightarrow\; N(0,1)$$

or that $\sqrt{W}$ follows the Standard Normal when T is large.

When a unit root exists, the appropriate standardisation is

This follows from the facts that y_t is a random walk and the starting value γ_0 can have only a transient effect on the distribution of $\hat{\psi}_{ml}$. The section concludes by proving that the property holds for finite T, too.

It is perhaps worth pointing out that the asymptotic unit-root distribution (3.11) applies to the MLE but the asymptotic unit-root distribution of the demeaned OLS estimate, which appears in Nabeya and Tanaka (1990), say, is (6.8) of Section 6.3.

$$T(\hat{\psi}_{ml} - 1)\sqrt{-T^{-2}\frac{\partial^2 l_p(\psi)}{\partial^2 \psi}\bigg|_{\psi=\hat{\psi}_{ml}}},$$

and it is shown in Appendix A3 that

$$\sqrt{W} \;\Rightarrow\; \frac{\int_0^1 W(r)dW(r)}{\sqrt{\int_0^1 [W(r)]^2\,dr}}.$$

This is distribution (3.12) which we have already encountered when the model is the simple AR(1) scrutinized in Sections 3.2 and 6.2.

It is quite remarkable that the asymptotic distributions of the MLE and the Wald statistic stick to those induced by the simple AR(1) model regardless of the order of integration of y_t (0 or 1). Estimation of a constant distorted quite severely the distributions under the $\mathrm{AR}_\mu(1)$ model with a unit root.

We continue with a derivation of the adjusted profile likelihood. Let us remark first that σ^2 is again orthogonal with respect to the parameters of the model:

$$i_{\psi\sigma^2} = E(j_{\psi\sigma^2}) = E\left\{\sigma^{-4}\sum_{t=2}^{T}\left[y_t - \gamma_0(1-\psi) - \psi y_{t-1}\right](y_{t-1} - \gamma_0)\right\} = 0$$

and

$$i_{\gamma_0\sigma^2} = E(j_{\gamma_0\sigma^2}) = E\left\{\sigma^{-4}\left\{y_1 - \gamma_0 + (1-\psi)\sum_{t=2}^{T}\left[y_t - \gamma_0(1-\psi) - \psi y_{t-1}\right]\right\}\right\} = 0.$$

Advantageously, also the constant γ_0 is orthogonal to ψ in this model:

$$\begin{aligned} i_{\psi\gamma_0} &= E(j_{\psi\gamma_0}) \\ &= E\Big\{\sigma^{-2}\textstyle\sum_{t=2}^{T}(y_t - \gamma_0) + \sum_{t=2}^{T}(y_{t-1} - \gamma_0) \\ &\qquad -2\psi\textstyle\sum_{t=2}^{T}(y_{t-1} - \gamma_0)\Big]\Big\} \\ &= 0. \end{aligned} \tag{6.54}$$

The last two results become obvious after observing that $E(y_t - \gamma_0) = 0$ for all t. The Cox–Reid (1987) formula is applicable as ψ is orthogonal with respect to the nuisance parameters. We go on to calculate

$$\begin{aligned} i_{\gamma_0\gamma_0} &= E(j_{\gamma_0\gamma_0}) = E\left\{\sigma^{-2}\left[1+(T-1)(1-\psi)^2\right]\right\} \\ &= \sigma^{-2}\left[1+(T-1)(1-\psi)^2\right]. \end{aligned} \tag{6.55}$$

The expected and observed information measures coincide here. The last required information measure is an observed one:

$$j_{\sigma^2\sigma^2} = \tfrac{-T}{2\sigma^4} + \sigma^{-6}\left\{(y_1-\gamma_0)^2 + \textstyle\sum_{t=2}^{T}\left[y_t - \gamma_0(1-\psi) - \psi y_{t-1}\right]^2\right\}.$$

The determinant term of the formula is

$$-\tfrac{1}{2}\log \mid \mathbf{J}_{\phi\phi}(\psi, \hat{\phi}_\psi) \mid = \tfrac{3}{2}\log\hat{\sigma}^2_\psi + \tfrac{1}{2}\log 2 - \tfrac{1}{2}\log T - \tfrac{1}{2}\log\left[1+(T-1)(1-\psi)^2\right]$$

and the adjusted profile log-likelihood is

$$\begin{aligned} l_{ad}(\psi) &= -\tfrac{T}{2}\log\hat{\sigma}^2_\psi \\ &\quad + \tfrac{3}{2}\log\hat{\sigma}^2_\psi + \tfrac{1}{2}\log 2 - \tfrac{1}{2}\log T - \tfrac{1}{2}\log\left[1+(T-1)(1-\psi)^2\right] \\ &\propto -\frac{(T-3)}{2}\log\hat{\sigma}^2_\psi - \frac{1}{2}\log\left[1+(T-1)(1-\psi)^2\right]. \end{aligned} \tag{6.56}$$

The term $-\frac{1}{2}\log\left[1+(T-1)(1-\psi)^2\right]$ pushes the AE to the right of the MLE on the real axis if $\hat{\psi}_{ml} < 1$ but shrinks the estimate downwards if $\hat{\psi}_{ml} > 1$. The term is nonpositive and penalizes large values of $(1-\psi)^2$ or small values of ψ when the parameter space is restricted to $(-1, 1]$. Arrestingly, the term fades as ψ approaches 1 and vanishes completely at ψ equal to 1. It is also interesting to note that values of ψ are penalized as a simple function of the amount of information of γ_0 (equation (6.55)). It was explained in

Section 4.2 that the Cox–Reid adjustment penalizes values of ψ which provide relatively more information on the nuisance parameters. The phenomenon is quite explicit here: a ψ equal to one would minimize the information on γ_0 the effect of which on estimation is depressed by the adjustment.

A kind of degrees-of-freedom correction arises, too. It appears to reflect the number of parameters in the model which is one more than is 'estimated' so far. With the present model the adjustment may increase or *decrease* the estimate compared to the case where $T-3$ is replaced by the usual T. An enlargement arises if the MLE is less than one while a subtraction results if the MLE is over one. Hence the correction strengthens the adjustment due to the term $-\frac{1}{2}\log[1+(T-1)(1-\psi)^2]$ (our reasoning is based on numerical experimentation).

The vigor of the Cox–Reid adjustment is more easily grasped from the adjusted likelihood equation:

$$\frac{\partial l_{ad}(\psi)}{\partial \psi} = -\frac{(T-3)}{2\,\hat{\sigma}^2_{\psi}}\frac{\partial\,\hat{\sigma}^2_{\psi}}{\partial \psi} + \frac{(T-1)(1-\psi)}{1+(T-1)(1-\psi)^2} = 0 \qquad (6.57)$$

The second term in equation (6.57) is $O(1)$ for all ψ and is asymptotically negligible in comparison to the first term. The corollary is that $T^{1/2}(\hat{\psi}_{ad}-\psi)$ or $T(\hat{\psi}_{ad}-1)$ if the model is stationary or not, respectively, share the same distribution with the correspondingly standardised $\hat{\psi}_{ml}$. Here $\hat{\psi}_{ad}$ is implicitly defined as the value of ψ fulfilling the adjusted likelihood equation. Comparison to the likelihood equation (6.16) reveals that the magnitudes of the adjustment terms are the same under stationarity so the adjustment should be $O_p(T^{-1/2})$ here, too, as it is for the $\mathrm{AR}_\mu(1)$ model under stationarity. Under nonstationarity the (additive) adjustment term is of smaller magnitude than under the $\mathrm{AR}_\mu(1)$ model which implies asymptotically the most modest modification yet.

Equation (6.57) is equivalent to a fifth order polynomial in ψ, and $\hat{\psi}_{ad}$ must be found by solving the equation numerically or by numerical maximization of the adjusted likelihood (6.56). As the usual likelihood equation, the adjusted likelihood equation is

guaranteed to possess a real root as the order of the polynomial is odd.

The second derivative of the adjusted profile log-likelihood is needed when constructing the adjusted version of the Wald test:

$$\frac{\partial^2 l_{ad}(\psi)}{\partial^2 \psi} = -\frac{(T-3)}{2}\left[\frac{\partial^2 \hat{\sigma}^2_\psi / \partial^2\psi}{\hat{\sigma}^2_\psi} - \frac{\left(\partial \hat{\sigma}^2_\psi / \partial\psi\right)^2}{\hat{\sigma}^4_\psi}\right] - \frac{(T-1)\left[1-(T-1)(1-\psi)^2\right]}{\left[1+(T-1)(1-\psi)^2\right]^2}. \tag{6.58}$$

The formula for $\partial^2 \hat{\sigma}^2_\psi / \partial^2\psi$ is supplied in Appendix A3.

When the model is stationary, the asymptotic distribution of $\sqrt{W_{ad}}$, as defined by equation (4.14), is

$$\sqrt{W_{ad}}$$
$$= T^{1/2}(\hat{\psi}_{ad} - \psi)\sqrt{-T^{-1}\left[\frac{\partial^2 l_{ad}(\psi)}{\partial^2\psi}\right]\Bigg|_{\psi=\hat{\psi}_{ad}}}$$
$$= T^{1/2}(\hat{\psi}_{ad} - \psi)\cdot$$
$$\sqrt{\left\{\frac{(T-3)}{2T}\left[\frac{\partial^2 \hat{\sigma}^2_\psi / \partial^2\psi}{\hat{\sigma}^2_\psi} - \frac{\left(\partial \hat{\sigma}^2_\psi / \partial\psi\right)^2}{\hat{\sigma}^4_\psi}\right] + \frac{(T-1)\left[1-(T-1)(1-\psi)^2\right]}{T\left[1+(T-1)(1-\psi)^2\right]^2}\right\}\Bigg|_{\psi=\hat{\psi}_{ad}}}$$
$$\Rightarrow N(0, 1-\psi^2)\sqrt{\frac{1}{2}\left[\frac{2\sigma^2/(1-\psi^2)}{\sigma^2} - \frac{0^2}{\sigma^4}\right] + 0} \quad \text{or to}$$
$$N(0,1).$$

The convergences of the derivatives are explained in Appendix A3. The convergence of the additive adjustment term $(T-1)\left[1-(T-1)(1-\psi)^2\right]/\left[1+(T-1)(1-\psi)^2\right]^2$, evaluated at the AE, to zero, is easy to see: $(T-1)(1-\psi)^2$ is $O_p(T)$ under $\mid \psi \mid < 1$ so the term is $O_p(1)$ but becomes divided by T yielding only a $O_p(T^{-1})$ modification underneath the square-root sign. The asymptotic distribution is hence not altered by the adjustment. We note in passing that also the adjusted Wald statistic can be calculated in a straightforward way even though an explicit analytic formula for it is not known.

In the nonstationary situation the additive adjustment term, evaluated at the AE, is divided by T but it still contributes an $O_p(T^{-1})$ component because the order of magnitude of the subtraction $1-\hat{\psi}_{ad}$ is then different. The outcome is again that the asymptotic distribution of the adjusted version of the Wald statistic is identical with the distribution for the non-adjusted version:

$$\sqrt{W_{ad}}$$

$$= T(\hat{\psi}_{ad}-1)\sqrt{-T^{-2}\left[\frac{\partial^2 l_{ad}(\psi)}{\partial^2\psi}\right]\Bigg|_{\psi=\hat{\psi}_{ad}}}$$

$$= T(\hat{\psi}_{ad}-\psi)\cdot$$

$$\sqrt{\left\{\frac{(T-3)}{2T}\left[\frac{T^{-1}\partial^2\,\hat{\sigma}^2_{\psi}\,/\partial^2\psi}{\hat{\sigma}^2_{\psi}}-\frac{\left(T^{-1/2}\partial\,\hat{\sigma}^2_{\psi}\,/\partial\psi\right)^2}{\hat{\sigma}^4_{\psi}}\right]+\frac{(T-1)\left[1-(T-1)(1-\psi)^2\right]}{T^2\left[1+(T-1)(1-\psi)^2\right]^2}\right\}\Bigg|_{\psi=\hat{\psi}_{ad}}}$$

$$\Rightarrow \frac{\int_0^1 W(r)dW(r)}{\int_0^1 [W(r)]^2\,dr}\sqrt{\frac{1}{2}\left\{\frac{2\sigma^2\int_0^1[W(r)]^2\,dr}{\sigma^2}-\frac{0}{\sigma^4}\right\}+0} \quad \text{or to}$$

$$\frac{\int_0^1 W(r)dW(r)}{\sqrt{\int_0^1 [W(r)]^2\,dr}}.$$

The reasoning beneath the convergences of the derivatives is again provided in the Appendix.

Like the AEs, the adjusted Wald statistics follow the same asymptotic distributions as their non-adjusted analogues.

Example. Model (6.47) simplifies to a random walk with starting value zero when $\psi = 1$ and $\gamma_0 = 0$. Hence the illustrative time series of the previous examples can be construed to have been generated from the present model. The MLEs from the unadjusted and adjusted profile likelihoods are 1.007 and 1.006, respectively, when the 25 first observations are considered. Interestingly, the AE is the smaller. The estimates are $\hat{\psi}_{ml}$= 0.989 and $\hat{\psi}_{ad}$= 0.990 for all the 100 observations. The AEs are slightly more accurate in both cases. An obvious reason for the similarity of the estimates is that $\hat{\psi}_{ml}$ is very close to

one in the first place (and consequently the penalty term of the adjusted profile likelihood is almost zero). The estimators draw a more accurate value by chance with the smaller sample.

The profile likelihoods feature a unique maximum on any scale we have experimented with. They are not peculiar in other respects either.

An arresting feature of the example is the efficiency with which the Bhargava model has utilized the information in the sample of 25 observations compared to the model of the simple AR(1) model of Section 6.2. All of the present estimates are comparable in precision with the estimates reported there. This is in line with asymptotic theory.□

Our analysis closes again with notes on the invariance of estimates. It will be shown that adding a constant to the observations (including y_1) has no effect on $\hat{\psi}_{ml}$ nor $\hat{\psi}_{ad}$. We remark first that $\hat{\gamma}_{0\psi}$ transfers to $\hat{\gamma}_{0\psi} + a$ after addition of a constant a. The residual variance

$$\begin{aligned}\hat{\sigma}^2_{\psi} &= T^{-1}\left\{\left[y_1 + a - (\hat{\gamma}_{0\psi} + a)\right]^2 + \sum_{t=2}^{T}\left[y_t + a - (\hat{\gamma}_{0\psi} + a)(1-\psi) - \psi(y_{t-1} + a)\right]^2\right\} \\ &= T^{-1}\left\{(y_1 - \hat{\gamma}_{0\psi})^2 + \sum_{t=2}^{T}\left[y_t - \hat{\gamma}_{0\psi}\,(1-\psi) - \psi y_{t-1})\right]^2\right\},\end{aligned}$$

remains unchanged. The result is that $\hat{\psi}_{ml}$ is tied to its original value regardless of any constant added. The adjustment term of the adjusted likelihood equation (6.57) stays unmodified, too, and so does $\hat{\psi}_{ad}$.

If $\psi = 1$ then the starting value of the disturbance series $x_0 \neq 0$ has an analogous effect on y_t as adding a constant to y_t after the generation of the series. It follows that $\hat{\psi}_{ml}$ is undisturbed by $x_0 \neq 0$ if $\psi = 1$, and one can allow for $x_0 \neq 0$ when testing for a unit root via $\hat{\psi}_{ml}$.[43]

[43]The entities x_0 and γ_0 are not identified separately under $\psi = 1$ so the value of γ_0 is irrelevant in the same sense.

Letting x_0 differ from zero almost bridges the gap between the Bhargava process (6.47) and the AR(1) process with constant analysed by Andrews (1993). If Y_0^* in that article was set equal to a fixed constant then the processes would agree.

As with the previous model, invariance of $\hat{\psi}_{ad}$ with respect to the value of x_0 under $\psi = 1$ is less obvious because some of the influence measures — on which the adjustments rest — are disturbed by the value of x_0.[44] However, again a sole information measure of relevance needs revision:

$$i_{\psi\gamma_0} = \begin{cases} \sigma^{-2}x_0(1+\psi-\psi^T) & \text{if } \mid\psi\mid < 1 \\ \sigma^{-2}x_0 & \text{if } \psi = 1. \end{cases}$$

The parameters are not orthogonal unless $x_0 = 0$. The term $\hat{c}\ (\psi - \hat{\psi}_{ml})$ should thus be derived and included in the adjusted profile log-likelihood. However, the term is superfluous as coefficient c equals zero:

$$c = \frac{\partial}{\partial\gamma_0}(i^{-1}_{\gamma_0\gamma_0} i_{\psi\gamma_0}) = 0.$$

The first equation arises just as demonstrated in equation (6.14) after substituting γ_0 in place of α. The second equation is simply due to the fact that neither of the information measures are functions of γ_0 (equations (6.54) and (6.55)). The conclusion is that $\hat{\psi}_{ad}$ is not affected by the value of x_0 under $\psi = 1$ so it can be used as a basis for a unit-root test even when allowing for $x_0 \neq 0$.

Moreover, $\hat{\psi}_{ml}$ and $\hat{\psi}_{ad}$ are in general invariant with respect to the magnitude of the constant (γ_0) under the maintained model. This follows by noting that a change in the constant implies a corresponding change in the observations y_t, $t = 1,\ldots,T$, regardless of the magnitude of ψ, by formula (6.48), and from the first argument that adding a constant to the observations has no effect on the estimates.

Power ranking of different unit root tests can vary according to the value of x_0 (Schmidt and Phillips (1992) and Section 7.3). It gives some motivation for allowing x_0 to differ from zero but otherwise the case $x_0 \neq 0$ is not empirically functional: the estimators $\hat{\sigma}^2_\psi$, $\hat{\gamma}_{0\psi}$, and $\hat{\psi}_{\gamma_0}$ should in general be modified to take into account the nonzero value of x_0 and that is not possible as the x_t series is unobservable.

[44]Formula (A5.2) for $\hat{\sigma}^2_\psi$ in Appendix A5 permits a non-zero x_0 and is the basis for the following calculations.

6.6 Adjusted Profile Likelihood for a Unit-Root AR(2) Model

The process we shall study in this section is the AR(2) or the Yule process (after Yule (1927)):

$$y_t = \rho_1 y_{t-1} + \rho_2 y_{t-2} + \epsilon_t,$$

where $\epsilon_t \sim \text{NID}(0, \sigma^2)$, $\sigma^2 > 0$, $t = 1, \ldots, T$, and $y_0 = y_{-1} = 0$. This can be reparameterised as

$$y_t = \psi y_{t-1} + \phi_1 \Delta y_{t-1} + \epsilon_t \tag{6.59}$$

where $\psi = \rho_1 + \rho_2$ and $\phi_1 = -\rho_2$ (as in equation (3.16)). The nuisance parameter vector $[\phi_1 \ \sigma^2]$ will be denoted by $\boldsymbol{\phi}'$. Our interest is focused on ψ. We shall assume a unit root or that $\psi = \rho_1 + \rho_2 = 1$ and also that $\mid \phi_1 \mid < 1$. (This parameter combination is indicated by the bold line in Figure 5.1 in Section 5.3.) As mentioned in the introduction, the assumption simplifies the algebra and is also motivated by the stress on testing in the unit-root literature. A similar *a priori* assumption worked in some respects very well for the $\text{AR}_\mu(1)$ model which gives justification for the method, too. If the assumption is not made in a derivation of an information measure, then we let 'ψ' appear in the derivation.

The time series y_t starts to resemble $\Delta^2 y_t = \epsilon_t$ or an $I(2)$ process as ϕ_1 approaches 1. It seems likely that the relevant distribution theory then becomes distorted in small samples motivating an adjustment. (I(2) processes can yield most peculiar distributions, see e.g. Section 4.1 in Hendry (1995).) Indeed, in the simulation of Dickey (1976, Table 6.3) the fractiles of $Tf(\hat{\psi}_{ml} - 1)$ differed notably from the asymptotic ones under a ϕ_1 equal to 0.8 (the largest experimented value) so there is room for adjustment in estimation. ($T(\hat{\psi}_{ml} - 1)$ needs to be multiplied by $f = (1 - \phi_1)^{-1}$ for the asymptotic DF_μ distribution (6.8) to hold as explained in Section 3.3.) The possibility of an autoregressive almost $I(2)$ process is also an extra interesting feature of the present model compared to the previous models we have studied. It was suggested in Chapter 5 that the bias in the estimation of $\rho_1 + \rho_2$ can be considerable or of the same magnitude as with the AR(1)

model with constant when ρ_1 is zero. Thus a bias correction might be appropriate in the unit-root case, too, though it will turn out later that the Cox–Reid adjustment, at least the way we have devised it, does not focus on the bias under the present model. On the other hand, it was also shown that under stationarity, the bias is nonnegligible or of the same order as with the simple AR(1) model when ρ_2 is zero or it is unnecessarily estimated and ρ_1 is close to unity.

Turning next to the τ statistic, the aforementioned simulations of Dickey showed that the fractiles of it do not depend much on the magnitude of ϕ_1.[45] The graphs in Cheung and Lai (1995a) confirm that the size distortions are in general small for this statistic, though they found some when ϕ_1 is very close to one. Relatedly, according to the simulations of Cheung and Lai (1995b), a single superfluous estimated autoregressive coefficient (i.e. ϕ_1 equals zero but is estimated) does not cause much distortion when the τ test is employed. The analysis of the present model may be thus seen also as an explanatory step towards employing the Cox–Reid adjustment when testing for unit roots in AR(p) models for which more serious distortions are likely to arise even when the first autoregressive parameter is close to unity.

The model is heavily employed in empirical research. Orcutt (1948) has suggested that the unit-root AR(2) process $y_t = 1.3y_{t-1} - 0.3y_{t-2} + \epsilon_t$ generates economic data in general in the US (after correcting for the mean). Interestingly, AR(2) models arise still when analysing US data (Nelson and Plosser (1982, footnote 12)) and Phillips (1991b, Table IV)) and even with similar estimates (with detrended or differenced GNP, Blanchard (1981) and Rudebusch (1992). The AR(2) model (with trend) is a possible description of the Finnish GNP, too (Linden (1995a, Chapter IV or 1995b), but with different parameter values from the above.

[45] Stronger small-sample effects emerge when a constant or a constant and a time trend are included in the estimation (Dickey (*op. cit.*, Tables 6.4 and 6.5)).

The log-likelihood function corresponding to the model is

$$l(\psi, \phi_1, \sigma^2; \mathbf{y}) = -\frac{T}{2}\log\sigma^2 - \frac{1}{2\sigma^2}\sum_{t=1}^{T}(y_t - \psi y_{t-1} - \phi_1 \Delta y_{t-1})^2$$

while the profile log-likelihood function is

$$l_p(\psi) = -\frac{T}{2}\log\hat{\sigma}^2_{\psi}\,.$$

Here

$$\hat{\sigma}^2_{\psi} = T^{-1}\sum_{t=1}^{T}(y_t - \psi y_{t-1} - \hat{\phi}_{1\psi}\,\Delta y_{t-1})^2$$

and

$$\hat{\phi}_{1\psi} = \frac{\sum_{t=1}^{T} y_t \Delta y_{t-1} - \psi \sum_{t=1}^{T} y_{t-1}\Delta y_{t-1}}{\sum_{t=1}^{T}(\Delta y_{t-1})^2}.$$

The MLE of ψ is

$$\begin{aligned}\hat{\psi}_{ml} &= \frac{\sum_{t=1}^{T}(y_t - d\Delta y_{t-1})(y_{t-1} - d_{-1}\Delta y_{t-1})}{\sum_{t=1}^{T}(y_{t-1} - d_{-1}\Delta y_{t-1})^2} \\ &= \frac{\sum_{t=1}^{T} y_t y_{t-1}\sum_{t=1}^{T}(\Delta y_{t-1})^2 - \sum_{t=1}^{T} y_{t-1}\Delta y_{t-1}\sum_{t=1}^{T} y_t \Delta y_{t-1}}{\sum_{t=1}^{T} y_{t-1}^2 \sum_{t=1}^{T}(\Delta y_{t-1})^2 - (\sum_{t=1}^{T} y_{t-1}\Delta y_{t-1})^2}\end{aligned} \tag{6.60}$$

where

$$d = \frac{\sum_{t=1}^{T} y_t \Delta y_{t-1}}{\sum_{t=1}^{T}(\Delta y_{t-1})^2} \tag{6.61}$$

and

$$d_{-1} = \frac{\sum_{t=1}^{T} y_{t-1}\Delta y_{t-1}}{\sum_{t=1}^{T}(\Delta y_{t-1})^2}. \tag{6.62}$$

The first form of the MLE is more intuitive and is useful for some derivations whereas the second is more convenient for evaluation and some asymptotic calculations. Under

stationarity, a conventional asymptotic Normal distribution is relevant. In contrast to the AR(1) dynamics, the asymptotic unit-root distribution depends on a nuisance parameter or ϕ_1:

$$T(\hat{\psi}_{ml} - 1) \;=\; (1-\phi_1)\frac{\int_0^1 W(r)dr}{\int_0^1 [W(r)]^2 dr} \tag{6.63}$$

(cf. Section 3.3 or Fuller (1976, p. 374), say). This is the *DF* distribution (3.11) — which springs from the simple AR(1) model — multiplied by $(1-\phi_1)$.[46]

The t-statistic for the null hypothesis $\psi = 1$ is

$$\frac{\hat{\psi}_{ml} - 1}{s\left\{\sum_{t=1}^T (\Delta y_{t-1})^2 / [\sum_{t=1}^T y_{t-1}^2 \sum_{t=1}^T (\Delta y_{t-1})^2 - (\sum_{t=1}^T y_{t-1}\Delta y_{t-1})^2]\right\}^{1/2}} \tag{6.64}$$

where s^2 is the degrees-of-freedom corrected estimate of variance or:

$$s^2 = (T-2)^{-1}\sum_{t=1}^T (y_t - \hat{\psi}_{ml}\, y_{t-1} - \hat{\phi}_{1_{ml}}\, \Delta y_{t-1})^2$$

and $\hat{\phi}_{1_{ml}}$ is $\hat{\phi}_{1_\psi}$ evaluated at $\hat{\psi}_{ml}$. It was explained in Section 3.3 that the asymptotic distribution (3.12) of this t-ratio is free of nuisance parameters and can be used for approximate inference in finite samples for the null of a unit root. Alternatively, we could just as well employ (the square root of) the Wald statistic with s^2 above replaced by $\hat{\sigma}^2_{ml}$ ($\hat{\sigma}^2_\psi$ evaluated at $\hat{\psi}_{ml}$ and $\hat{\phi}_{1_{ml}}$). If the model were stationary then, of course, the asymptotic distribution would be Standard Normal.

As with the previous models, the variance parameter σ^2 is orthogonal with respect to the other parameters:

$$i_{\psi\sigma^2} \;=\; E(j_{\psi\sigma^2}) \;=\; E\left[\sigma^{-4}\textstyle\sum_{t=1}^T (y_t - \psi y_{t-1} - \phi_1 \Delta y_{t-1}) y_{t-1}\right] \;=\; 0$$

[46] In the present section Δy_t follows an AR(1) process with autoregressive coefficient ϕ_1. This means that, for example, $T^{-2}\sum_{t=1}^T y_{t-1}^2$ converges to $\sigma^2(1-\phi_1)^{-2}\int_0^1 [W(r)]^2 dr$ instead of $\sigma^2 \int_0^1 [W(r)]^2 dr$ as when ϕ_1 equals zero or Δy_t is IID, cf. Hamilton (1994, Section 17.5), say.

and

$$i_{\phi_1\sigma^2} = E(j_{\phi_1\sigma^2}) = E\left[\sigma^{-4}\textstyle\sum_{t=1}^{T}(y_t - \psi y_{t-1} - \phi_1\Delta y_{t-1})\Delta y_{t-1}\right] = 0.$$

However,

$$\begin{aligned} i_{\psi\phi_1} &= E\left[\sigma^{-2}\textstyle\sum_{t=1}^{T} y_{t-1}\Delta y_{t-1}\right] \\ &= \sigma^{-2}\left[\textstyle\sum_{t=1}^{T-1} E(y_t^2) - \phi_1\sum_{t=1}^{T-2} E(y_t^2) - \sum_{t=1}^{T-2} E(y_t\sum_{i=1}^{t}\epsilon_i)\right] \\ &= \textstyle\sum_{t=1}^{T-1}\mathbf{f}_t'\mathbf{T}_t\mathbf{f}_t - \phi_1\sum_{t=1}^{T-2}\mathbf{f}_t'\mathbf{T}_t\mathbf{f}_t - \sum_{t=1}^{T-2}\mathbf{f}_t'\mathbf{t} \\ &= \frac{T}{(1-\phi_1)^2(1+\phi_1)} - \frac{\phi_1^2+\phi_1+1}{(1-\phi_1)^3(1+\phi_1)^2} - \frac{\phi_1^T(\phi_1^{T+1}-\phi_1^2-2\phi_1-1)}{(1-\phi_1)^3(1+\phi_1)^2} \end{aligned} \tag{6.65}$$

where

$$\begin{aligned} \mathbf{f}_t' &= [\ \phi_1^{t-1}\ \ \phi_1^{t-2}\ \ \cdots\ \ 1\ ] \\ \mathbf{t}' &= [\ \ 1\ \ \ \ 2\ \ \ \cdots\ \ t\ ] \end{aligned}$$

and

$$\mathbf{T}_t = \begin{bmatrix} 1 & 1 & 1 & \cdots & 1 \\ 1 & 2 & 2 & \cdots & 2 \\ 1 & 2 & 3 & \cdots & 3 \\ \vdots & \vdots & \vdots & & \vdots \\ 1 & 2 & 3 & \cdots & t \end{bmatrix}$$

which is positive definite. The third equality follows from the formulae

$$E(y_t^2) = \sigma^2\mathbf{f}_t'\mathbf{T}_t\mathbf{f}_t$$

and

$$E(y_t \sum_{i=1}^{t} \epsilon_i) = \sigma^2 \mathbf{f}_t' \mathbf{t},$$

$t = 1, 2, \ldots$ The Cholesky decomposition of $\mathbf{T}_t$ is:

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

It was used in the derivation of the final closed form. The outcome is that ψ and ϕ_1 are not orthogonal (even locally at $\psi = 1$).

We go on to evaluate the rest of the required information measures. The first information measure $i_{\psi\phi_1}$ needed to derive c was already found in equation (6.65). The second one is

$$\begin{aligned} i_{\phi_1\phi_1} &= E(j_{\phi_1\phi_1}) \\ &= E\left[\sigma^{-2} \textstyle\sum_{t=1}^{T} (\Delta y_{t-1})^2\right] \\ &= \sigma^{-2}\left[(1-\phi_1)^2 \textstyle\sum_{t=1}^{T-2} E(y_t^2) - 2(1-\phi_1) \sum_{t=1}^{T-2} E(y_t \sum_{i=1}^{t} \epsilon_i) + \sum_{t=1}^{T-1} t\sigma^2\right] \\ &= (1-\phi_1)^2 \textstyle\sum_{t=1}^{T-2} \mathbf{f}_t' \mathbf{T}_t \mathbf{f}_t - 2(1-\phi_1) \sum_{t=1}^{T-2} \mathbf{f}_t' \mathbf{t} + T(T-1)/2 \\ &= \frac{T}{1-\phi_1^2} - \frac{1}{(1-\phi_1^2)^2} + \frac{\phi_1^{2T}}{(1-\phi_1^2)^2}. \end{aligned}$$

Again, the helpful Cholesky decomposition of $\mathbf{T}_t$ was used in deriving the final closed form.

Orthogonality of ψ and σ^2 markedly facilitates derivation of c which simplifies to

$$c = \frac{\partial}{\partial \phi_1}(i_{\phi_1\phi_1}^{-1} i_{\psi\phi_1}).$$

(paralleling exactly the derivation of formula (6.12)). It follows by straightforward but lengthy algebra that c is given by

$$c = \frac{1}{(1-\phi_1)^2} + h(\phi_1, T) \tag{6.66}$$

where $h(\phi_1, T)$ is $o(1)$ given by:

$$\begin{aligned} h(\phi_1,T) \;=\; & \frac{(1-\phi_1^T)(1-\phi_1^{2T})(1+2\phi_1-\phi_1^2-2\phi_1^T)}{(1-\phi_1)^2\left[(1-\phi_1^2)T-(1-\phi_1^{2T})\right]^2} \\ & -\frac{(1-\phi_1^2)(1-\phi_1^T)\left\{(1+\phi_1)^2+\phi_1^{T-1}\left[1-\phi_1^T-\phi_1(1+2\phi_1+\phi_1^T)\right]\right\}}{(1-\phi_1)^2\left[(1-\phi_1^2)T-(1-\phi_1^{2T})\right]^2}T \\ & +\frac{\phi_1^{T-1}(1-\phi_1^2)^2(1+\phi_1-2\phi_1^T)}{(1-\phi_1)^2\left[(1-\phi_1^2)T-(1-\phi_1^{2T})\right]^2}T^2. \end{aligned}$$

Numerical analysis shows that $c > 0$ for $T \geq 3$ and $\mid \phi_1 \mid < 1$. It can be seen that c converges to $(1-\phi_1)^{-2} > 0$ as T tends to infinity. It follows, for large T, that $\hat{c}$ is inclined to lie close to zero for negative values of ϕ_1, and tend to infinity as ϕ_1 reaches one. Numerical experiments suggest that the approximation $c \approx (1-\phi_1)^{-2}$ performs quite well if ϕ_1 or T are not too extraordinary. As an example assume that $\phi_1 = 0 = \hat{\phi}_{1_{ml}}$. Then $\hat{c}\,|_{\hat{\phi}_{1_{ml}}=0} = 1 + (T-1)^{-1}$, and so $\hat{c}$ converges (in probability) rapidly to 1 as T increases. We will examine the finite-sample behaviour of adjusted statistics based on both the exact and the approximate form for c in Chapter 7.

The present model has produced a c which depends greatly on the magnitude of the nuisance parameter (ϕ_1). This is quite contrary to the case with the applications of the previous sections. The adjustment peaks (for a reasonable T) when ϕ_1 is close to 1 or the series starts to resemble an $I(2)$ process. Of course, ψ does not appear in the formula for c as $\psi = 1$ is assumed. Inspection of formula (6.65) of $i_{\psi\phi_1}$ reveals that $i_{\psi\phi_1}$ can be very far from zero and that it tends to infinity as the number of observations increases for ϕ_1 close to -1. However, coefficient c collapses, approximately, to tiny $1/4$ for such values of ϕ_1. Thus orthogonality considerations do not explain the behaviour of c. Further attempts to interpret c follow below.

The observed information measure for σ^2

$$j_{\sigma^2\sigma^2} = -\frac{T}{2\sigma^4} + \sigma^{-6}\sum_{t=1}^{T}(y_t - \psi y_{t-1} - \phi_1 \Delta y_{t-1})^2,$$

accompanied by the previous results yields:

$$-\tfrac{1}{2}\log \mid \mathbf{J}_{\phi\phi}(\psi, \hat{\phi}_\psi) \mid = \tfrac{3}{2}\log \hat{\sigma}^2_\psi - \tfrac{1}{2}\log\left[\tfrac{T}{2}\sum_{t=1}^{T}(\Delta y_{t-1})^2\right].$$

The Cox–Reid (1993) adjusted profile log-likelihood is:

$$\begin{aligned} l_{ad2}(\psi) &= -\tfrac{T}{2}\log \hat{\sigma}^2_\psi + \tfrac{3}{2}\log \hat{\sigma}^2_\psi - \tfrac{1}{2}\log\left[\tfrac{T}{2}\sum_{t=1}^{T}(\Delta y_{t-1})^2\right] + \hat{c}\,(\psi - \hat{\psi}_{ml}) \\ &\propto -\frac{(T-3)}{2}\log \hat{\sigma}^2_\psi + \hat{c}\,(\psi - \hat{\psi}_{ml}). \end{aligned} \tag{6.67}$$

The adjustment has produced a kind of degrees-of-freedom correction together with the $\hat{c}\,(\psi - \hat{\psi}_{ml})$ correction as with the previous models. The strength of the adjustment depends in general on the value of ϕ_1. As already stated, c is small for negative values of ϕ_1 and larger for values of ϕ_1 closer to one. If $\hat{\phi}_{1_{ml}} = 0$ then $\hat{c} \approx 1$ (cf. above). This does not imply collapsing of equation (6.67) to (6.2), the adjusted likelihood for the simple AR(1) model. Instead, the degrees-of-freedom correction is $T-3$, and also the $(\psi - \hat{\psi}_{ml})$ term remains. From the analysis of the parallel formula for the $\mathrm{AR}_\mu(1)$ model (p. 84) we can infer that the degrees-of-freedom correction fortifies the adjustment due to the term $\hat{c}\,(\psi - \hat{\psi}_{ml})$. The joint impact of the adjustments will be a local maximum to the right of $\hat{\psi}_{ml}$ or an estimate $\hat{\psi}_{ad2} > \hat{\psi}_{ml}$ as long as $\hat{c} > 0$. This makes sense in two ways. First, $\hat{\psi}_{ml}$ should be biased towards zero under a unit root according to the suggestive calculations of Section 5.3. Second, the power decline of the DF tests as the number of superfluous lags increases, contemplated in Section 3.5, is very possibly due to a shift of the distribution of $\hat{\psi}_{ml}$ to the left. Thus, when $\phi_1 = 0$, there is a further reason to pull the MLE to the right on the real axis.

Comparison of the formula for c with formula (5.10), expressing the approximate bias for stationary models in the estimation of $E(\rho_1 + \rho_2)$, yields a surprise. The formula suggests that the bias is zero along the line $\rho_2 = -(1+\rho_1)/3$. On the other hand, we are assuming that the parameters lie along the line $\rho_2 = 1 - \rho_1$. (Figure 5.1 in Section 5.3 may help in understanding the following discussion. The zero-bias line is indicated in the figure and the bolded line tracks the unit-root parameter combinations.) The bias should hence be puny for ρ_2 close to minus one and more substantial for ρ_2 large or for ρ_2 close to the zero bias line and for ρ_2 farther away from the line, respectively. Correspondingly, the bias and need for a bias correction seem to decrease with $\phi_1 = -\rho_2$. Puzzlingly, coefficient c increases with ϕ_1 or behaves in an exactly opposite fashion to the bias . The Cox–Reid adjustment seemed to have a modest bias increasing effect for a range of negative values of ψ as defined for the AR(1)$_\mu$ model. Here the approximative bias formula suggests that the magnitude of the Cox–Reid adjustment depends *inversely* on the bias for all parameter values under consideration. The explanation might lie in the fact that the approximation of the bias breaks down in the neighbourhood of nonstationary parameter values. However, it will be clear by the end of Section 7.7 that this explanation is not valid.

Another interpretation is that the adjustment focuses on other, possibly more important, properties of the distribution of the MLE. As already remarked, the distribution of the MLE should change radically as ϕ_1 tends to one or the process becomes $I(2)$. Let us consider the sum of the coefficients $\pi(1)$ of the Wold presentation for Δy_t,

$$\Delta y_t = \pi(B)\epsilon_t,$$

where $\pi(B) = 1 + \pi_1 B + \pi_2 B^2 + \ldots = (1 - \phi_1 B)^{-1}$. $\pi(1)$ is a measure of *persistence* or of the long-run effect of an individual innovation ϵ_t on the level of the series y_t. It derives from a decomposition of Beveridge and Nelson (1981), see also Campbell and Mankiw (1987), Rudebusch (1992) and Lippi and Reichlin (1992). An alternative explication of the measure is that it is the larger the more prevalently the nonstationary component

dominates y_t (cf. Beveridge and Nelson (*op. cit.*)). The reason to consider persistence here is that it is related in a straightforward fashion to the Cox–Reid adjustment: $\pi(1) = (1-\phi_1)^{-1}$ behaves very similarly as coefficient $c \approx (1-\phi_1)^{-2}$ for $\phi_1 \in (-1, 1)$, for example both peak for ϕ_1 close to unity. Hence the Cox–Reid adjustment increases, under a unit root, with the degree of persistence in y_t.

A further explanation might be that the magnitude of the Cox–Reid adjustment is not related in a straightforward fashion to the magnitude of coefficient c in the present context because of the dependence of the coefficient on a nuisance parameter. Namely, it can be shown that the Fisher information for ψ (under the assumption of a unit root) is strongly positively related to the magnitude of ϕ_1 as well. Thus, a sample is likely to be informative or the profile likelihood steeply curved when ϕ_1 is large. The magnitude of the adjustment depends on the interplay of the magnitude of coefficient c and the curviness of the profile log-likelihood (cf. equation (6.67)). The adjustment could tend to be smaller under large values of ϕ_1 if the profile log-likelihood becomes curved enough for such values.

We make a note on the order of magnitude of coefficient c for the stationary AR(2) model before commenting on the relative magnitude of the adjustment. Coefficient c would converge under $|\psi| < 1$: information measures divided by T converge if the model is stationary so a derivative of a ratio of the information measures (which are different from zero) like c for the stationary AR(2) model would converge as well.

The adjustment for the AR(2) model emerges similar in importance to the one for the Bhargava $\mathrm{AR}_\mu(1)$ model. The term $(\psi - \hat{\psi}_{ml})$ of the adjusted profile log-likelihood is multiplied by $\hat{c}$ which is $O_p(1)$ yielding a second term of the adjusted profile log-likelihood comparable in size with the second term in the adjusted profile log-likelihood for the Bhargava $\mathrm{AR}_\mu(1)$ model — whether the models are stationary or nonstationary. The influence of the first terms appears comparable for the models, too, whether they are stationary or not. That is to say, the adjustment is smaller for the unit-root model than for the stationary model. The outcome is quite contrary to the $\mathrm{AR}_\mu(1)$ model with which

the adjustment was more important (in magnitude) when the model had a unit root. The distinction to the $\mathrm{AR}_\mu(1)$ model will be explicit when an asymptotic approximation of the kind used in Section 6.3 is presented.

Adjusted profile log-likelihood (6.67) has the same appearance as its counterpart (6.15). So the derivation of the adjusted likelihood equation and its roots parallels exactly with that which was presented in Section 6.3. The sole difference is the new definition of $\hat{l}$:

$$\begin{aligned} \hat{l} &= \frac{\sum_{t=1}^{T}(y_t - d\Delta y_{t-1})^2}{\sum_{t=1}^{T}(y_{t-1} - d_{-1}\Delta y_{t-1})^2} \\ &= \frac{\sum_{t=1}^{T} y_t^2 \sum_{t=1}^{T}(\Delta y_{t-1})^2 - (\sum_{t=1}^{T} y_t \Delta y_{t-1})^2}{\sum_{t=1}^{T} y_{t-1}^2 \sum_{t=1}^{T}(\Delta y_{t-1})^2 - (\sum_{t=1}^{T} y_{t-1}\Delta y_{t-1})^2} \geq 0. \end{aligned} \tag{6.68}$$

Besides that the first expression is helpful in some derivations, it points out a similarity in the structure of $\hat{l}$ to the form it took under the $\mathrm{AR}_\mu(1)$ model with constant (equation (6.17)). $\hat{l}$ is a ratio of the sum of the squared residuals from y_t or y_{t-1} regressed on the variable associated with the nuisance parameter (Δy_{t-1} or the constant) in the numerator or in the denominator, respectively. The nonnegativeness of $\hat{l}$ is due to the Cauchy–Schwarz inequality.[47]

As already stated, analogous reasoning to that in Section 6.3 provides us with the roots

$$\hat{\psi}_{ml} + \frac{T-3}{2\hat{c}} \pm \sqrt{\frac{(T-3)^2}{4\hat{c}^2} + \hat{\psi}_{ml}^2 - \hat{l}}.$$

As with the $\mathrm{AR}_\mu(1)$ model it is beneficial to distinguish some cases:

[47]The entity $\hat{l}$ tends in probability to unity when the number of observations tends to infinity and ψ equals unity but not in general under stationarity in the present context, cf. Appendix A4.

i) $\hat{c} < 0$ and the roots are complex (with nonzero imaginary parts)

ii)* $\hat{c} < 0$, $\hat{\psi}_{ml}^2 - \hat{l} > 0$ and the roots are real (and unequal)

ii) $\hat{c} < 0$, $\hat{\psi}_{ml}^2 - \hat{l} < 0$ and the roots are real (and unequal)

iii) $\hat{c} = 0$ (in which case the adjusted likelihood equation implies a polynomial of order one)

iv) $\hat{c} > 0$, $\hat{\psi}_{ml}^2 - \hat{l} < 0$ and the roots are real (and unequal)

iv)* $\hat{c} > 0$, $\hat{\psi}_{ml}^2 - \hat{l} > 0$ and the roots are real (and unequal)

v) $\hat{c} > 0$ and the roots are complex (with nonzero imaginary parts).

The sign of the expression $\hat{\psi}_{ml}^2 - \hat{l}$ has to be taken into account here as we have not proven as for the $\mathrm{AR}_\mu(1)$ model (in Appendix A1) that it would always appear nonpositive. The cases *ii*)* and *iv*)* did not need consideration for this reason in Section 6.3. The AE is defined below in such a way that the local maximum of $l_{ad2}(\psi)$ emerges at the reported root (Appendix A4):

$$\hat{\psi}_{ad2} = \begin{cases} -1 & \text{if } i) \text{ applies} \\ \hat{\psi}_{ml} + \frac{T-3}{2\hat{c}} - \sqrt{\frac{(T-3)^2}{4\hat{c}^2} + \hat{\psi}_{ml}^2 - \hat{l}} & \text{if } ii^*) \text{ applies} \\ \hat{\psi}_{ml} + \frac{T-3}{2\hat{c}} + \sqrt{\frac{(T-3)^2}{4\hat{c}^2} + \hat{\psi}_{ml}^2 - \hat{l}} & \text{if } ii) \text{ applies} \\ \hat{\psi}_{ml} & \text{if } iii) \text{ applies} \\ \hat{\psi}_{ml} + \frac{T-3}{2\hat{c}} - \sqrt{\frac{(T-3)^2}{4\hat{c}^2} + \hat{\psi}_{ml}^2 - \hat{l}} & \text{if } iv) \text{ applies} \\ \hat{\psi}_{ml} + \frac{T-3}{2\hat{c}} + \sqrt{\frac{(T-3)^2}{4\hat{c}^2} + \hat{\psi}_{ml}^2 - \hat{l}} & \text{if } iv^*) \text{ applies} \\ 1 & \text{if } v) \text{ applies.} \end{cases} \tag{6.69}$$

The definition follows similar reasoning to that of the $\mathrm{AR}_\mu(1)$ model when the roots are complex. The cases are ordered in such a way that $\hat{\psi}_{ad2}$ increases relative to $\hat{\psi}_{ml}$ for the circumstances from *ii*)* to *iv*)*.

Parallel to the $\mathrm{AR}_\mu(1)$ model, the roots will be real for large T. Coefficient $\hat{c}$ tends again stochastically to $c > 0$, so the second term of $l_{ad2}(\psi)$ is multiplied by an $O_p(1)$

quantity. The order of magnitude would stay the same if the model were stationary (cf. the discussion on p. 133). The second term $-[(T-3)/2]\log\hat{\sigma}^2_{\psi}$ is the dominant one as it is at least or exactly $O_p(T)$ when the process is nonstationary or stationary, respectively. Hence $l_{ad2}(\psi)$ tracks curvature and a local maximum — an indication of real roots. Further, it is shown in Appendix A4 that case *iv*) applies asymptotically in general.[48]

Similar calculations as in Section 6.3 lead to the formula

$$\hat{\psi}_{ad2} = \hat{\psi}_{ml} - \frac{\hat{c}\,(\hat{\psi}^2_{ml} - \hat{l})}{T-3} + r_T, \qquad r_T = \begin{cases} O_p(T^{-3}) & \text{if } |\psi| < 1 \\ O_p(T^{-5}) & \text{if } \psi = 1. \end{cases} \tag{6.70}$$

If the process is stationary we have

$$T^{1/2}(\hat{\psi}_{ad2} - \psi) = T^{1/2}(\hat{\psi}_{ml} - \psi) - T^{1/2}\frac{\hat{c}\,(\hat{\psi}^2_{ml} - \hat{l})}{T-3} + r^*_T,$$

$$r^*_T = O_p(T^{-5/2}), \quad |\psi| < 1$$

(Appendix A4). The adjustment term on the right-hand side is $O_p(T^{-1/2})$ because both $\hat{c}$ and $(\hat{\psi}^2_{ml} - \hat{l})$ are $O_p(1)$ under $|\psi| < 1$ (Appendix A4) . The asymptotic distribution of the standardised AE or $T^{1/2}(\hat{\psi}_{ad2} - \psi)$ is hence the same as that of the ordinary MLE, so the adjustment fades when the process is stationary. The adjustment is of the same order as in the stationary $\mathrm{AR}_{\mu}(1)$ model.

We note in passing that if y_t were stationary but the unit-root formula for $\hat{c}$ (c defined as in equation (6.66)) were used (improperly) then the same limiting distribution would

---

[48]The following heuristic reasoning may be confirming. $\hat{\psi}^2_{ml} - \hat{l}$ converges stochastically to zero (as a unit root is assumed) so $\hat{\psi}_{ml} + \frac{T-3}{2\hat{c}} - \sqrt{\cdot} \approx \hat{\psi}_{ml}$ for large T and the other root $\hat{\psi}_{ml} + \frac{T-3}{2\hat{c}} + \sqrt{\cdot}$ diverges stochastically to infinity. Consequently, $\hat{\psi}_{ml} + \frac{T-3}{2\hat{c}} - \sqrt{\cdot}$ is the root which is likely to correspond to the local maximum. Again, asymptotically, $\partial^2 l_{adj2}(\psi)/\partial^2\psi < 0$ at the root (Appendix A4).

arise. The second term would still be $O_p(T^{-1/2})$.

The corresponding formula for the nonstationary case is

$$T(\hat{\psi}_{ad2} - 1) = T(\hat{\psi}_{ml} - 1) - T\frac{\hat{c}\,(\hat{\psi}_{ml}^2 - \hat{l})}{T-3} + r_T^*,$$

$$r_T^* = O_p(T^{-4}), \quad \psi = 1.$$

The adjustment term is here $O_p(T^{-1})$, so it vanishes asymptotically as $\hat{c}$ is $O_p(1)$ and $(\hat{\psi}_{ml}^2 - \hat{l})$ is $O_p(T^{-1})$ when $\psi = 1$ (Appendix A4). It follows that the asymptotic distribution is the same as that of the MLE. That is to say, $Tf(\hat{\psi} - 1)$, where $f = (1 - \phi_1)^{-1}$, follows the Dickey–Fuller distribution (3.11) (Section 3.3). The contrast to the order of the adjustment with the $\mathrm{AR}_\mu(1)$ model is now quite apparent.

The difference between the appropriately standardised estimates diminishes faster in the unit-root than in the stationary case. An adverse circumstance arose with the $\mathrm{AR}_\mu(1)$ model which generated a smaller adjustment (in terms of order of magnitude) for the stationary process.

Just as the roots of the adjusted likelihood equation have the same appearance for the present and the $\mathrm{AR}_\mu(1)$ model, so do the formulae for the adjusted Wald statistic:

$$W_{ad2} = (\hat{\psi}_{ad2} - \psi_0)^2 \left[\frac{-(T-3)(\psi^2 - \hat{l} - 2\psi\,\hat{\psi}_{ml} + 2\,\hat{\psi}_{ml}^2)}{\left(\hat{l} - 2\psi\,\hat{\psi}_{ml} + \psi^2\right)^2} \right]_{\psi = \hat{\psi}_{ad2}}. \tag{6.71}$$

The distinction is, of course, the revised definition of $\hat{l}$ (formula (6.68)). An approximation similar to the one in Section 6.3 is available, too, and is based on the following result proved in Appendix A4:

$$\left.\frac{\partial^2 l_{ad2}(\psi)}{\partial^2 \psi}\right|_{\psi=\hat{\psi}_{ad2}} = \frac{(T-3)\left(\hat{\psi}_{ml}^2 - \hat{l} + r_T^*\right)}{\left[-(\hat{\psi}_{ml}^2 - \hat{l}) + r_T^*\right]^2}, \quad r_T^* = \begin{cases} O_p(T^{-2}) \text{ for } |\psi| < 1 \\ O_p(T^{-4}) \text{ for } \psi = 1. \end{cases}$$

Here the evaluation at the AE assumes that T is large enough so that case *iv*) of definition (6.69) for the AE applies. It follows that the square root of the adjusted Wald statistic is

$$\sqrt{W_{ad2}} = (\hat{\psi}_{ad2} - \psi_0) \frac{\sqrt{-(T-3)\left(\hat{\psi}_{ml}^2 - \hat{l} + r_T^*\right)}}{-(\hat{\psi}_{ml}^2 - \hat{l}) + r_T^*}. \tag{6.72}$$

This expression allows us to infer the asymptotic distribution of the statistic.

Let us first consider the stationary case briefly. Useful results are then that

$$\frac{-T}{\hat{\psi}_{ml}^2 - \hat{l}} = \frac{\sum_{t=1}^T y_{t-1}^2 \sum_{t=1}^T (\Delta y_{t-1})^2 - (\sum_{t=1}^T y_{t-1}\Delta y_{t-1})^2}{\hat{\sigma}_{ml}^2 \sum_{t=1}^T (\Delta y_{t-1})^2} \geq 0 \tag{6.73}$$

and that $\hat{\psi}_{ml}^2 - \hat{l}$ is $O_p(1)$ as already remarked. The equality is easily found by straightforward algebra when the formulae

$$\begin{aligned} \hat{\sigma}_{ml}^2 &= T^{-1} \textstyle\sum_{t=1}^T (y_t - \hat{\psi}_{ml}\, y_{t-1} - \hat{\phi}_{1ml}\, \Delta y_{t-1})^2 \\ &= T^{-1} \Big[\textstyle\sum_{t=1}^T (y_t - d\Delta y_{t-1})^2 + \hat{\psi}_{ml}^2 \sum_{t=1}^T (y_{t-1} - d_{-1}\Delta y_{t-1})^2 \\ &\quad -2\, \hat{\psi}_{ml} \textstyle\sum_{t=1}^T (y_t - d\Delta y_{t-1})(y_{t-1} - d_{-1}\Delta y_{t-1}) \Big], \end{aligned}$$

$$\hat{\phi}_{1_{ml}} = \frac{\sum_{t=1}^T y_t \Delta y_{t-1} - \hat{\psi}_{ml} \sum_{t=1}^T y_{t-1}\Delta y_{t-1}}{\sum_{t=1}^T (\Delta y_{t-1})^2},$$

and the definitions (6.61) and (6.62) for d and d_{-1}, respectively, are taken account of. The inequality above is again due to the Cauchy-Schwarz result. We note in passing that

if the equality in (6.73) is employed in the construction of the Wald statistic then it can be expressed as a function of $\hat{\psi}_{ml}$ and $\hat{l}$ only just as when the model is the $\mathrm{AR}_{\mu}(1)$ (cf. p. 99).

Substituting the result (6.73) into the above approximation for $\sqrt{W_{ad2}}$ yields

$$\sqrt{W_{ad2}} \approx (\hat{\psi}_{ad2} - \psi_0) \sqrt{\frac{\sum_{t=1}^{T} y_{t-1}^2 \sum_{t=1}^{T} (\Delta y_{t-1})^2 - (\sum_{t=1}^{T} y_{t-1} \Delta y_{t-1})^2}{\hat{\sigma}_{ml}^2 \sum_{t=1}^{T} (\Delta y_{t-1})^2}}.$$

We have argued before that the AE shares the asymptotic distribution with the MLE. The square-root term is also essentially the same as the corresponding term of the usual t-ratio (6.64). Thus $\sqrt{W_{ad2}}$ follows the Standard Normal asymptotically just like the usual t-ratio.

The unit-root case is more interesting in the present application of the Cox–Reid theory. The first equality below can be inferred from the calculation detail of the appendix (u_t there stands for Δy_{t-1} here). The asymptotic distribution follows readily:

$$\begin{aligned}
T\left(\hat{\psi}_{ml}^2 - \hat{l}\right) &= T\left\{\frac{-(1-\phi_1^2)\sum_{t=1}^{T} y_{t-1}^2 \left[\sum_{t=1}^{T} (\Delta y_{t-1})^2\right]^3 + O_p(T^4)}{[\sum_{t=1}^{T} y_{t-1}^2 \sum_{t=1}^{T} (\Delta y_{t-1})^2 - (\sum_{t=1}^{T} y_{t-1} \Delta y_{t-1})^2]^2}\right\} \\
&= \frac{-(1-\phi_1^2) T^{-2} \sum_{t=1}^{T} y_{t-1}^2 \left[T^{-1} \sum_{t=1}^{T} (\Delta y_{t-1})^2\right]^3 + O_p(T^{-1})}{[T^{-2} \sum_{t=1}^{T} y_{t-1}^2 T^{-1} \sum_{t=1}^{T} (\Delta y_{t-1})^2 + O_p(T^{-1})]^2} \\
&\Rightarrow \frac{-(1-\phi_1^2)\sigma^2(1-\phi_1)^{-2} \int_0^1 [W(r)]^2 dr [\sigma^2(1-\phi_1^2)^{-1}]^3}{[\sigma^2(1-\phi_1)^{-2} \int_0^1 [W(r)]^2 dr \sigma^2(1-\phi_1^2)^{-1}]^2} \quad \text{or to} \\
&\quad \frac{(1-\phi_1)^2}{\int_0^1 [W(r)]^2 dr}.
\end{aligned}$$

This and the asymptotic unit-root distribution (6.63) for the standardised AE and the approximation (6.72) for $\sqrt{W_{ad2}}$ enable us to reason that

$$\sqrt{W_{ad2}} \Rightarrow (1-\phi_1)\frac{\int_0^1 W(r)dr}{\int_0^1[W(r)]^2dr}\sqrt{\frac{(1-\phi_1)^2/\int_0^1[W(r)]^2dr}{[(1-\phi_1)^2/\int_0^1[W(r)]^2dr]^2}} \quad \text{or to}$$

$$\frac{\int_0^1 W(r)dr}{\sqrt{\int_0^1[W(r)]^2dr}}.$$

The asymptotic distribution of the square root of the adjusted Wald statistic is free of nuisance parameters and coincides with the DF–τ distribution (3.12) for the conventional Wald statistic.

Example. The random-walk model results if $\psi = 1$ and $\phi_1 = 0$ so our exemplary time series can be interpreted to have been generated from the AR(2) model (6.59) as well. The estimates for the first 25 observations are $\hat{\psi}_{ml}$= 0.964 and $\hat{\psi}_{ad2}$= 0.978. Estimation with all the 100 data points yields $\hat{\psi}_{ml}$= 1.000 and $\hat{\psi}_{ad2}$= 1.000. The example suggests, in accordance with the result on the bias in the stationary case in Section 5.3 and the referred results of Cheung and Lai (1995), that a sole autoregressive parameter ϕ_1 is not a major nuisance if $\phi_1 = 0$. Namely, the estimates happen to be closer to 1 than the MLEs of Section 6.2 which used the *a priori* information $\phi_1 = 0$. The AE is still slightly better than the MLE if only the 25 first observations are analysed. It is also interesting that $\hat{\psi}_{ad2}$ does not overestimate ψ for the latter sample even though $\hat{\psi}_{ml}$ equaled 1 (within reporting accuracy).

The $\hat{c}$s are 0.964 and 0.762, for the shorter and longer time series, respectively. The numbers can be compared with the asymptotic theoretical value of c equal to 1. The ψ at which the minimum of $l_{ad2}(\psi)$ takes place, or the other root of the adjusted likelihood equation, diverges as suggested by asymptotic theory (Appendix A4): The roots were 24.993 and 128.348, respectively, for the two samples. Finally, we remark that the AEs would remain the same (within reporting accuracy) if the approximative formula $c \approx (1-\phi_1)^{-2}$ were utilised. □

Chapter 7
Finite-Sample Properties

7.1 Introductory Remarks

Fractiles of the finite-sample and the asymptotic distributions of the standardised adjusted coefficients $T(\hat{\psi}_{ad2} - 1)$, $T(\hat{\psi}_{ad2,i} - 1)$, and $T(\hat{\psi}_{ad2,ap} - 1)$ and of the adjusted Wald statistics $\sqrt{W_{ad2}}$ and $\sqrt{W_{ad2,ap}}$ for the $\mathrm{AR}_\mu(1)$ model are reported in Section 7.2. (As explained in Chapter 6 the appropriate standardising factor of the estimates is T when a unit root exists as opposed to $T^{1/2}$ when the process is stationary.) This makes it possible to use these statistics for testing of unit roots. The tests will be referred to as AD tests (AD for adjusted) for short.

Unless otherwise pointed out, it will be assumed that the data-based estimate of coefficient c is used (instead of the *a priori* formula $c = (T-1)/2$) in the construction of $\hat{\psi}_{ad2}$ and $\sqrt{W_{ad2}}$. The distribution of coefficient $\hat{c}$ (formula (6.14)) under $\psi = 1$ is considered in the section, too.

The power of the tests is focused on in Section 7.3. The usual DF test based on the statistic $T(\hat{\psi}_{ml} - 1)$ (when a constant is estimated) and the new tests based on the aforementioned SCs will be denoted ψ_μ, $\psi_{\mu,ad}$, $\psi_{\mu,ad,i}$ and $\psi_{\mu,ad,ap}$, respectively, or called jointly SC tests for short. The nomenclature ρ_μ was used by Dickey and Fuller (1979) of the first mentioned test but a substitute symbol is used here for consistency with the present notation. The corresponding (apart from the iterated variant) Wald tests will carry the labels τ_μ, $\tau_{\mu,ad}$, and $\tau_{\mu,ad,ap}$ in line with the original notation of Dickey and Fuller (*op. cit.*). A difference, though, is that τ_μ stands here for the Wald test which makes use of the MLE instead of the degrees-of-freedom corrected estimate for the variance of the

innovation.

The concern of Section 7.4 is the agreement of the finite-sample distributions of the statistics with the asymptotic Normal distribution under the unit-root $AR_\mu(1)$ with drift. The finite-sample fractiles for the MLE, the AE, and the corresponding Wald statistics for the Bhargava AR(1) model with constant are presented in Section 7.5. Powers of the respective tests, now denoted by ψ^B_μ, $\psi^B_{\mu,ad}$, τ^B_μ, and $\tau^B_{\mu,ad}$, are inspected in Section 7.6. The AR(2) model is scrutinised in Section 7.7. The relative accuracy of the AE and of the OMCE (cf. Section 5.4) for the different models is briefly compared in the final section.

Throughout, we shall first address the SCs and the Wald statistics which do not utilize *a priori* information of a unit root, and last the statistics which do (when such have been derived).

The effect of the starting value on the powers is highlighted. We consider also extreme values of it; such cases yield insight to the properties of the tests even if one did not consider them empirically relevant.

The results of this chapter are based on simulation experiments. The fractiles and empirical sizes are calculated from 100 000 and power estimates from 10 000 replications, respectively. Other details of the technique are provided in footnotes and in Appendix A6 where the programmes are also reported.

7.2 Finite-Sample Distributions for the Unit-Root AR(1) Model with Constant

Table 7.1 documents the fractiles of $T(\hat{\psi}_{ad2} - 1)$ for different sample sizes.[1] The fractiles move in general left-ward on the real axis with T which happens also with the fractiles

[1]Different random numbers were used to calculate each entry to avoid dependence. The entry corresponding to an infinite number of observations is simulated from formula (6.29). The Brownian motions were approximated by random walks of length 1 000 (cf. Appendix A6 for particulars).

For finite samples, the 95 per cent confidence intervals for the fractiles are at most of width 0.27 in the left half of the table but much smaller in general. In the right half of the table, the confidence intervals are at most of span 0.15 but usually much smaller.

of $T(\hat{\psi}_{ml} - 1)$.[2] The former lie consistently to the right of the latter (cf. Fuller (1976, p. 371)) so $\hat{\psi}_{ad2}$ appears less biased than $\hat{\psi}_{ml}$ for all sample sizes. The difference between the fractiles of $T(\hat{\psi}_{ad2} - 1)$ and $T(\hat{\psi}_{ml} - 1)$ is reported as Table 7.2 to ease reference.[3] It can be seen, among other things, that the difference between the medians remains at approximately 1.4 or the medians of $\hat{\psi}_{ad2}$ and $\hat{\psi}_{ml}$ deviate by about $1.4/T$. Other regular patterns exist between the distributions, at least for sample sizes less than 500. The difference is larger at the very right tail of the distribution than at the very left (as with fractiles 0.05 and 0.95, say) though the difference tends to diminish as the sample size increases. In other words, the adjustment stretches the right tail of the distribution and the more so the smaller the sample. The stretching phenomena was remarked on already when commenting upon Figure 6.8 portraying the asymptotic distribution, but it is interesting to note that the phenomenon is stronger in finite samples.[4] The left tail follows an opposite pattern as the difference tends to increase with the sample size. It is because the left tail of the distribution of $T(\hat{\psi}_{ml} - 1)$ bends more left-ward than that of $T(\hat{\psi}_{ad2} - 1)$. Hence the adjustment becomes slightly more effective in cutting the left tail of the MLE when the sample size increases. The adjustment seems to cut down skewness as well.

For larger sample sizes the above mentioned tendencies with increasing T are not so apparent, especially at the left tail of the distribution. The extreme fractiles of the right tail lie still relatively further right than the corresponding fractiles of the left tail. The

---

[2]The biases of the estimates are likely to decrease with T despite the leftward movement of the *standardised* estimates. Andrews (1993, p. 146) argues that it is curiously possible that a specific fractile does not increase with T.

[3]The comparison is based on the distribution of $T(\hat{\psi}_{ml} - 1)$ which was simulated simultaneously with the distribution of $T(\hat{\psi}_{ad2} - 1)$. The use of the same random numbers improves the accuracy of the comparison. The confidence intervals for the differences should be smaller than for the fractiles themselves because of the strong positive correlation between the fractiles of the two distributions. Furthermore, the fractiles reported in Fuller (1976) apply for a T one less than what we mean by T. He indexes time series so that the observations run from y_2 to y_T instead of from y_1 to y_T as in the present study.

As already remarked, Nabeya and Tanaka (1990a) have published accurate fractiles for the asymptotic distributions of the MLE of ψ for this and related models under a unit-root.

[4]A reason is that a substantial $\hat{c}$ is more probable when T is small (because the extreme fractiles, say the 0.975^{th} and 0.99^{th}, of $\hat{\psi}_{ml}$ decrease as T increases).

asymptotic distribution (6.29) emerges quite well with samples of size 1 000.[5]

T/Fr.	0.01	0.025	0.05	0.10	0.50	0.90	0.95	0.975	0.99
25	-15.23	-12.66	-10.54	-8.35	-2.63	0.99	2.19	3.48	5.41
50	-17.05	-13.83	-11.43	-8.94	-2.81	0.83	1.93	3.08	4.93
75	-17.63	-14.21	-11.71	-9.11	-2.85	0.78	1.86	3.02	4.60
100	-17.99	-14.52	-11.85	-9.26	-2.83	0.76	1.81	2.87	4.47
250	-18.38	-14.79	-12.12	-9.44	-2.90	0.69	1.71	2.77	4.30
500	-18.57	-14.92	-12.15	-9.44	-2.94	0.67	1.68	2.73	4.27
1000	-18.83	-15.31	-12.44	-9.60	-2.94	0.67	1.69	2.70	4.17
∞	-18.91	-15.26	-12.37	-9.53	-2.93	0.66	1.67	2.65	4.11

Table 7.1 Empirical fractiles of $T(\hat{\psi}_{ad2} - 1)$ for $\psi = 1$ ($\mathrm{AR}_\mu(1)$).

T/Fr.	0.01	0.025	0.05	0.10	0.50	0.90	0.95	0.975	0.99
25	1.37	1.42	1.49	1.52	1.45	1.70	2.16	2.86	4.03
50	1.60	1.58	1.61	1.59	1.43	1.63	2.01	2.58	3.75
75	1.65	1.67	1.64	1.62	1.43	1.58	1.95	2.54	3.45
100	1.69	1.66	1.67	1.64	1.43	1.57	1.91	2.42	3.37
250	1.72	1.72	1.73	1.67	1.43	1.52	1.85	2.32	3.24
500	1.77	1.77	1.72	1.68	1.42	1.51	1.84	2.31	3.23
1000	1.77	1.76	1.74	1.69	1.43	1.52	1.84	2.29	3.14
∞	1.82	1.78	1.74	1.69	1.43	1.51	1.83	2.24	3.10

Table 7.2 Differences between the empirical fractiles of $T(\hat{\psi}_{ad2} - 1)$ and $T(\hat{\psi}_{ml} - 1)$ for $\psi = 1$ ($\mathrm{AR}_\mu(1)$).

The fractiles for the iterated estimate, or for $T(\hat{\psi}_{ad2,i} - 1)$, are tabulated in Table 7.3, and the corresponding deviations from the fractiles for $T(\hat{\psi}_{ml} - 1)$ are laid out in Table 7.4.[6] The fractiles lie to the right of those of $T(\hat{\psi}_{ad2} - 1)$ due to the parallel shift in the distribution of the estimate of c (the shift was pointed out in Section 6.3 for the asymptotic case but it arises in finite samples, too). The fractiles up to the median move left-wards as T increases as do those of $T(\hat{\psi}_{ml} - 1)$ and $T(\hat{\psi}_{ad2} - 1)$. The movement is towards the right for the fractiles to the right of the median contrary to the behaviour

[5]The fractiles 0.025, 0.05, and 0.10 do not quite fit the general formation between the case of T equal to 1 000 and the asymptotic case. The fractiles for the former case are possibly overestimates because *i)*: Otherwise the asymptotic fractiles would fit the pattern and *ii)* A simulation experiment with the same seed but an approximation of Brownian motion by a random walk of length 10 000 does not alter the asymptotic fractiles much, so the asymptotic fractiles appear approximately correct. A simple reason for the overestimates may be Monte Carlo random error.

[6]The fractiles have been calculated from the same same random numbers which were used to construct Table 7.1. The simulation experiment was similar in other respects, too.

with the $T(\hat{\psi}_{ml} - 1)$ or $T(\hat{\psi}_{ad2} - 1)$ statistics. Moreover, the expansion to the right is fortified as T is increased. The medians for $T(\hat{\psi}_{ad2,i} - 1)$ and $T(\hat{\psi}_{ml} - 1)$ depart from each other by approximately 1.9, or the medians of $\hat{\psi}_{ad2,i}$ and $\hat{\psi}_{ml}$ lie about $1.9/T$ apart. The differences between the distributions are greatest at the right tail as with the $T(\hat{\psi}_{ad2} - 1)$ statistic, but here the departure multiplies as T increases. The difference increases also at the left tail but not by as much as at the right tail. Only the medians of the distributions approach each other somewhat as T is increased. The fractiles suggest that the variance of $T(\hat{\psi}_{ad2,i} - 1)$ is larger than that of $T(\hat{\psi}_{ml} - 1)$ and the more so the larger T is. The asymptotic distribution has not been reached by $T = 1000$. This is so because of the just referenced divergence of the fractiles with increasing T.[7]

The simulation results for the standardised *a priori* adjusted coefficient are documented in Table 7.5. The median is positive for small T unlike in the asymptotic case when the median is zero (within random error). The small-sample distributions can be also skewed unlike the asymptotic distribution: the right tail of the distribution is clearly longer than the left for T less than 100. The tails are reasonably balanced for T over 100, though.[8]

The distributions of $\hat{\psi}_{ml}$, $\hat{\psi}_{ad2}$, and $\hat{\psi}_{ad2,i}$ for $T = 100$ are illustrated in Figures 7.1, 7.2, and 7.3, respectively.[9] The histograms and the associated Monte Carlo statistics

---

[7]A reason for the increase in the departure could be that the term $\hat{c}_i$ $(\psi - \hat{\psi}_{ml})$ is more likely to dominate the adjusted likelihood if the sample is small than if it is large. This would lead to complex roots of the adjusted likelihood equation, and hence to $\hat{\psi}_{ad2,i} \equiv 1$ more often with a small, than a large, sample. In the latter case, a large $\hat{\psi}_{ad2}$ would be relatively more likely to arise, stretching the right tail relative to the small-sample case. This might explain as well the sudden lengthening of the extreme right-tail fractiles as T tends to infinity.

A similar phenomena could be in effect when estimating $\hat{\psi}_{ad2}$, though the fractiles behave in an opposite fashion in that case. The probability of complex roots is tiny when c is evaluated at $\hat{\psi}_{ml}$ so the phenomena does not play an appreciable role in the formation of the distribution.

[8]This simulation exercise does not enable us to comment on Rothenberg (1995) who argued that the statistic (5.15) should be approximately unbiased for small samples. The present adjustment and the one by Rothenberg agree asymptotically but not necessarily when T is finite. (Of course, the bias cannot be read from the fractiles either.)

[9]The scale of the graphs is fixed for the AR(1) models. The graphs for the AR(2) model share a separate scale because very different distributions can arise under this model (Figures 7.8–7.11).

T/Fr.	0.01	0.025	0.05	0.10	0.50	0.90	0.95	0.975	0.99
25	-15.11	-12.50	-10.35	-8.11	-2.15	1.41	2.89	4.39	6.31
50	-16.92	-13.67	-11.23	-8.68	-2.35	1.55	3.25	5.23	7.88
75	-17.50	-14.03	-11.50	-8.86	-2.40	1.61	3.34	5.54	8.81
100	-17.82	-14.34	-11.63	-9.00	-2.39	1.68	3.44	5.74	9.26
250	-18.23	-14.61	-11.90	-9.17	-2.46	1.69	3.59	6.12	10.94
500	-18.41	-14.74	-11.92	-9.18	-2.50	1.72	3.70	6.60	12.29
1000	-18.67	-15.13	-12.21	-9.34	-2.51	1.77	3.85	6.70	12.76
∞	-18.74	-15.06	-12.14	-9.26	-2.50	1.91	4.13	7.99	20.01

Table 7.3 Empirical fractiles of $T(\hat{\psi}_{ad2,i} - 1)$ for $\psi = 1$ ($\mathrm{AR}_\mu(1)$).

T/Fr.	0.01	0.025	0.05	0.10	0.50	0.90	0.95	0.975	0.99
25	1.48	1.57	1.67	1.76	1.93	2.12	2.86	3.77	4.94
50	1.73	1.74	1.81	1.85	1.88	2.35	3.34	4.73	6.69
75	1.77	1.86	1.85	1.88	1.88	2.41	3.42	5.06	7.66
100	1.86	1.84	1.89	1.90	1.87	2.49	3.54	5.29	8.16
250	1.87	1.89	1.94	1.93	1.87	2.52	3.73	5.68	9.87
500	1.93	1.96	1.95	1.94	1.86	2.56	3.85	6.18	11.25
1000	1.93	1.94	1.97	1.96	1.86	2.62	4.00	6.29	11.72
∞	1.99	1.98	1.97	1.96	1.86	2.76	4.29	7.59	19.00

Table 7.4 Differences between the empirical fractiles of $T(\hat{\psi}_{ad2,i} - 1)$ and $T(\hat{\psi}_{ml} - 1)$ for $\psi = 1$ ($\mathrm{AR}_\mu(1)$).

T/Fr.	0.01	0.025	0.05	0.10	0.50	0.90	0.95	0.975	0.99
25	-4.70	-3.59	-2.74	-1.86	0.47	3.35	4.70	6.13	8.10
50	-4.58	-3.52	-2.72	-1.92	0.25	2.55	3.55	4.61	6.14
75	-4.49	-3.49	-2.73	-1.95	0.17	2.31	3.22	4.18	5.37
100	-4.48	-3.48	-2.72	-1.93	0.13	2.21	3.07	3.92	5.04
250	-4.45	-3.50	-2.74	-1.96	0.05	2.05	2.83	3.62	4.60
500	-4.40	-3.43	-2.71	-1.96	0.03	2.00	2.77	3.56	4.51
1000	-4.48	-3.45	-2.74	-1.96	0.01	1.97	2.74	3.47	4.43
∞	-4.37	-3.45	-2.71	-1.94	0.00	1.94	2.69	3.41	4.33

Table 7.5 Empirical fractiles of $T(\hat{\psi}_{ad2,ap} - 1)$ for $\psi = 1$ ($\mathrm{AR}_\mu(1)$).

exemplify the above remarks.[10] It can be seen that the distributions of the AEs are more symmetric than those of the MLE.[11] The means of $\hat{\psi}_{ad2}$ and $\hat{\psi}_{ad2,i}$ (0.964 and 0.970, respectively) lie in between the mean of the MLE with constant (0.948) and the mean of the MLE of the simple AR(1) model of Section 6.2 (0.983 — not reported in the figures). Hence $\hat{\psi}_{ad2}$ and $\hat{\psi}_{ad2,i}$ are less biased than $\hat{\psi}_{ml}$. They are also more accurate in terms of MSE but their variance is larger than that of $\hat{\psi}_{ml}$ (at least for this sample size). The AE beats the MLE again by about 30 per cent in terms of MSE (according to the less rounded nonreported figures). The variance of $\hat{\psi}_{ad2,i}$ is especially inflated by a long right tail (the MSE of $\hat{\psi}_{ad2,i}$ is larger than the MSE of $\hat{\psi}_{ad2}$ though they are the same to three decimal places).

Another point of interest is that $\hat{\psi}_{ad2}$ equaled one, or the adjusted likelihood had complex roots (and $\hat{c}> 0$), only in about 0.02 per cent of the replications.[12] The simulations confirm that case *iv*) of p. 81 or the asymptotic formula for $\hat{\psi}_{ad2}$ is extremely likely to apply if $\psi = 1$, at least if the sample size is reasonable. The proportion of complex roots was higher at about 1.8 per cent when constructing the iterated estimate which is reflected in an increase of the height of the bar at unity in Figure 7.3.

Figure 7.4 and the accompanying statistics illuminate the distribution of $\hat{c}$ when $T = 100$. Both the median and the mean lie very far from the theoretical value of c equal to $(100-1)/2 = 49.5$. Furthermore, $\hat{c}$ faces substantial variation. The distribution is right skewed, too, for all the sample sizes examined. Pleasingly, there were no occurrences of a negative $\hat{c}$ neither with this exercise nor with those reported in Table 7.1.[13] Furthermore,

---

[10]The problem of extreme values of $\hat{\psi}_{ad2,i}$ and consequent rocketing of (Monte Carlo) moments in the asymptotic case has disappeared with $T = 100$. Here exaggerated values of $\hat{c}_i$ are likely to lead to complex roots and to $\hat{\psi}_{ad2,i}= 1$ (by definition) eliminating extreme values. As in the asymptotic case we did not enter the iteration if the original adjusted likelihood equation (6.16) had no real roots but defined $\hat{c}_i= \hat{c}$ and $\hat{\psi}_{ad2,i}=\hat{\psi}_{ad2}= 1$ (cf. definition (6.31)).

[11]The distribution of $\hat{\psi}_{ad2,i}$ would emerge even slightly more symmetric if $\hat{\psi}_{ad2,i}$ were defined to equal $\hat{\psi}_{ml}$ instead of unity in the case of a complex root. The same argument applies in principle to the distribution of $\hat{\psi}_{ad2}$, too, but in practice the effect would be minor.

[12]The proportion was 0.4 per cent when $T = 25$.

[13]A simulation exercise with $T = 5$ showed that it is quite unlikely to come across a $\hat{c}< 0$ under $\psi = 1$

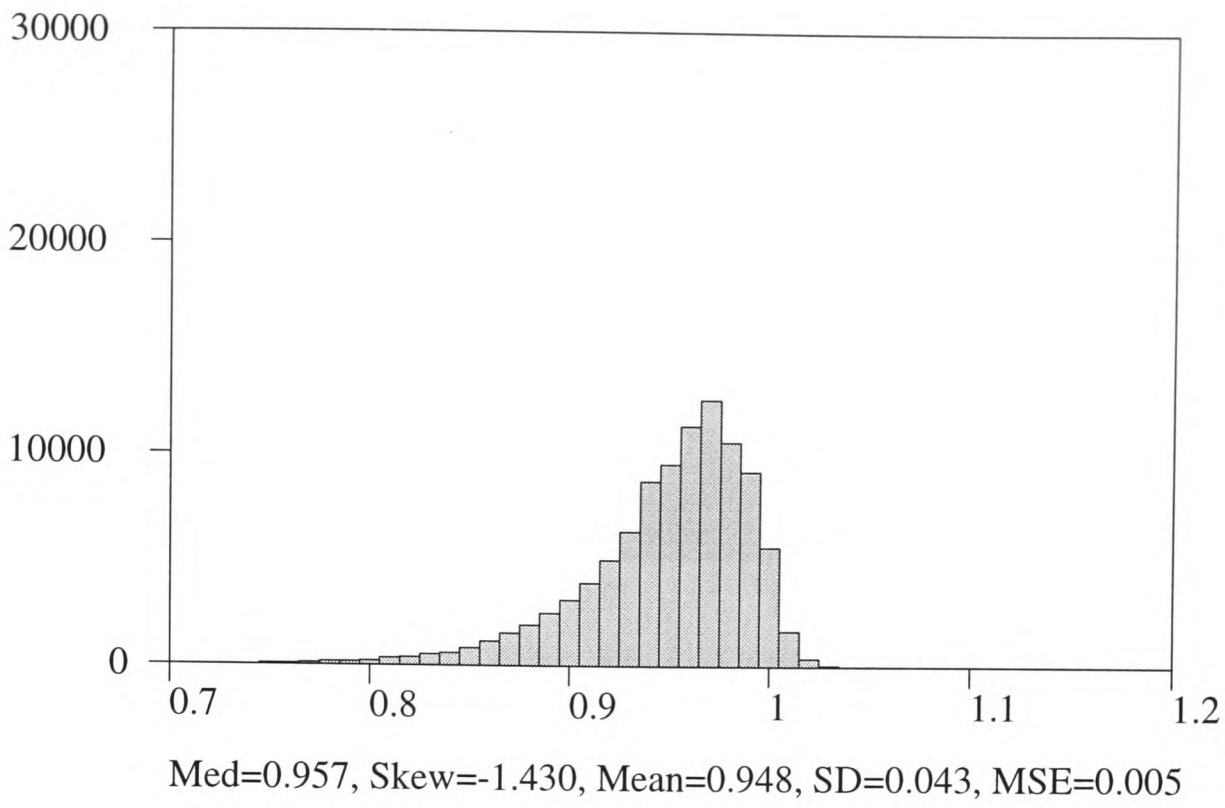


Figure 7.1 Empirical distribution of $\hat{\psi}_{ml}$ for $\psi = 1$ and $T = 100$ ($\mathrm{AR}_{\mu}(1)$).

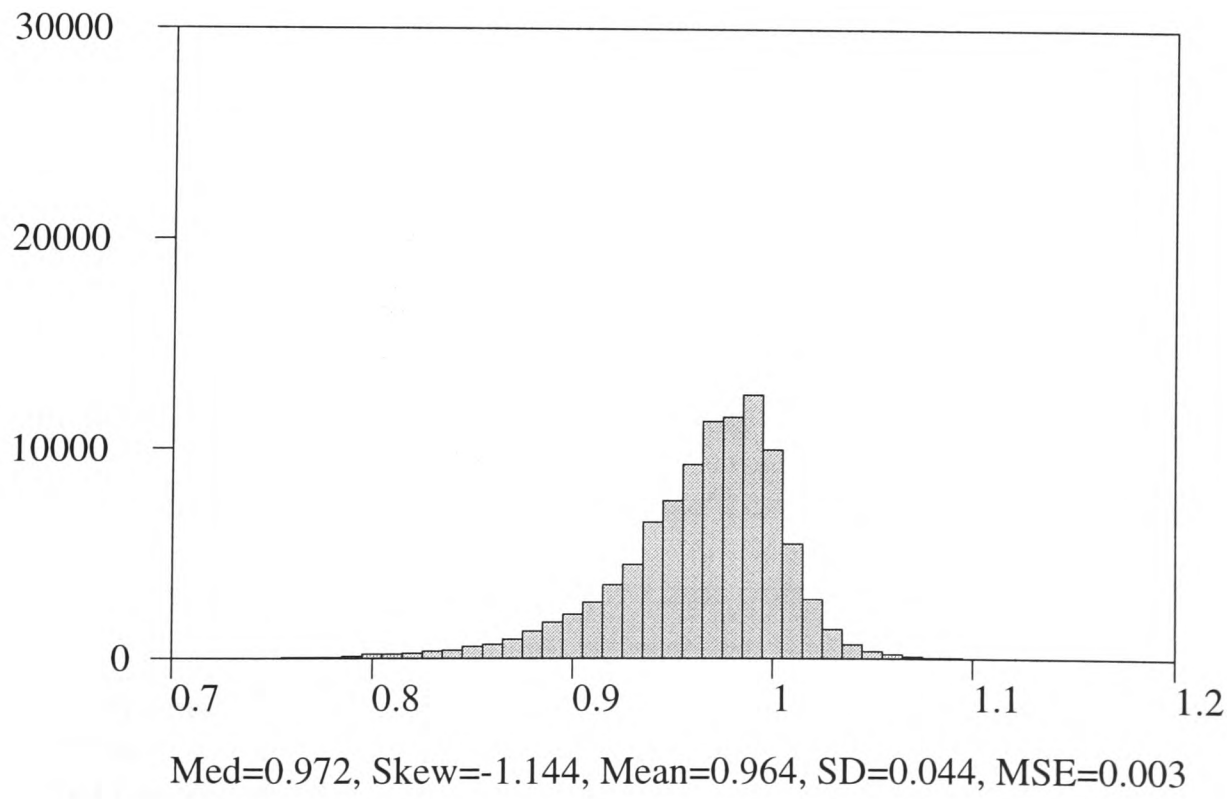


Figure 7.2 Empirical distribution of $\hat{\psi}_{ad2}$ for $\psi = 1$ and $T = 100$ ($\mathrm{AR}_{\mu}(1)$).

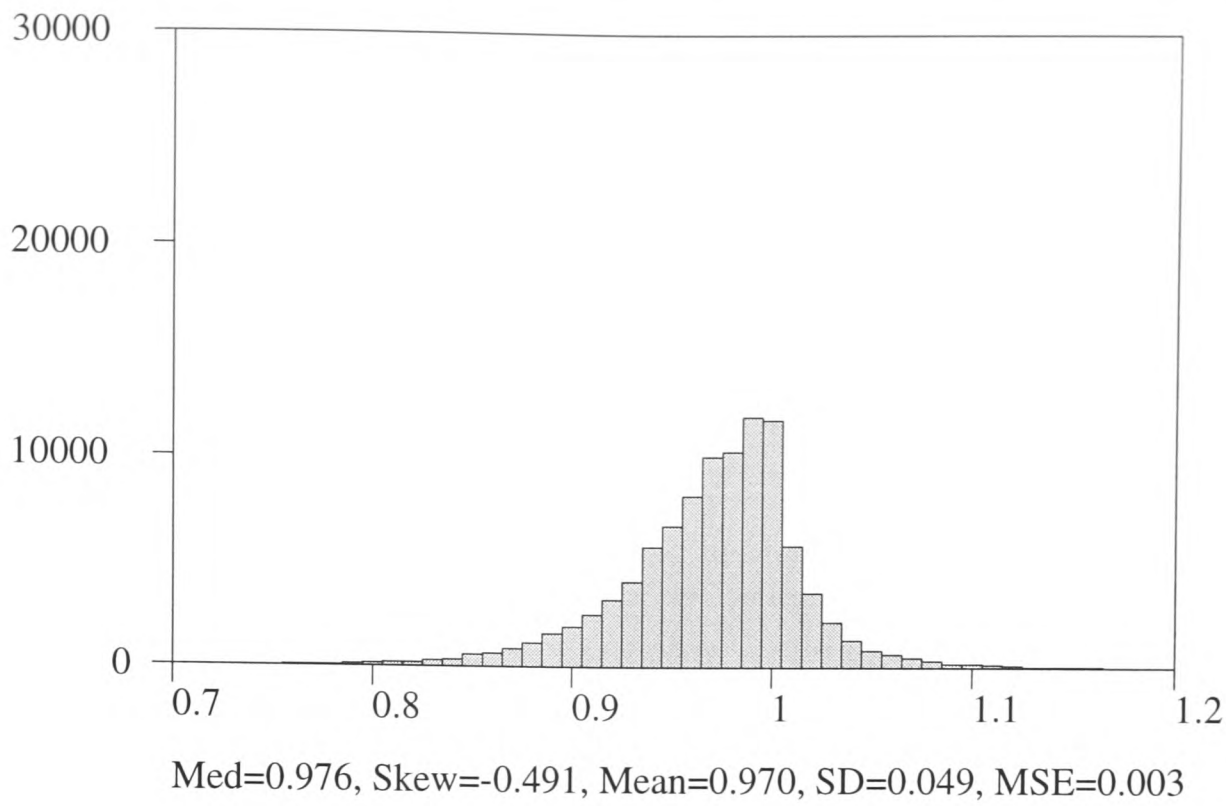


Figure 7.3 Empirical distribution of $\hat{\psi}_{ad2,i}$ for $\psi = 1$ and $T = 100$ ($\mathrm{AR}_\mu(1)$).

the shape of this small-sample distribution agrees with the asymptotic quite well (cf. Figure 6.7).[14]

The distribution of $\hat{c}_i$ is shifted to the right and has a longer right tail than the distribution of $\hat{c}$ (no figure is presented). The general pattern is the same, though. Also this distribution matches with the asymptotic counterpart of it quite well.[15]

The joint distribution of $T(\hat{\psi}_{ml} - 1)$ and $T(\hat{\psi}_{ad2} - 1)$ is plotted in Figure 7.5. The straight line tracks the points were the statistics would be equal. The general pattern resembles greatly the asymptotic one (Figure 6.9). However, a few points (21 out of the 100 000 draws) stand out from the figure. These are the sporadic occurrences where the adjusted likelihood equation had no real roots and the AE was defined to be one. The $\hat{c}\,(\psi - \hat{\psi}_{ml})$ term appears to be able to dominate the adjusted log-likelihood (leading to complex roots) only when the MLE is larger than one and $\hat{c}$ is consequently sizable.[16]

---

even with extremely small sample sizes: a negative estimate occurred in less than 0.2 per cent of the (10 000) replications.

[14]The estimated c exceeded 100 in about 0.3 per cent of the draws which agrees with the corresponding probability applying asymptotically (p. 90). (The scale of the horizontal axis in Figure 7.4 does not cover all of the variability of $\hat{c}$.) Complex roots appeared only in about 7 per cent of these cases so an excessive $\hat{c}$ does not necessarily lead to complex roots.

[15]The Monte Carlo median is 23.549 which agrees well (after division by 100) with the asymptotic one (0.231) reported in Section 6.3. The Monte Carlo moments are senseless due to the extreme values.

[16]The complex roots which appeared in the simulation with $T = 25$ transpired also solely with MLEs larger than one. Complex roots could possibly emerge with MLEs smaller than one if T were extremely

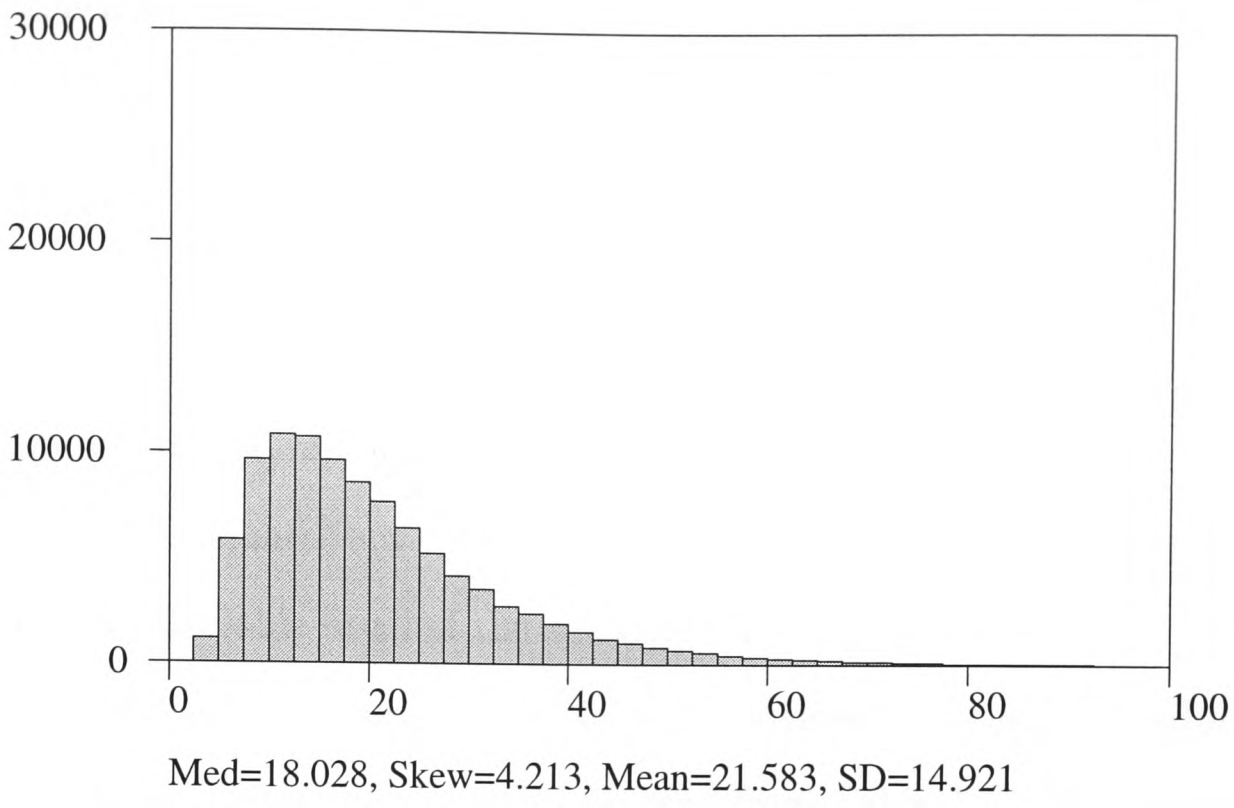


Figure 7.4 Empirical distribution of the data based $\hat{c}$ for $\psi = 1$ and $T = 100$ ($\mathrm{AR}_\mu(1)$).

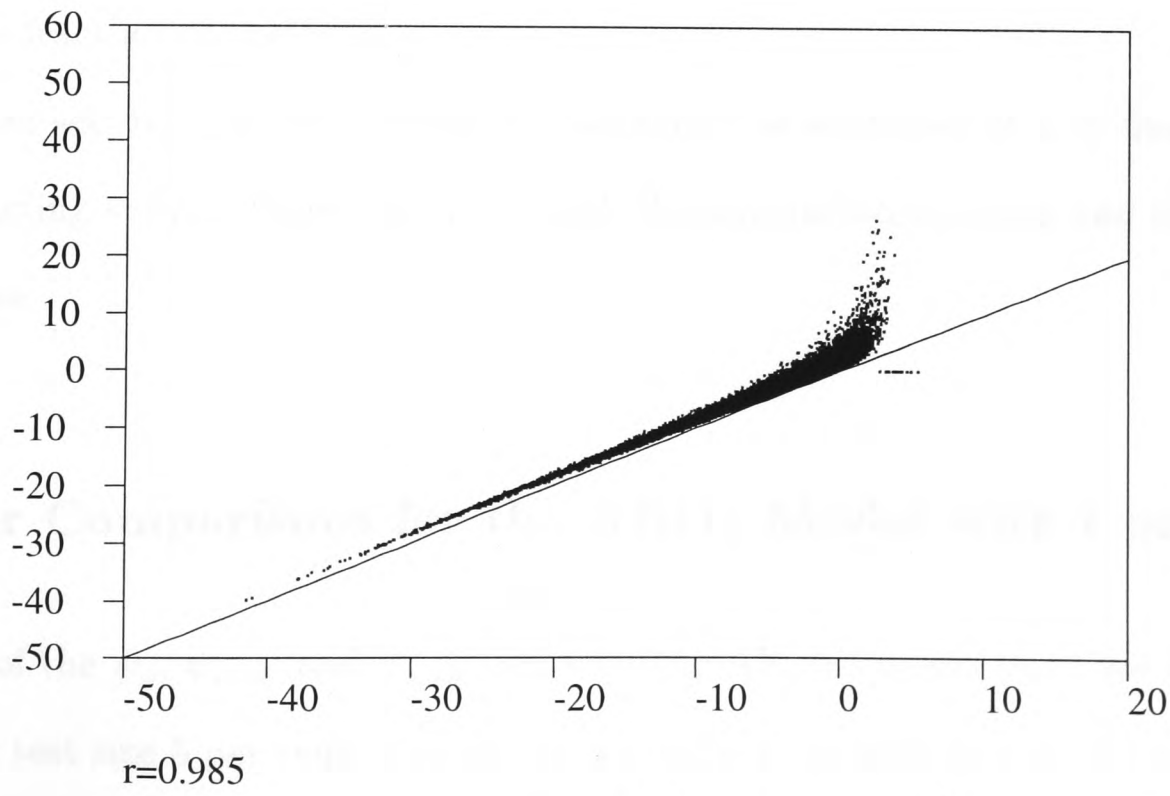


Figure 7.5 Empirical joint distribution of $T(\hat{\psi}_{ml} - 1)$ and $T(\hat{\psi}_{ad2} - 1)$ for $\psi = 1$ and $T = 100$ ($\mathrm{AR}_\mu(1)$).

Fractiles of the (data-based) adjusted Wald statistic are recorded in Table 7.6. The median of the distribution lies to the left of zero. In general, the right-tail fractiles, excluding the 0.99^{th} fractile but including the median, decrease and the left-tail fractiles increase or the distribution becomes more concise as T increases. Comparison to the fractiles of the ordinary Wald statistic would reveal that the adjustment has shifted the distribution uniformly to the right and especially so at the right tail. This is in line with the discussion on the asymptotic distribution (6.39) in Section 6.3. We note also that the argument of the square-root term of the formula (6.36) for $\sqrt{W_{ad2}}$ was nonnegative and the statistic calculable in all of the simulations.

The conciseness and fair symmetry of the distributions of the *a priori* adjusted Wald statistic for even quite small samples can be read from Table 7.7. The fractiles become even denser as T increases. The median is slightly positive for finite samples. The argument of the square-root term of the formula for $\sqrt{W_{ad2,ap}}$ was always nonnegative and the statistic calculable except when the samples were composed of 25 observations, in which case the term was negative for 0.02 per cent of the simulated time series.[17] Finally, we note that the fractiles of the Wald statistics are less affected by the sample size than the fractiles of the SCs.

In the next section, we shall discuss the behaviour of estimates of ψ in finite samples when the starting value is large relative to both the unconditional mean and the variance of the process.

7.3 Power Comparisons for the AR(1) Model with Constant

The powers of the ψ_μ, $\psi_{\mu,ad}$, and $\psi_{\mu,ad,i}$ tests for the $\mathrm{AR}_\mu(1)$ model (6.4) are reported in Table 7.8 for test size 5 per cent, ψ in the range 0.4 to 0.99, sample size (T) in the range

---

small. The role of the term $-[(T-3)/2]\log\hat{\sigma}^2_\psi$ would then be cut back relative to the term $\hat{c}\,(\psi - \hat{\psi}_{ml})$ when maximizing the adjusted profile log-likelihood.

[17]The statistic was appointed the value 10^6 in these cases so that the proportion of complex-root cases could be easily calculated. The impact on the estimated fractiles can only be minor.

T/Fr.	0.01	0.025	0.05	0.10	0.50	0.90	0.95	0.975	0.99
25	-3.37	-2.95	-2.60	-2.22	-0.98	0.44	0.94	1.37	1.80
50	-3.28	-2.88	-2.57	-2.21	-1.03	0.39	0.90	1.38	1.99
75	-3.24	-2.86	-2.56	-2.22	-1.05	0.37	0.89	1.38	2.01
100	-3.22	-2.86	-2.56	-2.21	-1.05	0.37	0.89	1.36	1.99
250	-3.17	-2.83	-2.55	-2.22	-1.07	0.34	0.85	1.35	2.00
500	-3.17	-2.83	-2.53	-2.21	-1.08	0.33	0.85	1.35	1.97
1000	-3.17	-2.84	-2.55	-2.23	-1.08	0.33	0.84	1.34	1.98
∞	-3.16	-2.82	-2.54	-2.21	-1.07	0.33	0.85	1.33	1.98

Table 7.6 Empirical fractiles of $\sqrt{W_{ad2}}$ for $\psi = 1$ ($\mathrm{AR}_\mu(1)$).

T/Fr.	0.01	0.025	0.05	0.10	0.50	0.90	0.95	0.975	0.99
25	-1.83	-1.39	-1.06	-0.71	0.14	0.84	1.11	1.36	1.68
50	-1.71	-1.33	-1.04	-0.73	0.08	0.80	1.09	1.35	1.66
75	-1.67	-1.32	-1.04	-0.73	0.05	0.79	1.07	1.33	1.65
100	-1.65	-1.31	-1.04	-0.73	0.04	0.77	1.05	1.31	1.62
250	-1.63	-1.31	-1.03	-0.73	0.02	0.75	1.03	1.30	1.61
500	-1.63	-1.29	-1.02	-0.73	0.01	0.74	1.02	1.29	1.60
1000	-1.62	-1.28	-1.03	-0.73	0.00	0.74	1.02	1.28	1.60
∞	-1.58	-1.28	-1.01	-0.73	0.00	0.74	1.02	1.28	1.59

Table 7.7 Empirical fractiles of $\sqrt{W_{ad2,ap}}$ for $\psi = 1$ ($\mathrm{AR}_\mu(1)$).

25 to 250, and $|y_0 - \mu| / \sigma$ in the range 0 to 250.[18] The quantity $|y_0 - \mu| / \sigma$ is the absolute value of the deviation from the standardised unconditional mean where $\mu = \alpha/(1-\psi)$ is the asymptotic mean of the process y_t and σ is the standard deviation of the innovation.[19] The blank entries in the table indicate that power equal to one (to three decimal places) was achieved already with a larger ψ than the ψ corresponding to the entry. We shall scrutinize first the power of the $\psi_{\mu,ad}$ test and then the power of the $\psi_{\mu,ad,i}$ test relative to the ψ_μ test.

The $\psi_{\mu,ad}$ test is slightly weaker than (or as powerful as) the ψ_μ test if the process has started from the unconditional mean.[20] Already a relatively small deviation in the

[18]The standard error of the power estimates in Table 7.8 was inspected and found to be less than 0.5 for all records. The same seed and hence random numbers were used for each experiment with an identical value of ψ to aid comparison. The power simulations employed critical values with three decimals instead of the rounded ones in Tables 7.1, 7.3, and 7.5.

[19]Deviation $-(y_0 - \mu)/\sigma$ would produce a similar but not exactly the same estimate of power as $(y_0 - \mu)/\sigma$. An exactly coinciding estimate would require reversing also the sign of the pseudo random innovations. The reason is that the statistics considered in this section remain numerically unchanged after such a reversal (cf. Davidson and MacKinnon (1993, p. 752), say).

[20]The figures for $\psi = 0.99$, $|y_0 - \mu| / \sigma = 0$, and $T = 25$ contradict the statement. However, a

starting value from the mean balances the powers, or converts the rank ordering in favour of the $\psi_{\mu,ad}$ test, cf. the sector for $T = 250$ and $T = 100$ under $|y_0 - \mu|/\sigma = 5$, respectively, in Table 7.8. If the departure is more substantial then the $\psi_{\mu,ad}$ test can be considerably more powerful, cf. the sector for $T = 25$ and $|y_0 - \mu|/\sigma = 5$ or $|y_0 - \mu|/\sigma = 25$ for the cases of $\psi = 0.5$ and $\psi = 0.6$ or the sector for $T = 100$ and $|y_0 - \mu|/\sigma = 25$ for the cases of $\psi = 0.85$ and $\psi = 0.9$, say.

An increasing divergence from the mean has a consistent impact on the powers of the tests for sample sizes up to 100, at least. The powers of both tests decrease with the divergence in the starting value for ψ larger than the threshold 0.5/0.85 and increases for ψ smaller than the threshold when the sample size is 25/100. The decrease is smaller and the increase is larger with the $\psi_{\mu,ad}$ test. The tests become biased and eventually completely powerless as the divergence becomes big enough when ψ is larger than the threshold. Relatedly, the power of the ψ_{μ} test at the threshold decreases first with the deviation but increases as the divergence becomes larger. The $\psi_{\mu,ad}$ test features monotonically increasing power at the threshold as the deviation increases. The outcome is that the power difference in favour of the $\psi_{\mu,ad}$ test peaks for ψ close to the threshold.[21]

A partial explanation for the above is that under a large $|y_0 - \mu|/\sigma$, the time series can be very autocorrelated, delusively suggesting $I(1)$ness, if ψ is large, but that the time series converges quickly, or even jumps to the mean, if ψ is smaller, in which case lots of evidence is gained on the magnitude of ψ. The threshold increases with T because a larger ψ is then needed to slow down stabilisation to the mean, which would reveal stationarity. The explication does not capture all aspects, though. A time series composed of e.g. 25 observations, an autoregressive coefficient 0.6, and which has started far from the mean, converges quickly to the mean level and reveals at least to the eye the nature of the set-up despite the lack of power of the tests. We shall resolve below why the tests are

simulation experiment with 100 000 replications indicates that the powers are the same in this case. To confirm the power difference for other values a further experiment with 100 000 replications was done for $\psi = 0.95$, $|y_0 - \mu|/\sigma = 0$, and $T = 100$. The power estimates were not changed essentially or the difference in power persisted.

[21]The above properties apply also for ψ in the span 0 to 0.3 and for the sample size $T = 50$ (for a different threshold) though these values are not included in the Table 7.8.

powerless under such cases.

A large starting value favours the $\psi_{\mu,ad}$ test when the sample size is 250 but otherwise the above patterns cannot be found. For example, a different threshold arises for the tests as the deviation becomes large (cf. the sector for $T = 250$ and $|y_0 - \mu|/\sigma = 250$; the power figure 87.2 would be replaced by 100.0 if the starting value equaled 1 000). Consequently, there are cases where the power of the ψ_μ test collapses to zero but the power of the $\psi_{\mu,ad}$ test ameliorates to one. The result suggests that a different threshold might be found already for the sample sizes 25 and 100 if a finer grid for ψ were used, and we shall confirm this below. Similar results are obtained when the size of the test is 1 per cent.[22]

The power of the $\psi_{\mu,ad,i}$ test behaves in essentially the same fashion as the power of the $\psi_{\mu,ad}$ test but the characteristics are magnified. Thus if the process starts from the mean then the power loss is moderately increased in comparison to the $\psi_{\mu,ad}$ test. On the other hand if the starting value does not match with the mean then the power gain over the ψ_μ test is multiplied. Combinations $\psi = 0.6$ and $T = 25$, $\psi = 0.9$ and $T = 100$, and $\psi = 0.95$ and $T = 250$ in the zone for $|y_0 - \mu|/\sigma = 25$ provide examples.

Interestingly, the power of the tests does not necessarily increase monotonically as ψ decreases if the starting value deviates enough from the mean (cf. the sectors for $T = 25$ and $T = 100$ with $|y_0 - \mu|/\sigma = 25$). This is actually intuitive because the tests are designed to keep their size for $\psi = 1$ regardless of the starting value so the power may remain closer to the size with a ψ very close to one than with a smaller ψ for which the power is negligible.

Example (Table 7.8). If $\psi = 0.6$, $T = 25$, and $|y_0 - \mu|/\sigma = 5$ then the powers ($*100$) of the ψ_μ, $\psi_{\mu,ad}$, and $\psi_{\mu,ad,i}$ tests are 40.9, 48.6, and 49.8, respectively. The corresponding figures for $\psi = 0.9$, $T = 100$, and $|y_0 - \mu|/\sigma = 25$ are 6.9, 19.8, and 22.4, respectively.□

[22]The thresholds move downward as the size of the tests is decreased to 1 per cent. The difference in the thresholds of the tests when $T = 250$ disappears — to the grid accuracy in Table 7.8 — as the threshold of the ψ_μ test remains the same but the threshold of the $\psi_{\mu,ad}$ test becomes 0.9.

A finding related to White (1961) explains the emergence of thresholds and why they differ across tests and sample sizes (and risk levels). White (*op. cit.*) proved the result, anticipated by Hurwicz (1950, p. 373), that if the constant is not estimated then the MLE becomes unbiased under (asymptotic) stationarity as the starting value diverges to infinity (as mentioned in Section 5.3). (See Appendix A5 for related results and references.) We prove in Appendix A5 that actually the MLE, the AE, and the AE_{ap} *converge* stochastically to the true value for $|\psi| < 1$ as the absolute value of the deviation of the starting value from the unconditional mean tends to infinity. (We also prove this for the MLE when the model does not encompass a constant and derive the asymptotic distributions of $y_0(\hat{\psi}_{ml} - \psi)$ for the models.) Some Monte Carlo experimentation suggests that also the iteratively calculated AE is similarly consistent. The connection to the thresholds is best explained by means of examples.

Example (Table 7.8). Let T equal 25 and ψ equal 0.6 or 0.5. Then the statistics $T(\hat{\psi}_{ml} - 1)$, $T(\hat{\psi}_{ad2} - 1)$ and $T(\hat{\psi}_{ad2,i} - 1)$ converge to $25*(0.6-1) = -10.0$, or to $25*(0.5-1) = -12.5$, respectively, as the deviation $|y_0 - \mu|/\sigma$ tends to infinity. The critical values for the statistics are -12.03, -10.54, and -10.35, respectively (by an unreported simulation experiment and Tables 7.1 and 7.3). The critical values lie between -10.0 and -12.5 so the test statistics do not reach the critical values under $\psi = 0.6$ but overshoot the critical values under $\psi = 0.5$ for a large enough $|y_0 - \mu|/\sigma$. A consequence is that the tests share a common rejection threshold at $\psi = 0.5$ for $T = 25$ (cf. the zones for T equal to 25 and $|y_0 - \mu|/\sigma$ equal to 250 in Table 7.8).□

Example (Table 7.8). Let T equal 250 and ψ equal 0.99, 0.95 or 0.9. Then the statistics $T(\hat{\psi}_{ml} - 1)$, $T(\hat{\psi}_{ad2} - 1)$ and $T(\hat{\psi}_{ad2,i} - 1)$ converge to $250 * (0.99 - 1) = -2.5$, $250 * (0.95 - 1) = -12.5$ or to $250*(0.9-1) = -25.0$, respectively, as the deviation $|y_0 - \mu|/\sigma$ tends to infinity. The critical value (-13.84 by an unreported simulation experiment) associated with the ψ_μ test is reached only by ψ equal to 0.9 so the test rejects the null of a unit root always under $\psi = 0.9$ but never for ψ equal to or larger than 0.95 (when $|y_0 - \mu|/\sigma$ is large enough). Instead, the critical values (-12.12 and -11.90 by Tables

7.1 and 7.3) related to the $\psi_{\mu,ad}$ and $\psi_{\mu,ad,i}$ tests are realized already by ψ equal to 0.95 but for ψ equal to 0.99 the critical values remain distant (for a large enough $|y_0 - \mu|/\sigma$). Different rejection thresholds at $\psi = 0.9$ and at $\psi = 0.95$ emerge for the tests (cf. the zones for T equal to 250 and $|y_0 - \mu|/\sigma$ equal to 250 in Table 7.8).□

By similar reasoning one could *a priori* locate a threshold of power for any of the above tests at an arbitrary sample size and risk level and spot more cases where the thresholds would differ for the tests.

A power comparison of the Wald tests τ_μ and $\tau_{\mu,ad}$ is given in Table 7.9.[23] The basic pattern is the same as above but the powers of the two tests differ less than the powers of the corresponding ψ_μ and $\psi_{\mu,ad}$ tests do. The adjusted version is still slightly less powerful if the process has started from the unconditional mean. If $|y_0 - \mu|/\sigma$ is five, then the powers are very much the same. As the deviation increases the adjusted version gains a power advantage but it is not as a profound one as what took place with the tests based on SCs.

A quick look at the column for $T = 25$ might suggest that thresholds for power would emerge again. The power of the tests seems to tend to zero at $\psi = 0.99$ as the starting value increases towards 25 when $T = 25$. However, the power decline of the $\tau_{\mu,ad}$ test converts into a power increase by $|y_0 - \mu|/\sigma = 250$ which throws the threshold in doubt. Indeed, a simulation experiment with $|y_0 - \mu|/\sigma = 1000$ (not reported in the table) gave rejection rates 60.3 and 70.3 for the τ_μ and $\tau_{\mu,ad}$ tests, respectively, so powers of both tests seem to converge to one implying no thresholds of the kind which arose with the ψ_μ, $\psi_{\mu,ad}$, and $\psi_{\mu,ad,i}$ tests. By studying the other two columns, it can be seen that the power

[23]Literally, we are studying the statistics $\sqrt{W}$ and $\sqrt{W_{ad2}}$ which employ the MLE of σ instead of the degrees-of-freedom corrected estimate s. The corresponding statistics differ only by a multiplying constant so the test outcomes and powers agree exactly.

The simulation algorithm accepts the null of a unit root if it comes across a noncalculable $\sqrt{W_{ad2}}$. Another alternative would be to exclude such instances which could only increase the power estimates for the $\tau_{\mu,ad}$ test. Our choice is probably more fair as in practise one does not throw away a time series under inspection only because of a difficulty in the calculation of the test statistic. Anyway, the practise does not matter according to the following check: The power estimates (*100) for the $\tau_{\mu,ad}$ test remain the same to two decimal places when the algorithm is changed to reject the null in a noncalculable case of $\sqrt{W_{ad2}}$ and the sample size is 25 and $|y_0 - \mu|/\sigma$ is 0 or 5.

of the τ_μ and $\tau_{\mu,ad}$ tests tends to one for all experimented values of T as the deviation $|y_0 - \mu|/\sigma$ increases. The power boost is not necessarily uniform for the tests, as just recognized.

Example (Table 7.9). The largest difference in power (*100) in favour of the τ_μ test is 0.6 and occurs in the zones for $|y_0 - \mu|/\sigma = 0$ at $\psi = 0.50$ and $T = 25$ (50.9 vs. 50.3) and at $\psi = 0.95$ and $T = 250$ (45.5 vs. 44.9). The largest difference in power (*100) in favour of the $\tau_{\mu,ad}$ test is 3.5 (77.0 vs. 80.5) which takes place at $\psi = 0.9$, $T = 25$, and $|y_0 - \mu|/\sigma = 25$.□

Tables 7.8 and 7.9 affirm the Monte Carlo results of Dickey *et al.* (1986, Table 2), Dickey and Fuller (1979, p. 430), and Schmidt and Phillips (1992): the ψ_μ test is more, and can be much more, powerful than the τ_μ test when the process has started from the unconditional mean, cf. the sectors for $|y_0 - \mu|/\sigma$ equal to zero in Tables 7.8 and 7.9.[24] However, the zones for $|y_0 - \mu|/\sigma$ equal to or larger than 25 reveal that the τ_μ test (or $\tau_{\mu,ad}$ test) can be far more powerful than the ψ_μ test (or $\psi_{\mu,ad}$ test) otherwise. In the case of $|y_0 - \mu|/\sigma = 5$ the power ranking depends on the sample size so that the τ_μ test (or $\tau_{\mu,ad}$ test) is more powerful than the ψ_μ test (or $\psi_{\mu,ad}$ test) for $T = 25$ or when $|y_0 - \mu|/\sigma$ is relatively notable and vice versa for $T = 250$ or when $|y_0 - \mu|/\sigma$ is relatively less notable.[25]

Example (Tables 7.8 and 7.9). When $\psi = 0.6$, $T = 25$, and $|y_0 - \mu|/\sigma = 0$ the powers (*100) of the ψ_μ and τ_μ tests are 54.4 and 34.8, respectively. The power of the former test decreases to 40.9 and the power of the latter test increases to 80.3 as $|y_0 - \mu|/\sigma$ is changed to 5. As $|y_0 - \mu|/\sigma$ is altered to 250 the power of the ψ_μ test fades to zero while the power of the τ_μ test reaches 100.□

[24]The simulation results of Dickey *et al.* (*op. cit.*) and Schmidt and Phillips (*op. cit.*) agree with the present estimates of power (for $T = 100$ and $|y_0 - \mu|/\sigma = 0$ in the case of the former article and for $T = 100$ and $|y_0 - \mu|/\sigma = 0$ or $|y_0 - \mu|/\sigma = 5$ in the case of the latter article). The power estimates of Dickey and Fuller (*op. cit.*) are not straightforwardly comparable as they employed a two-sided test.

[25]The result might be seen for the ψ_μ and τ_μ tests from Table 2 of Schmidt and Phillips (*op. cit.*) but they do not comment upon it. An exception to the rule is the case of $\psi = 0.99$, $|y_0 - \mu|/\sigma = 5$ and $T = 25$.

An interpretation of the differences in power between the SC tests and the Wald-kind of tests is simple. As already explained, standardised statistics like $T(\hat{\psi}_{ml} - 1)$ converge to $T(\psi - 1)$ or to a fixed number as $|y_0 - \mu|/\sigma$ tends to infinity which leads to an inconsistent test for many combinations of ψ and T. The Wald test can be interpreted as a departure of the null value of ψ (here unity) from the MLE multiplied by a measure of the curvature of the likelihood function (or the profile likelihood as in (4.4)), cf. Buse (1982), say. The departure tends to $\psi - 1 \neq 0$ by the consistency of the estimates with an increasing $|y_0 - \mu|/\sigma$. It is reasonable to assume that simultaneously the (possibly adjusted) profile likelihood develops an increasingly curved shape around the MLE. Hence the multiplicative factor in the formula for the Wald-family of tests accelerates forcing the Wald statistic to explode with an increasing $|y_0 - \mu|/\sigma$ inducing a consistent test for all ψ and T.

It seems fair to reason as is often done (cf. Section 3.5) that the (DF) unit-root tests are weak. For example, a researcher employing the ψ_μ test would make a type II error almost one out of seven times even when the autoregressive coefficient is as low as 0.40, $|y_0 - \mu|/\sigma$ equals zero, and the sample size is 25 (Table 7.8).

Quite extraordinary results arise for the *a priori* adjusted versions of the tests or $\psi_{\mu,ad,ap}$ and $\tau_{\mu,ad,ap}$ tests (Table 7.10). Both tests display uniformly close to zero power when $|y_0 - \mu|/\sigma = 0$. The intuition is the inconsistency of the *a priori* AE as rationalized on p. 88 and the aforementioned tendency for the power to stick to the size of the test in the neighbourhood of unity. The drop in power is in line, even though an extreme occurrence, with the previously observed pattern: if the process has started from the mean then the power of a test decreases when the adjustment intensifies. The power increases rapidly by $|y_0 - \mu|/\sigma = 5$ — except for $\tau_{\mu,ad,ap}$ under $T = 250$ — because, as we have pointed out, the AE_{ap} is consistent when $|y_0 - \mu|/\sigma$ tends to infinity. The emerging power appears a mixture of the depressing effect due to the inconsistency with the sample size and the gain in strength due to the departure of the starting value from the mean. For the zones below, the pattern of power matches again with the other two

tables: the powers increase in line with the strength of the adjustment (minor exceptions exist to this rule). The tests possess then power which is in some cases unachievable by the corresponding previously considered tests.

The power ranking of the *a priori* adjusted versions of the tests is simple though it does not follow the design laid out. The $\psi_{\mu,ad,ap}$ test is more powerful as long as $|y_0-\mu|/\sigma$ is five or smaller. The $\tau_{\mu,ad,ap}$ test is more powerful than the $\psi_{\mu,ad,ap}$ test for $|y_0-\mu|/\sigma$ equal to or larger than 25. The test performs very well already when $|y_0-\mu|/\sigma$ is 25 and it rejects the null of a unit root with probability close to one for all of the experimented values of ψ and T by $|y_0-\mu|/\sigma$ equal to 250.

Example (Table 7.10). The contrast in power between the $\psi_{\mu,ad,ap}$ and $\tau_{\mu,ad,ap}$ tests can be extreme. Two cases in point are $\psi = 0.7$, $T = 250$, $|y_0-\mu|/\sigma = 5$ and $\psi = 0.95$, $T = 25$, $|y_0-\mu|/\sigma = 250$ which imply departures in power (*100) close to 100. The departure is in favour of the former test in the first case (99.0 vs. 0.0) and in favour of the latter test in the second case (0.0 vs. 100.0).□

The materializing thresholds can again be explained by the convergence of the estimate, here $\hat{\psi}_{ad2,ap}$, to ψ as $|y_0-\mu|/\sigma$ tends to infinity. (The power figures 39.3 and 15.9 in the $|y_0-\mu|/\sigma = 250$ zone tend to 100.0 and 0.0, respectively, as $|y_0-\mu|/\sigma$ increases.)

Example (Table 7.10). Assume that the number of observations is 25 and that ψ equals 0.85 or 0.9. $T(\hat{\psi}_{ad2,ap}-1)$ tends then to $25(0.85-1) = -3.75$ or to $25(0.9-1) = -2.5$, respectively. When these values are compared to the 5 per cent critical value -2.74 (from Table 7.5) it can be seen that the limiting value under ψ equal to 0.85 leads to a rejection which does not happen under ψ equal to 0.9.□

		T								
		25			100			250		
$\lvert y_0 - \mu\rvert/\sigma$	ψ	ψ_μ	$\psi_{\mu,ad}$	$\psi_{\mu,ad,i}$	ψ_μ	$\psi_{\mu,ad}$	$\psi_{\mu,ad,i}$	ψ_μ	$\psi_{\mu,ad}$	$\psi_{\mu,ad,i}$
0	**0.99**	5.4	5.5	5.5	7.4	7.4	7.4	11.4	11.2	11.2
	0.95	7.8	7.8	7.7	19.3	18.6	18.4	63.9	62.6	62.2
	0.90	11.5	11.2	11.0	47.8	46.3	46.0	99.4	99.4	99.4
	0.85	15.4	14.8	14.7	79.3	78.2	78.0	100.0	100.0	100.0
	0.80	20.6	19.8	19.6	95.8	95.2	95.1			
	0.70	35.8	34.2	33.9	100.0	100.0	100.0			
	0.60	54.4	52.7	52.2						
	0.50	72.0	70.4	69.9						
	0.40	86.3	85.2	85.0						
5	**0.99**	5.3	5.2	5.2	6.9	6.9	6.8	10.4	10.4	10.4
	0.95	5.2	5.3	5.3	14.5	14.8	15.0	62.3	62.2	62.0
	0.90	4.3	4.8	4.9	42.5	43.6	43.9	99.6	99.5	99.5
	0.85	4.4	5.5	5.8	78.6	79.6	79.8	100.0	100.0	100.0
	0.80	7.1	8.9	9.3	96.7	96.9	97.0			
	0.70	17.5	22.4	23.4	100.0	100.0	100.0			
	0.60	40.9	48.6	49.8						
	0.50	68.1	75.1	76.4						
	0.40	89.5	92.8	93.3						
25	**0.99**	2.7	2.8	2.8	1.0	1.0	1.1	1.1	1.4	1.5
	0.95	0.0	0.0	0.0	0.1	0.2	0.3	44.8	59.3	62.1
	0.90	0.0	0.0	0.0	6.9	19.8	22.4	100.0	100.0	100.0
	0.85	0.0	0.0	0.0	82.5	95.6	96.6			
	0.80	0.0	0.0	0.0	100.0	100.0	100.0			
	0.70	0.0	0.1	0.1						
	0.60	1.9	28.2	35.6						
	0.50	73.3	99.0	99.5						
	0.40	100.0	100.0	100.0						
250	**0.99**	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
	0.95	0.0	0.0	0.0	0.0	0.0	0.0	0.0	87.3	96.3
	0.90	0.0	0.0	0.0	0.0	0.0	0.0	100.0	100.0	100.0
	0.85	0.0	0.0	0.0	100.0	100.0	100.0			
	0.80	0.0	0.0	0.0						
	0.70	0.0	0.0	0.0						
	0.60	0.0	0.0	0.0						
	0.50	100.0	100.0	100.0						
	0.40									

Table 7.8 Empirical power ($\times 100$) of the ψ_μ, $\psi_{\mu,ad}$, and $\psi_{\mu,ad,i}$ tests when size is 5 %.

		T					
		25		100		250	
$\lvert y_0 - \mu\rvert/\sigma$	ψ	τ_μ	$\tau_{\mu,\text{ad}}$	τ_μ	$\tau_{\mu,\text{ad}}$	τ_μ	$\tau_{\mu,\text{ad}}$
0	**0.99**	5.3	5.3	6.1	6.0	7.6	7.5
	0.95	6.1	6.0	11.5	11.3	45.5	44.9
	0.90	7.0	6.9	30.9	30.5	96.8	96.7
	0.85	8.6	8.4	62.1	61.4	100.0	100.0
	0.80	11.4	11.2	87.0	86.7		
	0.70	20.8	20.4	99.7	99.6		
	0.60	34.8	34.3	100.0	100.0		
	0.50	50.9	50.3				
	0.40	70.0	69.5				
5	**0.99**	4.9	4.9	5.7	5.7	7.9	7.8
	0.95	6.3	6.2	14.0	13.8	51.4	51.1
	0.90	8.6	8.7	42.7	42.4	98.7	98.6
	0.85	14.3	14.5	78.5	78.2	100.0	100.0
	0.80	24.6	24.8	96.7	96.6		
	0.70	51.6	52.0	100.0	100.0		
	0.60	80.3	80.4				
	0.50	95.0	95.0				
	0.40	99.3	99.3				
25	**0.99**	4.4	4.4	5.0	5.2	11.1	11.6
	0.95	12.2	14.1	87.7	89.2	99.9	99.9
	0.90	77.0	80.5	100.0	100.0	100.0	100.0
	0.85	99.8	99.8				
	0.80	100.0	100.0				
	0.70						
	0.60						
	0.50						
	0.40						
250	**0.99**	3.4	5.0	96.5	98.0	100.0	100.0
	0.95	100.0	100.0	100.0	100.0		
	0.90						
	0.85						
	0.80						
	0.70						
	0.60						
	0.50						
	0.40						

Table 7.9 Empirical power ($\times 100$) of the τ_μ and $\tau_{\mu,ad}$ tests when size is 5 %.

		T					
		25		100		250	
$\lvert y_0 - \mu\rvert/\sigma$	ψ	$\psi_{\mu,\text{ad,ap}}$	$\tau_{\mu,\text{ad,ap}}$	$\psi_{\mu,\text{ad,ap}}$	$\tau_{\mu,\text{ad,ap}}$	$\psi_{\mu,\text{ad,ap}}$	$\tau_{\mu,\text{ad,ap}}$
0	**0.99**	4.7	4.5	3.4	1.9	1.1	0.2
	0.95	3.3	1.8	0.1	0.0	0.0	0.0
	0.90	0.3	1.3	0.0	0.0	0.0	0.0
	0.85	0.4	0.1	0.0	0.0	0.0	0.0
	0.80	0.1	0.0	0.0	0.0	0.0	0.0
	0.70	0.0	0.0	0.0	0.0	0.0	0.0
	0.60	0.0	0.0	0.0	0.0	0.0	0.0
	0.50	0.0	0.0	0.0	0.0	0.0	0.0
	0.40	0.0	0.0	0.0	0.0	0.0	0.0
5	**0.99**	4.8	4.7	4.2	2.7	2.5	0.7
	0.95	7.0	6.3	7.9	1.1	3.9	0.0
	0.90	14.1	11.2	22.1	1.0	20.0	0.0
	0.85	26.5	19.3	49.5	1.6	59.6	0.0
	0.80	41.7	29.1	76.1	2.8	88.8	0.0
	0.70	71.9	50.9	97.1	9.9	99.0	0.0
	0.60	92.0	71.5	99.6	25.7	99.8	0.0
	0.50	98.1	85.2	99.9	50.8	99.9	0.4
	0.40	99.6	93.1	100.0	74.6	100.0	4.3
25	**0.99**	6.0	9.0	11.1	19.1	26.4	37.2
	0.95	5.9	57.2	92.4	99.8	100.0	100.0
	0.90	31.9	99.2	100.0	100.0		
	0.85	91.6	100.0				
	0.80	100.0					
	0.70						
	0.60						
	0.50						
	0.40						
250	**0.99**	0.0	39.3	0.0	100.0	15.9	100.0
	0.95	0.0	100.0	100.0		100.0	
	0.90	0.1					
	0.85	100.0					
	0.80						
	0.70						
	0.60						
	0.50						
	0.40						

Table 7.10 Empirical power ($\times 100$) of the $\psi_{\mu,ad,ap}$ and $\tau_{\mu,ad,ap}$ tests when size is 5 %.

We note a peculiar outcome: The ψ_μ test, say, is inconsistent for many combinations of ψ and $|y_0 - \mu|/\sigma$ though the estimate on which it is based, $\hat{\psi}_{ml}$, is consistent whether the number of observations or $|y_0 - \mu|/\sigma$ tends to infinity. On the other hand, the $\psi_{\mu,ad,ap}$ test, say, is consistent under many of the combinations of ψ and $|y_0 - \mu|/\sigma$ under which the ψ_μ test is inconsistent even though the estimate $\hat{\psi}_{\mu,ad,ap}$ underlying the $\psi_{\mu,ad,ap}$ test is in general consistent only as $|y_0 - \mu|/\sigma$ tends to infinity. The explanation lies in the more compact distribution of $\hat{\psi}_{\mu,ad,ap}$ over that of $\hat{\psi}_{ml}$.

A suggestion for the use of tests, conditional on the magnitude of $|y_0 - \mu|/\sigma$ or the standardised departure of the starting value from the unconditional mean and derived from Tables 7.8, 7.9, and 7.10, is outlined in Table 7.11.[26] The standard ψ_μ test is favoured when the process has started relatively close from the unconditional mean or $|y_0 - \mu|/\sigma$ is small relative to the sample size, the $\psi_{\mu,ad,i}$ test when the starting value deviates relatively somewhat more from the unconditional mean, and the $\psi_{\mu,ad,ap}$ and $\tau_{\mu,ad,ap}$ tests when the deviation is large or extreme. The most powerful test depends on both the number of observations and $|y_0 - \mu|/\sigma$ but not (or barely) on the autoregressive coefficient. No single test dominates the others. A further point which merits comment is that even though a Wald test often is more powerful than the corresponding SC test there is usually another SC test which is even more powerful for the circumstance.

Table 7.12 captures essentially the information in Table 7.11 but more neatly in terms of $|y_0 - \mu|/T\sigma$. A difference is that we have substituted the $\psi_{\mu,ad}$ test in place of the ψ_μ and $\psi_{\mu,ad,i}$ tests for the case $|y_0 - \mu|/T\sigma = 5$ for the sake of simplicity. Following this rule rather than the more complicated one in Table 7.11 would have very little effect on the power. The simple pattern in the table invites us to choose the test conditional on the value of $|y_0 - \mu|/T\sigma$ in general. However, we have not inspected whether the pattern remains the same when $|y_0 - \mu|/\sigma$ and/or T lie outside the range of experimented values or more generally do not match with them.

The above recommendation is quite theoretical because *i)* the starting value relative to

[26]There is a circumstance for which the recommendation does not point out the most powerful test: $\psi = 0.99$ and $|y_0 - \mu|/\sigma = 5$ in which case the ψ_μ test is the most powerful.

$\lvert y_0 - \mu\rvert/\sigma$	Suggested test
0	ψ_μ
5	ψ_μ for $T = 250$ $\psi_{\mu,ad,i}$ for $T = 100$ $\psi_{\mu,ad,ap}$ for $T = 25$
25–250	$\tau_{\mu,ad,ap}$

Table 7.11 Suggested test for different values of $|y_0 - \mu|/\sigma$ for $|y_0 - \mu|/\sigma$ in the range [0,250] and T in the range [25,250] ($AR_\mu(1)$).

$\lvert y_0 - \mu\rvert/T\sigma$	Suggested test
0	ψ_μ
0.02–0.05	$\psi_{\mu,ad}$
≥ 0.1	$\psi_{\mu,ad,ap}$ or $\tau_{\mu,ad,ap}$

Table 7.12 Suggested test for different values of $|y_0 - \mu|/T\sigma$ for $|y_0 - \mu|/\sigma$ in the range [0,250] and T in the range [25,250] ($AR_\mu(1)$).

the mean and the variance is seldom known *ii)* if one knew that the deviation is definitely non-zero (and one is using the tests properly) then one would also know that the process is stationary. (The present unit-root tests are valid only when $y_0 - \mu = 0$ under a unit root). A Bayesian might want to impose an *a priori* distribution on $|y_0 - \mu|/T\sigma$ and choose the test on the basis of it but such a distribution may be perplexing to determine. Less ambitiously, Table 7.12 can be helpful also if one has an opinion on the probable size of $|y_0 - \mu|/T\sigma$ under the alternative hypothesis or is concerned with an approximate range of values of it. The latter possibility appears especially relevant as such a divergence can lead to a notable decrease in power as we have seen.

The above described configuration of power may be somewhat informative when assessing the magnitude of $|y_0 - \mu|/\sigma$. For example, let us suppose that the τ_μ test rejects the null while the ψ_μ test does not and the sample size is less than or equal to one hundred. Even though this is not an exact test, the outcome is evidence in favour of a stationary process which has started from a large starting value (in absolute terms): a large starting value would shrink the power of the latter test for many values of ψ (i.e. for the values of ψ which are greater than the sample and the test specific thresholds discussed above) and a boost in power for the Wald-type of tests. Such an outcome is only suggestive, because, among other things, we have not investigated correlations

between the test statistics. Calculating the $\psi_{\mu,ad,ap}$ or the $\tau_{\mu,ad,ap}$ test statistics can be useful, too because their power functions take extreme values enabling one to exclude numerous parameter combinations of ψ and $|y_0-\mu|/\sigma$. For example, rejection of the null by either of the tests is strong evidence against the plausibility of $|y_0-\mu|/\sigma=0$.

Already a battery of two tests, the ψ_μ and the $\tau_{\mu,ad}$ tests, say, would assure in general, though not always, competent power across the scrutinized $(\psi,|y_0-\mu|/\sigma)$ set. Even though the latter test does not appear to be the most powerful in any circumstance it performs well in general. If the circumstance $|y_0-\mu|/\sigma\approx 0$ were deemed unrealistic then a useful battery of tests would be $\tau_{\mu,ad}$ and $\tau_{\mu,ad,ap}$ which do not imply drastic losses of power for any of the experimented combinations of $(\psi,|y_0-\mu|/\sigma)$ with a relatively small $|y_0-\mu|/\sigma$ but achieve the highest powers for situations with a substantial $|y_0-\mu|/\sigma$. Of course, a problem is how to fix the size if two tests are employed.

If a sole test would have to be chosen for general use then the τ_μ or the $\tau_{\mu,ad}$ test would be a reasonable choice as they are the only tests which are consistent whether it is T or $|y_0-\mu|/\sigma$ which tends to infinity. An important drawback would be that quite large losses in power (relative to any of the SC tests) would arise when the process has started from the mean (cf. Tables 7.8 and 7.9). Empiricists often calculate only the Wald statistic. We have given some theoretical justification for this practise but pointed out a flaw with it, too. A single test (of the considered ones) is not capable of handling all the cases we have pondered.

An important recommendation is the usual one of plotting the data. A graph of a case under which the SC tests are powerless may well reveal that the process appears to converge to a constant level suggesting gradual stationarity.

7.4 Finite-Sample Distributions for the Unit-Root AR(1) Model with Drift

It was proved in Section 6.4 that the asymptotic distributions of the SCs and the corresponding Wald tests are Normal when the process is a random walk with drift. We

study in this section how well the results hold in finite samples and whether the adjusted statistics follow the asymptotic Normals better than their non-adjusted counterparts.

The distributions of the SCs are addressed in Table 7.13 (the seeds beneath the tables in this section are documented in Table A6.3 in Appendix A6). The entries were constructed by simulating the empirical fractiles, dividing them by $\sqrt{12\alpha^{-2}}$, and recording the deviations of these figures from the fractiles of the Standard Normal. The division by $\sqrt{12\alpha^{-2}}$ was done because the asymptotic variance of the SCs is $12(\sigma/\alpha)^2$ (formulae (6.41) and (6.45)) and the simulations assumed a σ equal to one throughout. An entry with a dash only indicates that the deviation was zero to two decimal places. The shorthand notation for each SC is used so that AE_i, say, points out that the fractiles relate to those of $T^{3/2}(\hat{\psi}_{ad2,i} - 1)$. Sample sizes 25, 100, and 250 and drifts equal to 0.1, 0.5, 1.0, and 2.5 are considered. The drift can be interpreted to be the constant relative to the standard deviation because of the just mentioned formula for the asymptotic variance.

In practice, one should use an estimated constant in place of the theoretical one when standardizing the distribution of the SCs. Hence the results on the distributions of the SCs may give too optimistic a view of the fit with the Standard Normal. On the other hand, the results should demonstrate how good a fit at best may be achieved when SCs are used.

We start by commenting on the zones with a drift equal to or larger than 0.5. Evidently, the finite sample records depend on the drift even though asymptotically the entries follow the Standard Normal distribution. Another outstanding feature is that the fractiles of the MLE lie uniformly left of the Standard Normal. The departure can be considerable but decreases when the drift or the sample size increases. The AE follows the Normal much better in general, especially at the more vital left tail (measured by the absolute value of the deviations). The left-tail fractiles tend to be yet too small whereas the right-tail fractiles overshoot. At the very right tail or mainly at the 0.975th and 0.99th fractiles, the fit of the AE can be worse.

The fit at the left tail may be considered more vital in two senses: *i*) If the Normal

approximation were to be used for testing of a unit root then most researches would probably be interested in alternatives with a smaller coefficient than larger and employ critical values from the left tail. *ii*) A departure of a fractile in the left tail implies a larger deviation in probabilistic terms than at the right tail when the true distribution is shifted to the left of the approximating Normal distribution used for inference.

The properties of the AE_i are those of the AE but amplified, replicating the pattern in Section 7.3. That is, the fit is in general somewhat better at the left tail but somewhat worse at the right tail compared to the AE.

The *a priori* AE gives a fairly good fit around the 0.05th fractile whenever the drift is 0.5 or larger. The fit is not always satisfying at the other fractiles but yet in general by far the best.

The information in the uppermost zone for a drift equal to 0.1 is counterintuitive. First of all, the left-tail fractiles of the MLE match surprisingly well already at the sample size 25. This appears to be a coincidence as the rest of the empirical fractiles miss the true marks badly and the fit becomes *worse* at the left tail as the sample size increases. Contrary to the previously commented cases, the distributions of the AE and the AE_i do not fit quite as well at the left tail but fit much better at the median and at the right tail than the distribution of the MLE. The AE_{ap} is the sole statistic which behaves sensibly in general or all of the fractiles of it approach those of the Normal as the number of observations increases. The medians and the right-tail fractiles of the other two AEs converge too, though. Two unreported simulation experiments revealed that these trends continue at least up to a sample size of 500 but that general convergence of the fractiles towards the Normal of the MLE, the AE, and the AE_i arises by a sample size of 1 000.

A conclusion from the table is that the sample size and especially the drift need to be substantial for the asymptotic Normal distribution to hold in practice (which is in line with the results of Hylleberg and Mizon (*op. cit.*) for the usual statistics). The AEs, particularly the *a priori* AE, achieve then a better fit at the more important left tail of the distributions.

The Wald statistics behave more consistently than the SCs, cf. Table 7.14. When the drift is small, or the asymptotics do not apply, the distributions are more symmetric and the deviations of the fractiles from the Standard Normal vary less than in Table 7.13. The deviations decrease in general, with minor exceptions, as the drift or the sample size increases albeit the convergence is slow when the drift is 0.1. Obviously, the finite-sample distributions depend on the size of the drift even though the asymptotic distribution is Standard Normal and hence free of nuisance parameters. The adjusted Wald statistic follows the Normal almost uniformly better than the usual Wald statistic. It contrasts with the case of the SCs for which the non-adjusted version (estimate) tended to fit better at the right tail than the adjusted version. $\sqrt{W_{ad2,ap}}$ follows the Normal even better in general except at the very right-tail fractiles under a drift equal to 0.1. Indeed, the match of the distribution of the $\sqrt{W_{ad2,ap}}$ is in general excellent whenever the drift is 0.5 or larger.

The general tendency at the left tail is that the SCs follow the Normal better than the Wald statistics when the drift (relative to the standard deviation) is 2.5 and *vice versa* when the drift is equal to or smaller than 0.5 (with the exception of the coincidental case of $\alpha = 0.1$ and $T = 25$). A drift equal to one is a borderline case for the standard statistics (MLE and W) but the better fit of the Wald statistics can be expanded to this case, too, for the adjusted statistics (AE and W_{ad2}). The Wald statistics follow the Normal about as well or better than the SCs at the right tail. These are rules of thumb; minor exceptions exist.

As a whole, the size of the drift is probably more important in determining the match with the Standard Normal: In the case of the SCs the Normal is not a sensible approximation when the drift is 0.1 and in the case of the Wald statistics the deviation from the Normal fades faster in the direction of an increased drift than in the direction of a larger sample. Indeed, the asymptotic local-drift theory (the drift α is devided by $T^{1/2}$ so that the drift fades) of Haldrup and Hylleberg (1994) suggests that the distribution would remain (approximately) the same when $\alpha T^{1/2}$ is fixed.

We state next a few results on the moments of the statistics which we have inspected only under a unit drift. (No table is presented.) Both the usual standardised MLE and the Wald statistic feature a negative mean but the adjusted versions are much less biased, indeed essentially unbiased in the case of the adjusted Wald statistics. Surprisingly, the least biased estimate is the AE instead of the AE_{ap}. The variance is the smallest for the MLE with a marginal difference to the AE_{ap}, then follow the AE and the AE_i. The variance of the *a priori* adjusted Wald statistic is the smallest of the Wald statistics. The variance of the adjusted Wald statistic is larger than that of the non-adjusted but the difference is very small. The AE and the AE_i are more skewed and feature more kurtosis than their non-adjusted counterparts. The AE_{ap} appears the best in these respects, too. On the other hand, the adjusted Wald statistics beat the usual ones in these respects. Having described these results, we point out that the match of the fractiles is more important than the match of the third or fourth moments as the former determine the location of the distribution on top of the shape of it.

Empirically a drift relative to the standard deviation is not likely to be as large as 2.5 but 1.0 is already plausible (cf. the discussion on p. 110). The adjusted statistics follow the Normal distribution quite well even at fairly small samples if the drift is of this size. The distribution of the adjusted Wald statistic can be very close to the Standard Normal already under a drift of 0.5. The *a priori* adjusted statistics, especially $\sqrt{W_{ad2,ap}}$, follow the Normal distribution very well even for small samples like 25 when the drift is 0.5. The statistic is in this sense most apt for testing of unit roots.

Example (Table 7.14). $\sqrt{W_{ad2}}$ deviates only by -0.12 from the Standard Normal at the 0.05th fractile when the drift is 0.5 and the number of observations is 100. The departure of the 0.05th fractile of the distribution of $\sqrt{W_{ad2,ap}}$ is 0.06 when the sample size is 25.□

		T											
		25				100				250			
Drift	**Fr.**	**MLE**	**AE**	**AE$_i$**	**AE$_{ap}$**	**MLE**	**AE**	**AE$_i$**	**AE$_{ap}$**	**MLE**	**AE**	**AE$_i$**	**AE$_{ap}$**
0.1	**0.01**	-0.03	0.17	0.19	1.65	-2.85	-2.37	-2.32	1.03	-4.98	-4.20	-4.11	0.37
	0.025	-0.03	0.17	0.20	1.44	-2.23	-1.76	-1.70	0.95	-3.75	-2.99	-2.89	0.42
	0.05	-0.07	0.15	0.18	1.25	-1.82	-1.36	-1.29	0.85	-2.87	-2.16	-2.04	0.42
	0.10	-0.10	0.12	0.15	1.01	-1.42	-0.97	-0.89	0.70	-2.08	-1.43	-1.30	0.38
	0.50	-0.55	-0.34	-0.27	0.07	-0.89	-0.54	-0.43	0.03	-0.91	-0.48	-0.34	0.01
	0.90	-1.35	-1.11	-1.05	-0.80	-1.32	-0.90	-0.65	-0.64	-1.16	-0.54	-0.11	-0.35
	0.95	-1.62	-1.31	-1.20	-0.98	-1.51	-0.96	-0.48	-0.77	-1.29	-0.47	0.43	-0.38
	0.975	-1.85	-1.44	-1.30	-1.08	-1.68	-0.96	-0.12	-0.85	-1.40	-0.29	1.35	-0.36
	0.99	-2.11	-1.53	-1.39	-1.16	-1.89	-0.84	0.57	-0.88	-1.52	0.08	3.27	-0.26
0.5	**0.01**	-4.34	-3.40	-3.26	-0.31	-0.91	-0.54	-0.51	-0.24	-0.46	-0.26	-0.25	-0.17
	0.025	-2.98	-2.10	-1.94	-0.07	-0.67	-0.34	-0.32	-0.12	-0.31	-0.12	-0.11	-0.05
	0.05	-2.02	-1.25	-1.11	0.07	-0.53	-0.22	-0.19	-0.05	-0.27	-0.08	-0.07	-0.02
	0.10	-1.30	-0.67	-0.55	0.13	-0.43	-0.13	-0.11	–	-0.23	-0.04	-0.04	–
	0.50	-0.64	-0.18	-0.06	0.07	-0.34	-0.02	–	0.01	-0.22	-0.01	–	–
	0.90	-0.79	0.04	0.18	0.15	-0.37	0.12	0.20	0.03	-0.22	0.05	0.06	0.01
	0.95	-0.83	0.32	0.56	0.39	-0.35	0.24	0.37	0.09	-0.21	0.11	0.12	0.04
	0.975	-0.83	0.75	1.05	0.75	-0.30	0.42	0.63	0.19	-0.18	0.16	0.18	0.08
	0.99	-0.79	1.52	1.72	1.45	-0.23	0.73	1.13	0.34	-0.14	0.25	0.28	0.13
1.0	**0.01**	-0.83	-0.47	-0.45	-0.21	-0.26	-0.10	-0.10	-0.05	-0.15	-0.05	-0.05	-0.03
	0.025	-0.60	-0.28	-0.27	-0.09	-0.21	-0.06	-0.06	-0.02	-0.13	-0.03	-0.03	-0.02
	0.05	-0.47	-0.16	-0.14	-0.01	-0.20	-0.05	-0.04	-0.01	-0.12	-0.02	-0.02	-0.01
	0.10	-0.38	-0.09	-0.07	0.02	-0.18	-0.03	-0.03	–	-0.11	-0.01	-0.01	–
	0.50	-0.32	–	0.02	0.03	-0.17	–	–	–	-0.11	–	–	–
	0.90	-0.34	0.16	0.23	0.08	-0.17	0.04	0.05	0.02	-0.11	0.01	0.02	–
	0.95	-0.31	0.32	0.44	0.17	-0.15	0.07	0.08	0.03	-0.10	0.03	0.03	0.01
	0.975	-0.24	0.53	0.72	0.29	-0.14	0.10	0.11	0.06	-0.09	0.04	0.05	0.03
	0.99	-0.13	0.94	1.30	0.53	-0.11	0.16	0.17	0.09	-0.07	0.06	0.06	0.04
2.5	**0.01**	-0.16	-0.04	-0.04	-0.01	-0.09	-0.03	-0.03	-0.02	-0.05	-0.01	-0.01	-0.01
	0.025	-0.15	-0.03	-0.03	-0.01	-0.08	-0.02	-0.02	-0.01	-0.04	–	–	–
	0.05	-0.14	-0.01	-0.01	0.01	-0.07	-0.01	-0.01	–	-0.05	-0.01	-0.01	-0.01
	0.10	-0.13	–	–	0.01	-0.07	-0.01	-0.01	–	-0.04	–	–	–
	0.50	-0.13	0.01	0.01	0.01	-0.07	–	–	–	-0.04	–	–	–
	0.90	-0.12	0.04	0.04	0.02	-0.06	0.01	0.01	0.01	-0.04	–	–	–
	0.95	-0.10	0.07	0.07	0.05	-0.06	0.02	0.02	0.01	-0.04	0.01	0.01	0.01
	0.975	-0.09	0.10	0.10	0.07	-0.06	0.02	0.02	0.01	-0.04	0.01	0.01	0.01
	0.99	-0.04	0.14	0.15	0.11	-0.06	0.02	0.02	0.01	-0.03	0.01	0.01	0.01

Table 7.13. Deviations of the standardised empirical fractiles of the SCs from the Standard Normal (unit-root $AR_\mu(1)$ with drift).

		T								
		25			100			250		
Drift	Fr.	$\sqrt{W}$	$\sqrt{W_{ad2}}$	$\sqrt{W_{ad2,ap}}$	$\sqrt{W}$	$\sqrt{W_{ad2}}$	$\sqrt{W_{ad2,ap}}$	$\sqrt{W}$	$\sqrt{W_{ad2}}$	$\sqrt{W_{ad2,ap}}$
0.1	**0.01**	-1.52	-1.02	0.42	-1.13	-0.81	0.45	-0.95	-0.66	0.33
	0.025	-1.47	-0.96	0.51	-1.16	-0.81	0.45	-1.00	-0.68	0.31
	0.05	-1.45	-0.93	0.52	-1.18	-0.82	0.43	-1.02	-0.68	0.29
	0.10	-1.44	-0.91	0.52	-1.22	-0.83	0.39	-1.05	-0.68	0.25
	0.50	-1.55	-0.93	0.14	-1.38	-0.85	0.04	-1.13	-0.60	0.01
	0.90	-1.56	-0.73	-0.40	-1.37	-0.54	-0.37	-1.10	-0.25	-0.24
	0.95	-1.53	-0.60	-0.49	-1.35	-0.36	-0.42	-1.09	-0.08	-0.28
	0.975	-1.49	-0.49	-0.55	-1.34	-0.18	-0.46	-1.08	0.12	-0.31
	0.99	-1.45	-0.47	-0.61	-1.34	0.07	-0.52	-1.09	0.42	-0.35
0.5	**0.01**	-1.11	-0.64	-0.06	-0.41	-0.16	-0.01	-0.24	-0.07	–
	0.025	-1.00	-0.53	0.01	-0.40	-0.14	–	-0.23	-0.06	–
	0.05	-0.94	-0.47	0.06	-0.38	-0.12	–	-0.24	-0.06	-0.01
	0.10	-0.88	-0.40	0.09	-0.38	-0.10	0.01	-0.23	-0.04	0.01
	0.50	-0.75	-0.19	0.07	-0.36	-0.03	0.01	-0.22	-0.01	–
	0.90	-0.68	0.02	-0.07	-0.35	0.08	-0.01	-0.22	0.04	–
	0.95	-0.64	0.09	-0.09	-0.34	0.13	-0.01	-0.21	0.06	0.01
	0.975	-0.62	0.12	-0.11	-0.34	0.17	-0.02	-0.21	0.07	–
	0.99	-0.57	0.15	-0.11	-0.33	0.24	-0.02	-0.20	0.10	0.01
1.0	**0.01**	-0.66	-0.27	-0.14	-0.24	-0.07	-0.03	-0.14	-0.03	-0.01
	0.025	-0.57	-0.20	-0.08	-0.23	-0.06	-0.02	-0.12	-0.02	–
	0.05	-0.51	-0.15	-0.03	-0.22	-0.05	-0.02	-0.13	-0.02	-0.01
	0.10	-0.45	-0.10	–	-0.20	-0.03	-0.01	-0.12	-0.01	–
	0.50	-0.35	–	0.03	-0.17	–	–	-0.11	–	–
	0.90	-0.28	0.09	–	-0.15	0.03	–	-0.10	0.01	–
	0.95	-0.23	0.13	0.01	-0.14	0.04	0.01	-0.09	0.02	0.01
	0.975	-0.19	0.18	0.03	-0.14	0.05	0.01	-0.09	0.02	–
	0.99	-0.13	0.25	0.06	-0.13	0.07	0.01	-0.08	0.03	0.01
2.5	**0.01**	-0.43	-0.15	-0.13	-0.14	-0.04	-0.04	-0.07	-0.02	-0.01
	0.025	-0.34	-0.09	-0.07	-0.12	-0.03	-0.03	-0.06	-0.01	–
	0.05	-0.29	-0.05	-0.03	-0.11	-0.02	-0.01	-0.06	-0.01	-0.01
	0.10	-0.24	-0.03	-0.01	-0.09	-0.01	–	-0.05	–	–
	0.50	-0.13	0.01	0.01	-0.07	–	–	-0.05	–	–
	0.90	-0.05	0.03	0.01	-0.05	0.01	0.01	-0.04	–	–
	0.95	–	0.04	0.03	-0.03	0.02	0.01	-0.03	0.01	0.01
	0.975	0.06	0.08	0.06	-0.03	0.02	0.01	-0.03	0.01	0.01
	0.99	0.14	0.13	0.10	-0.01	0.03	0.02	-0.01	0.03	0.02

Table 7.14. Deviations of the standardised empirical fractiles of the Wald statistics from the Standard Normal (unit-root $AR_{\mu}(1)$ with drift).

The empirical sizes calculated from the simulated distributions of Tables 7.13 and 7.14 are reported in Tables 7.15 and 7.16. The tables exemplify properties already commented on: the magnitude of the drift is more important in determining the match with the Standard Normal than the number of observations, especially the fit with the Standard Normal is in general bad when the drift is 0.1 or small (apart from the accidental case when $T = 25$); the adjusted statistics follow the Standard Normal much better than the standard counterparts; the Wald statistics follow the Normal better than the SCs when the drift equals one half, and *vice versa* when the drift is very large or over two. In the last respect, the differences are smaller for the adjusted statistics. The empirical size of the usual statistics can be unacceptable even under a unit drift but remains close to the nominal size when the statistics are adjusted.

	T								
	25			**100**			**250**		
Drift	**MLE**	**AE**	**AE$_{ap}$**	**MLE**	**AE**	**AE$_{ap}$**	**MLE**	**AE**	**AE$_{ap}$**
0.1	5.9	3.5	0.0	25.6	17.3	0.3	30.8	21.0	2.1
0.5	20.0	12.5	4.2	11.2	7.1	5.5	8.1	5.6	5.1
1.0	10.1	6.5	5.1	7.3	5.6	5.3	6.1	5.0	5.0
2.5	6.6	5.2	5.0	5.9	5.2	5.2	5.5	5.0	4.9

Table 7.15 Empirical size of the SC tests when nominal size is 5 % (unit-root AR$_\mu$(1) with drift).

	T								
	25			**100**			**250**		
Drift	$\sqrt{\mathbf{W}}$	$\sqrt{\mathbf{W}_{\mathbf{ad2}}}$	$\sqrt{\mathbf{W}_{\mathbf{ad2,ap}}}$	$\sqrt{\mathbf{W}}$	$\sqrt{\mathbf{W}_{\mathbf{ad2}}}$	$\sqrt{\mathbf{W}_{\mathbf{ad2,ap}}}$	$\sqrt{\mathbf{W}}$	$\sqrt{\mathbf{W}_{\mathbf{ad2}}}$	$\sqrt{\mathbf{W}_{\mathbf{ad2,ap}}}$
0.1	46.0	23.1	1.7	38.1	20.9	1.6	30.3	17.2	2.5
0.5	20.1	10.0	4.7	10.3	6.3	5.0	7.8	5.3	4.9
1.0	11.8	6.4	5.3	7.6	5.6	5.3	6.3	5.1	4.9
2.5	8.3	5.5	5.4	6.3	5.3	5.3	5.6	5.1	5.0

Table 7.16 Empirical size of the Wald tests when nominal size is 5 % (unit-root AR$_\mu$(1) with drift).

We expressed caveats on p. 110 on the applicability of the asymptotic normality results on testing for unit roots. The caveats were the poor fit of the empirical distributions with the Standard Normal, inconsistency against trend-stationary alternatives (as the sample size tends to infinity), and the incredibility of the stationary alternative

if a drifted random walk is considered the null. We suggested that the result might be more useful when the alternative is a gradually stationary process or a stationary process which has started from a value very far away from the unconditional mean but a trend-stationary process can be excluded from the set of alternative processes. The basis of our statement is that a process which has started well off the unconditional mean can be trended, not necessarily linearly, and may sometimes be a more reasonable alternative to the null than a stationary process (stationary apart from start-up effects). The suggestion gains some more merit from the fact that the asymptotic normality approximation is more realistic after adjusting the statistics, especially at the important left tail.

As a general note, asymptotic normality of the Wald statistics seems attractive because the Standard Normal is more compact than the DF distribution which results when a constant and a time trend are included in the model. This is the distribution to refer to when the null hypothesis is a drifted random walk (Section 3.2). Hence a test based on asymptotic normality should be more powerful than tests based on the DF distribution. As regards to SCs, all the contemplated estimates of ψ converge to the true value as the divergence of the starting value from the unconditional mean increases. When the null is a drifted random walk, the SC is multiplied, after subtracting one, by $T^{3/2}$ instead of T. It implies larger values, in absolute terms, of the SC statistics, which combined with smaller, in absolute terms, critical values indicates a more powerful test in general. We note in passing that the distribution of $\sqrt{W_{ad2,ap}}$ is the most compact but that does not necessarily imply a powerful test.

On the other hand, in small samples the adjusted statistics should display smaller power relative to the unadjusted as the former, or at least the AEs, are in general larger than the latter in absolute value and the same critical values from the Normal distribution are used for both kind of statistics. The power difference is not surprising after noting the many analogical cases in econometrics where the statistic with empirical size close to the nominal size is less powerful than another statistic which is less faithful to the nominal size (recall the argument of Blough on p. 25, say, on the relation between size

and power for unit-root tests). Stock (1994) reports simulation evidence on this issue in the context of testing for unit roots.

In conclusion, allowing for a gradually stationary process and the Cox–Reid adjustments expands somewhat the scope of functionality of unit-root tests based on asymptotic normality but probably by not as much as many econometricians would desire.

7.5 Fractiles for the Unit-Root Bhargava AR(1) Model with Constant

The (empirical) fractiles of the standardised MLE for the unit-root $\mathrm{AR}_{\mu}^{B}(1)$ model have not been studied or documented elsewhere so we report them on top of the empirical fractiles of the standardised AE (Tables 7.17 and 7.20). The fractiles apply regardless of the constant (γ_0) or starting value (x_0) as proved at the end of Section 6.6. There is extra uncertainty associated with the fractiles in this section because the estimates have been solved numerically, see Appendix A6 for details (especially Table A6.5).[27]

The fractiles for the MLE show, as expected, that the median is negative and the left tail of the distribution is longer than the right suggesting a biased estimate for all sample sizes. As the number of observations increases the fractiles become less scattered. They move essentially monotonically towards zero apart from the two left-most fractiles. (The fractiles for the asymptotic case relate to the previously simulated distribution portrayed in Figure 6.1.)

Comparison to the (unreported) distribution of the MLE under the simple AR(1) model of Section 6.3 yields a surprise (Table 7.18).[28] When the sample size is small the left tail fractiles of Table 7.17 lie far left of those of the distribution of the aforementioned MLE increasing the spread of the present MLE. Moreover, the distribution converges to

[27]The simulations employed the same seeds and random numbers which were used when analysing the $\mathrm{AR}_{\mu}(1)$ model (Table A6.1).

[28]Such a comparison is meaningful under a unit root and a zero starting value because then the processes models (6.4) and (6.47) generate agree (cf. formulae (6.11) and (6.48) and the discussion on p. 112).

the asymptotic one slowly. Thus the model does not yield as compact small-sample distributions as the asymptotic distribution suggests. The Bhargava model establishes invariance with respect to the starting value at the cost of inaccuracy in estimation in comparison to the simple AR(1) model.

The small-sample left tail fractiles of the present distribution fit better with the ones applying under the $AR_\mu(1)$ model (Table 7.19) than the ones which hold under the simple AR(1) model. (In large samples the fractiles naturally differ from those applying under the former model.) The variance of the present MLE appears to exceed that of the MLE under the $AR_\mu(1)$ model when the sample size is 25.[29] However, calculating the Monte Carlo moments reveals that the MLE based on the $AR_\mu^B(1)$ model is yet more accurate in terms of MSE (the MSEs of the MLEs are 0.044 and 0.060).

The adjustment has cut both tails of the distribution of the MLE but especially the left tail (Tables 7.20 and 7.21). Moreover, in essence all the fractiles of the distribution of the AE (excluding the asymptotic ones which have been copied from Table 7.17) lie closer both to zero and to the asymptotic fractiles than those for the MLE. That is, the AE appears less biased and less variable, and hence more accurate in terms of MSE than the MLE, and is found to follow the common asymptotic distribution better than the MLE.

[29]The north-east-most figure in Table 7.19 is positive which implies that the corresponding fractile of the present MLE lies to the left of the one associated with the MLE under the $AR_\mu(1)$ model. A negative number at the north-west-most figure indicates that the fractile of the present MLE lies further to the right.

T/Fr.	0.01	0.025	0.05	0.10	0.50	0.90	0.95	0.975	0.99
25	-17.01	-14.39	-12.07	-9.42	-1.01	1.09	1.51	1.91	2.45
50	-18.05	-14.76	-11.74	-8.44	-0.94	1.00	1.37	1.73	2.21
75	-18.38	-14.25	-11.10	-7.58	-0.90	0.97	1.34	1.69	2.14
100	-18.00	-13.79	-10.44	-7.10	-0.88	0.96	1.32	1.66	2.10
250	-16.00	-11.97	-9.02	-6.20	-0.86	0.95	1.30	1.64	2.06
500	-15.01	-11.27	-8.50	-5.95	-0.87	0.93	1.29	1.62	2.04
1000	-14.30	-10.86	-8.25	-5.84	-0.86	0.93	1.29	1.62	2.05
∞	-13.68	-10.41	-7.97	-5.65	-0.87	0.92	1.28	1.60	2.00

Table 7.17 Empirical fractiles of $T(\hat{\psi}_{ml} - 1)$ for $\psi = 1$ ($\mathrm{AR}_{\mu}^{B}(1)$).

T/Fr.	0.01	0.025	0.05	0.10	0.50	0.90	0.95	0.975	0.99
25	-5.09	-5.00	-4.71	-4.11	-0.22	0.07	0.09	0.11	0.11
50	-5.25	-4.91	-4.05	-2.93	-0.11	0.03	0.03	0.03	0.04
75	-5.40	-4.21	-3.28	-1.99	-0.07	0.01	0.01	0.02	0.02
100	-4.70	-3.55	-2.49	-1.41	-0.04	0.01	0.00	0.02	0.02
250	-2.58	-1.63	-1.05	-0.56	-0.00	0.02	0.01	0.02	0.01
500	-1.36	-0.87	-0.52	-0.22	-0.01	0.00	0.00	0.01	0.01
1000	-0.50	-0.38	-0.15	-0.11	0.00	0.00	0.00	0.01	0.02
∞	0	0	0	0	0	0	0	0	0

Table 7.18 Differences between the empirical fractiles of the MLEs under the AR(1) and $\mathrm{AR}_{\mu}^{B}(1)$ models for $\psi = 1$.

T/Fr.	0.01	0.025	0.05	0.10	0.50	0.90	0.95	0.975	0.99
25	0.41	0.32	0.04	-0.46	-3.07	-1.80	-1.49	-1.29	-1.07
50	-0.60	-0.65	-1.30	-2.09	-3.30	-1.80	-1.46	-1.23	-1.02
75	-0.90	-1.63	-2.25	-3.15	-3.37	-1.77	-1.43	-1.21	-0.99
100	-1.67	-2.39	-3.08	-3.80	-3.38	-1.77	-1.42	-1.20	-1.00
250	-4.10	-4.53	-4.82	-4.90	-3.47	-1.78	-1.44	-1.20	-1.00
500	-5.33	-5.42	-5.37	-5.17	-3.50	-1.77	-1.44	-1.20	-1.00
1000	-6.29	-6.21	-5.93	-5.46	-3.51	-1.78	-1.44	-1.21	-1.02
∞	-7.05	-6.63	-6.14	-5.57	-3.49	-1.77	-1.44	-1.20	-0.99

Table 7.19 Differences between the empirical fractiles of the MLEs under the $\mathrm{AR}_{\mu}(1)$ and $\mathrm{AR}_{\mu}^{B}(1)$ models for $\psi = 1$.

The distributions of the estimates look by and large similar when the sample size is hundred (Figures 7.6 and 7.7). A sharp-eyed reader may also be able to detect the relatively shorter tails of the distribution in Figure 7.6 compared to the one in Figure 7.7 and see that the former distribution is also otherwise slightly more tight. The accompanying statistics give more detailed information and support our surmise above. Pleasingly, the

T/Fr.	0.01	0.025	0.05	0.10	0.50	0.90	0.95	0.975	0.99
25	-15.35	-12.49	-9.99	-7.10	-0.84	1.03	1.41	1.76	2.24
50	-16.15	-12.72	-9.65	-6.58	-0.87	0.97	1.33	1.67	2.11
75	-16.42	-12.28	-9.23	-6.30	-0.86	0.95	1.31	1.65	2.08
100	-16.02	-11.94	-8.95	-6.18	-0.85	0.94	1.30	1.63	2.05
250	-14.72	-11.06	-8.42	-5.88	-0.85	0.94	1.29	1.63	2.04
500	-14.27	-10.81	-8.21	-5.80	-0.86	0.92	1.28	1.62	2.04
1000	-13.93	-10.64	-8.11	-5.76	-0.86	0.93	1.29	1.62	2.04
∞	-13.68	-10.41	-7.97	-5.65	-0.87	0.92	1.28	1.60	2.00

Table 7.20 Empirical fractiles of $T(\hat{\psi}_{ad} - 1)$ for $\psi = 1$ $(\mathrm{AR}_{\mu}^{B}(1))$.

T/Fr.	0.01	0.025	0.05	0.10	0.50	0.90	0.95	0.975	0.99
25	1.66	1.90	2.08	2.32	0.17	-0.06	-0.10	-0.15	-0.21
50	1.90	2.04	2.09	1.86	0.07	-0.03	-0.04	-0.06	-0.10
75	1.96	1.97	1.87	1.28	0.04	-0.02	-0.03	-0.04	-0.06
100	1.98	1.85	1.49	0.92	0.03	-0.02	-0.02	-0.03	-0.05
250	1.28	0.91	0.60	0.32	0.01	-0.01	-0.01	-0.01	-0.02
500	0.74	0.46	0.29	0.15	0.01	-0.01	-0.01	0	0
1000	0.37	0.22	0.14	0.08	0	0	0	0	-0.01
∞	0	0	0	0	0	0	0	0	0

Table 7.21 Differences between the empirical fractiles of $T(\hat{\psi}_{ml} - 1)$ and $T(\hat{\psi}_{ad} - 1)$ for $\psi = 1$ $(\mathrm{AR}_{\mu}^{B}(1))$.

comparison proves that the adjustment can yield a reduction in both bias and SD. (The MSEs appear the same only because of the reporting accuracy. Actually, the comparison on p. 118 already yielded the result.) Perhaps the sole negative aspect is that the adjustment seems to increase skewness. Recalling the corresponding statistics for the estimates for the $AR_\mu(1)$ model (from Figures 7.1, 7.2, and 7.3) we find that both the present MLE and the present AE are less biased and scattered and much more accurate in terms of MSE for hundred observations.

T/Fr.	0.01	0.025	0.05	0.10	0.50	0.90	0.95	0.975	0.99
25	-3.45	-3.03	-2.65	-2.19	-0.51	1.02	1.45	1.82	2.28
50	-3.17	-2.75	-2.35	-1.88	-0.50	0.94	1.36	1.72	2.12
75	-3.05	-2.59	-2.19	-1.74	-0.50	0.92	1.33	1.69	2.11
100	-2.93	-2.48	-2.09	-1.67	-0.50	0.92	1.32	1.66	2.08
250	-2.66	-2.26	-1.93	-1.58	-0.50	0.90	1.29	1.64	2.05
500	-2.57	-2.21	-1.91	-1.60	-0.50	0.89	1.29	1.63	2.03
1000	-2.55	-2.20	-1.92	-1.61	-0.50	0.89	1.28	1.63	2.03
∞	-2.56	-2.23	-1.93	-1.61	-0.51	0.88	1.27	1.61	1.98

Table 7.22 Empirical fractiles of $\sqrt{W}$ for $\psi = 1$ ($AR_\mu^B(1)$).

T/Fr.	0.01	0.025	0.05	0.10	0.50	0.90	0.95	0.975	0.99
25	-2.76	-2.31	-1.91	-1.48	-0.43	0.93	1.33	1.67	2.11
50	-2.63	-2.21	-1.85	-1.48	-0.47	0.90	1.31	1.65	2.04
75	-2.58	-2.16	-1.83	-1.49	-0.48	0.89	1.29	1.65	2.06
100	-2.53	-2.13	-1.82	-1.49	-0.48	0.90	1.29	1.63	2.04
250	-2.47	-2.12	-1.84	-1.53	-0.49	0.89	1.28	1.63	2.03
500	-2.48	-2.15	-1.87	-1.57	-0.50	0.88	1.28	1.63	2.02
1000	-2.51	-2.17	-1.90	-1.60	-0.50	0.89	1.28	1.63	2.02
∞	-2.56	-2.23	-1.93	-1.61	-0.51	0.88	1.27	1.61	1.98

Table 7.23 Empirical fractiles of $\sqrt{W_{ad}}$ for $\psi = 1$ ($AR_\mu^B(1)$).

Zivot (1994, p. 564), who focused on the unit-root case, reasoned that the MLE takes similar values to the OLS estimate from the simple AR(1) model when the OLS estimate is calculated for observations from y_2 to y_T (y_0 is undefined in the Bhargava model so the OLS estimate is uncalculable for observations from y_1 to y_T, cf. formula (6.3)).[30] Unreported simulation experimentation reveals that this holds only under a

[30] Formula (37) of Zivot for $S^2(\rho, x_0)$ agrees with our formula (6.50) for $\hat{\sigma}^2_\psi$ after substituting ψ in

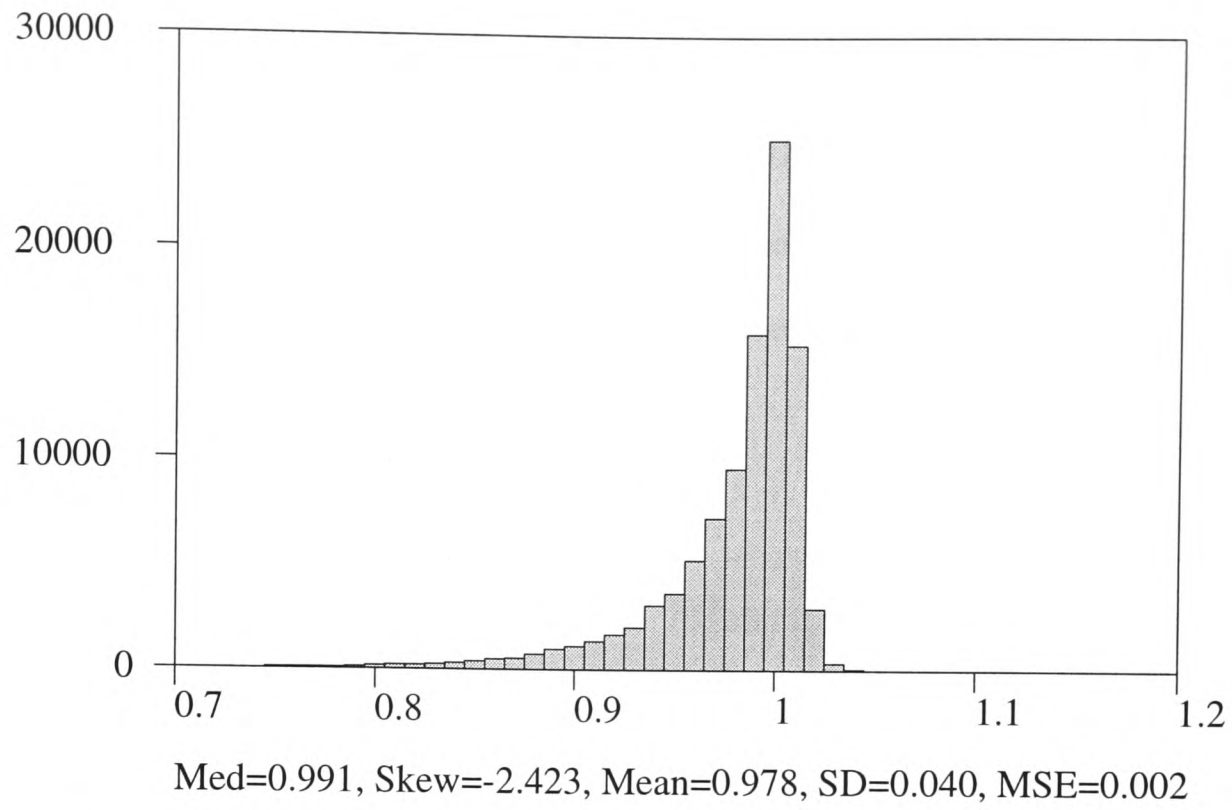


Figure 7.6 Empirical distribution of $\hat{\psi}_{ml}$ for $\psi = 1$ and $T = 100$ ($\mathrm{AR}_{\mu}^{B}(1)$).

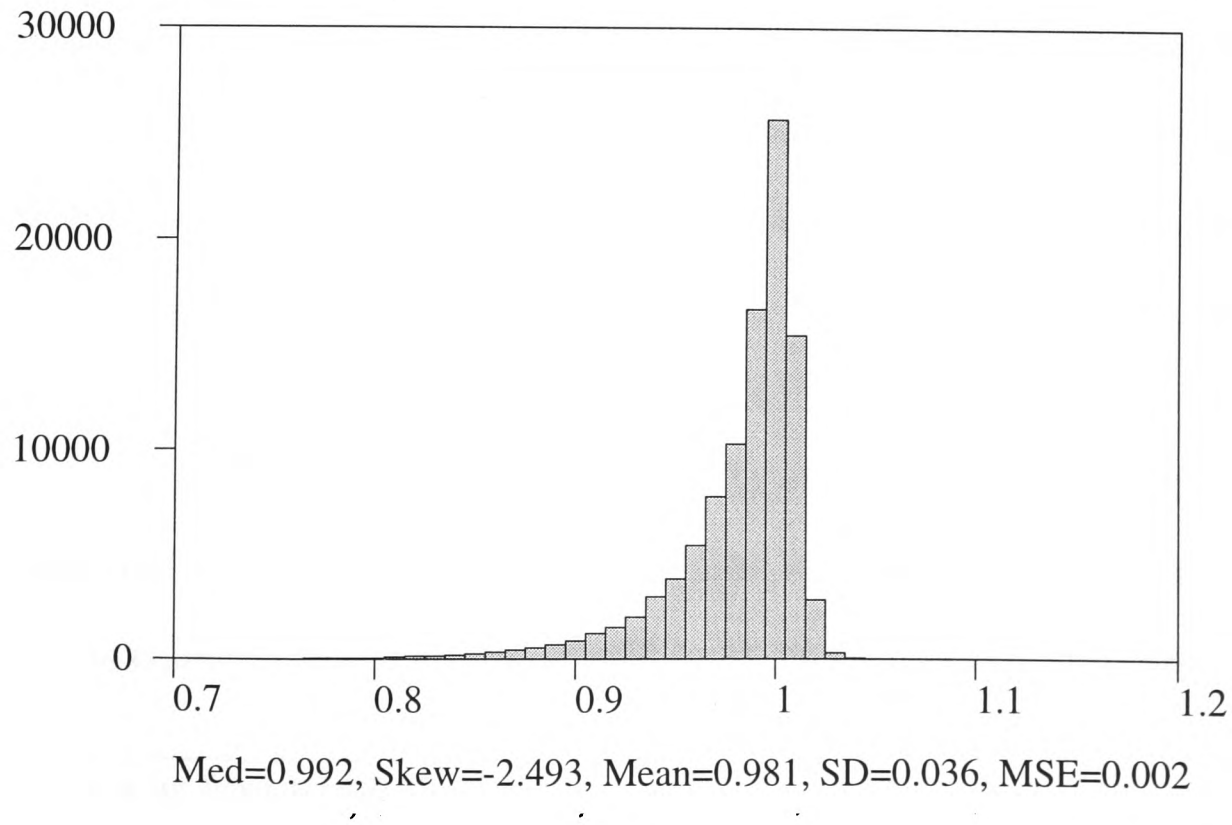


Figure 7.7 Empirical distribution of $\hat{\psi}_{ad}$ for $\psi = 1$ and $T = 100$ ($\mathrm{AR}_{\mu}^{B}(1)$).

unit root and a fairly large sample size (Table 7.18 gives some information on this issue). Otherwise the estimates can take very different values.

The fractiles of the Wald statistics can be read from Tables 7.22 and 7.23. The location and shape of the distributions vary less with the sample size than those of the estimates. This applies especially to the adjusted statistic which follows the asymptotic distribution better than the unadjusted one. The distribution of the adjusted statistic is also more concentrated in general.

7.6 Power Comparisons for the Bhargava AR(1) Model with Constant

The simulated powers of the ψ_{μ}^{B} and $\psi_{\mu,ad}^{B}$ tests are documented in Table 7.24. Again the test size is 5 per cent, ψ in the range 0.4 to 0.99, and sample size (T) in the range 25 to 250.[31] Even though our analysis has assumed that the starting value of the unobservable process x_t (beneath the y_t process) equals zero it is allowed to differ from zero in the simulation experiment. The quantity $|x_0|/\sigma$ is the standardised absolute value of the deviation of the starting value of the x_t process from zero or the unconditional mean of the x_t process (recall equation (6.47)). The deviation of interest does not relate to the unconditional mean of the y_t process as in Section 7.3. The unconditional mean of y_t or γ_0 can be imagined to have been added to the x_t process afterwards and hence be irrelevant as regards to the autocorrelation structure.[32] The following argumentation will refer to the y_t process and its properties (not to the unconditional mean of the x_t, say).

The profile and adjusted profile log-likelihoods which were derived in Section 6.6 assumed that x_0 equals zero. The term should be incorporated into the likelihoods when

---

place of ρ and ignoring fixed terms. Zivot (*op. cit.*) writes on p. 564 that '$S^2(\mu, x_0)$...is minimized near $\widetilde{\rho}$, the OLS estimate of ρ in the simple AR(1) model without an intercept.' Apparently Zivot refers to $S^2(\rho, x_0)$ as $S^2(\mu, x_0)$ is not defined in the article and ρ is not an argument in it.

[31] Tables 7.24 and 7.25 are based on the same seeds or random numbers which were used when studying power under the $\mathrm{AR}_{\mu}(1)$ model (Table A6.2).

[32] Deviation $-(x_0)/\sigma$ would produce a similar but not exactly the same estimate of power as $(x_0)/\sigma$. An exactly coinciding estimate would require reversing also the sign of the pseudo-random innovations.

the assumption is in doubt but a problem is that the time series x_t is unobservable.

The results of the zones where x_0 is not zero yield information on two kinds of misspecifications. In the first instance, we learn about the performance of the Bhargava model when the assumption $x_0 = 0$ is not met even though it is employed in the (possibly adjusted) profile likelihood. Second, we learn how the present tests statistics cope with a time series the unconditional mean of which is zero and has started from $y_0 = x_0/\sigma$ by the $\mathrm{AR}_\mu(1)$ model, or more generally when a time series generated by the $\mathrm{AR}_\mu(1)$ model may possess a non-zero mean but the starting value is connected to the mean so that $x_0/\sigma = [y_0 - \alpha/(1-\psi)]/\sigma$. This interpretation stems from the fact that the time series the two models generate coincide in these circumstances (p. 112). For simplicity, we shall refer to MLE and AE also when $x_0 \neq 0$ even though the estimates are then based on a misleading likelihood.

The blank entries in the table indicate that power equal to one (to three decimal places) was achieved already with a larger ψ than the ψ corresponding to the entry. There is more uncertainty associated with some of the power figures than with the fractiles due to numerical solving of the estimates. We shall return to this issue shortly (see also Appendix A6 for details and especially Table A6.6 there).

For the tests based on the SCs the results are to an extent a mirror image of those found under the $\mathrm{AR}_\mu(1)$ model. The $\psi^B_{\mu,ad}$ test based on the AE is more or essentially as powerful as the ψ^B_μ test based on the MLE when the time series has started from the unconditional mean (Table 7.24). The power figures favour the adjusted statistic for most scrutinized values of ψ when the sample size is hundred or less. There is no notable difference in power when there are 250 observations. A gain in power thus can result despite the asymptotically negligible adjustment to the MLE.

A starting value different from zero (five) turns the power ordering clearly in favour of the ψ^B_μ test. An intuitive explanation is that the adjustment makes use of information measures which mistakenly do not take x_0 into account. Both ψ^B_μ and $\psi^B_{\mu,ad}$ display in general little power when $|x_0|/\sigma$ is large. This suggests that the MLE and AE do not

converge to the true value when x_0 is not incorporated in the (possibly adjusted) profile likelihood and $|x_0|/\sigma$ tends to infinity. This can be easily proved, see Appendix A5.

Example (Table 7.24). The ψ_μ^B and $\psi_{\mu,ad}^B$ tests reject the null with probabilities 64.2 and 65.1, respectively, when ψ is 0.9, $|x_0|/\sigma$ equals zero and the sample size is hundred. Tuning $|x_0|/\sigma$ to five gives the corresponding figures 12.1 and 9.3.□

Example (Table 7.24). The largest difference in power in the table takes place at ψ equal to 0.8, $|x_0|/\sigma$ equal to five and a sample size of one hundred. The powers of the ψ_μ^B and $\psi_{\mu,ad}^B$ tests are then 49.8 and 39.0, correspondingly.□

The results for the Wald tests τ_μ^B and $\tau_{\mu,ad}^B$ are fairly similar (Table 7.25). The adjusted version is more powerful when the sample size is 25. For larger sample sizes the unadjusted version is more often more powerful than the adjusted version. Again the unadjusted version dominates in power when $|x_0/\sigma|$ is five. The powers of both tests shrink to zero for most cases in the lower part of the table where the starting value is very large. An intuitive explanation is the inconsistency of the MLE and the AE.

Example (Table 7.25). The τ_μ^B and $\tau_{\mu,ad}^B$ tests display powers 20.8 and 21.8, respectively, when ψ is 0.8, $|x_0|/\sigma$ is zero, and the number of observations is 25. The corresponding powers are 6.0 and 4.7 when $|x_0|/\sigma$ is five.□

There is no notable difference in power between the tests based on the SCs relative to the Wald tests when $|x_0|/\sigma$ is greater or equal to 25. In the upper part of the tables the relative powers of the two type of tests behave quite differently, mirrorwise to an extent, from the circumstance under the $AR_\mu(1)$ model. The SC tests are more powerful, and often quite notably so, when the starting value *differs* from zero. The Wald tests are inclined to be more powerful when the starting value is zero and there are a hundred observations or less.[33] The rank ordering of the types of tests depends on the size of ψ

[33]The circumstance where ψ equals 0.95 and T equals a hundred makes an exception.

in the case of a zero starting value and sample size 250. As a whole differences exist but they are not drastic.

Detailed comparison of the power figures in Tables 7.24 and 7.25 to those in Tables 7.8 and 7.9 (for the $AR_\mu(1)$ model) cannot be done straightforwardly. Even though the null models agree, the models under the alternatives against which the powers are evaluated differ in general except when $x_0/\sigma = [y_0 - \alpha/(1-\psi)]/\sigma$ (cf. formulae (6.4) and (6.47) and the discussion on p. 112). Comparison is thus valid under this additional assumption.

The uppermost zone for $|x_0|/\sigma$ equal to zero transfers into the assumption that the time series has started from the unconditional mean under the $AR_\mu(1)$ model. If one were unaware of the distributional results in Table 7.19 it would be incredible that the powers of the ψ_μ^B and ψ_μ tests are pretty similar under this assumption (and especially so for the $\psi_{\mu,ad}^B$ and ψ_μ tests which are the most powerful for each model) when there are 25 observations. However, the tests gain power in the present context much faster than under the $AR_\mu(1)$ model as the number of observations increases which leads to a power ranking which is in a better accordance with our expectations. In the case of the Wald statistics the powers emerge larger under the Bhargava model throughout (apart from a case with equal power) and especially so for distant alternatives (when a zero starting value is assumed). This is largely due to a major difference between the models in terms of the power of the Wald tests: For the present model the power of the Wald tests does not fall short of the power of the SC tests when the starting value is zero which happens under the $AR_\mu(1)$ model.

Example (Tables 7.8 and 7.24). The power figures for the ψ_μ test under $\psi = 0.95$ and a zero starting value at sample sizes 25 and 250 are 7.8 and 63.9, respectively. The corresponding figures for the ψ_μ^B test are 7.7 and 85.4. The corresponding figures for the Wald tests are 6.1 and 45.5 and 7.8 and 83.5. The difference in the gain in power is even larger than for the SC tests.□

		T					
		25		100		250	
$\|x_0\|/\sigma$	ψ	ψ_μ^B	$\psi_{\mu,ad}^B$	ψ_μ^B	$\psi_{\mu,ad}^B$	ψ_μ^B	$\psi_{\mu,ad}^B$
0	**0.99**	5.3	5.2	8.0	7.9	14.2	14.2
	0.95	7.7	7.7	26.8	27.7	85.4	85.3
	0.90	11.3	11.7	64.2	65.1	99.9	99.9
	0.85	15.2	15.8	90.6	91.0	100.0	100.0
	0.80	19.8	20.8	98.8	98.7		
	0.70	34.8	36.6	100.0	100.0		
	0.60	52.8	54.6				
	0.50	70.6	72.2				
	0.40	85.4	86.5				
5	**0.99**	5.7	5.7	7.1	7.0	10.7	10.7
	0.95	5.2	5.1	8.0	7.2	21.1	19.8
	0.90	4.4	3.8	12.1	9.3	45.1	40.7
	0.85	4.6	3.8	28.2	20.6	77.3	70.6
	0.80	7.0	5.5	49.8	39.0	94.9	92.2
	0.70	16.3	12.8	87.9	81.2	99.9	99.8
	0.60	36.4	30.3	98.7	97.5	100.0	100.0
	0.50	61.0	54.7	99.9	99.9		
	0.40	82.2	78.4	100.0	100.0		
25	**0.99**	2.9	2.9	0.5	0.4	0.0	0.0
	0.95	0.0	0.0	0.0	0.0	0.0	0.0
	0.90	0.0	0.0	0.0	0.0	0.0	0.0
	0.85	0.0	0.0	0.0	0.0	0.0	0.0
	0.80	0.0	0.0	0.0	0.0	0.0	0.0
	0.70	0.0	0.0	0.0	0.0	0.0	0.0
	0.60	0.0	0.0	0.0	0.0	0.0	0.0
	0.50	0.0	0.0	0.0	0.0	15.1	9.8
	0.40	0.0	0.0	19.0	10.6	97.7	95.9
250	**0.99**	0.0	0.0	0.0	0.0	0.0	0.0
	0.95	0.0	0.0	0.0	0.0	0.0	0.0
	0.90	0.0	0.0	0.0	0.0	0.0	0.0
	0.85	0.0	0.0	0.0	0.0	0.0	0.0
	0.80	0.0	0.0	0.0	0.0	0.0	0.0
	0.70	0.0	0.0	0.0	0.0	0.0	0.0
	0.60	0.0	0.0	0.0	0.0	0.0	0.0
	0.50	0.0	0.0	0.0	0.0	0.0	0.0
	0.40	0.0	0.0	0.0	0.0	0.0	0.0

Table 7.24 Empirical power ($\times 100$) of the ψ_μ^B and $\psi_{\mu,ad}^B$ tests when size is 5 %.

		T					
		25		100		250	
$\lvert x_0\rvert/\sigma$	ψ	τ_μ^B	$\tau_{\mu,ad}^B$	τ_μ^B	$\tau_{\mu,ad}^B$	τ_μ^B	$\tau_{\mu,ad}^B$
0	**0.99**	5.3	5.4	8.0	8.0	14.6	14.7
	0.95	7.8	7.9	28.8	29.4	83.5	82.7
	0.90	11.7	12.1	66.3	66.0	99.4	99.2
	0.85	15.8	16.6	90.5	89.3	100.0	100.0
	0.80	20.8	21.8	98.1	97.5		
	0.70	36.5	37.9	100.0	99.9		
	0.60	55.1	56.6		100.0		
	0.50	72.4	73.6				
	0.40	86.4	87.0				
5	**0.99**	5.4	5.3	7.2	7.3	10.8	10.8
	0.95	5.1	5.2	6.4	5.9	12.5	10.9
	0.90	3.9	3.4	8.3	5.9	31.6	25.5
	0.85	3.9	3.2	21.0	15.2	73.4	65.0
	0.80	6.0	4.7	42.7	32.7	94.7	91.6
	0.70	13.8	11.0	86.5	79.3	99.9	99.8
	0.60	32.1	26.8	98.7	97.5	100.0	100.0
	0.50	57.6	51.2	99.9	99.9		
	0.40	79.6	75.5	100.0	100.0		
25	**0.99**	2.9	2.9	0.4	0.3	0.0	0.0
	0.95	0.0	0.0	0.0	0.0	0.0	0.0
	0.90	0.0	0.0	0.0	0.0	0.0	0.0
	0.85	0.0	0.0	0.0	0.0	0.0	0.0
	0.80	0.0	0.0	0.0	0.0	0.0	0.0
	0.70	0.0	0.0	0.0	0.0	0.0	0.0
	0.60	0.0	0.0	0.0	0.0	0.0	0.0
	0.50	0.0	0.0	0.0	0.0	15.1	9.8
	0.40	0.0	0.0	19.0	10.6	97.7	95.9
250	**0.99**	0.0	0.0	0.0	0.0	0.0	0.0
	0.95	0.0	0.0	0.0	0.0	0.0	0.0
	0.90	0.0	0.0	0.0	0.0	0.0	0.0
	0.85	0.0	0.0	0.0	0.0	0.0	0.0
	0.80	0.0	0.0	0.0	0.0	0.0	0.0
	0.70	0.0	0.0	0.0	0.0	0.0	0.0
	0.60	0.0	0.0	0.0	0.0	0.0	0.0
	0.50	0.0	0.0	0.0	0.0	0.0	0.0
	0.40	0.0	0.0	0.0	0.0	0.0	0.0

Table 7.25 Empirical power ($\times 100$) of the τ_μ^B and $\tau_{\mu,ad}^B$ tests when size is 5 %.

The present SC tests are somewhat more powerful than the SC tests for the $AR_\mu(1)$ model for some local alternatives for $|x_0|/\sigma$ different from zero especially in the zone for $|x_0|/\sigma$ equal to five. However, in most cases the latter tests dominate in power. A similar statement holds for the Wald tests as well.

It is more difficult to summarise the pattern of power in the present context than under the $AR_\mu(1)$ model because the power ranking of the tests is here less systematic. Apart from the case x_0 equal to 0 and T equal to 25, no test dominates the others consistently in any of the zones for pairs of $(|x_0|/\sigma, T)$ and it seems legitimate to reason that no simple rank ordering exists. To clarify ideas, we yet lay out Table 7.26 which depicts the formation of power fairly well.[34] The table is a summary; we are not recommending application of the ψ_μ^B test when $x_0 \geq 0$. The model is then misspecified and x_0 should be merged into the likelihood function. In practise this may be impossible because x_0 is probably unobserved.

$\|x_0\|/T\sigma$	Suggested test
0	$\tau_{\mu,ad}^B$ or ψ_μ^B
≥ 0.02	ψ_μ^B

Table 7.26 Suggested test for different values of $|x_0|/T\sigma$ for $|x_0|/\sigma$ in the range [0,250] and T in the range [25,250] ($AR_\mu^B(1)$).

A question which is likely to arise is which model, $AR_\mu^B(1)$ or $AR_\mu(1)$, is to be to employed for testing? As already noted, the question is well posed only when the alternatives are such that $x_0/\sigma = [y_0 - \alpha/(1-\psi)]/\sigma$. Table 7.27 gives a recommendation for such a case. When the conditions take value zero, the time series engendered by the $AR_\mu(1)$ model has started from the unconditional mean. Tests based on the Bhargava model are then favoured. Otherwise the Bhargava model, as we have devised it, is misspecified and tests based on the $AR_\mu(1)$ model should be used. As before, it is difficult

[34]The ψ_μ^B or $\tau_{\mu,ad}^B$ test is favoured when the starting value (of x_t) equals zero. If this is the case and the sample size is 25 then the latter test dominates strictly in power. For larger sample sizes one or the other is the most powerful test or does not cause a large power loss relative to the most powerful test. (Picking a single test for the case would be more difficult; τ_μ^B would be a possible choise, too, say.) Contrary to the case under the $AR_\mu(1)$ model, the ψ_μ^B test emerges as the most solid test. Choosing a single test of choice is harder. The ψ_μ^B test is a good candidate as it is fairly clearly the best test whenever x_0 deviates from zero and does not do too badly when x_0 matches with zero.

to follow the recommendation because in practise x_0 is unobservable.

$\lvert x_0\rvert/T\sigma = [y_0 - \alpha/(1-\psi)]/T\sigma$	Suggested test
0	$\tau^B_{\mu,ad}$ or ψ^B_μ
0.02–0.05	$\psi_{\mu,ad}$
≥ 0.1	$\psi_{\mu,ad,ap}$ or $\tau_{\mu,ad,ap}$

Table 7.27 Suggested test for different values of $|x_0|/T\sigma = [y_0 - \alpha/(1-\psi)]/T\sigma$ for $|x_0|/\sigma$ in the range [0,250] and T in the range [25,250] ($\mathrm{AR}_\mu(1)$ and $\mathrm{AR}^B_\mu(1)$).

7.7 Finite-Sample Results for the Unit-Root AR(2) Model

The sizes of (the square roots of) the Wald statistics are evaluated below because the distribution of the SCs depends on the nuisance parameter ϕ_1 (Section 3.3). The statistic $W_{ad2,s}$ (s for shortcut) is an adjusted Wald statistic in the calculation of which coefficient c has been estimated in a shortcut way or by using the formula $c \approx (1-\phi_1)^{-2}$.

The empirical sizes reported in Table 7.28 make it clear that the adjustment does not yield an improved fit with the asymptotic unit-root distribution (3.11) the fractiles of which are used as critical values. The five models considered are: *i)* $\phi_1 = 0.95$, *ii)* $\phi_1 = 0.5$, *iii)* $\phi_1 = 0$, *iv)* $\phi_1 = -0.5$, and *v)* $\phi_1 = -0.95$. Under all models the sum of the AR coefficients is restricted to unity or $\psi = 1$ in the notation of Section 6.6 (e.g. ρ_1 is 1.95 and ρ_2 is -0.95 under model *i)*).

The adjustment does not calibrate the small-sample distribution with the asymptotic distribution when the latter is nonstandard. It may be worth recalling that an improvement arose under the unit-root $\mathrm{AR}_\mu(1)$ model with drift or when Normality applied asymptotically. The empirical sizes — which are smaller than the nominal for the adjusted statistics — suggest instead that the adjustment has shifted or otherwise moved the distribution to the right. We turn our attention to estimation as the adjustment is not helpful from the point of view of testing in this context (if one wants to base inference on the asymptotic distribution (3.11)).

Tables 7.29, 7.30, and 7.31 record the bias, the SD, and the MSE of the MLE, the AE and the shortcut AE or the SAE. We pointed out in Section 5.3 that the process starts to

	T								
	25			100			250		
ϕ_1	$\sqrt{W}$	$\sqrt{W_{ad2}}$	$\sqrt{W_{ad2,s}}$	$\sqrt{W}$	$\sqrt{W_{ad2}}$	$\sqrt{W_{ad2,s}}$	$\sqrt{W}$	$\sqrt{W_{ad2}}$	$\sqrt{W_{ad2,s}}$
0.95	7.41	3.91	0.33	6.08	3.57	3.55	5.37	2.21	1.76
0.50	6.29	1.17	0.73	6.08	3.57	3.55	4.79	4.71	4.71
0.00	6.18	2.53	2.38	5.39	4.54	4.53	5.26	4.92	4.92
−0.50	6.18	3.26	3.21	5.32	4.68	4.68	5.23	4.94	4.94
−0.95	6.08	3.57	3.55	5.33	4.71	4.71	5.23	5.00	5.00

Table 7.28 Empirical size of the Wald statistics when nominal size is 5 % (AR(2)).

resemble $\Delta^2 y_t = \epsilon_t$ or an $I(2)$ process when ϕ_1 tends to unity. The bias appeared to fade for such values whereas a bias comparable to the bias under the $\text{AR}_\mu(1)$ seemed to arise for ϕ_1 close to minus unity. The biases in Table 7.29 depend on ϕ_1 as suggested by the theory in Section 5.3 (which is really valid for stationary values of ψ only, though). The simulations confirm a radical convertion in the distribution also otherwise as anticipated in the beginning of Section 6.6.

Example (Table 7.29 and other simulation experiments). The bias of the MLE is −0.001 when ϕ_1 equals 0.95 and the sample size is 100 (note that the figures in the table have been multiplied by ten). If ϕ_1 is −0.95 then the bias explodes by more than 30 fold or to −0.034. Figures 7.8 and 7.9 illustrate the staggering differences between the distributions. (Note that the scale of the graphs is different from the scale used for the AR(1) models.) An increase in spread arises as ϕ_1 is tuned towards larger values but also the increase in bias should be visible. The difference in bias is in line with our previous reasoning when discussing Figure 5.1 in Section 5.3 on bias and interpreting the coefficient c in Section 6.6 (p. 132).□

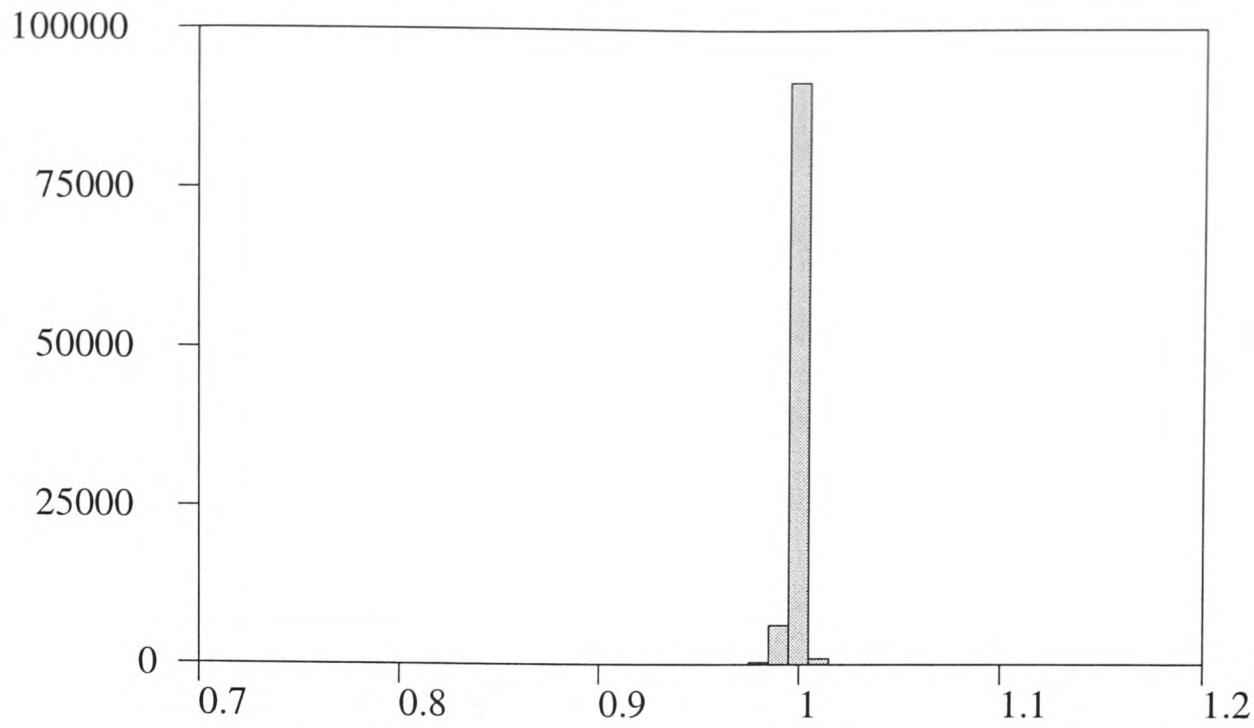


Figure 7.8 Empirical distribution of $\hat{\psi}_{ml}$ for $\psi = 1$, $\phi_1 = 0.95$, and $T = 100$ (AR(2)).

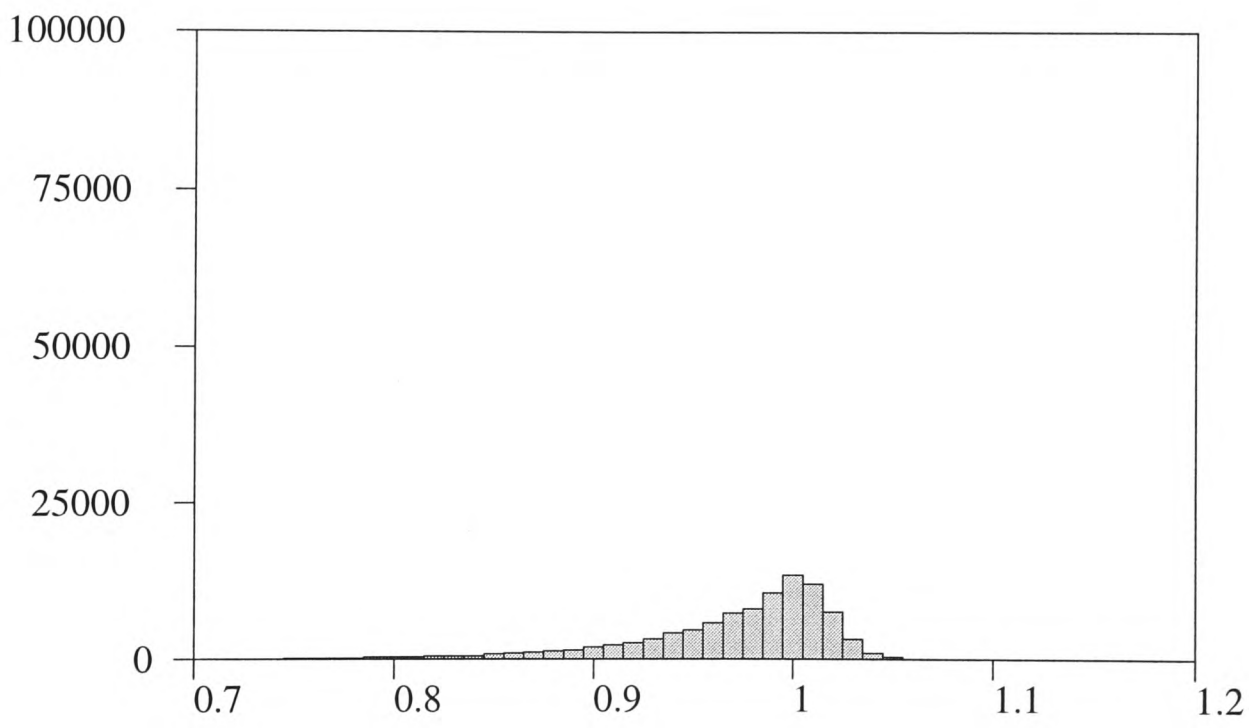


Figure 7.9 Empirical distribution of $\hat{\psi}_{ml}$ for $\psi = 1$, $\phi_1 = -0.95$, and $T = 100$ (AR(2)).

Example (Table 7.29 and other simulation experiments). The bias -0.034 under ϕ_1 equal to -0.95 and T equal to 100 is about twice as big as the bias under the simple AR(1) model (-0.0174 by Table 7.34 in Section 7.8) and about two thirds of the bias under the $\mathrm{AR}_\mu(1)$ model (-0.052 from Figure 7.1) for the same sample size. The formulae (5.10) and (5.11) hint that the bias should tend to the bias valid under the $\mathrm{AR}_\mu(1)$ model as ϕ_1 tends to -1 (cf. also the example on p. 58). Our findings are fairly well in line with this theory.□

Example (Table 7.29 and other simulation experiments). The midpoint case of $\phi_1 = 0$ results in a bias equal to -0.0175 when T is 100. The bias is only slightly larger (in absolute value) than the bias (-0.0174) when the parameter is not estimated or the simple AR(1) model is utilized. The outcome is in agreement with the theoretical formula (5.12) for the bias in the two circumstances (the formula assumes a ψ smaller than unity but the unit-root circumstance appears to behave similarly by and large).□

The adjustment decreases bias in all of the cases we have investigated. In absolute terms, the adjustment (for bias) is in general larger for larger values of ϕ_1 (except at ϕ_1 equal to 0.95) when the time series is of length 25 but for longer spans the adjustment (for bias) is essentially fixed (at 0.06 or 0.01 for T equal to 100 and 250, respectively). The relative decrease is hence more substantial for larger values of ϕ_1 for all sample sizes. Actually, in the case of $\phi_1 = 0.95$ and $T = 25$ the AEs tend to overestimate ψ. The upward biases are likely to be due to inflated estimates of c for values of ϕ_1 close to unity. The largest biases arise under the smaller values of ϕ_1 so the adjustment is relatively less beneficial when the bias is more severe. The adjustment diminishes as T takes larger values because the MLE and the AE share the same asymptotic distribution.

Example (Table 7.29 and other simulation experiments). The bias of the MLE and the AE are -0.067 and -0.055 respectively, when $\phi_1 = 0$ or the second autoregressive coefficient is unnecessarily estimated and the sample size is 25. The adjustment has cut the bias by about 18 per cent. This is enough to bring down the bias even below the MSE of

the MLE when the estimated model is the simple AR(1) (the bias is then -0.063 by an unreported simulation experiment). For data length 100, the biases are -0.018 and -0.017, respectively. A decrease in bias is found again but it is relatively smaller at about three per cent. Again the bias of the AE is smaller than the bias of the MLE under the simple AR(1) model (-0.018 by Table 7.34), and the statement holds also when the length of the data is 250 (by an unreported simulation experiment).□

An astonishing feature is that the SAE appears to be the least biased estimate in most cases. On the other hand, the shortcut adjustment overshoots for extreme values of ϕ_1 when the sample size is small (the case of $\phi_1 = 0.95$ and $T = 25$ in Table 7.29).

With the benefit of hindsight given the large variation in bias, and indeed in the distribution, with the size of ϕ_1, it is not amazing that the size of coefficient c of the Cox–Reid adjustment depends greatly on ϕ_1. However, it is astonishing that the simulations confirm that the theoretical formulae of Section 5.3 suggest a correct pattern for the bias despite our caveats. It means that coefficient c and the bias relate mirrorwise to each other, i.e. c decreases as the bias increases. Because the SD and the MSE behave similarly to the bias, i.e. decrease with ϕ_1, the explanation for the behaviour of c is not that the adjustment would focus on SD or MSE instead of bias as was suggested on p. 132 in Section 6.6. The third explanation provided on the aforementioned page remains valid. Thus a probable explanation for the increase in c with ϕ_1 and the fact that the adjustment yet does not necessarily increase with ϕ_1 (or can even decrease with it) is the following. The Fisher information for ψ (under the assumption of a unit root, cf. p. 132) increases with ϕ_1. Consequently, the profile likelihood is likely to be more steeply curved under a large ϕ_1 and the adjustment can remain the same (or become smaller) despite the larger coefficient c.

In general, the adjustments cut down SD as well, and, as expected, the smaller the sample is the larger the decrease tends to be (Table 7.30). Like the decrease in bias also the decrease in SD is relatively more considerable for large values of ϕ_1. Exceptions occur at ϕ_1 equal to 0.95 and T equal to 25 and 100 in which cases the SD of the AE is larger

than the SD of the MLE. The explanation is probably that the estimate of coefficient c may increase rapidly for such values which results in an exaggerated adjustment. At first sight, the SAE seems to do better in this respect but Table A6.8 in Appendix A6 reveals that the SAE took a value of unity by substitution (or the adjusted profile likelihood possessed complex roots and a positive c) in about 36 per cent of the draws for T equal to 25 and in about 5 per cent for T equal to 100. The superiority of the SAE is hence an unreliable issue. Like the bias also the SD of the AE (and the SAE) is smaller under a zero ϕ_1 than the SD of the MLE calculated from the simple AR(1) model (Table 7.34 and unreported simulation exercises). According to the theory of MacKinnon and Smith (1996) on bias corrections (cf. Section 5.5) and the tentative formula (5.11) (implying a bias function with a negative slope for the unit-root AR(2) model) elimination of bias should have a variance decreasing impact, too. We have not eliminated the bias altogether but the decrease in variance is in line with their more general results.

The MSEs of the estimates are arranged as Table 7.31. The pattern of the previous two tables is reinforced: the AEs are favoured in general and especially when ϕ_1 is large or the sample size is small. The SAE compares the best in general with a small margin over the AE but — as just pointed out — the relatively good performance of it is somewhat artificial under a ϕ_1 equal to 0.95.

The gain in MSE in terms of percentage is presented in Table 7.32. The decrease can be fairly large especially when the time series is short. Two entries for ϕ_1 equal to 0.95 lack a figure because in those cases the MSE increases contrary to the general pattern. The increase is probably due to overshooting of the estimate of coefficient c and thus of the AE (recall that the bias of the AE is *positive* for $T = 25$ and the SD of the AE is *larger* than that of the MLE when $T = 25$ or $T = 100$ under a ϕ_1 equal to 0.95). In practise, such cases would probably be allocated the value of unity and the MSE after such a transformation should be more favourable to the AE.

Example (7.32). The decrease in MSE due to the adjustment is 21.8 per cent when ϕ_1 is zero and T is 25. The adjustment is so effective that the MSE of the MLE from the simple

AR(1) model is larger (this MSE is 0.0172 by an unreported simulation experiment). This holds for the other two sample sizes as well (Table 7.34 and an unreported Monte Carlo experiment).□

In terms of estimation accuracy, as measured by MSE, the adjustment can be very beneficial. Not only is the MSE smaller for the AEs in general but it is also smaller under a ϕ_1 equal to zero. In other words, if the MSE is the yardstick then it seems that one should never estimate a simple AR(1) model under a unit-root because the AEs for the unit-root AR(2) model are more accurate also when the simple AR(1) model is valid.[35] The heuristic for the result is that the adjustments for the AR(2) model have been derived under the assumption of a unit root.

It is disappointing that the adjustment is modest when the bias (in absolute terms), the SD, and the MSE peak or ϕ_1 is small. However, it is useful that larger values of ϕ_1, for which the adjustment is more effective, are empirically relevant. For example, it was pointed out in the beginning of Section 6.6 that the process $y_t = 1.3y_{t-1} - 0.3y_{t-2} + \epsilon_t$ ($\phi_1 = 0.3$) is the time-series process for US economic data in general (after trend removal) according to Orcutt (1948) and others.

	T								
	25			100			250		
ϕ_1	MLE	AE	SAE	MLE	AE	SAE	MLE	AE	SAE
0.95	-0.046	0.028	0.118	-0.010	-0.004	0.003	-0.004	-0.003	-0.002
0.50	-0.344	-0.215	-0.184	-0.089	-0.083	-0.082	-0.035	-0.035	-0.035
0.00	-0.667	-0.549	-0.541	-0.175	-0.169	-0.169	-0.071	-0.070	-0.070
−0.50	-0.990	-0.875	-0.872	-0.262	-0.256	-0.256	-0.106	-0.105	-0.105
−0.95	-1.269	-1.165	-1.155	-0.341	-0.335	-0.335	-0.138	-0.137	-0.137

Table 7.29 Empirical bias (×10) of the MLE, the AE, and the SAE (AR(2)).

The biases, the SDs, and the MSEs of the AEs under the extreme processes with $\phi_1 = -0.95$ and $\phi_1 = 0.95$ for $T = 100$ are replicated in Table 7.33. It also reports the corresponding but most extreme statistics out of the pool of the other models we have

[35]To improve comparability, we have also conducted the simulations of the AR(2) model with $\phi_1 = 0$ with the same seeds which were used to simulate the simple AR(1) model (the ones in Table A6.1). The outcome was essentially the same or the difference in MSEs in favour of the AE from the AR(2) model was slightly larger than reported here.

	T								
	25			100			250		
ϕ_1	MLE	AE	SAE	MLE	AE	SAE	MLE	AE	SAE
0.95	0.343	0.362	0.284	0.033	0.069	0.030	0.009	0.008	0.008
0.50	0.758	0.642	0.621	0.166	0.160	0.160	0.065	0.064	0.064
0.00	1.267	1.141	1.131	0.317	0.310	0.310	0.127	0.126	0.126
−0.50	1.784	1.658	1.653	0.467	0.461	0.461	0.190	0.189	0.189
−0.95	2.250	2.119	2.121	0.603	0.597	0.597	0.246	0.245	0.245

Table 7.30 Empirical SD (×10) of the MLE, the AE, and the SAE (AR(2)).

	T								
	25			100			250		
ϕ_1	MLE	AE	SAE	MLE	AE	SAE	MLE	AE	SAE
0.95	1.1947	1.3154	0.9438	0.0115	0.0482	0.0091	0.0010	0.0008	0.0007
0.50	6.9218	4.5874	4.1927	0.3540	0.3238	0.3228	0.0548	0.0529	0.0529
0.00	20.5105	16.0307	15.7270	1.3097	1.2504	1.2497	0.2120	0.2083	0.2082
−0.50	41.6427	35.1349	34.9351	2.8720	2.7834	2.7832	0.4720	0.4664	0.4664
−0.95	66.7221	58.4786	58.3235	4.8004	4.6860	4.6857	0.7933	0.7861	0.7861

Table 7.31 Empirical MSE (×1000) of the MLE, the AE, and the SAE (AR(2)).

	T		
ϕ_1	25	100	250
0.95			20.0
0.50	33.7	8.5	3.5
0.00	21.8	4.5	1.8
−0.50	15.6	3.1	1.2
−0.95	12.4	2.4	0.9

Table 7.32 The empirical reduction in MSE (per cent) of the AE relative to the MLE (AR(2)).

considered in the thesis (AR(1), $AR_\mu(1)$, and $AR_\mu^B(1)$) when the process is a random walk. (The latter figures are extracted from Table 7.34 in Section 7.8.) The richness of the AR(2) model is reflected in the fact that the range of the statistics for this model nests the ranges of the SDs and MSEs for the rest of the models. The range of the biases under the other models is not quite nested by the range for the AR(2) model but the range is yet far wider under it. Moreover, the range for the AR(2) model should expand if more extreme values of ϕ_1 ($0.95 < |\phi_1| < 1$) were included in the comparison.

	AR(2)	**The other models**
Bias	$-0.04 \rightarrow -3.35$	$-1.75 \rightarrow -3.65$
SD	$0.69 \rightarrow 5.97$	$3.12 \rightarrow 4.38$
MSE	$0.00 \rightarrow 0.47$	$0.13 \rightarrow 0.33$

Table 7.33 The range of the biases, the SDs, and the MSEs ($\times$ 100) of the AEs for $T = 100$ under the AR(2) model (with $\phi_1 \in [-0.95, 0.95]$) and the other models (when the process is a random walk).

Figures 7.10 and 7.11 illustrate the distributions of the MLE and the AE, respectively, for ϕ_1 equal to 0.5 and one hundred observations. The differences are not obvious but actually the bars at the left are shorter and the bar at unity, say, are higher for the AE. The magnitude of ϕ_1 seems to be much more important in determining the shape of the distribution (of the MLE or the AE) than whether the estimate is the usual MLE or the AE (cf. Figures 7.9 and 7.8).

Example (Tabled 7.31 and 7.32 and Figures 7.9 and 7.8). Changing the parameter ϕ_1 from 0.95 to -0.95 causes an over 500 fold increase in the MSE (Figures 7.9 and 7.8). The adjustment can yield at best only a reduction of about a third in the MSE of the MLE (Table 7.32).□

7.8 Comparison to the Optimally Multiplicatively Corrected Estimates

The bias, the SD and the MSE of the AEs under the models studied are printed in the first column in Table 7.34 so that the relative accuracy of the AEs under different

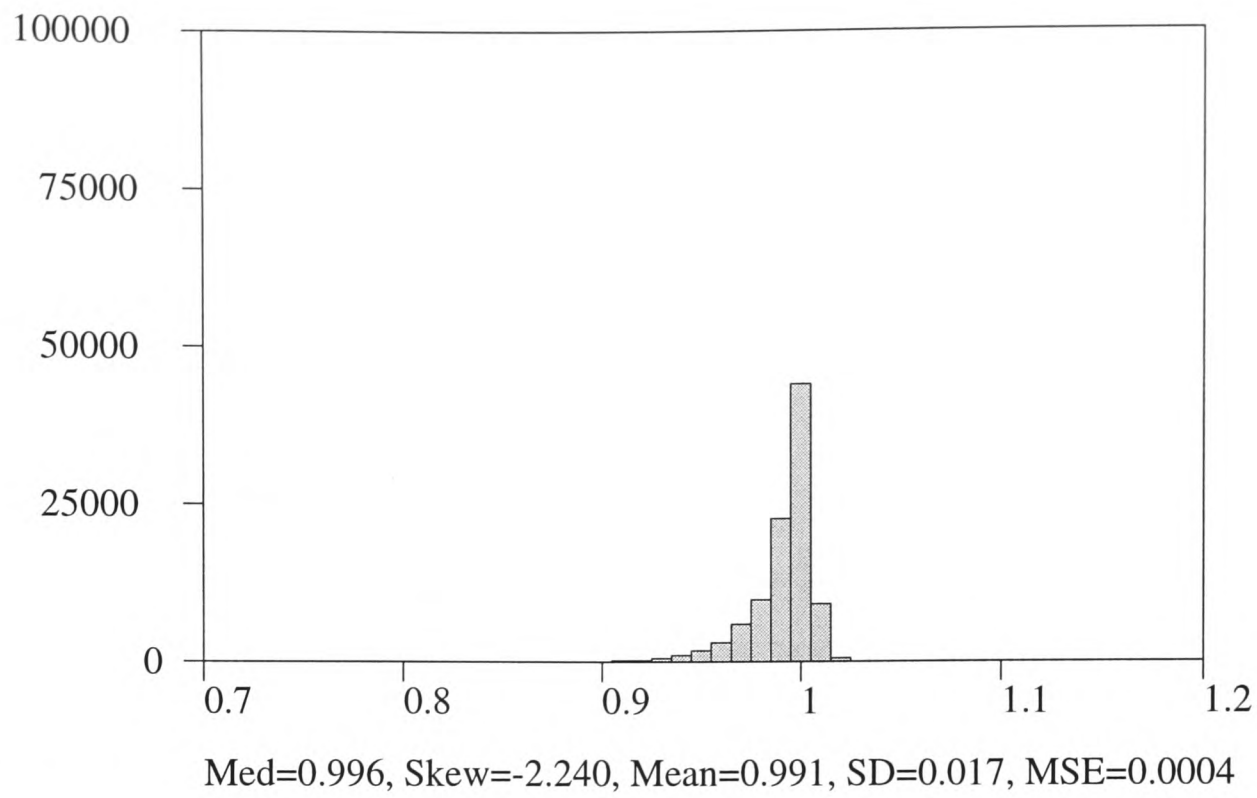


Figure 7.10 Empirical distribution of $\hat{\psi}_{ml}$ for $\psi = 1$, $\phi_1 = 0.5$, and $T = 100$ (AR(2)).

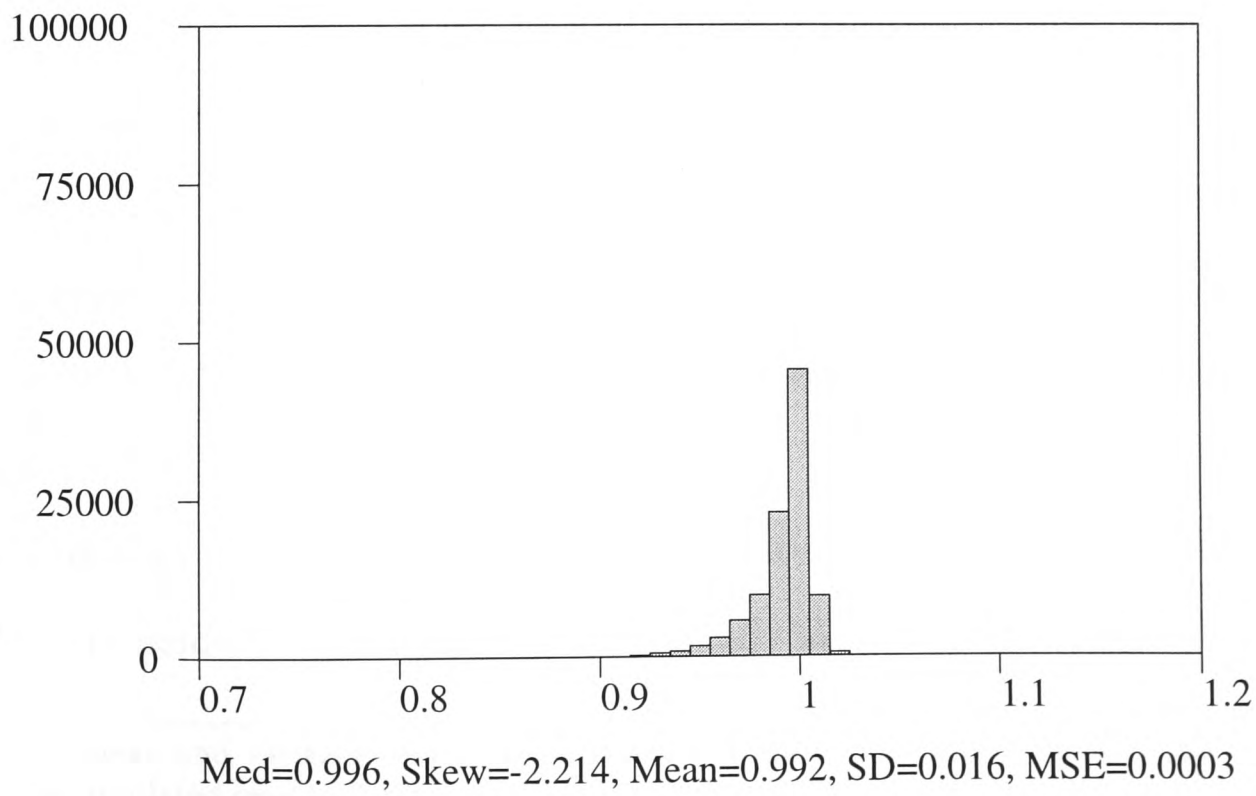


Figure 7.11 Empirical distribution of $\hat{\psi}_{ad}$ for $\psi = 1$, $\phi_1 = 0.5$, and $T = 100$ (AR(2)).

models can be compared. We shall consider these results to clarify the more detailed analyses which are scattered in the previous sections. The statistics have been calculated from simulated generations of a random walk which is a process encompassed by all of the models in the table (the statistics for the $AR_\mu(1)$ and $AR_\mu^B(1)$ models have been reported already in Figures 7.2 and 7.7, respectively). The simple AR(1) model is included in the comparison, too. It serves as a benchmark because it does not encompass any nuisance parameters on top of the variance parameter.[36] The AE and the MLE match or there is no adjustment under this model.

For all of the models the bias, the SD, and the MSE of the AE decrease uniformly from the upper-most row to the last row. The $AR_\mu(1)$ model is in a category of its own, of course for worse, with respect to each of the quantities reported. The statistics for the $AR_\mu^B(1)$ model are clearly smaller but not yet quite at the level of the statistics for the simple AR(1) and AR(2) models. The latter yield fairly similar and the most pleasing estimates. It is noteworthy that the AE is more accurate under the AR(2) model than under the simple AR(1) as pointed out in the previous section (Table 7.32).

The corresponding optimally multiplicatively corrected estimator (OMCE) $a_T^m\,\hat{\psi}_{ml}$ is reported in the second column of the table.[37] We shall employ the OMCE as a benchmark to throw light on the performance of the AEs in relation to another correction which is valid under a unit root, too. Abadir (1995) did not suggest that the OMCE should be used in general so no contest of estimates is taking place. (His main point was that minimising MSE by a multiplicative correction returns an almost unbiased estimate under a unit root.)

Table 7.34 does not have a column for the statistics of the MLE, the AE_i, or the AE_p for the $AR_\mu(1)$ model but we shall comment briefly on them, too. The statistics for the

[36]The exact mean and variance under the simple AR(1) model are known (Evans and Savin (1981)). We report the simulated ones to facilitate comparison with the other statistics so that all of the statistics (except the ones for the AR(2) model) are based on the same random numbers.

[37]Formulae (5.14) and (5.16) for the bias (b_T^m), the SD (SD_T^m), and the minimum achievable MSE (MSE_T^m) of the OMCE hold in general so the OMCE is a sensible estimate for all of the models. The multiplying factors for the four models from the uppermost to the lowest in Table 7.34 are 1.053, 1.021, 1.017, and 1.017, respectively. We have not included the unit-root $AR_\mu(1)$ model in the comparison because the asymptotic distribution of the MLE varies with the magnitude of the drift.

MLE and AE_i have been given in Figures 7.1 and 7.3 but the figures for the AE_p referred to below are not documented in the thesis.

The OMCE is slightly more dispersed than the AE (or the MLE and as skewed as the MLE) but the bias and the MSE are clearly the smallest under a unit root, at least when the model is $AR_\mu(1)$. The AE_i features the largest variance. It might be more fair to compare the OMCE to the AE_{ap} because both employ information of a unit root. The latter estimate is better, indeed the best estimate, by any of the criteria: the bias, the variance and the MSE of it are all zero to three decimal places. (The MSE is about one fifteenth of the MSE of the MLE.)

When the model is $AR^B_\mu(1)$ the OMCE again possesses by far the smallest bias but the AE is distributed more compactly leading to a slightly smaller SD and MSE. The Cox–Reid adjustment can thus beat the OMCE — or at least achieve comparable accuracy as the difference in the MSEs is minor — by the criterion the latter focuses on or MSE.

The OMCE for the AR(1) model is less biased and more variable than the corresponding AE (the MLE). It is the most accurate estimate in terms of MSE in the table.

Finally, when the model is AR(2) with a zero ϕ_1 the bias of the OMCE is far smaller but the SD of it is larger than of the AE. The MSE of the OMCE is yet smaller.

For all of the models studied, the OMCE is the least biased and the most dispersed estimate (in terms of variance). In three cases out of the four examined, the MSE of the OMCE is much smaller than that of the AE. In one case ($AR^B_\mu(1)$) the AE is the more accurate estimate in terms of MSE but the difference is marginal. The OMCE obviously has merit in the unit-root circumstance. On the other hand, even though the AE is not as accurate an estimate under a unit root it is valid under stationarity, too.

	AE			OMCE		
Model	**Bias**	**SD**	**MSE**	**Bias**	**SD**	**MSE**
$\mathbf{AR}_{\mu}(1)$	-3.649	4.384	0.325	-0.206	4.539	0.206
$\mathbf{AR}_{\mu}^{\mathbf{B}}(1)$	-1.936	3.568	0.165	-0.168	4.101	0.168
$\mathbf{AR}(1)$	-1.746	3.116	0.128	-0.100	3.168	0.100
$\mathbf{AR}(2)$	-1.694	3.103	0.125	-0.104	3.219	0.104

Table 7.34 The bias, the SD, and the MSE ($\times$ 100) of the AE and the OMCE for a random walk and $T = 100$.

Chapter 8

Discussion

8.1 General Comments

In many cases, the Cox–Reid adjustment appears to focus on the bias of the MLE. The models we have studied are categorised below in terms of existence of bias and of the adjustment. Category *A* is composed of models where the existence of the adjustment integrates with the existence of bias. Models for which the adjustment fails to account for the bias — asymptotically or entirely — are allocated to Category *B*. The possibility that the bias may disappear for a specific value of the parameter is ignored in the categorisation.

CATEGORY *A*.

A1) A bias and an adjustment exist in both finite and large samples in the

- estimation of the variance parameter in the presence of incidental mean parameters
- estimation of the autoregressive coefficient of the nonstationary AR(1) model with (zero) constant.

A2) A bias and an adjustment exist only in finite samples in the

- estimation of the variance parameter of the AR(1) model
- estimation of the variance parameter of the AR(1) model with constant
- estimation of the constant of the Bhargava AR(1) model with constant
- estimation of the autoregressive coefficient of the stationary AR(1) model with constant
- estimation of the autoregressive coefficient of the stationary Bhargava AR(1) model with constant.

- estimation of the autoregressive coefficient of the unit-root AR(1) model with drift.

A3) No bias nor an adjustment arise at all in the

- estimation of the coefficient of the inverse of the time trend.

CATEGORY *B.*

B1) A bias exists in finite and large samples but the adjustment fades asymptotically in the

- estimation of the autoregressive coefficient of the unit-root Bhargava AR(1) model with constant.
- estimation of the sum of the autoregressive coefficients of the unit-root AR(2) model.

B2) A bias exists in finite and large samples and yet no adjustment arises at all in the

- estimation of the autoregressive coefficient of the unit-root AR(1) model.

B3) A bias exists in finite though not in large samples but no adjustment arises at all in the

- estimation of the autoregressive coefficient of the stationary AR(1) model.

Models and parameters of interest vary in Category *A*. Instead, Category *B* is dominated by unit-root models with the autoregressive coefficient or the sum of them as the parameter of interest. However, also the stationary AR(1) model falls into this category which suggests that the failure to account for the bias may relate to the parameter of interest (the autoregressive coefficient) rather than to the stationarity status of the model. On the other hand, the bias is not too severe for this model in general (Section 5.3). A bias correction (for median-unbiasedness) for the AR(1) model did not seem feasible in Andrews (1993) either, despite that a correction was presented for an AR(1) model with constant. Likewise the innovation permutation argument of Rothenberg (1995), referred to in Section 5.4, fails to modify the estimate (personal communication). Relatedly but not indicated in either of the categories, the adjustment seems to ignore the bias when the

autoregressive coefficient of the AR(1) model with constant lies in the stationary region $(-1, -1/3]$. The adjustment is often linked to existence of bias but other properties are involved, too.

In the case of the simple AR(1) another interpretation on the focus of the adjustment is that the existence of it is related rather to problems in estimation than to problems in testing. An adjustment arises when the nuisance parameter is the autoregressive coefficient but not when the nuisance parameter is the variance. The former is a nuisance in both respects. The latter is a scale parameter and as such not a problem from the point of view of estimation but is a nuisance from the point of view of testing.

A drawback of the adjustment is that an increase of the variance of the estimate or Wald statistic can arise both asymptotically and especially in finite samples as with estimation of the autoregressive coefficient of the AR(1) model with constant, say. The AEs can still remain more accurate in terms of MSE. Delightfully, the unit-root Bhargava AR(1) with constant demonstrates that a simultaneous cutback in both bias and variance of the MLE can be achieved with the adjustment. This can happen also with the unit-root AR(1) with constant model, and the unit-root AR(2) model when *a priori* information is made use of in the estimation of the unit root. The decrease and increase in variance under the AR(1) model with constant and the AR(2) model, respectively, fit the the arguments of MacKinnon and Smith (1996) on the effect of an bias correction when they are combined with the suggestive bias formulae of Section 5.3.

In a nutshell, if an adjustment exists then it is useful in reducing bias for most of the models we have studied. It can reduce MSE, too, even asymptotically, but this is necessarily not the case. The properties of the adjustment in terms of estimation and testing are considered in more depth in the following two sections.

We have uncovered three models under which coefficient c of the Cox–Reid adjustment (of the 1993 article) is zero though the parameters are not orthogonal. Such cases arise in the

- estimation of the coefficient of the inverse of the time trend

- estimation of the autoregressive coefficient of the Bhargava AR(1) model with constant when the starting value x_0 of the x_t process beneath the y_t process is not zero (and x_0 is properly taken account of in the likelihood).
- estimation of the constant of the Bhargava AR(1) model (with constant) when the starting value x_0 of the x_t process beneath the y_t process is not zero (and x_0 is properly taken account of in the likelihood).

We have not reported the calculation for the last case but it should seem fairly natural by the second case. If the Bhargava models possess a unit root then a common feature with these models is that the parameter of interest is not identified asymptotically. However, under stationarity there is no identification problem with the parameters of the Bhargava models so that is not the reason. Relatedly, we have found that the coefficient can behave mirror wise with the orthogonality measure under the AR(2) model.

It is also interesting that coefficient c, appropriately standardised, converges to a distribution instead of a constant when the model is the unit-root AR(1) with zero constant. The property derives from the fact that information measures evaluated at the MLE, after standardisation, do not necessarily tend to constants either when the model has a unit root.

Some researchers might favour a non-symmetrical criterion instead of MSE to measure accuracy of estimates of a unit root because empiricists often have little faith in explosive processes. Our results should appear then the more favourable when the model is AR(1) with constant or AR(2). If estimates larger than unity were transferred to unity then the MSEs of the AEs would decrease more than the MSE of the MLE because the former take larger values with greater frequency. In the case of the Bhargava AR(1) model with constant the effect of such a convertion is not clear on the relative superiority of the AE because the adjustment cuts both tails in finite samples under this model. The AE would of course remain the more accurate in absolute terms in finite samples because of the shortened left tail.

A related issue which we have not pondered is how good or bad a property is skewness

of the distributions of estimates when the focus is on an autoregressive parameter close to unity. The adjustment seems to tend to increase skewness but in the light of the previous paragraph some may not consider this a problem.

A possible weakness of the adjustment, which has only been touched on, is the effect of misspecification of the model. The adjusted statistics faced somewhat more serious trouble than the unadjusted when the effect of ignoring the magnitude of the starting value x_0 was scrutinized under the Bhargava AR(1) model with constant. The finding may apply more generally because the adjustment makes use of information measures which are adequate in general only when the model is correctly specified.

We have considered Wald tests instead of LM or LR tests. The LM test may run into serious trouble when the likelihood is multimodal as it may be in the applications of adjusted profile likelihood we have looked at. The LM test is based on the slope of the likelihood function at the null value of the parameter under test and an artificial mode may occur at that value causing a modest slope and a weak or inconsistent test. A Wald test is based on the 'horizontal' distance of the MLE and the value of the parameter under the null hypothesis so many modes are unconsequential from the point of view of this test as long as the highest mode occurs at the true value of the parameter. The LR test behaves, at the outset, similarly to the Wald test, except that it should be weaker under some circumstances.

The Cox–Reid adjustment is applicable under a unit-root time-series model even though it was not originally designed for such a case. We shall list a few potential applications in the final section.

8.2 Discussion on Estimation and Related Distributional Results

The adjustment to the profile likelihood fades or persists asymptotically when the model is the stationary or nonstationary (undrifted) AR(1) model with constant, respectively, and the parameter of interest is the autoregressive coefficient. We have focused on the

nonstationary case. The asymptotic distribution is then shifted to the right on the real axis, and the right tail is stretched, among other things. The outcome is that the AE is less biased, more accurate in terms of MSE, features a less skewed distribution than the MLE, but also a slightly larger variance. The distribution is modified relatively more strongly when the sample size is finite. The AE is easy to calculate as it depends solely on two simple statistics: $\hat{\psi}_{ml}=\hat{\psi}_{ols}$ and $\hat{l}= [\sum_{t=1}^{T}(y_{t-1}- \bar{y}_{-1})^2]^{-1} \sum_{t=1}^{T}(y_t- \bar{y})^2$.

A very concise distribution would result if the *a priori* information $\psi = 1$, or that the unit-root parameter ψ equals unity, were used in the estimation of coefficient c of the adjusted profile likelihood. A simple way to describe this adjustment, for large T, is that $T\sigma^2/2$ is added to the numerator of the ratio defining the MLE. Such an adjustment has been independently discovered by Rothenberg (1995). Unfortunately this route would lead to inconsistent estimates in general or when $\mid \psi \mid < 1$.

It is remarkable that the AE features a smaller MSE than the MLE not only in finite samples but also asymptotically under a unit root. The usual regularity conditions do not apply under nonstationarity which makes this possible. Relatedly, the AE achieves a decrease in bias without a notable increase in variance over that of the MLE (at least when $\psi = 1$). Cox and Llatas (1991) and Lucas (1996, Chapter 7) have pointed out other (nearly) nonstationary cases in which the MLE can be beaten in terms of MSE.

Departing from classical analysis, Zivot (1994) has derived the posterior distribution of the Bayes estimate for ψ based on the Jeffreys prior $|I(\omega^*)|^{1/2}$ where $I(\omega^*)$ is the information matrix for $\omega^{*\prime} = [\psi\ \alpha\ \sigma]$. The distribution features a long right tail. There is some parallelism to the Cox–Reid adjustment which is based on the information matrix, too, and which stretches the right tail of the distribution of the MLE as well.

We note in passing that the Cox–Reid adjustment for the MLE of the autoregressive coefficient seems to perform relatively better, in terms of MSE, than the adjustment for the MLE of the variance parameter. In the latter case the bias decreases too but apparently at the cost of an increase in MSE (the examples on pp. 41–43).

The idea of calculating the Cox–Reid AE iteratively is introduced. Iteration makes

sense only, at least in asymptotic terms, if the model is nonstationary so the adjustment does not fade with T. Otherwise the iterated estimate would feature the same asymptotic distribution as the original standardised AE or the standardised MLE. The median of the iteratively calculated AE for the unit-root AR(1) model with constant lies closer to unity than that of the original Cox–Reid AE. The iterated estimate seems to lack moments asymptotically when it is calculated without restricting coefficient c to the theoretical range $(0, (T-1)/2]$. The variance is inflated in finite samples, too, but not severely. Nevertheless, the MSE of the iterated estimate is smaller than the MSE of the MLE for a small sample size like one hundred.

There is no benefit asymptotically of using the Cox–Reid adjustment when the model is the unit-root AR(1) model with drift but nor is there any bias to correct when T is large. The small-sample distributions of $T^{3/2}(\hat{\psi}_{ml} - 1)$ and corresponding Wald statistic for the null of a unit root can be very distorted though. The AEs and the companion Wald statistics achieve often a much better match with the asymptotic Normals expanding the usefulness of the result but probably not by as much as we would like to. It is pleasing that the adjustment yields a better match with the asymptotic Normal also under a unit root just as it does under IID or stationary models which have been previously considered in the literature.

The Bhargava AR(1) model with constant presents an elegant way to treat nuisance parameters and deserves attention as such. The more simple adjusted profile log-likelihood $l_{ad}(\psi)$ is serviceable because coefficient c equals zero under this model. It is interesting that the asymptotic distribution of $T(\hat{\psi}_{ml} - 1)$ (and of $T(\hat{\psi}_{ad} - 1)$) is the most simple DF distribution (3.11) if $\psi = 1$. That is to say, estimation of the constant does not affect the asymptotic distribution of the standardised MLE (or standardised $\hat{\psi}_{ad}$) even in the unit-root case. There is hence less need and room for correcting bias with this model than for the standard AR(1) with constant but the emerging adjustment is lucid and intuitive. The adjustment can reduce both the the bias and the variance of the MLE under this model. Interestingly, the adjustment increases or decreases the

MLE if it is smaller or larger than one, respectively. The associated sort of degrees-of-freedom correction reinforces the adjustment which means that it reduces the estimate in the latter case — contrary to the standard increasing impact when estimating residual variance. The adjusted estimate and the adjusted Wald statistic follow the asymptotic distribution better than their unadjusted counterparts.

The magnitude of the adjustment is not as easily grasped with this model because we have not developed an analytic expression for the AE. Scrutinisation of the adjusted profile log-likelihood suggests that the adjustment is comparable with the one for the ordinary AR(1) model with constant when the models are stationary. A unit root seems to cut down the adjustment with the Bhargava model, though the opposite happened with the ordinary AR(1) model with constant.

Our analysis of the AR(2) model concentrates on the unit-root circumstance but asymptotics enables us to comment on the stationary case, too. The adjustment appears asymptotically comparable in size to the case with the Bhargava AR(1) model with constant in both situations but in small samples the magnitude of the adjustment depends greatly on the ϕ_1 parameter of the AR(2) process. The adjustment is able to decrease the bias, the SD and the MSE for most of the parameter space under a unit root. (If ϕ_1 is very large then MSE may be boosted by overadjusting the MLE but in practice estimates larger than unity would probably be transferred to unity.) Simulations showed that the bias follows the pattern suggested by the tentative unit-root formula (5.11) derived from the expressions for the stationary AR(2) model. As already noted, the adjustment yields a decrease in variance (for most of the experimented values) in line with the theory of MacKinnon and Smith (1996) on bias corrections (after taking account of the aforementioned formula for the bias under a unit-root AR(2) model). The benefits of the adjustment are larger, of course, when the sample size is relatively small because the adjustment fades asymptotically. The magnitude of the adjustment apparently depends on an interplay of the size of coefficient c and of the Fisher information for the autoregressive parameter which have an opposite effect. It is delightful that the decrease in MSE is

yet in general more forceful for positive values of ϕ_1 because such values should be empirically relevant, e.g. the process $y_t = 1.3y_{t-1} - 0.3y_{t-2} + \epsilon_t$ ($\phi_1 = 0.3$) is the time-series process for US data in general (after trend removal) according to Orcutt (1948).

Moreover, the adjustment is so effective that the AE beats the MLE from the simple AR(1) model in these respects. The result potentially implies that one should never calculate the MLE from the simple AR(1) because the AE from the (unit-root) AR(2) model is more accurate even when the autoregression is first order only and because the model allows for greater generality via the second order autoregression. To justify the claim one should inspect the behaviour of the MSE, say, of the AE under stationarity. As pointed out in Section 6.6 the adjustment — even though derived under a unit root — behaves sensibly also under stationarity so the claim has merit on *a priori* grounds. On the other hand, the magnitude of the ϕ_1 parameter is far more important in determining the distribution of the estimate than whether the estimate is adjusted or not which is less pleasing.

The adjustment is easily calculated if the formula $c \approx (1 - \phi_1)^{-2}$ is employed. In general the AE calculated in this shortcut way (SAE) performs even better than the AE. However, complex roots of the adjusted likelihood equation or overshooting estimates can arise when the ϕ_1 parameter of the AR(2) model is large but the sample size is not. The adjustment which employs the exact form of c seems thus advisable.

The models are ranked in decreasing order in terms of the vigour of the adjustment in Table 8.1.[1] The model for which the Cox–Reid adjustment persists asymptotically is the AR(1) model with constant under a unit root and zero constant. (The original, *a priori*, and the iterative adjustment for the model are not differentiated in the table as they are of the same order of magnitude.) This is the model where the MLE of the unit-root parameter is the most biased, for which coefficient c of the Cox–Reid adjustment (equation (4.11)) does not tend to a constant (even after a standardisation), and in general the sole model where a superfluous nuisance parameter (the constant) is estimated. The

[1]The placement of the Bhargava models is somewhat tentative as explained above.

Magnitude	Model
$O_p(1)$	The unit-root $\mathrm{AR}_\mu(1)$ (with zero constant)
$O_p(T^{-1/2})$	The stationary $\mathrm{AR}_\mu(1)$ model The stationary $\mathrm{AR}_\mu^B(1)$ The stationary AR(2) The unit-root $\mathrm{AR}_\mu(1)$ with drift
$O_p(T^{-1})$	The unit-root $\mathrm{AR}_\mu^B(1)$ The unit-root AR(2)
No adjustment	The simple AR(1) (stationary or not)

Table 8.1 The order of magnitude of the Cox–Reid adjustment to the standardised MLE.

models for which asymptotic normality applies to the MLE of the unit-root parameter (the stationary models and the unit-root AR(1) model with drift) feature an $O_p(T^{-1/2})$ adjustment term. The adjustment appears the smallest for the unit-root Bhargava AR(1) model with constant and the AR(2).

The rate of convergence of the MLE does not alone explain the strength of the adjustment term. There are unit-root models — for which the rate of convergence of the MLE is at least $O_p(T^{-1})$ instead of the more universal $O_p(T^{-1/2})$ — in all categories of Table 8.1. The extreme magnitudes of the adjustment arise under unit-root models, though.

The analysis of the unit-root models points out how different in nature the nuisance parameters α (the constant) of the AR(1) model with constant and ϕ_1 (the second autoregressive parameter) of the AR(2) model are. The adjustment peaks for the former parameter (under $\alpha = 0$) but reaches the bottom value for the latter parameter (unrestricted) at $\psi = 1$ (according to asymptotics). Moreover, the adjustment does not depend on the size of α with the AR(1) model with constant but varies greatly with ϕ_1 with the AR(2) model.

We cannot compare the adjustment for the Bhargava AR(1) model with constant with the others in this fashion because coefficient c equals zero for the model. However, the relatively small magnitude of the adjustment for that model and the AR(2) model might be due to the fact that both can be written as a *common-factor model* (Sargan (1964), Hendry and Mizon (1978)). This can be seen from the representations

$$(1 - \psi B)y_t = (1 - \psi B)\gamma_0 + \epsilon_t$$

for the Bhargava process and

$$(1 - \psi B)y_t = \phi_1(1 - B)y_{t-1} + \epsilon_t$$

for the AR(2) process, for $t = 2, ..., T$. A common-factor restriction applies to the former equation but also to the latter when $\psi = 1$. This kind of a representation does not exist for the ordinary AR(1) model with constant.

It is of interest, too, that the statistic $\hat{c}\ (\hat{\psi}^2_{ml} - \hat{l})/(T - 3)$ follows asymptotically — for the two models where we have studied its behaviour — the same distribution as the AE (after standardisation). These models are the unit-root AR(1) model with constant equal to zero and the unit-root AR(1) model with drift. The asymptotic distribution of the former is nonstandard; the latter shares the asymptotic Normal distribution with the MLE.

The most compact distribution we have come across — unless distributions which emerge by letting the starting value tend to infinity (Appendix A5) are taken account of — is the asymptotic distribution (6.38) of the *a priori* adjusted Wald statistic. The variance of this distribution is much less than one.

The OMCE or optimally multiplicatively corrected estimate of the form $a_T\ \hat{\psi}_{ml}$ by Abadir (1995) is clearly less biased than the AE in the four cases we have looked at: the simple AR(1) (for which the AE is the MLE), the AR(1) model with constant, the Bhargava AR(1) model with constant and the simple AR(2) model when the sample size is a hundred (and a unit root exists). The OMCE possesses inevitably a larger variance but it is still more accurate in terms of MSE under the first, second, and fourth models. However, under the Bhargava model the AE is not only less variable but also slightly more accurate in terms of MSE. The OMCE employs information of a unit root and appears valid only under it which contrasts with the AE, so the comparison is not too

unfavourable for the AE. Indeed, the AE_{ap} — which is also a reasonable estimate only under a unit root — features the smallest bias, variance, and MSE of any of the estimates we have considered (when a unit root exists in the AR(1) model and the sample size is a hundred).

The MLEs and the AEs are consistent as the starting value tends to infinity under all of the AR(1) models when the models are correctly specified. (We have not examined the AE_i for the $\mathrm{AR}_\mu(1)$ model in this sense so the statement does not necessarily apply to it.) An exception is that the MLE of the autoregressive coefficient is not uniquely defined when the model is Bhargava AR(1) with constant, the starting value x_0 is incorporated in the likelihood and the starting value tends to infinity. It is not surprising that if x_0 is not taken account of in the likelihood but is let to tend to infinity then the MLE or the AE are in general inconsistent. Despite the simplicity of the proofs, the results appear new except for the simple AR(1) model.

8.3 Discussion on Testing

Returning to the AR(1) model with (zero) constant, the $\psi_{\mu,ad}$ test, or the test based on the statistic $T(\hat{\psi}_{ad2} - 1)$, is comparable in power though slightly weaker than the ψ_μ test, based on the DF test statistic $T(\hat{\psi}_{ml} - 1)$, if the process starts from the unconditional mean. If the standardised starting value or the deviation of the starting value from the unconditional mean (of the time series) divided by the standard deviation (of the innovation) is moderate then the $\psi_{\mu,ad}$ test achieves greater power than the ψ_μ test, and the difference in power can be large if the deviation is sizable. Indeed, combinations of parameter values and standardised starting values exist for which the power of the ψ_μ test collapses to zero but the power of the $\psi_{\mu,ad}$ test is boosted to one. The properties of the $\psi_{\mu,ad}$ test are amplified if the test is based on the iterative estimate or the $\psi_{\mu,ad,i}$ test is implemented. This test is the weakest (but not by much) if the starting value is the mean, but the gain in power in other cases is magnified.

The same pattern of power emerges with the Wald-type-of statistics, though the dif-

ferences in power are smaller than for the corresponding tests based on the standardised coefficients. The adjusted versions of both tests perform relatively better for large standardised starting values.

The power of the τ_μ test (or of the $\tau_{\mu,ad}$ test) tends to one regardless of the value of ψ (for the experimented values) whereas the power of the ψ_μ test (or of the $\psi_{\mu,ad}$ test) fades to zero, for many values of ψ, as the starting value diverges from the unconditional mean of the time series. This difference in the relative performance of the two standard tests appears to have been disregarded in the literature. Instead, the power advantage of the ψ_μ test over the τ_μ test, when the starting value and unconditional mean concur, has been pointed out.

The results on power can be neatly rationalised to a notable extent by the already referred finding that the estimates converge to the true values as the starting value relative to the unconditional mean tends to infinity. The property peculiarly leads to inconsistent (standardised coefficient) tests in many cases.

There are other circumstances too when the power of the Wald tests is known to vary mirror-wise with the power of the tests based on the standardised coefficient. For example, the power of the ρ_τ test decreases with the magnitude of the drift in the process whereas the power of the τ_τ test increases as the drift increases (when the drift is positive, cf. Guilkey and Schmidt (1991, Table 1)).

The tests utilizing *a priori* information of a unit root display zero power when the process has started from the unconditional mean which is not surprising after recalling that the *a priori* AE is inconsistent in general. A more surprising feature is that these tests can display power equal to one, even when the power of the corresponding non-adjusted statistic is zero, when the starting value deviates notably from the unconditional mean. The explanation lies in the aforementioned convergence of the estimates — including the AE_{ap} — and test statistics based on them as the divergence of the starting value from the unconditional mean tends to infinity, and by the smaller critical values in absolute terms of the adjusted forms.

One may have an idea on the possible size of the standardised starting value or be concerned with a range of values of it, because of the deterioration in power for a large deviation, say. Then a test can be selected according to Table 7.12 reproduced here for convenience:

$\|y_0 - \mu\|/T\sigma$	**Suggested test**
0	ψ_μ
0.02–0.05	$\psi_{\mu,ad}$
≥ 0.1	$\psi_{\mu,ad,ap}$ or $\tau_{\mu,ad,ap}$

Table 7.12 Suggested test for different values of $|y_0 - \mu|/T\sigma$ for $|y_0 - \mu|/\sigma$ in the range $[0, 250]$ and T in the range $[25, 250]$ ($\mathrm{AR}_\mu(1)$).

Already a battery of two tests, the ψ_μ and the $\tau_{\mu,ad}$ tests, say, would assure in general, though not always, competent power across the scrutinized $(\psi, |y_0 - \mu|/\sigma)$ set. If the circumstance $|y_0 - \mu|/\sigma \approx 0$ were deemed unrealistic, then a useful battery of tests would be $\tau_{\mu,ad}$ and $\tau_{\mu,ad,ap}$ which do not imply drastic losses of power for any of the experimented combinations of $(\psi, |y_0 - \mu|/\sigma)$ with a relatively small $|y_0 - \mu|/\sigma$ but achieve the highest powers for situations with a substantial $|y_0 - \mu|/\sigma$. Of course, a problem is how to fix the size if two tests are employed. If a sole test would have to be chosen for general use then the τ_μ or the $\tau_{\mu,ad}$ test would be a reasonable choice as they are the only tests which are consistent whether it is T or $|y_0 - \mu|/\sigma$ which tends to infinity. An important drawback would be that quite large losses in power (relative to any of the standardised coefficient tests) would arise when the process has started from the mean (Tables 7.8 and 7.9). A single test out of the considered ones is not capable of handling all the cases we have pondered. Nevertheless, empiricists often calculate only the Wald statistic. We have given some theoretical justification for this practice but pointed out the just mentioned flaw with it.

The formation of power is less systematic for the Bhargava AR(1) model with constant. Table 7.26, repeated below, captures much of the information in the power simulations. If the model is correctly specified, or the starting value x_0 is zero, then either the

ψ_μ^B test or the $\tau_{\mu,ad}^B$ test are recommended. (The latter test dominates strictly in power when the sample size is 25.) Thus despite the asymptotically negligible adjustment, a gain in power can result. All the tests perform poorly when the starting value differs from zero and the model is misspecified. The powers of the tests decrease then, and the ψ_μ^B test is favoured instead of a Wald test as with the $\mathrm{AR}_\mu(1)$ model.

$\lvert x_0\rvert/T\sigma$	**Suggested test**
0	$\tau_{\mu,ad}^B$ or ψ_μ^B
≥ 0.02	ψ_μ^B

Table 7.26 Suggested test for different values of $|x_0|/T\sigma$ for $|x_0|/\sigma$ in the range $[0, 250]$ and T in the range $[25, 250]$ ($\mathrm{AR}_\mu^B(1)$).

As remarked, the estimates do not in general converge to the true values as the starting value (here x_0) diverges to infinity under the then misspecified Bhargava AR(1) model with constant. The (possibly adjusted) profile likelihood develops a local maximum around unity. This leads to (in general) biased estimates and powerless tests. It also means that care is needed in the estimation of the parameters of the model. Plotting the (perhaps adjusted) profile likelihood can reveal that the numerical algorithm has converged to a local maximum even though the likelihood takes larger values in the range of values of interest or that the shape of the likelihood is otherwise curious.

A natural question is which model, $\mathrm{AR}_\mu^B(1)$ or $\mathrm{AR}_\mu(1)$, is to be to employed for testing? The question is well posed only when the alternatives are such that $x_0/\sigma = [y_0 - \alpha/(1-\psi)]/\sigma$ in which case the time series under the two different models agree (for $t = 1, 2, \ldots$, p. 112). Table 7.27, repeated below, gives a recommendation for such a case. When the conditions take value zero, the time series generated by the $\mathrm{AR}_\mu(1)$ model has started from the unconditional mean. Tests based on the Bhargava model are then favoured. Otherwise the Bhargava model, as we have devised it, is misspecified and tests based on the $\mathrm{AR}_\mu(1)$ model should be used. The power differences in favour of one or the other model can be considerable. However, x_0 is unobservable and implementing the suggestion in practice is hard.

In the case of a zero starting value and 25 observations, it is surprising that the power of the ψ_μ^B test does not necessarily exceed the power of the ψ_μ test. The power increases more rapidly with the number of observations under the Bhargava model so for larger sample sizes the corresponding powers are nevertheless greater under it. Wald statistics behave more consistently with expectations as the powers emerge larger under the Bhargava model throughout (apart from a case with equal power) and especially so for distant alternatives.

$\lvert x_0\rvert/T\sigma = [y_0 - \alpha/(1-\psi)]/T\sigma$	**Suggested test**
0	$\tau_{\mu,ad}^B$ or ψ_μ^B
0.02–0.05	$\psi_{\mu,ad}$
≥ 0.1	$\psi_{\mu,ad,ap}$ or $\tau_{\mu,ad,ap}$

Table 7.27 Suggested test for different values of $|x_0|/T\sigma = [y_0 - \alpha/(1-\psi)]/T\sigma$ for $|x_0|/\sigma$ in the range $[0, 250]$ and T in the range $[25, 250]$ ($\mathrm{AR}_\mu(1)$ and $\mathrm{AR}_\mu^B(1)$).

The adjustment does not improve the fit (of the Wald statistics) with the asymptotic distribution when the model is AR(2) so it is not helpful from the point of view of testing under this model.

8.4 Suggestions for Theoretical Research

Most importantly, the performance of the derived AEs under stationarity should be checked. There is good reason to expect useful results because that is the case for which the Cox–Reid adjustment has been constructed in the first place. The variant of the adjustment which we derived for the unit-root AR(2) model should also perform reasonably under stationarity by the arguments presented in Section 6.6.

It might be beneficial to restrict the estimate of coefficient c to lie in the theoretical range $(0, (T-1)/2]$ when calculating $\hat{\psi}_{ad2,i}$. The modification should cut the increase in variance from the occurring exaggerated estimates of c due to the iteration. On the other hand, the asymptotic distribution (6.34) would not hold then. Another counterargument

is why not restrict the space of values which the estimates themselves are allowed to take. The two other ways to iterate an AE for the AR(1) model with constant — mentioned in Section 6.3 — could be studied, as well.

An omnibus-test statistic could be constructed by combining, say linearly, a standardised coefficient test statistic and a Wald test statistic into a new statistic. The fractiles of the new statistic could be simulated and employed as critical values of the new test. The test should possess two desirable features: i) It should be more powerful than the Wald test when the time series has started from the unconditional mean and ii) Its power should tend to unity as the standardised starting value tends to infinity (thus eliminating the failure with the standardised coefficient test statistic).

An obvious way to continue with the theoretical analysis would be to add a time trend to the model or apply the adjusted profile likelihood to model (3.6). The parameters γ_0 and γ_1 together increase even more the bias and the variance of $\hat{\psi}_{ml}=\hat{\rho}_{ml}$ than when only a constant is included in the model.[2] The simulations of Section 7.3 provide further reason for optimism with respect to the analysis of this model. The power of the ρ_τ test or the test based on the statistic $T(\hat{\psi}_{ml}-1)$ (when a constant and a time trend are included in the regression) depends on the magnitude of the drift in the process as pointed out above (a further reference is Evans and Savin (1984)). If the result for the AR(1) model with constant — that the AE achieves greater power for most values of the entity with respect to which the test is invariant — then the AE-based test would be more powerful for most values of the drift parameter. (ψ_{ad} tests are invariant with respect to the starting value and achieve greater power than the MLE-based counterpart for most starting values or their departures from the long run mean.)

The complicated algebra in the derivation of the adjustment does not encourage one to generalize the Cox–Reid technique to AR(p) models. On the basis of the asymptotic form of coefficient c, $(1-\phi_1)^{-2}$, one might speculatively examine whether an adjusted profile likelihood like (6.67) with $c=(1-\phi_1-\cdots-\phi_p)^{-2}$ would be useful for the AR(p)

[2]Helpfully, Phillips (1991) has derived some of the related information measures.

model, though. The most promising path for expanding the scope of our results to AR(p) models might be to study whether the present techniques could be applied to residuals whitened by OLS from AR dynamics over that of an AR(1).

The nuisance parameter θ in the ARMA(1,1) model (3.21) almost ruins the present tests so adjusted profile likelihood has potential in this framework as well.[3] However, the asymptotic results of Faust (1993, 1994), referred to in Section 3.5, throw some doubt on the usefulness of adjusting the MLE in this context.

As regards univariate stationary models, it would be interesting to compare the performance of the Cox–Reid AE with the bias-corrected estimate $\hat{\rho}_b$ (equation (5.6)). The null of (trend) stationarity could also be pondered. Evans and Savin (1984) studied the performance of the statistic $(\hat{\psi}_{ml} - \psi)/[T/(1-\psi^2)]$ which is asymptotically Standard Normal under the null of (trend) stationarity, and found the Normal approximation very poor in small samples. It could be worthwhile to study how much better the AE follows the asymptotic Normal as it has been used in this way before and even in this thesis in the context of a drifted random walk.

A further theoretical advancement would be to generalise the analysis to the multivariate case. An interesting possibility is *seemingly unrelated regression* (SUR) where there are a number of AR(1) relations with potentially varying coefficients and errors which correlate across equations. Unit-root testing in the context of SUR models can lead to large gains in power (Abuaf and Jorion (1990)) so the case merits attention. Such an application would lead us to study if we could generalise the Cox–Reid theory to the case of many parameters of interest (the AR coefficients) in this particular case. Other instances are the *fixed effects model* and the *random effects model* where the constants may vary across equations which share a common autoregressive coefficient but the disturbances are independent across equations. Such panel data can lead to asymptotic

[3] On the unpleasant side, likelihood functions of ARMA processes with MA terms are complicated. Also information measures are intricate if models involve many AR and/or MA terms. An idea worth investigating might be to apply computerized algebra (Wolfram (1991)) if the analysis were expanded to more complex ARMA(p,q) models. The Mathematica package for analysis of ARIMA models provided by Stine (1992) could be useful when proceeding in this direction.

normality of the estimates (Levin and Lin (1992)). Power gains for testing of unit roots can be expected from the relatively more informative panel data as such. On the other hand, the Cox-Reid adjustment could again serve to increase the match with the asymptotic Normal distributions. As mentioned in Section 4.2, Cruddas *et al.* (1989) achieved good results in this sense with the adjustment for stationary panel data.

8.5 Suggestions for Empirical Research

Potential empirical applications for unit-root tests, including for the ones we have devised from the Bhargava AR(1) model, are numerous and need not be indicated. Instead, the case for unit-root tests, the power function of which is especially favourable against gradually stationary alternatives (asymptotically stationary processes which have started fairly far from the unconditional mean), is not obvious. The adjusted tests we have derived from the AR(1) model with constant fit into this class. Gradually stationary behaviour might result from convergence to a saturation or minimum, perhaps in some sense equilibrium, level of the variable. The examples below aim to make the point concrete.

- Saturation. The starting value of a time series of sales (or inventories) of a product equals zero at the time of the introduction of the commodity, and there may well be a saturation level of sales (or inventories) for the commodity (cf. with the concept of a life cycle of a product). A service industry like tourism may be limited in both ways, say. The product could also be of financial character. Consumption of a good might possess a saturation level, too. An example is consumption of food or calories which many people try to limit to an appropriate non-excessive amount. There might also exist also a saturation level for social phenomena like crime or unemployment which a society is willing to tolerate.
- Bounded variables provide a further motivation for the relevance of saturation. Numerous economic variables are proportions or more generally ratios of two underlying economic factors so that the denominator is larger than the numerator.

Such rations are bounded and strictly speaking they cannot be integrated. An ARIMA model may yet describe such a variable adequately in small samples (in terms of distributions of statistics, say). Trend-like behaviour of such a variable might be explained theoretically more satisfactorily as convergence to a long run or saturation level from a small (or large) starting value. Specific potential applications follow. Wagner's hypothesis asserts that the share of government expenditure as a proportion of GNP tends to increase with the level of the GNP (cf. Henrekson (1990) or Koop and Poirier (1995), say). The share may well converge to a saturation level and the tests developed in this thesis should be more prone to detect this than the DF tests presently in use. An application similar in spirit would be a study of inequality measures. Hayes *et al.* (1989) were unable to reject a unit root in the inequality measure they inspected. Again, convergence to a saturation level could be a reasonable alternative and would motivate the use of the tests developed here. An example different in spirit is migration on which limitations are sometimes imposed by the receiving countries.

- Convergence to a minimum or towards zero. Consumption of an inferior good in a growing economy is a natural example; another is (a potentially large) decrease in consumption because of emergence of close substitutes or a change in preferences. A further example occurs when public authorities, say, try to push occurrence of an undesired thing or event like accidents or infant deaths towards zero. Eliminating the event altogether may be impossible but convergence to a (stochastic) minimum level may be achievable.
- Convergence in general. Behaviour of relative wages of women related to those of men and internationalisation of consumption patterns could exemplify convergence behaviour.[4] The currently popular growth theory provides an interesting example. An economy with an increasing *per capita* stock of capital faces decreasing returns in production according to typical neoclassical growth theory. Consequently coun-

[4]According to Maula (1996), convergence between the eating habits between Europe and the USA takes place slowly. The study of Heikkinen and Maula (1996) inspired the examples related to food.

tries with a relatively smaller capital stock should grow faster than more opulent countries leading to convergence of the relative GNPs. Incompatibility of data with the hypothesis might be regarded as evidence in favour of modern endogenous growth theory. In a number of studies, the null of no convergence among a set of countries has been tested by regressing average growth of GNP on the initial level of it (in logarithms). Quah (1993a,b) and Bernard and Durlauf (1996) have severely criticised the interpretability of such a regression. Instead, the null of no convergence could be tested for by studying if unit roots existed in time series of pairwise ratios of (logarithms of) GNPs of different countries (perhaps after choosing a country as a benchmark). Ben-David and Rahman (1996) and Bernard and Jones (1993), among others, have conducted unit-root tests in a related fashion with panel data. According to Bernard and Durlauf (*op. cit.*, p. 163) such tests have in general lead to acceptance of the null of no convergence. They also suspect (*op. cit.*, p. 171) that the tests are likely to face poor power when the economies inspected are in a state of transition. (Our simulations show — when transition is interpreted to mean convergence to the mean level — that this applies to the ψ_μ test though not to the τ_μ test.) The adjusted versions of the DF tests we have developed display especially large power in such a case so they might prove useful in the empirical analysis of growth.

Appendix A1
Proofs for the AR(1) Model with Constant

We shall first prove the small-sample property that:

- the local maxima (if they exist) of $l_{ad2}(\psi)$ arise at the points expressed in definition (6.18) of $\hat{\psi}_{ad2}$.

For this purpose, we shall also prove that:

- $\hat{\psi}_{ml}^2 - \hat{l} \leq 0$ for all sample sizes.

Next, we focus on asymptotics and show that

- $\hat{c} = \begin{cases} O_p(1) & \text{if } \mid \psi \mid < 1 \\ O_p(T) & \text{if } \psi = 1 \end{cases}$
- $\hat{\psi}_{ml}^2 - \hat{l} = (\hat{\psi}_{ml} - 1)(\hat{\psi}_{ml} + 1) + O_p(T^{-1}) = \begin{cases} O_p(1) & \text{if } \mid \psi \mid < 1 \\ O_p(T^{-1}) & \text{if } \psi = 1 \end{cases}$ and
- $T\left(\hat{\psi}_{ml}^2 - \hat{l}\right) \Rightarrow -\left\{\left[\int_0^1 W_*(r)\right]^2 dr\right\}^{-1}$.

These results enable us to:

- derive the approximation $\hat{\psi}_{ad2} \approx \hat{\psi}_{ml} - \hat{c}\,(\hat{\psi}_{ml}^2 - \hat{l})/(T-3)$, the order of accuracy of it, and confirm consistency, or that $\hat{\psi}_{ad2} \xrightarrow[p]{} \psi$
- prove that the root $\hat{\psi}_{ml} + \frac{T-3}{2\hat{c}} + \sqrt{\cdot}$ is $O_p(T)$ or $O_p(1)$ and lies within $(1, \infty)$, if $\mid \psi \mid < 1$ or $\psi = 1$, respectively
- prove that $l_{ad2}(\psi)$ has asymptotically a (local) maximum at $\hat{\psi}_{ad2}$, or at the root $\hat{\psi}_{ml} + \frac{T-3}{2\hat{c}} - \sqrt{\cdot}$ by verifying that $\partial^2 l_{ad2}(\psi)/\partial^2 \psi \xrightarrow[p]{} -\infty$ at the root
- prove that $l_{ad2}(\psi)$ is asymptotically flat or has a (local) minimum at the root $\hat{\psi}_{ml} + \frac{T-3}{2\hat{c}} + \sqrt{\cdot}$ by verifying that $\partial^2 l_{ad2}(\psi)/\partial^2 \psi \xrightarrow[p]{} 0$ or $\xrightarrow[p]{} \infty$ if $\mid \psi \mid < 1$ or $\psi = 1$, respectively, at the root.

The last result is supplementary: The finite sample argument for the local maximum of $l_{ad2}(\psi)$ and the fact that $\hat{c}$ is asymptotically positive determine already the local maximum, if it exists, to occur asymptotically at $\hat{\psi}_{ml} + \frac{T-3}{2\hat{c}} - \sqrt{\cdot}$ (as defined in case *iv*) below). The second last finding is not needed in the determination of the root corresponding to the maximum either, because of the just stated reason, but it proves that a (local) maximum exists asymptotically.

The derivations allow for $y_0 \neq 0$; the accuracy of some of the approximations would improve if $y_0 = 0$ were assumed. The fact that $\hat{\psi}_{ml}$ tends in probability to ψ is taken as granted in this and the appendices that follow. The appendix concludes with a summary of the findings.

Local maximum of $l_{ad2}(\psi)$ in finite samples

For convenience, cases *i)* to *v)* as defined on p. 81 are repeated here:

i) $\hat{c} < 0$ and the roots are complex (with nonzero imaginary parts)
ii) $\hat{c} < 0$ and the roots are real (and unequal)
iii) $\hat{c} = 0$ (in which case the adjusted likelihood equation implies a polynomial of order one)
iv) $\hat{c} > 0$ and the roots are real (and unequal)
v) $\hat{c} > 0$ and the roots are complex (with nonzero imaginary parts).

Circumstances *ii)* and *iv)* are considered here as the reasoning for the other cases was provided already in the main text.

The adjusted likelihood equation (6.16) is also repeated here:

$$\frac{\partial l_{ad2}(\psi)}{\partial \psi} = \frac{(T-3)(\hat{\psi}_{ml} - \psi)}{\hat{l} - 2\psi\hat{\psi}_{ml} + \psi^2} + \hat{c} = 0$$

where $\hat{l} = \left[\sum_{t=1}^{T}(y_{t-1} - \bar{y}_{-1})^2\right]^{-1} \sum_{t=1}^{T}(y_t - \bar{y})^2$. Evaluating the derivative $\partial l_{ad2}(\psi)/\partial\psi$ at the MLE of ψ gives

$$\left[\frac{\partial l_{ad2}(\psi)}{\partial \psi}\right]_{\psi=\hat{\psi}_{ml}} = \hat{c}$$

assuming that $\hat{l} - \hat{\psi}_{ml}^2 \neq 0$ which is the case with probability one for finite T. The adjusted log-likelihood is hence downward or upward sloping at $\psi = \hat{\psi}_{ml}$ if $\hat{c} < 0$ or $\hat{c} > 0$, respectively.

Note next that definitions (6.6) and (6.17) and the Cauchy–Schwarz inequality imply that

$$\begin{aligned} \hat{\psi}_{ml}^2 - \hat{l} &= \frac{\left[\sum_{t=1}^{T}(y_t - \bar{y})(y_{t-1} - \bar{y}_{-1})\right]^2 - \sum_{t=1}^{T}(y_t - \bar{y})^2 \sum_{t=1}^{T}(y_{t-1} - \bar{y}_{-1})^2}{\left[\sum_{t=1}^{T}(y_{t-1} - \bar{y}_{-1})^2\right]^2} \\ &\leq 0. \end{aligned} \tag{A1.1}$$

(The '<' sign applies with probability one for finite T.)

It follows from inequality (A1.1) that the roots

$$\hat{\psi}_{ml} + \frac{T-3}{2\hat{c}} \pm \sqrt{\frac{(T-3)^2}{4\hat{c}^2} + \hat{\psi}_{ml}^2 - \hat{l}}$$

of equation (6.16) lie both to the left of $\hat{\psi}_{ml}$ or to the right of $\hat{\psi}_{ml}$ if $\hat{c} < 0$ or $\hat{c} > 0$, respectively. If $\hat{c} < 0$ then the root $\hat{\psi}_{ml} + (T-3)/2\hat{c} - \sqrt{\cdot}$ is the left-most root. If $\hat{c} > 0$ then $\hat{\psi}_{ml} + (T-3)/2\hat{c} + \sqrt{\cdot}$ is the right-most root.

These results enable us to differentiate which of the two roots corresponds to the local maximum and which to the local minimum. The adjusted log-likelihood $l_{ad2}(\psi)$ implies a cubic equation, the derivative of which has two different real roots by assumption. The adjusted log-

likelihood must then feature a local maximum and a local minimum. The two possible cases are *ii*) which assumes $\hat{c}< 0$ and *iv*) which assumes $\hat{c}> 0$. Let *ii*) apply. The roots of the adjusted likelihood equation lie then to the left of $\hat{\psi}_{ml}$. Because $l_{ad2}(\psi)$ is downward sloping at $\hat{\psi}_{ml}$ if $\hat{c}< 0$, the local maximum must relate to the root closer to $\hat{\psi}_{ml}$ or the root $\hat{\psi}_{ml} +(T-3)/2\ \hat{c} +\sqrt{\cdot}$. Let *iv*) apply. The roots of the adjusted likelihood equation are then larger than $\hat{\psi}_{ml}$. Because $l_{ad2}(\psi)$ is upward sloping at $\hat{\psi}_{ml}$ if $\hat{c}> 0$, the local maximum must associate with the root closer to $\hat{\psi}_{ml}$ or the root $\hat{\psi}_{ml} +(T-3)/2\ \hat{c} -\sqrt{\cdot}$. Of course, the local minima connect to the other roots, respectively. It is shown below that asymptotically only case *iv*) is relevant.

If $\psi = 1$ then asymptotically $\hat{\psi}^2_{ml} - \hat{l}$ equals zero in probability (though the expression must be negative for finite samples), $\hat{c}$ tends stochastically to infinity with T (cf. below), the root $\hat{\psi}_{ml} +(T-3)/2\ \hat{c} -\sqrt{\cdot}$ is asymptotically appropriate, and converges stochastically to $\hat{\psi}_{ml}$, and $[\partial^2 l_{ad2}(\psi)/\partial^2\psi]_{\psi=\hat{\psi}_{ml}}$ tends stochastically to minus infinity (cf. below). The fact that the adjusted log-likelihood becomes infinitely edged at $\psi =\hat{\psi}_{ml}$ provides intuition to the simultaneous phenomena of the root approaching $\hat{\psi}_{ml}$ and the first derivative at $\psi =\hat{\psi}_{ml}$ seeming to become infinitely large.

Basic asymptotic results

The coefficient c evaluated at $\hat{\omega}_{ml}= [\hat{\psi}_{ml}\ \hat{\phi}_{ml}]'$ is:

$$\begin{aligned}\hat{c} &= \frac{1}{1-\hat{\psi}_{ml}} - \frac{1-\hat{\psi}^T_{ml}}{T(1-\hat{\psi}_{ml})^2} \\ &= \begin{cases} O_p(1) & \text{if } |\psi|<1 \\ \frac{T(1-\hat{\psi}_{ml})-1+\hat{\psi}^T_{ml}}{T(1-\hat{\psi}_{ml})(1-\hat{\psi}_{ml})} = \frac{O_p(1)}{O_p(1)O_p(T^{-1})} = O_p(T) & \text{if } \psi = 1 \end{cases}\end{aligned} \tag{A1.2}$$

(cf. formula (6.14) and p. 89 where it is shown that $\hat{\psi}^T_{ml}$ is $O_p(1)$). Furthermore, $\hat{c}> 0$ for T large, as $c > 0$ for the parameter space of ψ and $\hat{\psi}_{ml} \xrightarrow[p]{} \psi$ as shown later.

We go on to our second task. First note that:

$$\hat{\psi}^2_{ml} - \hat{l} = (\hat{\psi}_{ml} -1)(\hat{\psi}_{ml} +1) - (\hat{l} -1).$$

The last term can be written as (cf. definition (6.17)):

$$\hat{l} -1 = \frac{\sum_{t=1}^T (y_t - \bar{y})^2 - \sum_{t=1}^T (y_{t-1} - \bar{y}_{-1})^2}{\sum_{t=1}^T (y_{t-1} - \bar{y}_{-1})^2} = \frac{\sum_{t=1}^T y_t^2 - T(\bar{y})^2 - \left[\sum_{t=1}^T y_{t-1}^2 - T(\bar{y}_{-1})^2\right]}{\sum_{t=1}^T (y_{t-1} - \bar{y}_{-1})^2}.$$

It is well known that the denominator is $O_p(T)$ or $O_p(T^2)$ if $|\psi|< 1$ or $\psi = 1$, respectively. Some cancellations have to be taken into account when considering the order of magnitude of the numerator. It equals

$$\sum_{t=1}^{T} y_{t-1}^2 + y_T^2 - y_0^2 - T\left[\bar{y}_{-1} + T^{-1}(y_T - y_0)\right]^2 - \sum_{t=1}^{T} y_{t-1}^2 + T(\bar{y}_{-1})^2$$
$$= y_T^2 - y_0^2 - 2\,\bar{y}_{-1}\,(y_T - y_0) - T^{-1}(y_T - y_0)^2$$
$$= y_T(y_T - 2\,\bar{y}_{-1}) - y_0(y_0 - 2\,\bar{y}_{-1}) - T^{-1}(y_T - y_0)^2.$$

If $|\psi| < 1$ then the numerator is clearly $O_p(1)$. If $\psi = 1$ then it can be expressed as

$$y_T(y_T - 2\,\bar{y}_{-1}) + O_p(T^{1/2}).$$

The term in the first parentheses appears $O_p(T)$ but we shall confirm that no further cancellations occur. We note first that:

$$\begin{aligned}\bar{y}_{-1} &= T^{-1}\sum_{t=2}^{T}\sum_{s=1}^{t-1}\epsilon_s + T^{-1}y_0 \\ &= \sum_{t=1}^{T-1}(T-t)\epsilon_t + T^{-1}y_0 \\ &= \sum_{t=1}^{T-1}\epsilon_t - \sum_{t=1}^{T-1}(t/T)\epsilon_t + T^{-1}y_0.\end{aligned}$$

The formula enables us to write

$$\begin{aligned}T^{-1/2}(y_T - 2\,\bar{y}_{-1}) &= T^{-1/2}\left[\left(y_0 + \sum_{t=1}^{T}\epsilon_t\right) - 2\left(\sum_{t=1}^{T-1}\epsilon_t - \sum_{t=1}^{T-1}(t/T)\epsilon_t + T^{-1}y_0\right)\right] \\ &= -\left[T^{-1/2}\sum_{t=1}^{T-1}\epsilon_t - 2T^{-1/2}\sum_{t=1}^{T-1}(t/T)\epsilon_t\right] + O_p(T^{-1/2}). \\ &\Rightarrow N(0, \sigma^2/3)\end{aligned}$$

as

$$\begin{bmatrix} T^{-1/2}\sum_{t=1}^{T-1}\epsilon_t \\ T^{-1/2}\sum_{t=1}^{T-1}(t/T)\epsilon_t \end{bmatrix} \Rightarrow N(\mathbf{0}, \sigma^2\mathbf{Q})$$

where

$$\mathbf{Q} = \begin{bmatrix} 1 & 1/2 \\ 1/2 & 1/3 \end{bmatrix}.$$

(E.g. Hamilton (1994, pp. 458–60) gives a derivation for the convergence to $N(\mathbf{0}, \sigma^2\mathbf{Q})$). It follows that the numerator is $O_p(T)$.

Orders of magnitude of the numerator and the denominator imply that $\hat{l} - 1 = O_p(T^{-1})$ for both circumstances $|\psi| < 1$ and $\psi = 1$. Hence it appears that

$$\begin{aligned}\hat{\psi}_{ml}^2 - \hat{l} &= (\hat{\psi}_{ml} - 1)(\hat{\psi}_{ml} + 1) - O_p(T^{-1}) \\ &= \begin{cases} O_p(1) & \text{if } |\psi| < 1 \\ O_p(T^{-1}) & \text{if } \psi = 1. \end{cases}\end{aligned} \tag{A1.3}$$

In principle, subtraction could diminish the order of magnitude of the individual terms in the nonstationary case. We confirm below that this does not happen. Such a cancellation of orders of magnitude takes place under $\alpha \neq 0$ and $\psi = 1$ (Appendix A2).

We shall assume in the rest of the section that $\psi = 1$ and effectively give a separate proof for the statement $\hat{\psi}_{ml}^2 - \hat{l}$ is $O_p(T^{-1})$ under a unit root. The former arguments are needed, too, for the analysis of the asymptotic form of $T(\hat{l} - 1)$ which is considered in Section 6.3.

A first-order Taylor-series expansion for $\hat{\psi}_{ml}^2$ around unity is

$$\hat{\psi}^2_{ml} = 2\hat{\psi}_{ml} - 1 + r_{1,T}, \quad r_{1,T} = O_p(T^{-2}),$$

where the exact form of $r_{1,T}$ is $(\hat{\psi}_{ml} - 1)^2$. It follows, by recalling definitions (6.6) and (6.17), that

$$\begin{aligned}
\hat{\psi}^2_{ml} - \hat{l} &= 2\hat{\psi}_{ml} - 1 - \hat{l} + r_{1,T} \\
&= \frac{2(\sum_{t=1}^T y_t y_{t-1} - T\,\bar{y}_{-1}\bar{y}) - \left[\sum_{t=1}^T y_{t-1}^2 - T(\bar{y}_{-1})^2\right] - \left[\sum_{t=1}^T y_t^2 - T(\bar{y})^2\right]}{\sum_{t=1}^T (y_{t-1} - \bar{y}_{-1})^2} \\
&\quad + r_{1,T}
\end{aligned}$$

where we have expressed sums like $\sum_{t=1}^T (y_{t-1} - \bar{y}_{-1})^2$ as $\sum_{t=1}^T y_{t-1}^2 - T(\bar{y}_{-1})^2$. We shall consider first the order of magnitude of the numerator, next the magnitude of the denominator, and finally the magnitude of the ratio itself.

We shall apply the algebra in Appendix A2 where $\alpha \neq 0$ but the model is otherwise the same. Formula (A2.4) gives the exact form for the above numerator under $\alpha \neq 0$. The formula applies here as well if we set $\alpha = 0$. This yields the expression

$$-2\left(\frac{1}{2}y_T^2 - \sum_{t=1}^T y_{t-1}\epsilon_t\right) + T^{-1}(y_T - y_0)^2 + y_0^2$$

for the numerator. This simplifies to

$$-\sum_{t=1}^T \epsilon_t^2 + O_p(1)$$

after substituting $\sum_{t=1}^T y_{t-1}\epsilon_t = (1/2)\left(y_T^2 - y_0^2 - \sum_{t=1}^T \epsilon_t^2\right)$ (equation (A2.3) under $\alpha = 0$) and conducting some cancellations. We can now see that the numerator divided by T is, and tends to,

$$\begin{aligned}
&T^{-1}\left\{2(\textstyle\sum_{t=1}^T y_t y_{t-1} - T\,\bar{y}_{-1}\bar{y}) - \left[\sum_{t=1}^T y_{t-1}^2 - T(\bar{y}_{-1})^2\right] - \left[\sum_{t=1}^T y_t^2 - T(\bar{y})^2\right]\right\} \\
&= T^{-1}\left[-\textstyle\sum_{t=1}^T \epsilon_t^2 + O_p(1)\right] \\
&\underset{p}{\longrightarrow} -\sigma^2 < 0.
\end{aligned}$$

It is well known that the denominator

$$T^{-2}\textstyle\sum_{t=1}^T (y_{t-1} - \bar{y}_{-1})^2 \;\Rightarrow\; \sigma^2 \int_0^1 [W_*(r)]^2\,dr$$

(as in formula (6.8)). It follows that

$$T\left(\hat{\psi}^2_{ml} - \hat{l}\right) \;\Rightarrow\; \frac{-\sigma^2}{\sigma^2 \int_0^1 [W_*(r)]^2\,dr} = -\left\{\int_0^1 [W_*(r)]^2\,dr\right\}^{-1} < 0.$$

$\hat{\psi}^2_{ml} - \hat{l}$ is hence $O_p(T^{-1})$ under $\psi = 1$.

Additionally, we note that

$$\hat{\psi}_{ml}^2 - \hat{l} \xrightarrow[p]{} \begin{cases} \psi^2 - 1 < 0 & \text{if } |\psi| < 1 \\ 0 & \text{if } \psi = 1. \end{cases}$$

Derivation of the approximation $\hat{\psi}_{\mathbf{ad2}} \approx \hat{\psi}_{\mathbf{ml}} - \hat{\mathbf{c}}\,(\hat{\psi}_{\mathbf{ml}}^2 - \hat{\mathbf{l}})/(\mathbf{T-3})$

The previous results enable us to derive the expression (6.19). It is based on the first order Maclaurin expansion of $\sqrt{1+x} = 1 + \frac{1}{2}x + r_1$, which is convergent for $|x| \leq 1$ and where the remainder term $r_1 = -\frac{1}{8}(1+ax)^{-3/2}x^2$ where $a \in (0,1)$ by Taylor's theorem. This and the previous results applied to definition (6.18) case *iv)* (which applies for large T, cf. the discussion at the beginning of this appendix) give:

$$\begin{aligned}
\hat{\psi}_{ad2} &= \hat{\psi}_{ml} + \frac{T-3}{2\hat{c}} - \sqrt{\frac{(T-3)^2}{4\hat{c}^2} + \hat{\psi}_{ml}^2 - \hat{l}} \\
&- \hat{\psi}_{ml} + \frac{T-3}{2\hat{c}} - \frac{T-3}{2\hat{c}}\sqrt{1 + \frac{4\hat{c}^2}{(T-3)^2}(\hat{\psi}_{ml}^2 - \hat{l})} \\
&= \hat{\psi}_{ml} + \frac{T-3}{2\hat{c}} - \frac{T-3}{2\hat{c}}\left[1 + \frac{2\hat{c}^2}{(T-3)^2}(\hat{\psi}_{ml}^2 - \hat{l}) + r_{1,T}\right], \\
&\qquad r_{1,T} = \begin{cases} O_p(T^{-4}) & \text{if } |\psi| < 1 \\ O_p(T^{-2}) & \text{if } \psi = 1 \end{cases} \\
&= \hat{\psi}_{ml} - \frac{\hat{c}\,(\hat{\psi}_{ml}^2 - \hat{l})}{T-3} + r_T, \qquad r_T = \begin{cases} O_p(T^{-3}) & \text{if } |\psi| < 1 \\ O_p(T^{-2}) & \text{if } \psi = 1 \end{cases} \\
&\xrightarrow[p]{} \psi.
\end{aligned}$$

The convergence is due to $\hat{c}\,(\hat{\psi}_{ml}^2 - \hat{l})/(T-3)^2$ being $O_p(T^{-1})$ (regardless of whether $|\psi| < 1$ or $\psi = 1$) and the consistency of $\hat{\psi}_{ml}$. We note that $4\,\hat{c}^2\,(\hat{\psi}_{ml}^2 - \hat{l})/(T-3)^2$ is $O_p(T^{-2})$ or $O_p(T^{-1})$ if $|\psi| < 1$ or $\psi = 1$, respectively. Consequently, the convergence condition of the Maclaurin expansion holds for large T.

Asymptotic behaviour of the root $\hat{\psi}_{\mathbf{ml}} + \frac{\mathbf{T-3}}{\mathbf{2\hat{c}}} + \sqrt{\cdot}$

It is easy to see, by the above Maclaurin expansion, say, that $\sqrt{\frac{(T-3)^2}{4\hat{c}^2} + \hat{\psi}_{ml}^2 - \hat{l}}$ is $O_p(T)$ if $|\psi| < 1$ but is $O_p(1)$ if $\psi = 1$. The previous results make it then obvious that $\hat{\psi}_{ml} + \frac{T-3}{2\hat{c}} + \sqrt{\frac{(T-3)^2}{4\hat{c}^2} + \hat{\psi}_{ml}^2 - \hat{l}}$ is $O_p(T)$ and that it diverges to infinity if $|\psi| < 1$ but is $O_p(1)$ if $\psi = 1$. More specifically, it can be seen in the unit-root case that

$$\hat{\psi}_{ml} + \frac{T-3}{2\hat{c}} + \sqrt{\frac{(T-3)^2}{4\hat{c}^2} + \hat{\psi}_{ml}^2 - \hat{l}} \;\Rightarrow\; 1 + \left[\frac{\exp(DF_\mu) - (1+DF_\mu)}{(DF_\mu)^2}\right]^{-1} \in (1,\infty)$$

as

$$\hat{c}\,/(T-3) \;\Rightarrow\; \frac{\exp(DF_\mu) - (1+DF_\mu)}{(DF_\mu)^2} \in (0,\infty)$$

(Section 6.3), $\hat{\psi}_{ml} \xrightarrow[p]{} 1$, and $\hat{\psi}_{ml}^2 - \hat{l} \xrightarrow[p]{} 0$. (If the a priori formula $\hat{c} = (T-1)/2$ were used then $\sqrt{\frac{(T-3)^2}{4\hat{c}^2} + \hat{\psi}_{ml}^2 - \hat{l}}$ and $\hat{\psi}_{ml} + \frac{T-3}{2\hat{c}} + \sqrt{\frac{(T-3)^2}{4\hat{c}^2} + \hat{\psi}_{ml}^2 - \hat{l}}$ would tend in probability to 1 and to $1+1+1=3$, respectively, if $\psi = 1$.) In conclusion, the root lies outside the parameter space of interest $(-1, 1]$ whether $|\psi| < 1$ or $\psi = 1$ (and whether the data based or ***a priori*** formula for $\hat{c}$ is used).[1]

Asymptotic behaviour of $\partial^2 l_{ad2}(\psi)/\partial^2\psi$ at the root $\hat{\psi}_{ml} + \frac{T-3}{2\hat{c}} - \sqrt{\ }$.

We have now the tools to prove the final two of the statements with which we started the appendix. It follows from equation (6.16) that

$$\frac{\partial^2 l_{ad2}(\psi)}{\partial^2\psi} = \frac{(T-3)(\psi^2 - \hat{l} - 2\psi\hat{\psi}_{ml} + 2\hat{\psi}_{ml}^2)}{\left(\hat{l} - 2\psi\hat{\psi}_{ml} + \psi^2\right)^2}. \tag{A1.4}$$

Evaluating equation (A1.4) at the root

$$\hat{\psi}_{ml} + \frac{T-3}{2\hat{c}} - \sqrt{\frac{(T-3)^2}{4\hat{c}^2} + \hat{\psi}_{ml}^2 - \hat{l}}$$

[1]The convergence of the other root to infinity in the stationary case contrasts with the well-known result for the fully stationary AR(1) model. The three roots of the likelihood equation for that model converge to constants (White (1961)).

produces (after some algebra)[2]

$$\left[\frac{\partial^2 l_{ad2}(\psi)}{\partial^2 \psi}\right]_{\psi=\hat{\psi}_{ml}+\frac{T-3}{2\hat{c}}-\sqrt{\cdot}}$$

$$= \frac{(T-3)\left\{2(\hat{\psi}_{ml}^2-\hat{l})+\frac{T-3}{\hat{c}}\left[\frac{T-3}{2\hat{c}}-\sqrt{\frac{(T-3)^2}{4\hat{c}^2}+\hat{\psi}_{ml}^2-\hat{l}}\right]\right\}}{\left\{\frac{T-3}{\hat{c}}\left[\frac{T-3}{2\hat{c}}-\sqrt{\frac{(T-3)^2}{4\hat{c}^2}+\hat{\psi}_{ml}^2-\hat{l}}\right]\right\}^2}$$

$$= \frac{(T-3)\left\{2(\hat{\psi}_{ml}^2-\hat{l})+\frac{T-3}{\hat{c}}\left[-\frac{\hat{c}(\hat{\psi}_{ml}^2-\hat{l})}{T-3}+r_T\right]\right\}}{\left\{\frac{T-3}{\hat{c}}\left[-\frac{\hat{c}(\hat{\psi}_{ml}^2-\hat{l})}{T-3}+r_T\right]\right\}^2}$$

$$= \frac{(T-3)\left[2(\hat{\psi}_{ml}^2-\hat{l})-(\hat{\psi}_{ml}^2-\hat{l})+r_T^*\right]}{\left[-(\hat{\psi}_{ml}^2-\hat{l})+r_T^*\right]^2}, \qquad r_T^* = O_p(T^{-2}) \text{ for } |\psi|<1 \text{ and } \psi=1$$

$$= \frac{(T-3)\left(\hat{\psi}_{ml}^2-\hat{l}+r_T^*\right)}{\left[-(\hat{\psi}_{ml}^2-\hat{l})+r_T^*\right]^2}$$

$$\xrightarrow[p]{} -\infty.$$

The second equality arises from the approximation derived above. The order of the remainder term r_T^* becomes apparent after observing that $r_T^* = \hat{c}^{-1}(T-3)r_T$. The divergence to minus infinity is due to $\hat{\psi}_{ml}^2 - \hat{l}$ being negative for finite T and being of larger order of magnitude than r_T^*. The divergence to infinity is boosted if $\psi = 1$ because the denominator of the last ratio tends stochastically to zero.

It has been proven that $[\partial^2 l_{ad2}(\psi)/\partial^2\psi]_{\psi=\hat{\psi}_{ml}+\frac{T-3}{2\hat{c}}-\sqrt{\cdot}} \xrightarrow[p]{} -\infty$. Accordingly, $l_{ad2}(\psi)$ has asymptotically an infinitely sharp (local) maximum at

$$\hat{\psi}_{ad2} = \hat{\psi}_{ml}+\frac{T-3}{2\hat{c}}-\sqrt{\frac{(T-3)^2}{4\hat{c}^2}+\hat{\psi}_{ml}^2-\hat{l}} \xrightarrow[p]{} \psi$$

(as shown above).

Asymptotic behaviour of $\partial^2 l_{ad2}(\psi)/\partial^2\psi$ at the root $\hat{\psi}_{ml}+\frac{T-3}{2\hat{c}}+\sqrt{\cdot}$

Evaluating equation (A1.4) at the root

$$\hat{\psi}_{ml}+\frac{T-3}{2\hat{c}}+\sqrt{\frac{(T-3)^2}{4\hat{c}^2}+\hat{\psi}_{ml}^2-\hat{l}}$$

[2]The notation '$\xrightarrow[p]{} -\infty$' stands for divergence in probability to minus infinity, cf. the notation list in the beginning of the thesis.

yields

$$\left[\frac{\partial^2 l_{ad2}(\psi)}{\partial^2 \psi}\right]_{\psi=\hat{\psi}_{ml}+\frac{T-3}{2\hat{c}}+\sqrt{\cdot}}$$
$$= \frac{(T-3)\left[2(\hat{\psi}^2_{ml}-\hat{l})+\frac{(T-3)^2}{2\hat{c}^2}+\frac{T-3}{\hat{c}}\sqrt{\frac{(T-3)^2}{4\hat{c}^2}+\hat{\psi}^2_{ml}-\hat{l}}\right]}{\left[\frac{(T-3)^2}{2\hat{c}^2}+\frac{T-3}{\hat{c}}\sqrt{\frac{(T-3)^2}{4\hat{c}^2}+\hat{\psi}^2_{ml}-\hat{l}}\right]^2}. \qquad \text{(A1.5)}$$

The asymptotics are straightforward. As argued above $\sqrt{\frac{(T-3)^2}{4\hat{c}^2}+\hat{\psi}^2_{ml}-\hat{l}}$ is $O_p(T)$ if $\mid \psi \mid < 1$ but tends in probability to 1 if $\psi = 1$. This, and the previous order calculations, enable us to express the right-hand side of equation (A1.5) as

$$\begin{cases} \frac{O_p(T^3)}{O_p(T^4)} = O_p(T^{-1}) & \text{if } \mid \psi \mid < 1 \\ \frac{O_p(T)}{O_p(1)} = O_p(T) & \text{if } \psi = 1. \end{cases}$$

Our calculations tell us that $[\partial^2 l_{ad2}(\psi)/\partial^2\psi]_{\psi=\hat{\psi}_{ml}+\frac{T-3}{2\hat{c}}+\sqrt{\cdot}}$ converges in probability to zero under $\mid \psi \mid < 1$ but diverges under $\psi = 1$. The divergence is to infinity because

$$(T-3)/\hat{c} \;\Rightarrow\; \left[\frac{\exp(DF_\mu)-(1+DF_\mu)}{(DF_\mu)^2}\right]^{-1} > 0$$

(Section 6.3),

$$\frac{(T-3)^2}{2\hat{c}^2}+\frac{T-3}{\hat{c}}\sqrt{\frac{(T-3)^2}{4\hat{c}^2}+\hat{\psi}^2_{ml}-\hat{l}} \;\Rightarrow\; \left[\frac{\exp(DF_\mu)-(1+DF_\mu)}{(DF_\mu)^2}\right]^{-2} > 0,$$

and so the ratio of the square bracket terms in equation (A1.5) is asymptotically composed of positive terms ($\hat{\psi}^2_{ml}-\hat{l}$ tends in probability to zero). Correspondingly, $l_{ad2}(\psi)$, evaluated at $\hat{\psi}_{ml}+\frac{T-3}{2\hat{c}}+\sqrt{\cdot}$, has asymptotically no curvature if the model is stationary but has an infinitely sharp minimum if a unit root exists.

Summary

The local maximum, if it exists, of $l_{ad2}(\psi)$ occurs at the root $\hat{\psi}_{ml}+(T-3)/2\,\hat{c}+\sqrt{\cdot}$ if $\hat{c} < 0$ or at the root $\hat{\psi}_{ml}+(T-3)/2\,\hat{c}-\sqrt{\cdot}$ if $\hat{c} > 0$. The latter case applies for large T.

The adjusted estimate converges in probability to the true ψ and the adjusted likelihood has asymptotically at that point an infinitely sharp (local) maximum. The behaviour of the other root, and the second derivative of the adjusted profile likelihood, depend on whether $\mid \psi \mid < 1$ or $\psi = 1$. In the first case, the root diverges stochastically to infinity and the adjusted profile likelihood becomes infinitely flat at the root. In the latter case, the root is $O_p(1)$ but lies in the set $(1, \infty)$, at which point the adjusted profile likelihood has asymptotically an infinitely sharp minimum. The root lies hence outside the parameter space of ψ or the set $(-1, 1]$ whether the model is stationary or not. The difference in the behaviour of the roots reflects the fact that the adjustment term of the profile likelihood fades in the first case but persists in the latter case. It is fairly intuitive that $l_{ad2}(\psi)$ becomes flat for faraway values of ψ under $\mid \psi \mid < 1$ as the discrepancy between $l_{ad2}(\psi)$ and $l_p(\psi)$ tends to become relatively smaller as T increases.

Appendix A2
Proofs for the Unit-Root AR(1) Model with Drift

The following two results were proved in Appendix A1:

- the local maxima (if they exist) of $l_{ad2}(\psi)$ arise at the points expressed in definition (6.18) of $\hat{\psi}_{ad2}$
- $\hat{\psi}^2_{ml} - \hat{l} \leq 0$ for all sample sizes.

The proofs apply as such for the drifted unit-root AR(1) model (6.40), as well, which is inspected in this appendix (i.e. it is assumed that $\alpha \neq 0$ and $\psi = 1$ throughout this appendix). The model is distinct asymptotically but not in exact small-sample results or algebra, from the $AR_{\mu}(1)$ model of Appendix A1.

We begin by proving the following asymptotic statements:

- $\hat{c}\ /T \underset{p}{\longrightarrow} 1/2$ or $\hat{c} = O_p(T)$
- $T^2(\hat{\psi}^2_{ml} - \hat{l}) \underset{p}{\longrightarrow} -12\sigma^2/\alpha^2 < 0$ or $\hat{\psi}^2_{ml} - \hat{l} = O_p(T^{-2})$.

The outcome $\hat{c}\ /T \underset{p}{\longrightarrow} 1/2$, or that $\hat{c}$ is asymptotically positive, together with the first two results above imply that the maximum of the adjusted profile likelihood takes place at the root $\hat{\psi}_{ml} + \frac{T-3}{2\hat{c}} - \sqrt{\cdot}$ (as defined in (6.18), case *iv*), see p. 81 for the definitions of the cases) for large T.

In the course of the derivations, we shall prove and make use of the supplementary results:

- $\hat{\psi}^T_{ml} = T(\hat{\psi}_{ml} - 1) + \left[T(\hat{\psi}_{ml} - 1)\right]^2 /2 + O_p(T^{-3/2})$ (implying that $\hat{\psi}^T_{ml} \underset{p}{\longrightarrow} 1$)
- $\hat{\psi}^2_{ml} = 2\,\hat{\psi}_{ml} - 1 + O_p(T^{-3})$.

The third and the fourth (bulleted) findings enable us to derive the approximation

- $\hat{\psi}_{ad2} \approx \hat{\psi}_{ml} - \hat{c}\,(\hat{\psi}^2_{ml} - \hat{l})/(T-3)$,

the order of accuracy of it, and confirm consistency, or that

- $\hat{\psi}_{ad2} \underset{p}{\longrightarrow} \psi$.

Finally, we prove that

- $T^{3/2}(\hat{l} - 1) \Rightarrow N(0, 48\sigma^2/\alpha^2)$.

The calculations allow for $y_0 \neq 0$.[1]

Basic asymptotic results

By a simple first-order Taylor-series expansion of $\log \hat{\psi}_{ml}$ around unity:

$$\log \hat{\psi}_{ml} = \hat{\psi}_{ml} - 1 + r_{1,T}, \quad r_{1,T} = O_p(T^{-3}),$$

where $r_{1,T} = -(1/2)\left[1 + a(\hat{\psi}_{ml} - 1)\right]^{-2} (\hat{\psi}_{ml} - 1)^2$ and $a \in (0,1)$. We have taken here — and will take in what follows — for granted that $\hat{\psi}_{ml} - 1$ is $O_p(T^{-3/2})$ (cf. formula (6.41)). Consequently,

$$\log \hat{\psi}_{ml}^{T} = T \log \hat{\psi}_{ml} = T(\hat{\psi}_{ml} - 1) + r^*_{1,T}, \quad r^*_{1,T} = O_p(T^{-2}).$$

On the other hand, a second-order Maclaurin-series expansion of e^x is $1 + x + x^2/2 + r_2$ where $r_2 = e^{ax}x^3/6$ and $a \in (0,1)$. Hence

$$\begin{aligned} \hat{\psi}_{ml}^{T} &= \exp(\log \hat{\psi}_{ml}^{T}) \\ &= 1 + \left[T(\hat{\psi}_{ml} - 1) + r^*_{1,T}\right] + \left[T(\hat{\psi}_{ml} - 1) + r^*_{1,T}\right]^2 /2 + r_{2,T}, \quad r_{2,T} = O_p(T^{-3/2}), \end{aligned}$$

or

$$\hat{\psi}_{ml}^{T} = 1 + T(\hat{\psi}_{ml} - 1) + \left[T(\hat{\psi}_{ml} - 1)\right]^2 /2 + r_T, \quad r_T = O_p(T^{-3/2}). \tag{A2.1}$$

$T(\hat{\psi}_{ml} - 1)$ is $O_p(T^{-1/2})$ under the present assumptions so we find that

$$\hat{\psi}_{ml}^{T} \xrightarrow[p]{} 1.$$

This contrasts sharply with the outcome for the unit-root $\mathrm{AR}_\mu(1)$ model with $\alpha = 0$ in which case $\hat{\psi}_{ml}^{T}$ converges weakly to the random variate $\exp(DF_\mu)$ (formula (6.27)).

Approximation (A2.1) enables us to consider the asymptotic behaviour of $\hat{c}\,/T$. Coefficient c evaluated at $\hat{\omega}_{ml} = [\hat{\psi}_{ml} : \hat{\phi}'_{ml}]'$ and standardized by T tends stochastically to one half:

$$\begin{aligned} \hat{c}\,/T &= \frac{-T(\hat{\psi}_{ml} - 1) - 1 + \hat{\psi}_{ml}^{T}}{\left[T(\hat{\psi}_{ml} - 1)\right]^2} \\ &= \frac{\left[T(\hat{\psi}_{ml} - 1)\right]^2 /2 + O_p(T^{-3/2})}{\left[T(\hat{\psi}_{ml} - 1)\right]^2} \qquad \text{(A2.2)} \\ &\xrightarrow[p]{} \frac{1}{2} > 0 \end{aligned}$$

[1] The curvature of the adjusted profile log-likelihood at the roots is not considered.

by formulae (6.14) and (A2.1). $\hat{c}\,/T$ tends thus to the theoretical limit of c/T (c evaluated at $\psi = 1$ is $(T-1)/2$). The asymptotics are quite different than under $\alpha = 0$ and $\psi = 1$ when $\hat{c}\,/T$ converged weakly to the random variate $[\exp(DF_\mu) - (1 + DF_\mu)]\,/\,(DF_\mu)^2$ (formula (6.28)).

We go on to prove that $T^2(\hat{\psi}^2_{ml} - \hat{l}) \underset{p}{\longrightarrow} -12\sigma^2/\alpha^2 < 0$ or that $\hat{\psi}^2_{ml} - \hat{l}$ is $O_p(T^{-2})$. To this end, we note two auxiliary formulae. First, we expand $\hat{\psi}^2_{ml}$ around unity:

$$\hat{\psi}^2_{ml} = 2\,\hat{\psi}_{ml} - 1 + r_{1,T}, \quad r_{1,T} = O_p(T^{-3}),$$

where the exact form of $r_{1,T}$ is $(\hat{\psi}_{ml} - 1)^2$ as in Appendix A1. Second, we find an alternative expression for the sum $\sum_{t=1}^{T} y_{t-1}\epsilon_t$. Note that

$$\begin{aligned} \textstyle\sum_{t=1}^{T} y_t^2 &= \textstyle\sum_{t=1}^{T}(\alpha + y_{t-1} + \epsilon_t)^2 \\ &= T\alpha^2 + \textstyle\sum_{t=1}^{T} y_{t-1}^2 + \sum_{t=1}^{T} \epsilon_t^2 + 2\alpha \sum_{t=1}^{T} y_{t-1} + 2\sum_{t=1}^{T} y_{t-1}\epsilon_t + 2\alpha \sum_{t=1}^{T} \epsilon_t. \end{aligned}$$

It follows that

$$y_T^2 - y_0^2 = T\alpha^2 + \textstyle\sum_{t=1}^{T} \epsilon_t^2 + 2\alpha \sum_{t=1}^{T} y_{t-1} + 2\sum_{t=1}^{T} y_{t-1}\epsilon_t + 2\alpha \sum_{t=1}^{T} \epsilon_t$$

or that

$$\textstyle\sum_{t=1}^{T} y_{t-1}\epsilon_t = \frac{1}{2}(y_T^2 - y_0^2) - \frac{1}{2}\left(T\alpha^2 + \sum_{t=1}^{T} \epsilon_t^2\right) - \alpha\left(\sum_{t=1}^{T} y_{t-1} + \sum_{t=1}^{T} \epsilon_t\right). \qquad \text{(A2.3)}$$

We are ready to analyze the formula $\hat{\psi}^2_{ml} - \hat{l}$. Substituting the above Taylor-series expansion for $\hat{\psi}^2_{ml}$, definitions (6.6) and (6.17) of $\hat{\psi}_{ml}$ and $\hat{l}$, respectively, to the subtraction $\hat{\psi}^2_{ml} - \hat{l}$ yields

$$\begin{aligned} \hat{\psi}^2_{ml} - \hat{l} &= 2\,\hat{\psi}_{ml} - 1 - \hat{l} + r_{1,T} \\ &= \frac{2(\sum_{t=1}^{T} y_t y_{t-1} - T\,\bar{y}_{-1}\bar{y}) - \left[\sum_{t=1}^{T} y_{t-1}^2 - T(\bar{y}_{-1})^2\right] - \left[\sum_{t=1}^{T} y_t^2 - T(\bar{y})^2\right]}{\sum_{t=1}^{T}(y_{t-1} - \bar{y}_{-1})^2} \\ &\quad + r_{1,T} \end{aligned}$$

where we have expressed sums like $\sum_{t=1}^{T}(y_{t-1} - \bar{y}_{-1})^2$ as $\sum_{t=1}^{T} y_{t-1}^2 - T\left(\bar{y}_{-1}\right)^2$. We shall first consider the numerator of this expression.

Substituting the formulae $y_t = \alpha + y_{t-1} + \epsilon_t$ and $\bar{y}_{-1} = \bar{y} + T^{-1}(y_T - y_0)$ into the numerator gives

$$-2\left(\frac{1}{2}y_T^2 - \sum_{t=1}^{T} y_{t-1}\epsilon_t - \alpha\sum_{t=1}^{T} y_{t-1}\right) + T^{-1}(y_T - y_0)^2 + y_0^2 \qquad \text{(A2.4)}$$

after carrying out some cancellations. This equals, by formula (A2.3),

$$-\left(T\alpha^2 + \sum_{t=1}^{T} \epsilon_t^2\right) + 2\alpha\sum_{t=1}^{T} \epsilon_t + T^{-1}\,(y_T - y_0)^2$$

after some cancellation. This expression, or the numerator, are $O_p(T)$ as

$$\begin{aligned} & T^{-1}\left[-\left(T\alpha^2+\textstyle\sum_{t=1}^{T}\epsilon_t^2\right)+2\alpha\textstyle\sum_{t=1}^{T}\epsilon_t+T^{-1}\left(y_T-y_0\right)^2\right] \\ & =-\alpha^2-T^{-1}\textstyle\sum_{t=1}^{T}\epsilon_t^2+T^{-2}y_T^2+O_p(T^{-1/2}) \\ & \underset{p}{\longrightarrow} -\alpha^2-\sigma^2+\alpha^2=-\sigma^2<0 \end{aligned}$$

where the convergence of $T^{-2}y_T^2$ to α^2 is obvious after writing $y_T = y_0 + \alpha T + \sum_{t=1}^{T}\epsilon_t$.

The denominator of the above ratio is $O_p(T^{-3/2})$ because

$$\begin{aligned} T^{-3}\left[\textstyle\sum_{t=1}^{T}(y_{t-1}-\bar{y}_{-1})^2\right] &= T^{-3}\textstyle\sum_{t=1}^{T}y_{t-1}^2-T^{-2}(\bar{y}_{-1})^2 \\ &\underset{p}{\longrightarrow} \alpha^2/3-(\alpha/2)^2 \end{aligned}$$

or

$$T^{-3}\left[\textstyle\sum_{t=1}^{T}(y_{t-1}-\bar{y}_{-1})^2\right] \underset{p}{\longrightarrow} \alpha^2/12$$

(Hamilton (1994, p. 460), say, proves the convergences).

It follows from the above analysis that

$$\begin{aligned} & T^2\left(\hat{\psi}_{ml}^2-\hat{l}\right) \\ & = \frac{T^{-1}\left\{2(\sum_{t=1}^{T}y_ty_{t-1}-T\,\bar{y}_{-1}\bar{y})-\left[\sum_{t=1}^{T}y_{t-1}^2-T(\bar{y}_{-1})^2\right]-\left[\sum_{t=1}^{T}y_t^2-T(\bar{y})^2\right]\right\}}{T^{-3}\left[\sum_{t=1}^{T}(y_{t-1}-\bar{y}_{-1})^2\right]} \\ & \quad +T^{-2}r_{1,T} \\ & \underset{p}{\longrightarrow} \frac{-\sigma^2}{\alpha^2/12} \end{aligned}$$

or that

$$T^2\left(\hat{\psi}_{ml}^2-\hat{l}\right) \underset{p}{\longrightarrow} -12\sigma^2/\alpha^2<0.$$

Thus $\hat{\psi}_{ml}^2-\hat{l}$ is $O_p(T^{-2})$. Note that in the alternative expression $(\hat{\psi}_{ml}-1)(\hat{\psi}_{ml}+1)-(\hat{l}-1)$ of $\hat{\psi}_{ml}^2-\hat{l}$ the first term is $O_p(T^{-3/2})$ as is the second (see the end of this section for the latter result). Some cancellation must hence occur between the terms to bring down the order of magnitude of $\hat{\psi}_{ml}^2-\hat{l}$ to $O_p(T^{-2})$. Such a cancellation does not arise under $\alpha = 0$ and $\psi = 1$ (cf. Appendix A1).

Derivation of the approximation $\hat{\psi}_{\mathbf{ad2}} \approx \hat{\psi}_{\mathbf{ml}} - \hat{\mathbf{c}}\,(\hat{\psi}_{\mathbf{ml}}^2 - \hat{\mathbf{1}})/(\mathbf{T}-\mathbf{3})$

The previous results enable us to find the order of magnitude to which approximation (6.19) holds under $\alpha \neq 0$ and $\psi = 1$. The derivation is very similar to the corresponding one in Appendix A1, only the remainder term is of different order of magnitude.

We recall from Appendix A1 that $\sqrt{1+x} = 1+\frac{1}{2}x+r_1$, which is convergent for $|\,x\,| \leq 1$ and where the remainder term $r_1 = -\frac{1}{8}(1+ax)^{-3/2}x^2$ where $a \in (0,1)$. When the expansion is applied to definition (6.18), case *iv)*, which applies for large T (cf. the discussion at the beginning of this appendix), and the previous results are accounted for, we get:

$$
\begin{aligned}
\hat{\psi}_{ad2} &= \hat{\psi}_{ml} + \frac{T-3}{2\,\hat{c}} - \sqrt{\frac{(T-3)^2}{4\,\hat{c}^2} + \hat{\psi}_{ml}^2 - \hat{l}} \\
&= \hat{\psi}_{ml} + \frac{T-3}{2\,\hat{c}} - \frac{T-3}{2\,\hat{c}}\sqrt{1 + \frac{4\,\hat{c}^2}{(T-3)^2}(\hat{\psi}_{ml}^2 - \hat{l})} \\
&= \hat{\psi}_{ml} + \frac{T-3}{2\,\hat{c}} - \frac{T-3}{2\,\hat{c}}\left[1 + \frac{2\,\hat{c}^2}{(T-3)^2}(\hat{\psi}_{ml}^2 - \hat{l}) + r_{1,T}\right], \quad r_{1,T} = O_p(T^{-3}) \\
&= \hat{\psi}_{ml} - \frac{\hat{c}\,(\hat{\psi}_{ml}^2 - \hat{l})}{T-3} + r_T, \qquad r_T = O_p(T^{-3}) \\
&\xrightarrow[p]{} \psi.
\end{aligned}
$$

The convergence is due to $\hat{c}\,(\hat{\psi}_{ml}^2 - \hat{l})/(T-3)^2$ being $O_p(T^{-2})$ and the consistency of $\hat{\psi}_{ml}$. We note that $4\,\hat{c}^2\,(\hat{\psi}_{ml}^2 - \hat{l})/(T-3)^2$ is $O_p(T^{-2})$. Consequently, the convergence condition of the Maclaurin expansion holds for large T.

Proof of the convergence $\mathbf{T^{3/2}(\hat{l}-1) \Rightarrow N(0, 48\sigma^2/\alpha^2)}$
We recall from definition (6.17) that

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

The numerator of this ratio equals

$$y_T(y_T - 2\,\bar{y}_{-1}) - y_0(y_0 - 2\,\bar{y}_{-1}) - T^{-1}(y_T - y_0)^2$$

(Appendix A1). It is easy to see that this can be simplified to

$$y_T(y_T - 2\,\bar{y}_{-1}) + O_p(T).$$

Substitution of $y_T = y_0 + \alpha T + \sum_{t=1}^{T}\epsilon_t$ and $\bar{y}_{-1} = \sum_{t=0}^{T-1} y_0 + \sum_{t=1}^{T}\alpha(t-1) + \sum_{t=1}^{T} S_{t-1}$, where $S_{t-1} = \sum_{i=1}^{t-1}\epsilon_i$ and $S_0 = 0$, yields

$$y_T(S_T - 2T^{-1}\sum_{t=1}^{T} S_{t-1}) + O_p(T).$$

We find that the numerator divided by $T^{3/2}$ is, and converges weakly to,

$$
\begin{aligned}
T^{-3/2}\left[\sum_{t=1}^{T}(y_t - \bar{y})^2 - \sum_{t=1}^{T}(y_{t-1} - \bar{y}_{-1})^2\right] &= T^{-1}y_T(T^{-1/2}S_T - 2T^{-3/2}\textstyle\sum_{t=1}^{T} S_{t-1}) \\
&\quad + O_p(T^{-1/2}) \\
&\Rightarrow \alpha N(0, \sigma^2/3).
\end{aligned}
$$

The convergence of $T^{-1}y_T$ to α has already been indicated. The convergence of the term in parenthesis to the Normal distribution is proved in Appendix A1 (y_t with $y_0 = 0$ there corresponds to S_t here).

The denominator of the above ratio form of $\hat{l} - 1$ has already been shown, after division by T^3, to tend in distribution to $\alpha^2/12$. We are hence ready to state that

$$T^{3/2}\left(\hat{l}-1\right) = \frac{T^{-3/2}\left[\sum_{t=1}^{T}(y_t-\bar{y})^2 - \sum_{t=1}^{T}(y_{t-1}-\bar{y}_{-1})^2\right]}{T^{-3}\left[\sum_{t=1}^{T}(y_{t-1}-\bar{y}_{-1})^2\right]}$$

$$\Rightarrow \frac{\alpha N(0,\sigma^2/3)}{\alpha^2/12}$$

or that

$$T^{3/2}(\hat{l}-1) \Rightarrow N(0, 48\sigma^2/\alpha^2).$$

Appendix A3
Proofs for the Bhargava AR(1) Model with Constant

We consider here:

- a proof revealing that $\hat{\gamma}_{0ml}$ is $O_p(1)$ and equals asymptotically y_1 under a unit root
- the expressions for the derivatives of $\hat{\sigma}^2_{\psi}$ which are a bit involved and are hence left to the appendix
- proofs for the asymptotic distributions or limits of the Wald statistic and the relevant derivatives, first under stationarity and next under a unit root, and
- the asymptotics for the adjusted Wald statistic.

The result on the $O_p(1)$ness of $\hat{\gamma}_{0ml}$ is needed in Section 6.6 when deriving the asymptotic distribution of $T(\hat{\psi}_{ml} - 1)$. The expressions for the derivatives of $\hat{\sigma}^2_{\psi}$ are presented simultaneously with the analysis of the Wald statistics.

Proving that $\hat{\gamma}_{0ml}$ is $O_p(1)$ and equals y_1 asymptotically under a unit root

An analytical formula for $\hat{\gamma}_{0ml}$ is not available as explained in Section 4.3. We can yet prove conveniently by means of the profile likelihood that $\hat{\gamma}_{0ml}$ is $O_p(1)$ and that it tends stochastically to y_1.

The residual variance, of which the profile log-likelihood for γ_0 is in essence composed of, equals:

$$\hat{\sigma}^2_{\gamma_0} = T^{-1}\left\{\sum_{t=1}^{T}(y_t-\gamma_0)^2 - \frac{\left[\sum_{t=2}^{T}(y_t-\gamma_0)(y_{t-1}-\gamma_0)\right]^2}{\sum_{t=2}^{T}(y_{t-1}-\gamma_0)^2}\right\},$$

cf. formulae (4.16) and (4.17) (with a different notation) in Section 4.3. As explained there, an analytical solution for the minimum of this equation is not available. Some algebra, substitution of the relation $y_t = y_{t-1}+\epsilon_t$, and well-known orders of magnitude of sums involving I(1) variables imply that

$$
\begin{aligned}
\hat{\sigma}^2_{\gamma_0} &= T^{-1}\left\{ (y_1-\gamma_0)^2 - \right. \\
&\quad \left. \frac{\sum_{t=2}^T (y_t-\gamma_0)^2 \sum_{t=2}^T (y_{t-1}-\gamma_0)^2 - \left[\sum_{t=2}^T (y_t-\gamma_0)(y_{t-1}-\gamma_0)\right]^2}{\sum_{t=2}^T (y_{t-1}-\gamma_0)^2} \right\} \\
&= T^{-1}\left\{ (y_1-\gamma_0)^2 + \sum_{t=2}^T \epsilon_t^2 - \right. \\
&\quad \left. \frac{(\sum_{t=2}^T y_{t-1}\epsilon_t)^2 - 2\gamma_0 \sum_{t=2}^T y_{t-1}\epsilon_t \sum_{t=2}^T \epsilon_t + \gamma_0^2 \left(\sum_{t=2}^T \epsilon_t\right)^2}{\sum_{t=2}^T y_{t-1}^2 - 2\gamma_0 \sum_{t=2}^T y_{t-1} + (T-1)\gamma_0^2} \right\} \\
&= T^{-1}\left\{ (y_1-\gamma_0)^2 + \sum_{t=2}^T \epsilon_t^2 - \right. \\
&\quad \left. \frac{[O_p(T)]^2 - 2\gamma_0 O_p(T) O_p(T^{1/2}) + \gamma_0^2 [O_p(T^{1/2})]^2}{O_p(T^2) - 2\gamma_0 O_p(T^{3/2}) + \gamma_0^2 O(T)} \right\} \\
&\propto (y_1-\gamma_0)^2 - \frac{O_p(T^2) - 2\gamma_0 O_p(T^{3/2}) + \gamma_0^2 O_p(T)}{O_p(T^2) - 2\gamma_0 O_p(T^{3/2}) + \gamma_0^2 O(T)}
\end{aligned}
$$

Assume for the moment that the MLE for γ_0 or the solution to the minimisation problem

$$
\inf_{\gamma_0} \left[(y_1-\gamma_0)^2 - \frac{O_p(T^2) - 2\gamma_0 O_p(T^{3/2}) + \gamma_0^2 O_p(T)}{O_p(T^2) - 2\gamma_0 O_p(T^{3/2}) + \gamma_0^2 O(T)} \right]
$$

diverges. Inspection reveals that the ratio above would continue to be bounded but the first term would explode to infinity. However, the solution cannot be unbounded because the function within the square brackets can take a finite value, too: If γ_0 is replaced by the potential solution y_1 — a constant from the point of view of the minimisation problem — then the first term collapses to zero, the second term or the ratio persists to be bounded, and the whole expression takes a finite value, too. Because the solution is even asymptotically a finite value we can express the minimisation problem as

$$
\inf_{\gamma_0} \left[(y_1-\gamma_0)^2 - \gamma_0 O_p(T^{-1/2}) \right].
$$

The second term is asymptotically negligible (the minimising value of γ_0 being bounded) whereas the first term does not diminish with T. The minimum is thus determined by the first term. The minimum or the MLE locates at y_1 at which value of γ_0 the first term reduces to zero. Of course, from the point of view of estimation the asymptotic MLE or y_1 is $O_p(1)$ instead of a constant as it was treated above.

Construction and asymptotics of the ordinary Wald statistic

The (square root of the) Wald statistic was defined to be

$$(\hat{\psi}_{ml} - \psi)\sqrt{-\frac{\partial^2 l_p(\psi)}{\partial^2 \psi}\bigg|_{\psi=\hat{\psi}_{ml}}}$$

in Section 6.6. We did not present the components of which $\partial^2 l_p(\psi)/\partial^2\psi$ is composed of. They are (cf. equation (6.52)):

$$\frac{\partial^2 l_p(\psi)}{\partial^2 \psi} = -\frac{T}{2}\left[\frac{\partial^2 \hat{\sigma}^2_\psi / \partial^2\psi}{\hat{\sigma}^2_\psi} - \frac{\left(\partial \hat{\sigma}^2_\psi / \partial\psi\right)^2}{\hat{\sigma}^4_\psi}\right]$$

where

$$\frac{\partial \hat{\sigma}^2_\psi}{\partial \psi} = 2T^{-1}\sum_{t=2}^{T}\left[y_t - \hat{\gamma}_{0\psi}(1-\psi) - \psi y_{t-1}\right]\left(\hat{\gamma}_{0\psi} - y_{t-1}\right), \tag{A3.1}$$

$$\begin{aligned}\frac{\partial^2 \hat{\sigma}^2_\psi}{\partial^2 \psi} &= 2T^{-1}\Bigg\{\sum_{t=2}^{T} y_{t-1}^2 + \\ &\quad \left[-2\sum_{t=2}^{T} y_{t-1} + (T-1)\hat{\gamma}_{0\psi} - (T-1)(1-\psi)\frac{\partial \hat{\gamma}_{0\psi}}{\partial\psi}\right]\hat{\gamma}_{0\psi} + \\ &\quad \left[\sum_{t=2}^{T}(y_t - \psi y_{t-1}) + (1-\psi)\sum_{t=2}^{T} y_{t-1} - (T-1)(1-\psi)\hat{\gamma}_{0\psi}\right]\frac{\partial \hat{\gamma}_{0\psi}}{\partial\psi}\Bigg\} \\ &= 2T^{-1}\Bigg\{\sum_{t=2}^{T}\left(y_{t-1} - \hat{\gamma}_{0\psi}\right)^2 - \\ &\quad \frac{\left\{\sum_{t=2}^{T}\left[y_t - \hat{\gamma}_{0\psi}(1-\psi) - \psi y_{t-1}\right] + (1-\psi)\sum_{t=2}^{T}\left(y_{t-1} - \hat{\gamma}_{0\psi}\right)\right\}^2}{1 + (T-1)(1-\psi)^2}\Bigg\},\end{aligned}$$

and

$$\frac{\partial \hat{\gamma}_{0\psi}}{\partial\psi} = -\frac{\sum_{t=2}^{T}\left[y_t - \hat{\gamma}_{0\psi}(1-\psi) - \psi y_{t-1}\right] + (1-\psi)\sum_{t=2}^{T}\left(y_{t-1} - \hat{\gamma}_{0\psi}\right)}{1 + (T-1)(1-\psi)^2}.$$

The asymptotic distribution of the statistic depends, of course, whether the process is stationary or not. In the stationary circumstance, the appropriate standardisation is

$$T^{1/2}(\hat{\psi}_{ml} - \psi)\sqrt{-T^{-1}\frac{\partial^2 l_p(\psi)}{\partial^2 \psi}\bigg|_{\psi=\hat{\psi}_{ml}}}$$

because the MLE is $O_p(T^{-1/2})$. We shall lay out the essentials of the derivation for the stationary case, too, because the model is not as familiar as the other contemplated models.

By definition, $\partial \hat{\sigma}^2_\psi / \partial\psi$, evaluated at the MLE, equals zero. The asymptotic limit of the second derivative, evaluated at the MLE, is:

$$\left(\frac{\partial^2 \hat{\sigma}^2_\psi}{\partial^2\psi}\right)\Bigg|_{\psi=\hat{\psi}_{ml}} = 2T^{-1}\left\{\sum_{t=2}^{T}\left(y_{t-1}-\hat{\gamma}_{0ml}\right)^2 - \frac{\left[\sum_{t=2}^{T}\hat{\epsilon}_t+(1-\hat{\psi}_{ml})\sum_{t=2}^{T}\left(y_{t-1}-\hat{\gamma}_{0ml}\right)\right]^2}{1+(T-1)(1-\hat{\psi}_{ml})^2}\right\} \tag{A3.2}$$

$$= 2T^{-1}\sum_{t=2}^{T}\left(y_{t-1}-\hat{\gamma}_{0ml}\right)^2 - \frac{[O_p(1)+O_p(1)\cdot O_p(1)]^2}{1+O(T)O_p(1)}$$

$$= 2T^{-1}\sum_{t=2}^{T}\left(y_{t-1}-\hat{\gamma}_{0ml}\right)^2 - O_p(T^{-1})$$

$$\xrightarrow[p]{} \frac{2\sigma^2}{1-\psi^2}$$

where $\hat{\epsilon}_t = y_t - \hat{\gamma}_{0ml}\,(1-\hat{\psi}_{ml}) - \hat{\psi}_{ml}\,y_{t-1}$, $t = 2,\ldots,T$.[1] The stochastic convergence presupposes that $T^{-1}\sum_{t=2}^{T}\left(y_{t-1}-\hat{\gamma}_{0ml}\right)^2$ converges to the familiar AR(1) population (asymptotic) counterpart as usual.

We can now see that

$$\sqrt{W} = T^{1/2}(\hat{\psi}_{ml}-\psi)\sqrt{-T^{-1}\frac{\partial^2 l_p(\psi)}{\partial^2\psi}\Bigg|_{\psi=\hat{\psi}_{ml}}}$$

$$= T^{1/2}(\hat{\psi}_{ml}-\psi)\sqrt{T^{-1}\frac{T}{2}\left[\frac{\partial^2\hat{\sigma}^2_\psi/\partial^2\psi}{\hat{\sigma}^2_\psi}-\frac{\left(\partial\hat{\sigma}^2_\psi/\partial\psi\right)^2}{\hat{\sigma}^4_\psi}\right]\Bigg|_{\psi=\hat{\psi}_{ml}}}$$

$$\Rightarrow N(0,1-\psi^2)\sqrt{\frac{1}{2}\left[\frac{2\sigma^2/(1-\psi^2)}{\sigma^2}-\frac{0^2}{\sigma^4}\right]} \quad \text{or to}$$

$$N(0,1)$$

so $\sqrt{W}$ follows the Standard Normal when T is large.

When a unit root exists, the appropriate standardisation is

$$T(\hat{\psi}_{ml}-1)\sqrt{-T^{-2}\frac{\partial^2 l_p(\psi)}{\partial^2\psi}\Bigg|_{\psi=\hat{\psi}_{ml}}}.$$

The derivatives of $\hat{\sigma}^2_\psi$ (of which $\partial^2 l_p(\psi)/\partial^2$ is composed of) become multiplied by powers of T this time. The asymptotic distribution of the standardised second derivative, evaluated at the MLE, is non-degenerate:

---

[1] It may be helpful to note that $\sum_{t=2}^{T}\hat{\epsilon}_t$ does not necessarily equal zero but $\hat{\epsilon}_{1\psi}+(1-\psi)\sum_{t=2}^{T}\hat{\epsilon}_{t\psi} \equiv y_1-\hat{\gamma}_{0\psi}+(1-\psi)\sum_{t=2}^{T}\left[y_t-\hat{\gamma}_{0\psi}\,(1-\psi)-\psi y_{t-1}\right]$ does. $\sum_{t=1}^{T}\hat{\epsilon}_{t\psi}$ is zero in general only when evaluated at $\psi = 0$.

$$T^{-1}\left(\frac{\partial^2 \hat{\sigma}^2_{\psi}}{\partial^2 \psi}\right)_{\psi=\hat{\psi}_{ml}}$$

$$= 2T^{-2}\ \Big\{\ \sum_{t=2}^{T} y_{t-1}^2 - 2\,\hat{\gamma}_{0ml}\sum_{t=2}^{T} y_{t-1} + (T-1)\,\hat{\gamma}^2_{0ml} -$$

$$\left[1+(T-1)(1-\hat{\psi}_{ml})^2\right]^{-1}\cdot$$

$$\left\{\left(\sum_{t=2}^{T}\hat{\epsilon}_t\right)^2 + 2(1-\hat{\psi}_{ml})\sum_{t=2}^{T}\hat{\epsilon}_t\sum_{t=2}^{T}\left(y_{t-1}-\hat{\gamma}_{0ml}\right) +\right.$$

$$\left.(1-\hat{\psi}_{ml})^2\left[\sum_{t=2}^{T}\left(y_{t-1}-\hat{\gamma}_{0ml}\right)\right]^2\right\}\ \Big\}$$

$$= 2T^{-2}\left[\sum_{t=2}^{T} y_{t-1}^2 - O_p(T^{3/2}) + O_p(T) -\right.$$

$$\left.\frac{O_p(T)+O_p(T^{-1})O_p(T^{1/2})O_p(T^{3/2})+O_p(T^{-2})O_p(T^3)}{1+O(T)O_p(T^{-2})}\right]$$

$$\Rightarrow 2\sigma^2\int_0^1 [W(r)]^2\,dr$$

where the first equality is based on the uppermost line of expression (A3.2). The asymptotic distribution of $\sqrt{W}$ under a unit root can be now found to be:

$$\sqrt{W} = T(\hat{\psi}_{ml}-1)\sqrt{-T^{-2}\frac{\partial^2 l_p(\psi)}{\partial^2\psi}\Big|_{\psi=\hat{\psi}_{ml}}}$$

$$= T(\hat{\psi}_{ml}-1)\sqrt{\frac{T}{2T}\left[\frac{T^{-1}\partial^2\,\hat{\sigma}^2_{\psi}\,/\partial^2\psi}{\hat{\sigma}^2_{\psi}} - \frac{\left(T^{-1/2}\partial\,\hat{\sigma}^2_{\psi}\,/\partial\psi\right)^2}{\hat{\sigma}^4_{\psi}}\right]\Bigg|_{\psi=\hat{\psi}_{ml}}}$$

$$\Rightarrow \frac{\int_0^1 W(r)dW(r)}{\int_0^1 [W(r)]^2\,dr}\sqrt{\frac{1}{2}\left\{\frac{2\sigma^2\int_0^1[W(r)]^2\,dr}{\sigma^2} - \frac{0}{\sigma^4}\right\}} \quad \text{or}$$

$$\frac{\int_0^1 W(r)dW(r)}{\sqrt{\int_0^1 [W(r)]^2\,dr}}.$$

Construction and asymptotics of the adjusted Wald statistic

The adjusted Wald statistic is:

$$\sqrt{W_{ad}}$$
$$= (\hat{\psi}_{ad} - 1)\sqrt{-\frac{\partial^2 l_{ad}(\psi)}{\partial^2 \psi}\bigg|_{\psi=\hat{\psi}_{ad}}}$$
$$= (\hat{\psi}_{ad} - \psi)\cdot$$
$$\sqrt{\left\{\frac{(T-3)}{2}\left[\frac{\partial^2 \hat{\sigma}^2_{\psi}/\partial^2\psi}{\hat{\sigma}^2_{\psi}} - \frac{\left(\partial \hat{\sigma}^2_{\psi}/\partial\psi\right)^2}{\hat{\sigma}^4_{\psi}}\right] + \frac{(T-1)\left[1-(T-1)(1-\psi)^2\right]}{\left[1+(T-1)(1-\psi)^2\right]^2}\right\}\bigg|_{\psi=\hat{\psi}_{ad}}}.$$

$(\partial^2 l_{ad}(\psi)/\partial^2\psi$ is defined in equation (6.58).) An additional complexity compared to the previous section is that $\partial\ \hat{\sigma}^2_{\psi}\ /\partial\psi$ evaluated at the AE is not zero. Asymptotically it is still zero:

$$\begin{aligned}
\left(\frac{\partial\ \hat{\sigma}^2_{\psi}}{\partial\psi}\right)\bigg|_{\psi=\hat{\psi}_{ad}} &= 2T^{-1}\sum_{t=2}^{T}\left[y_t - \hat{\gamma}_{0ad}\,(1-\hat{\psi}_{ad}) - \hat{\psi}_{ad}\,y_{t-1}\right]\left(\hat{\gamma}_{0ad} - y_{t-1}\right) \\
&= 2T^{-1}\left[-\sum_{t=2}^{T}\left(y_t - \hat{\gamma}_{0ad}\right)\left(y_{t-1} - \hat{\gamma}_{0ad}\right) + \hat{\psi}_{ad}\sum_{t=2}^{T}\left(y_{t-1} - \hat{\gamma}_{0ad}\right)^2\right] \\
&\xrightarrow[p]{} 2\left(-\frac{\psi\sigma^2}{1-\psi^2} + \psi\frac{\sigma^2}{1-\psi^2}\right) \\
&= 0,
\end{aligned}$$

As before, the convergencies are based on the presupposed consistency of sample moments and the consistency of the AE.[2]

By parallel reasoning as above, the asymptotic limit of $\partial^2\ \hat{\sigma}^2_{\psi}\ /\partial^2\psi$ evaluated at the AE is the same as when it is evaluated at the MLE. The asymptotic distribution of $\sqrt{W_{ad}}$ under stationarity is then readily found:

---

[2] A computationally useful form for the derivative of $\hat{\sigma}^2_{\psi}$ with respect to ψ is $\partial\ \hat{\sigma}^2_{\psi}\ /\partial\psi = 2T^{-1}$.

$$\left\{ -\sum_{t=2}^{T} y_t y_{t-1} + \psi\sum_{t=2}^{T} y_{t-1}^2 + \frac{(1-\psi)\left[y_1 + (1-\psi)\left(\sum_{t=2}^{T} y_t - \psi\sum_{t=2}^{T} y_{t-1}\right)\right]\sum_{t=2}^{T} y_{t-1}}{1+(T-1)(1-\psi)^2} + \frac{\left[y_1 + (1-\psi)\left(\sum_{t=2}^{T} y_t - \psi\sum_{t=2}^{T} y_{t-1}\right)\right]\left[\left(\sum_{t=2}^{T} y_t - \psi\sum_{t=2}^{T} y_{t-1}\right) - (T-1)(1-\psi)y_1\right]}{\left[1+(T-1)(1-\psi)^2\right]^2} \right\}.$$

$$\sqrt{W_{ad}}$$

$$= T^{1/2}(\hat{\psi}_{ad} - \psi)\sqrt{-T^{-1}\frac{\partial^2 l_{ad}(\psi)}{\partial^2 \psi}\bigg|_{\psi=\hat{\psi}_{ad}}}$$

$$= T^{1/2}(\hat{\psi}_{ad} - \psi)\cdot$$

$$\sqrt{\left\{\frac{(T-3)}{2T}\left[\frac{\partial^2 \hat{\sigma}^2_{\psi}/\partial^2\psi}{\hat{\sigma}^2_{\psi}} - \frac{\left(\partial \hat{\sigma}^2_{\psi}/\partial\psi\right)^2}{\hat{\sigma}^4_{\psi}}\right] + \frac{(T-1)\left[1-(T-1)(1-\psi)^2\right]}{T\left[1+(T-1)(1-\psi)^2\right]^2}\right\}\bigg|_{\psi=\hat{\psi}_{ad}}}$$

$$\Rightarrow N(0, 1-\psi^2)\sqrt{\frac{1}{2}\left[\frac{2\sigma^2/(1-\psi^2)}{\sigma^2} - \frac{0^2}{\sigma^4}\right] + 0} \quad \text{or to}$$

$$N(0,1).$$

We have made use above also of the result explained in Section 6.6: the additive adjustment term $(T-1)\left[1-(T-1)(1-\psi)^2\right] / \left[1+(T-1)(1-\psi)^2\right]^2$, evaluated at the AE, becomes divided by T when constructing the statistic so the term yields only a $O_p(T^{-1})$ modification underneath the square-root sign.

In the nonstationary situation the adjustment term, evaluated at the AE, becomes multiplied by T^{-2} but still contributes an $O_p(T^{-1})$ component because the order of magnitude of the subtraction $1-\hat{\psi}_{ad}$ is then different. The asymptotic limit of the standardised first derivative, evaluated at the MLE, is again zero:

$$T^{-1/2}\left(\frac{\partial \hat{\sigma}^2_{\psi}}{\partial\psi}\right)\bigg|_{\psi=\hat{\psi}_{ad}}$$

$$= 2T^{-3/2}\left\{ -\sum_{t=2}^{T}(y_{t-1}+\epsilon_t)y_{t-1} + \hat{\psi}_{ad}\sum_{t=2}^{T} y_{t-1}^2 + \right.$$
$$\left[\sum_{t=2}^{T} y_t - \hat{\psi}_{ad}\sum_{t=2}^{T} y_{t-1} + \left(1-\hat{\psi}_{ad}\right)\sum_{t=2}^{T} y_{t-1} - \right.$$
$$\left.\left.(T-1)\left(1-\hat{\psi}_{ad}\right)\hat{\gamma}_{0ad}\right]\hat{\gamma}_{0ad}\right\}$$

$$= 2T^{-3/2}\left\{ \left(\hat{\psi}_{ad}-1\right)\sum_{t=2}^{T} y_{t-1}^2 - \sum_{t=2}^{T} y_{t-1}\epsilon_t + \right.$$
$$\left.\left[(y_T - y_1) - 2\left(\hat{\psi}_{ad}-1\right)\sum_{t=2}^{T} y_{t-1} - (T-1)\left(1-\hat{\psi}_{ad}\right)\hat{\gamma}_{0ad}\right]\hat{\gamma}_{0ad}\right\}$$

$$= 2T^{-3/2}\left\{ O_p(T^{-1})O_p(T^2) - O_p(T) + \right.$$
$$\left.\left[O_p(T^{1/2}) + O_p(T^{-1})\cdot O_p(T^{3/2}) - O(T)O_p(T^{-1})O_p(1)\right]O_p(1)\right\}$$

$$\underset{p}{\longrightarrow} 0.$$

The first equality follows from formula (A3.1) and the unit-root relation $y_t = y_{t-1} + \epsilon_t$.

By similar reasoning as above, $T^{-1}\partial^2 \hat{\sigma}^2_{\psi}/\partial^2\psi$ evaluated at the AE follows asymptotically the same distribution as when the quantity is evaluated at the MLE. The asymptotics are then straightforward:

$$\sqrt{W_{ad}}$$

$$= T(\hat{\psi}_{ad} - 1)\sqrt{-T^{-2}\frac{\partial^2 l_{ad}(\psi)}{\partial^2 \psi}\bigg|_{\psi=\hat{\psi}_{ad}}}$$

$$= T(\hat{\psi}_{ad} - \psi)\cdot$$

$$\sqrt{\left\{\frac{(T-3)}{2T}\left[\frac{T^{-1}\partial^2\,\hat{\sigma}^2_{\psi}\,/\partial^2\psi}{\hat{\sigma}^2_{\psi}} - \frac{\left(T^{-1/2}\partial\,\hat{\sigma}^2_{\psi}\,/\partial\psi\right)^2}{\hat{\sigma}^4_{\psi}}\right] + \frac{(T-1)\left[1-(T-1)(1-\psi)^2\right]}{T^2\left[1+(T-1)(1-\psi)^2\right]^2}\right\}\bigg|_{\psi=\hat{\psi}_{ad}}}$$

$$\Rightarrow \frac{\int_0^1 W(r)dW(r)}{\int_0^1 [W(r)]^2\,dr}\sqrt{\frac{1}{2}\left\{\frac{2\sigma^2\int_0^1 [W(r)]^2\,dr}{\sigma^2} - \frac{0}{\sigma^4}\right\} + 0} \quad \text{or to}$$

$$\frac{\int_0^1 W(r)dW(r)}{\sqrt{\int_0^1 [W(r)]^2\,dr}}.$$

Appendix A4
Proofs for the AR(2) Model

This appendix follows the structure of Appendix A1 to a great extent and makes use of the results presented there. A difference is that the characteristic $\hat{\psi}^2_{ml} - \hat{l} \leq 0$ is proven only for T large (the proof for the stationary case presupposes an appropriate convergence theorem). We start by establishing the small-sample properties that

- the local maxima (if they exist) of $l_{ad2}(\psi)$ arise at the points expressed in definition (6.69) of $\hat{\psi}_{ad2}$.

We proceed to derive the asymptotic results

- $\hat{c} = O_p(1)$ whether $\mid \psi \mid < 1$ or $\psi = 1$,
- $\hat{\psi}^2_{ml} - \hat{l} = (\hat{\psi}_{ml} - 1)(\hat{\psi}_{ml} + 1) + O_p(T^{-1}) = \begin{cases} O_p(1) & \text{if } \mid \psi \mid < 1 \\ O_p(T^{-1}) & \text{if } \psi = 1, \end{cases}$ and
- $\hat{\psi}^2_{ml} - \hat{l} \leq 0$ for T large.

These conclusions enable us to:

- derive the approximation $\hat{\psi}_{ad2} \approx \hat{\psi}_{ml} - \hat{c}\,(\hat{\psi}^2_{ml} - \hat{l})/(T-3)$, the order of accuracy of it, and confirm consistency, or that $\hat{\psi}_{ad2} \xrightarrow[p]{} \psi$
- prove that the root $\hat{\psi}_{ml} + \frac{T-3}{2\hat{c}} + \sqrt{\cdot}$ is $O_p(T)$ whether $\mid \psi \mid < 1$ or $\psi = 1$
- prove that $l_{ad2}(\psi)$ has asymptotically a (local) maximum at $\hat{\psi}_{ad2}$ or at the root $\hat{\psi}_{ml} + \frac{T-3}{2\hat{c}} - \sqrt{\cdot}$ by verifying that $\partial^2 l_{ad2}(\psi)/\partial^2 \psi \xrightarrow[p]{} -\infty$ at the root (here we rely on the result that $\hat{\psi}^2_{ml} - \hat{l} \leq 0$ for T large the proof for which assumes an appropriate convergence theorem)
- prove that $l_{ad2}(\psi)$ is asymptotically flat at the root $\hat{\psi}_{ml} + \frac{T-3}{2\hat{c}} + \sqrt{\cdot}$ by verifying that $\partial^2 l_{ad2}(\psi)/\partial^2 \psi \xrightarrow[p]{} 0$ at the root.

A subresult or an asymptotic expression for the numerator of $\hat{\psi}^2_{ml} - \hat{l}$ is employed also in Section 6.6 when the asymptotic distribution of $\sqrt{W_{ad2}}$ is derived. As in Appendix A1, the last result is supplementary: The finite sample argument for the local maximum of $l_{ad2}(\psi)$, the fact that $\hat{c}$ is asymptotically positive and the property that $\hat{\psi}^2_{ml} - \hat{l} \leq 0$ for large T restrict already the local maximum to occur at $\hat{\psi}_{ml} + \frac{T-3}{2\hat{c}} - \sqrt{\cdot}$ (as defined in case *iv*) below) for large T. The point of the second last passage is that a (local) maximum exists asymptotically and that it

is the root $\hat{\psi}_{ml} + \frac{T-3}{2\hat{c}} - \sqrt{.}$ which determines the asymptotic distribution of the estimate. The appendix concludes with a summary of the findings.

Local maximum of $l_{ad2}(\psi)$ in finite samples

For convenience, cases *i*) to *v*) as defined on p. 134 are repeated here:

i) $\hat{c} < 0$ and the roots are complex (with nonzero imaginary parts)

*ii**) $\hat{c} < 0$, $\hat{\psi}_{ml}^2 - \hat{l} > 0$ and the roots are real (and unequal)

ii) $\hat{c} < 0$, $\hat{\psi}_{ml}^2 - \hat{l} < 0$ and the roots are real (and unequal)

iii) $\hat{c} = 0$ (in which case the adjusted likelihood equation implies a polynomial of order one)

iv) $\hat{c} > 0$, $\hat{\psi}_{ml}^2 - \hat{l} < 0$ and the roots are real (and unequal)

*iv**) $\hat{c} > 0$, $\hat{\psi}_{ml}^2 - \hat{l} > 0$ and the roots are real (and unequal)

v) $\hat{c} > 0$ and the roots are complex (with nonzero imaginary parts).

where $\hat{\psi}_{ml}$ and $\hat{l}$ are defined by equations (6.60) and (6.68), respectively. Only incidents *ii**) and *iv**) are examined here as the argumentation for the other situations parallels with the reasoning promoted for the $AR_\mu(1)$ model (Appendix A1).

Just as with the $AR_\mu(1)$ model

$$\left[\frac{\partial l_{ad2}(\psi)}{\partial \psi}\right]_{\psi=\hat{\psi}_{ml}} = \hat{c}$$

which assumes that $\hat{l} - \hat{\psi}_{ml}^2 \neq 0$ (the condition applies with probability one for finite T). We see that $l_{ad2}(\psi)$ is downward ($\hat{c} < 0$) or upward ($\hat{c} > 0$) sloping at $\psi = \hat{\psi}_{ml}$ according to the sign of $\hat{c}$.

The Cauchy–Schwarz inequality does not apply as such to the expression $\hat{\psi}_{ml}^2 - \hat{l}$ when the model is AR(2) which is why an analysis of the cases with $\hat{\psi}_{ml}^2 - \hat{l} > 0$ is needed.

Let us assume now that $\hat{\psi}_{ml}^2 - \hat{l} > 0$. The root of the adjusted likelihood equation

$$\hat{\psi}_{ml} + \frac{T-3}{2\hat{c}} - \sqrt{\frac{(T-3)^2}{4\hat{c}^2} + \hat{\psi}_{ml}^2 - \hat{l}}$$

lies then to the left of $\hat{\psi}_{ml}$ regardless of whether $\hat{c} < 0$ or $\hat{c} > 0$. The other root

$$\hat{\psi}_{ml} + \frac{T-3}{2\hat{c}} + \sqrt{\frac{(T-3)^2}{4\hat{c}^2} + \hat{\psi}_{ml}^2 - \hat{l}}$$

places to the right of $\hat{\psi}_{ml}$ regardless of the sign of $\hat{c}$ as well. This differs from the case $\hat{\psi}_{ml}^2 - < 0$ which implies that both roots are together smaller or larger than $\hat{\psi}_{ml}$ (depending on the sign of $\hat{c}$).

It will now be assumed that the adjusted likelihood equation has two different real roots or that the adjusted log-likelihood $l_{ad2}(\psi)$ is (equivalent to) a cubic equation which traces a local maximum and a local minimum. Let us study first the situation *ii**) when $\hat{c} < 0$ and $\hat{\psi}_{ml}^2 - \hat{l} > 0$. The local maximum must then be placed to the left of $\hat{\psi}_{ml}$ or relate to the root

$\hat{\psi}_{ml} + (T-3)/2\ \hat{c} - \sqrt{\cdot}$ as $l_{ad2}(\psi)$ is downward sloping at $\hat{\psi}_{ml}$ and the local extremes occur around $\hat{\psi}_{ml}$. Next, let *iv**) or the sample distinguish by $\hat{c} > 0$ and $\hat{\psi}_{ml}^2 - \hat{l} > 0$. Now the root $\hat{\psi}_{ml} + (T-3)/2\ \hat{c} + \sqrt{\cdot}$ is the one to the right of $\hat{\psi}_{ml}$ and must link to the local maximum: $l_{ad2}(\psi)$ is upward sloping at $\hat{\psi}_{ml}$ and the local extremes occur around $\hat{\psi}_{ml}$. Obviously, the local minimums connect to the other roots, respectively. It is shown below that only circumstance *iv*) is relevant for large T. The incident $\hat{\psi}_{ml}^2 - \hat{l} = 0$ is excluded from the preceding list as it can happen only with probability zero if the sample is finite.[1]

Basic asymptotic results

The coefficient c evaluated at $\hat{\phi}_{1_{ml}}$ is (ψ does not appear in the formula for c as it was assumed that $\psi = 1$ when it was derived):

$$\hat{c} = \frac{1}{(1-\hat{\phi}_{1_{ml}})^2} + h(\hat{\phi}_{1_{ml}}, T) = O_p(1)$$

(formula (6.66)). The coefficient is positive for large T as $\hat{\phi}_{1_{ml}} \xrightarrow[p]{} \phi_1$ and c is positive for the values of ϕ_1 in the parameter space of ϕ_1. The coefficient should be of the same order if $|\psi| < 1$ (cf. the argumentation on p. 133).

We proceed to the second task. Write

$$\hat{\psi}_{ml}^2 - \hat{l} = (\hat{\psi}_{ml} - 1)(\hat{\psi}_{ml} + 1) - (\hat{l} - 1).$$

The first term is obviously $O_p(1)$ or $O_p(T^{-1})$ if $|\psi| < 1$ or $\psi = 1$, respectively. Recall that

$$\hat{l} = \frac{\sum_{t=1}^T y_t^2 \sum_{t=1}^T (\Delta y_{t-1})^2 - (\sum_{t=1}^T y_t \Delta y_{t-1})^2}{\sum_{t=1}^T y_{t-1}^2 \sum_{t=1}^T (\Delta y_{t-1})^2 - (\sum_{t=1}^T y_{t-1} \Delta y_{t-1})^2}.$$

Hence

$$\begin{aligned} \hat{l} - 1 = & \left[\sum_{t=1}^T y_{t-1}^2 \sum_{t=1}^T (\Delta y_{t-1})^2 - (\sum_{t=1}^T y_{t-1} \Delta y_{t-1})^2\right]^{-1} \cdot \\ & \left\{\left[\sum_{t=1}^T y_t^2 \sum_{t=1}^T (\Delta y_{t-1})^2 - (\sum_{t=1}^T y_t \Delta y_{t-1})^2\right] - \right. \\ & \left. \left[\sum_{t=1}^T y_{t-1}^2 \sum_{t=1}^T (\Delta y_{t-1})^2 - (\sum_{t=1}^T y_{t-1} \Delta y_{t-1})^2\right]\right\}. \end{aligned}$$

It is not difficult to see that the denominator is $O_p(T^2)$ under $|\psi| < 1$ but is $O_p(T^3)$ if $\psi = 1$. Similarly, it can be shown that the numerator is $O_p(T^2)$ whether $|\psi| < 1$ or $\psi = 1$.

We shall depict the proof for the numerator in the unit-root case. After substituting $\sum_{t=1}^T y_t^2 = y_T^2 + \sum_{t=1}^T y_{t-1}^2$ ($y_0 = 0$) and $y_t = y_{t-1} + \phi_1 \Delta y_{t-1} + \epsilon_t$ into the expression for the numerator and carrying out some cancellations, we can express the numerator as

$$\left[y_T^2 - 2\phi_1 \sum_{t=1}^T y_{t-1} \Delta y_{t-1} - \phi_1^2 \sum_{t=1}^T (\Delta y_{t-1})^2\right] \sum_{t=1}^T (\Delta y_{t-1})^2 + O_p(T^{3/2}).$$

A further substitution of $u_{t-1} \equiv \Delta y_{t-1}$ ($u_t \equiv \phi_1 \Delta y_{t-1} + \epsilon_t$), $y_T^2 = 2\sum_{t=1}^T y_{t-1} u_t + \sum_{t=1}^T u_t^2$ (Banerjee *et al.* (1993, p. 90)), and $u_t = \phi_1 u_{t-1} + \epsilon_t$ ($u_0 = 0$) gives

---

[1]The local maximums under *ii**) or *iv**) could occur at quite extreme values of ψ. The asymptotic property $\hat{\psi}_{ml}^2 - \hat{l} \leq 0$ meaningfully excludes those cases.

$$\left[y_T^2 - 2\phi_1 \sum_{t=1}^T y_{t-1}u_{t-1} - \phi_1^2 \sum_{t=1}^T u_{t-1}^2\right] \sum_{t=1}^T u_{t-1}^2 + O_p(T^{3/2})$$
$$= \left[2\sum_{t=1}^T y_{t-1}u_t + \sum_{t=1}^T u_t^2 - 2\phi_1 \sum_{t=1}^T y_{t-1}u_{t-1} - \phi_1^2 \sum_{t=1}^T u_{t-1}^2\right] \sum_{t=1}^T u_{t-1}^2 + O_p(T^{3/2})$$
$$= \left[\sum_{t=1}^T (2y_{t-1} + \epsilon_t)\epsilon_t\right] \sum_{t=1}^T u_{t-1}^2 + O_p(T^{3/2}).$$

The expected value of $(2y_{t-1}+\epsilon_t)\epsilon_t$ is non-zero so the term written out is $O_p(T)O_p(T) = O_p(T^2)$, as was to be confirmed.

It follows from the preceding analysis that

$$\hat{\psi}_{ml}^2 - \hat{l} = \begin{cases} O_p(1) + O_p(T^2)/O_p(T^2) = O_p(1) & \text{if } |\psi| < 1 \\ O_p(T^{-1}) + O_p(T^2)/O_p(T^3) = O_p(T^{-1}) & \text{if } \psi = 1. \end{cases}$$

We shall next show that $\hat{\psi}_{ml}^2 - \hat{l} \leq 0$ for large T with the equality applying only asymptotically under $\psi = 1$ (the proof for the stationary case presumes an appropriate convergence theorem). We assume first that the process is stationary.

It is immediate that $\hat{\psi}_{ml}^2$ tends in probability to $(\rho_1 + \rho_2)^2$ from the fact that $\hat{\psi}_{ml}$ tends in probability to $(\rho_1 + \rho_2)$. The asymptotics of $\hat{l}$ are considered next. The terms in the expression of $\hat{l}$ — divided by T — tend in probability to the following limiting values:

$$\begin{aligned}
T^{-1}\textstyle\sum_{t=1}^T y_t^2 &\xrightarrow[p]{} \sigma_y^2, \\
T^{-1}\textstyle\sum_{t=1}^T y_{t-1}^2 &\xrightarrow[p]{} \sigma_y^2, \\
T^{-1}\textstyle\sum_{t=1}^T (\Delta y_{t-1})^2 &= T^{-1}\textstyle\sum_{t=1}^T y_{t-1}^2 - 2T^{-1}\textstyle\sum_{t=1}^T y_{t-1}y_{t-2} + T^{-1}\textstyle\sum_{t=1}^T y_{t-1}^2 \\
&\xrightarrow[p]{} 2\sigma_y^2(1 - r_1), \\
T^{-1}\textstyle\sum_{t=1}^T y_t \Delta y_{t-1} &= T^{-1}\textstyle\sum_{t=1}^T y_t y_{t-1} - T^{-1}\textstyle\sum_{t=1}^T y_t y_{t-2} \\
&\xrightarrow[p]{} \sigma_y^2(r_1 - r_2), \qquad \text{and} \\
T^{-1}\textstyle\sum_{t=1}^T y_{t-1} \Delta y_{t-1} &= T^{-1}\textstyle\sum_{t=1}^T y_{t-1}^2 - T^{-1}\textstyle\sum_{t=1}^T y_{t-1}y_{t-2} \\
&\xrightarrow[p]{} \sigma_y^2(1 - r_1).
\end{aligned}$$

Here σ_y^2 is the asymptotic variance of y_t (or the variance of y_t after the start-up effects have died out) and r_i stands for the i^{th} asymptotic autocorrelation of y_t (or the autocorrelation of y_t after the start-up effects have faded). The convergences assume that an appropriate convergence theorem is in operation. It follows that

$$\hat{l} \xrightarrow[p]{} \frac{2(1 - r_1) - (r_1 - r_2)^2}{1 - r_1^2}.$$

The autocorrelations can be expressed (after the start-up effects have disappeared) in terms of the parameters ρ_1 and ρ_2 (the parameterisation in levels) as follows:

$$\begin{aligned}
r_1 &= \frac{\rho_1}{1 - \rho_2} \qquad \text{and} \\
r_2 &= \rho_2 + \frac{\rho_1^2}{1 - \rho_2}
\end{aligned}$$

(Box and Jenkins (1976, p. 60)). Substituting these formulae into the asymptotic expression for $\hat{l}$ gives

$$\hat{l} \xrightarrow[p]{} \frac{2(1-\rho_2)-(\rho_2-\rho_1)^2(1-\rho_1-\rho_2)}{1-(\rho_2-\rho_1)}$$

after some algebra.

We are now ready to derive the asymptotic limit in probability of $\hat{\psi}^2_{ml} - \hat{l}$ under stationarity:

$$\begin{aligned}\hat{\psi}^2_{ml} - \hat{l} &\xrightarrow[p]{} (\rho_1+\rho_2)^2 + \frac{2(1-\rho_2)-(\rho_2-\rho_1)^2(1-\rho_1-\rho_2)}{1-(\rho_2-\rho_1)} \\ &= -2(1+\rho_2)(1-\rho_1-\rho_2) \\ &< 0.\end{aligned}$$

The inequality derives from the stationarity conditions for the process (formulae (5.8)).

The proof for the unit-root circumstance $\psi = 1$ closes the section. In this case the appropriately standardised sums of cross products of y_ts and the lags of it do not converge to the autocorrelations (Bierens (1993), Hasza (1980)) and a different proof from the above is needed. We start from the exact expression for $\hat{\psi}^2_{ml} - \hat{l}$:

$$\begin{aligned}&\hat{\psi}^2_{ml} - \hat{l} = \\ &\quad \left[\textstyle\sum_{t=1}^T y_{t-1}^2 \sum_{t=1}^T (\Delta y_{t-1})^2 - (\sum_{t=1}^T y_{t-1}\Delta y_{t-1})^2\right]^{-2} \cdot \\ &\quad \Big\{\left[\textstyle\sum_{t=1}^T y_t y_{t-1} \sum_{t=1}^T (\Delta y_{t-1})^2 - \sum_{t=1}^T y_{t-1}\Delta y_{t-1} \sum_{t=1}^T y_t \Delta y_{t-1}\right]^2 \\ &\qquad - \left[\textstyle\sum_{t=1}^T y_t^2 \sum_{t=1}^T (\Delta y_{t-1})^2 - (\sum_{t=1}^T y_t \Delta y_{t-1})^2\right] \cdot \\ &\qquad \left[\textstyle\sum_{t=1}^T y_{t-1}^2 \sum_{t=1}^T (\Delta y_{t-1})^2 - (\sum_{t=1}^T y_{t-1}\Delta y_{t-1})^2\right]\Big\} \cdot\end{aligned} \tag{A4.1}$$

The numerator, which determines the sign of the expression, can be written as:

$$\begin{aligned}&\left[\left(\textstyle\sum_{t=1}^T y_t y_{t-1}\right)^2 - \sum_{t=1}^T y_t^2 \sum_{t=1}^T y_{t-1}^2\right]\left[\textstyle\sum_{t=1}^T (\Delta y_{t-1})^2\right]^2 + \\ &\Big[-2\textstyle\sum_{t=1}^T y_t y_{t-1} \sum_{t=1}^T y_{t-1}\Delta y_{t-1} \sum_{t=1}^T y_t \Delta y_{t-1} + \sum_{t=1}^T y_t^2 \left(\sum_{t=1}^T y_{t-1}\Delta y_{t-1}\right)^2 + \\ &\textstyle\sum_{t=1}^T y_{t-1}^2 \left(\sum_{t=1}^T y_t \Delta y_{t-1}\right)^2\Big] \sum_{t=1}^T (\Delta y_{t-1})^2.\end{aligned}$$

The terms in the first square brackets are $O_p(T^4)$ and of larger magnitude than the rest of the terms.[2] One might hence be tempted to jump to the conclusion that $\hat{\psi}^2_{ml} - \hat{l} \leq 0$ for large T as the Cauchy–Schwarz inequality applies to the difference of the terms of largest magnitude. This would not be legitimate as $\left[\left(\sum_{t=1}^T y_t y_{t-1}\right)^2 - \sum_{t=1}^T y_t^2 \sum_{t=1}^T y_{t-1}^2\right] \cdot \left[\sum_{t=1}^T (\Delta y_{t-1})^2\right]^2$ is not $O_p(T^6)$ but $O_p(T^5)$ or of the same magnitude as the rest of the terms. A way to detect this is laid out below.

Substitute $\sum_{t=1}^T y_t^2 = y_T^2 + \sum_{t=1}^T y_{t-1}^2$ (employing the condition $y_0 = 0$), $y_t = y_{t-1} + u_t$ and $\Delta y_t = u_t$ where $u_t \equiv \phi_1 \Delta y_{t-1} + \epsilon_t$ into the above expression for the numerator. It then appears as

[2] Terms like $\sum_{t=2}^T y_t y_{t-1} \sum_{t=2}^T (\Delta y_{t-1})^2$ are $O_p(T^2)O_p(T) = O_p(T^3)$ as opposed to terms like $\sum_{t=2}^T y_{t-1}\Delta y_{t-1} \sum_{t=2}^T y_t \Delta y_{t-1}$ which are $O_p(T)O_p(T) = O_p(T^2)$.

$$\left[-\left(y_T^2 - 2\sum_{t=1}^T y_{t-1}u_t\right)\sum_{t=1}^T y_{t-1}^2 + \left(\sum_{t=1}^T y_{t-1}u_t\right)^2\right]\left(\sum_{t=1}^T u_{t-1}^2\right)^2 +$$
$$\left\{-2\left[\sum_{t=1}^T y_{t-1}u_t\left(\sum_{t=1}^T y_{t-1}u_{t-1}\right)^2 + \sum_{t=1}^T y_{t-1}u_t\sum_{t=1}^T y_{t-1}u_{t-1}\sum_{t=1}^T u_t u_{t-1}\right] +\right.$$
$$\left. y_T^2\left(\sum_{t=1}^T y_{t-1}u_{t-1}\right)^2 + \sum_{t=1}^T y_{t-1}^2\left(\sum_{t=1}^T u_t u_{t-1}\right)^2\right\}\sum_{t=1}^T u_{t-1}^2$$

after carrying out some cancellations. We can see now that the product on the first line cannot be of larger order than $O_p(T^5)$. Replace next y_T^2 by $2\sum_{t=1}^T y_{t-1}u_t + \sum_{t=1}^T u_t^2$ (see Banerjee *et al.* (1993, p. 90) or equation (A2.3) with $\alpha = y_0 = 0$) to find

$$-\left[\sum_{t=1}^T u_t^2\sum_{t=1}^T u_{t-1}^2 - \left(\sum_{t=1}^T u_t u_{t-1}\right)^2\right]\sum_{t=1}^T y_{t-1}^2\sum_{t=1}^T u_{t-1}^2 + O_p(T^4)$$

after taking account of some cancellations and suppressing a part of the expression to the remainder term. The term written out is nonpositive by the Cauchy–Schwarz inequality. The final substitution of $\sum_{t=1}^T u_t^2 = u_T^2 + \sum_{t=1}^T u_{t-1}^2$ and $u_t = \phi_1 u_{t-1} + \epsilon_t$ ($u_0 = 0$) confirms that the term dominates the numerator which becomes

$$-(1-\phi_1^2)\sum_{t=1}^T y_{t-1}^2\left(\sum_{t=1}^T u_{t-1}^2\right)^3 + O_p(T^4)$$

after some algebra. The first term is negative as $|\ \phi_1\ | < 1$ by assumption and rules the numerator for large T as it is $O_p(T^5)$. It follows that the numerator is negative (with probability one) for large T.

The denominator in equation (A4.1) is $O_p(T^6)$ so $\hat{\psi}_{ml}^2 - \hat{l} \xrightarrow[p]{} 0$ and takes negative values for large enough T before convergence. Hence $\hat{\psi}_{ml}^2 - \hat{l} \leq 0$ for large T also under a unit root.

Derivation of the approximation $\hat{\psi}_{ad2} \approx \hat{\psi}_{ml} - \hat{c}\,(\hat{\psi}_{ml}^2 - \hat{l})/(T-3)$

Following the reasoning in Appendix A1 we find the formulae

$$\begin{aligned}
\hat{\psi}_{ad2} &= \hat{\psi}_{ml} + \frac{T-3}{2\hat{c}} - \frac{T-3}{2\hat{c}}\left[1 + \frac{2\hat{c}^2}{(T-3)^2}(\hat{\psi}_{ml}^2 - \hat{l}) + r_{1,T}\right], \\
& \qquad r_{1,T} = \begin{cases} O_p(T^{-4}) & \text{if } |\ \psi\ | < 1 \\ O_p(T^{-6}) & \text{if } \psi = 1 \end{cases} \\
&= \hat{\psi}_{ml} - \frac{\hat{c}\,(\hat{\psi}_{ml}^2 - \hat{l})}{T-3} + r_T, \qquad r_T = \begin{cases} O_p(T^{-3}) & \text{if } |\ \psi\ | < 1 \\ O_p(T^{-5}) & \text{if } \psi = 1 \end{cases} \\
&\xrightarrow[p]{} \psi.
\end{aligned}$$

The convergence is due to $\hat{c}\,(\hat{\psi}_{ml}^2 -)/(T-3)^2$ being $O_p(T^{-2})$ if $|\ \psi\ | < 1$ or $O_p(T^{-3})$ if $\psi = 1$ and the consistency of $\hat{\psi}_{ml}$. We note that the convergence condition of the Maclaurin expansion holds for large T ($4\hat{c}^2\,(\hat{\psi}_{ml}^2 - \hat{l})/(T-3)^2$ is likewise $O_p(T^{-2})$ or $O_p(T^{-3})$ if $|\ \psi\ | < 1$ or $\psi = 1$, respectively).

Asymptotic behaviour of the root $\hat{\psi}_{\mathbf{ml}}+\frac{\mathbf{T-3}}{\mathbf{2\hat{c}}}+\sqrt{\cdot}$

It is easy to see, by the foregoing Maclaurin expansion, say, that $\sqrt{\frac{(T-3)^2}{4\hat{c}^2}+\hat{\psi}_{ml}^2-\hat{l}}$ is $O_p(T)$ whether $\mid\psi\mid<1$ or $\psi=1$. The preceding results imply that $\hat{\psi}_{ml}+\frac{T-3}{2\hat{c}}+\sqrt{\frac{(T-3)^2}{4\hat{c}^2}+\hat{\psi}_{ml}^2-\hat{l}}$ is $O_p(T)$, too, and that it diverges stochastically to ∞ in both cases $\mid\psi\mid<1$ and $\psi=1$.

Asymptotic behaviour of $\partial^2\mathbf{l_{ad2}}(\psi)/\partial^2\psi$ at the root $\hat{\psi}_{\mathbf{ml}}+\frac{\mathbf{T-3}}{\mathbf{2\hat{c}}}-\sqrt{\cdot}$

The adjusted likelihood equation and the second derivative have the same appearance as those with the $\mathrm{AR}_\mu(1)$ model. The difference is that the symbol $\hat{l}$ stands here as defined by equation (6.68). The second derivative of $l_{ad2}(\psi)$ evaluated at the root

$$\hat{\psi}_{ml}+\frac{T-3}{2\hat{c}}-\sqrt{\frac{(T-3)^2}{4\hat{c}^2}+\hat{\psi}_{ml}^2-\hat{l}}$$

and algebra like that in Appendix A1 yields

$$\begin{aligned}
&\left[\frac{\partial^2 l_{ad2}(\psi)}{\partial^2\psi}\right]_{\psi=\hat{\psi}_{ml}+\frac{T-3}{2\hat{c}}-\sqrt{\cdot}}\\
&=\frac{(T-3)\left\{2(\hat{\psi}_{ml}^2-\hat{l})+\frac{T-3}{\hat{c}}\left[-\frac{\hat{c}(\hat{\psi}_{ml}^2-\hat{l})}{T-3}+r_T\right]\right\}}{\left\{\frac{T-3}{\hat{c}}\left[-\frac{\hat{c}(\hat{\psi}_{ml}^2-\hat{l})}{T-3}+r_T\right]\right\}^2},\\
&\qquad r_T=\begin{cases}O_p(T^{-3}) & \text{if } \mid\psi\mid<1\\ O_p(T^{-5}) & \text{if } \psi=1\end{cases}\\
&=\frac{(T-3)\left[2(\hat{\psi}_{ml}^2-\hat{l})-(\hat{\psi}_{ml}^2-\hat{l})+r_T^*\right]}{\left[-(\hat{\psi}_{ml}^2-\hat{l})+r_T^*\right]^2},\\
&\qquad r_T^*=\begin{cases}O_p(T^{-2}) & \text{if } \mid\psi\mid<1\\ O_p(T^{-4}) & \text{if } \psi=1\end{cases}\\
&=\frac{(T-3)\left(\hat{\psi}_{ml}^2-\hat{l}+r_T^*\right)}{\left[-(\hat{\psi}_{ml}^2-\hat{l})+r_T^*\right]^2}\\
&\xrightarrow[p]{}-\infty.
\end{aligned}$$

The divergence is to $-\infty$ because $\hat{\psi}_{ml}^2-\hat{l}\leq 0$ for large T (the equality sign can apply only if $\psi=1$ and T is infinite) and it is of larger order than r_T^*. The divergence to infinity is faster if $\psi=1$ in which case the denominator tends stochastically to zero. The outcome is that $[\partial^2 l_{ad2}(\psi)/\partial^2\psi]_{\psi=\hat{\psi}_{ml}+\frac{T-3}{2\hat{c}}-\sqrt{\cdot}}$ tends stochastically to minus infinity with T.

Accordingly $l_{ad2}(\psi)$ has asymptotically an infinitely sharp maximum at $\hat{\psi}_{ad2}=\hat{\psi}_{ml}+\frac{T-3}{2\hat{c}}-$

$$\sqrt{\frac{(T-3)^2}{4\hat{c}^2}+\hat{\psi}_{ml}^2-\hat{l}} \underset{p}{\longrightarrow} \psi.$$

Asymptotic behaviour of $\partial^2 l_{ad2}(\psi)/\partial^2\psi$ at the root $\hat{\psi}_{ml}+\frac{T-3}{2\hat{c}}+\sqrt{\cdot}$

Evaluating the second derivative of $l_{ad2}(\psi)$ at the root

$$\hat{\psi}_{ml}+\frac{T-3}{2\hat{c}}+\sqrt{\frac{(T-3)^2}{4\hat{c}^2}+\hat{\psi}_{ml}^2-\hat{l}}$$

gives

$$\begin{aligned}
&\left[\frac{\partial^2 l_{ad2}(\psi)}{\partial^2\psi}\right]_{\psi=\hat{\psi}_{ml}+\frac{T-3}{2\hat{c}}+\sqrt{\cdot}} \\
&= \frac{(T-3)\left[2(\hat{\psi}_{ml}^2-\hat{l})+\frac{(T-3)^2}{2\hat{c}^2}+\frac{T-3}{\hat{c}}\sqrt{\frac{(T-3)^2}{4\hat{c}^2}+\hat{\psi}_{ml}^2-\hat{l}}\right]}{\left[\frac{(T-3)^2}{2\hat{c}^2}+\frac{T-3}{\hat{c}}\sqrt{\frac{(T-3)^2}{4\hat{c}^2}+\hat{\psi}_{ml}^2-\hat{l}}\right]^2} \qquad \text{(A4.2)} \\
&= \frac{O_p(T^3)}{O_p(T^4)} = O_p(T^{-1}) \qquad \text{for } |\psi|<1 \text{ and } \psi=1.
\end{aligned}$$

The asymptotics follows from the previous results and especially from the fact that $\sqrt{\cdot}$ is $O_p(T)$ whether $|\psi|<1$ or $\psi=1$. We have found that $[\partial^2 l_{ad2}(\psi)/\partial^2\psi]_{\psi=\hat{\psi}_{ml}+\frac{T-3}{2\hat{c}}+\sqrt{\cdot}}$ converges in probability to zero under both cases $|\psi|<1$ and $\psi=1$. In other words, $l_{ad2}(\psi)$, evaluated at $\hat{\psi}_{ml}+\frac{T-3}{2\hat{c}}+\sqrt{\cdot}$, has asymptotically no curvature regardless of the stationarity status of the model.

Summary

The adjusted estimate converges in probability to the true ψ and $l_{ad2}(\psi)$ has asymptotically at that point an infinitely sharp (local) maximum. The other root diverges stochastically to infinity and $l_{ad2}(\psi)$ becomes infinitely flat at the root. This happens whether the model is stationary or not. It contrasts with the $AR_\mu(1)$ model in which case the other root was $O_p(1)$ and $l_{ad2}(\psi)$ developed a sharp minimum at the root if the model was nonstationary. It is easy to comprehend that $l_{ad2}(\psi)$ becomes flat for remote values of ψ as the discrepancy between $l_{ad2}(\psi)$ and $l_p(\psi)$ tends to become relatively smaller as T increases. The behaviour of $l_{ad2}(\psi)$ under the unit-root AR(2) model resembles the behaviour of $l_{ad2}(\psi)$ under the stationary $AR_\mu(1)$ model. The explanation lies in the fact that the adjustment becomes asymptotically negligible in relative terms in both cases (cf. Sections 6.3 and 6.6).

Appendix A5
Starting Value Asymptotics

We shall first

- prove that $\hat{\psi}_{ml}$ is consistent for $\psi \in (-1, 1]$ when y_0 or the starting value tends to infinity and
- derive the asymptotic distribution of $y_0(\hat{\psi}_{ml} - \psi)$ for $\psi \in (-1, 1]$

when the model is the simple AR(1) of Section 6.2.
Next we

- accomplish the same tasks when the model includes a constant with the modification that $\psi \in (-1, 1)$ is assumed

and prove that

- $\hat{\psi}_{ad2}$ and $\hat{\psi}_{ad2,ap}$ are consistent in the above sense for $\psi \in (-1, 1)$.

We go on to show that

- $\hat{\psi}_{ml}$ and $\hat{\psi}_{ad}$ converge to unity when $\psi > 0$ and the starting value tends to infinity

under the $\mathrm{AR}^B_\mu(1)$ model which erroneously assumes a zero starting value.
In contrast, we conclude by showing that

- $\hat{\psi}_{ml}$ and $\hat{\psi}_{ad}$ are consistent in the above sense for $\psi \in (-1, 1)$ except perhaps for $\psi = 0$

and that

- the MLE may not be uniquely defined under $\psi = 0$ and an infinitely large starting value because the profile likelihood possesses then two local maxima

when the model is $\mathrm{AR}^B_\mu(1)$ which allows for a non-zero starting value. The proofs for the last two results are conditional on the assumption that the MLE and the AE are bounded.

The estimates are not consistent as $|y_0|$ tends to infinity when the model is $\mathrm{AR}_\mu(1)$ and $\mathrm{AR}^B_\mu(1)$ with a unit root: It was proved on p. 103 that the estimates are invariant with respect to y_0 when a unit root exists. This is why the unit-root case is excluded from the analysis of the third, fourth and sixth bulleted assignments.

Interestingly, it can be shown that information on ψ as measured by the expected information measure in general tends to infinity under the models studied which supplies intuition for the consistency results. It is however worth remarking on the peculiar behaviour of the expected information measure under the unit-root $\mathrm{AR}_\mu(1)$ model (with zero constant). The measure is then

$$i_{\psi\psi}|_{\psi=1} = \frac{T(T-1)}{2} + \sigma^{-2} T y_0^2.$$

This tends to infinity with $|y_0|$ even though the MLE of ψ is invariant with respect to the starting value y_0. The extraordinary behaviour of expected information measures under a unit root has been noted already on p. 39.

Algebra for a derivation of the asymptotic distribution of the AE seems involved, and is not of vital importance from the point of view of this thesis, so we have not accomplished such a derivation.

We refer to the above kind of analysis as 'starting value asymptotics' to differentiate it from the usual 'T large asymptotics'. The circumstance of y_0 tending to infinity is considered below but similar results would hold for $|y_0|$ tending to infinity. In this appendix orders of magnitude terms like $O_p(1)$ express magnitude with respect to y_0 and consistency refers to consistency as the starting value tends to infinity.

Starting value asymptotics is connected to the *small-σ asymptotics* in Evans and Savin (1984), say, in the sense that σ relative to y_0 tends to zero in both (assuming a non-zero y_0). In the present context, we could as well standardise the estimates by y_0/σ and let σ tend to zero which procedure would give otherwise the same asymptotic distributions except that they would be free of σ. The distribution which would arise by letting σ tend to zero in the first place would be different.[1] Another related development is *continuous record asymptotics* of Phillips (1987) and Perron (1991) who let the sampling interval tend to zero. The last two authors consider only the simple AR(1) model.

There is also parallelism to a result of Lucas (1996, p. 102). He shows that $\hat{\psi}_{ols}$ for the simple AR(1) model converges to the true value if there is an outlier in the innovation process and the outlier tends (in absolute value) to infinity.

Starting value asymptotics for the simple AR(1) model (Section 6.2)

The analysis is very straightforward for this model. Substitution of $y_t = \psi y_{t-1} + \epsilon_t$ into formula (6.3) for the MLE yields the familiar result

$$\hat{\psi}_{ml} - \psi = \frac{\sum_{t=1}^{T} y_{t-1}\epsilon_t}{\sum_{t=1}^{T} y_{t-1}^2}.$$

Formula (6.11) holds for the present model, too, when the constant is set to zero. The formula implies that

$$y_t = \psi^t y_0 + O_p(1). \tag{A5.1}$$

Substituting this into the numerator above yields

$$\begin{aligned} \sum_{t=1}^{T} y_{t-1}\epsilon_t &= y_0 \sum_{t=0}^{T-1} \psi^t \epsilon_{t+1} + O_p(1) \\ &= O_p(y_0). \end{aligned}$$

Note that the last equality holds also for $\psi = 0$. The denominator of the MLE becomes

$$\begin{aligned} \sum_{t=1}^{T} y_{t-1}^2 &= y_0^2 \sum_{t=0}^{T-1} \psi^{2t} + O_p(y_0) \\ &= \begin{cases} y_0^2 \dfrac{1-\psi^{2T}}{1-\psi^2} + O_p(y_0) & \text{if } |\psi| < 1 \\ y_0^2 T & \text{if } \psi = 1 \end{cases} \\ &= O_p(y_0^2) \end{aligned}$$

after a similar substitution and a bit of algebra. It is found that

$$\hat{\psi}_{ml} - \psi = \frac{O_p(y_0)}{O_p(y_0^2)} = O_p(y_0^{-1}), \quad \psi \in (-1, 1]$$

so the MLE converges to the true value at a rate of $O_p(y_0)$.

---

[1] The theorem of Evans and Savin (1984) does not apply to the simple AR(1) model but applies to the $AR_\mu(1)$ model when the constant is different from zero. In this case, the theorem easily yields $\alpha[(1-\psi)\sigma]^{-1}[2(T-1)]^{1/2}(\hat{\psi}_{ml} - \psi) \sim N(0,1)$.

The asymptotic distribution of $y_0(\hat{\psi}_{ml} - \psi)$ is Normal:

$$\begin{aligned} y_0(\hat{\psi}_{ml} - \psi) &= \frac{y_0 \sum_{t=1}^{T} y_{t-1}\epsilon_t}{\sum_{t=1}^{T} y_{t-1}^2} \\ &= \frac{y_0 \left[y_0 \sum_{t=0}^{T-1} \psi^t \epsilon_{t+1} + O_p(1) \right]}{y_0^2 \frac{1-\psi^{2T}}{1-\psi^2} + O_p(y_0)} \\ &= \frac{\sum_{t=0}^{T-1} \psi^t \epsilon_{t+1} + O_p(y_0^{-1})}{\frac{1-\psi^{2T}}{1-\psi^2} + O_p(y_0^{-1})} \\ &\Rightarrow \frac{N\left(0, \sigma^2 \frac{1-\psi^{2T}}{1-\psi^2}\right)}{\frac{1-\psi^{2T}}{1-\psi^2}} \qquad \text{or to} \\ &\begin{cases} N\left(0, \sigma^2 \frac{1-\psi^2}{1-\psi^{2T}}\right) & \text{if } |\psi| < 1 \\ N\left(0, \sigma^2/T\right) & \text{if } \psi = 1 \end{cases} \end{aligned}$$

by the above results. For simplicity of presentation, we have not analysed the unit-root case separately. The specific result for it can be found by, say, employing l'Hospital's rule. It may be worth stressing that the MLE is consistent in the present context also under $\psi = 1$.

The variance takes an inverse U-shape as a function of ψ so the variance decreases as the nonstationary boundaries are approached. It is in accordance with the assertions at the end of Section 5.3 according to which the impact of y_0 on the distribution strengthens as the nonstationary boundaries are approached. As a straightforward calculation reveals, the asymptotic variance of the standardised MLE is exactly the inverse of the variance of the process.

If we standardised $(\hat{\psi}_{ml} - \psi)$ by y_0/σ (it is the standardised deviation which is in operation as noted in the introduction to this appendix) then the variance would approximately equal the conventional $(1 - \psi^2)$ for T large and ψ in the stationary region. Thus the present asymptotics can return a similar asymptotic distribution as the conventional asymptotics.

Starting value asymptotics for $\hat{\psi}_{\mathrm{ml}}$ under the AR(1) model with constant (Section 6.3)

It follows from the definitions for the MLE and the $\mathrm{AR}_\mu(1)$ process (formulae (6.6) and (6.4)), and the property $\sum_{t=1}^{T}(y_{t-1} - \bar{y}_{-1}) = 0$ that

$$\hat{\psi}_{ml} = \psi + \frac{\sum_{t=1}^{T}(y_{t-1} - \bar{y}_{-1})\epsilon_t}{\sum_{t=1}^{T} y_{t-1}^2 - T(\bar{y}_{-1})^2} = \psi + \frac{\sum_{t=1}^{T} y_{t-1}\epsilon_t - \bar{y}_{-1}\sum_{t=1}^{T}\epsilon_t}{\sum_{t=1}^{T} y_{t-1}^2 - T(\bar{y}_{-1})^2}.$$

As in the previous section,

$$y_t = \psi^t y_0 + O_p(1),$$

and

$$\sum_{t=1}^{T} y_{t-1}^2 = y_0^2 \frac{1-\psi^{2T}}{1-\psi^2} + O_p(y_0) = O_p(y_0^2)$$

by formula (6.11). It also implies that

$$\sum_{t=1}^{T} y_{t-1} = y_0 \frac{1-\psi^T}{1-\psi} + O_p(1) = O_p(y_0),$$

$$\begin{aligned}\sum_{t=1}^{T} y_{t-1}^2 - T(\bar{y}_{-1})^2 &= y_0^2 \frac{1-\psi^{2T}}{1-\psi^2} + O_p(y_0) - T\left\{T^{-1}\left[y_0 \frac{1-\psi^T}{1-\psi} + O_p(1)\right]\right\}^2 \\ &= y_0^2\left[\frac{1-\psi^{2T}}{1-\psi^2} - T^{-1}\left(\frac{1-\psi^T}{1-\psi}\right)^2\right] + O_p(y_0) \\ &= O_p(y_0^2),\end{aligned}$$

$$\sum_{t=1}^{T} y_{t-1}\epsilon_t = y_0 \sum_{t=0}^{T-1} \psi^t \epsilon_{t+1} + O_p(1),$$

$$\bar{y}_{-1} \sum_{t=1}^{T} \epsilon_t = y_0 T^{-1} \frac{1-\psi^T}{1-\psi} \sum_{t=1}^{T} \epsilon_t + O_p(1),$$

and

$$\begin{aligned}\sum_{t=1}^{T} y_{t-1}\epsilon_t - \bar{y}_{-1} \sum_{t=1}^{T} \epsilon_t &= y_0 \sum_{t=0}^{T-1} \psi^t \epsilon_{t+1} + O_p(1) - \left[y_0 T^{-1} \frac{1-\psi^T}{1-\psi} \sum_{t=1}^{T} \epsilon_t + O_p(1)\right] \\ &= y_0\left(\sum_{t=0}^{T-1} \psi^t \epsilon_{t+1} - T^{-1} \frac{1-\psi^T}{1-\psi} \sum_{t=1}^{T} \epsilon_t\right) + O_p(1) \\ &= O_p(y_0).\end{aligned}$$

Substituting these results into the above expression for the MLE delivers

$$\hat{\psi}_{ml} - \psi = \frac{O_p(y_0)}{O_p(y_0^2)} = O_p(y_0^{-1})$$

proving consistency. As noted in the beginning of this appendix, consistency does not take place under a unit root. Indeed, the above stated orders of magnitude would not hold if we calculated them under a unit root.

The previous results enable us to derive the asymptotic distribution, too:

$$
\begin{aligned}
y_0(\hat{\psi}_{ml} - \psi) &= \frac{y_0\left[y_0\left(\sum_{t=0}^{T-1}\psi^t\epsilon_{t+1} - T^{-1}\frac{1-\psi^T}{1-\psi}\sum_{t=1}^{T}\epsilon_t\right) + O_p(1)\right]}{y_0^2\left[\frac{1-\psi^{2T}}{1-\psi^2} - T^{-1}\left(\frac{1-\psi^T}{1-\psi}\right)^2\right] + O_p(y_0)} \\
&= \frac{\sum_{t=0}^{T-1}\psi^t\epsilon_{t+1} - T^{-1}\frac{1-\psi^T}{1-\psi}\sum_{t=1}^{T}\epsilon_t + O_p(y_0^{-1})}{\left[\frac{1-\psi^{2T}}{1-\psi^2} - T^{-1}\left(\frac{1-\psi^T}{1-\psi}\right)^2\right] + O_p(y_0^{-1})} \\
&\Rightarrow \frac{N\left(0, \sigma^2\left[\frac{1-\psi^{2T}}{1-\psi^2} - T^{-1}\left(\frac{1-\psi^T}{1-\psi}\right)^2\right]\right)}{\frac{1-\psi^{2T}}{1-\psi^2} - T^{-1}\left(\frac{1-\psi^T}{1-\psi}\right)^2} \quad \text{or to} \\
& N\left(0, \sigma^2\left[\frac{1-\psi^{2T}}{1-\psi^2} - T^{-1}\left(\frac{1-\psi^T}{1-\psi}\right)^2\right]^{-1}\right).
\end{aligned}
$$

The term $-T^{-1}[(1-\psi^T)/(1-\psi)]^2$ is extra compared to the formula for the variance under the simple AR(1) model. The variances take similar values for $\psi < -0.5$, say, but for larger values of ψ the present formula gives a larger value. The variance traces a rough inverse 'U' as a function of ψ still, though this pattern brakes down in the neighbourhood of unity. The variance explodes for such values which reflects the invariance of the MLE with respect to y_0 under a unit root. Figure A5.1 exemplifies the property.

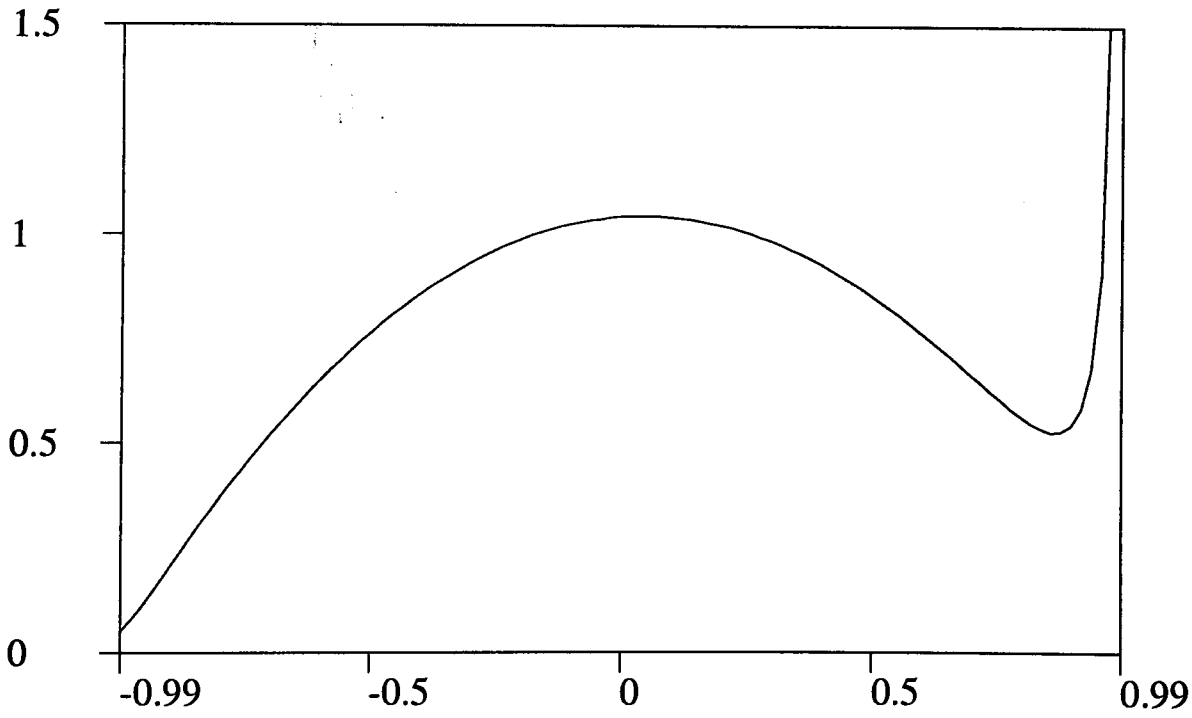


Figure A5.1 Variance of the standardised MLE as a function of ψ for $|y_0|$ large and $T = 25$ ($\mathrm{AR}_\mu(1)$).

The additional term — relative to the corresponding variance for the simple AR(1) — in the formula for the variance becomes negligible when T is large. Thus an asymptotic distribution similar to the conventional one with variance $1-\psi^2$ emerges again for $\psi \in (-1, 1)$ (cf. the previous section). The variance of the MLE is not strictly related to the variance of the process y_t — which remains at $(1-\psi^{2T})/(1-\psi^2)$ — contrary to the case with the simple AR(1) model.

Starting value asymptotics for $\hat{\psi}_{ad2}$ under the AR(1) model with constant (Section 6.3)

The first thing to note is that the coefficient c evaluated at the MLE converges to c (formula (6.14)):

$$\begin{aligned}\hat{c} &= \frac{1}{1-\hat{\psi}_{ml}} - \frac{1-\hat{\psi}_{ml}^{T}}{T(1-\hat{\psi}_{ml})^2} \\ &\xrightarrow[p]{} \frac{1}{1-\psi} - \frac{1-\psi^T}{T(1-\psi)^2} \quad > \quad 0 \quad \text{for } T \geq 2,\end{aligned}$$

where the convergence is due to the consistency of the MLE as y_0 tends to infinity. This implies — if we take for granted that a local maximum exists for y_0 large — that only case *iv)* of the definition of the AE (p. 81) is relevant when y_0 is large. We shall thus analyse the stochastic limit of the root

$$\hat{\psi}_{ml} + \frac{T-3}{2\hat{c}} - \sqrt{\frac{(T-3)^2}{4\hat{c}^2} + \hat{\psi}_{ml}^2 - \hat{l}}.$$

The sole quantity the asymptotic limit of which is unknown in this expression is $\hat{l}$ or

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

(definition (6.17)). The stochastic limit of the denominator has been derived above. The numerator is related, for large y_0, to the denominator in a simple way as we shall now show.

$$\begin{aligned}\sum_{t=1}^{T} y_t^2 &= \sum_{t=1}^{T} \left[\psi^t y_0 + O_p(1)\right]^2 \\ &= y_0^2 \sum_{t=1}^{T} \psi^{2t} + O_p(y_0) \\ &= y_0^2 \psi^2 \frac{1-\psi^{2T}}{1-\psi^2} + O_p(y_0).\end{aligned}$$

The first equality arises from formula (6.11). Other implications of it are that

$$\sum_{t=1}^{T} y_t = y_0 \psi \frac{1-\psi^T}{1-\psi} + O_p(1)$$

and that

$$\begin{aligned}\sum_{t=1}^{T} y_t^2 - T(\bar{y})^2 &= y_0^2 \psi^2 \frac{1-\psi^{2T}}{1-\psi^2} + O_p(y_0) - T\left\{T^{-1}\left[y_0 \psi \frac{1-\psi^T}{1-\psi} + O_p(1)\right]\right\}^2 \\ &= y_0^2 \psi^2 \left[\frac{1-\psi^{2T}}{1-\psi^2} - T^{-1}\left(\frac{1-\psi^T}{1-\psi}\right)^2\right] + O_p(y_0).\end{aligned}$$

The numerator is hence, for large y_0, ψ^2 times the denominator and the asymptotic limit of the ratio is thus ψ^2:

$$\begin{aligned}
\hat{l} &= \frac{\sum_{t=1}^{T} y_t^2 - T(\bar{y})^2}{\sum_{t=1}^{T} y_{t-1}^2 - T(\bar{y}_{-1})^2} \\
&= \frac{y_0^2\psi^2\left[\frac{1-\psi^{2T}}{1-\psi^2} - T^{-1}\left(\frac{1-\psi^T}{1-\psi}\right)^2\right] + O_p(y_0)}{y_0^2\left[\frac{1-\psi^{2T}}{1-\psi^2} - T^{-1}\left(\frac{1-\psi^T}{1-\psi}\right)^2\right] + O_p(y_0)} \\
&\xrightarrow[p]{} \psi^2.
\end{aligned}$$

Our calculations imply that the AE is consistent as y_0 tends to infinity:

$$\begin{aligned}
\hat{\psi}_{ml} + \frac{T-3}{2\,\hat{c}} - \sqrt{\frac{(T-3)^2}{4\,\hat{c}^2} + \hat{\psi}_{ml}^2 - \hat{l}} &\xrightarrow[p]{} \psi + \frac{T-3}{2c} - \sqrt{\frac{(T-3)^2}{4c^2} + \psi^2 - \psi^2} \\
&= \psi.
\end{aligned}$$

Moreover, AE_{ap} is found consistent, too, by replacing $\hat{c}$ and c above by $(T-1)/2$ which is the theoretical value of c under a unit root (formula (6.14)).

Starting value asymptotics for $\hat{\psi}_{\mathrm{ml}}$ and $\hat{\psi}_{\mathrm{ad}}$ under the Bhargava AR(1) model with constant when $\mathbf{x_0 = 0}$ is assumed (Section 6.6)

We shall analyse the residual variance, of which the profile and adjusted profile likelihoods are essentially composed, because we do not have an analytical formula for the MLE or the AE. We shall assume that the starting value of the unobservable time series x_0 tends to infinity in absolute value which implies that $|x_0/\sigma|$ tends to infinity, too, for a fixed σ. In contrast, to operationalise the model we assumed in Section 6.6 that x_0 takes value zero. The analysis below reveals that the MLE and the AE, derived from a likelihood function which employs the erroneous assumption $x_0 = 0$, are not consistent in general as $|x_0/\sigma|$ tends to infinity. We shall assume a stationary process because we know already from the end of Section 6.6 that the estimates are invariant with respect to the starting value and hence inconsistent (as $|x_0/\sigma|$ tends to infinity) when a unit root exists. We shall also assume that the MLE and the AE are bounded.

A preliminary result is (from equation (6.48)):

$$y_t = \psi_*^t x_0 + \gamma_0 + \sum_{i=1}^{t} \psi_*^{t-i}\epsilon_i = \psi_*^t x_0 + O_p(1)$$

for ψ_* — the true value — in the range $(-1, 1)$. The true value of the autoregressive coefficient needs to be earmarked to avoid confusion with the parameter ψ with respect to which we are about to minimise a function. Substituting the above expression and formula (6.47) into the formula (6.50) for the residual variance yields

$$\begin{aligned}
\hat{\sigma}^2_{\psi} \propto\ & \sum_{t=1}^{T} y_t^2 - 2\psi \sum_{t=2}^{T} y_t y_{t-1} + \psi^2 \sum_{t=2}^{T} y_{t-1}^2 \\
& - \frac{\left[y_1 + (1-\psi)\left(\sum_{t=2}^{T} y_t - \psi \sum_{t=2}^{T} y_{t-1} \right) \right]^2}{1 + (T-1)(1-\psi)^2} \\
= & \left\{ \frac{\psi_*^2(1-\psi_*^{2T}) - 2\psi\psi_*^3(1-\psi_*^{2T-2}) + \psi^2\psi_*^2(1-\psi_*^{2T-2})}{1-\psi_*^2} \right. \\
& \left. - \frac{\left\{ \psi_* + (1-\psi)\left[\frac{\psi_*^2(1-\psi_*^{T-1})}{1-\psi_*} - \psi \frac{\psi_*(1-\psi_*^{T-1})}{1-\psi_*} \right] \right\}^2}{1 + (T-1)(1-\psi)^2} \right\} x_0^2 + O_p(x_0)
\end{aligned}$$

after some algebra.[2] The first term dominates the behaviour of the residual variance when x_0 is large. Numerical analysis reveals that a local or a global minimum (within a reasonable range) arises around unity unless the autoregressive coefficient ψ_* is very close to minus unity or the sample size is very small. Evaluation of the asymptotic expression for a range of values of ψ and T suggests that the minimum is global if $\psi_* > 0$ and local if $\psi_* < 0$ in which case a global minimum emerges around the true value ψ_*. A local minimum around — but not necessarily exactly at — ψ_* can arise also when $\psi_* > 0$. If ψ_* equals zero then the coefficient of x_0^2 above collapses to zero and the $O_p(x_0)$ term comes into play. A few simulation experiments suggest that the MLE takes then values randomly around zero.

The coefficient of x_0^2 is graphed in Figure A5.2 for $\psi_* = -0.3$ and $\psi_* = 0.3$ to exemplify these remarks. In the former case a curve results with a global minimum around -0.3 and a local minimum around 1; in the latter case the global minimum takes place around 1 and the local minimum around 0.3.

Numerical experimentation exposes also that the minima would occur at the true values if the term with $1 + (T-1)(1-\psi)^2$ in the denominator or just ψ_* of $\psi_* + (1-\psi)[\cdot]$ in the numerator of the term were excluded from the above formula. Existence of the term derives from the special treatment of the starting value under the $\mathrm{AR}^B_\mu(1)$ model.

Similar remarks apply to adjusted profile log-likelihood because the adjustment is additive and does not depend on x_0 in the present context. The findings provide a theoretical explanation for the simulation results reported in Section A6.5 (especially Figure A6.1).

Some heuristics for the distortion towards unity is that trend like behaviour results under $|\psi_*| < 1$ excluding the circumstance $\psi_* = 0$ when x_0 lies far from zero. Such a configuration should in general result only under a unit root when the model is $\mathrm{AR}^B_\mu(1)$ with $x_0 = 0$. Other effects are involved, too, as the unity value maintains attraction even when x_0 is explicitly taken account of, a situation explored in the next section.

[2] If we had not excluded the possibility that the MLE could diverge then the remainder term should be $O_p(\psi x_0)$ instead of $O_p(x_0)$. This would complicate the analysis, cf. Appendix A3 where a corresponding situation is handled without assuming boundedness of the MLE.

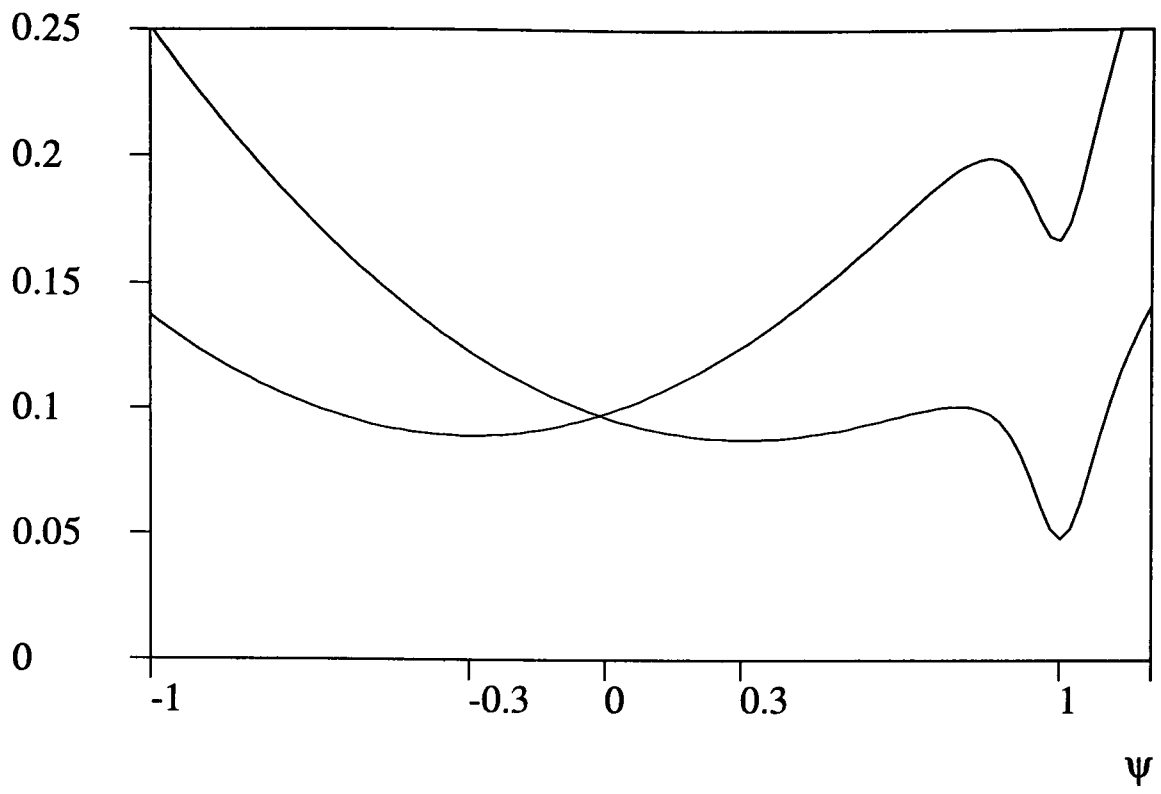


Figure A5.2 Residual variances with two minima for $|x_0|$ large, $T = 100$, and $\psi_0 = -0.3$ or $\psi_0 = 0.3$ ($\mathrm{AR}_{\mu}^{B}(1)$).

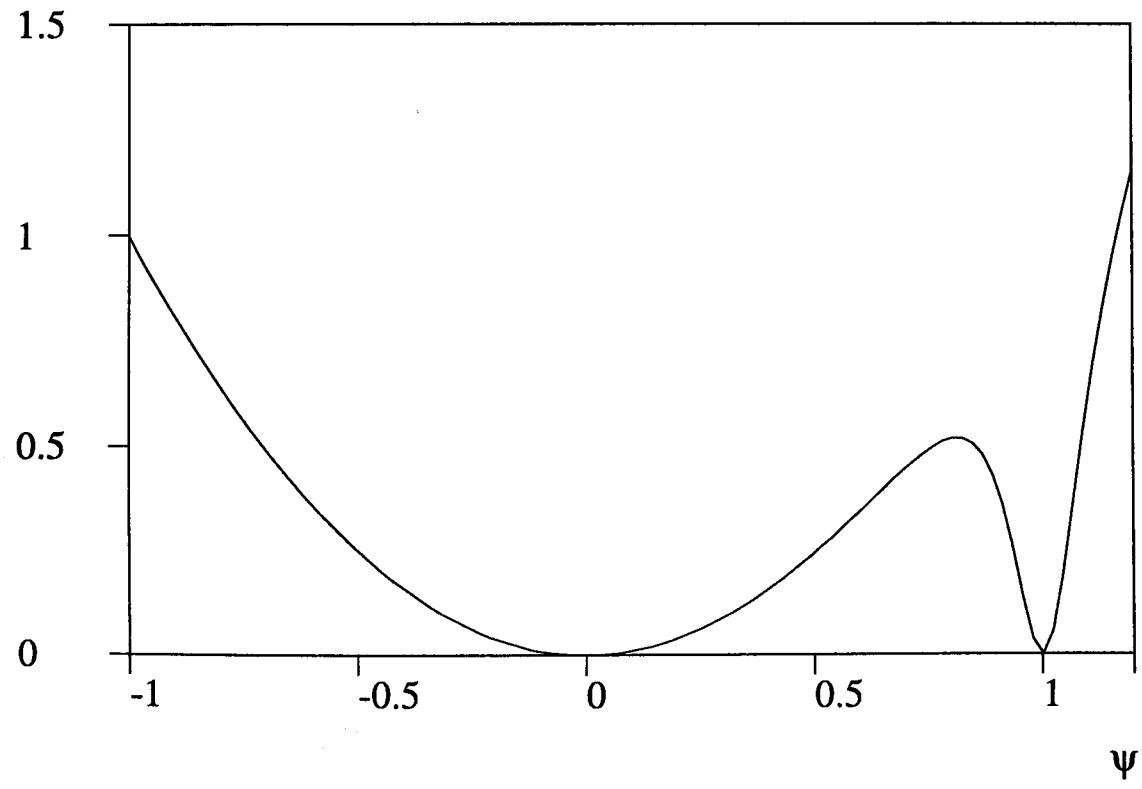


Figure A5.3 Residual variance with two minima for $|x_0|$ large, $T = 100$, and $\psi_0 = 0$ ($\mathrm{AR}_{\mu}^{B}(1)$ with x_0 incorporated in the model).

Starting value asymptotics for $\hat{\psi}_{ml}$ and $\hat{\psi}_{ad}$ under the Bhargava AR(1) model with constant when $x_0 \neq 0$ is allowed for

When x_0 is explicitly allowed to differ from zero in the $AR_{\mu}^{B}(1)$ model the residual variance and the approximate form of it for x_0^2 large become:[3]

$$\begin{aligned}
\hat{\sigma}_{\psi}^{2} \propto\ & \textstyle\sum_{t=1}^{T} y_t^2 - 2\psi\left(y_1 x_0 + \sum_{t=2}^{T} y_t y_{t-1}\right) + \psi^2\left(x_0^2 + \sum_{t=2}^{T} y_{t-1}^2\right) \\
& - \frac{\left[y_1 - \psi x_0 + (1-\psi)\left(\sum_{t=2}^{T} y_t - \psi \sum_{t=2}^{T} y_{t-1}\right)\right]^2}{1 + (T-1)(1-\psi)^2} \\
=\ & \left\{ \frac{\psi_*^2(1-\psi_*^{2T}) - 2\psi\psi_*(1-\psi_*^{2T}) + \psi^2(1-\psi_*^{2T})}{1-\psi_*^2} \right. \\
& \left. - \frac{\left\{\psi_* - \psi + (1-\psi)\left[\frac{\psi_*^2(1-\psi_*^{T-1})}{1-\psi_*} - \psi\frac{\psi_*(1-\psi_*^{T-1})}{1-\psi_*}\right]\right\}^2}{1 + (T-1)(1-\psi)^2} \right\} x_0^2 \\
& + O_p(x_0)
\end{aligned} \tag{A5.2}$$

Numerical evaluation of this expression establishes that the MLE based on the above expression is consistent or that a global minimum of the formula arises at ψ_* for ψ_* in the range $(-1, 1)$ excluding 0. Moreover, the minimum is sharper than for the corresponding expression in the previous section when both functions yield consistent estimates or when $\psi_* < 0$. Increased curviness indicates faster convergence so it pays to include x_0 in the likelihood even if ψ_* were known to be negative and x_0 to be large in absolute value.

The model shows still, with the magnitude of x_0 taken account of, an attraction towards unity, as a local minimum tends to arise at it. When ψ_* equals zero the function takes value zero at both zero and unity and the MLE does not seem to be uniquely defined. The incident is illustrated in Figure A5.3. The residual variance appears to decrease to zero at the two indicated values of ψ. Intuitively this should not happen and indeed we have not inspected whether the smaller order terms shrink to zero, too, at these values. However, such an artificial phenomenon is possible in principle and occurs for example in the estimation of the shifted power transformation (Atkinson (1985, Section 9.3)). The MLE would be uniquely defined if the term $O_p(x_0)$ contributed to $\hat{\sigma}_{\psi}^{2}$ less at zero than at unity but still the local minimum at unity would persist.

[3] Again boundedness of the MLE is assumed as explained in the footnote to the corresponding formula in the previous section.

Appendix A6
Simulation Techniques

A6.1 General Notes

Details of the simulation procedures are given in this appendix. Monte Carlo Theory is not considered in any depth as the purpose is simply to document the methods used. Hendry (1984) and Davidson and MacKinnon (1994) provide surveys on simulation.[1]

The codes for simulation of the finite sample fractiles of the statistics under the $AR_\mu(1)$, $AR_\mu^B(1)$, and AR(2) models are also presented to exemplify the programmes we have created for the GAUSS language and to ease repetition of our results. The simulations for the $AR_\mu(1)$ model employed version 3.0, the simulations for the $AR_\mu^B(1)$ model version 3.1.1 and the simulations for the AR(2) model version 3.2.15. The power evaluations can be easily added to these pieces of code. The programmes generate and save the statistics but do not calculate the fractiles which is accomplished by another (unreported) programme. A further (unreported) programme generates draws from the asymptotic distributions. The simulation experiments were performed on 486 and Pentium processor based PCs.

A6.2 Generation of the Random Numbers

The *multiplicative congruential generator*

$$e_t = \frac{z_t}{m}, \qquad z_t = \lambda z_{t-1}(mod\ m)$$

was used to produce the pseudo-random numbers. Here e_t is the t^{th} pseudo-random number, $t = 1, ..., T$, z_t is a positive integer, z_0 is the *seed*, λ is called the *multiplier*, $(mod\ m)$ is the *modulus function* m being the *modulus*. The resulting numbers are pseudo U(0,1) (Uniform) but can be converted to pseudo $N(\mu,\sigma^2)$ in GAUSS. The default values in GAUSS, $\lambda = 397204094$ and $m = 2^{31} - 1$, were used. The choice of a good pair of λ and m is crucial for a successful simulation experiment. The indicated pair has been reported suitable (Quandt (1983)).[2]

The mean of the pseudo-Normal random numbers is set to zero and the variance to one without loss of generality with all of the simulations.[3] The seeds are extracted runs from the thousand first decimals of π. The estimation of the fractiles and construction of the histograms of the statistics make use of 100 000 replications. The estimates of power are based on 10 000 draws.

The estimation of the asymptotic fractiles in Table 7.1 and of the (unreported) benchmark fractiles of the DF distributions are based on approximating the Brownian motions with random walks of length 1 000 (y_t, $t = 1, ..., 1000$). For this purpose, the functionals in the numerators of formulae (3.11) and (6.8) were replaced by equivalent but computationally more useful forms, namely $\int_0^1 W(r)dW(r)$ by $\frac{1}{2}[W(1)^2 - 1]$, $\int_0^1 W_*(r)dW(r)$ by $\frac{1}{2}[W(1)^2 - 1] - W(1)\int_0^1 W(r)dr$

---

[1]The latter argue (p. 754) that variance reduction techniques (or control variates) are not useful when the processes feature a unit root or are very autocorrelated. This influenced our decision not to apply such methods.

[2]Quandt (*op. cit.*) and Dudewicz and Mishra (1988, p. 191) recite other proper pairs.

[3]The mean and the variance are referred to here in the sense pseudo-random variables can be considered to have such characteristics.

and $\int_0^1 W_*(r)^2dr$ by $\int_0^1 W(r)^2dr - \left[\int_0^1 W(r)dr\right]^2$. In the simulation $W(1)$ is approximated by $y_{1000}/\sqrt{1000}$, $\int_0^1 W(r)dr$ by $\bar{y}\,/\sqrt{1000}$ and $\int_0^1 W(r)^2dr$ by $\sum_{t=1}^{1000} y_t^2/1000^2$.

The seeds used to generate the unit-root distributions (including the unreported benchmark fractiles of the *DF* distributions) are reported in Table A6.1. Different seeds are used for each entry of the table to avoid dependence between them. The seeds for generating power estimates can be found from Table A6.2. The same seeds and hence random numbers are used to facilitate comparison of the power of the tests. The estimation of the deviations of the fractiles of the unit root with drift distributions was based on the seeds documented in Table A6.3. The seeds for the simulations on the AR(2) model can be found from Table A6.4.

T	seed
25	83026425
50	481117450
75	10597
100	27038857
250	78577
500	219608
1000	507992
∞	8925903

Table A6.1 The seeds used in the simulation of the fractiles in Tables 7.1, 7.3, 7.5, 7.17, 7.20, 7.33 and 7.34 (excluding the results for the AR(2) model).

ψ	seed
0.99	98336
0.95	875937
0.90	18577
0.85	812848111
0.80	146844090
0.70	6286208
0.60	46908302
0.50	710100031
0.40	1320005

Table A6.2 The seeds used in the simulation of powers in Tables 7.8, 7.9, 7.10, 7.24, and 7.25.

T	seed
25	20005681
100	44622948
250	1027019

Table A6.3 The seeds used in the simulation of the entries in Tables 7.13 and 7.14.

A6.3 Calculation of the Confidence Intervals

We report a few confidence intervals for the simulated fractiles in Chapter 7 to get an idea of the reliability of our estimates in general. The calculation of the intervals is based on the following well-known result (cf. Dudewicz and Mishra (1988, p. 660), say):

T	seed
25	1195909
100	415926
250	2164201

Table A6.4 The seeds used in the simulation of the entries in Tables 7.28–7.34 for the AR(2) model.

$$\begin{aligned} P\left[z_{(r)} < \xi_p < z_{(s)}\right] &= \textstyle\sum_{i=r}^{s-1} \binom{T}{i} p^i (1-p)^{T-i} \\ &\approx \Phi\left[\frac{s-1-Tp}{\sqrt{Tp(1-p)}}\right] - \Phi\left[\frac{r-Tp}{\sqrt{Tp(1-p)}}\right] \end{aligned}$$

where $z_{(k)}$, $k = 1, ..., T$, denotes the kth order statistic of a random sample $\mathbf{z} = \{z_1, ..., z_T\}$ from the distribution of a random variable Z, ξ_p is the pth fractile of the distribution of Z, $P[\cdot]$ denotes probability, and $\Phi[\cdot]$ is the value of the Standard Normal distribution function at $[\cdot]$. The second line of the above formula is due to the normal approximation of a binomial sum (the approximation is legitimate if $Tp(1-p)$ is not too small). The approximation should hold well in the present case as the simulations comprised 100 000 ($= T$) replications.

A 95 % confidence interval for the fractile ξ_p ensues (approximately) by setting the terms inside the square brackets equal to 1.96 and -1.96, in this order, as then $P\left[z_{(r)} < \xi_p < z_{(s)}\right] \approx 0.95$. (The numbers 1.96 and -1.96 are the 0.975^{th} and 0.025^{th} fractiles, respectively, of the Standard Normal distribution.) This implies that the confidence interval can be determined to be $(z_{(\hat{r})}, z_{(\hat{s})})$ where

$$\begin{aligned} \hat{r} &\equiv 1.96\sqrt{Tp(1-p)} + Tp + 1 \qquad \text{and} \\ \hat{s} &\equiv -1.96\sqrt{Tp(1-p)} + Tp. \end{aligned}$$

The confidence intervals are of different width when the distribution is skewed. This is the reason for giving the maximum spans separately for the left and right tails of the distributions reported in Chapter 7.

The powers were estimated by denoting rejections by one and nonrejections by zero and calculating the average ($\hat{p}$). Each observation is a Bernoulli experiment with variance $p(1-p)$ so the variance of $\hat{p}$ is $p(1-p)/T$. Asymptotic normality applies to $\hat{p}$ (which is an average) so the evaluation of accuracy of it can be based on the estimated standard error $\sqrt{\hat{p}\,(1-\hat{p})/T}$ where $T = 10\ 000$.

A6.4 A Simulation Programme for the AR(1) Model with Constant

The symbols are explained on comment lines within the code or are self-explanatory. The Wald statistics are allocated a value of 1 000 000 000 when the argument of the square-root term is negative. This is done only to enable a check if such cases have occurred. Such an allocation can take place only when the adjusted profile log-likelihood is convex at the AE or when the adjusted profile log-likelihood does not possess a local maximum within $(-1, 1]$ and the AE is allocated a value of unity or minus unity. We note that in the first case a value of zero might seem more justified. In the latter circumstance it is more difficult to specify a value which the statistic should take. One could argue for minus infinity because the adjusted likelihood is decreasing in ψ but this would ignore the curvature of the adjusted likelihood at the adjusted estimate. If the reader plans to examine very small samples he may want to write a few more lines of code for a separate check and e.g. substitute zero in place of 1 000 000 000 when the

first mentioned case arises.

The multiplying factors of the SCs should be changed from T to $T^{3/2}$ when a non-zero drift is assumed ('alfa' below is set to a non-zero value). It takes less than nine minutes to run the programme (with 100 000 replications) on a PC with a Pentium 133 MHz processor and 16 megabytes of RAM. Before running the programme one should create a subdirectory called c:\simu (and c:\gauss if such does not exist) on the hard disk for saving the results.

```
@ Remember to change the seed and the output file below (s100ml, say)
accordingly with T. Keep track also of the starting value, alfa (the
constant) and a (the autoregressive coefficient). @

t1 = hsec;     /* start timer */

@ The programme starts by spesifying the lenght (T) of the time series, the
autoregressive coefficient, and the constant (alfa). @

T = 100; a = ones(1,1); alfa = 0.0;

@ Next the seed is specified and the time series (y) is initialized (with
zeros). @

rndseed 27038857; y = zeros(T,1);

@ The files in which the simulated rejection outcomes will be saved, one by
one (c.f. p. 1116 of the manual). Remember to change the names of the
variables accordingly with T! @

let vnames1 = ml nml;
create fh1 = c:\simu\s100ml with ^vnames1,0,8;
let vnames2 = mlc nmlc;
create fh2 = c:\simu\s100mlc with ^vnames2,0,8;
let vnames3 = c;
create fh3 = c:\simu\s100c with ^vnames3,0,8;
let vnames4 = adc nadc;
create fh4 = c:\simu\s100adc with ^vnames4,0,8;
let vnames5 = ic;
create fh5 = c:\simu\s100ic with ^vnames5,0,8;
let vnames6 = iadc niadc;
create fh6 = c:\simu\s100iadc with ^vnames6,0,8;
let vnames7 = wc;
create fh7 = c:\simu\s100wc with ^vnames7,0,8;
let vnames8 = awpc;
create fh8 = c:\simu\s100awpc with ^vnames8,0,8;
let vnames9 = awc;
create fh9 = c:\simu\s100awc with ^vnames9,0,8;
let vnames10 = adcp nadcp;
create fh10 = c:\simu\s100adcp with ^vnames10,0,8;

@ The number of replications N is indicated, the starting value is
specified, the first random vector is drawn and y[1] is set (effectively)
equal to a times the starting value+constant+e[1]: The recserar command
would ignore e[1] if y0=zeros(1,1), say, were specified. By defining
y0=a*stvalue+e[1] we actually get an y[1] equal to a times the starting
value + constant (alfa)+e[1] and a time series with starting value (y[0]) as
specified. @
```

```
N = 100000; stvalue = 0;
e = alfa+rndn(T,1); y0 = a*stvalue+e[1];

@ The time series is constructed from e. @

y = recserar(e,y0,a);

@ y is lagged by a period. @
y1 = lag1(y); y1[1] = stvalue; /* y1[1] would otherwise be missing */

@ The numerator and the denominator of the MLE and the MLE itself are
evaluated. @

yy1 = y.*y1;
y1sq = (y1)^2;
cyy1 = sumc(yy1); /* c as cumulative */
cy1sq = sumc(y1sq);
ml = cyy1/cy1sq;
nml = T*(ml-1);  /* n as normed */

@ Next, the numerator and the denominator of the MLE with constant and the
MLE with constant itself are generated. @

my = meanc(y);
my1 = meanc(y1);
crp = (y-my).*(y1-my1);     /* crp as cross product */
syy1 = sumc(crp);
lsq = (y1-my1)^2;           /* l as lagged */
sy1y1 = sumc(lsq);
mlc = syy1/sy1y1;
nmlc = T*(mlc-1);

@ The adjusted MLEs are calculated (or adc: ad as adjusted, c as (with)
constant. The coefficient c is evaluated first and then the other sample
values. Next it is tested if the square root term of adc is real and adc is
allocated a value accordingly. @

c = 1/(1-mlc)-(1-mlc^T)/(T*(1-mlc)^2);
sq = (y-my)^2;
syy = sumc(sq);
one = syy/sy1y1;     /* l hat in the text */
sqterm = sqrt(((T-3)^2)/(4*c^2)+mlc^2-one);
im = imag(sqterm);
if c > 0 and im == 0;
  adc = mlc+(T-3)/(2*c)-sqterm;
elseif c > 0;
  adc = 1;
elseif c < 0 and im == 0;
  adc = mlc+(T-3)/(2*c)+sqterm;
elseif c < 0;
  adc = -1;
else;
  adc = mlc;
endif;
```

```
nadc = T*(adc-1);

@ The one time iterated adjusted estimate (iadc) is evaluated next. The
coefficient c has to be evaluated first at the adjusted estimate (ic). If
the adjusted estimate equaled one then the revised estimate is defined to
equal one, too, and the coefficient c is defined to equal its previous
value. @

if adc == 1;
 iadc = 1; ic=c; goto label1;
endif;
ic = 1/(1-adc)-(1-adc^T)/(T*(1-adc)^2);
isqterm = sqrt(((T-3)^2)/(4*ic^2)+mlc^2-one);
im = imag(isqterm);
if ic > 0 and im == 0;
  iadc = mlc+(T-3)/(2*ic)-isqterm;
elseif ic > 0;
  iadc = 1;
elseif ic < 0 and im == 0;
  iadc = mlc+(T-3)/(2*ic)+isqterm;
elseif ic < 0;
  iadc = -1;
else;
  iadc = mlc;
endif;
label1:
niadc = T*(iadc-1);

@ The a priori adjusted estimate is calculated. The coefficient c is now
always positive so a shorter piece of code than above (for adc) is needed
for the specification of the a priori estimate of psi (adcp). @

cp = (T-1)/2;      /* cp for c a priori */
sqterm = sqrt(((T-3)^2)/(4*cp^2)+mlc^2-one);
im = imag(sqterm);
if cp > 0 and im == 0;
  adcp = mlc+(T-3)/(2*cp)-sqterm;
else;
  adcp = 1;
endif;
nadcp = T*(adcp-1);

@ Next, (square roots of) the Wald statistics are calculated. The
calculation of awpc and awc is based on the derivations on pp. 1C.6 and
1C.23 of my hand written calculation detail. The programme is told to
substitute 10000000000 for awpc or awc if the square root term of them is
not positive. (The adjusted likelihood is U-shaped at places.) @

wc = (mlc-1)*(-T/((mlc^2)-one))^0.5;   /* the root term is >0 */

if -(T-3)*(-one+mlc^2+(adcp-mlc)^2)>=0;
 awpc =
 (adcp-1)*((-(T-3)*(-one+mlc^2+(adcp-mlc)^2))^0.5)/(one-2*adcp*mlc+adcp^2);
else;
 awpc = 10000000000;
endif;
```

```
if -(T-3)*(-one+mlc^2+(adc-mlc)^2)>=0;
 awc =
 (adc-1)*((-(T-3)*(-one+mlc^2+(adc-mlc)^2))^0.5)/(one-2*adc*mlc+adc^2);
else;
 awc = 10000000000;
endif;

@ The estimates of a and c are saved to files fh1=s100ml, fh2=s100mlc,
fh3=s100c, fh4=s100adc, fh5=s100ic, fh6=s100iadc, fh7=s100wc, fh8=s100awpc,
fh9=s100awc, fh10=s100adcp. @

data1 = ml~nml; writer(fh1,data1);
data2 = mlc~nmlc; writer(fh2,data2);
data3 = c; writer(fh3,data3);
data4 = adc~nadc; writer(fh4,data4);
data5 = ic; writer(fh5,data5);
data6 = iadc~niadc; writer(fh6,data6);
data7 = wc; writer(fh7,data7);
data8 = awpc; writer(fh8,data8);
data9 = awc; writer(fh9,data9);
data10 = adcp~nadcp; writer(fh10,data10);

@ The simulation starts. @

screen off;        /* to speed the simulation */
i = 2;
do until i > N;
 e = alfa+rndn(T,1); y0 = a*stvalue+e[1];
 y = recserar(e,y0,a);
 y1 = lag1(y); y1[1] = stvalue;
 yy1 = y.*y1;
 y1sq = (y1)^2;
 cyy1 = sumc(yy1);
 cy1sq = sumc(y1sq);
 ml = cyy1/cy1sq;
 nml = T*(ml-1);
 my = meanc(y);
 my1 = meanc(y1);
 crp = (y-my).*(y1-my1);
 syy1 = sumc(crp);
 lsq = (y1-my1)^2;
 sy1y1  = sumc(lsq);
 mlc = syy1/sy1y1;
 nmlc = T*(mlc-1);
 c = 1/(1-mlc)-(1-mlc^T)/(T*(1-mlc)^2);
 sq = (y-my)^2;
 syy = sumc(sq);
 one = syy/sy1y1;
 sqterm = sqrt(((T-3)^2)/(4*c^2)+mlc^2-one);
 im = imag(sqterm);
 if c > 0 and im == 0;
   adc = mlc+(T-3)/(2*c)-sqterm;
 elseif c > 0;
   adc = 1;
 elseif c < 0 and im == 0;
```

```
   adc = mlc+(T-3)/(2*c)+sqterm;
 elseif c < 0;
   adc = -1;
 else;
   adc = mlc;
 endif;
 nadc = T*(adc-1);
 if adc == 1;
  iadc = 1; ic=c; goto label2;
 endif;
 ic = 1/(1-adc)-(1-adc^T)/(T*(1-adc)^2);
 isqterm = sqrt(((T-3)^2)/(4*ic^2)+mlc^2-one);
 im = imag(isqterm);
 if ic > 0 and im == 0;
   iadc = mlc+(T-3)/(2*ic)-isqterm;
 elseif ic > 0;
   iadc = 1;
 elseif ic < 0 and im == 0;
iadc = mlc+(T-3)/(2*ic)+isqterm;
 elseif ic < 0;
   iadc = -1;
 else;
   iadc = mlc;
 endif;
 label2:
 niadc = T*(iadc-1);
  cp = (T-1)/2;      /* cp for c a priori */
sqterm = sqrt(((T-3)^2)/(4*cp^2)+mlc^2-one);
im = imag(sqterm);
if cp > 0 and im == 0;
  adcp = mlc+(T-3)/(2*cp)-sqterm;
else;
  adcp = 1;
endif;
nadcp = T*(adcp-1);
wc = (mlc-1)*(-T/((mlc^2)-one))^0.5;
 if -(T-3)*(-one+mlc^2+(adcp-mlc)^2)>=0;
 awpc =
 (adcp-1)*((-(T-3)*(-one+mlc^2+(adcp-mlc)^2))^0.5)/(one-2*adcp*mlc+adcp^2);
else;
 awpc = 10000000000;
endif;
if -(T-3)*(-one+mlc^2+(adc-mlc)^2)>=0;
 awc =
 (adc-1)*((-(T-3)*(-one+mlc^2+(adc-mlc)^2))^0.5)/(one-2*adc*mlc+adc^2);
else;
 awc = 10000000000;
endif;
data1 = ml~nml; writer(fh1,data1);
data2 = mlc~nmlc; writer(fh2,data2);
data3 = c; writer(fh3,data3);
data4 = adc~nadc; writer(fh4,data4);
data5 = ic; writer(fh5,data5);
data6 = iadc~niadc; writer(fh6,data6);
data7 = wc; writer(fh7,data7);
data8 = awpc; writer(fh8,data8);
```

```
 data9 = awc; writer(fh9,data9);
 data10 = adcp~nadcp; writer(fh10,data10);
 i = i+1;
endo;
fh1 = close(fh1);
fh2 = close(fh2);
fh3 = close(fh3);
fh4 = close(fh4);
fh5 = close(fh5);
fh6 = close(fh6);
fh7 = close(fh7);
fh8 = close(fh8);
fh9 = close(fh9);
fh10 = close(fh10);
screen on;

@ Timer stops. This gives a proper time unless midnight has been passed: @

elapsed = (hsec - t1)/100;
minutes = elapsed/60;

print "The simulation took"; print minutes; print "minutes.";
print "The simulation employed starting value"; print stvalue;
print "and constant"; print alfa;

print "The estimates are (ml, mlc, adcp, and adc):"; print ml mlc adcp adc;
print "The coefficient c and the iterated version of it are:"; print c ic;
print "The quantity called one is:"; print one;
print "The Wald statistics are (wc, awpc, and awc):"; print wc awpc awc;
end;
```

The last lines tell GAUSS to print the time the exercise took and a few statistics from the last simulation. The latter are printed only to get a hold of the entities we are dealing with. The printed statistics are not in any way different from the rest of the simulated statistics and the programme would function just as well with out these last lines of code.

A6.5 A Simulation Programme for the Bhargava AR(1) Model with Constant

Conducting the simulations for the $\mathrm{AR}_{\mu}^{B}(1)$ model took more effort than for the analytically simpler $\mathrm{AR}_{\mu}(1)$ model. The calculations entail numerical maximisation for which purpose the default optimisation algorithm in GAUSS (the method of Broyden, Fletcher, Goldfarb, and Shanno) is employed. The accuracy of the stopping criteria of the algorithm was set to 10^{-6} for the sample sizes up to 100, to 10^{-5} for the sample sizes 250 and 500, and to 10^{-4} for the sample size 1 000 when simulating the unit-root distributions. Increasing the accuracy with the larger sample sizes led to convergence problems with some draws. The power calculations employed accuracy 10^{-5} throughout. The criteria should be changed in the code according to the sample size to repeat the simulation outcomes exactly.

All simulations employed an initial value of unity for the autoregressive coefficient because using the true value of the autoregressive coefficient as the starting value would lead sometimes to convergence to a subordinate local maximum in the power simulations. As pointed out in Section 7.6 the (misspecified) profile likelihoods and adjusted profile likelihoods emerge sometimes bimodal when the autoregressive coefficient lies in the stationary region and the starting value is large. In such a case a smaller local maximum appears to take place around the true value of the autoregressive coefficient and the larger local maximum around unity.

A way of tracking a case where the algorithm has converged to the smaller maximum instead of the larger one is to calculate the residual variance under the MLE and the AE. The residual

variance should, of course, be smaller under the MLE. The proportion of cases where this was not the case — or the residual variance took a larger value under the AE than under the MLE as located by the search algorithm — was very small for all sample sizes when simulating the fractiles of the unit-root distributions (Table A6.5). We inspected by random a few of these occurrences. It turned out that numerical inaccuracy in locating the maximum had led to a slightly smaller residual variance under the AE than under the MLE in these cases. The estimates were essentially the same as they differed in the eighth decimal place or so.

T	%
25	0.10
50	0.07
75	0.04
100	0.03
250	0.02
500	0.04
1000	0.16

Table A6.5 Per cents of the cases where the residual variance took a smaller value under the AE than the MLE when simulating the unit-root distributions ($AR_{\mu}^{B}(1)$).

Values of the constant and the starting value of the time series were set to zero in the simulations. Theoretically the values are irrelevant but a large constant turned out to prolong the simulation or decelerate convergence, sometimes apparently indefinitely. The figures in Table A6.5 proved also to depend on the value of the constant indicating changes in convergence properties. In the cases we have looked at the fractiles themselves remained the same to reporting accuracy but the estimated powers — which are based on a smaller number of replications — change a bit. The explanation is that the algorithm is more likely to get into problems with converging when the constant differs from zero. When it does, a random search process for a maximum starts which means that GAUSS (or the OPTMUM module) takes for itself some of the random numbers in the random number sequence, leaving a different sequence for future observations. (I thank the technical support personnel of Aptech for the information.)

Table A6.6 reports the corresponding proportion when simulating the powers of the tests. (Blank entries indicate cases where the proportion was zero to two decimal places.) The proportion takes the largest values when the starting value is five (the maximum figure 4.57 arises under $\psi = 0.85$, $T = 250$, and $|x_0|/\sigma = 5$). The algorithm appears to converge well (from the starting value of unity) in other cases or when the starting value is zero or very large.

The problem of convergence to the wrong local maximum is more likely to arise with the profile likelihood than with the adjusted profile likelihood. Figures A6.1 and A6.2 clarify the issue. They are based on an autoregressive coefficient of 0.5, starting value 25, and sample size 25. The former figure displays a local maximum around the true value 0.5 and a larger local maximum around unity. It is obvious that an algorithm looking for a maximum may easily converge to the smaller local maximum if the search starts from the true value. Instead, the adjustment often smooths the likelihood so that the adjusted profile log-likelihood features a unique maximum in the reasonable range of parameters. Such cases can arise easily when the starting value is large. The proportion of cases where the residual variance took a smaller value under the MLE (as located by the search algorithm) than under the AE was in our (unreported) simulations at worst around 75 per cent when the initial value for the search algorithm was the true value. Such distortions would favor artificially the tests based on the MLE or their power would appear artificially large. In empirical work the log-likelihoods should always be looked at!

It is shown in Appendix A5 that both the unadjusted and adjusted profile likelihoods possess an artificial local maximum at unity when the starting value tends to infinity whatever the true value of the autoregressive coefficient. The emergence of a maximum around unity in the figures is explained by this result.

		T		
x_0/σ	ψ	**25**	**100**	**250**
0	**0.99**	0.11	0.05	0.01
	0.95	0.13	0.02	
	0.90	0.17	0.02	
	0.85	0.12	0.00	
	0.80	0.12	0.01	
	0.70	0.08		
	0.60	0.06		
	0.50	0.06		
	0.40	0.03		
5	**0.99**	0.12	0.07	
	0.95	0.18	0.10	0.10
	0.90	0.33	0.39	1.49
	0.85	0.33	1.90	4.57
	0.80	0.65	3.30	1.63
	0.70	1.09	1.80	
	0.60	1.51	0.17	
	0.50	0.81		
	0.40	0.40		
25	**0.99**	0.16		0.02
	0.95			0.01
	0.90			
	0.85			0.01
	0.80			0.01
	0.70			0.01
	0.60			
	0.50			0.07
	0.40		0.01	0.01
250	**0.99**			
	0.95			
	0.90			
	0.85			
	0.80			
	0.70			
	0.60			
	0.50			
	0.40			

Table A6.6 Per cents of the cases where the residual variance took a larger value under the AE than the MLE when simulating powers of the tests ($AR_{\mu}^{B}(1)$).

When the code is run the first time in a GAUSS session under GAUSS 3.1 and OPTMUM 3.0 module for GAUSS an error message on duplicate definitions of locals occurs. The message is due to using a newer version of GAUSS than OPTMUM. The message can be avoided by deleting the local variables from the referred lines in the file OPTMUM.SRC (they appear as arguments in the corresponding procedures so they do not need to be included in the locals). It does not occur if the latest versions of GAUSS (3.2.15 at the time of writing this) and OPTMUM are used. (I thank the technical support personnel of Aptech for the advice.)

The programme assumes that subdirectories c:\simu and c:\gauss exist. The programme runs (with 100 000 replications) for about an hour and 50 minutes on a Pentium 133 MHz based PC with 16 megabytes of RAM.

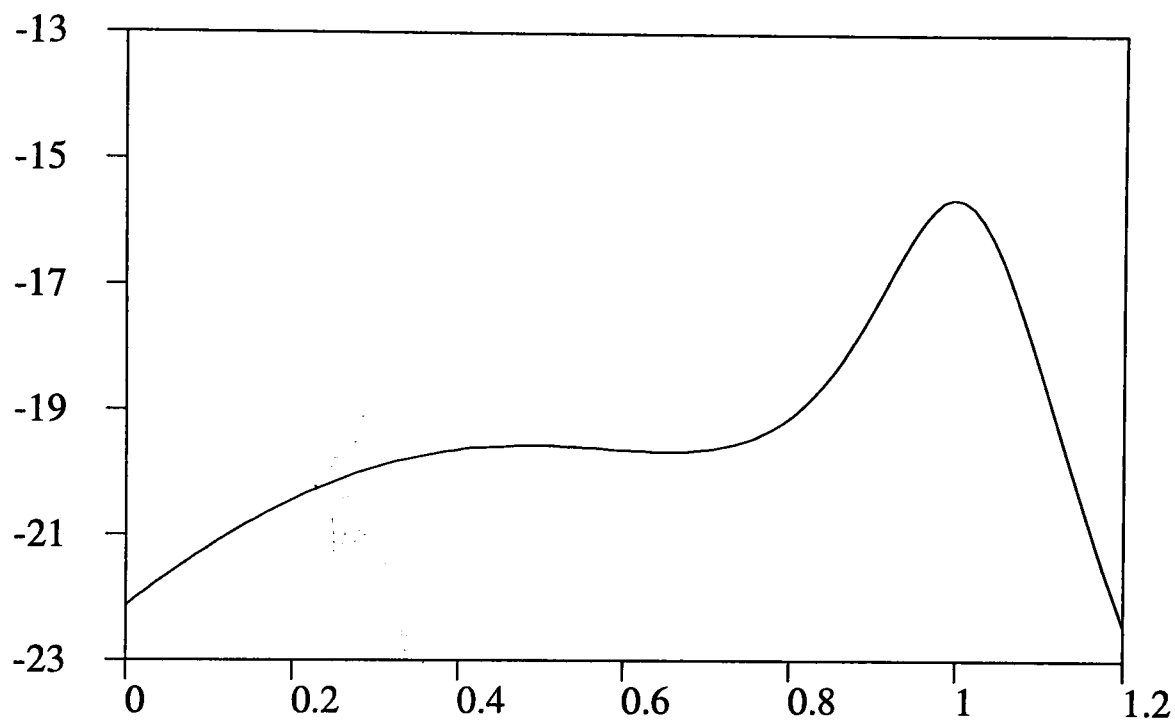


Figure A6.1 An example of a bimodal profile log-likelihood ($AR_{\mu}^{B}(1)$).

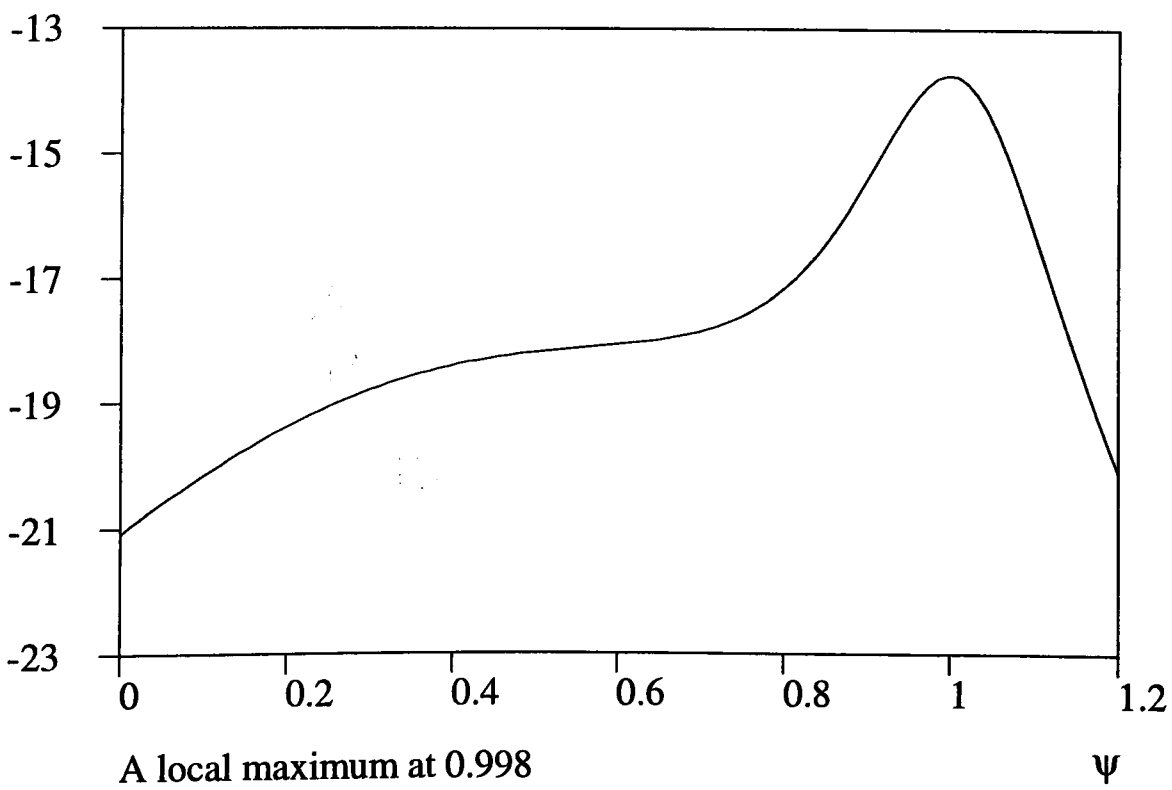


Figure A6.2 An example of adjusted profile log-likelihood ($AR_{\mu}^{B}(1)$).

```
@ Small-sample fractiles are simulated when the model is the Bhargava
AR(1) with constant. Remember to change the seed according to the
autoregressive coefficient, to define the starting value (stvalue) and the
constant (g0), to change the name of the files (b100ols, say) where the
results are saved, and to change the stopping criteria (_opgtol) according
to the sample size. One may want to change also the initial value in the
optimisation algorithm defined (some way) below (ip). Note though that the
likelihood fuction is valid only when the starting value is zero. @

t1 = hsec;     /* start timer */

@ It is a good programming practise to set the default values of optmum in
the beginning of the programme even though they were changed later: @

optset;

@ The programme starts by spesifying the lenght (T) of the time series,
the autoregressive coefficient (a), the constant (g0), the starting value
(stvalue, of x), the seed (rndseed), and the number of replications (N). @

T = 100; a = 1; g0 = 0; stvalue = 0; rndseed 27038857; N = 100000;

@ The progress of the numerical maximisation (actually minimisation)
procedure is not shown on the screen when the parameter below is set
equal to zero speeding the simulation. @

__output = 0;

@ The optimisation method is chosen (2 is the default). @

_opalgr = 2;

@ Criterion according to which OPTMUM exists the iterations is tuned (10^-4
is the default). @

_opgtol = 10^(-6);

@ The files in which the simulated statistics will be saved, one by one, are
created. The derivative of the residual variance evaluated at the MLE - which
should equal zero by definition - will be saved into the file b100c1 for a check-
up and another checking procedure, defined below, into the file b100c2. @

let vnames1 = ols nols;
create fh1 = c:\simu\b100ols with ^vnames1,0,8;
let vnames2 = mlc nmlc;
create fh2 = c:\simu\b100mlc with ^vnames2,0,8;
let vnames3 = adc nadc;
create fh3 = c:\simu\b100adc with ^vnames3,0,8;
let vnames4 = w;
create fh4 = c:\simu\b100w with ^vnames4,0,8;
let vnames5 = wc;
create fh5 = c:\simu\b100wc with ^vnames5,0,8;
let vnames6 = awc;
create fh6 = c:\simu\b100awc with ^vnames6,0,8;
let vnames7 = ds2mlc;
create fh7 = c:\simu\b100c1 with ^vnames7,0,8;
```

```
let vnames8 = err;
create fh8 = c:\simu\b100c2 with ^vnames8,0,8;

@ The simulation starts. @

i = 1;
do until i > N;

@ The following piece of code generates a time series x (and the lagged
value of it) which follows an AR(1) process with (asymptotically) zero mean
and with starting value (x0) as specified above, and a time series y which
equals x+g0 and for which the starting value is undefined (and the lagged
value of y). The time series are first initialized (with zeros). @

x = zeros(T,1); y = zeros(T,1);
e = rndn(T,1);
x0 = a*stvalue+e[1];
x = recserar(e,x0,a);
x1 = lag1(x); x1[1] = stvalue; /* x1[1] would otherwise be missing */
y = g0 + x;
y1 = lag1(y);  y1[1] = stvalue;

@ y1[1] would otherwise be missing. In fact, y1[1] is supposed to be missing
because y[0] is undefined. However, the construction of some cumulative sums
below require a value for y1[1]. We will deduct it from the sums wherever
needed. @

ysq = y^2;
cysq = sumc(ysq);  /* c as cumulative */
yy1 = y.*y1;
cy1sq = cysq-ysq[T];

@ The contribution of the initial value is eliminated in the following
command because the initial value of y is theoretically undefined. @

cyy1 = sumc(yy1)-y[1]*stvalue;

@ The ols estimate is calculated next for comparison purposes. It can be
calculated sensibly only for observations two onwards because y[0] or the
starting value of y is undefined. The estimate can be expressed similarly as
in the other programmes we have created because of the above revised
definitions of cyy1 and cysq (note that also cy1sq excludes y[0] above). @

ols = cyy1/cy1sq;
nols = T*(ols-1);     /* n as normed */

@ The sum of the yts and yts lagged are needed in the construction of the
log-likelihood. @

cy = sumc(y);
cy1 = cy-y[T];

@ Procedures for numerical maximization of the profile and adjusted
log-likelihood are created. The negatives of the log-likelihoods are
specified here because optmum looks for the minimum of the function
inspected. @
```

```
proc bhmlc(p);   /* p is a local variable */
 retp((T/2)*ln((cysq-2*p*cyy1+(p^2)*cy1sq-
               (y[1]+(1-p)*(cy-y[1]-p*cy1))^2/(1+(T-1)*(1-p)^2))/T));
endp;

proc bhadc(p);   /* p is a local variable */
 retp(((T-3)/2)*ln((cysq-2*p*cyy1+(p^2)*cy1sq-
                   (y[1]+(1-p)*(cy-y[1]-p*cy1))^2/(1+(T-1)*(1-p)^2))/T)
                  +0.5*ln(1+(T-1)*(1-p)^2));
endp;

@ The initial value for the estimate of the autoregressive parameter is
defined to be unity (the true value as an initial value would often lead to
an ill-converged estimate from the profile log-likelihood when the starting
value of x is large). @

ip = 1;

@ fmlc, gmlc, and rcmlc below are quantities created by optmum
automatically. @

{ mlc,fmlc,gmlc,rcmlc }  = optmum(&bhmlc,ip);

nmlc = T*(mlc-1);

{ adc,fadc,gadc,rcadc }  = optmum(&bhadc,ip);

nadc = T*(adc-1);

@ The residual variances under different estimates are calculated. The first
observation of y squared is deducted when evaluating the variance under the
ols estimate because it is calculated for observations two onwards. @

extramlc = ((y[1]+(1-mlc)*(cy-y[1]-mlc*cy1))^2)/(1+(T-1)*(1-mlc)^2);
extraadc = ((y[1]+(1-adc)*(cy-y[1]-adc*cy1))^2)/(1+(T-1)*(1-adc)^2);
s2ols = (cysq-ysq[1]-2*ols*cyy1+(ols^2)*cy1sq)/T;
s2mlc = (cysq-2*mlc*cyy1+(mlc^2)*cy1sq-extramlc)/T;
s2adc = (cysq-2*adc*cyy1+(adc^2)*cy1sq-extraadc)/T;

@ The figure one is allocated for the variable err if the residual variance
is larger under the MLE than under the AE. A value of one would indicate a
failure in the maximisation exercise. @
if s2mlc > s2adc;
 err = 1;
else;
 err = 0;
endif;

@ The constants evaluated at the two estimates are: @

g0mlc = (y[1]+(1-mlc)*(cy-y[1]-mlc*cy1))/(1+(T-1)*(1-mlc)^2);
g0adc = (y[1]+(1-adc)*(cy-y[1]-adc*cy1))/(1+(T-1)*(1-adc)^2);

@ The relevant derivatives of the residual variances, evaluated at the two
estimates, are calculated. @

ds2adc = (2/T)*(-cyy1+adc*cy1sq+
```

```
     (1-adc)*(y[1]+(1-adc)*(cy-y[1]-adc*cy1))*cy1/(1+(T-1)*(1-adc)^2)+
     ((1-adc)*(cy-y[1]-adc*cy1)+y[1])*(cy-y[1]-adc*cy1-(T-1)*(1-adc)*y[1])
     /(1+(T-1)*(1-adc)^2)^2);

@ ds2mlc is calculated only as a check-up: it should equal zero at the MLE
or mlc). @

ds2mlc = (2/T)*(-cyy1+mlc*cy1sq+
     (1-mlc)*(y[1]+(1-mlc)*(cy-y[1]-mlc*cy1))*cy1/(1+(T-1)*(1-mlc)^2)+
     ((1-mlc)*(cy-y[1]-mlc*cy1)+y[1])*(cy-y[1]-mlc*cy1-(T-1)*(1-mlc)*y[1])
     /(1+(T-1)*(1-mlc)^2)^2);

dds2mlc = (2/T)*(cy1sq-2*g0mlc*cy1+(T-1)*g0mlc^2-
         ((cy-y[1]-(T-1)*g0mlc*(1-mlc)-mlc*cy1+(1-mlc)*(cy1-(T-1)*g0mlc))^2)
         /(1+(T-1)*(1-mlc)^2));

dds2adc = (2/T)*(cy1sq-2*g0adc*cy1+(T-1)*g0adc^2-
         ((cy-y[1]-(T-1)*g0adc*(1-adc)-adc*cy1+(1-adc)*(cy1-(T-1)*g0adc))^2)
         /(1+(T-1)*(1-adc)^2));

@ The corresponding Wald statistics are calculated. @

w = (ols-1)/sqrt((s2ols/cy1sq));
wc = (mlc-1)*sqrt((T/2)*dds2mlc/s2mlc);
awc = (adc-1)*sqrt(((T-3)/2)*(dds2adc/s2adc-(ds2adc/s2adc)^2)
                  +(T-1)*(1-(T-1)*(1-adc)^2)/((1+(T-1)*(1-adc)^2)^2));
@ The statistics are saved to files b100ols, b100mlc, b100adc, b100w,
b100wc, and b100awc. The screen is turned off. @

screen off;
data1 = ols~nols; writer(fh1,data1);
data2 = mlc~nmlc; writer(fh2,data2);
data3 = adc~nadc; writer(fh3,data3);
data4 = w; writer(fh4,data4);
data5 = wc; writer(fh5,data5);
data6 = awc; writer(fh6,data6);
data7 = ds2mlc; writer(fh7,data7);
data8 = err; writer(fh8,data8);

i=i+1;
endo;
output off;
fh1 = close(fh1);
fh2 = close(fh2);
fh3 = close(fh3);
fh4 = close(fh4);
fh5 = close(fh5);
fh6 = close(fh6);
fh7 = close(fh7);
fh8 = close(fh8);
screen on;

@ The maximum deviation of ds2mlc from zero is checked and if err took
values greater than zero (or unity). Finally, the proportion of cases where
err took a value of unity is calculated. @
```

```
open fh7 = c:\simu\b100c1;
z1 = readr(fh7,N);
ds2mlcmx = maxc(z1);
open fh8 = c:\simu\b100c2;
z2 = readr(fh8,N);
errmax = maxc(z2);
z3 = meanc(z2)*100;
fh7 = close(fh7);
fh8 = close(fh8);

@ Timer stops. This gives a proper time unless midnight has been passed: @
elapsed = (hsec - t1)/100;
minutes = elapsed/60;

output file = c:\simu\temp.txt reset;

@ Reporting accuracy is set to three decimal places: @

format /m1 /rd 16,3;

print "The simulation took"; print minutes; print "minutes.
                                                          ";
print "The simulation employed initial value " ip " in the optimisation
algorithm";
print "If the sample size is" T ", the starting
value is" stvalue ", the constant is " g0 ", and the autoregressive
coefficient is" a "then:";
print "The estimates from the last iteration are (ols, mlc, and adc):";
print ols mlc adc;
print "The Wald statistics from the last iteration are (w, wc, and awc):";
print w wc awc;
print "The residual variances are (s2ols, s2mlc, s2adc):";
print s2ols s2mlc s2adc;
print "The constants are (g0mlc and g0adc):";
print g0mlc g0adc;

@ Reporting accuracy is tuned back to eight decimal places: @

format /m1 /rd 16,8;

print "The maximum deviation of the derivative of the residual variance,
evaluated at the MLE, from zero was:";
print ds2mlcmx;

@ Reporting accuracy is tuned to three decimal places: @

format /m1 /rd 16,3;

print "The number below is one if the residual variance were larger under
the MLE than AE in any of the simulations:";
print errmax;
print "In" z3 "per cent of the" N "cases the residual variance took a larger
value under the MLE than under the AE.";

output off;
```

```
closeall;
end;
```

A6.6 A Simulation Programme for the Unit-Root AR(2) Model

Tables A6.7 and A6.8 document the proportions of draws in which the adjusted profile likelihood equation featured complex roots, and a value of minus unity or unity, respectively, was allocated for the adjusted estimate(s). A blank entry indicates that such cases did not arise at all in the simulation exercise, and the figure '0.00' means that the proportion was zero to two decimal places. Substitutions emerged mainly for the SAE when ϕ_1 was large.

	T					
	25		100		250	
ϕ_1	AE	SAE	AE	SAE	AE	SAE
0.95	0.01		0.00			
0.50	0.00					
0.00						
−0.50						
−0.95						

Table A6.7 Per cent of the draws in which the AE or the SAE were substituted a value of minus unity (AR(2)).

	T					
	25		100		250	
ϕ_1	AE	SAE	AE	SAE	AE	SAE
0.95	0.06	35.91		5.05		0.02
0.50	0.04	0.50				
0.00		0.00				
−0.50						
−0.95	0.00					

Table A6.8 Per cent of the draws in which the AE or the SAE were substituted a value of unity (AR(2)).

The programme with which we conducted the simulations is given below. It runs under GAUSS 3.2.15, say, but does not under GAUSS 3.1.1.[4] The latter version appears to have a fallacy in the algorithm in evaluating large powers of negative numbers.[5] Such evaluations are needed when calculating the estimate of the coefficient c (the 'exact' version).

The version 3.2.15 will typically give an error message of an underflow when the programme is run. This should be due to the fact that the computing accuracy of the Intel processor of the PC has been reached in the evaluation of a large power of a figure close to zero. The processor substitutes zero in such a case in place of the correct figure. This should be quite unharmful from the point of view of the simulation experiments.

[4]The programme will not run under GAUSS 3.2.15 either if the number of replications, N in the programme, is specified to be four or less (it may run under GAUSS 3.1.1, though). This is a bug in the 3.2.15 version and we have been informed from Aptech that the bug will be fixed in the near future.

[5]For example, the following small piece of code does not run under GAUSS 3.1.1 (though it does under 3.2.15): $T = 40$; $a = -1$; $b = .a^T$; print b;.

There are a few long lines in the programme. If one wants to avoid splitting them and a warning message on the split lines then one should edit the programme first in the DOS editor (instead of the GAUSS editor), say, and run it under GAUSS by giving the command 'run <file_name>'.

The Wald statistics are allocated a value of 1 000 000 000 when the argument of the square-root term is negative as in the programme for the $AR_{\mu}^{B}(1)$ model. This is done only to enable a check if such cases have occurred.

The time to complete a simulation with 100 000 replications is about 15 minutes on a PC with a 133 MHz Pentium processor and 16 megabytes of RAM. Existence of the subdirectories c:\simu and c:\gauss is required.

```
Remember to change the seed and the output file below (s100ml, say)
accordingly with T! Keep track also of the vector a of autoregressive
coefficients! @

t1 = hsec;     /* start timer */

@ The programme starts by spesifying the lenght (T) of the time series and
the vector a of autoregressive coefficients. Note that the autoregressive
coefficients relate here to the usual parameterisation. We employ an
alternative parameterisation in the thesis where the first coefficient is the
sum of the autoregressive parameters and the second equals in absolute
value the original second but is of opposite sign.

The seed is specified, the number of replications N is indicated, and the
vector of starting values is specified. @

T = 100; a = 1.0|0.0; rndseed 415926; N = 100000;
stvalue = 0|0;     /* the value on the left is the more recent one */

@ The files in which the simulated rejection outcomes will be
saved, one by one (cf. p. 1116 of the manual) are created. (An
"a" at the end of a name stands for approximate.) Remember to
change the names of the variables accordingly with T! @

let vnames1 = mlp nmlp;
create fh1 = c:\simu\s100ml with ^vnames1,0,8;
let vnames2 = c;
create fh2 = c:\simu\s100c with ^vnames2,0,8;
let vnames3 = ca;
create fh3 = c:\simu\s100ca with ^vnames3,0,8;
let vnames4 = ad nad;
create fh4 = c:\simu\s100ad with ^vnames4,0,8;
let vnames5 = ada nada;
create fh5 = c:\simu\s100ada with ^vnames5,0,8;
let vnames6 = w;
create fh6 = c:\simu\s100w with ^vnames6,0,8;
let vnames7 = aw;
create fh7 = c:\simu\s100aw with ^vnames7,0,8;
let vnames8 = awa;
create fh8 = c:\simu\s100awa with ^vnames8,0,8;

screen off;     /* to speed the simulation */

@ The simulation starts. @
```

```
i = 1;
do until i > N;

@ The time series (y) is initialized (with zeros), the first random vector
is drawn and y[1] is set (effectively) equal to a' times the starting
values+e[1]: The recserar command would ignore e[1] if y0=zeros(1,1), say. @

y = zeros(T,1); e = rndn(T,1);
y0 = (a'stvalue+e[1])|(a'((a'stvalue+e[1])|stvalue[1,.])+e[2]);
     /* the value on the right is the more recent one! */

@ The time series is constructed from e. @

y = recserar(e,y0,a);


@ y is lagged by a period. y1[1], y2[1], and y2[2] would otherwise be
missing. @

y1 = lag1(y); y2 = lagn(y,2);
y1[1] = stvalue[1,.]; y2[1] = stvalue[2,.]; y2[2] = stvalue[1,.];

@ New variables and cumulative sums of them are calculated. @

yy1 = y.*y1;
ysq = y^2;
y1sq = (y1)^2;
dy1 = lag1(y-y1);
dy1[1] = stvalue[1]-stvalue[2];
dy1sq = (dy1)^2;
ydy1 = y.*dy1;
y1dy1 = y1.*dy1;

cyy1 = sumc(yy1); /* c as cumulative */
cydy1 = sumc(ydy1);
cy1dy1 = sumc(y1dy1);
cysq = sumc(ysq);
cy1sq = sumc(y1sq);
cdy1sq =sumc(dy1sq);

@ The MLEs for the sum of the autoregressive parameters (mlp) and for the
second coefficient (mlf, in the parameterisation which is employed in the
thesis) are evaluated. @

mlp = (cyy1*cdy1sq-cy1dy1*cydy1)/(cy1sq*cdy1sq-(cy1dy1)^2);
nmlp = T*(mlp-1);  /* n as normed */
mlf = (cydy1-mlp*cy1dy1)/cdy1sq;

@ The two versions of the coefficient c, the exact form (c) and the
approximate asymptotic form are evaluated at the appropriate MLE. @

c =
(1-mlf)^(-2)+((1-mlf^T)*(1-mlf^(2*T))*(1+2*mlf-mlf^2-2*mlf^T)-((1-mlf^2)*(1-
mlf^T)*((1+mlf)^2+mlf^(T-1)*(1-mlf^T-mlf*(1+2*mlf+mlf^T))))*T+(mlf^(T-1)*((1
-mlf^2)^2)*(1+mlf-2*mlf^T))*T^2)/(((1-mlf)^2)*((1-mlf^2)*T-(1-mlf^(2*T)))^2)
;
```

```
ca = (1-mlf)^(-2);

@ The statistic "l hat" denoted here by "one" is calculated. @

one = (cysq*cdy1sq-(cydy1)^2)/(cy1sq*cdy1sq-(cy1dy1)^2);

@ The first version of the adjusted MLE or ad based on c is calculated. For
this purpose, it is tested whether the square root term of the adjusted
estimates (AEs) are real and the sign of "mlp^2-one" is inspected. ad is
allocated a value accordingly. @

sqterm = sqrt(((T-3)^2)/(4*c^2)+mlp^2-one);
im = imag(sqterm);
if c > 0 and im == 0 and mlp^2-one < 0;
  ad = mlp+(T-3)/(2*c)-sqterm;
elseif c > 0 and im == 0 and mlp^2-one > 0;
  ad = mlp+(T-3)/(2*c)+sqterm;
elseif c > 0;
  ad = 1;

elseif c < 0 and im == 0 and mlp^2-one < 0;
  ad = mlp+(T-3)/(2*c)+sqterm;
elseif c < 0 and im == 0 and mlp^2-one > 0;
  ad = mlp+(T-3)/(2*c)-sqterm;
elseif c < 0;
  ad = -1;
else;
  ad = mlp;
endif;
nad = T*(ad-1);

@ The second version of the adjusted MLE or ada based on ca is calculated.
For this purpose, it is tested whether the square root term of the AEs are real
and the sign of "mlp^2-one" is inspected. ada is allocated a value accordingly. @

sqterm = sqrt(((T-3)^2)/(4*ca^2)+mlp^2-one);
im = imag(sqterm);
if ca > 0 and im == 0 and mlp^2-one < 0;
  ada = mlp+(T-3)/(2*ca)-sqterm;
elseif ca > 0 and im == 0 and mlp^2-one > 0;
  ada = mlp+(T-3)/(2*ca)+sqterm;
elseif ca > 0;
  ada = 1;
elseif ca < 0 and im == 0 and mlp^2-one < 0;
  ada = mlp+(T-3)/(2*ca)+sqterm;
elseif ca < 0 and im == 0 and mlp^2-one > 0;
  ada = mlp+(T-3)/(2*ca)-sqterm;
elseif ca < 0;
  ada = -1;
else;
  ada = mlp;
endif;
nada = T*(ada-1);

@ Next, (square roots of) the Wald statistics are calculated. The statistics
aw and awa are based on the estimated coefficients c and ca, respectively.
The programme is told to substitute 0 or -10000000000 for aw or  awa if the
```

```
square root term of them is not positive. Zero is substituted when the roots
of the adjusted likelihood equation are complex and the adjusted likelihood
is increasing or the corresponding estimate takes exactly the value unity,
and -10000000000 when the roots are complex but the adjusted likelihood is
decreasing. The code replicates that for the AR(1) model with constant
because the algebra is exactly the same for that model apart from the
different composition of the adjusted estimate and the entity "one". @

w = (mlp-1)*(-T/((mlp^2)-one))^0.5;   /* the root term is >0 */

if -(T-3)*(-one+mlp^2+(ad-mlp)^2)>=0;
 aw =
 (ad-1)*((-(T-3)*(-one+mlp^2+(ad-mlp)^2))^0.5)/(one-2*ad*mlp+ad^2);
elseif ad == 1;
 aw = 0;
else;
 aw = -10000000000;

endif;
if -(T-3)*(-one+mlp^2+(ada-mlp)^2)>=0;
 awa =
 (ada-1)*((-(T-3)*(-one+mlp^2+(ada-mlp)^2))^0.5)/(one-2*ada*mlp+ada^2);
elseif ada == 1;
 awa = 0;
 else;
 awa = -10000000000;
endif;
@ The estimates of a, c and ca are saved to files fh1=s100ml, fh2=s100c,
fh3=s100ca, fh4=s100ad, fh5=s100ada, fh6=s100w, fh7=s100aw, fh8=s100awa. @

data1 = mlp~nmlp; writer(fh1,data1);
data2 = c; writer(fh2,data2);
data3 = ca; writer(fh3,data3);
data4 = ad~nad; writer(fh4,data4);
data5 = ada~nada; writer(fh5,data5);
data6 = w; writer(fh6,data6);
data7 = aw; writer(fh7,data7);
data8 = awa; writer(fh8,data8);
i=i+1;
endo;
output off;

fh1 = close(fh1);
fh2 = close(fh2);
fh3 = close(fh3);
fh4 = close(fh4);
fh5 = close(fh5);
fh6 = close(fh6);
fh7 = close(fh7);
fh8 = close(fh8);

@ The proportions of cases where the two AEs take value unity or minus unity
or the roots of the adjusted likelihood equation are complex are calculated.
@

open fh4 = c:\simu\s100ad;
```

```
u = readr(fh4,N);
u1 = u[.,1];      /* The first column or ad is read */
@ A vector composed of the order numbers of elements equal to unity is
created. @
u2 = indexcat(u1,1);
u3 = u2./u2;      /* the division yields a vector composed only of ones */
u4 = u3'u3;       /* the number of ones is calculated */
u5 = u4/N;        /* the proportion of ones is calculated */
@ A vector composed of the order numbers of elements equal to minus unity is

created. @
u6 = indexcat(u1,-1);
u7 = u6./u6;      /* the division yields a vector composed only of ones */
u8 = u7'u7;       /* the number of minus ones is calculated */
u9 = u8/N;        /* the proportion of minus ones is calculated */
fh4 = close(fh4);

open fh5 = c:\simu\s100ada;
v = readr(fh5,N);
v1 = v[.,1];      /* The first column or ad is read */

@ A vector composed of the order numbers of elements equal to unity is
created. @
v2 = indexcat(v1,1);
v3 = v2./v2;      /* the division yields a vector composed only of ones */
v4 = v3'v3;       /* the number of ones is calculated */
v5 = v4/N;        /* the proportion of ones is calculated */
@ A vector composed of the order numbers of elements equal to minus unity is
created. @
v6 = indexcat(v1,-1);
v7 = v6./v6;      /* the division yields a vector composed only of ones */
v8 = v7'v7;       /* the number of minus ones is calculated */
v9 = v8/N;        /* the proportion of minus ones is calculated */
fh5 = close(fh5);

screen on;
elapsed = (hsec - t1)/100; /* timer stops */
minutes = elapsed/60;

output file = c:\gauss\temp.txt reset;

print "The simulation took"; print minutes; print "minutes.";
print "The sample size and the autoregressive coefficients were:";
print T a';
print "respectively.";
print "The estimates are (mlf, mlp, ad, and ada):";
print mlf mlp ad ada;
print "The coefficient c and the approximate version of it are:";
print c ca;
print "The quantity called one is:"; print one;
print "The Wald statistics are (w, aw, and awa):"; print w aw awa;
print "The AE was substituted a value of unity in:";
print u5*100 "per cent of the draws.";
print "The AE was substituted a value of minus unity in:";
print u9*100 "per cent of the draws.";
print "The approximate AE was substituted a value of unity in:";
```

```
print v5*100 "per cent of the draws.";
print "The approximate AE was substituted a value of minus unity in:";
print v9 "per cent of the draws.";

output off;
closeall;
end;
```

References

Abadir, K.M. (1993): OLS Bias in a Nonstationary Autoregression, *Econometric Theory*, 9, 81–93.

---- (1995): Unbiased Estimation as a Solution to Testing for Random Walks, *Economics Letters*, 47, 263–68.

Abadir, K.M. and Hadri, K. (1996): Bias Nonmonotonicity in Stochastic Difference Equations, Discussion Papers in Economics 96/15, University of York.

Adda, J. and Gonzalo, J. (1996): P-Values for Non-Standard Distributions with an Application to the DF Test, *Economics Letters*, 50, 155–60.

Agbeyegbe, T.D. (1993): The Stochastic Behavior of Mineral-Commodity Prices, in Phillips, P.C.B. (ed.): *Models, Methods, and Applications of Econometrics — Essays in Honor of A.R. Bergstrom*, Basil Blackwell, Cambridge, Massachusetts.

Agiakloglou, C. and Newbold, P. (1992): Empirical Evidence on Dickey-Fuller-Type Tests, *Journal of Time Series Analysis*, 13, 471–83.

Ahn, S.K. (1993): Some Tests for Unit Roots in Autoregressive-Integrated-Moving-Average Models with Deterministic Trends, *Biometrika*, 80, 855–68.

An, S. and Bloomfield, P. (1993): Cox and Reid's Modification in Regression Models with Correlated Errors, North Carolina State University, manuscript.

Andrews, D.W.K. (1993): Exactly Median-Unbiased Estimation of First Order Autoregressive/Unit Root Models, *Econometrica*, 61, 139–65.

Andrews, D.W.K. and Chen, H-Y. (1994): Approximately Median-Unbiased Estimation of Autoregressive Models, *Journal of Business & Economic Statistics*, 12, 187–204.

Atkinson, A. (1985): *Plots, Transformations, and Regression*, Clarendon Press, Oxford.

Bachelier, L. (1900): Theory of Speculation, *Annales Scientifiques de l'Ecole Normale Superieure*, Sup (3), no. 1018. Reprinted in English (translated by Boness, A.J.) in Cootner, P.H. (1964, ed.): *The Random Character of Stock Market Prices*, The MIT Press.

Balke, N.S. and Fomby, T.B. (1991a): Infrequent Permanent Shocks and the Finite-Sample Performance of Unit Root Tests, *Economics Letters*, 36, 269–73.

---- (1991b): Shifting Trends, Segmented Trends and Infrequent Permanent Shocks, *Journal of Monetary Economics*, 28, 61–85.

---- (1994):Large Shocks, Small Shocks, and Economic Fluctuations: Outliers in Macroeconomic Time Series, *Journal of Applied Econometrics*, 9, 181–200.

Banerjee, A. and Hendry, D.F.(1992): An Overview, Special Issue of *Oxford Bulletin of Economics and Statistics*, 54, 225–51.

Banerjee, A., Dolado, J., Galbraith, J.W. and Hendry, D.F. (1993): *Co-Integration, Error-Correction, and the Econometric Analysis of Non-Stationary Data*, Oxford University Press, Oxford.

Banerjee, A., Lumsdaine, R.L. and Stock, J.H. (1992): Recursive and Sequential Tests of the Unit Root and Trend Break Down Hypothesis: Theory and International Evidence, *Journal of Business & Economic Statistics*, 10, 271–87.

Barndorff-Nielsen, O.E. (1983): On a Formula for the Distribution of the Maximum Likelihood Estimator, *Biometrika*, 70, 343–65.

---- (1988): *Parametric Statistical Models and Likelihood*, Lecture Notes in Statistics, Springer-Verlag, Heidelberg.

---- (1991): Likelihood Theory, in Hinkley, D.V., Reid, N. and Snell, E.J. (eds.): *Statistical Theory and Modelling — In Honour of Sir David Cox*, FRS, St. Edmundsbury Press, Suffolk.

Barndorff-Nielsen, O.E. and McCullagh, P. (1993): A Note on the Relation between Modified Profile Likelihood and the Cox-Reid Adjusted Profile Likelihood, *Biometrika*, 80, 321–8.

Bartlett, M.S. (1937): Properties of Sufficiency and Statistical Tests, *Proceedings of the Royal Society (London)*, A, 160, 268–82.

---- (1954): A Note on the Multiplying Factors for Various χ^2 Approximations, *Journal of the Royal Statistical Society*, B, 16, 296–8.

Beaudry, P. and Koop, G. (1993): Do Recessions Permanently Change Output?, *Journal of Monetary Economics*, 31, 149–163.

Ben-David, D. and Rahman, A.K.M.A. (1996): Technological Convergence and International Trade, Discussion Paper, 1359, Centre for Economy Policy Research.

Bernard, A.B. and Durlauf, S.N. (1996): Interpreting Tests of the Convergence Hypothesis, *Journal of Econometrics*, 71, 161–73.

Bernard, A.B. and Jones, C.I. (1993): Productivity across Industries and Countries: Time Series Theory and Evidence, Massachusetts Institute of Technology, Working Paper, 93-17.

Beveridge, S. and Nelson, C.R. (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, 151–74.

Bhargava, A. (1986): On the Theory of Testing for Unit Roots in Observed Time Series, *Review of Economic Studies*, 52, 369–84.

Bierens, H.J. (1993): Higher Order Sample Autocorrelations and the Unit Root Hypothesis, *Journal of Econometrics*, 57, 137–60.

Blanchard, O.J. (1981): What Is Left of the Multiplier Accelerator?, *American Economic Review*, Papers and Proceedings, 150–154.

Blanchard, O.J. and Summers, L.H. (1986): Hysteresis and the European Unemployment Problem, *NBER Macroeconomics Annual*, MIT Press, Cambridge. Reprinted with revisions in Summers, L.H. (1990): *Understanding Unemployment*, MIT Press, Cambridge.

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.

Boswijk, P. and Franses, P.H. (1992): Dynamic Specification and Cointegration, *Oxford Bulletin of Economics and Statistics*, 54, 369–381.

Boswijk, P. and Franses, P.H. (1996): Unit Roots in Periodic Autoregressions, *Journal of Time Series Analysis*, 17, 221–45.

Box, G.E.P. and Jenkins, G.M. (1970, 1976, 2nd ed.): *Time Series Analysis — Forecasting and Control*, Holden-Day, Oakland, California.

Buse, A. (1982): The Likelihood, Wald, and Lagrange Multiplier Tests: An Expository Note, *The American Statistician*, 36, 153–7.

Campbell, J.Y. and Mankiw, N.G. (1987): Are Output Fluctuations Transitory?, *The Quarterly Journal of Economics*, CII, 857–80.

Campbell, J.Y. and Mankiw, N.G. (1989): International Evidence on the Persistence of Economic Fluctuations, *Journal of Monetary Economics*, 23, 319–33.

Campbell, J.Y. and Perron, P. (1991): Pitfalls and Opportunities: What Macroeconomists Should Know about Unit Roots, in Blanchard, O.J. and Fischer, S. (eds.): *NBER Macroeconomics Annual*, 141–201.

Caporale, G.M. and Pittis, N. (1993): Common Stochastic Trends and Inflation Convergence in the EMS, *Weltwirtschaftliches Archiv*, 129, 207–15.

Chen, C. and Tiao, G.C. (1990): Random Level-Shift Time Series Models, ARIMA Approximations, and Level Shift Detection, *Journal of Business and Economics Statistics*, 8, 83–97.

Cheung, Y-W. (1993): Long-Memory in Foreign-Exchange Rates, *Journal of Business and Economics Statistics*, 11, 93–102.

---- (1994): Aggregate Output Dynamics in the Twentieth Century, *Economics Letters*, 45, 15–22.

Cheung, Y-W. and Lai, K.S. (1993): A Fractional Cointegration Analysis of Purchasing Power Parity, *Journal of Business and Economics Statistics*, 11, 103–112.

---- (1995a): Estimating Finite Sample Critical Values for Unit Root Tests Using Pure Random Walk Processes: A Note, *Journal of Time Series Analysis*, 16, 493–8.

---- (1995b): Lag Order and Critical Values of the Augmented Dickey–Fuller Test, *Journal of Business and Economics Statistics*, 13, 277-80.

Chiang, A.C. (1984): *Fundamental Methods of Mathematical Economics*, 3rd ed., McGraw-Hill, Tokyo.

Choi, I. and Phillips, P.C.B. (1993): Testing for a Unit Root by Frequency Domain Regression, *Journal of Econometrics*, 59, 263–286.

Christensen, B.J. and Kiefer, N.M. (1993): Local Cuts and Separate Inference, *Scandinavian Journal of Statistics*, 21, 389–401.

Christiano, L.J. (1992): Searching for a Break in GNP, *Journal of Business & Economic Statistics*, 10, 237–50.

Christiano, L.J. and Eichenbaum, M. (1990): Unit Roots in Real GNP: Do We Know, And Do We Care?, *Carnegie-Rochester Conference Series on Public Policy*, 32, 7–62.

Clements, M.P. and Mizon, G.E. (1991): Empirical Analysis of Macroeconomic Time Series — VAR and Structural Models, *European Economic Review*, 35, 887–932.

Cochrane, J.H. (1991a): A Critique of the Application of Unit Root Tests, *Journal Of Economic Dynamics and Control*, 15, 275–84.

---- (1991b): Comment, in Blanchard, O.J. and Fischer, S. (eds.): *NBER Macroeconomics Annual*, 201–10.

Copas, J.B. (1966): Monte Carlo Results for Estimation in a Stable Markov Time Series, *Journal of the Royal Statistical Society*, A, 129, 110–6.

Copeland, L.S. (1991): Cointegration Tests with Daily Exchange Rate Data, *Oxford Bulletin of Economics And Statistics*, 53, 185–98.

Cox, D.D. and Llatas, I. (1991): Maximum Likelihood Type Estimation for Nearly Nonstationary Autoregressive Time Series, *The Annals of Statistics*, 19, 1109–28.

Cox, D.R. and Reid, N. (1987): Parameter Orthogonality and Approximate Conditional Inference, *Journal of the Royal Statistical Society*, B, 49, 1–39.

---- (1989): On the Stability of Maximum-Likelihood Estimators of Orthogonal Parameters, *Canadian Journal of Statistics*, 17, 229–33.

---- (1992): A Note on the Difference between Profile and Modified Profile Likelihood, *Biometrika*, 79, 408–11.

---- (1993): A Note on the Calculation of Adjusted Profile Likelihood, *Journal of the Royal Statistical Society*, B, 55, 467–71.

Cramer, H. (1946): *Mathematical Methods of Statistics*, Almqvist and Wicksell.

Crato, N. and de Lima, P.J.F. (1994): Long-Range Dependence in the Conditional Variance of

Stock Returns, *Economics Letters*, 45, 281–85.

Crato, N. and Rothman, P. (1994): Fractional Integration Analysis of Long-Run Behaviour for US Macroeconomic Time Series, *Economics Letters*, 45, 287–91.

Cribari-Neto, F. (1996): On Time-Series Econometrics, forthcoming in *Quarterly Review of Economics and Finance* (also available from http://www.siu.edu/ econ/fc-publications.html).

Cribari-Neto, F. and Cordeiro, G.M. (1997): On Bartlett and Bartlett-Type Corrections, forthcoming in *Econometric Reviews* (also available from http://www.siu.edu/ econ/fc-publications.html).

Cruddas, A.M., Reid, N. and Cox, D.R. (1989): A Time Series Illustration of Approximate Conditional Likelihood, *Biometrika*, 76, 231–7.

Davidson, J., Hendry, D.F., Srba, F. and Yeo, S. (1978): Econometric Modelling of the Aggregate Time-Series Relationship between Consumer's Expenditure and Income in the United Kingdom, *Economic Journal*, 88, 661–92.

Davidson, R. and MacKinnon, J. G. (1993): *Estimation and Inference in Econometrics*, Oxford University Press, New York.

Deaton, A. (1986): Life-Cycle Models of Consumption: Is the Evidence Consistent with the Theory?, NBER Working Paper Series, 1910, published later in Bewley, T.F (1987, ed.): *Advances in Econometrics, Fifth World Congress*, Vol. 2, Cambridge University Press, Cambridge.

---- (1992): *Understanding Consumption*, Clarendon Lectures in Economics, Clarendon Press, Oxford.

DeJong, D.N., Nankervis, J.C., Savin, N.E. and Whiteman, C.H. (1992): The Power Problems of Unit Root Tests in Time Series with Autoregressive Errors, *Journal of Econometrics*, 53, 323–43.

De Long, J.B. and Lang, K. (1992): Are All Economic Hypothesis False?, *Journal of Political Economy*, 100, 1257–72.

Dickey, D.A. (1976): Estimation and Hypothesis Testing in Nonstationary Time Series, Doctor of Philosophy Thesis, Iowa State University (unpublished).

Dickey, D.A., Bell, W.R. and Miller, R.B. (1986): Unit Roots in Time Series Models: Tests and Implications, *The American Statistician*, 40, 12–26.

Dickey, D.A. and Fuller, W.A. (1979): Distribution of the Estimators for Autoregressive Time Series with a Unit Root, *Journal of the American Statistical Association*, 74, 427–31.

---- (1981): The Likelihood Ratio Statistics for Autoregressive Time Series with a Unit Root, *Econometrica*, 49, 1057–72.

Dickey, D.A. and Pantula, S.G. (1987): Determining the Order of Differencing in Autoregressive Processes, *Journal of Business & Economic Statistics*, 5, 455–61.

Diebold, F.X. (1988): Testing for Bubbles, Reflecting Barriers and Other Anomalies, *Journal of Economic Dynamics & Control*, 12, 63–70.

---- (1989): Random Walks versus Fractional Integration: Power Comparisons of Scalar and Joint Tests of the Variance Time Function, in Raj, B. (ed.): *Advances in Econometrics and Modelling*, 2945, Kluwer Academic Publishers, Needham, Mass.

Diebold, F.X. and Rudebusch, G.D. (1989): Long Memory and Persistence in Aggrerate Output, *Journal of Monetary Economics*, 24, 189–209.

---- (1991a): Is Consumption too Smooth? Long Memory and the Deaton Paradox, *The Review of Economics and Statistics*, 78, 1–9.

---- (1991b): On the Power of Dickey-Fuller Tests against Fractional Altarnatives, *Economics Letters*, 35, 155–60.

Dolado, J.J., Jenkinson, T., and Sosvilla-Rivero, S. (1990): Cointegration and Unit Roots,

Journal of Economic Surveys, 4, 249–73.

Dolado, J.J. and Maravall, A. (1989): Comment on Intertemporal Consumer Behaviour Under Structural Changes on Income, *Econometric Reviews*, 8, 101–9.

Duck, N.W. (1992): UK Evidence on Breaking Trend Functions, *Oxford Economic Papers*, 44, 426–39.

Dudewicz, E.J. and Mishra, S.N. (1988): *Modern Mathematical Statistics*, John Wiley, New York.

Edwards, A.W.F. (1974): The History of Likelihood, *International Statistical Review*, 42, 9–15. Reprinted in ____ (1992): *Likelihood*, Expanded ed., The Johns Hopkins University Press, Baltimore.

Elliot, G., Rothenberg, T.J. and Stock, J.H. (1996): Efficient Tests for an Autoregressive Unit Root, *Econometrica.* 4, 813-36.

Engel, C. and Hamilton, J.D. (1990): Long Swings in the Dollar: Are They in the Data and Do Markets Know It? *American Economic Review*, 80, 689–713.

Engle, R.F. and Granger, C.W.J. (1987): Co-Integration and Error Correction: Representation, Estimation, and Testing, *Econometrica*, 55, 251–76.

Engsted, T. (1993): Cointegration and Cagan's Model of Hyperinflation under Rational Expectations, *Journal of Money, Credit and Banking*, 25, 350–60.

Evans, G.B.A. and Savin, N.E. (1981): Testing for Unit Roots: 1, *Econometrica*, 49, 753–79.

____ (1984): Testing for Unit Roots: 2, *Econometrica*, 52, 1241–69.

Faust, J. (1992): When Are Variance Ratio Tests for Serial Dependence Optimal?, *Econometrica*, 60, 1215–26.

____ (1993): Near Observational Equivalence and Unit Root Processes: Formal Concepts and Implications, Board of Governors of the Federal Reserve System, manuscript.

____ (1994): Small and Large Sample Size Problems with Conventional Long-Run Inferences, Board of Governors of the Federal Reserve System, manuscript.

Ferguson, H., Reid, N. and Cox, D.R. (1991): Estimating Equations from Modified Profile Likelihood, in Godambe, V.P. (ed.): *Estimating Functions*, Clarendon Press, Oxford.

Franses, P.H. and Boswijk, H.P. (1991): Testing for Periodic Cointegration, Econometric Institute, University of Rotterdam, manuscript.

Franses, P.H. and Haldrup, N. (1994): The Effects of Additive Outliers on Tests for Unit Roots and Cointegration, *Journal of Business & Economic Statistics*, 12, 471–8.

Fuller, W.A. (1976): *Introduction to Statistical Time Series*, John Wiley, New York.

Gallant, A.R. and Fuller, W.A. (1973): Fitting Segmented Polynomial Regression Models Whose Join Points Have to Be Estimated, *Journal of the American Statistical Association*, 68, 144–47.

The GAUSS System Version 3.0 (1992), Aptech Systems, Washington.

Geweke, J. and Porter-Hudak, S. (1983): The Estimation and Application of Long Memory Time Series Models, *Journal of Time Series Analysis*, 4, 221–38.

Ghysels, E. (1990): Unit-Root Tests and the Statistical Pitfalls of Seasonal Adjustment: The Case of U.S. Postwar Real Gross National Product, *Journal of Business & Economic Statistics*, 8, 145–52.

Ghysels, E. (1993, ed.): Breakpoints and Unit Roots, Special Issue of *Journal of Business & Economic Statistics*, 10, 3.

Ghysels, E. and Perron, P. (1993): The Effect of Seasonal Adjustment Filters on Tests for a Unit Root, *Journal of Econometrics*, 55, 57–98.

Gouriéroux, C., Monfort, A., and Trognon, A. (1984): Pseudo Maximum Likelihood Methods:

Theory, *Econometrica*, 52, 681–720.

Granger, C.W.J. (1981): Some Properties of Time Series Data and Their Use in Econometric Model Specification, *Journal of Econometrics*, 16, 121–30.

---- (1986): Developments in the Study of Cointegrated Variables, *Oxford Bulletin of Economics and Statistics*, 48, 213–28.

Granger, C.W.J. and Joyeux, R. (1980): An Introduction to Long Memory Time Series Models and Fractional Differencing, *Journal of Time Series Analysis*, 1, 15–29.

Granger, C.W.J. and Morgenstern, O. (1970): *Predictability of Stock Market Prices*, Heath Lexington Books, Lexington, Massachusetts.

Granger, C.W.J. and Newbold, P. (1974): Spurious Regressions in Econometrics, *Journal of Econometrics*, 2, 111–20.

---- (1975): Economic Forecasting: The Atheist's Viewpoint, in Renton, G.A. (ed.): *Modelling the Economy*, Heinemann Educational Books, London, 131–47.

---- (1986): *Forecasting Economic Time Series*, 2nd ed., Academic Press, Orlando.

Granger, C.W.J. and Teräsvirta, T. (1993): *Modelling Nonlinear Economic Relationships*, Advanced Texts in Econometrics, Oxford University Press, Oxford.

Guilkey, D.K. and Schmidt, P. (1989): Extended Tabulations for Dickey–Fuller Tests, *Economics Letters*, 31, 355–57.

---- (1991): An Interpolation Test for a Unit Root in the Presence of Drift, *Biometrika*, 78, 809–15.

Haldrup, N. and Hylleberg, S. (1994): A Note on the Distribution of the Least Squares Estimator of a Random Walk with Drift: Some Analytical Evidence, *Economics Letters*, 48, 221–8.

Hall, R.E. (1978): Stochastic Implications of the Life Cycle - Permanent Income Hypothesis: Theory and Evidence, *Journal of Political Economy*, 86, 971–87.

Hall, S.G. (1986): An Application of the Granger & Engle Two-Step Estimation Procedure to United Kingdom Aggregate Wage Data, *Oxford Bulletin of Economics and Statistics*, 48, 229–39.

Hamilton, J.D. (1989): A New Approach to the Economic Analysis of Nonstationary Time Series and the Business Cycle, *Econometrica*, 57, 357–84.

---- (1994): *Time Series Analysis*, Princeton University Press, Princeton.

Harris, R.I.D. (1992): Small Sample Testing for Unit Roots, *Oxford Bulletin of Economics and Statistics*, 54, 615–25.

Harvey, A. C. (1981): *Time Series Models*, Philip Allan, Deddington.

---- (1990): *Forecasting, Structural Time Series Models and the Kalman Filter*, Cambridge University Press, Cambridge.

Hassler, U. and Wolters, J. (1994): On the Power of Unit Root Tests against Fractional Alternatives, *Economics Letters*, 45, 1–5.

Hasza, D.P. (1980): The Asymptotic Distribution of the Sample Autocorrelations for an Integrated ARMA Process, *Journal of the American Statistical Association*, 75, 349–52.

Haubrich, J.G. and Lo, A.W. (1989): The Sources and Nature of Long Term Memory in the Business Cycle, NBER Working Paper Series, 2951.

Haug, A.A. (1993): Residual Based Tests for Cointegration – A Monte Carlo Study of Size Distortions, *Economics Letters*, 41, 345–51.

Hayes, K., Slottje, D.J., Porter-Hudak, S. and Scully, G. (1990): Is the Size Distribution of Income a Random Walk?, *Journal of Econometrics*, 43, 213–226.

Heikkinen, S. and Maula, J. (1996): Finnish Food Consumption 1860-1993, National Consumer

Research Centre, Publications 1/1996.

Hendry, D.F. (1984): Monte Carlo Experimentation in Economics, Ch. 16 in Griliches, Z. and Intriligator, M.D. (eds.): *Handbook of Econometrics*, II, North-Holland, Amsterdam.

____ (1995): *Dynamic Econometrics*, Oxford University Press, Oxford.

Hendry, D.F. and Mizon, G.E. (1978): Serial Correlation as a Convenient Simplification Not a Nuisance: A Comment on a Study of the Demand for Money by the Bank of England, *Economic Journal*, 88, 549–63.

Hendry, D.F. and Neale, A.J. (1991): A Monte Carlo Study of the Impact of Structural Breaks on Tests for Unit Roots, in Hackl, P. and Westlund, A:H. (eds.): *Economic Structural Change, Analysis and Forecasting*, pp. 95–119, Springer-Verlag, Berlin.

Hendry, D.F., Neale, A.J. and Ericsson, N.R. (1990): *PC-NAIVE — An Interactive Program for Monte Carlo Experimentation in Econometrics*, Biddles, Guilford.

Henrekson, M. (1990): Wagner's Law — A Spurious Relationship?, Working Paper, 75, Fackföreningsrörelsens Institut för Ekonomisk Forskning.

Hooker, R.H. (1901): Correlation of the Marriage-Rate with Trade, *Journal of the Royal Statistical Society*, 64, 485–92. Reprinted in Hendry, D.F. and Morgan, M.S. (1995): The Foundations of Econometric Analysis, Cambridge University Press, Cambridge.

Hosking, J.R.M. (1981): Fractional Differencing, *Biometrika*, 68, 165–76.

Hurwicz, L. (1950): Least-Squares Bias in Time Series, in Koopmans, T.C. (ed.): *Statistical Inference in Dynamic Models*, Cowles Commission Monograph 10, John Wiley, New York.

Hylleberg, S. and Mizon, G.E. (1989): A Note on the Distribution of the Least Squares Estimator of a Random Walk with Drift, *Economics Letters*, 29, 225–230.

Jager, A. and Kunst, R.M. (1990): Seasonal Adjustment and Measuring Persistence in Output, *Journal of Applied Econometrics*, 5, 47–58.

Jorion, P. (1989): On Jump Processes in the Foreign Exchange and Stock Markets, *Review of Financial Studies*, 1, 427–37.

Judge, G.G. and Bock, M.E. (1983): Biased Estimation, in Griliches, Z. and Intriligator, M.D. (eds.): *Handbook of Econometrics*, Vol. 1, North-Holland, Amsterdam.

Juselius, K. (1991): On the Duality between Long Run Relations and Common Trends in an Empirical Analysis of Aggregate Money Holdings, Discussion Paper, 91–15, Institute of Economics, University of Copenhagen.

Kendall, M.G: (1954): Note on Bias in the Estimation of Autocorrelation, *Biometrika*, 41, 403–4.

Kiviet, J.F. and Phillips, G.D.A. (1992): Exact Similar Tests for Unit Roots and Cointegration, *Oxford Bulletin of Economics and Statistics*, 54, 349–67.

Knight, K. (1989): Limit Theory for Autoregressive-Parameter Estimates in an Infinite-Variance Random Walk, *Canadian Journal of Statistics*, 17, 261–78.

Koop, G. and Poirier, D.J. (1995): An Empirical Investigation of Wagner's Hypothesis by Using a Model Occurrence Framework, *Journal of the Royal Statistical Society*, A, 158, 123–41.

Koopmans, T.C. and Hood, Wm.C. (1953): The Estimation of Simultaneous Linear Economic Relationships, in Hood, Wm.C. and Koopmans, T.C. (eds.): *Studies in Econometric Method*, Cowles Foundation Monograph 14, 112-99, Yale University Press, New Haven.

Lam, P. (1990): The Hamilton Model with a General Autoregressive Component, *Journal of Monetary Economics*, 26, 409–32.

Layton, A.P. (1994): Further on the Nature of the Australian Business Cycle, *Economic Record*, 70, 12–18.

Lawley, D.N. (1956): A General Method for Approximating to the Distribution of Likelihood Ratio Criteria, *Biometrika*, 43, 295–303.

Lawrance, A.J. (1991): Directionality and Reversibility in Time Series, *International Statistical Review*, 59, 67–79.

Leitch, G. and Tanner, J.E. (1995): Professional Economic Forecasts: Are They Worth Their Costs?, *Journal of Forecasting*, 14, 143–57.

LeRoy, S.F. (1989): Efficient Capital Markets and Martingales, *Journal of Economic Literature*, 27, 1583–621.

Leyborne, S.J., McCabe, B.P.M., and Mills, T.C. (1996): Randomized Unit Root Processes for Modelling and Forecasting Financial Time Series: Theory and Applications, Journal of Forecasting, 15, 253–70.

Levin, A. and Lin, C. (1992): Unit Root Tests in Panel Data: Asymptotic and Finite-Sample Properties, Discussion Paper, 92-23, Department of Economics, University of California, San Diego.

Li, W.K. and McLeod, A.I. (1986): Fractional Time Series Modelling, *Biometrika*, 73, 217–21.

Lin, C-F. J. and Teräsvirta, T. (1994): Testing the Constancy of Regression Parameters against Continuous Structural Change, *Journal of Econometrics*, 62, 211–28.

Linden, M. (1995a): Studies in Integrated and Cointegrated Economic Time Series, Doctoral Thesis, Statistical Research Reports 15, The Finnish Statistical Society, Helsinki.

Linden, M. (1995b): Finnish GNP-Series 1954/I-1990/IV: Small Shock Persistence or Trend Stationarity? Some Evidence with Variance Ratio Estimates, forthcoming in *Empirical Economics.*

Lippi, M. and Reichlin, L. (1992): On Persistence of Shocks to Economic Variables — A Common Misconception, *Journal of Monetary Economics*, 29, 87–93.

Lo, A.W. and MacKinley, A.C. (1989): The Size and Power of the Variance Ratio Test in Finite Samples: A Monte Carlo Investigation, *Journal of Econometrics*, 40, 203–38.

Longbottom, J.A. and Holly, S. (1985): The Role of Time Series Analysis in the Evaluation of Econometric Models, *Journal of Forecasting*, 4, 75–87.

Lucas, A. (1996): Outlier Robust Unit Root Analysis, Ph.D. Thesis, Thesis Publishers, Amsterdam.

Luukkonen, R. and Saikkonen, P. (1996): Power of the Lagrange Multiplier Test for Testing an Autoregressive Root, *Economics Letters*, 51, 27–36.

McCullagh, P. and Tibshirani, R. (1990): A Simple Method for the Adjustment of Profile Likelihoods, *Journal of the Royal Statistical Society*, 52, B, 325–44.

MacKinnon, J.G. (1994): Approximate Asymptotic Distribution Functions for Unit-Root and Cointegration Tests, *Journal of Business & Economic Statistics*, 12, 167–76.

MacKinnon, J.G. and Smith, A.A. (1996): Approximate Bias Correction in Econometrics, manuscript (available from http://qed.econ.queensu.ca:80/pub/faculty/mackinnon/).

Maddala, G. (1977): *Econometrics*, McGraw-Hill, New York.

Mann, H.B. and Wald, A. (1943): On Stochastic Limit and Order Relationships, *Annals of Mathematical Statistics*, 14, 217–26.

Maravall, A. and Mathis, A. (1994): Encompassing Univariate Models in Multivariate Time Series, *Journal of Econometrics*, 61, 197–233.

Marriot, F.H.C. and Pope, J.A: (1954): Bias in the Estimation of Autocorrelations, *Biometrika*, 41, 390–402.

Maula, J. (1996): Differences in Food Consumption between Europe and USA 1961-1993 (in Finnish with an abstract in English), National Consumer Research Centre, Publications 2/1996.

Mills, T.C. (1990): *Time Series Techniques for Economists*, Cambridge University Press, Cambridge.

Minozzo, M. and Azzalini, A. (1993): On the Unimodality of the Exact Likelihood Function for Normal AR(2) Series, *Journal of Time Series Analysis*, 14, 497–509.

Miron, J.A. (1991): Comment, in Blanchard, O.J. and Fischer, S. (eds.): *NBER Macroeconomics Annual*, 211–8.

Mocan, H.N. (1994): Is There a Unit Root in U.S. Real GNP?, *Economics Letters*, 45, 23–31.

Molinas, C. (1986): A Note on Spurious Regressions with Integrated Moving Average Errors, *Oxford Bulletin of Economics and Statistics*, 48, 279–82.

Mustonen, S. (1992): *SURVO — An Integrated Environment for Statistical Computing and Related Areas*, Helsinki University Printing House, Helsinki.

Nabeya, S. and Tanaka, K. (1990a): A General Approach to the Limiting Distribution for Estimators in Time Series Regression with Nonstable Autoregressive Errors, *Econometrica*, 58, 145–63.

---- (1990b): Limiting Power of Unit-Root Tests in Time-Series Regression, *Journal of Econometrics*, 46, 247–71.

Nankervis, J.C. and Savin, N.E. (1985): Testing the Autoregressive Parameter with the t Statistic, *Journal of Econometrics*, 27, 143–61.

---- (1988): The Exact Moments of the Least-Squares Estimator for the Autoregressive Model — Corrections and Extensions, *Journal of Econometrics*, 37, 381–88.

Naylor, T.H., Seaks, T.G. and Wichern, D.W. (1972): Box-Jenkins Methods: An Alternative to Econometric Models, *International Statistical Review*, 40, 113–37.

Nelson, C.R. and Plosser, C.I. (1982): Trends and Random Walks in Macroeconomic Time Series — Some Evidence and Implications, *Journal of Monetary Economics*, 10, 139–62.

Newbold, P. (1992) and Agiakloglou, C. (1993): Bias in the Sample Autocorrelations of Fractional Noise, *Biometrika*, 80, 698–702.

Neyman, J. and Scott, E.L. (1948): Consistent Estimates Based on Partially Consistent Observations, *Econometrica*, 16, 1–32.

Ng, S. and Perron. P. (1995): Unit Root Tests in ARMA Models with Data Dependent Methods for the Selection of the Truncation Lag, *Journal of the American Statistical Association*, 90, 268–81.

Olekalns, N. (1994): Testing for Unit Roots in Seasonally Adjusted Data, *Economics Letters*, 45, 273–9.

Orcutt, G.H. (1948): A Study of the Autoregressive Nature of the Time Series Used for Tinbergen's Model of the Economic System of the United States, 1919–32 (with discussion), *Journal of the Royal Statistical Society*, B, 10, 1–53.

Orcutt, G.H. and Winokur, H.S. Jr. (1969): First Order Autoregression: Inference, Estimation, and Prediction, *Econometrica*, 37, 1–14.

Osborn, D.R., Chui, A.P.L., Smith, J.P. and Birchenhall, C.R. (1988): Seasonality and the Order of Integration for Consumption, *Oxford Bulletin of Economics and Statistics*, 50, 361–77.

Osborn, D.R. and Smith, J.P. (1989): The Performance of Periodic Autoregressive Models in Forecasting Seasonal U.K. Consumption, *Journal of Business & Economic Statistics*, 7, 117–127.

Pantula, S.G. (1991): Asymptotic Distributions of Unit-Root tests When the Process Is Nearly Stationary, *Journal of Business & Economic Statistics*, 9, 63–71.

Pantula, S.G., Gonzalez–Farias, G. and Fuller, W.A. (1994): A Comparison of Unit-Root Test Criteria, *Journal of Business & Economic Statistics*, 12, 449–59.

Pantula, S.G. and Hall, A. (1991): Testing for Unit Roots in Autoregressive Moving Average Models — An Instrumental Variable Approach, *Journal of Econometrics*, 48, 325–53.

Parzen, E. (1982): ARARMA Models for Time Series Analysis and Forecasting, *Journal of Forecasting*, 1, 67–82.

Pearson, K. (1897): Mathematical Contributions to the Theory of Evaluation — On a Form of Spurious Correlation Which May Arise When Indices Are Used in the Measurement of Organs, *Proceedings of the Royal Society*, 60, 489–98.

Peiris, M.S. and Perera, B.J.C. (1988): On Prediction with Fractionally Differenced ARIMA Models, *Journal of Time Series Analysis*, 9, 215–20.

Perron, P. (1988): Trends and Random Walks in Macroeconomic Time Series — Further Evidence from a New Approach, *Journal of Economic Dynamics & Control*, 12, 297–332.

---- (1989): The Great Crash, the Oil Price Shock, and the Unit Root Hypothesis, *Econometrica*, 57, 1361–401.

---- (1990): Testing for a Unit Root in a Time Series with a Changing Mean, *Journal of Business & Economic Statistics*, 8, 153–62.

---- (1991): A Continuous Time Approximation to the Unstable First-Order Autoregressive Process: The Case without an Intercept, *Econometrica*, 59, 211–36.

Perron, P. and Vogelsang, T.J. (1991): Nonstationarity and Level Shifts with an Application to Purchasing Power Parity, Econometric Research Program, Research Memorandum No. 359, Princeton University.

Phillips, A.W. (1954): Stabilisation Policy in a Closed Economy, *Economic Journal*, 64, 290–323.

---- (1957): Stabilisation Policy and the Time Form of Lagged Responses, *Economic Journal*, 67, 265–77.

Phillips, P.C.B. (1977): Approximations to Some Finite Sample Distributions Associated with a First-Order Stochastic Difference Equation, *Econometrica*, 45, 463–85.

---- (1978): Edgeworth and Saddlepoint Approximations in the First-Order Noncircular Autoregression, *Biometrika*, 65, 91–8.

---- (1986): Understanding Spurious Regressions in Econometrics, *Journal of Econometrics*, 33, 311–40.

---- (1987): Time Series Regression with a Unit Root, *Econometrica*, 55, 277–301.

---- (1991a): To Criticize the Critics: An Objective Bayesian Analysis of Stochastic Trends, Themed Issue (Classical and Bayesian Methods of Testing for Unit Roots) of *Journal of Applied Econometrics*, 6, 333–64.

---- (1991b): Bayesian Routes and Unit Roots: De Rebus Prioribus Semper Est Disputandum, Themed Issue (Classical and Bayesian Methods of Testing for Unit Roots) of *Journal of Applied Econometrics*, 6, 435–73.

---- (1995): Nonstationary Time Series and Cointegration, *Journal of Applied Econometrics*, 10, 87–94.

Phillips, P.C.B. and van Dijk, H.K. (1994): Bayes Methods and Unit Roots — Editors' Introduction, Symposium Double Issue (Bayes Methods and Unit Roots) of *Econometric Theory*, 10, 453–60.

Phillips, P.C.B. and Perron, P. (1988): Testing for a Unit Root in Time Series Regression, *Biometrika*, 75, 335–46.

Pippenger, M.K. and Goering, G.E. (1993): A Note on the Empirical Power of Unit Root Tests under Threshold Processes, *Oxford Bulletin of Economics and Statistics*, 55, 4, 473–81.

Porter-Hudak, S. (1990): An Application of the Seasonal Fractionally Differenced Model to the Monetary Aggregates, *Journal of the American Statistical Association*, 85, 338–44.

Quah, D. (1993a): Galton's Fallacy and Tests of the Convergence Hypothesis, *Scandinavian Journal of Economics*, 95, 427–43.

Quah, D. (1993b): Empirical Cross-Section Dynamics in Economic Growth, *European Economic Review*, 37, 426–34.

Quandt, R.E: (1983): Computational Problems and Methods, in Griliches, Z. and Intriligator, M.D. (eds.): *Handbook of Econometrics*, Vol. 1, North-Holland, Amsterdam.

Quenouille, M.H. (1949): Approximate Tests of Correlation in Time Series, *Journal of the Royal Statistical Society*, B, 68–83.

Rappoport, P. and Reichlin, L. (1989): Segmented Trends and Non-Stationary Time Series, *Economic Journal*, 99, 168–77.

Reichlin, L. (1989): Structural Change and Unit Root Econometrics, *Economics Letters*, 31, 231–33.

Reid, N. (1992): Aspects of Modified Profile Likelihood, in Saleh, A.K.Md.E. (ed.): *Nonparametric Statistics and Related Topics*, Elsevier Science Publishers.

Rothenberg, T. (1995): Some Distribution Theory for an Autoregression Fitted to a Random Walk, a paper given on the 20th of January, 1995 at the Research Seminar on Economic Theory and Econometrics at Nuffield College, University of Oxford.

Rubin, H. (1950): Consistency of Maximum-Likelihood Estimates in the Explosive Case, in Koopmans, T.C. (ed.): *Statistical Inference in Dynamic Economic Models*, Cowles Commission Monograph 10, John Wiley, New York.

Rudebusch, G.D. (1992): The Uncertain Unit Root in Real GNP, *American Economic Review*, 83, 264–72.

Said, E.S. and Dickey, D.A. (1984): Testing for Unit Roots in Autoregressive-Moving Average Models of Unknown Order, *Biometrika*, 71, 599–607.

Saikkonen, P. (1993): A Note on a Lagrange Multiplier Test for Testing an Autoregressive Unit Root, *Econometric Theory*, 9, 494–98.

---- (1995): Problems with the Asymptotic Theory of Maximum Likelihood Estimation in Integrated and Cointegrated Systems, *Econometric Theory*, 11, 888–911.

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. (1964): Wages and Prices in the United Kingdom: A Study in Econometric Methodology, in Hart, P.E., Millsand, G. and Whitaker, J.K. (eds.): *Econometric Analysis for National Economic Planning*, Butterworths, London. Reprinted in Hendry, D.F. and Wallis, K.F. (eds., 1984): *Econometrics and Quantitative Economics*, Basil Blackwell, Padstow.

Schmidt, P. and Phillips, P.C.B. (1992): LM Tests for a Unit Root in the Presence of Deterministic Time Trends, *Oxford Bulletin of Economics and Statistics*, 54, 257–87.

Schwert, G.W. (1987): Effects of Model Specification on Tests for Unit Roots in Macroeconomic Data, *Journal Of Monetary Economics*, 20, 73–103.

---- (1989): Tests for Unit Roots: A Monte Carlo Investigation, *Journal of Business & Economic Statistics*, 7, 147–59.

Seber, G.A.F. and Wild, C.J. (1989): Nonlinear Regression, John Wiley, New York.

Serletis, A. (1992): The Random Walk in Canadian Output, *Canadian Journal of Economics*, XXV, 393–406.

Shaman, P. and Stine, R.A. (1988): The Bias of Autoregressive Coefficient Estimators, *Journal of the American Statistical Association*, 83, 842–8.

Shenton, L.R. and Johnson, W.L. (1965): Moments of a Serial Correlation Coefficient, *Journal of the Royal Statistical Society*, B, 27, 308–20.

Sims, C.A. (1988): Bayesian Scepticism on Unit Root Econometrics, *Journal of Economic*

Dynamics and Control, 12, 463–74.

---- (1991): Comment by Christopher A. Sims on 'To Criticize the Critics', by Peter C.B. Phillips, Themed Issue (Classical and Bayesian Methods of Testing for Unit Roots) of *Journal of Applied Econometrics*, 6, 423–34.

Sims, C.A. and Uhlig, H. (1991): Understanding Unit Rooters: A Helicopter Tour, *Econometrica*, 59, 1591–600.

Solo, V. (1984): The Order of Differencing in ARIMA Models, *Journal of the American Statistical Association*, 79, 916–21.

Sowell, F. (1990): The Fractional Unit Root Distribution, *Econometrica*, 58, 495–505.

---- (1992a): Maximum Likelihood Estimation of Stationary Univariate Fractionally Integrated Time Series Models, *Journal of Econometrics*, 53, 165–88.

---- (1992b): Modeling Long-Run Behaviour with the Fractional ARIMA Model, *Journal of Monetary Economics*, 29, 277–302.

Stine, R. (1992): Time Series Models and Mathematica, in Varian, H. (ed.): *Economic and Financial Modeling with Mathematica*, Springer-Verlag, New York.

Stine, R.A. and Shaman, P. (1989): A Fixed Point Characterization for Bias of Autoregressive Estimators, *The Annals of Statistics*, 17, 1275–84.

Stock, J.H. (1987): Asymptotic Properties of Least Squares Estimators of Cointegrating Vectors, *Econometrica*, 55, 1035–56.

---- (1990): A Comment (On Christiano and Eichenbaum (1990)), *Carnegie-Rochester Conference Series on Public Policy*, 32, 63–82.

---- (1994): Unit Roots, Structural Breaks and Trends, in Griliches, Z. and Intriligator, M.D. (eds.): *Handbook of Econometrics*, Vol. 4, North-Holland, Amsterdam.

Stuart, A. and Ord, K. (1991): *Kendall's Advanced Theory of Statistics*, Vol. 2, 5th ed., Edward Arnold, London.

Taylor, M.P. (1991): The Hyperinflation Model of Money Demand Revisited, *Journal of Money, Credit and Banking*, 23, 327–51.

Teräsvirta, T. and Anderson, H.M. (1993): Characterizing Nonlinearities in Business Cycles Using Smooth Transition Autoregressive Models, in Pesaran, M.H. and Potter, S.M. (eds.): *Nonlinear Dynamics, Chaos and Econometrics*, John Wiley & Sons, Chichester.

Tiao, G.C. and Grupe, M.R. (1980): Hidden Periodic Autoregressive Moving Average Time Series Models in Time Series Data, *Biometrika*, 67, 365–73.

Tjøstheim, D. and Paulsen, J. (1983): Bias of Some Commonly-Used Time Series Estimates, *Biometrika*, 70, 389–99. Corrigendum (1984): *Biometrika*, 71, 656.

Tuan, D.P. (1992): Approximate Distribution of Parameter Estimators for First-Order Autoregressive Models, *Journal of Time Series Analysis*, 13, 147–70.

Vecchia, A.V. (1985): Maximum Likelihood Estimation for Periodic Autoregressive Moving Average Models, *Technometrics*, 27, 375–84.

Vecchia, A.V. and Ballerini, R. (1991): Testing for Periodic Autocorrelations in Seasonal Time Series Data, *Biometrika*, 78, 53–63.

West, K.D. (1987): A Note on the Power of Least Squares Tests for a Unit Root, *Economics Letters*, 24, 249–52.

---- (1988): On the Interpretation of Near Random-Walk Behaviour in GNP, *American Economic Review*, 78, 202–9.

---- (1988b): Asymptotic Normality, When Regressors Have a Unit Root, *Econometrica*, 56, 1397–1417.

White, H. (1982): Maximum Likelihood Estimation of Misspecified Models, *Econometrica*, 50,

1–25.

White, J.S. (1958): The Limiting Distribution of the Serial Correlation Coefficient in the Explosive Case, *Annals of Mathematical Statistics*, 29, 1188–97.

---- (1961): Asymptotic Expansions for the Mean and Variance of the Serial Correlation Coefficient, *Biometrika*, 48, 85–94.

Wold, H. (1938, 1954 2nd ed.): *A Study in the Analysis of Stationary Time Series*, Almqvist and Wicksell, Uppsala.

Wolfram, S. (1991): *Mathematica — A System for Doing Mathematics by Computer*, 2nd ed., Addison Wesley, Redwood City.

Working, H. (1934): A Random-Difference Series for Use in the Analysis of Time Series, *Journal of the American Statistical Association*, 29, 11–24. Reprinted (data cut) in Hendry, D.F. and Morgan, M.S. (1995): The Foundations of Econometric Analysis, Cambridge University Press, Cambridge.

---- (1949): The Investigation of Economic Expectations (with discussion), *American Economic Review*, 39, 150–66.

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. Reprinted in Stuart, A. and Kendall, M.G. (eds., 1971): *Statistical Papers of George Udny Yule*, Charles Griffin & Co., London, and (in part) in Hendry, D.F. and Morgan, M.S. (1995): The Foundations of Econometric Analysis, Cambridge University Press, Cambridge.

---- (1927): On a Method of Investigating Periodicities in Disturbed Series, with Special Reference to Wolfer's Sunspot Numbers, *Philosophical Transactions of the Royal Society of London*, A, 226, 267–98. Reprinted in Stuart, A. and Kendall, M.G. (eds., 1971): *Statistical Papers of George Udny Yule*, Charles Griffin & Co., London and (in part) in Hendry, D.F. and Morgan, M.S. (1995): The Foundations of Econometric Analysis, Cambridge University Press, Cambridge.

Zivot, E. (1994): A Bayesian Analysis of the Unit Root Hypothesis within an Unobserved Components Model, *Econometric Theory*, 10, 552–78.