

Bootstrap Tests for Regression Models

Palgrave Texts in Econometrics

General Editors: Kerry Patterson

Titles include:

Simon P. Burke and John Hunter

MODELLING NON-STATIONARY TIME SERIES

Michael P. Clements

EVALUATING ECONOMETRIC FORECASTS OF ECONOMIC AND
FINANCIAL VARIABLES

Leslie Godfrey

BOOTSTRAP TESTS FOR REGRESSION MODELS

Terence C. Mills

MODELLING TRENDS AND CYCLES IN ECONOMETRIC TIME SERIES

Palgrave Texts in Econometrics

Series Standing Order ISBN 978-1-4039-0172-9 (hardcover)

Series Standing Order ISBN 978-1-4039-0173-6 (paperback)

(outside North America only)

You can receive future titles in this series as they are published by placing a standing order. Please contact your bookseller or, in case of difficulty, write to us at the address below with your name and address, the title of the series and the ISBN quoted above.

Customer Services Department, Macmillan Distribution Ltd, Houndmills, Basingstoke, Hampshire RG21 6XS, England

Bootstrap Tests for Regression Models

Leslie Godfrey

palgrave
macmillan



© Leslie Godfrey 2009

All rights reserved. No reproduction, copy or transmission of this publication may be made without written permission.

No paragraph of this publication may be reproduced, copied or transmitted save with written permission or in accordance with the provisions of the Copyright, Designs and Patents Act 1988, or under the terms of any licence permitting limited copying issued by the Copyright Licensing Agency, Saffron House, 6–10 Kirby Street, London EC1N 8TS.

Any person who does any unauthorised act in relation to this publication may be liable to criminal prosecution and civil claims for damages.

The author has asserted his right to be identified as the author of this work in accordance with the Copyright, Designs and Patents Act 1988.

First published in 2009 by
PALGRAVE MACMILLAN

PALGRAVE MACMILLAN in the UK is an imprint of Macmillan Publishers Limited, registered in England, company number 785998, of Houndmills, Basingstoke, Hampshire RG21 6XS.

PALGRAVE MACMILLAN in the US is a division of St Martin's Press LLC, 175 Fifth Avenue, New York, NY 10010.

PALGRAVE MACMILLAN is the global academic imprint of the above companies and has companies and representatives throughout the world.

Palgrave® and Macmillan® are registered trademarks in the United States, the United Kingdom, Europe and other countries.

ISBN-13: 978-0-230-20230-6 hardback

ISBN-10: 0-230-20230-6 hardback

ISBN-13: 978-0-230-20231-3 paperback

ISBN-10: 0-230-20231-4 paperback

This book is printed on paper suitable for recycling and made from fully managed and sustained forest sources. Logging, pulping and manufacturing processes are expected to conform to the environmental regulations of the country of origin.

A catalogue record for this book is available from the British Library.

A catalog record for this book is available from the Library of Congress.

10 9 8 7 6 5 4 3 2 1
18 17 16 15 14 13 12 11 10 09

Printed and bound in Great Britain by
CPI Antony Rowe, Chippenham and Eastbourne

To Ben

This page intentionally left blank

Contents

<i>Preface</i>	xi
1 Tests for Linear Regression Models	1
1.1. Introduction	1
1.2. Tests for the classical linear regression model	3
1.3. Tests for linear regression models under weaker assumptions: random regressors and non-Normal IID errors	10
1.4. Tests for generalized linear regression models	14
1.4.1. HCCME-based tests	18
1.4.2. HAC-based tests	21
1.5. Finite-sample properties of asymptotic tests	25
1.5.1. Testing the significance of a subset of regressors	27
1.5.2. Testing for non-Normality of the errors	31
1.5.3. Using heteroskedasticity-robust tests of significance	33
1.6. Non-standard tests for linear regression models	35
1.7. Summary and concluding remarks	42
2 Simulation-based Tests: Basic Ideas	44
2.1. Introduction	44
2.2. Some key concepts and simple examples of tests for IID variables	46
2.2.1. Monte Carlo tests	47
2.2.2. Bootstrap tests	50
2.3. Simulation-based tests for regression models	55
2.3.1. The classical Normal model	55
2.3.2. Models with IID errors from an unspecified distribution	59
2.3.3. Dynamic regression models and bootstrap schemes	64
2.3.4. The choice of the number of artificial samples	67
2.4. Asymptotic properties of bootstrap tests	69
2.5. The double bootstrap	72
2.6. Summary and concluding remarks	77

3	Simulation-based Tests for Regression Models with IID Errors: Some Standard Cases	81
3.1.	Introduction	81
3.2.	A Monte Carlo test of the assumption of Normality	83
3.3.	Simulation-based tests for heteroskedasticity	88
3.3.1.	Monte Carlo tests for heteroskedasticity	91
3.3.2.	Bootstrap tests for heteroskedasticity	94
3.3.3.	Simulation experiments and tests for heteroskedasticity	95
3.4.	Bootstrapping F tests of linear coefficient restrictions	101
3.4.1.	Regression models with strictly exogenous regressors	101
3.4.2.	Stable dynamic regression models	109
3.4.3.	Some simulation evidence concerning asymptotic and bootstrap F tests	110
3.5.	Bootstrapping LM tests for serial correlation in dynamic regression models	118
3.5.1.	Restricted or unrestricted estimates as parameters of bootstrap worlds	119
3.5.2.	Some simulation evidence on the choice between restricted and unrestricted estimates	123
3.6.	Summary and concluding remarks	132
4	Simulation-based Tests for Regression Models with IID Errors: Some Non-standard Cases	134
4.1.	Introduction	134
4.2.	Bootstrapping predictive tests	136
4.2.1.	Asymptotic analysis for predictive test statistics	136
4.2.2.	Single and double bootstraps for predictive tests	139
4.2.3.	Simulation experiments and results	144
4.2.4.	Dynamic regression models	148
4.3.	Using bootstrap methods with a battery of OLS diagnostic tests	149
4.3.1.	Regression models and diagnostic tests	151
4.3.2.	Bootstrapping the minimum p-value of several diagnostic test statistics	152
4.3.3.	Simulation experiments and results	155
4.4.	Bootstrapping tests for structural breaks	160
4.4.1.	Testing constant coefficients against an alternative with an unknown breakpoint	162

4.4.2.	Simulation evidence for asymptotic and bootstrap tests	166
4.5.	Summary and conclusions	173
5	Bootstrap Methods for Regression Models with Non-IID Errors	177
5.1.	Introduction	177
5.2.	Bootstrap methods for independent heteroskedastic errors	178
5.2.1.	Model-based bootstraps	181
5.2.2.	Pairs bootstraps	183
5.2.3.	Wild bootstraps	185
5.2.4.	Estimating function bootstraps	188
5.2.5.	Bootstrapping dynamic regression models	190
5.3.	Bootstrap methods for homoskedastic autocorrelated errors	193
5.3.1.	Model-based bootstraps	194
5.3.2.	Block bootstraps	198
5.3.3.	Sieve bootstraps	201
5.3.4.	Other methods	205
5.4.	Bootstrap methods for heteroskedastic autocorrelated errors	207
5.4.1.	Asymptotic theory tests	207
5.4.2.	Block bootstraps	210
5.4.3.	Other methods	213
5.5.	Summary and concluding remarks	214
6	Simulation-based Tests for Regression Models with Non-IID Errors	218
6.1.	Introduction	218
6.2.	Bootstrapping heteroskedasticity-robust regression specification error tests	221
6.2.1.	The forms of test statistics	221
6.2.2.	Simulation experiments	226
6.3.	Bootstrapping heteroskedasticity-robust autocorrelation tests for dynamic models	231
6.3.1.	The forms of test statistics	232
6.3.2.	Simulation experiments	235
6.4.	Bootstrapping heteroskedasticity-robust structural break tests with an unknown breakpoint	241

6.5.	Bootstrapping autocorrelation-robust Hausman tests	247
6.5.1.	The forms of test statistics	247
6.5.2.	Simulation experiments	254
6.6.	Summary and conclusions	262
7	Simulation-based Tests for Non-nested Regression Models	266
7.1.	Introduction	266
7.2.	Asymptotic tests for models with non-nested regressors	268
7.2.1.	Cox-type LLR tests	269
7.2.2.	Artificial regression tests	273
7.2.3.	Comprehensive model F -test	274
7.2.4.	Regularity conditions and orthogonal regressors	274
7.2.5.	Testing with multiple alternatives	275
7.2.6.	Tests for model selection	277
7.2.7.	Evidence from simulation experiments	279
7.3.	Bootstrapping tests for models with non-nested regressors	281
7.3.1.	One non-nested alternative regression model: significance levels	281
7.3.2.	One non-nested alternative regression model: power	289
7.3.3.	One non-nested alternative regression model: extreme cases	290
7.3.4.	Two non-nested alternative regression models: significance levels	293
7.3.5.	Two non-nested alternative regression models: power	295
7.4.	Bootstrapping the LLR statistic with non-nested models	297
7.5.	Summary and concluding remarks	300
8	Epilogue	303
	<i>Bibliography</i>	305
	<i>Author Index</i>	319
	<i>Subject Index</i>	323

Preface

My wife is past-President of the Society for the Study of Addiction, but I suspect that even she finds it difficult to understand why I have not been able to free myself from an obsession with tests for econometric models in the last 30 years. My only defence is that I hoped that these tests would be useful to applied workers. Like many other researchers in the area, I had to make use of asymptotic theory when deriving tests. I now believe that the application of appropriate bootstrap techniques can greatly increase the usefulness of asymptotic test procedures at low cost and so I have a new obsession to combine with the old one.

Two types of problems associated with using asymptotic analysis to obtain tests are often mentioned. First, even when theory is tractable and leads to asymptotically valid critical values from standard distributions like Normal and Chi-Squared, which are convenient to use, there may be serious approximation errors in finite samples. In particular, the critical values implied by asymptotic theory may produce finite sample significance levels that are not close to the desired probabilities. Second, there are important test procedures for which asymptotic theory is intractable and does not provide a standard distribution from which critical values can be taken. The bootstrap has been used to tackle both types of problem. When a standard asymptotic test is available, the corresponding bootstrap test is often found to provide a better finite sample approximation and the improvement is sometimes remarkable. When no standard asymptotic test can be derived, the bootstrap can sometimes produce a test that is easy to carry out and has significance levels that are reasonably close to the desired values.

The bootstrap approach involves using computer programs to generate many samples from an artificial model that is intended to approximate the process assumed to generate the actual data. The values of test statistics calculated from these bootstrap samples can then be used to assess the statistical significance of the corresponding test statistic derived from the real observations. Given that many artificial samples are generated and each is subjected to the same statistical analysis as the genuine sample, there might be concerns about the computational costs of bootstrap tests. However, given the amazing increases in the power of personal computers, the real cost of the bootstrap approach is often very small in absolute terms, for example, the waiting time for results to appear. The

costs of bootstrapping are, therefore, often small and there is a great deal of evidence to suggest that the benefits can be very large.

The examples in this book that illustrate the value of the bootstrap and the dangers of relying upon asymptotically justified critical values are in the familiar framework of ordinary least squares (OLS) procedures for a linear regression model. The regression model is central to econometrics and its familiarity allows the reader to concentrate on the bootstrap techniques. The level of discussion is at an intermediate textbook standard and the aim has been to write a book that is useful to a fairly wide audience. However, references that cover more complicated models and more technical analyses of bootstrap procedures are provided.

Chapter 1 contains a discussion of regression models and OLS-based tests in order to summarize key results, to provide details of notation and to motivate going beyond conventional asymptotic theory as a basis for inference. The second chapter covers some basic ideas of simulation-based tests, with bootstrap procedures being given prominence but other approaches also being discussed. The application of simulation-based tests in regression models, under the assumption of independently and identically distributed (IID) errors, is examined in Chapters 3 and 4. The first of these two chapters covers test statistics that have standard asymptotic distributions, for example, Chi-Squared, when the null hypothesis is true. Chapter 4 is devoted to examples of situations of importance to empirical workers in which the bootstrap can be applied to statistics that, under the null hypothesis, have non-standard asymptotic distributions.

While the assumption that regression models have IID errors has often been made in the past when explaining results concerning the asymptotic properties of OLS estimators and test statistics, there has been a growing body of opinion that it is too restrictive. There are, of course, many ways in which data can be modelled using regression models with non-IID errors. The bootstrap world must mimic the process that is assumed to generate actual data under the null hypothesis. Consequently there is a need for bootstrap methods that allow for departures from the assumption of IID errors that are of interest to applied workers. Some of these methods are discussed in Chapter 5.

When the errors are not restricted to be IID, they can be assumed to be autocorrelated or heteroskedastic, according to precisely defined parametric models or in unspecified ways. The basic position taken in Chapter 5 is that there is rarely very clear guidance about the specification of parametric error models. There is, therefore, an emphasis on bootstrap methods that are designed to be asymptotically valid under unknown forms of autocorrelation and/or heteroskedasticity. Some examples of

the applications of these methods are examined in Chapter 6, which contains results on the finite sample behaviour of autocorrelation-robust and heteroskedasticity-robust bootstrap tests.

All of the tests discussed in Chapters 1 to 6 are based upon the assumption that the null-hypothesis model is a special case of the alternative-hypothesis model, that is, the former is nested in the latter. This assumption is required for much of the standard asymptotic theory of testing statistical hypotheses. However, competing specifications of linear regression models in applied econometric work are sometimes not nested and there is a considerable literature on tests for non-nested relationships. Chapter 7 contains a discussion of asymptotic and bootstrap tests for non-nested regression models. This discussion indicates how the bootstrap can help to overcome both of the above-mentioned general types of problem associated with reliance upon asymptotic theory when implementing tests of non-nested hypotheses. Finally, Chapter 8 contains an epilogue.

In the discussions of the application of bootstrap methods to OLS-based tests in regression analysis, I have used some examples from articles that I have written with various coauthors. I owe many debts to Chris Orme, Hashem Pesaran, Joao Santos Silva, Andy Tremayne and Mike Veall. It was a pleasure to work with these fine researchers and Mike Veall deserves special acknowledgment because he introduced me to the bootstrap during his first visit to York. I am very much indebted to Kerry Patterson, editor of this series, for his careful and constructive comments on my drafts. I am also grateful to Taiba Batool, commissioning editor at Palgrave Macmillan, for her encouragement and help, and to Alina Spiru for her assistance with the indexes. Finally, my thanks go to Christine who probably never realized that marriage might lead to the burden of helping me to sort out my ideas about this book during our lunchtime walks around the York campus.

L. G. Godfrey

This page intentionally left blank

1

Tests for Linear Regression Models

1.1. Introduction

The linear regression model is often used to study economic relationships and is familiar from standard intermediate and introductory courses at the level of, for example, Greene (2008), Gujarati (2003) and Wooldridge (2006). In such courses, considerable emphasis is usually placed on the important topic of testing hypotheses about the values of the parameters of the model. The text-book tests for regression models are developed using very strong auxiliary assumptions that simplify teaching but are of limited relevance in practical situations. As a consequence, applied workers often have to replace procedures that are exactly valid in finite samples under strong assumptions by tests that are based on weaker assumptions but are only asymptotically valid.

It is also often necessary to rely upon asymptotic, rather than finite sample, results when carrying out tests for misspecification of a regression model. It is now commonplace for the results of estimation to be accompanied by checks of the assumptions required to validate standard empirical analysis. Even under the restrictive assumptions of the classical textbook model, many of these checks have to be carried out using critical values that are only asymptotically valid. When these assumptions are relaxed, there is an even greater need to use asymptotic theory.

The problem for the empirical researcher is that asymptotic theory sometimes provides a poor approximation to the actual distribution of test statistics; so that the use of asymptotic critical values may lead to misleading inferences. Moreover, there is a second type of problem associated with the standard approach to deriving asymptotically valid tests. In some situations of importance, this approach is not capable of providing a usable tool for the applied worker. This failure can occur with some

2 Bootstrap Tests for Regression Models

tests when classical assumptions are relaxed, or when several separate large sample tests are being applied.

The purpose of this book is to explain how computers and appropriate software can be combined to tackle these problems. More precisely, the use of procedures involving the simulation of artificial sets of data is examined and some important cases are discussed in detail. The various computationally intensive simulation techniques, collectively known as *bootstrap methods*, provide:

1. ways to improve the finite-sample performance of well-known and widely-used large sample tests for regression models; and
2. new tests that can be employed when conventional asymptotic theory does not lead to a test statistic that can be compared with critical values from some standard distribution.

The reason for believing that it is worth providing a concise, but quite extensive, account of bootstrap tests in regression analysis is that, in recent years, personal computers have become so powerful and relatively cheap that it is now feasible to implement bootstrap procedures as part of routine econometric analysis. Also the linear multiple regression model provides a very useful framework for introducing ideas that can be used in more complicated models that are of interest to applied workers, students and others who carry out empirical econometric analyses.

The emphasis is on practical applications of bootstrap methods in regression models. There are many excellent treatments of theoretical issues associated with the validity and properties of bootstrap techniques in quite general settings. References to such technical material will be provided and key results will be summarized.

This chapter is intended to give an outline of the various frameworks for which results about regression model tests are available and widely used. The foundations required for the detailed treatments contained in later chapters are provided, along with notation. More thorough coverage of tests for regression models, including numerical examples, can be found in many text books, for example, J. Davidson (2000, chs 2 and 3) and Greene (2008, ch 5). The discussion in Davidson and MacKinnon (2004, ch 4) links the statistical underpinnings of tests with the use of simulation methods and so is especially useful for the purpose of this book.

The important problem of testing linear restrictions in the classical Normal linear regression model is covered in Section 1.2, which includes much of the required notation. Section 1.2 provides key results that are exactly valid under the very strong assumptions of the textbook classical

model. It is argued that, despite their value in simplifying the teaching of econometric tests, these assumptions should not be regarded as suitable for practical applications. Section 1.3 contains comments on carrying out tests under weaker assumptions about the error terms and explanatory variables of the regression model. However, the analysis of Section 1.3 is based upon the assumptions of independence and homoskedasticity. In Section 1.4, tests that are asymptotically valid in the presence of autocorrelation and/or heteroskedasticity are described. The tests of linear restrictions that are covered in Sections 1.3 and 1.4 are only asymptotically valid. Applied workers have to use data sets with a finite number of observations and may be concerned about relying on results that only hold as the sample size tends to infinity. Some examples are provided in Section 1.5 that illustrate the problems of inadequate approximations derived from asymptotic theory. Section 1.6 contains examples of situations in which it is not possible to derive an asymptotic test that permits reference to a standard distribution to assess statistical significance. A summary and some concluding remarks are given in Section 1.7.

1.2. Tests for the classical linear regression model

As in many texts, the starting point is the *classical linear regression model*

$$y_i = \sum_{j=1}^k x_{ij}\beta_j + u_i, \quad (1.1)$$

in which: y_i is a typical observation on the dependent variable; the terms x_{i1}, \dots, x_{ik} are the nonrandom values of a typical observation on the k regressors; the unknown regression coefficients to be estimated are β_1, \dots, β_k ; and the unobservable errors, with typical term u_i , are independently and identically distributed (IID), each having the Normal distribution with zero mean and variance σ^2 . The classical assumptions concerning the error term will sometimes be written using the notation $\text{NID}(0, \sigma^2)$, with NID standing for "Normally and independently distributed."

Suppose that there are $n > k$ observations for statistical analysis. It follows from (1.1) that the random variables y_1, \dots, y_n are independently distributed, with individual distributions being given by

$$y_i \sim N\left(\sum_{j=1}^k x_{ij}\beta_j, \sigma^2\right), i = 1, \dots, n. \quad (1.2)$$

The system of n equations with typical member (1.1) can be written in matrix-vector notation as

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{u}, \quad (1.3)$$

in which: \mathbf{y} and \mathbf{u} are the n -dimensional vectors with typical elements y_i and u_i , respectively; \mathbf{X} is the n by k matrix with typical element x_{ij} , which is assumed to have rank equal to k , that is, there is no perfect multicollinearity; and $\boldsymbol{\beta}$ is the k -dimensional vector with typical element β_j .

The classical assumptions about the errors imply that their joint distribution can be written in the form

$$\mathbf{u} \sim N(\mathbf{0}_n, \sigma^2 \mathbf{I}_n), \quad (1.4)$$

in which: $N(\boldsymbol{\mu}, \boldsymbol{\Omega})$ denotes the multivariate Normal distribution with *mean vector* $\boldsymbol{\mu}$ and *covariance matrix* $\boldsymbol{\Omega}$; $\mathbf{0}_n$ is the n -dimensional column vector with every element equal to zero; and \mathbf{I}_n denotes the $n \times n$ identity matrix. These assumptions, combined with those made about the regressor terms, also imply that the joint distribution of the elements of \mathbf{y} is given by

$$\mathbf{y} \sim N(\mathbf{X}\boldsymbol{\beta}, \sigma^2 \mathbf{I}_n). \quad (1.5)$$

The parameters to be estimated are, therefore, the elements of $\boldsymbol{\theta}' = (\boldsymbol{\beta}', \sigma^2)$.

Under classical assumptions, there are strong incentives to use the *ordinary least squares* (OLS) estimator for $\boldsymbol{\beta}$ because it is best unbiased and also the maximum likelihood estimator (MLE). The OLS estimator of $\boldsymbol{\beta}$ is

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}, \quad (1.6)$$

and so (1.5) implies that

$$\hat{\boldsymbol{\beta}} \sim N(\boldsymbol{\beta}, \sigma^2 (\mathbf{X}'\mathbf{X})^{-1}), \quad (1.7)$$

with $\hat{\boldsymbol{\beta}} = (\hat{\beta}_1, \dots, \hat{\beta}_k)'$. The implied vector of OLS predicted values is denoted by

$$\hat{\mathbf{y}} = \mathbf{X}\hat{\boldsymbol{\beta}} = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}, \quad (1.8)$$

using (1.6).

In (1.8), pre-multiplication of \mathbf{y} by $\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'$ produces $\hat{\mathbf{y}}$, which is read as “y-hat.” The n by n matrix $\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'$ is sometimes referred to as the *hat-matrix* and is denoted by \mathbf{H} , that is,

$$\mathbf{H} = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}' \quad (1.9)$$

The diagonal elements of \mathbf{H} , denoted by $h_{ii}, i = 1, \dots, n$, are called the *leverage values* in the literature on diagnostics for regression models. By combining (1.7) and (1.8), it can be seen that, in the classical framework,

$$\hat{\mathbf{y}} \sim N(\mathbf{X}\boldsymbol{\beta}, \sigma^2\mathbf{H}), \quad (1.10)$$

in which \mathbf{H} is a matrix that is symmetric and idempotent, having rank equal to k .

It remains to estimate the error variance σ^2 . The errors are not observed but their variability can be estimated by using the OLS residuals as proxies. The OLS residuals are the elements of

$$\hat{\mathbf{u}} = \mathbf{y} - \hat{\mathbf{y}} = (\mathbf{I}_n - \mathbf{H})\mathbf{y} = \mathbf{M}\mathbf{y} = \mathbf{M}\mathbf{u}, \quad (1.11)$$

in which $\mathbf{M} = \mathbf{I}_n - \mathbf{H} = (\mathbf{I}_n - \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}')$ has rank equal to $(n - k)$. Like \mathbf{H} , \mathbf{M} is a symmetric, idempotent matrix; so that (1.4) implies

$$\hat{\mathbf{u}} \sim N(\mathbf{0}_n, \sigma^2(\mathbf{I}_n - \mathbf{H})), \quad (1.12)$$

with a typical OLS residual having a Normal distribution according to

$$\hat{u}_i \sim N(0, \sigma^2(1 - h_{ii})). \quad (1.13)$$

The residual sum of squares (RSS) from OLS estimation is

$$RSS = \sum_{i=1}^n \hat{u}_i^2 = \hat{\mathbf{u}}'\hat{\mathbf{u}}. \quad (1.14)$$

Under the assumptions of the classical regression model, the distribution of RSS is given by

$$RSS \sim \sigma^2\chi^2(n - k), \quad (1.15)$$

in which $n - k$ is the number of degrees of freedom associated with the estimation of (1.3). It follows from properties of the χ^2 distribution that if s^2 is defined by

$$s^2 = \frac{RSS}{(n - k)}, \quad (1.16)$$

then s^2 is unbiased and consistent for σ^2 . The MLE of σ^2 is given by

$$\hat{\sigma}^2 = \frac{RSS}{n} = \frac{(n-k)}{n} \cdot \frac{RSS}{(n-k)}, \quad (1.17)$$

and so is consistent, but not unbiased.

It will be assumed that $\hat{\beta}$ of (1.6) and s^2 of (1.16) are to be used for the estimation of $\theta' = (\beta', \sigma^2)$, whether or not the restrictive assumptions of nonrandom regressors and Normally distributed errors are made. In addition to the unrestricted estimation of the elements of θ , there is often interest in testing restrictions that reduce the number of elements of β that require estimation. Such restrictions can take many forms. If the restrictions are linear, that is, they specify the values of known linear combinations of the regression coefficients, the assumptions of the classical model permit the application of tests that are exactly valid. In such a case, let the restrictions to be tested be written as the null hypothesis

$$H_0 : \mathbf{R}\beta = \mathbf{r}, \quad (1.18)$$

in which \mathbf{R} is a known q by k , $q \leq k$, matrix with rank equal to q and \mathbf{r} is a known q -dimensional vector.

The alternative hypothesis is assumed to be

$$H_1 : \beta_1, \dots, \beta_k \text{ are unrestricted.}$$

The OLS estimator $\hat{\beta}$ of (1.6) minimizes the residual sum of squares

$$Q(\beta) = (\mathbf{y} - \mathbf{X}\beta)'(\mathbf{y} - \mathbf{X}\beta),$$

under H_1 , and will be called the *unrestricted estimator*. The elements of $\hat{\mathbf{u}}$ will be referred to as the *unrestricted residuals*. It is convenient to add to the notation by using $RSS(H_1)$ to stand for the unrestricted residual sum of squares, that is, the quantity defined by (1.14) and to denote the number of degrees of freedom for the unrestricted model by $df(H_1)$.

The estimator that minimizes $Q(\beta)$ subject to H_0 , that is, subject to the restrictions of $\mathbf{R}\beta = \mathbf{r}$, will be called the *restricted estimator* and is denoted by $\tilde{\beta}$. The *restricted residuals* are defined by

$$\tilde{\mathbf{u}} = \mathbf{y} - \mathbf{X}\tilde{\beta}, \quad (1.19)$$

and the restricted residual sum of squares is written as

$$RSS(H_0) = \tilde{\mathbf{u}}'\tilde{\mathbf{u}}.$$

The standard F -statistic for testing H_0 against H_1 can then be calculated as

$$F = \frac{RSS(H_0) - RSS(H_1)}{RSS(H_1)} \cdot \frac{df(H_1)}{q}, \quad (1.20)$$

where, in this case, $df(H_1) = n - k$. When the null hypothesis is true and the classical assumptions are satisfied, F of (1.20) has the F distribution with q and $df(H_1)$ degrees of freedom. This result, which is exactly valid, is written as

$$F \sim F(q, df(H_1)),$$

under H_0 . Large values of the test statistic in (1.20) indicate that there is strong evidence against H_0 , so that a one-sided test should be conducted. If the required significance level is α , the decision rule can be written as:

$$\text{reject } H_0 \text{ if } F \geq f(\alpha; q, df(H_1)), \quad (1.21)$$

in which the *critical value* $f(\cdot)$ is determined by

$$\Pr(F(q, df(H_1)) \leq f(\alpha; q, df(H_1))) = 1 - \alpha.$$

If there is a single linear restriction to be tested, there is an alternative to calculating the F -statistic of (1.20). Suppose that the null hypothesis has the form $H_0 : \mathbf{R}\boldsymbol{\beta} = r_1$, where \mathbf{R} is the row vector (R_{11}, \dots, R_{1k}) and r_1 is a specified scalar, and the alternative hypothesis is $H_1 : \mathbf{R}\boldsymbol{\beta} \neq r_1$. With this combination of a single restriction in H_0 and a two-sided alternative, the reference distribution for the F -test is $F(1, df(H_1))$. A random variable with the same distribution is the square of a random variable that has the Student t distribution with $df(H_1)$ degrees of freedom. This relationship is denoted by

$$F(1, df(H_1)) = [t(df(H_1))]^2.$$

It follows that a test of a single restriction against a two-sided alternative can be based upon the t -ratio defined by

$$t = \frac{\mathbf{R}\hat{\boldsymbol{\beta}} - r_1}{SE(\mathbf{R}\hat{\boldsymbol{\beta}} - r_1)}, \quad (1.22)$$

in which $SE(\cdot)$ denotes the estimated standard error, that is,

$$SE(\mathbf{R}\hat{\boldsymbol{\beta}} - r_1) = \sqrt{s^2 \mathbf{R}(\mathbf{X}'\mathbf{X})^{-1} \mathbf{R}'}$$

Under the assumptions of the classical regression model, a two-sided t -test with significance level α can be based upon the decision rule

$$\text{reject } H_0 \text{ if } |t| \geq t(\alpha/2; df(H_1)), \quad (1.23)$$

in which $\Pr(t(df(H_1)) \leq t(\alpha/2; df(H_1))) = 1 - \alpha/2$. This two-sided t -test is equivalent to the F -test, with the sample values of test statistics obeying $t^2 = F$.

If there is a priori (non-sample) information about the sign of $\mathbf{R}\boldsymbol{\beta} - r_1$ when H_0 is false, a one-sided t -test can be applied in the usual way. With $H_1^+ : \mathbf{R}\boldsymbol{\beta} > r_1$, the decision rule is

$$\text{reject } H_0 \text{ if } t \geq t(\alpha; df(H_1)), \quad (1.24)$$

and with $H_1^- : \mathbf{R}\boldsymbol{\beta} < r_1$, it is

$$\text{reject } H_0 \text{ if } -t \geq t(\alpha; df(H_1)), \quad (1.25)$$

where $\Pr(t(df(H_1)) \leq t(\alpha; df(H_1))) = 1 - \alpha$.

Rules (1.21), (1.23), (1.24) and (1.25) have all been written so that the rejection region is in the right-hand tail of the relevant reference distribution. It is convenient, for the subsequent discussions, to assume that all tests are set up in this form. Some diagnostic checks, for example, the widely-used *test for heteroskedasticity* proposed in Breusch and Pagan (1979), involve the use of criteria that are asymptotically distributed as χ^2 under the null hypothesis. The rejection region for such tests are, as with those given above, in the right-hand tail.

It is worth noting that, as an alternative to a χ^2 -form, many diagnostic checks can be computed as seemingly conventional tests of the significance of artificial (constructed) variables that are added to the regressors of (1.1). For example, tests for autocorrelation, structural change, errors-in-variables etcetera can be computed using standard formulae for F or t statistics, which are applied to an appropriate artificial regression model; see Davidson and MacKinnon (2004, section 15.2) for a general discussion. In such cases, (1.1) is viewed as the *restricted (null) model*. The nature of the *unrestricted (alternative) model*, which contains the restricted model (1.1) as a special case, has important implications for the properties of the test of the latter against the former. The unrestricted model required for the convenient calculation of a diagnostic check is often such that

the F and t tests are not exactly valid even when the classical Normal regression model (1.1) is the correct specification.

The problems associated with appealing to classical finite sample theory in the context of testing for misspecification can be illustrated by considering the well-known *Breusch-Godfrey Lagrange Multiplier* (LM) test for autocorrelation; see Breusch (1978) and Godfrey (1978). Suppose that quarterly data are being used and that the researcher believes that it is useful to test the assumption that the errors are independent against the fourth-order alternative

$$u_i = \phi_1 u_{i-1} + \dots + \phi_4 u_{i-4} + \epsilon_i,$$

with the variates ϵ_i being $\text{NID}(0, \sigma_\epsilon^2)$. The required Breusch-Godfrey test can be implemented by applying the F -test of the four linear restrictions $\phi_1 = \phi_2 = \phi_3 = \phi_4 = 0$ in the augmented version of (1.1) given by

$$y_i = \sum_{j=1}^k x_{ij} \beta_j + \sum_{j=1}^4 \phi_j \hat{u}_{i-j} + u_i, \quad (1.26)$$

where terms \hat{u}_{i-j} with $i \leq j$ are set equal to zero. Even under the restrictive assumption that the errors u_i are $\text{NID}(0, \sigma^2)$, the F -test of (1.1) against (1.26) is not exactly valid, but does have a significance level that tends to the required level α as $n \rightarrow \infty$, that is, it is asymptotically valid.

The failure of standard finite sample theory to apply to the F -test of (1.1) against (1.26) might be anticipated on the grounds that the regressors of the latter, which serves as the alternative or unrestricted model, include random variables, namely, the lagged residuals. However, there are cases of diagnostic checks in which F -tests are exact even though the regressors of the alternative model include random variables. An important example is the *RESET test* proposed in Ramsey (1969).

The RESET test provides a check of the specification of the mean function of (1.1), with the OLS predicted values from estimation of this model being employed to obtain the additional regressors required for the alternative model. More precisely, in the formula for the RESET F -statistic with q test variables, $RSS(H_0)$ is derived from OLS estimation of (1.1), that is, it is given by (1.14), and $RSS(H_1)$ is obtained after estimation of the artificial model

$$y_i = \sum_{j=1}^k x_{ij} \beta_j + \sum_{j=1}^q \hat{y}_i^{j+1} \delta_j + u_i, \quad i = 1, \dots, n, \quad (1.27)$$

with $df(H_1) = n - k - q$. The F -statistic for testing $\delta_1 = \dots = \delta_q = 0$ in (1.27) is denoted by F_R and

$$F_R \sim F(q, n - k - q),$$

under the null hypothesis, when the assumptions of the classical model concerning \mathbf{X} and \mathbf{u} hold. Consequently, under these assumptions, it is possible to have perfect control of finite sample significance levels of the RESET test. This result follows from a general property of tests involving functions of $\hat{\mathbf{y}}$; see Milliken and Graybill (1970).

Notwithstanding the interest to theorists of results such as those in Milliken and Graybill (1970) and also in Stewart (1997), there is a need to weaken the assumptions of the classical model and to see what can be established about the properties of tests under more general conditions.

1.3. Tests for linear regression models under weaker assumptions: random regressors and non-Normal IID errors

From the viewpoint of the applied econometrician, the results concerning the exact validity of the F and t tests in the classical linear regression model are of doubtful relevance. The assumption that the regressors are non-random and would be fixed if repeated sampling were possible may well be appropriate for the analysis of data obtained, for example, from experiments in a laboratory. However, in econometric models, the regressors will usually include economic variables that are properly regarded as random. Thus, in general, the regressor set will include both random and non-random terms. The applicability of the results of the previous section is now open to question.

Suppose first that the following two conditions hold: the regressors are such that any random term x_{ij} is independent of u_m for $i, m = 1, \dots, n$; and the errors u_m are $NID(0, \sigma^2)$ for $m = 1, \dots, n$. When the first of these conditions is satisfied, the regressors are said to be *strictly exogenous* or, less precisely, *exogenous*. The complete independence of errors and regressors implies that conditioning on regressor values has no impact on the distribution of the errors. Consequently, given the two conditions, we can write the conditional error distribution as

$$\mathbf{u}|\mathbf{X} \sim N(\mathbf{0}_n, \sigma^2 \mathbf{I}_n), \quad (1.28)$$

and for the conditional distribution of the dependent variable, given the regressor values, we have

$$\mathbf{y}|\mathbf{X} \sim N(\mathbf{X}\boldsymbol{\beta}, \sigma^2\mathbf{I}_n). \quad (1.29)$$

Comparison of (1.4) and (1.5) with (1.28) and (1.29), respectively, indicates how results given in the previous section for the former pair of equations will now apply in a conditional sense under the latter pair. In particular, when testing restrictions of the form (1.18), the F statistic of (1.20) will have the conditional distribution

$$F|\mathbf{X} \sim F(q, df(H_1)), \quad (1.30)$$

under the null hypothesis. This conditional distribution is completely characterized by the values of q and $df(H_1)$, but neither of these items depends upon \mathbf{X} . Hence, when the null hypothesis is true, the unconditional distribution is the same as the conditional distribution in (1.30), which is the same as the reference distribution appropriate for the classical model. The F -test is, therefore, exactly valid. Similar arguments apply to the t -test.

However, the conditions that underpin this argument are very restrictive. The assumption that all regressors are strictly exogenous is inconsistent with the common practice of including lagged values of the dependent variable as explanatory variables when estimating regression models using time series data. For example, the standard partial adjustment model leads to the inclusion of y_{i-1} as a regressor and this regressor cannot be independent of all past errors (obviously $E(y_{i-1}u_{i-1}) \neq 0$). Moreover, there is rarely precise information available about the shape of the error distribution and, in particular, there seems little reason to believe that the errors are Normally distributed, even if they are assumed to be IID.

If the assumption of Normally distributed errors is relaxed, tests involving the use of critical values from standard distributions must, in general, be based upon asymptotic theory. Appeal has to be made to versions of a Law of Large Numbers and a Central Limit Theorem (CLT). Discussions of these topics and their application to tests for regression models can be rather technical and readers are referred to Davidson (1994), McCabe and Tremayne (1993), and White (1984) for detailed treatments. For the purpose of providing an outline of the relevant arguments of asymptotic theory, it is useful to introduce the ideas of *orders of magnitude* for random variables due to Mann and Wald (1943).

Given a sequence of real random variables, denoted by $\{S_{(n)}\}$, and some real number a , we say that $S_{(n)}$ is of order of probability n^a if for any $\varepsilon > 0$ there exists $b_\varepsilon > 0$ such that

$$\Pr(-b_\varepsilon \leq n^{-a}S_{(n)} \leq b_\varepsilon) \geq 1 - \varepsilon,$$

for all n . The standard notation for such a variable is to write $S_{(n)} = O_p(n^a)$. If, for some real number c , $p \lim n^{-c}S_{(n)} = 0$, we say that $S_{(n)}$ is of smaller order of probability than n^c and write $S_{(n)} = o_p(n^c)$.

For example, assume that the observations y_1, \dots, y_n are NID(μ, σ^2) and $S_{(n)} = y_1 + \dots + y_n$. Since, in this case, $S_{(n)}$ is $N(n\mu, n\sigma^2)$, it follows that: (i) $S_{(n)} = O_p(n)$ with $p \lim n^{-1}S_{(n)} = \mu$; (ii) $S_{(n)} - n\mu$ is $O_p(n^{1/2})$ with $n^{-1/2}(S_{(n)} - n\mu)$ being $N(0, \sigma^2)$; and (iii) $S_{(n)} - n\mu$ is $o_p(n)$ with $n^{-1}(S_{(n)} - n\mu)$ being $N(0, n^{-1}\sigma^2)$ so that $p \lim n^{-1}(S_{(n)} - n\mu) = 0$.

In standard textbook discussions of linear regression models, assumptions are made that imply that $\hat{\beta} = \beta + O_p(n^{-1/2})$, with $n^{1/2}(\hat{\beta} - \beta)$ being asymptotically Normally distributed with zero mean vector and finite, positive-definite covariance matrix. Strictly speaking, the notation used in the discussion of asymptotic theory for regression models should reflect the dependence of estimators and test statistics on the sample size n , for example, $\hat{\beta}_{(n)}$ rather than $\hat{\beta}$. However, no confusion should be caused by adopting the less cluttered style employed above and the key results can be summarized as follows. First, when using F of (1.20) to test the null hypothesis of (1.18), asymptotic theory predicts that, when the null is true, F is $O_p(1)$ with

$$F \sim_a \frac{\chi^2(q)}{q},$$

in which \sim_a is used as a shorthand for “is asymptotically distributed as”. Second, if $q = 1$, the t -ratio of (1.22) is $O_p(1)$ and is asymptotically distributed as $N(0, 1)$ when the null hypothesis is true.

Asymptotic theory can also be used as a source of approximations to the behaviour of test statistics when the null hypothesis is false. Consider the case of testing a single restriction, which is written as $H_0 : \mathbf{R}\beta = r_1$, as above. The relevant t -statistic can be written as

$$\frac{\mathbf{R}\hat{\beta} - r_1}{SE(\mathbf{R}\hat{\beta} - r_1)} = \frac{(\mathbf{R}\hat{\beta} - \mathbf{R}\beta)}{SE(\mathbf{R}\hat{\beta} - r_1)} + \frac{(\mathbf{R}\beta - r_1)}{SE(\mathbf{R}\hat{\beta} - r_1)}, \quad (1.31)$$

in which the first term on the right-hand side of (1.31) tends to $N(0, 1)$, whether or not H_0 is true, but the asymptotic behaviour of the second

term does depend upon the value of $\mathbf{R}\boldsymbol{\beta} - r_1$. If H_0 is true, $\mathbf{R}\boldsymbol{\beta} - r_1 = 0$ and the second term vanishes. If $\mathbf{R}\boldsymbol{\beta} - r_1$ is a fixed nonzero number (so that H_0 is untrue), the second term on the right-hand side of (1.31) is $O_p(n^{1/2})$, under the standard assumptions of asymptotic theory for regression models. (The standard errors of OLS estimators are $O_p(n^{-1/2})$, given these assumptions.) Hence, as $n \rightarrow \infty$, the t -statistic goes to $\pm\infty$, according to the sign of the nonzero constant $\mathbf{R}\boldsymbol{\beta} - r_1$. Thus, with *fixed alternatives* $H_1 : \mathbf{R}\boldsymbol{\beta} - r_1 \neq 0$, asymptotic theory cannot lead to the limit of a proper distribution with finite mean and variance as a basis for approximating the behaviour of the t -statistic. A device known as a *sequence of local alternatives*, or as *Pitman drift*, does allow asymptotic theory to provide such an approximation for the study of power; see, for example, Godfrey and Tremayne (1988).

The device is to introduce the sequence of alternatives

$$H_{1n} : \mathbf{R}\boldsymbol{\beta} - r_1 = \frac{\lambda}{\sqrt{n}}, |\lambda| < \infty, \tag{1.32}$$

which clearly tends to the null hypothesis as n increases. The second term of (1.31) is, under (1.32), given by

$$\frac{\lambda}{\sqrt{n}SE(\mathbf{R}\hat{\boldsymbol{\beta}} - r_1)},$$

which tends to a finite constant, say μ_λ . Consequently, under the local alternatives assumption, the asymptotic distribution of the t -ratio can be written as

$$\frac{\mathbf{R}\hat{\boldsymbol{\beta}} - r_1}{SE(\mathbf{R}\hat{\boldsymbol{\beta}} - r_1)} \sim_a N(\mu_\lambda, 1),$$

and this distribution satisfies the requirements to have finite mean and variance. Local alternatives are often used when researchers seek to choose between two or more asymptotically valid tests on the basis of their sensitivity to departures from the null hypothesis. A similar result is available when the F -test is used to check several restrictions.

Several researchers, while acknowledging a reliance on asymptotic theory, prefer to use the conventional $F(q, df(H_1))$ and $t(df(H_1))$ distributions for critical values, rather than the corresponding limiting forms of $\chi^2(q)/q$ and $N(0, 1)$. There may be reason to be concerned about the relevance of asymptotic theory if $df(H_1)$ is not large enough for the choice between, for example, $t(df(H_1))$ and $N(0, 1)$ to be unimportant. Indeed, from a practical point of view, a question of real interest is how large does

the sample size have to be before a CLT will give a useful approximation for controlling the significance level when testing the null hypothesis. Unfortunately there is no generally valid answer.

The robustness of the standard regression F -test to *non-Normality* of the errors is investigated in Ali and Sharma (1996). In addition to the sample size, degrees of freedom and the actual distribution of the errors, important determinants of the robustness to non-Normality are the non-Normality of the regressors and the presence of observations with relatively high leverage values. The relevance of such characteristics of the regressor set is not surprising, given the dependence of the test statistic on OLS residuals and the form of (1.11). In view of the uncertain quality of the approximation provided by asymptotic theory in the case of a linear regression model with IID, but non-Normal, errors and the evidence that the approximation is sometimes poor, it is natural to look for an alternative approach to testing. Chapter 2 contains a discussion of a simulation-based approach that can be applied in the context of linear regression models with IID errors. However, like the standard t and F tests, these simulation techniques may produce misleading inferences when the errors of (1.1) are not IID, that is, the data are generated by a *generalized regression model*.

1.4. Tests for generalized linear regression models

The generalized regression model with exogenous regressors is derived by combining the model in (1.3), that is,

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{u},$$

with

$$E(\mathbf{u}|\mathbf{X}) = \mathbf{0},$$

and

$$E(\mathbf{u}\mathbf{u}'|\mathbf{X}) = \sigma^2\boldsymbol{\Omega}, \tag{1.33}$$

in which $\boldsymbol{\Omega}$ is an n by n matrix that is symmetric and positive definite. If the errors are independent but heteroskedastic, the elements of $\boldsymbol{\Omega}$ are such that: $\omega_{ij} = 0$ if $i \neq j$; and $\omega_{ii} \neq \omega_{jj}$ for some $i \neq j$. If the errors are correlated but homoskedastic, the elements of $\boldsymbol{\Omega}$ are such that: $\omega_{ii} = \omega_{jj} = 1$, say, for all i and j ; and $\omega_{ij} \neq 0$ for some $i \neq j$. In the latter case, it is assumed that there are time series data and the errors are autocorrelated.

(Tests can be developed for models with spatial correlation; see Anselin, 2006.) It will be assumed that autocorrelated errors are generated by (weakly) stationary processes so that ω_{ij} depends upon $|i - j|$, rather than on i and j separately. For example, if the errors were generated by a stationary first-order autoregression

$$u_i = \phi u_{i-1} + \epsilon_i, |\phi| < 1, \epsilon_i \sim NID(0, \sigma_\epsilon^2),$$

a typical element of $\mathbf{\Omega}$ in (1.33) would be $\omega_{ij} = \phi^{|i-j|}$.

The OLS estimator of β , under the assumptions of the generalized regression model, has conditional mean vector

$$E(\hat{\beta}|\mathbf{X}) = \beta,$$

and conditional covariance matrix given by

$$V_G(\hat{\beta}|\mathbf{X}) = \sigma^2(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{\Omega}\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}. \tag{1.34}$$

In general, the matrix of (1.34) is not equal to the one that appears in (1.7) and so the tests described above cannot be expected to be asymptotically valid.

In some special models, the elements of $\mathbf{\Omega}$ are known constants. For example, if each element of \mathbf{u} is the sum of a known number of basic IID disturbances, $\mathbf{\Omega}$ can be calculated very simply; see Rowley and Wilton (1973) for an example based upon the “four-quarter overlapping-change” model in wage analysis. When $\mathbf{\Omega}$ is known, the OLS estimator can be replaced by the more efficient Generalized Least Squares (GLS) estimator

$$\check{\beta} = (\mathbf{X}'\mathbf{\Omega}^{-1}\mathbf{X})^{-1}\mathbf{X}'\mathbf{\Omega}^{-1}\mathbf{y}, \tag{1.35}$$

which has conditional covariance matrix equal to $\sigma^2(\mathbf{X}'\mathbf{\Omega}^{-1}\mathbf{X})^{-1}$. The estimator of σ^2 is no longer given by s^2 of (1.16) but is now defined by

$$\check{\sigma}^2 = \frac{(\mathbf{y} - \mathbf{X}\check{\beta})'\mathbf{\Omega}^{-1}(\mathbf{y} - \mathbf{X}\check{\beta})}{(n - k)}.$$

Given $\check{\beta}$ and $\check{\sigma}^2$, an asymptotically valid test of $H_0 : \mathbf{R}\beta = \mathbf{r}$ in (1.18) can be based upon the result that, when H_0 is true, the *Wald statistic*

$$W_{GLS} = (\mathbf{R}\check{\beta} - \mathbf{r})' \left[\check{\sigma}^2 \mathbf{R}(\mathbf{X}'\mathbf{\Omega}^{-1}\mathbf{X})^{-1}\mathbf{R}' \right]^{-1} (\mathbf{R}\check{\beta} - \mathbf{r}), \tag{1.36}$$

is asymptotically distributed as $\chi^2(q)$; see, for example, Greene (2008, ch. 8, section 8.3.1) for the corresponding asymptotically valid *F*-statistic.

Significantly large values of W_{GLS} indicate that the restrictions of $H_0 : \mathbf{R}\boldsymbol{\beta} = \mathbf{r}$ are not consistent with the sample data.

Unfortunately, the test statistic of (1.36) is rarely available in practical situations because, in general, $\boldsymbol{\Omega}$ is unknown and it is not feasible to calculate the GLS estimator $\hat{\boldsymbol{\beta}}$.

When the elements of $\boldsymbol{\Omega}$ are continuous functions of the elements of an unknown parameter vector $\boldsymbol{\psi}$, estimates of the parameters of the generalized regression model can be obtained by minimizing the *Nonlinear Least Squares* (NLS) criterion

$$Q_{NLS}(\boldsymbol{\beta}, \boldsymbol{\psi}) = (\mathbf{y} - \mathbf{X}\boldsymbol{\beta})' [\boldsymbol{\Omega}(\boldsymbol{\psi})]^{-1} (\mathbf{y} - \mathbf{X}\boldsymbol{\beta}),$$

with respect to both $\boldsymbol{\beta}$ and $\boldsymbol{\psi}$. Alternatively, if some consistent estimator of $\boldsymbol{\psi}$, denoted by $\hat{\boldsymbol{\psi}}$, is available and necessary regularity conditions are satisfied, $\boldsymbol{\beta}$ can be estimated by minimizing the *Feasible Generalized Least Squares* (FGLS) function

$$Q_{FGLS}(\boldsymbol{\beta}) = (\mathbf{y} - \mathbf{X}\boldsymbol{\beta})' [\boldsymbol{\Omega}(\hat{\boldsymbol{\psi}})]^{-1} (\mathbf{y} - \mathbf{X}\boldsymbol{\beta}).$$

However, both of these estimation methods are based upon the assumptions that: (i) there is a parametric model that determines the structure of $\boldsymbol{\Omega}$; and (ii) the general form of $\boldsymbol{\Omega}(\boldsymbol{\psi})$ is known, with only its finite-dimensional parameter vector $\boldsymbol{\psi}$ being unknown. While economics might be a source of useful information about the mean function of \mathbf{y} , there is little reason to suppose that applied workers will know the form of, for example, heteroskedasticity. Thus it will often be difficult to have confidence in an assumed error model.

Misspecification of the model for autocorrelation and/or heteroskedasticity will, in general, lead to an inconsistent estimator of the covariance matrix of the minimizers of $Q_{NLS}(\boldsymbol{\beta}, \boldsymbol{\psi})$ and $Q_{FGLS}(\boldsymbol{\beta})$. Hence errors made in modelling $\boldsymbol{\Omega}$ may imply misleading outcomes of tests of hypotheses such as (1.18), because such tests use the estimated covariance matrix to assess the significance of sample outcomes. An investigation of the effects of misspecifying the model for heteroskedasticity is reported in Belsley (2002). It is found that effects can be serious and Belsley concludes that

Correction for heteroskedasticity clearly does best when both the proper arguments and the proper form of the skedasticity function are known. But this is an empty conclusion since misspecification is probably the rule. (Belsley, 2002, p. 1396)

Moreover, it has been argued that, even with correct specification of the model underlying Ω , it is not clear that FGLS is superior to OLS in finite samples because of the extra variability associated with the estimation of ψ ; see, for example, Greene (2008, p. 158).

In view of these findings, it is not surprising that there has been an interest in deriving tests using the uncorrected OLS estimator $\hat{\beta}$ and an appropriate estimator of its covariance matrix, which is no longer given by the *IID-valid* formula $\sigma^2(\mathbf{X}'\mathbf{X})^{-1}$ used in (1.7). If the errors are assumed to be independent and heteroskedastic, a *Heteroskedasticity-Consistent Covariance Matrix Estimator* (usually denoted by *HCCME*) is required. If the errors are heteroskedastic and autocorrelated, a *Heteroskedasticity and Autocorrelation Consistent* (usually denoted by *HAC*) estimator is needed. The former provides standard errors that are *heteroskedasticity-robust*. The latter provides standard errors that are *heteroskedasticity and autocorrelation robust*.

Many computer programs offer users the chance to use robust standard errors from either some HCCME or some HAC estimate, rather than relying on the traditional IID-valid standard errors given by the matrix $s^2(\mathbf{X}'\mathbf{X})^{-1}$. However, the traditional standard errors are often provided as the default and this approach has been criticized. Stock and Watson remark that

In econometric applications, there is rarely a reason to believe that the errors are homoskedastic and normally distributed. Because sample sizes are typically large, however, inference can proceed... by first computing the heteroskedasticity-robust standard errors. (Stock and Watson, 2007, p. 171)

Similarly, it is argued in Hansen (1999) that a modern approach should involve the use of test statistics that are valid under heteroskedasticity and do not require the assumption of Normality. (It is also suggested in Hansen (1999) that applied workers should think about using the bootstrap for inference, rather than relying on asymptotic theory. Much of what follows in this book is concerned with presenting evidence to support this suggestion and to help empirical researchers to select the appropriate form of the bootstrap.)

Since the use of procedures based upon HCCME and HAC estimates offers the chance to derive tests that are asymptotically valid in the presence of unspecified forms of departure from the assumption of IID errors, such robust tests are of real interest in practical applications. Moreover, the availability of suitable software means that there is no important

obstacle to hinder the use of robust tests. Given the potential importance of these alternatives to the conventional IID-based asymptotic t and F tests, each will be discussed.

1.4.1. HCCME-based tests

Suppose first that the errors u_i are independently distributed with zero means and variances $\sigma_i^2, i = 1, \dots, n$, with all variances being finite and positive. It is not assumed that there is any precise information available to support the specification of a parametric model of the heteroskedasticity. The tests that are to be discussed are asymptotically robust to heteroskedasticity of unspecified form and are also asymptotically valid under the classical assumption of homoskedasticity. The key results for HCCME-based inference in linear regression models will now be discussed.

If the regressors were not random and the errors were Normally distributed, the OLS estimator would, in the presence of unspecified forms of heteroskedasticity, have the following distribution

$$\hat{\beta} \sim N(\beta, (X'X)^{-1}X'\Sigma X(X'X)^{-1}),$$

or equivalently,

$$n^{1/2}(\hat{\beta} - \beta) \sim N(\mathbf{0}_k, n(X'X)^{-1}X'\Sigma X(X'X)^{-1}), \quad (1.37)$$

in which Σ is the n by n diagonal matrix with the variances $\sigma_i^2, i = 1, \dots, n$, as the nonzero elements on its leading diagonal. The random vector $n^{1/2}(\hat{\beta} - \beta)$ is $O_p(1)$, with the covariance matrix that appears in (1.37) being assumed to tend to a finite positive-definite matrix as $n \rightarrow \infty$. This property of the covariance matrix is more easily seen when it is noted that

$$n(X'X)^{-1}X'\Sigma X(X'X)^{-1} = \left(\frac{X'X}{n}\right)^{-1} \left(\frac{X'\Sigma X}{n}\right) \left(\frac{X'X}{n}\right)^{-1}.$$

The covariance matrix that appears in (1.37) is sometimes referred to as a *sandwich covariance matrix*; the term depending on error variances, that is, $X'\Sigma X$, being sandwiched between the two terms equal to $(X'X)^{-1}$. The problem of finding useful estimates for the sandwich form in order to develop methods for feasible inference was studied in the statistics literature, for example, Eicker (1967). However, interest and applications in econometrics were stimulated by an important paper by White who relaxed the assumptions of fixed regressors and Normally distributed errors; see White (1980).

White showed that, under suitable regularity conditions, the OLS estimator $\hat{\beta}$ is consistent for β , with $(\hat{\beta} - \beta)$ being $O_p(n^{-1/2})$ and $n^{1/2}(\hat{\beta} - \beta)$ having a limiting distribution (as $n \rightarrow \infty$) given by

$$n^{1/2}(\hat{\beta} - \beta) \sim_a N(\mathbf{0}_k, \text{plim } n(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\Sigma\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}). \quad (1.38)$$

It is (1.38) that provides the basis for asymptotically valid heteroskedasticity-robust tests. If the null hypothesis $H_0 : \mathbf{R}\beta = \mathbf{r}$ is to be tested, we can use the result that (1.38) implies that

$$n^{1/2}\mathbf{R}(\hat{\beta} - \beta) \sim_a N(\mathbf{0}_q, \text{plim } n\mathbf{R}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\Sigma\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{R}'),$$

and so, if the null hypothesis is true,

$$n^{1/2}(\mathbf{R}\hat{\beta} - \mathbf{r}) \sim_a N(\mathbf{0}_q, \text{plim } n\mathbf{R}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\Sigma\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{R}').$$

Consequently, if the restrictions of $\mathbf{R}\beta = \mathbf{r}$ are valid, standard asymptotic theory implies that

$$n(\mathbf{R}\hat{\beta} - \mathbf{r})'[\text{plim } n\mathbf{R}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\Sigma\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{R}']^{-1}(\mathbf{R}\hat{\beta} - \mathbf{r}) \sim_a \chi^2(q).$$

However, this result does not yield a feasible test procedure because it concerns a random variable that depends upon the probability limit of a matrix that is, in part, determined by the unknown matrix Σ .

White provided a very simple and convenient solution to the problem of deriving a feasible large sample test. In White (1980), it is shown that, under certain regularity conditions that place mild restrictions on the behaviour of errors and random regressors,

$$\text{plim } n^{-1}\mathbf{X}'\dot{\Sigma}\mathbf{X} = \text{plim } n^{-1}\mathbf{X}'\Sigma\mathbf{X}, \quad (1.39)$$

in which $\dot{\Sigma}$ is obtained from Σ by replacing the unknown variance σ_i^2 by the calculable squared OLS residual $\hat{u}_i^2, i = 1, \dots, n$. Consequently feasible and asymptotically robust tests can be derived by using the heteroskedasticity-consistent estimator

$$HCO = n(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\dot{\Sigma}\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}, \quad (1.40)$$

for the covariance matrix that appears in (1.38). A heteroskedasticity-robust test of $H_0 : \mathbf{R}\beta = \mathbf{r}$ can then be based upon the statistic

$$W_{HCO} = n(\mathbf{R}\hat{\beta} - \mathbf{r})' \left[n\mathbf{R}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\dot{\Sigma}\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{R}' \right]^{-1} (\mathbf{R}\hat{\beta} - \mathbf{r}) \sim_a \chi^2(q),$$

with significantly large values of W_{HC0} indicating the data inconsistency of the linear restrictions of H_0 .

There is a large literature on the construction and analysis of heteroskedasticity-robust tests for regression models and a summary will be given in Chapter 6. However, it is worth noting that statistics that are asymptotically equivalent to W_{HC0} , that is, differ from it by terms that are $o_p(1)$, can be obtained by modifying $HC0$ of (1.40). Three modifications are often discussed. First, a simple degrees-of-freedom adjustment is employed, which leads to

$$HC1 = (n - k)(X'X)^{-1}X'\check{\Sigma}X(X'X)^{-1}. \quad (1.41)$$

The second and third standard modifications both involve taking the leverage values h_{ii} (see (1.9) above) into account, with the estimators being defined by

$$HC2 = n(X'X)^{-1}X'\check{\check{\Sigma}}X(X'X)^{-1}, \quad (1.42)$$

and

$$HC3 = n(X'X)^{-1}X'\check{\check{\check{\Sigma}}}X(X'X)^{-1}, \quad (1.43)$$

in which $\check{\check{\Sigma}}$ and $\check{\check{\check{\Sigma}}}$ are derived from $\check{\Sigma}$ by replacing the terms \hat{u}_i^2 by $(1 - h_{ii})^{-1}\hat{u}_i^2$ and $(1 - h_{ii})^{-2}\hat{u}_i^2$, $i = 1, \dots, n$, respectively. Clearly $HC0$ and $HC1$ have the same probability limit, with

$$HC1 = \frac{(n - k)}{n} \cdot HC0 = HC0 + O_p(n^{-1}),$$

so that $(HC1 - HC0)$ is asymptotically negligible relative to $HC0$. Similarly the differences $(HC2 - HC0)$ and $(HC3 - HC0)$ are also asymptotically negligible since each term h_{ii} is $O_p(n^{-1})$, with $h_{11} + \dots + h_{nn} = k$ for all $n \geq k$. An examination of these variants is provided in, for example, Long and Ervin (2000) and MacKinnon and White (1985).

Many textbooks point out that heteroskedasticity could be present when regression models are estimated using cross-section data. It is, therefore, not surprising that the assumption that the regressors are independently distributed over the observations is made in White (1980). However, while this assumption concerning the behaviour of regressors may often be appropriate for cross-section applications, it is too restrictive when time series regressions are estimated and heteroskedasticity certainly cannot be ruled out in such cases. Fortunately, it is possible to extend White's results by establishing that a HCCME can be obtained

for time series regressions by using one of the estimators given in (1.40), (1.41), (1.42) and (1.43); see Hsieh (1983).

If time series data are being used in OLS analysis, it does not seem sufficient to make tests robust to heteroskedasticity because it seems reasonable to consider the possibility of error autocorrelation. Consequently there is a need to examine the possibility of deriving procedures that are asymptotically valid in the presence of unspecified forms of both autocorrelation and heteroskedasticity. The key step to the construction of such tests is the derivation of a heteroskedasticity and autocorrelation consistent estimator of the covariance matrix of the OLS estimator.

1.4.2. HAC-based tests

Since the use of HAC-based tests seems especially useful in the context of time series regressions, it will be convenient to rewrite a typical observation for the linear regression model as

$$y_t = \sum_{j=1}^k x_{tj} \beta_j + u_t, \quad (1.44)$$

that is, the observation subscript is now t for time period. It is assumed that the errors can exhibit both autocorrelation and heteroskedasticity, but there are no unit roots. Details of the regularity conditions needed to support asymptotic analysis are given in the literature, for example, see Andrews (1991), Andrews and Monahan (1992), Kiefer and Vogelsang (2002), Kiefer, Vogelsang and Bunzel (2000) and Newey and West (1987).

As before, it is assumed that it is required to test linear restrictions on the regression coefficients using OLS estimators. Let the k -dimensional vector \mathbf{x}_t be defined by

$$\mathbf{x}_t = (x_{t1}, \dots, x_{tk})',$$

so that (1.44) can be rewritten as

$$y_t = \mathbf{x}_t' \boldsymbol{\beta} + u_t.$$

The OLS estimator can then be expressed as

$$\hat{\boldsymbol{\beta}} = \left(\sum_{t=1}^n \mathbf{x}_t \mathbf{x}_t' \right)^{-1} \sum_{t=1}^n \mathbf{x}_t y_t,$$

so that we have

$$n^{1/2}(\hat{\beta} - \beta) = \left(n^{-1} \sum_{t=1}^n \mathbf{x}_t \mathbf{x}'_t \right)^{-1} n^{-1/2} \sum_{t=1}^n \mathbf{x}_t u_t, \quad (1.45)$$

in which all terms are scaled to be $O_p(1)$. As with the simpler case of HCCME, the basic problem is to obtain a consistent estimator of the covariance matrix of the asymptotic distribution of $n^{1/2}(\hat{\beta} - \beta)$. Asymptotic tests of linear restrictions can then be readily found.

Under standard assumptions, the first term on the right-hand side of (1.45) tends to a finite positive definite matrix and the second term is asymptotically Normally distributed. The asymptotic covariance of $n^{1/2}(\hat{\beta} - \beta)$ is then given by

$$plim \left(n^{-1} \sum_{t=1}^n \mathbf{x}_t \mathbf{x}'_t \right)^{-1} \times \Phi \times plim \left(n^{-1} \sum_{t=1}^n \mathbf{x}_t \mathbf{x}'_t \right)^{-1},$$

in which Φ is the asymptotic covariance of $n^{-1/2} \sum_{t=1}^n \mathbf{x}_t u_t$. Consistent estimators of the matrices on the outside of this sandwich form are obtained simply by dropping the *plim* operator, so the real issue is how to estimate Φ .

It is standard in work on HAC estimation of covariance matrices to introduce the *autocovariance matrices* of the vectors $\mathbf{x}_t u_t$, which are defined as follows:

$$\begin{aligned} \Gamma(j) &= n^{-1} \sum_{t=j+1}^n E(u_t u_{t-j} \mathbf{x}_t \mathbf{x}'_{t-j}) \text{ for } j \geq 0, \\ &= n^{-1} \sum_{t=-j+1}^n E(u_{t+j} u_t \mathbf{x}_{t+j} \mathbf{x}'_t) \text{ for } j < 0. \end{aligned} \quad (1.46)$$

Now Φ is the limit, as $n \rightarrow \infty$, of the sum

$$\mathbf{J} = \sum_{j=-n+1}^{n-1} \Gamma(j);$$

see, for example, Andrews (1991, eq. 2.3, p. 820) or Hamilton (1994, pp. 279–283). Several estimators for Φ have been discussed and there are many articles of relevance; see, for example, Andrews (1991), Andrews

and Monahan (1992), Kiefer and Vogelsang (2002), Kiefer, Vogelsang and Bunzel (2000) and Newey and West (1987). Computer programs for econometric estimation and testing often provide the *Newey-West estimator*, which is

$$\hat{\mathbf{J}} = \hat{\mathbf{\Gamma}}(0) + \sum_{j=1}^l \left(1 - \frac{j}{l+1}\right) (\hat{\mathbf{\Gamma}}(j) + \hat{\mathbf{\Gamma}}(j)'), \quad (1.47)$$

where $\hat{\mathbf{\Gamma}}(0) = n^{-1} \mathbf{X}' \hat{\Sigma} \mathbf{X}$, as in (1.39), and

$$\hat{\mathbf{\Gamma}}(j) + \hat{\mathbf{\Gamma}}(j)' = n^{-1} \sum_{t=j+1}^n \hat{u}_t \hat{u}_{t-j} (\mathbf{x}_t \mathbf{x}'_{t-j} + \mathbf{x}_{t-j} \mathbf{x}'_t) \text{ for } j = 1, \dots, l, \quad (1.48)$$

where l denotes the lag truncation value that allows all asymptotically relevant autocorrelations to be taken into account, with $l \rightarrow \infty$ and $l/n \rightarrow 0$, as $n \rightarrow \infty$.

A test of the usual set of linear restrictions (1.18) can now be obtained, since

$$W_{HAC} = n(\mathbf{R}\hat{\boldsymbol{\beta}} - \mathbf{r})' \left[n\mathbf{R}(\mathbf{X}'\mathbf{X})^{-1} \hat{\mathbf{J}}(\mathbf{X}'\mathbf{X})^{-1} \mathbf{R}' \right]^{-1} (\mathbf{R}\hat{\boldsymbol{\beta}} - \mathbf{r}) \sim_a \chi^2(q), \quad (1.49)$$

when the null hypothesis is true and l is selected in an appropriate way. In the analysis of Newey and West, l is $o(n^{-1/4})$, but it is remarked that, under different assumptions, l being $o(n^{-1/2})$ would be appropriate for asymptotically valid inference; see Newey and West (1987, p. 705). Clearly rules about the asymptotic order of magnitude in n of the truncation value l do not provide real guidance about the choice of this parameter in a practical situation in which a finite sample is employed.

There are several problems for the practitioner who has to choose the form of the HAC estimator to derive an asymptotic test that is robust to unspecified forms of heteroskedasticity and autocorrelation; see Den Hann and Levin (1997). The choice can have an important impact on the finite sample performance of the test and some variants do not guarantee that the calculated test statistic will be positive, which is inconsistent with the reference distribution of $\chi^2(q)$.

There is also an important limitation on the types of model to which OLS-based HAC tests can be applied. A fundamental requirement is that $E(u_t | \mathbf{x}_t) = 0$, which excludes the possibility of having lagged dependent

variables as regressors in the presence of unspecified forms of autocorrelation. It is very often the case that applied workers use OLS to estimate time series regression equations with lagged dependent variables as regressors and HAC tests should not be employed in such situations. It may be possible to obtain useful HAC tests for dynamic models by using *instrumental variable* estimation, rather than OLS. However, instrumental variable estimation is sometimes associated with difficulties. Some problems encountered in using instrumental variable methods to estimate dynamic regression models with autocorrelated errors are discussed in Chapter 3.

If the asymptotic theory that underpins HAC tests were to provide a good approximation to actual finite sample behaviour, there would be no need to carry out tests for either autocorrelation or heteroskedasticity. However, there remain other types of departures from regularity assumptions that must be considered. It has been mentioned that it is vital that $E(u_t | \mathbf{x}_t) = 0$. Consequently it is important to test for endogeneity/errors-in-variables using the *Hausman test* and for omitted variables/incorrect functional form using the RESET test; see Hausman (1978) for the former procedure and Ramsey (1969) for the latter check. Both of these tests can be implemented by testing the *significance of a subset of regressors* in an expanded version of (1.44). The unrestricted model for computing RESET statistics is given by (1.27) and the corresponding model for the Hausman test is discussed in Krämer and Sonnberger (1986). The restricted model is (1.44) and the test of the restrictions that yield this model as a special case of the relevant alternative model should be based upon an HAC estimator of the covariance matrix.

The application of the HAC estimator of the covariance matrix in the context of RESET-type procedures is discussed in Godfrey (1987). In such applications, the unrestricted (alternative) model can be written as

$$y_t = \sum_{j=1}^k x_{tj} \beta_j + \sum_{j=1}^q z_{tj} \gamma_j + u_t, \quad (1.50)$$

in which the terms z_{tj} represent the test variables, for example, $z_{tj} = \hat{y}_t^{j+1}$ for the RESET check. When testing the null hypothesis that $\gamma_1 = \dots = \gamma_q = 0$, a HAC estimator of the covariance matrix is to be constructed using OLS residuals, for example, as in (1.47) and (1.48). However, it is not clear at first sight whether the OLS residuals used should be those from the estimation of (1.44) or those from the estimation of (1.50). The former choice leads to the use of *restricted residuals* and the latter choice

to the construction of the HAC using *unrestricted residuals*. Under the null hypothesis, either choice is asymptotically valid.

It is sometimes argued that there may be something to be gained by using the unrestricted residuals if they are closer to the true errors than the restricted residuals when the null model is invalid. However, in the context of RESET-type and Hausman tests, models like (1.50) do not correspond to genuine alternative explanations of the determination of the dependent variable. Instead such relationships are simply artificial (auxiliary) regressions designed to allow the convenient computation of a diagnostic check. The issue of whether to use restricted or unrestricted residuals when calculating robust tests will be discussed in the next section and will be considered in Chapter 3 in the context of *simulation-based tests*.

1.5. Finite-sample properties of asymptotic tests

It has been pointed out that there is rarely precise information available about the form of the error distribution and, in particular, there seems little reason to believe that the errors are Normally distributed. Moreover, even if Normality is assumed, there are many cases in which this assumption does not imply the existence of exact finite sample tests, for example, in cross-section studies in which heteroskedasticity-consistent standard errors are used and in time-series studies in which the regressors include lagged values of the dependent variable. Consequently it seems reasonable to suggest that, in most cases of practical relevance, the standard OLS tests and confidence intervals must be justified by appeal to large sample theory, rather than to exact results. It is, therefore, important to obtain evidence about the usefulness of the approximations derived from asymptotic theory.

Much of the available evidence has been obtained from studies that use simulation experiments. In such studies, computer programs are written that allow the specification of a *data generation process* (DGP), generate many artificial samples of the required size from the DGP, calculate the test statistics of interest for each generated sample and use these sample values to learn about the finite sample distribution of the test statistics; see Davidson and MacKinnon (1993, ch. 21) for a useful discussion. The increasing power and falling price of modern computers make accurate and extensive simulation analysis feasible. This subsection contains results from simulation experiments that indicate that asymptotic theory does not always lead to reliable inferences. It will be argued that there is, therefore, a need to go beyond the use of conventional first-order

asymptotic theory when applying tests in econometrics, despite the convenience of the associated standard distributions (namely, $N(0, 1)$, t , F and χ^2).

Clearly it would not be expected that the predictions from asymptotic theory about, for example, finite sample significance levels would be perfectly accurate. What is of practical importance is how close the asymptotic values are to the actual finite sample values in an interesting variety of situations. Opinions will differ about close an approximation should be to be viewed as useful. In applications in psychology, *criteria for robustness* of tests have been proposed; see, for example, Bradley (1978) and Serlin (2000). The stringent criterion for robustness is that the actual significance level should be in the range $0.9\alpha_d$ to $1.1\alpha_d$, where α_d denotes the desired (asymptotic) significance level. The range $0.8\alpha_d$ to $1.2\alpha_d$ is sometimes regarded as implying a moderate degree of robustness. A (very) liberal criterion is that the actual rejection probability should be in the range $0.5\alpha_d$ to $1.5\alpha_d$. Whatever, the precise definition of robustness, it is obviously important to know if, for sample sizes of interest to empirical researchers, there is evidence that a test using asymptotically valid critical values leads to the rejection of a true null hypothesis either far too infrequently or far too frequently. In the former case, rejections may also be infrequent in the presence of departures from the null that are important, that is, there might be problems of low power. In the latter case, valid models will be discarded far more often than intended.

The results from simulation experiments, which are reported in this subsection, are intended to supplement the predictions of conventional asymptotic and finite sample theory. It is hoped that the results obtained from simulation experiments are of value in practical situations, but this property can never be guaranteed. It is very common to refer to such experiments as Monte Carlo experiments; see Davidson and MacKinnon (1993, ch. 21) for an excellent discussion. However, given that "Monte Carlo" will often be used below to refer to a particular type of simulation-based test, this terminology will not be adopted in order to avoid any confusion.

There are three sets of simulation experiments, each reported in its own subsection. The first subsection contains results about the performance of F -statistics as checks of the joint significance of a subset of regressors and it is found that asymptotic theory provides a fair approximation overall. The second of the subsections provides examples of widely-used tests for which actual significance levels seem likely to be rather smaller than the values predicted by asymptotic theory. The third subsection shows

how, for other tests, asymptotic critical values can lead to finite sample significance levels that are larger than the desired (nominal) values.

1.5.1. Testing the significance of a subset of regressors

The first case to be examined is one in which, after any necessary renumbering, the parameter vector of the regression model can be written as $\beta = (\beta'_1, \beta'_2)'$ and the null hypothesis is $\beta_2 = \mathbf{0}_q$, with $\mathbf{0}_q$ denoting the q -dimensional vector with every element equal to zero. This example is familiar from textbook discussions and, if the assumptions of the classical linear regression model were satisfied, the relevant F -test would be exactly valid. However, suppose that the errors, while IID, are not Normally distributed. As a result, the F -test is only asymptotically valid. White and MacDonald remark that the significance levels of F -tests and t -tests seem robust to non-Normality and give references to previous research on this matter; see White and MacDonald (1980). It seems worthwhile to provide some simulation-based evidence and two small scale experiments will be used.

In the first experiment, data for the estimation of production functions are taken from Greene (2008). These data come from a study of production in the US. There are 27 statewide observations on value added, VA , labour input, L , and capital stock, K . The production function relationship includes an intercept and is specified in logs of the basic economic variables, with $x_{i1} = 1$, $x_{i2} = \log(L_i)$, and $x_{i3} = \log(K_i)$.

In order to obtain evidence on the behaviour of F -tests with $q > 1$ restrictions, the problem of testing the Cobb-Douglas form

$$y_i = \beta_1 + \beta_2 x_{i2} + \beta_3 x_{i3} + u_i, \tag{1.51}$$

against the more general translog model

$$y_i = \beta_1 + \beta_2 x_{i2} + \beta_3 x_{i3} + \beta_4 x_{i2}^2 + \beta_5 x_{i2} x_{i3} + \beta_6 x_{i3}^2 + u_i, \tag{1.52}$$

is considered. For this case, the null hypothesis is $H_{CD} : \beta_4 = \beta_5 = \beta_6 = 0$, so that $q = 3$. The conventional F -test therefore uses critical values taken from the $F(3, 27 - 6)$ distribution. Finite sample significance levels of the F -test under Normal and non-Normal IID errors are estimated by generating data with (1.51) for specified values of the coefficients β_j and error processes that are described below.

Let \mathbf{X}_1 be the 27 by 3 matrix with typical row (x_{i1}, x_{i2}, x_{i3}) and \mathbf{X}_2 be the 27 by 3 matrix with typical row $(x_{i2}^2, x_{i2} x_{i3}, x_{i3}^2)$. The regressor matrix for the unrestricted model, $\mathbf{X} = (\mathbf{X}_1, \mathbf{X}_2)$, is then 27 by 6 with typical

row $(x_{i1}, x_{i2}, x_{i3}, x_{i2}^2, x_{i2}x_{i3}, x_{i3}^2)$. Given $\beta_2 = (\beta_4, \beta_5, \beta_6)' = \mathbf{0}_3$, the F -test statistic is invariant with respect to the value of $\beta_1 = (\beta_1, \beta_2, \beta_3)$ because

$$RSS(H_1) = \hat{\mathbf{u}}' \hat{\mathbf{u}} = \mathbf{u}' \mathbf{M}' \mathbf{M} \mathbf{u} = \mathbf{u}' \mathbf{M} \mathbf{u},$$

and

$$RSS(H_0) = \tilde{\mathbf{u}}' \tilde{\mathbf{u}} = \mathbf{u}' \mathbf{M}'_1 \mathbf{M}_1 \mathbf{u} = \mathbf{u}' \mathbf{M}_1 \mathbf{u},$$

in which $\mathbf{u}' = (u_1, \dots, u_{27})$ and $\mathbf{M}_1 = (\mathbf{I}_{27} - \mathbf{X}_1(\mathbf{X}'_1 \mathbf{X}_1)^{-1} \mathbf{X}'_1)$; see Davidson and MacKinnon (2004, pp. 141–142) for a detailed explanation in a more general setting. Consequently, for the purpose of investigating the significance levels of the F -test under non-Normality, the elements of β can all be set equal to zero, without any loss of generality.

The relevant F -statistic can be expressed as

$$\frac{\mathbf{u}' \mathbf{M}_1 \mathbf{u} - \mathbf{u}' \mathbf{M} \mathbf{u}}{\mathbf{u}' \mathbf{M} \mathbf{u}} \times \frac{21}{3} = \frac{(\mathbf{c}\mathbf{u})' \mathbf{M}_1 (\mathbf{c}\mathbf{u}) - (\mathbf{c}\mathbf{u})' \mathbf{M} (\mathbf{c}\mathbf{u})}{(\mathbf{c}\mathbf{u})' \mathbf{M} (\mathbf{c}\mathbf{u})} \times \frac{21}{3},$$

for any nonzero constant c . It follows that the F -statistic is unaffected by the choice of the error variance in the simulation experiment. The choice can be based simply on convenience, given the set of error distributions to be used.

The distributions that are used to obtain the terms of (u_1, \dots, u_{27}) are: $N(0, 1)$ as a benchmark; $t(5)$ as an example of a symmetric distribution that is far from being Normal; and the heavily-skewed $\chi^2(2)$ distribution. Drawings from these distributions are adjusted, when necessary, so that, without loss of generality, they come from a population with zero mean and variance equal to one. (In fact, for the simple case under consideration, the invariance results given above imply that there is no need to adjust drawings, but this result does not hold in several of the experiments discussed later.)

In order to derive fairly precise estimates of rejection probabilities, 25,000 sets of artificial data with $n = 27$ are generated for each error distribution, that is, the number of *replications* is $R = 25,000$. The regressors values are held fixed over replications, which corresponds to the classical assumption of nonrandom regressors. The results of Ali and Sharma (1996) point to the relevance of leverage values, that is, the diagonal elements of the hat matrix, to the impact of non-Normality on the F -test. High leverage values may lead to a lack of robustness. A leverage value can be viewed as high if it exceeds two or three times the average of the leverage values; see Belsley et al. (1980) for further discussion of leverage

Table 1.1 Estimates of significance levels for F -test of (1.51) against (1.52)

Error distribution	$\alpha_d = 1\%$	2%	3%
Normal	1.01	4.98	9.79
Student $t(5)$	1.35	5.76	10.62
$\chi^2(2)$	2.18	6.51	10.94

Note: $n = 27$

in regression analysis. For tests of H_{CD} , the maximum of the diagonal elements of the unrestricted regression hat matrix is 3.37 times the average, with the corresponding figure being 2.79 for the restricted regression hat matrix. Consequently there are no grounds for suspecting that the experiments will produce over-optimistic evidence about the robustness of F -tests to non-Normality.

The results from the experiments are summarized in Table 1.1, in which estimates of null rejection probabilities are given as percentages and the desired significance levels are the conventional values of 1 per cent, 5 per cent and 10 per cent. As expected, use of a Normal error distribution produces estimates that are very well behaved because the test is exact and the number of replications is quite large. It is also unsurprising that the distortion of significance levels associated with the markedly skewed $\chi^2(2)$ error distribution is greater than that observed with the use of $t(5)$ errors. With a desired significance level of 1 per cent, the estimate for $\chi^2(2)$ errors does not even satisfy the liberal criterion of robustness given above.

It could be objected that the estimates contained in Table 1.1 overstate the robustness of the F -test to non-Normality because the regressors are not random. Given the popularity of *dynamic models*, it seems useful to combine departures from Normality with the inclusion of lagged dependent variables as regressors to provide a more stringent check of the quality of approximation provided by asymptotic theory. A second small-scale simulation experiment is constructed by considering the problem of testing

$$y_t = \beta_1 + \beta_2 x_t + \beta_3 y_{t-1} + \beta_4 x_{t-1} + u_t, \tag{1.53}$$

against

$$y_t = \beta_1 + \beta_2 x_t + \beta_3 y_{t-1} + \beta_4 x_{t-1} + \beta_5 y_{t-2} + \beta_6 x_{t-2} + u_t, \tag{1.54}$$

in which the errors u_t are $\text{IID}(0, \sigma^2)$. The error distributions are the same as are used in the production function example. Drawings from these distributions are transformed to obtain variates ε_t , which have zero mean and unit variance, and then the errors u_t are calculated as $u_t = \sigma \varepsilon_t$, using a specified value of σ .

The exogenous variable x_t is generated according to a first-order autoregressive (AR(1)) scheme with

$$x_t = 0.8x_{t-1} + v_t,$$

in which the terms v_t are $\text{NID}(0, \sigma_v^2)$ and σ_v^2 is selected so that the variance of x_t is 1. The initial value x_1 is taken from the exogenous variable's stationary distribution, that is, $\text{N}(0, 1)$.

The null and alternative specifications are *Autoregressive Distributed Lag* (ADL) models. Parameter values under the null hypothesis $H_{ADL} : \beta_5 = \beta_6 = 0$ are set as follows: $\beta_1 = 1$; $\beta_2 = 0.6$; $\beta_3 = 0.7$; and $\beta_4 = 0.4$. The value of σ^2 is selected by trial and error to give an average value of the R^2 statistic after OLS estimation of (1.53) close to 0.8. The simple invariance results used in the fixed regressor production function experiment cannot be assumed to hold in the experiment with ADL models.

The dynamic structure of the model under the null hypothesis requires that start-up values be provided. The start-up values $y_s, s \leq 0$, are set equal to the unconditional mean of y_t and, given the parameter values, 90 observations are generated, using (1.53). The first 50 of these observations are then discarded to reduce the impact of using fixed start-up values. There remain $n = 40$ observations for estimation and testing $H_{ADL} : \beta_5 = \beta_6 = 0$. The inclusion of lagged dependent variables in the regressor set implies that the F -test is only asymptotically valid, even when the errors are NID. The estimated significance levels, derived from 25,000 replications, are given in Table 1.2. These estimates show that

Table 1.2 Estimates of significance levels for F-test of (1.53) against (1.54)

Error distribution	$\alpha_d = 1\%$	5%	10%
<i>Normal</i>	1.06	5.30	10.56
<i>Student t(5)</i>	1.06	5.08	9.95
$\chi^2(2)$	1.02	4.68	9.70

Note: $n = 40$

there is evidence that the stringent criterion for robustness is satisfied since all are in the range $\alpha_d \pm 0.1\alpha_d$.

Overall, the results from these two small experiments do not suggest major failings are likely to be common when F -tests are implemented in the presence of non-Normal IID errors. Thus the remarks in White and MacDonald (1980) are corroborated by the estimates contained in Tables 1.1 and 1.2.

1.5.2. Testing for non-Normality of the errors

Given the major role played by the Normal distribution in discussions of inference in the context of linear regression models, it is not surprising that tests of the assumption of Normality have been developed; see White and MacDonald (1980) for an important contribution. A test proposed by Jarque and Bera has become widely used and is discussed in many textbooks; see Jarque and Bera (1980, 1987).

The basic idea of the *Jarque-Bera test* is derived from the result that

$$z_t = \frac{u_t}{\sigma} \sim N(0, 1),$$

if the errors u_t are $NID(0, \sigma^2)$. It follows that, under the null hypothesis H_{JB} : *the errors are* $NID(0, \sigma^2)$, the moments of z_t are those of a standard Normal distribution. In particular, when H_{JB} is true, the third and fourth moments are such that $E(z_t^3) = 0$ and $E(z_t^4 - 3) = 0$. If the errors could be observed and had a known variance, then the terms $z_t, t = 1, \dots, n$, could be found and used in an appropriate test in which the researcher investigated the joint significance of

$$n^{-1} \sum_{t=1}^n z_t^3 = n^{-1} \sum_{t=1}^n \left(\frac{u_t^3}{\sigma^3} \right),$$

and

$$n^{-1} \sum_{t=1}^n (z_t^4 - 3) = n^{-1} \sum_{t=1}^n \left(\frac{u_t^4}{\sigma^4} - 3 \right);$$

so that claims about expected values would be tested using the corresponding sample means, as is standard. However, the errors are unobservable and their variance is unknown.

Jarque and Bera show that the effects of replacing u_t by the corresponding OLS residual \hat{u}_t and $\sigma = \sqrt{\sigma^2}$ by the consistent estimator $\sqrt{\hat{\sigma}^2}$

(or $\sqrt{s^2}$) are asymptotically unimportant and that, when H_{JB} is true,

$$JB = n \left[(\sqrt{b_1})^2 / 6 + (b_2 - 3)^2 / 24 \right] \sim_a \chi^2(2), \quad (1.55)$$

in which

$$\sqrt{b_1} = n^{-1} \sum_{t=1}^n \left(\frac{\hat{u}_t^3}{\hat{\sigma}^3} \right),$$

and

$$b_2 = n^{-1} \sum_{t=1}^n \left(\frac{\hat{u}_t^4}{\hat{\sigma}^4} \right);$$

see Jarque and Bera (1987). The null hypothesis of Normality is rejected for significantly large values of the statistic JB defined in (1.55), with the asymptotically valid reference distribution being $\chi^2(2)$. The test statistic JB has become part of standard regression output in econometric programs and so the adequacy of the asymptotic $\chi^2(2)$ distribution is of considerable interest.

Several studies have provided evidence that the asymptotic distribution does not provide an adequate approximation to the finite sample distribution of JB when n is not very large. Jarque and Bera (1987) report poor approximations even with $n = 500$. Further evidence is discussed in Deb and Sefton (1996) and Urzúa (1996). The picture that emerges very clearly is that, if sample values of JB are compared with critical values from the $\chi^2(2)$ distribution, the actual significance levels are much smaller than those predicted by asymptotic theory. For example, Deb and Sefton remark that, when $n = 20$, the actual significance levels of tests using asymptotic critical values for 5 per cent and 10 per cent levels are 2.41 per cent and 3.71 per cent, respectively; see Deb and Sefton (1996, p. 125).

The basic form of the JB statistic (but not its interpretation as a Lagrange Multiplier criterion) was derived in Bowman and Shenton (1975). However, as pointed out in Urzúa (1996, p. 248), its deficiencies had been recognized and summarized as follows in D'Agostino (1986, p. 391): "Due to the slow convergence of b_2 to normality this test is not useful."

Attempts have been made to adjust the test statistic in order to obtain better control of finite sample significance levels. However, as will be seen in Chapter 3, simulation methods make it possible to obtain an exact test of Normality when the regressors are exogenous.

It might be argued that the attempt to derive an asymptotic test of Normality is misplaced because one would be able to rely upon a CLT to justify standard t and F tests if the sample size were large enough to support the use of the test. However, some tests require the assumption of Normality even for asymptotic validity. An important example of such a procedure is the well-known test for heteroskedasticity proposed in Breusch and Pagan (1979). The *Breusch-Pagan test*, like the Jarque-Bera test, is described in many textbooks and is calculated by several programs for least squares estimation of regression models. However, as with the Jarque-Bera procedure, it seems that the actual finite sample significance levels of the Breusch-Pagan test are smaller than those given by asymptotic theory. Estimated rejection rates of the Breusch-Pagan test under the assumption of homoskedasticity are obtained using simulation experiments in Godfrey and Orme (1999). Godfrey and Orme use the conventional values of 1 per cent, 5 per cent and 10 per cent for asymptotic significance levels. They find that the ratio of estimated significance level to the corresponding asymptotic significance level is about 3/10 when $n = 40$ and about 5/10 when $n = 80$. Thus there is strong evidence that this very well-established test for heteroskedasticity will underreject relative to the nominal significance level.

1.5.3. Using heteroskedasticity-robust tests of significance

As remarked above, the use of heteroskedasticity-robust tests is now common and recommended in several leading textbooks. However, such tests rely on asymptotic results for their justification. The production data used in Subsection 1.5.1 are employed in Godfrey and Orme (2002b) to examine the small sample properties of a heteroskedasticity-robust t -test. The unrestricted (alternative) model is the Cobb-Douglas relationship given in (1.51) and the single restriction to be tested is $H_{CR} : \beta_2 + \beta_3 = 1$, that is, there are constant returns to scale. The heteroskedasticity-robust standard error of $(\hat{\beta}_2 + \hat{\beta}_3 - 1)$ can be obtained from the HCCME of White (1980), denoted by HCO , and used to obtain a t -test that is asymptotically valid in the presence of unspecified forms of heteroskedasticity.

Godfrey and Orme use the original data to generate an extended data set of 54 observations by setting

$$x_{i+27,j} = x_{ij} \text{ for } i = 1, \dots, 27 \text{ and } j = 1, 2, 3.$$

They study the behaviour of the HCCME-based test of H_{CR} under homoskedasticity and various forms of heteroskedasticity, with errors being defined by $u_i = \sigma_i \epsilon_i$, in which the terms ϵ_i are independent

drawings from standardized versions of the distributions used in the experiments of Subsection 1.5.1, $i = 1, \dots, 54$. Godfrey and Orme find that, when the asymptotic significance level is 5 per cent, the estimates are in the range 15.60 per cent to 20.25 per cent.

Estimates of the actual significance levels of a heteroskedasticity-robust t -test are also reported in Horowitz (1997). Horowitz carries out simulation experiments with $n = 25$ and not surprisingly finds even stronger evidence of the poorness of the asymptotic approximation than Godfrey and Orme. With a nominal significance level of 5 per cent, Horowitz obtains estimates in the range 15.6 per cent to 44.1 per cent when asymptotically valid critical values are used.

These results are obviously disappointing to applied workers who are interested in using heteroskedasticity-robust procedures. Chapter 6 contains a detailed discussion of tests for heteroskedastic regression models and evidence that a version of the bootstrap can lead to reliable robust tests. However, there is also evidence that asymptotic theory can provide a reasonable approximation to the finite sample significance levels of a heteroskedasticity-robust t -test, provided that the HCCME of (1.40) is modified. The modification is suggested in Davidson and MacKinnon (1985a) and simply involves replacing the diagonal matrix

$$\hat{\Sigma} = \text{diag}(\hat{u}_1^2, \dots, \hat{u}_n^2), \quad (1.56)$$

in which \hat{u}_i denotes a typical OLS residual from the unrestricted (alternative) model, by the diagonal matrix

$$\hat{\Sigma}_R = \text{diag}(\tilde{u}_1^2, \dots, \tilde{u}_n^2), \quad (1.57)$$

in which \tilde{u}_i denotes a typical residual from the estimation of the restricted (null) model, that is, \tilde{u}_i^2 is the square of the i th element of the vector in (1.19).

In the experiments of Davidson and MacKinnon (1985a), the use of HCO calculated as in (1.40) with $\hat{\Sigma}$ produces t -tests that reject true null hypotheses far too frequently. This finding is consistent with those mentioned above. When $\hat{\Sigma}$ is replaced by $\hat{\Sigma}_R$, the implied t -tests no longer suffer from this problem. In fact, Davidson and MacKinnon point out that if there is a problem associated with the use of $\hat{\Sigma}_R$ it is that significance levels may be a little too low when nominal values are small; see Davidson and MacKinnon (1985a, p. 214). For example, in one of the experiments that Davidson and MacKinnon carry out with $n = 50$, asymptotic critical values are used that correspond to the widely-used

significance levels of 10 per cent, 5 per cent and 1 per cent. The estimates from t -tests based upon (1.40) and (1.56), with the unrestricted squared residuals, are 19.22 per cent, 13.34 per cent and 4.70 per cent. These estimates indicate that asymptotic theory gives a very poor approximation, with marked overrejection. In contrast, the estimates from t -tests based upon a modified version of (1.40) using (1.57), that is, with the restricted squared residuals, are 12.77 per cent, 4.94 per cent and 0.68 per cent, indicating that asymptotic theory gives a much better approximation. The marked differences in the behaviour of tests based upon restricted and unrestricted residuals are not predicted by standard asymptotic theory because

$$plim \ n^{-1}X'\dot{\Sigma}_R X = plim \ n^{-1}X'\dot{\Sigma} X = plim \ n^{-1}X'\Sigma X,$$

so that differences are asymptotically negligible.

There are two limitations on the usefulness of this finding about the relative merits of restricted and unrestricted residuals when constructing the HCCME as a tool for inference. First, there is the practical problem that the matrix $\dot{\Sigma}_R$ must be recalculated each time the null hypothesis is changed. For example, it is conventional to look at the individual significance of each of the regressors after OLS estimation. In order to make use of restricted residuals for heteroskedasticity-robust t -tests for this purpose, it would be necessary to estimate the k restricted models (each with $k - 1$ coefficients to be estimated) and each heteroskedasticity-consistent standard error for an estimated coefficient would be based upon a different value of

$$HCO_R = n(X'X)^{-1}X'\dot{\Sigma}_R X(X'X)^{-1},$$

or of the corresponding variant of one of the HCCME defined in (1.41)–(1.43). Second, it has been found that, even with the use of restricted residuals in the HCCME, asymptotic theory does not provide good control of finite sample significance levels when several linear restrictions are being tested; see Godfrey and Orme (2004, p. 286). Fortunately, as will be argued in Chapter 5, there are types of bootstrap procedures that can lead to reliable heteroskedasticity-robust tests of null hypotheses of the general form (1.18) in OLS-based regression analysis.

1.6. Non-standard tests for linear regression models

In the discussion above, it has been assumed that the applied worker is using a test statistic that, when the null hypothesis is true, has a

known distribution, at least asymptotically. In conventional asymptotically valid tests, the sample values of test statistics can be compared with critical values from standard distributions. These standard distributions are: $N(0, 1)$ and t when the null hypothesis imposes one restriction; and χ^2 and F when several restrictions are to be tested. It might be thought odd that critical values from t and F distributions are used in asymptotic theory tests because these critical values depend upon the actual finite sample size, via the term $df(H_1)$. The justification for using $t(df(H_1))$ and $F(q, df(H_1))$ is usually that they give better finite sample approximations than the corresponding limiting forms of $N(0, 1)$ and $\chi^2(q)/q$, respectively; see, for example, Kiviet (1986). However, there are important tests for which asymptotic theory fails to provide a standard reference distribution that allows asymptotically valid inferences. The purpose of this section is to illustrate this failure by considering three important examples of *non-standard tests*.

One of the best known tests for regression models is the test of the hypothesis of constant coefficients described in Chow (1960). The null hypothesis that each of the coefficients β_j is invariant over the n observations is tested against the alternative that is made up of the following assumptions: (i) two sets of coefficients apply; (ii) it is known which set of coefficients is relevant for each observation, so that the researcher can identify two sub-samples, each characterized by its own set of coefficients; and (iii) if the two sub-samples contain, say, n_1 and n_2 ($n_1 + n_2 = n$) observations, the inequalities $n_1 > k$ and $n_2 > k$ are satisfied, so that sub-sample estimation by OLS is feasible. In order to simplify the form of the test, it is usual to make the auxiliary assumptions that the regressors are strictly exogenous and the errors are $NID(0, \sigma^2)$.

The *Chow test* can then be viewed as a test of the classical Normal linear regression model (1.1) against the alternative model

$$y_i = \sum_{j=1}^k x_{ij}\beta_j + \sum_{j=1}^k (d_i x_{ij})\gamma_j + u_i, \quad (1.58)$$

in which d_i is a dummy variable that takes the value 0 for all observations in the sub-sample with n_1 observations and the value 1 for all observations in the sub-sample with n_2 observations. Thus, under the alternative hypothesis, the coefficient vector for the former sub-sample has elements β_j and the coefficient vector for the latter sub-sample has elements $\beta_j + \gamma_j$, $j = 1, \dots, k$. The null hypothesis of constant coefficients can, therefore, be written as $H_{CC} : \gamma_1 = \dots = \gamma_k = 0$. Under the null

hypothesis H_{CC} , the standard F test yields a statistic, denoted by F_{CC} , which is distributed as $F(k, n - 2k)$.

While this textbook version of the Chow test may sometimes be applicable, it has been recognized that it is often the case that there is uncertainty about the break in coefficient values. Suppose that a researcher is using time series data to estimate a regression relationship and there is concern that there was a single change in coefficients but it is not known exactly when it occurred. If the researcher believes that the break was no earlier than period i_0 and no later than period i_1 , $i_1 > i_0$, a modified version of the Chow test can be applied. The researcher simply computes the Chow statistic F_{CC} for each period in the range i_0 to i_1 and bases the test of constant coefficients on the maximum of these test statistics. In Stock and Watson (2007), this procedure is called the Quandt likelihood ratio (QLR) method; see Quandt (1960). In more general settings, this type of check is known as a *Sup-test*; see, for example, Andrews (1993) and the references that it contains.

While, under the null hypothesis, each separate Chow statistic has the $F(k, n - 2k)$ distribution, it is clear that their maximum will not have the same distribution. It is not possible to use one of the standard distributions as the source of asymptotically valid critical values. It is pointed out in Stock and Watson (2007, p. 569) that the large sample null distribution depends upon the number of restrictions being tested (which here equals k) and the fractions i_0/n and i_1/n . Asymptotic critical values have been provided, using simulation, so that tests can be applied; see Andrews (1993, 2003a). However, there is evidence from simulation experiments which indicates that these asymptotic critical values may not be accurate approximations to actual finite sample values; see, for example, Diebold and Chen (1996).

The second example of a non-standard procedure is also based upon a test described in Chow (1960). As well as proposing a test of the claim that all regression coefficients are constant, Chow explains how to carry out a test of the hypothesis that prediction errors have a zero mean. Greene refers to such procedures as *predictive tests*; see Greene (2008, pp. 121–122). Predictive tests are used in many applied studies involving the least squares estimation of a linear regression model, with the number of predictions often being smaller than the number of regressors. Hendry has obtained an asymptotically valid simplification of Chow's test, which he refers to as a test for *predictive failure*; see Hendry (1980).

In a predictive test, estimated parameters derived from an *estimation sample* are used to generate predicted values for a *prediction sample*. Let the former contain $n_1 > k$ observations and the latter contain n_2

observations, with $n_2 < k$ being the usual case. For simplicity of exposition, it is assumed that the first n_1 observations comprise the estimation sample. In the standard Chow test for prediction residuals, the null hypothesis is that (1.1) is valid for $i = 1, \dots, n$ and, under the alternative, this model is only assumed to hold for $i = 1, \dots, n_1$, that is, the estimation sample.

Let $\check{\beta}$ be the OLS estimator derived from the estimation sample. The prediction residuals for the remaining n_2 observations are denoted by

$$\check{e}_{i-n_1} = y_i - \sum_{j=1}^k x_{ij}\check{\beta}_j, i = n_1 + 1, \dots, n, \quad (1.59)$$

in which $n = n_1 + n_2$. Chow proposes testing the joint significance of the prediction residuals of (1.59) using

$$P = \frac{(\hat{\mathbf{u}}'\hat{\mathbf{u}} - \check{\mathbf{u}}'\check{\mathbf{u}})/n_2}{(\check{\mathbf{u}}'\check{\mathbf{u}})/(n_1 - k)}, \quad (1.60)$$

in which $\hat{\mathbf{u}}$ is the n -dimensional OLS residual vector obtained when all observations are used for estimation and $\check{\mathbf{u}}$ is the n_1 -dimensional OLS residual vector when only the estimation sample is used. Under the null hypothesis that the same classical assumptions apply to all n observations, P has the $F(n_2, n_1 - k)$ distribution with large values of this test statistic indicating predictive failure. Thus, when the significance of (1.60) is assessed using right-hand tail critical values of the $F(n_2, n_1 - k)$ distribution, the null model under test includes the assumption that the errors u_i are Normally distributed.

There are alternatives to Chow's test. Hendry has proposed a large sample test for predictive failure that, like Chow's procedure, requires the errors to be Normally distributed. Under Normality, it is asymptotically valid (as $n_1 \rightarrow \infty$, with n_2 fixed) to compare sample values of

$$H = \frac{\sum_{i=1}^{n_2} \check{e}_i^2}{(\check{\mathbf{u}}'\check{\mathbf{u}})/(n_1 - k)} \quad (1.61)$$

to right-hand-tail critical values of the $\chi^2(n_2)$ distribution; see Hendry (1980, p. 222) and Kiviet (1986, section 4).

The assumption of Normality that justifies both the exact test P and the large sample test H is convenient, but it may not provide a very good approximation in practical situations and the effects of non-Normality

merit consideration. The assumption that errors are NID is, therefore, replaced by the weaker assumption that they are IID.

In order to derive asymptotically valid results for the case of IID errors with an unspecified common CDF \mathcal{F} , it is necessary to decide what to assume about the separate behaviour of n_1 and n_2 as their sum n tends to infinity. It is often the case that n_2 is small relative to n_1 and is smaller than k . In order to generate approximations relevant to such cases, Godfrey and Orme carry out an asymptotic analysis in which $n_1 \rightarrow \infty$ and n_2 is fixed; see Godfrey and Orme (2000) for details of regularity conditions.

Under the null hypothesis of model constancy and the assumptions of Godfrey and Orme (2000), Chow's prediction error test statistic P of (1.60) is asymptotically equivalent to

$$P^* = \frac{\sum_{i=1}^{n_2} u_{i+n_1}^2}{n_2 \sigma^2}, \quad (1.62)$$

when $n_1 \rightarrow \infty$ and n_2 is fixed. Equation (1.62) implies that, when the null hypothesis is true, the conventional Chow test statistic P is, under the assumptions of Godfrey and Orme (2000), asymptotically equivalent to the average squared value of the last n_2 elements of the standardized vector $\boldsymbol{\varepsilon} = \sigma^{-1} \mathbf{u}$. Similarly Hendry's test statistic H is, under these assumptions and the null hypothesis, asymptotically equivalent to the sum of the squared values of the last n_2 elements of the standardized vector $\boldsymbol{\varepsilon} = \sigma^{-1} \mathbf{u}$. Thus, when the null hypothesis that the same regression model holds for all n observations is true, P of (1.60) and H of (1.61) both have asymptotic distributions that depend upon \mathcal{F} , the unknown error CDF, but not upon either β or σ^2 .

The final example of the failure of asymptotic theory to provide a convenient reference distribution has its origins in the common practice of reporting a collection of diagnostic checks to accompany the usual results (point estimates, standard errors, goodness of fit measures) after OLS estimation. Many computer programs that are used for regression analysis provide a wide range of standard tests for specification errors and also allow users to set up their own case-specific diagnostics. In these programs, the standard tests are calculated separately for each of the specification errors that is taken into account. In a typical case, these specification errors might include autocorrelation, heteroskedasticity, non-Normality, incorrect functional form and omitted variables. Unfortunately, even if asymptotic theory were to provide a very accurate guide to the actual significance level of each separate test,

there would not be guidance about the precise magnitude of the *overall significance level*.

In order to outline the relevant theory, suppose that the researcher is using a program that gives J separate diagnostic checks. Let the null hypotheses and asymptotic significance levels of the separate misspecification tests be denoted by H_{0j} and π_j , respectively, $j = 1, \dots, J$. If all the null hypotheses H_{0j} under test are true, the overall asymptotic probability of rejecting the claim that H_{01}, H_{02}, \dots and H_{0J} are simultaneously true on the basis of the outcomes of the J separate diagnostic checks is, in general, unknown. However, the results in Darroch and Silvey (1963) imply that it cannot be smaller than $\max(\pi_1, \dots, \pi_J)$ and it cannot be larger than $(\pi_1 + \dots + \pi_J)$. In the special case in which the test statistics are asymptotically independent, it is possible to be more precise, since the overall asymptotic significance level is

$$1 - \prod_{j=1}^J (1 - \pi_j).$$

For example, suppose cross-section production data are being used to estimate the regression equation of (1.51), that is, the model to be checked for misspecification is

$$y_i = \beta_1 + \beta_2 x_{i2} + \beta_3 x_{i3} + u_i, \quad i = 1, \dots, n.$$

Let the OLS predicted values and residuals be, as usual, denoted by \hat{y}_i and \hat{u}_i , respectively, $i = 1, \dots, n$. Next suppose that $J = 3$ diagnostic checks are used. First, a RESET procedure with \hat{y}_i^2 as the only test variable in the artificial model corresponding to (1.27). Second, the Jarque-Bera (1980, 1987) test for non-Normality. Third, a test for heteroskedasticity, as proposed in White (1980), which is derived from the OLS estimation of the artificial regression

$$\hat{u}_i^2 = \gamma_1 + \gamma_2 x_{i2} + \gamma_3 x_{i3} + \gamma_4 x_{i2}^2 + \gamma_5 x_{i2} x_{i3} + \gamma_6 x_{i3}^2 + v_i, \quad (1.63)$$

with the test statistic being the product of the sample size n and the R^2 statistic from (1.63). If each test has an asymptotic significance level of 5 per cent, the general inequality of Darroch and Silvey (1963) implies that the overall asymptotic significance level is between 5 per cent and 15 per cent. The problem for the applied worker who wishes to have some control over the actual overall significance level is, of course, exacerbated by the fact that the asymptotic significance levels of the separate tests

may not be close to the actual values. It has already been noted that asymptotic theory can provide a poor approximation to the finite sample distribution of the Jarque-Bera statistic. There is also evidence that the asymptotic χ^2 distribution of White's test for heteroskedasticity is quite different from the actual distribution under homoskedastic errors unless the sample size is very large; see Jeong and Lee (1999).

It should be noted that, in contrast to the problem of testing for coefficient constancy with an unknown break point, it would not be sensible to try to derive a test by finding the maximum of the separate diagnostic checks. In general, statistics designed to detect different misspecifications correspond to null hypotheses that have different numbers of restrictions. In the example of the previous paragraph, the RESET, Jarque-Bera and White tests have asymptotic reference distributions of $\chi^2(1)$, $\chi^2(2)$ and $\chi^2(5)$, respectively. Consequently, under the assumption of no misspecification, asymptotic theory predicts that, if the sample value of the RESET statistic is 4.000, the probability of obtaining a statistic that is at least as large as the observed value is

$$\Pr(\chi^2(1) \geq 4.000) = 4.55 \text{ per cent.}$$

This result would be regarded as leading to rejection of the null hypothesis of correct specification at nominal significance levels of 5 per cent and 10 per cent since 4.00 must be larger than the corresponding asymptotic critical values. However, if the sample value of the White statistic is $4.351 > 4.000$, the corresponding asymptotic probability is

$$\Pr(\chi^2(5) \geq 4.351) = 50 \text{ per cent, approximately,}$$

indicating that the null hypothesis would not be rejected at any conventional significance level.

In the context of tests with the decision rule being that the null hypothesis is rejected when the sample value of the test statistic is unusually large, the probability of observing a value of the test statistic that is greater than or equal to the sample value is known as the *p-value*. When the actual distribution of the test statistic is approximated by its limiting form, as in the previous paragraph, it is possible to obtain the *asymptotic p-value* as an approximation to the actual *p-value*. Given an observed test statistic, denoted by $\hat{\tau}$, its asymptotic *p-value*, $p^a(\hat{\tau})$, can, in many cases, be obtained from a standard continuous distribution, for example, χ^2 . The null hypothesis is then rejected at a desired (asymptotic) significance level α_d if $p^a(\hat{\tau}) \leq \alpha_d$. A simulation-based (bootstrap) approach to

tackling the problem of controlling the overall significance level when several diagnostic checks are applied is discussed in Chapter 4. As will be seen, this approach involves applying a bootstrap technique to estimates of p -values, rather than to the corresponding observed values of the diagnostic checks.

1.7. Summary and concluding remarks

It is very often the case that applied researchers wish to apply t -tests and F -tests to investigate the validity of null hypotheses that impose restrictions on regression coefficients. This chapter has contained an outline of such tests. It has been emphasized that, in general, they are only exactly valid under the very strong assumptions that the regressors are strictly exogenous and the errors are $NID(0, \sigma^2)$. If, in time series regressions, lagged dependent variables are included as regressors or, more generally, the restrictive assumption of Normality is relaxed, the use of t -tests and F -tests has to be based upon asymptotic theory. Similarly, when, as is now common, checks for misspecification are carried out, the sample values of the diagnostics are compared with critical values that are typically only asymptotically valid.

Moreover, it has been pointed out above that many econometricians now argue that it is inappropriate even to assume that the errors are IID with an unspecified common distribution. Instead it is believed that, whenever it is possible, best practice methods should involve the use of OLS-based tests that are asymptotically robust to autocorrelation and/or heteroskedasticity. Thus, for example, Hansen (1999) urges the use of heteroskedasticity-consistent covariance matrices for OLS point estimators of coefficients, rather than the standard textbook IID-valid estimates, when computing test statistics. The finite sample distributions of these robust test statistics are usually unknown and once again inferences have to be based upon asymptotic theory.

Unfortunately, as illustrated by the examples provided in Section 1.5, asymptotic theory cannot be relied upon to provide an acceptable approximation for all tests of interest. There is evidence that the actual significance levels associated with asymptotic critical values may not be close to the desired (nominal) values. For some tests, actual significance levels appear to be too small, while, for others, estimates of actual significance levels are much greater than the desired values.

Consequently there is a need to look for a better foundation for tests in regression models. This need is underlined by the failure of asymptotic

theory to provide a feasible procedure for some situations of real importance in applied work; see Section 1.6. The remainder of this book is devoted to explaining various simulation-based bootstrap methods that can be used either to improve the finite sample behaviour of existing tests that use asymptotic critical values or to derive feasible asymptotically valid tests when conventional asymptotic theory is not capable of providing such procedures.

2

Simulation-based Tests: Basic Ideas

2.1. Introduction

The merits of tests are usually discussed by considering their behaviour under both null and alternative hypotheses. In the former situation, attention is drawn to the problem of devising a decision rule that permits the probability of rejecting a true null hypothesis to be controlled, at least approximately, for example, (1.21) of Chapter 1. In the latter situation, the probability of detecting a departure from the null hypothesis, that is, the power of the test, is emphasized. Given the decision rule, these probabilities are implied by the sampling distributions of test statistics under null and alternative hypotheses, respectively. As discussed in the previous chapter, the exact form of a sampling distribution under the null hypothesis can sometimes be derived for certain tests, under very restrictive assumptions. However, it is much more common in econometrics to admit that the assumptions required for exact knowledge are not satisfied in many cases of practical relevance and instead to use approximate sampling distributions that are asymptotically valid under relatively weak assumptions. A matter of real concern to the applied worker is then the quality of the asymptotically justified approximation to the sampling distribution of the test statistic that is being calculated.

Suppose that the null hypothesis that is being tested is true. If it were possible to take a very large number of samples, it would be feasible to calculate the test statistic for each of the samples. Given the evidence contained in the large number of calculated test statistics, the adequacy of the approximation based upon asymptotic theory could be investigated. In effect, the researcher would have drawn a large sample of values from the finite sample distribution of the test statistic and, for example, the observed proportion of test statistics deemed “statistically significant”

according to the asymptotically valid decision rule could be compared with the nominal significance level. Unfortunately few econometricians are able to draw a large number of samples. In most situations, just one set of observed data is available, implying a sample of size one from the sampling distribution of the test statistic. However, it has been shown by several leading statisticians that things are not as bleak as they first appear and that simulation-based tests derived by obtaining artificial samples from the observed data are often worth serious consideration; see the references and discussion in, for example, Davison and Hinkley (1997), Efron and Tibshirani (1993) and Hall (1992). The process of making pseudo-samples from the actual sample is sometimes called *resampling*.

The combination of a personal computer and appropriate software makes it possible to simulate playing sports, flying planes and various activities not for those of a nervous disposition. Similarly econometricians are now able to use computers to generate artificial samples, each of which gives an opportunity to apply the simulation world counterpart of the test of interest. Many artificial test statistic values can be found and used in order to obtain potentially relevant information about the sampling distribution of the test statistic obtained from the actual data. The simulation world is based upon the estimated model and so the single body of, say, n actual observations is being used to generate many artificial samples of size n . Despite first appearances that an attempt is being made to get something for nothing, this sort of activity is justifiable in terms of statistical theory; see, e.g., Hall (1992), Mammen (1992) and Mammen and Nandi (2004). The idea of deriving information by resampling the original data in some way has been likened to the attempt to pull oneself up by one's bootstraps and the associated tests are often referred to as *bootstrap tests*; see Efron (1979).

The purpose of this chapter is to describe some of the basic ideas that underpin simulation-based tests. For the sake of exposition, these ideas are first outlined in the context of using a sample of IID variables to test an hypothesis about the mean of their common distribution. Having outlined the basic ideas of simulation-based tests for IID data in Section 2.2, their application to linear regression models with IID errors is considered in Section 2.3. In general, the simulation-based tests, like their more conventional counterparts, are only asymptotically valid. The asymptotic properties of the former are discussed in Section 2.4 and the potential benefits of bootstrapping are examined. Section 2.5 contains remarks on an extension of the original bootstrap approach, which is known as the double bootstrap. The contents of this chapter are summarized in Section 2.6, which also includes some concluding remarks.

2.2. Some key concepts and simple examples of tests for IID variables

It is useful to start by considering a problem for which the solution is known from any basic course in statistics. It is assumed that a simple random sample is available from a Normal population and it is desired to test the claim that the mean of the population equals zero, without any restriction on the variance σ^2 , except that it be finite and positive.

Suppose then that the random variables y_1, y_2, \dots, y_n are Normally, identically and independently distributed. The null hypothesis to be tested is $H_0 : E(y) = 0$. As a starting point, it is assumed that the alternative hypothesis is $H_1^+ : E(y) > 0$. A one-sided test is, therefore, required. Clearly H_0 only determines one of the two parameters that together define a single member of the family of Normal distributions. The parameter σ^2 remains unknown whether or not H_0 is true.

Unknown parameters that are not determined by the null hypothesis are sometimes called *nuisance parameters*. Nuisance parameters are often present in econometric applications because the null hypothesis rarely specifies the values of all of the parameters. A common strategy for dealing with nuisance parameters is to replace these unknown terms by consistent estimators. For the example being discussed, it is very convenient to estimate σ^2 by

$$s_y^2 = \frac{1}{(n-1)} \sum_{i=1}^n (y_i - \bar{y})^2,$$

in which

$$\bar{y} = \frac{1}{n} \sum_{i=1}^n y_i,$$

denotes the sample mean which is unbiased for $E(y)$, whether or not H_0 is true.

It is well known that, whatever the true value of $E(y)$,

$$\frac{\bar{y} - E(y)}{\sqrt{s_y^2/n}} \sim t(n-1).$$

Hence, when $H_0 : E(y) = 0$ is true, the test statistic given by

$$\hat{t} = \frac{\bar{y}}{\sqrt{s_y^2/n}}, \tag{2.1}$$

also has the $t(n - 1)$ distribution. Clearly the sampling distribution of $\hat{\tau}$, when H_0 is true, does not depend upon any unknown parameters. In particular, the exact sampling distribution of $\hat{\tau}$ does not depend upon the value of σ^2 . This distribution does, of course, depend upon n . Given the form of the alternative hypothesis H_1^+ , large values of $\hat{\tau}$ are viewed as providing strong evidence against H_0 .

2.2.1. Monte Carlo tests

A test statistic, like $\hat{\tau}$ of (2.1), that, under its associated null hypothesis, has a distribution that does not depend upon any unknown parameters is called a *pivotal statistic* or a *pivot*. When the test statistic is pivotal, it is possible to obtain an exact test using simulation methods. In such cases, it is usual to refer to the procedures as *Monte Carlo tests*; see Dwass (1957) and Barnard (1963) for early contributions and Dufour and Khalaf (2001) for a discussion of Monte Carlo tests in the context of econometric settings.

Since the value of σ^2 is of no consequence, it is possible to generate a sample of n independent drawings from any Normal distribution with zero mean and the corresponding test statistic will have the same distribution as $\hat{\tau}$ has when $E(y) = 0$. Given a suitable computer program, data for B artificial samples of size n , denoted by y_{bi}^\dagger , can be generated from the standard Normal distribution, that is, $y_{bi}^\dagger \sim N(0, 1)$, $b = 1, \dots, B$ and $i = 1, \dots, n$. The simulation world counterparts of the real world sample mean and variance estimator are given by

$$\bar{y}_b^\dagger = \frac{1}{n} \sum_{i=1}^n y_{bi}^\dagger, b = 1, \dots, B,$$

and

$$s_{yb}^{\dagger 2} = \frac{1}{(n-1)} \sum_{i=1}^n (y_{bi}^\dagger - \bar{y}_b^\dagger)^2, b = 1, \dots, B,$$

respectively. The implied artificial test statistics are

$$\tau_b^\dagger = \frac{\bar{y}_b^\dagger}{\sqrt{s_{yb}^{\dagger 2}/n}}, b = 1, \dots, B, \quad (2.2)$$

and each of them has the $t(n - 1)$ distribution.

Hence, when $H_0 : E(y) = 0$ is true, $\tau_1^\dagger, \dots, \tau_B^\dagger$ form a sample of IID random variables possessing the same finite sample distribution as $\hat{\tau}$. Thus $(\hat{\tau}, \tau_1^\dagger, \dots, \tau_B^\dagger)$ is a simple random sample of $B+1$ random variables, under the null, and the *Monte Carlo p-value* of the observed test statistic is

$$PV_{MC} = \frac{\sum_{b=1}^B \mathbf{1}(\tau_b^\dagger \geq \hat{\tau}) + 1}{B+1}, \quad (2.3)$$

in which $\mathbf{1}(A)$ is the indicator variable that is equal to 1 if the event A is true and is otherwise equal to zero. The Monte Carlo test rejection rule is

$$\text{Reject } H_0 \text{ if } PV_{MC} \leq \alpha, \quad (2.4)$$

with α denoting the required significance level. Given that the alternative hypothesis is $H_1^+ : E(y) > 0$, the rule (2.4) is appropriate since large values of $\hat{\tau}$ imply small values of PV_{MC} . Under regularity conditions provided in Dufour et al. (2004), this rule provides an exact test when $\alpha(B+1)$ is an integer.

The above example has used the one-sided alternative $H_1^+ : E(y) > 0$. Modifications for other alternatives are straightforward. If the alternative hypothesis were $H_1^- : E(y) < 0$, the test statistics would be defined by modified versions of (2.1) and (2.2) in which \bar{y} and \bar{y}_b^\dagger were replaced by $-\bar{y}$ and $-\bar{y}_b^\dagger$, respectively, before calculating PV_{MC} and applying (2.4). This treatment corresponds to that implied by (1.24) and (1.25) in Chapter 1. If the alternative hypothesis were two-sided, that is, $H_1 : E(y) \neq 0$, it would be possible to treat positive and negative values of sample means (observed or artificial) symmetrically. This strategy would lead to the test statistics being defined in modified versions of (2.1) and (2.2) in which \bar{y} and \bar{y}_b^\dagger were replaced by $|\bar{y}|$ and $|\bar{y}_b^\dagger|$, respectively. However, as pointed out in Cox and Hinkley (1974), positive and negative values of the test statistics of (2.1) and (2.2) need not be treated in this way when there is no obvious way of defining what are equally important departures from the null hypothesis. When a symmetric treatment is not imposed, it is possible to proceed as follows: let

$$PV_{MC} = 2 \min \left(\frac{\sum_{b=1}^B \mathbf{1}(\tau_b^\dagger \geq \hat{\tau}) + 1}{B+1}, 1 - \frac{\sum_{b=1}^B \mathbf{1}(\tau_b^\dagger \geq \hat{\tau}) + 1}{B+1} \right),$$

and now apply (2.4); see Cox and Hinkley (1974, p. 79).

Two points should be made about the exact test that has been obtained for the simple example. First, although like the standard t -test for the

above example, the Monte Carlo test gives exactly the desired significance level α , provided regularity conditions are satisfied and $\alpha(B + 1)$ is an integer, there is clearly an important difference between the standard rule that uses a specific critical value from the $t(n - 1)$ distribution and the Monte Carlo rule of (2.4). In the latter approach, there is no fixed rejection region for values of the test statistic. In Marriott (1979), this feature of the Monte Carlo test is called the blurring of the critical (that is, rejection) region. As remarked by Marriott, this blurring effect leads to loss of power but can be reduced by increasing the value of B , subject to the restriction that $\alpha(B + 1)$ is an integer; see Marriott (1979) and Section 2.3.4 below for comments on the choice of the value of B .

Second, it is not essential to assume Normality in applications of Monte Carlo tests. Indeed, when Normality is assumed and suitable tables for the t -distribution are available, there is little point in using simulation-based methods to test $E(y) = 0$. As will be seen below, the key requirements for the valid application of a Monte Carlo test are that the correct distribution, whether it is Normal or not, be specified and that the test statistic be exactly pivotal. However, given the possibility of uncertainty about the correct form of the distribution, it is reasonable to be concerned about the robustness of Monte Carlo tests.

The robustness of Monte Carlo tests to misspecification of the distribution is examined in Godfrey et al. (2006). It is found that a Monte Carlo test will be asymptotically valid when it is derived using an incorrect distribution if it is based on a statistic that is *asymptotically pivotal*. A statistic is said to be asymptotically pivotal (or equivalently, an *asymptotic pivot*) if, when the null hypothesis is true, it has an asymptotic distribution that does not depend upon any unknown parameters. In the definition of this important property, the word “parameters” has a more general meaning than it has in standard introductory econometrics texts and now the parameter vector includes a characterization of the relevant distribution. It is conventional to use the cumulative density function (CDF), denoted by \mathcal{F} , as the parameter to represent the general shape of the distribution.

Many test statistics that are used with regression models are asymptotically pivotal, so that corresponding Monte Carlo tests derived under an incorrect distributional assumption will be asymptotically valid. However, when the test statistic is asymptotically pivotal, applied workers may do better to switch from a Monte Carlo test based upon a choice of distribution, which is very likely to be wrong, and to use instead bootstrap tests that do not impose a specific type of distribution; see Godfrey et al. (2006). If the test statistic of interest is not asymptotically pivotal,

the analysis in Godfrey et al. (2006) indicates that there are even stronger reasons to use a bootstrap approach. The basic ideas of bootstrap tests will now be discussed.

2.2.2. Bootstrap tests

In the Monte Carlo test, the simulation scheme which satisfies the null hypothesis and generates the artificial samples can be constructed without using sample information. For the problem discussed in Section 2.2.1, these samples are obtained using the $N(0, 1)$ distribution, which is not regarded as an approximation to the true null hypothesis distribution of $N(0, \sigma^2)$. When the sample data are used in one way or another to derive the simulation scheme, the artificial samples will be referred to as *bootstrap samples* and the simulation scheme will be called the *bootstrap data generation process* or *bootstrap DGP*. One way in which to introduce a dependence of simulation schemes on actual sample information is to allow for unspecified forms of non-Normality in the example.

Consider regarding the random variables y_1, y_2, \dots, y_n as IID, with mean $E(y)$ and variance σ^2 . Suppose, as seems reasonable, that there is no precise information available that allows the specification of the form of their common distribution. An exact Monte Carlo test of $H_0 : E(y) = 0$ against $H_1^+ : E(y) > 0$ is no longer feasible. An asymptotic test is, however, readily obtained. Under mild restrictions, appeal can be made to a Central Limit Theorem and so

$$\hat{\tau} = \frac{\bar{y}}{\sqrt{s_y^2/n}} \sim_a N(0, 1), \quad (2.5)$$

when the null hypothesis is true. Consequently, in this more general version of the original example, it is asymptotically valid to use critical values from the standard Normal distribution. Given the form of the alternative hypothesis, the null hypothesis is to be rejected when values of $\hat{\tau}$ are judged to be sufficiently large. (The modification of the rejection rule to take account of another form of the alternative hypothesis is straightforward; see the discussion of this issue in the context of Monte Carlo tests above.)

Note that it is only the asymptotic distribution of the test statistic $\hat{\tau}$ that is independent of unknown parameters and so $\hat{\tau}$ is asymptotically pivotal, but is not an exact pivot. The finite sample distribution of $\hat{\tau}$ under the null hypothesis depends, in part, upon the unknown distribution of a typical term y_i . If simulation methods are to be used in an

attempt to get better control of finite sample significance levels than is provided by asymptotic theory, the unknown distribution of each y_i will have to be mimicked in the simulation scheme.

As mentioned above, it is convenient to use \mathcal{F} , the CDF of the IID terms y_i , to represent the unknown distribution parameter. It has also been remarked that nuisance parameters are often replaced by consistent estimators and this strategy was applied in Section 2.2.1 to deal with σ^2 when Normality was assumed. The problem, therefore, is how to apply this strategy when the nuisance parameter is the CDF \mathcal{F} .

Now the CDF \mathcal{F} evaluated at some number c is defined as

$$\mathcal{F}(c) = \Pr(y \leq c),$$

which is estimated consistently by the corresponding sample proportion

$$\frac{1}{n} \sum_{i=1}^n \mathbf{1}(y_i \leq c) = \frac{\#(y_i \leq c)}{n}, \quad (2.6)$$

where $\#(A)$ denotes the number of times that event A occurs. This sample proportion can be reinterpreted as the CDF for an artificial random variable y° , defined conditionally upon the observed data, with

$$\Pr(y^\circ = y_i) = \frac{1}{n}, i = 1, \dots, n, \quad (2.7)$$

since, with this probability distribution,

$$\Pr(y^\circ \leq c) = \mathcal{F}^\circ(c) = \frac{\#(y_i \leq c)}{n},$$

which is often referred to as the *empirical probability distribution* or the *empirical distribution function* (EDF) of the actual data and denoted by $\hat{\mathcal{F}}$; see, for example, the discussions given in Davison and Hinkley (1997, ch. 2) and Efron and Gong (1983).

In view of the above reinterpretation of the sample proportion (2.6), it is tempting to think of using (2.7) to derive artificial data. In the simulation scheme based upon (2.7), each observed value of the real sample is allocated equal probability. Thus artificial samples of size n , that is, $(y_1^\circ, y_2^\circ, \dots, y_n^\circ)$, are obtained by simple random sampling, with replacement, from the original data. However, there is a problem with this artificial data generation process.

The expected value of y° in the simulation world, conditional upon observed data, is, using an obvious notation,

$$E^\circ(y^\circ) = \sum_{i=1}^n y_i \Pr(y^\circ = y_i) = \frac{1}{n} \sum_{i=1}^n y_i = \bar{y},$$

but the purpose is to approximate the behaviour of $\hat{\tau}$ under the null hypothesis, which specifies a zero population mean. In other words, the simulation scheme of (2.7) does not belong to the family of distributions in which the null hypothesis is true.

The adjustment that makes the bootstrap population satisfy the null hypothesis is simple. The original data are *recentred* by subtracting \bar{y} from each value. Given the recentred observed data, B simple random samples of size n , denoted by $[y_{b1}^*, y_{b2}^*, \dots, y_{bn}^*; b = 1, \dots, B]$, can be drawn, with replacement, from the bootstrap probability model defined by

$$\Pr(y^* = y_i - \bar{y}) = \frac{1}{n}, i = 1, \dots, n, \quad (2.8)$$

for which $E^*(y^*) = 0$, where $E^*(\cdot)$ denotes an expectation taken under the *bootstrap law* of (2.8). (The bootstrap world counterparts of the standard items of notation for asymptotic analysis will be written as $o^*(\cdot)$, $O^*(\cdot)$, $o_p^*(\cdot)$, $O_p^*(\cdot)$ and \sim_a^* , when they are required.)

The bootstrap counterparts of the sample mean and variance estimator from actual data are given by

$$\bar{y}_b^* = \frac{1}{n} \sum_{i=1}^n y_{bi}^*, b = 1, \dots, B, \quad (2.9)$$

and

$$s_{yb}^{*2} = \frac{1}{(n-1)} \sum_{i=1}^n (y_{bi}^* - \bar{y}_b^*)^2, b = 1, \dots, B, \quad (2.10)$$

respectively. Similarly the counterparts of $\hat{\tau}$, as given by (2.5), in bootstrap samples are

$$\tau_b^* = \frac{\bar{y}_b^*}{\sqrt{s_{yb}^{*2}/n}} \sim_a^* N(0, 1), b = 1, \dots, B.$$

The values of these bootstrap statistics are now to be used, in place of a critical value from the asymptotic reference distribution of $N(0, 1)$, to

judge the strength of the evidence that $\hat{\tau}$ provides against $H_0 : E(y) = 0$, with $H_1^+ : E(y) > 0$. As in the Monte Carlo approach, a p -value is calculated and compared with the desired significance level. There are differences in opinion about how the *bootstrap p-value* should be calculated and two methods are both quite popular.

One of these methods, described in Davison and Hinkley (1997), is based upon applying a formula like (2.3) to estimate the p -value of an observed test statistic $\hat{\tau}$. Thus, in this approach, the bootstrap p -value of $\hat{\tau}$ is computed using

$$\tilde{p} = \frac{\#\{\tau_b^* \geq \hat{\tau}\} + 1}{B + 1}, \quad (2.11)$$

where the terms τ_b^* , $b = 1, \dots, B$, are the bootstrap realizations of the test statistic. The second approach, given in Efron and Tibshirani (1993), does not use the same formula as a Monte Carlo test and estimated bootstrap p -values are instead obtained using

$$\hat{p} = \frac{\#\{\tau_b^* \geq \hat{\tau}\}}{B}. \quad (2.12)$$

It is not clear why the same formula should be used for Monte Carlo and bootstrap tests. As stressed previously, if the test statistic being considered were to have a finite sample null distribution that did not depend upon unknown parameters, it would be possible to obtain exact Monte Carlo tests. In this special case, $\hat{\tau}$ and τ_b^* , $b = 1, \dots, B$, would constitute $B + 1$ independent drawings from the same sampling distribution, under the null hypothesis, and using (2.11) with the critical region $\tilde{p} \leq \alpha$ would give an exact test, given standard conditions; see, for example, Dufour et al. (2004). However, the test statistics do not have this convenient finite sample property when a bootstrap approach has to be adopted because the distribution of y is unspecified.

In the bootstrap tests based upon comparing $\hat{\tau}$ with τ_b^* , $b = 1, \dots, B$, an artificial bootstrap world is constructed, conditional on the observed data, in order to approximate the finite sample null distribution of test statistics that are only asymptotically pivotal. The use of \hat{p} of (2.12) reflects the conditioning on the observed test statistic, with the B bootstrap statistics being used to obtain the classical sample proportion estimator, under the bootstrap law, which provides the approximation to the true p -value. It is remarked in Broman and Caffo (2003) that “Evaluating p -value estimates conditionally on the observed data is widely accepted” and that “it is immaterial whether one uses \hat{p} or \tilde{p} ”. Given

that, in empirical applications, B is likely to be 1,000 or 2,000 and that

$$0 \leq \bar{p} - \hat{p} \leq \frac{1}{B+1}, \quad (2.13)$$

it seems difficult to disagree with the latter comment.

The formula for the bootstrap p -value is not the only source of differences between statisticians working on bootstrap methods. While there is general agreement that it is important to carry out resampling in a way that reflects the null hypothesis, there are differences about how to achieve this aim. In the discussion above, the same null hypothesis is tested using actual and bootstrap data, with the latter being drawn with replacement from recentred versions of the former. In Hall and Wilson (1991), it is proposed that the bootstrap data should be obtained using (2.7), not (2.8), and that the null hypothesis tested in the (conditional) bootstrap world should be $H_0^\circ : E^\circ(y^\circ) = \bar{y}$, not the restriction $E^\circ(y^\circ) = 0$. For the simple example under consideration, the two methods are exactly equivalent; see the comments in Tibshirani (1992) and the replies in Hall and Wilson (1992). In what follows, the approach adopted will be to use the same null hypothesis for tests applied to actual and bootstrap data and, when appropriate, to recentre the terms which are to be resampled to generate the required bootstrap samples. Thus the first golden rule of bootstrap hypothesis testing given in Davidson (2007) is followed, with the bootstrap DGP belonging to a general model in which the null hypothesis is true.

What is important is that bootstrap hypothesis testing should not be implemented using either the combination of (2.7) and the false restriction $E^\circ(y^\circ) = 0$ as the bootstrap world null hypothesis, or the combination of (2.8) and the false restriction $E^*(y^*) = \bar{y}$ as the bootstrap world null hypothesis. The point is clear for control of significance levels and several authors point out that using inappropriate combinations can have a serious impact on power; see, for example, Hall and Wilson (1991).

There are other guidelines and golden rules that have been discussed for bootstrap hypothesis testing. As will be seen, some key results are based upon the assumption that nuisance parameters are estimated consistently. It would be possible to require that estimators for nuisance parameters should be selected by considering not only consistency, but also asymptotic variances. Davidson suggests that it is desirable for the bootstrap DGP to be based upon estimators that are asymptotically efficient when the null hypothesis is true; see Davidson (2007).

Also, there seems to be a widespread (but not unanimous) agreement that, if possible, a test statistic that is asymptotically pivotal should be bootstrapped. Thus, for the above example, it is $\hat{\tau}$ of (2.5), not $\sqrt{n}\bar{y}$, that is bootstrapped. The statistic $\sqrt{n}\bar{y}$ could be used in a bootstrap test but it is not asymptotically pivotal since

$$\sqrt{n}\bar{y} \sim_a N(0, \sigma^2), \quad (2.14)$$

under the null hypothesis, and the limit distribution in (2.14) depends upon the unknown parameter σ^2 . The benefits associated with the use of asymptotically pivotal statistics are discussed in, for example, Beran (1988) and Hall and Titterton (1989). The general results will be outlined below.

One final definition can be provided before moving on to consider the application of bootstrap methods to tests for regression models. The bootstrap schemes of (2.7) and (2.8) are derived from the empirical distribution functions of uncentred and recentred actual observations, respectively, and are called *nonparametric bootstraps*. In some econometric models, the general form of the relevant CDF is assumed but there are nuisance parameters that require estimation before a bootstrap scheme can be obtained. An important group of such models consists of the standard microeconomic specifications that are estimated by MLE, for example, logit, probit and Tobit. When the bootstrap DGP is derived by combining an assumed family of distributions with estimates of nuisance parameters, the resampling scheme is called a *parametric bootstrap*.

2.3. Simulation-based tests for regression models

As in Chapter 1, the discussion will start with the classical Normal regression model and then the strong assumptions of this model will be relaxed. For each of the sequence of regression models considered, bootstrap techniques are examined. For the main part, only general issues are discussed and specific examples of bootstrap tests for regression models are the subject matter of later chapters.

2.3.1. The classical Normal model

The regressors are assumed to be strictly exogenous, which is a minor modification of the classical assumptions in which these variables are taken to be nonrandom. The relevant conditional results for the model

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{u} \quad (2.15)$$

are then

$$\mathbf{u}|\mathbf{X} \sim N(\mathbf{0}_n, \sigma^2 \mathbf{I}_n), \quad (2.16)$$

and

$$\mathbf{y}|\mathbf{X} \sim N(\mathbf{X}\boldsymbol{\beta}, \sigma^2 \mathbf{I}_n). \quad (2.17)$$

As explained in Chapter 1, there are several tests that are exactly valid under these strong assumptions about the regressors and error terms. However, even in this restrictive framework, other widely-used tests are only asymptotically valid, e.g., the Lagrange Multiplier (LM) tests of Breusch (1978), Breusch and Pagan (1979) and Godfrey (1978). Hence, there is still some incentive for considering the use of resampling methods to improve the approximation to finite sample distributions that is provided by asymptotic theory. Artificial data which are derived under the assumption that the error distribution is known, apart from certain parameters, are denoted by \dagger .

Let \mathbf{S} denote the sample data on the regressand and regressors of (2.15). In view of the assumption of conditional Normality, B bootstrap samples of size n can be obtained, given \mathbf{S} , from the bootstrap DGP of

$$\mathbf{y}^\dagger|\mathbf{S} \sim N(\mathbf{X}\hat{\boldsymbol{\beta}}, s^2 \mathbf{I}_n), \quad (2.18)$$

in which $\hat{\boldsymbol{\beta}}$ and s^2 are the usual OLS estimators of $\boldsymbol{\beta}$ and σ^2 , respectively. More explicitly, a typical bootstrap sample consists of the n artificial observations

$$y_{bi}^\dagger = \hat{y}_i + u_{bi}^\dagger, \quad i = 1, \dots, n, \quad (2.19)$$

in which \hat{y}_i is a typical predicted value from the OLS estimation of (2.15) and the terms u_{bi}^\dagger are n independent drawings from the $N(0, s^2)$ distribution; $b = 1, \dots, B$ and $i = 1, \dots, n$. There are many programs that allow the latter terms to be obtained for a specified value of s^2 . The scheme given in (2.18) is an example of a parametric bootstrap DGP.

In order to illustrate how the parametric bootstrap might be used to improve on reliance on conventional asymptotic theory, consider the problem of testing the assumption of homoskedasticity against the alternative that variances are determined by

$$\text{Var}(u_i) = \sigma_i^2 = \exp\left(\gamma_0 + \sum_{j=1}^q z_{ij}\gamma_j\right), \quad i = 1, \dots, n,$$

in which the terms z_{ij} are observations on strictly exogenous variables that satisfy the regularity conditions of Breusch and Pagan (1979). The commonly used LM statistic tests the q restrictions of $H_0 : \gamma_1 = \dots = \gamma_q = 0$, which imply $\sigma_i^2 = \exp(\gamma_0) = \sigma^2, i = 1, \dots, n$.

The test statistic, denoted by $\hat{\tau}_{BP}$, is one-half of the explained sum of squares from the OLS estimation of the artificial regression

$$\frac{\hat{u}_i^2}{\hat{\sigma}^2} = \gamma_0 + \sum_{j=1}^q z_{ij}\gamma_j + \text{residual}. \quad (2.20)$$

If the null hypothesis is true, $\hat{\tau}_{BP}$ is asymptotically distributed as $\chi^2(q)$, with the rejection region being in the right-hand side of this reference distribution. Thus the asymptotically valid rejection rule can be written, in terms of *asymptotic p-values*, as

$$\text{Reject } H_0 \text{ if } \Pr(\chi^2(q) \geq \hat{\tau}_{BP}) \leq \alpha, \quad (2.21)$$

in which α is the nominal significance level. (The effects of replacing $\hat{\sigma}^2 = [(n-k)/n]s^2$ by s^2 in (2.20) are asymptotically irrelevant.)

As an alternative to using the asymptotic reference distribution of $\chi^2(q)$, consider computing B artificial values of the Breusch-Pagan statistic, denoted by $\hat{\tau}_{BPb}^\dagger, b = 1, \dots, B$, and using the empirical distribution function (EDF) of these statistics to assess the statistical significance of the actual value of $\hat{\tau}_{BP}$. For each value of b , the vector of n observations of \mathbf{y}_b^\dagger , with typical element y_{ib}^\dagger given by (2.19), is used in an OLS regression with regressor matrix \mathbf{X} to obtain a residual vector $\hat{\mathbf{u}}_b^\dagger$, with typical element denoted by \hat{u}_{bi}^\dagger . Given the latter vector, the required variance estimate can be computed as

$$\hat{\sigma}_b^{2\dagger} = n^{-1}(\hat{\mathbf{u}}_b^\dagger)'(\hat{\mathbf{u}}_b^\dagger).$$

The bootstrap counterpart of the artificial regression (2.20) can be estimated by OLS to obtain the corresponding Breusch-Pagan statistic $\hat{\tau}_{BPb}^\dagger$ for $b = 1, \dots, B$.

The bootstrap p -value \hat{p} of (2.12) can then be derived, with \hat{p} being the proportion of bootstrap Breusch-Pagan statistics that are greater than or equal to the Breusch-Pagan statistic that was calculated from the actual data. The rule based upon the asymptotic p -value, that is, (2.21), is now replaced by a rule that uses the bootstrap estimate of the p -value, that is,

$$\text{Reject } H_0 \text{ if } \hat{p} \leq \alpha. \quad (2.22)$$

The bootstrap version of the Breusch-Pagan test is still only asymptotically valid. For reasons that are discussed later, there are grounds for believing that it gives a better approximation to finite sample significance levels than the well-established use of asymptotic critical values from a χ^2 distribution. However, there is an alternative approach that can give perfect control of finite sample significance levels, i.e., an exact test, under the classical assumptions.

The Breusch-Pagan statistic $\hat{\tau}_{BP}$, like many other statistics, is a function of the OLS residuals, which are the elements of

$$\hat{\mathbf{u}} = (\mathbf{I}_n - \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}')\mathbf{u},$$

and observations on test variables. Let \mathbf{Z} denote the $n \times q$ matrix of observations on test variables that are used in (2.20). The Breusch-Pagan statistic can then be written as

$$\hat{\tau}_{BP} = h(\mathbf{u}, \mathbf{X}, \mathbf{Z}),$$

and, given the form of the dependent variable in (2.20), it is clear that

$$\hat{\tau}_{BP} = h(\mathbf{u}, \mathbf{X}, \mathbf{Z}) = h(c\mathbf{u}, \mathbf{X}, \mathbf{Z}), \quad (2.23)$$

for any constant $c > 0$. It follows that, when \mathbf{X} and \mathbf{Z} are both strictly exogenous and the errors of (2.15) are, under H_0 , $\text{NID}(0, \sigma^2)$ variates, a Monte Carlo test is easily obtained. First, choose B so that $\alpha(B+1)$ is an integer (so $B = 99$ would be valid, given standard choices for the required significance level). Next, setting $c = \sigma^{-1}$ has no impact in view of (2.23) and so simulation values of the Breusch-Pagan statistic with the same finite sample distribution as $\hat{\tau}_{BP}$, under H_0 , can be derived using

$$\hat{\tau}_{BPb}^\dagger = h(\mathbf{u}_b^\dagger, \mathbf{X}, \mathbf{Z}), \mathbf{u}_b^\dagger \sim N(\mathbf{0}_n, \mathbf{I}_n), b = 1, \dots, B. \quad (2.24)$$

The n independent drawings from a standard Normal distribution that make up a typical vector \mathbf{u}_b^\dagger can be obtained from a number of widely available computer programs.

In fairness to proponents of Monte Carlo tests, it should be pointed out there is no need to assume Normality in order to derive the Monte Carlo test of the assumption of homoskedasticity. Suppose that the errors of (2.15) are $\text{IID}(0, \sigma^2)$ and have a known non-Normal common distribution. Also suppose that a computer program can be used to obtain

independent drawings from this distribution; there are many *random number generators* that cover a large number of standard distributions. If (2.16) is replaced by

$$\mathbf{u}|\mathbf{X} \sim D_0(\mathbf{0}_n, \sigma^2 \mathbf{I}_n),$$

in which D_0 denotes the known non-Normal distribution, it is only necessary to replace (2.24) by

$$\hat{\tau}_{BPb}^\dagger = h(\mathbf{u}_b^\dagger, \mathbf{X}, \mathbf{Z}), \mathbf{u}_b^\dagger \sim D_0(\mathbf{0}_n, \mathbf{I}_n), b = 1, \dots, B.$$

Provided that D_0 is the correct choice for the error distribution, a Monte Carlo p -value can be calculated as in (2.3) and an exact test obtained using (2.4), provided regularity conditions are satisfied and $\alpha(B+1)$ is an integer; see, for example, Dufour et al. (2004).

However, it is not clear that applied workers will have information that permits the specification of the correct error distribution, whether or not it is the Normal distribution. Consequently it seems reasonable to acknowledge that, in practice, there will be limited opportunity for the application of exact Monte Carlo tests, except when the null hypothesis specifies a particular form for the CDF of the error distribution; see Neumeyer et al. (2004) for a discussion of tests of null hypotheses that specify the parametric form of the CDF. Tests that are instead based upon nonparametric bootstrap methods are now considered.

2.3.2. Models with IID errors from an unspecified distribution

As in the previous subsection, the linear regression function $\mathbf{X}\boldsymbol{\beta}$ of (2.15) is regarded as the conditional mean function $E(\mathbf{y}|\mathbf{X})$; so that $E(\mathbf{u}|\mathbf{X}) = \mathbf{0}_n$. Also, the regressors of (2.15) are again assumed to be strictly exogenous. However, the restrictive assumption of Normality is now abandoned. Instead it is assumed that the error terms in \mathbf{u} are, conditionally and unconditionally, IID with zero mean and finite variance σ^2 . The general family to which the common distribution of the errors belongs is not assumed to be known. Thus (2.16) is replaced by the weaker assumption that

$$\mathbf{u}|\mathbf{X} \sim D(\mathbf{0}_n, \sigma^2 \mathbf{I}_n), \tag{2.25}$$

with the CDF for the unspecified distribution $D(.,.)$ being treated as a parameter and denoted by \mathcal{F} . The unknown parameter vector $\boldsymbol{\theta}$ now contains \mathcal{F} , as well as the elements of $\boldsymbol{\beta}$, that is, $\boldsymbol{\theta} = (\boldsymbol{\beta}', \mathcal{F})$.

In order to set up a bootstrap world, it is necessary to use the actual data to derive a consistent estimator of θ . Conditional upon the actual observations, the artificial DGP that is characterized by the sample value of this estimator can then be used to produce as many bootstrap samples as are required. Observations derived from the artificial DGP are denoted by $*$.

For the moment, the emphasis on tests will be put to one side and instead it is assumed that the researcher simply wants to apply the bootstrap to simulate the distribution of OLS estimators. In the same way that the assumed model for actual data is defined by the parameter vector $\theta = (\beta', \mathcal{F})$, the bootstrap world is characterized by $\check{\theta} = (\check{\beta}', \check{\mathcal{F}})$, with $\check{\theta} - \theta$ being $O_p(n^{-1/2})$. It is assumed that $\hat{\beta}$ of (1.6) is used as the consistent estimator of β . It remains to choose a consistent estimator of \mathcal{F} . As in the simpler example of Section 2.2, for any constant c , the proportion of errors less than or equal to c , that is,

$$\frac{\#(u_i \leq c)}{n}, \quad (2.26)$$

has probability limit equal to $\mathcal{F}(c)$. Thus, if the errors were observed, the EDF of (2.26) would be a consistent estimator of \mathcal{F} . In the absence of observations on errors, it is natural to think of using the corresponding OLS residuals, given by (1.11). If the EDF of the residuals is adopted, the estimator $\hat{\mathcal{F}}$ is determined according to

$$\hat{\mathcal{F}}(c) = \frac{\#(\hat{u}_i \leq c)}{n}.$$

This estimator is often written in the form

$$\hat{\mathcal{F}} : \text{probability } \frac{1}{n} \text{ on } \hat{u}_i, i = 1, \dots, n. \quad (2.27)$$

With the parameter vector θ of the model for actual data estimated by $\hat{\beta}$ of (1.6) and $\hat{\mathcal{F}}$ of (2.27), the artificial data for bootstrap sample b are generated using

$$y_{bi}^* = \hat{y}_i + u_{bi}^*, i = 1, \dots, n, \quad (2.28)$$

in which \hat{y}_i is a typical predicted value from the OLS estimation of (2.15) and the terms u_{bi}^* are sampled randomly, with replacement, using (2.27), $i = 1, \dots, n$. Thus the actual residuals are resampled to obtain bootstrap world errors. However, this scheme is only appropriate if the model for

the bootstrap data matches a key assumption of the model for actual data. This assumption is that $E(\mathbf{u}) = E(\mathbf{u}|\mathbf{X}) = \mathbf{0}_n$. The corresponding expectation for a typical bootstrap error is

$$E^*(u^*) = \sum_{i=1}^n \hat{u}_i \Pr(u^* = \hat{u}_i) = \frac{1}{n} \sum_{i=1}^n \hat{u}_i = \mu^*, \text{ say.} \quad (2.29)$$

When the regression model includes an intercept term, as is usually the case, the first-order conditions for OLS imply that $\mu^* = 0$; so that no adjustment is required to the bootstrap scheme that uses (2.27) and (2.28). If, however, there is no intercept, the OLS residuals will, for almost all samples, sum to a nonzero value and the OLS residuals must be recentered before being used in (2.27). Thus, when there is no intercept term in (2.15), (2.27) is replaced by

$$\hat{F}: \text{probability } \frac{1}{n} \text{ on } (\hat{u}_i - \frac{1}{n} \sum_{j=1}^n \hat{u}_j), i = 1, \dots, n. \quad (2.30)$$

It will assumed from now on that an intercept is included in the vector of regression coefficients and that the adjustment given in (2.30) is unnecessary. However, it should be noted that other modifications of the OLS residuals are sometimes suggested.

One common adjustment is based upon consideration of the variance of the bootstrap errors of (2.27). This variance is

$$E^*(u^{*2}) = \sum_{i=1}^n \hat{u}_i^2 \Pr(u^* = \hat{u}_i) = \frac{1}{n} \sum_{i=1}^n \hat{u}_i^2 = \hat{\sigma}^2,$$

as defined in (1.17) in Chapter 1. Under the probability model for the actual data, $\hat{\sigma}^2$ is consistent for σ^2 but it is not the standard unbiased estimator s^2 of (1.16). A simple adjustment based upon the number of degrees of freedom implies that the bootstrap model variance is s^2 . More precisely, the bootstrap error distribution is now given by

$$\hat{F}: \text{probability } \frac{1}{n} \text{ on } \sqrt{\frac{n}{(n-k)}} \hat{u}_i, i = 1, \dots, n, \quad (2.31)$$

rather than (2.27). Efron and Tibshirani suggest that the modification used in (2.31) is only likely to be important if $k > 0.25n$; see Efron and Tibshirani (1993, p. 112). Others express a stronger preference for (2.31).

A second adjustment of OLS residuals is based upon consideration of the variances of OLS residuals, under the probability model for the actual data. As in (1.13) of Chapter 1, the variance of a typical OLS residual is

$$\text{Var}(\hat{u}_j|X) = \sigma^2(1 - h_{jj}),$$

in which h_{ii} is a leverage value, being the i^{th} diagonal element of H of (1.9). Some argue that the OLS residuals are more like the corresponding errors if they are replaced by transformed versions that are homoskedastic. The transformation is to divide by $\sqrt{(1 - h_{ii})}$; but the transformed residuals do not sum to zero and so recentered values must be obtained for resampling. The effect of combining actual-world variance and bootstrap-world mean adjustments is to have

$$\hat{F} : \text{probability } \frac{1}{n} \text{ on } \left(\frac{\hat{u}_i}{\sqrt{(1 - h_{ii})}} - \frac{1}{n} \sum_{j=1}^n \frac{\hat{u}_j}{\sqrt{(1 - h_{jj})}} \right), i = 1, \dots, n, \quad (2.32)$$

as the bootstrap world CDF.

The evidence reported below, like that obtained in some other studies, suggests that the choice of \hat{F} from (2.27), (2.31) and (2.32) does not have a major impact on the finite sample behaviour of bootstrap tests for regression models. However, it may sometimes be valuable to use adjusted OLS residuals for resampling when k/n is not small and/or some of the leverage values are very large relative to others.

When tests are carried out, there are two methods available for calculating parameter estimates and residuals. The imposition of the null hypothesis leads to a restricted estimator and associated residuals, whereas, under the alternative, the unrestricted estimator produces its own set of residuals. For example, when the null hypothesis consists of the q linear restrictions of (1.18), the restricted estimator of β is $\tilde{\beta}$ and the restricted residuals \tilde{u}_i are the elements of $\tilde{\mathbf{u}} = \mathbf{y} - \mathbf{X}\tilde{\beta}$. The restricted estimator $\tilde{\beta}$ always satisfies (1.18) but the unrestricted OLS estimator $\hat{\beta}$ does not, since $\mathbf{R}\hat{\beta} \neq \mathbf{r}$ for almost all samples. Consequently, for the bootstrap DGP to satisfy the null hypothesis, that is, for the first golden rule of bootstrap tests to be followed, the bootstrap model can be defined using $\tilde{\beta}$ and the EDF of restricted residuals \tilde{u}_i . An appropriate scheme can, therefore, be written as

$$y_{bi}^* = \tilde{y}_i + u_{bi}^*, b = 1, \dots, B \text{ and } i = 1, \dots, n, \quad (2.33)$$

in which \tilde{y}_i is a typical predicted value from restricted estimation and, assuming that the restricted model has an unknown intercept term, the bootstrap errors u_{bi}^* are obtained by random sampling, with replacement, from

$$\tilde{\mathcal{F}} : \text{probability } \frac{1}{n} \text{ on } \tilde{u}_i, i = 1, \dots, n. \quad (2.34)$$

(Details of the counterparts of (2.31) and (2.32) are omitted to focus on (2.34).)

The discussion of the simple example of Section 2.2 suggests an alternative to using $\hat{\beta}$ and $\tilde{\mathcal{F}}$ when bootstrapping, say, the F -test of $H_0 : \mathbf{R}\beta = \mathbf{r}$. In the simple example, it was pointed out that, rather than making the bootstrap model satisfy the restriction placed on the actual model by the null hypothesis, a test could be obtained by changing the null hypothesis. Applying a similar strategy would lead to using artificial data generated from a bootstrap DGP with parameters $\hat{\beta}$ and $\tilde{\mathcal{F}}$ to obtain bootstrap F -tests of $H_0 : \mathbf{R}\beta = \mathbf{R}\hat{\beta}$.

Comparisons of bootstrap tests derived from artificial DGPs defined by restricted and unrestricted estimates are provided in the next chapter. However, at this stage, it is worth noting a finding that emerges from a study of bootstrap tests in a simple regression model; see Paparoditis and Politis (2005). A simple regression with IID errors can be written as

$$y_i = \beta_1 + \beta_2 x_i + u_i, i = 1, \dots, n.$$

Paparoditis and Politis consider testing $H_0 : \beta_2 = 0$ against $H_1 : \beta_2 \neq 0$. As well as examining the usual t -ratio $\hat{\beta}_2/SE(\hat{\beta}_2)$, the test statistic $\sqrt{n}\hat{\beta}_2$ is considered. The former statistic is asymptotically pivotal, with a limit null distribution of $N(0, 1)$. The latter statistic is not asymptotically pivotal, and, as $n \rightarrow \infty$, tends to $N(0, \text{var}(\sqrt{n}\hat{\beta}_2))$, when $\beta_2 = 0$, with $\text{var}(\sqrt{n}\hat{\beta}_2)$ depending on the unknown parameter σ^2 .

When the t -ratio is used, Paparoditis and Politis find that performance is quite robust to the choice of restricted or unrestricted residuals for resampling. However, when the asymptotically non-pivotal test using $\sqrt{n}\hat{\beta}_2$ is investigated, it is found that resampling unrestricted residuals leads to an increase in power compared with resampling restricted residuals. However, the practical relevance of this finding for those who wish to conduct tests of restrictions on regression coefficients may be limited. It is standard practice to use asymptotically pivotal statistics, for example, t and F , to test such restrictions and so the choice between restricted and unrestricted residuals may be of little consequence.

The emphasis so far has been on specifying a bootstrap DGP that can be viewed as consisting of two parts; see, for example, (2.33). First, there is a conditional mean function that has components, equal to corresponding actual predicted values, which are fixed in repeated sampling in the bootstrap world. Second, random bootstrap errors are added to the conditional means, with the former being drawn randomly, with replacement, from a set of (possibly modified) actual residuals, each of which is assigned probability $1/n$ in the bootstrap world. This approach is in keeping with the interpretation of a regression model as representing the conditional distribution of the dependent variable, given the values of strictly exogenous regressors. It also imposes the required IID structure on errors. There is, however, an alternative approach to generating bootstrap data from the observed data.

In the alternative to the residual resampling method of (2.28) and (2.33), it is the terms (y_i, \mathbf{x}'_i) that are randomly resampled, with replacement, from

$$[(y_1, \mathbf{x}'_1), (y_2, \mathbf{x}'_2), \dots, (y_n, \mathbf{x}'_n)].$$

This approach is known as the *pairs bootstrap*. It is argued in Hall (1992, section 4.3.2) that, while such an approach is appropriate when correlation analyses for multivariate data are considered, the residual resampling scheme is preferred in the context of regression modelling with IID errors. Indeed, as will be explained in Chapter 5, the pairs bootstrap does not impose homoskedasticity and allows for heteroskedasticity of unspecified form. Several authors have argued against the use of the pairs bootstrap and for the use of residual resampling when attempting to bootstrap tests for regression models with IID errors; see, for example, Davidson and MacKinnon (2006, pp. 822, 835), Davison and Hinkley (1997, p. 264), Hall (1992, p. 170) and McQuarrie and Tsai (1998, p. 265). The use of the pairs bootstrap for regression models with heteroskedastic errors will be discussed in Chapters 5 and 6.

2.3.3. Dynamic regression models and bootstrap schemes

It was remarked in Chapter 1 that applied workers sometimes use lagged values of the dependent variable as regressors. In such situations, the regression equation is said to be a *dynamic model*. For the purpose of asymptotic analysis, the assumption of strictly exogenous regressors must be replaced by the assumption that all regressors are predetermined.

Using t as the subscript for time series observations, a dynamic regression model can be written as

$$y_t = \mathbf{y}'_{t(p)}\boldsymbol{\alpha} + \mathbf{x}'_t\boldsymbol{\beta} + u_t = \mathbf{w}'_t\boldsymbol{\gamma} + u_t, t = 1, \dots, n, \quad (2.35)$$

in which: $\mathbf{y}'_{t(p)} = (y_{t-1}, \dots, y_{t-p})$, $p \geq 1$, contains the lagged values of the dependent variable; $\boldsymbol{\alpha}' = (\alpha_1, \dots, \alpha_p)$ has elements such that the roots of

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

are all strictly inside the unit circle (for dynamic stability); \mathbf{x}_t is the k -dimensional vector holding a typical observation on the strictly exogenous regressors of the model; $\boldsymbol{\beta}$ is a k -dimensional vector of coefficients; $\mathbf{w}'_t = (\mathbf{y}'_{t(p)}, \mathbf{x}'_t)$; $\boldsymbol{\gamma}' = (\boldsymbol{\alpha}', \boldsymbol{\beta}')$; and the errors u_t are IID(0, σ^2). Let \mathcal{F} denote the CDF of the error term.

Suppose that a null hypothesis is to be tested using results from the estimation of (2.35). A null hypothesis might consist of linear restrictions on the elements of $\boldsymbol{\gamma}$, or it might reflect an attempt to test for misspecification by nesting (2.35) in some more general model. In the former case, the restricted least squares estimator for (2.35), denoted by $\tilde{\boldsymbol{\gamma}}$, and the EDF of the associated residuals $\tilde{u}_t = y_t - \mathbf{w}'_t\tilde{\boldsymbol{\gamma}}$, $t = 1, \dots, n$, can be used as the parameters of a bootstrap DGP. In the latter case, the unrestricted least squares estimator for (2.35), denoted by $\hat{\boldsymbol{\gamma}}$, and the EDF of the associated residuals $\hat{u}_t = y_t - \mathbf{w}'_t\hat{\boldsymbol{\gamma}}$, $t = 1, \dots, n$, are the bootstrap counterparts of $\boldsymbol{\gamma}$ and \mathcal{F} . (For convenience, it is assumed that an intercept term is estimated in both cases, which implies that residuals from both restricted and unrestricted estimation of (2.35) sum to zero and that there is no need to recentre before resampling.)

The details of generating bootstrap data will be considered for the first type of test, that is, when the null hypothesis specifies the values of $q < p + k$ linear combinations of the elements of $\boldsymbol{\gamma}$. The modifications required when dealing with tests for misspecification are obvious. (All that needs to be done is to replace $\tilde{\boldsymbol{\gamma}}$ by $\hat{\boldsymbol{\gamma}}$ and (2.34) by (2.27) in the schemes below.) Given the form of (2.35), an obvious way in which to simulate bootstrap data y_t^* for B samples, each of size n , is to use an *autoregressive (recursive) bootstrap scheme* defined by

$$y_t^* = \mathbf{y}'_{t(p)^*}\tilde{\boldsymbol{\alpha}} + \mathbf{x}'_t\tilde{\boldsymbol{\beta}} + u_t^*, t = 1, \dots, n, \quad (2.36)$$

in which $\mathbf{y}'_{t(p)^*} = (y_{t-1}^*, \dots, y_{t-p}^*)$ and the terms u_t^* are sampled randomly, with replacement, from the EDF of (2.34). Clearly, in order to start up

the simulation engine represented by (2.36), values must be provided for y_s^* , $s = 0, -1, \dots, 1 - p$. The obvious solution is to use the actual values y_s , $s = 0, -1, \dots, 1 - p$. This treatment of initial values implies a conditioning that is of no consequence asymptotically.

There is one possible difficulty with the use of (2.36) that may occur, albeit with low probability for moderately large sample sizes. Although the elements of α are assumed to satisfy conditions for dynamic stability, it is possible that sampling fluctuations could produce a value of $\tilde{\alpha}$ such that (2.36) is dynamically explosive. For example, with $p = 1$, the true coefficient is assumed to be such that $|\alpha_1| < 1$, but the corresponding estimate $\tilde{\alpha}_1$ might be observed to be greater than 1 or less than -1 . When the null hypothesis is true, the consistency of $\tilde{\alpha}$ implies that, as $n \rightarrow \infty$, (2.36) will be an appropriate dynamically stable model that satisfies the null hypothesis, with IID errors, for almost all samples. However, in applied work, it would seem sensible to check by examining the roots of

$$z^p - \tilde{\alpha}_1 z^{p-1} - \dots - \tilde{\alpha}_p = 0,$$

to see if any are on or outside the unit circle.

The combination of the autoregressive (recursive) bootstrap (2.36) with the residual EDF of (2.34) seems a natural way in which to mimic the DGP assumed to hold for real data, viz., (2.35). Some researchers have considered a different approach in which the lagged dependent variables, as well as the exogenous variables, in the regressor set of (2.35) are treated as fixed in the bootstrap world. In this *fixed regressor bootstrap*, (2.36) is replaced by

$$y_t^* = \mathbf{y}'_{t(p)} \tilde{\alpha} + \mathbf{x}'_t \tilde{\beta} + u_t^* = \mathbf{w}'_t \tilde{\gamma} + u_t^*, t = 1, \dots, n, \quad (2.37)$$

with the bootstrap errors u_t^* being sampled randomly, with replacement, from the EDF of (2.34). At first sight, treating all regressors as fixed in (2.37) seems inconsistent with the general notion that the bootstrap DGP should be as close as possible to the DGP assumed to yield the actual observed data, that is, the autoregressive regression model (2.35). It, therefore, seems reasonable to be concerned about the validity of using the fixed regressor bootstrap when the assumed model has lagged values of the dependent variable included as regressors.

The use of autoregressive (recursive) bootstraps and fixed regressor bootstraps for time series models is examined in Franke et al. (2002). It is established that, provided regularity conditions are satisfied, both schemes are asymptotically valid. However, Franke et al. provide evidence from simulation experiments, as well as asymptotic theory, and

this evidence leads them to the conclusion that the “autoregression resampling scheme, which takes full account of the dependence structure, results in much better approximations”; see Franke et al. (2002, p. 18). In the applications to tests for dynamic regression models to be discussed in later chapters, autoregressive bootstrap schemes like (2.36) will, therefore, be adopted.

2.3.4. The choice of the number of artificial samples

Various schemes for generating B artificial samples, each with n observations, have been described above. However, the choice of B has not been considered so far. In order to fill this gap in the exposition, some comments on the importance of this choice and on available evidence that guides practical implementation are provided in this subsection.

If computing were a free good and computers were infinitely fast, the obvious choice would be the *ideal bootstrap* with B being infinitely large. However, neither of these conditions will ever apply and a finite value of B must be selected. If a Monte Carlo test is being used, the comments in Sections 2.2.1 and 2.3.1 indicate that, in order to derive an exact test, B should be chosen so that, for a required significance level of α , $0 < \alpha < 1$, $\alpha(B + 1)$ is an integer. But this rule is not sufficient to guide the choice of the specific value of B . After discussing the blurring of the critical region mentioned above and the associated loss of power, Marriott suggests that, when carrying out a Monte Carlo test, arranging for the value of $\alpha(B + 1)$ to be 5 might often be suitable; see Marriott (1979) and also Hall and Titterton (1989) and Jöckel (1986) for related theoretical analyses.

Given existing computer resources, using $B = 999$ would be feasible for Monte Carlo tests for regression models in a very large number of cases and would satisfy Marriott’s rule of thumb for all standard choices of α . However, there have been major advances in computing power since 1979; see the illustration based upon the estimation of regression equations given in MacKinnon (2002, p. 616), which can be usefully combined with reported computing times for simulation-based tests in McCullough and Vinod (1993, section 6). Consequently, unless the test under consideration is exceptionally difficult to compute, Marriott’s rule might be updated to choosing B so that $\alpha(B + 1)$ is 15 or 20.

Turning to bootstrap tests, there will not be perfect control of significance levels in finite samples because the artificial test statistics only share the null distribution of the actual test statistic asymptotically. Evidence concerning the typical value of B that gives a useful degree of control for sample sizes of relevance to applied workers is obviously of

interest. In an important early contribution to the econometrics literature on bootstrap tests, Horowitz reports that “satisfactory results can be obtained with as few as 100 bootstrap samples”; see Horowitz (1994, p. 404). However, MacKinnon points out that the use of such a small value for B can have costs; see MacKinnon (2002, p. 619).

Using a small number of artificial samples clearly affects the precision of estimation of p -values under the null hypothesis. One way in which to take precision into account is to construct confidence interval estimates of the p -value. Hedges argues, on the basis of a common desired width of the approximate 95 per cent confidence interval for the actual null rejection probability, α_a , in favour of using $B = 400$ with $\alpha_a = 0.01$ and $B = 2,000$ with $\alpha_a = 0.05$; see Hedges (1992). However, others have suggested that it is proportionate, rather than absolute, accuracy that is relevant for significance levels. For example, the stringent criterion of robustness given in Serlin (2000) is that, with a desired significance level of α_d , null hypothesis rejection probabilities should be in the range $\alpha_d \pm 0.1\alpha_d$. Also, as for Monte Carlo tests, behaviour under the alternative hypothesis should be considered when selecting the value of B . Several authors have drawn attention to the loss of power associated with the use of relatively small values of B ; see, for example, Davidson and MacKinnon (2000) and Hall and Titterton (1989).

As well as being informed by theoretical analyses and simulation studies of significance level and power, recommendations about suitable values for B are likely to take into account computing power and the nature of the calculations required to obtain the test statistic of interest. The estimation of linear regressions is not onerous with modern computers, especially when the model is not dynamic so that it is not necessary to invert repeatedly the matrix $(X'X)$, which is fixed for each bootstrap sample. If the updated version of Marriott's rule for Monte Carlo tests were applied to bootstrap tests, with bootstrap p -values determined by (2.12), the value of B would be selected so that $\alpha_d B = 15$ or 20, given the prespecified value of α_d . Using the more demanding version of this rule would imply $B = 400$ when $\alpha_d = 0.05$ and $B = 2,000$ when $\alpha_d = 0.01$. These combinations of α_d and B are reasonably similar to those identified by Davidson and MacKinnon as likely to avoid a power loss of more than 0.01 (1 percentage point); see Davidson and MacKinnon (2000, p. 60).

In many applications, it would probably be reasonable to use $B = 1,000$ for bootstrap tests, unless the null is only to be rejected for very strong evidence (e.g., $\alpha_d = 0.01$) in which case $B = 2,000$ might be preferred; see Davidson and MacKinnon (2000). Whichever of these two values is

adopted, the associated waiting time for the user with access to a typical personal computer will be of no practical importance. The real obstacle to the use of the bootstrap as a standard tool is not the required computing time but the absence of suitable code in commercial estimation packages. This difficulty should be eliminated as the usefulness of the bootstrap becomes widely known and discussions of simulation-based tests become part of econometrics text books and lecture courses.

Finally, it should be noted that B need not be prespecified. The value of B can instead be selected by pretest methods with the aim of getting close to the performance of ideal bootstrap tests; see Davidson and MacKinnon (2000). A three-step method for selecting B when constructing bootstrap confidence intervals is proposed in Andrews and Buchinsky (2002) and its application to bootstrap tests is examined in Davidson and MacKinnon (2000).

2.4. Asymptotic properties of bootstrap tests

Test statistics that are used in applied analyses are very rarely exactly pivotal. Consequently, in most cases, the comparison of a test statistic calculated from the actual data with the values of test statistics calculated from bootstrap samples is only justified asymptotically. Thus the bootstrap approach will usually only produce an asymptotically valid test. It is, therefore, reasonable to ask what is gained by using bootstrap methods, rather than making use of the critical values implied by asymptotic theory and the choice of the nominal significance level. There are two types of gain that can be identified.

First, as illustrated by the examples of Section 1.6, the standard first-order asymptotic theory does not always permit the derivation of a feasible test. Horowitz (2003, p. 211) contains the following remarks:

Many important statistics in econometrics have complicated asymptotic distributions that depend on nuisance parameters and, therefore, cannot be tabulated. ... The bootstrap and related resampling techniques provide practical methods for estimating the asymptotic distributions of such statistics.

The first type of gain is, therefore, that the bootstrap can fill important gaps in the toolkit of the applied worker. However, even when asymptotic theory leads to a feasible test for the problem at hand, it may be beneficial to use a bootstrap procedure.

The second type of gain from using bootstrap samples may be available when asymptotic theory, although applicable, provides an inadequate approximation to finite sample behaviour. As will be shown in later chapters, there is a great deal of evidence that bootstrapping an asymptotic test can lead to an improved degree of control of significance levels and that the improvement is substantial in some cases.

This second type of gain is often discussed in terms of the difference between the nominal (asymptotically achieved) significance level and the actual significance level. This difference is called the *error in rejection probability* (ERP). The ability of the bootstrap to reduce the ERP, relative to the use of standard asymptotic theory, is discussed by Beran, who refers to the *asymptotic refinements* associated with bootstrapping; see Beran (1988).

Beran's asymptotic analysis is frequently mentioned in the econometrics literature on bootstrap tests. It is based upon an assumption that, under the null hypothesis, the finite sample CDF of the test statistic, denoted by $\mathcal{G}_n(\cdot; \theta)$, can be written in the form of an expansion, with

$$\mathcal{G}_n(a; \theta) = \mathcal{G}_\infty(a; \theta) + n^{-j/2}g(a; \theta) + O(n^{-(j+1)/2}), \quad (2.38)$$

in which $\mathcal{G}_\infty(a; \theta)$ denotes the limit null distribution, which is continuous and strictly monotonic, $j \geq 1$ is an integer and $g(\cdot; \theta)$ is continuous; see Beran (1988, p. 690). The bootstrap world is defined using the estimator $\check{\theta}$ which (i) satisfies the null hypothesis and (ii) differs from θ by terms that are $O_p(n^{-1/2})$ when the null hypothesis is true. The bootstrap counterpart of (2.38) is

$$\mathcal{G}_n(a; \check{\theta}) = \mathcal{G}_\infty(a; \check{\theta}) + n^{-j/2}g(a; \check{\theta}) + O^*(n^{-(j+1)/2}),$$

and $\mathcal{G}_n(a; \check{\theta})$ can be approximated using the EDF of the B values of the bootstrap statistics. Beran uses his expansions to obtain the following results on the relative magnitudes of the ERP terms for asymptotic and bootstrap tests.

First, if the test statistic is not asymptotically pivotal, in other words, $\mathcal{G}_\infty(a; \theta)$ really depends on at least one term in θ , the bootstrap test has an ERP of the same order in the sample size n as the asymptotic test. Second, if the test statistic is asymptotically pivotal, that is, $\mathcal{G}_\infty(a; \theta)$ is independent of all terms in θ and so can be written as $\mathcal{G}_\infty(a)$, the bootstrap test has an ERP of smaller order in the sample size n than does the asymptotic test. A detailed discussion of these findings is given in section 3 of Beran (1988). The result for an asymptotically pivotal statistic is widely applicable in econometrics because so many tests are carried out

using critical values from asymptotically valid standard distributions, such as $N(0, 1)$ and χ^2 .

In fact, the result given by Beran can be strengthened for many econometric tests of regression models that are based upon asymptotic pivots. Davidson and MacKinnon show that a further asymptotic refinement, in addition to the one established by Beran, is available when the estimator $\ddot{\theta}$, which defines the bootstrap DGP, is asymptotically independent of the test statistic; see Davidson and MacKinnon (1999, section 4). In regression models of the type discussed in the previous section there is asymptotic independence of these objects, given regularity conditions, when $\ddot{\theta} = (\ddot{\beta}, \ddot{\mathcal{F}})$ is derived by using appropriate restricted extremum estimators, with the restrictions imposed on the estimators being those that make up the null hypothesis; see Davidson and MacKinnon (1999, p. 369).

The results given in Beran (1988) and Davidson and MacKinnon (1999) on the asymptotic refinements associated with bootstrapping tests can be linked to those of analytical approaches that provide the same orders of improvement of ERP relative to an unadjusted asymptotic theory test, for example, see Hall (1992) for an examination of the bootstrap and Edgeworth expansions. Horowitz provides a useful summary of results on the orders of magnitude in the sample size n for the ERP terms of asymptotic and bootstrap tests, drawing on large sample theory contained in Hall (1992); see Horowitz (2001, section 3). For example, he mentions that for test statistics that are asymptotically distributed as χ^2 , when the null hypothesis is true, the ERP terms for asymptotic and bootstrap tests are $O(n^{-1})$ and $O(n^{-2})$, respectively; see Horowitz (2001, p. 3183). Many standard econometric test statistics fall into this category, each being compared with an asymptotic critical value in the right-hand tail of the relevant χ^2 distribution.

However, in practical situations with limited sample sizes, it is not clear how much weight should be placed upon formal results about such orders of magnitude. These results indicate the relative asymptotic orders of magnitude of ERP terms, but, as noted in Davidson and MacKinnon (2006), it is possible that there are cases in which, when n is small, the asymptotic test works very well but the bootstrap test is a little inferior; also see MacKinnon (2002, section 4). However, while such cases cannot be ruled out as impossible, it is important to stress that there is a great deal of evidence, covering a number of applications, which indicates that asymptotically valid critical values can be far from the true values for sample sizes of relevance to the empirical worker. Moreover,

this evidence suggests that the application of the bootstrap does not simply improve the situation but gives good management of finite sample significance levels.

The analysis given in Beran (1988) is based on high-level assumptions, that is, the validity of (2.38), and is an outline of a general approach. While in no way wishing to detract from the importance of the asymptotic analysis in Beran (1988), it must be acknowledged that a full proof of the asymptotic validity of bootstrap methods for specific applications usually has to be carried out on a case-by-case basis. Freedman has provided details for linear regression models and shows the asymptotic validity of some bootstrap tests, under weak regularity conditions; see Freedman (1981, 1984). However, as remarked by Freedman, his analysis is confined to asymptotic theory and he is not addressing the problem of showing that the bootstrap test is better behaved in finite samples than the asymptotic test; see, for example, Freedman (1984, p. 827). Fortunately, in recent years, many researchers have carried out simulation studies of the relative finite sample performance of asymptotic and bootstrap tests for regression models. Some of these studies will be discussed below and references to others will be provided.

2.5. The double bootstrap

As well as comparing tests based upon a limit null distribution with those derived using the bootstrap, Beran considers a third general method in which the bootstrap samples are themselves bootstrapped; see Beran (1988, section 2). In this third approach, the test statistic is said to be the subject of *prepivoting*. Whether or not the original test statistic is asymptotically pivotal, its p -value is asymptotically uniformly distributed between 0 and 1, when the null hypothesis is true, and so is an asymptotic pivot. The second level of bootstrapping, therefore, involves using the first-level bootstrap p -value as the test statistic of interest. (An equivalent approach would be to work with the CDF of the bootstrap null distribution, rather than the p -value; see Beran (1988, section 2.3). The implied adjustments, for example, to rejection rules are straightforward.)

The application of the double bootstrap to tests for regression models can be described as follows. Suppose that the model used to approximate the real world is the regression model (2.15), with strictly exogenous regressors and IID errors having a conditional distribution given by (2.25) and common cdf \mathcal{F} . Also suppose that the number of samples in the first and second stages of bootstrapping are B and C , respectively. As in Beran's analysis, an estimator of $\theta = (\beta, \mathcal{F})$ is required which satisfies the null

hypothesis and differs from θ by $O_p(n^{-1/2})$ when this null hypothesis is true; see Beran (1988, section 3.1). Let $\hat{\theta} = (\hat{\beta}, \hat{\mathcal{F}})$ be such an estimator and $\hat{u}_i, i = 1, \dots, n$, denote the residuals associated with $\hat{\beta}$.

If linear restrictions on the elements of β are under test, with (1.18) as the null hypothesis, $\hat{\theta}$ can consist of the restricted least squares estimator $\hat{\beta}$ and the EDF of the restricted residuals $\hat{u}_i, i = 1, \dots, n$. If, on the other hand, (2.15) is regarded as the restricted model for the purpose of deriving a test for misspecification, for example, a check for heteroskedasticity, the OLS estimator $\hat{\beta}$ and the EDF of the OLS residuals $\hat{u}_i, i = 1, \dots, n$, can be used to obtain $\hat{\theta}$. Whatever the nature of the null hypothesis and the form of the associated restricted estimator, the test statistic calculated from actual data is denoted by \tilde{r} .

Consider then the application of a double bootstrap to a test based upon \tilde{r} . The first level of the double bootstrap involves deriving B artificial samples of size n , using

$$\mathbf{y}_b^* = \mathbf{X}\tilde{\beta} + \mathbf{u}_b^*, b = 1, \dots, B,$$

in which: $\mathbf{y}_b^* = (y_{b1}^*, \dots, y_{bn}^*)'$; and $\mathbf{u}_b^* = (u_{b1}^*, \dots, u_{bn}^*)'$ is obtained by simple random sampling, with replacement, from the EDF defined by

$$\tilde{\mathcal{F}} : \text{probability } \frac{1}{n} \text{ on } \hat{u}_i, i = 1, \dots, n.$$

(For simplicity, it is assumed that $\tilde{\beta}$ includes an estimator of an intercept, so that the residuals need not be recentred before resampling takes place.) By applying the formulae that are used with actual data to these artificial data, the first-level bootstrap values of coefficient estimates, residual vectors and test statistics are found. These objects are, using the obvious notation, denoted by $\tilde{\beta}_b^*, \tilde{\mathbf{u}}_b^* = (\tilde{u}_{b1}^*, \dots, \tilde{u}_{bn}^*)'$ and τ_b^* , for $b = 1, \dots, B$.

In the second level of the double bootstrap, C artificial samples of size n are derived using the bootstrap DGP defined by $\tilde{\theta}_b^* = (\tilde{\beta}_b^*, \tilde{\mathcal{F}}_b^*)$, where $\tilde{\mathcal{F}}_b^*$ is the EDF of the residuals of the vector $\tilde{\mathbf{u}}_b^* = (\tilde{u}_{b1}^*, \dots, \tilde{u}_{bn}^*)'$, for $b = 1, \dots, B$. More precisely, for each value of b , the second-level bootstrap data are generated using

$$\mathbf{y}_{bc}^{**} = \mathbf{X}\tilde{\beta}_b^* + \mathbf{u}_{bc}^{**}, c = 1, \dots, C,$$

in which: $\mathbf{y}_{bc}^{**} = (y_{bc1}^{**}, \dots, y_{bcn}^{**})'$; and $\mathbf{u}_{bc}^{**} = (u_{bc1}^{**}, \dots, u_{bcn}^{**})'$ is obtained by simple random sampling, with replacement, from the EDF defined by

$$\tilde{\mathcal{F}}_b^* : \text{probability } \frac{1}{n} \text{ on } \tilde{u}_{bi}^*, i = 1, \dots, n.$$

The simulated data from the BC second-level samples are used to calculate the bootstrap test statistics τ_{bc}^{**} , $b = 1, \dots, B$ and $c = 1, \dots, C$.

In the single bootstrap test, the estimated p -value of $\check{\tau}$, that is,

$$\check{p} = \frac{\sum_{b=1}^B \mathbf{1}(\tau_b^* \geq \check{\tau})}{B} = \frac{\#\{\tau_b^* \geq \check{\tau}\}}{B}, \quad (2.39)$$

is used in the rejection rule

$$\text{Reject } H_0 \text{ if } \check{p} \leq \alpha_d,$$

in which α_d is the desired significance level. The bootstrap p -value of (2.39) is obtained by using the B first-level bootstrap statistics as a reference set for the actual statistic. In the double bootstrap method, a bootstrap p -value for each of the first-level statistics τ_b^* is obtained by using the corresponding C second-level statistics $(\tau_{b1}^{**}, \tau_{b2}^{**}, \dots, \tau_{bC}^{**})$ as the reference set. Thus, the estimated p -value of τ_b^* is given by

$$\check{p}_b^* = \frac{\sum_{c=1}^C \mathbf{1}(\tau_{bc}^{**} \geq \tau_b^*)}{C} = \frac{\#\{\tau_{bc}^{**} \geq \tau_b^*\}}{C}, b = 1, \dots, B.$$

When \check{p} is viewed as an asymptotically pivotal statistic, it is to be compared with the reference set that consists of the terms $(\check{p}_1^*, \check{p}_2^*, \dots, \check{p}_B^*)$. The *double bootstrap p-value* is then

$$\check{p}_D = \frac{\sum_{b=1}^B \mathbf{1}(\check{p}_b^* \leq \check{p})}{B} = \frac{\#\{\check{p}_b^* \leq \check{p}\}}{B}, \quad (2.40)$$

with small values indicating a large amount of evidence against the null hypothesis. Given a desired significance level of α_d , the double bootstrap leads to the decision rule

$$\text{Reject } H_0 \text{ if } \check{p}_D \leq \alpha_d.$$

The double bootstrap p -value \check{p}_D is sometimes referred to as an *adjusted p-value* and some authors denote it by p_{adj} ; see Davison and Hinkley (1997, p. 175).

By using the asymptotic expansion (2.38), Beran shows that, if the ERP for the single bootstrap test is $O(n^{-j/2})$, the ERP for the double bootstrap test is $O(n^{-(j+1)/2})$, for some integer $j \geq 1$. This additional asymptotic refinement may be especially useful when the original test statistic is not asymptotically pivotal, for example, as in the cases discussed in Section 1.6. It is, however, clear that this potential gain in accuracy must be

bought at the price of a higher computational cost than the standard single bootstrap discussed in the previous section. The total number of bootstrap samples required is $B(1 + C)$. There is, however, no need to select C so that it is equal to B (or of similar size) and $C = O(B^{1/2})$ is sometimes recommended; see Booth and Hall (1994). Moreover, there are several results that show how the computational costs of the double bootstrap can be greatly reduced compared with direct repetition of the calculations carried out with the actual sample for all $B(1 + C)$ bootstrap samples.

It is pointed out by Godfrey and Orme that, when the regressors are strictly exogenous and so are held fixed over bootstrap samples, time consuming operations like matrix inversion need only be carried out once, not $1 + B(1 + C)$ times, and that fixed projection matrices, like the hat-matrix of (1.9), can be used in some applications; see Godfrey and Orme (2002a, p. 433). Godfrey and Orme also refer to results on the usefulness of *stopping rules* for the double bootstrap that are given in Horowitz et al. (2006). Horowitz et al. provide details of stopping rules for double bootstrap tests and report that the combined effect of the rules is that the number of second-level samples required in their experiments is only between $BC/11$ and $BC/15$; see Horowitz et al. (2006, appendix B, p. 861). Computational savings of this size are clearly very useful, but even more impressive savings are associated with a technique proposed by Davidson and MacKinnon; see Davidson and MacKinnon (2002b, 2007).

Davidson and MacKinnon describe a *fast double bootstrap* (FDB) technique. In the notation of this section, the FDB test uses $2B$ artificial samples, with each first-level bootstrap sample being bootstrapped to give just one second-level sample, in other words, $C = 1$. Consequently, there is a first-level test statistic τ_b^* and a second-level test statistic τ_{b1}^{**} , for $b = 1, \dots, B$. The FDB p -value, as discussed in Davidson and MacKinnon (2002b), is defined by

$$\ddot{p}_F = \frac{\sum_{b=1}^B \mathbf{1}(\tau_b^* > \ddot{Q}_B^{**})}{B} = \frac{\#(\tau_b^* > \ddot{Q}_B^{**})}{B}, \quad (2.41)$$

in which \ddot{Q}_B^{**} is selected to satisfy

$$\frac{\sum_{b=1}^B \mathbf{1}(\tau_{b1}^{**} > \ddot{Q}_B^{**})}{B} = \frac{\#(\tau_{b1}^{**} > \ddot{Q}_B^{**})}{B} = \ddot{p}, \quad (2.42)$$

and \check{p} is given by (2.39). For any finite number of bootstrap samples, \check{Q}_B^{**} must be selected from an interval of possible values implied by (2.42); see, for example, (2.43) below.

Davidson and MacKinnon draw attention to the need for asymptotic independence of the bootstrap DGP and the test statistic if the FDB procedure is to have a smaller order of ERP than the single-bootstrap test based upon comparing \check{p} with the desired significance level; see Davidson and MacKinnon (2002b, p. 423). The standard double bootstrap test achieves the asymptotic refinement without imposing this assumption, but it has a much higher computational cost. If the smallest possible value of \check{Q}_B^{**} that satisfies (2.42) is selected, a simple link between the FDB method and the standard double bootstrap can be shown.

In order to illustrate the general rule that the smallest possible value of \check{Q}_B^{**} be used, suppose that $B = 1,000$ and $\check{p} = 0.05$. Let $\tau_{[1]1}^{**}, \dots, \tau_{[1,000]1}^{**}$ denote the ordered values of the second-level statistics τ_{b1}^{**} , with $\tau_{[1]1}^{**}$ being the largest value. From (2.42), $1,000 \times 0.05 = 50$ of the terms τ_{b1}^{**} must be greater than \check{Q}_{1000}^{**} , so that

$$\tau_{[50]1}^{**} > \check{Q}_{1000}^{**} \geq \tau_{[51]1}^{**}. \quad (2.43)$$

The smallest suitable value of \check{Q}_{1000}^{**} is, therefore, $\tau_{[51]1}^{**}$.

A link between fast and standard versions of the double bootstrap can be shown by introducing the terms \check{p}_b^* , which are the proportions of the second-level bootstrap statistics $\tau_{11}^{**}, \dots, \tau_{B1}^{**}$ not less than τ_b^* , i.e.,

$$\check{p}_b^* = \frac{\sum_{j=1}^B \mathbf{1}(\tau_{j1}^{**} \geq \tau_b^*)}{B} = \frac{\#(\tau_{j1}^{**} \geq \tau_b^*)}{B}, \quad (2.44)$$

for $b = 1, \dots, B$. It then follows that $\mathbf{1}(\tau_b^* > \check{Q}_B^{**}) = \mathbf{1}(\check{p}_b^* \leq \check{p})$ and so, considering the numerator of (2.41),

$$\#(\tau_b^* > \check{Q}_B^{**}) = \sum_{b=1}^B \mathbf{1}(\tau_b^* > \check{Q}_B^{**}) = \sum_{b=1}^B \mathbf{1}(\check{p}_b^* \leq \check{p}) = \#(\check{p}_b^* \leq \check{p}).$$

Hence, given the rule on selecting \check{Q}_B^{**} , the FDB p -value of (2.41) can be rewritten as

$$\check{p}_F^Q = \frac{\sum_{b=1}^B \mathbf{1}(\check{p}_b^* \leq \check{p})}{B} = \frac{\#(\check{p}_b^* \leq \check{p})}{B}, \quad (2.45)$$

which can be compared with the conventional double bootstrap p -value of (2.40).

The comparison of (2.40) and (2.45) reveals that the computational savings of the fast double bootstrap, relative to the conventional version, result from using the same set of statistics $(\tau_{11}^{**}, \dots, \tau_{B1}^{**})$ with each first-level statistic τ_b^* , $b = 1, \dots, B$, in order to build up an estimated reference distribution, that is, the EDF of $(\check{p}_1^*, \dots, \check{p}_B^*)$, with which to assess \check{p} . However, each term τ_{b1}^{**} is calculated using an artificial sample from its own bootstrap population defined by a first-level bootstrap estimate $\check{\theta}_b^*$, which varies with b . For an asymptotic refinement of the fast double bootstrap relative to the single bootstrap, it is required that the variations in bootstrap DGPs have negligible effects when estimating a p -value for τ_b^* using (2.44). This requirement highlights the importance of the asymptotic independence of the bootstrap DGP and the test statistic that is stressed in Davidson and MacKinnon (2007). The FDB test will be asymptotically valid under more general conditions; see Davidson and MacKinnon (2002b).

2.6. Summary and concluding remarks

The conventional approach to testing a null hypothesis involves the comparison of the sample value of a test statistic with a critical value from an appropriate standard distribution (such as $N(0, 1)$, t , F or χ^2). In many cases, this approach can only be justified by appealing to asymptotic theory. There are some tests for which it is possible to find sufficiently strong assumptions to obtain exact, rather than asymptotic, validity. However, in such cases, it is reasonable to be concerned about the validity of the restrictive assumptions, for example, Normality, that permit perfect control of the finite sample significance level. When these assumptions do not hold, the tests are typically only asymptotically valid. There is, therefore, considerable reliance on a body of results that only hold as the sample size tends to infinity.

In practical situations, the sample size is finite and may not be large. The quality of the approximation provided by asymptotic theory is open to question and more reliable approaches to judging the statistical significance of sample outcomes are of interest. This chapter has contained descriptions of simulation-based methods that replace a tabulated reference distribution by the EDF of a set of test statistics calculated from artificial samples of the same size as the genuine one. The actual test statistic is compared with these artificial test statistics to see if it can be viewed as being so unusual that the null hypothesis can be rejected.

It is obviously important that the artificial model used to generate the simulated data should be specified according to proper guidelines. It cannot be guaranteed that, for the given actual sample, the application of the guidelines will produce a good approximation, but that is not a good reason not to adopt sensible rules. The first rule identified in Davidson (2007) is that the artificial model, also known as the bootstrap DGP, should satisfy the null hypothesis that is to be tested. This rule means that restricted estimates calculated from the actual data are used for bootstrap world parameters. It is assumed that the restricted estimators are consistent when the null hypothesis is true. Moreover, as argued in Davidson (2007), it would seem sensible to use restricted estimators that are asymptotically efficient.

The focus is on tests for linear regression models. As argued in Hansen (1999, p. 195), the “interpretation of the regression function as a conditional expectation has no relationship to an auxiliary assumption of normality” and there seems no good reason to specify the Normal or any other distribution for the errors, which have been assumed simply to be IID in this chapter. The absence of an assumption that specifies the general family of the distribution of an error term leads to an emphasis in this book on nonparametric bootstraps, rather than parametric bootstraps and Monte Carlo methods of the type discussed in Dufour and Khalaf (2001).

The IID terms required for bootstrap world errors are obtained by simple random sampling, with replacement, from the actual (possibly modified and/or recentred) residuals. Hence, in a selected set of bootstrap errors, some residuals will appear more than once and others will not appear at all. The bootstrap errors are added to predicted values from actual data, which serve as conditional expectations in the bootstrap world, in order to derive the required artificial data. This process is repeated many times and a bootstrap test statistic is calculated each time an artificial sample is drawn.

The comparison of the actual sample value of the test statistic with the artificially produced values enables tests to be carried out when there is no feasible procedure based upon the conventional combination of asymptotic theory and critical values from a standard distribution. Thus new tests can be provided for routine application in empirical work, simply by modifying estimation programs to allow bootstrapping.

When a test can be carried out on the basis of asymptotic theory and critical values from a standard tabulated distribution, in other words, the test statistic is an asymptotic pivot, the bootstrap yields asymptotic refinements relative to this conventional approach. Indeed Beran

suggests that classical analytic modifications of asymptotic tests can be regarded as approximations to bootstrap tests; see Beran (1988, p. 687) for general comments and Navidi (1989) for a detailed examination of bootstrap procedures in regression models. In view of the results on relative orders of magnitude of the errors in rejection probabilities, it is usually recommended that, when possible, an asymptotically pivotal test statistic should be bootstrapped.

As well as considering bootstrapping the actual data, the possibility of bootstrapping the bootstrap data, that is, double bootstrapping, has been discussed. The theoretical results on any improvements in errors in rejection probabilities associated with single and double bootstraps have been outlined. These improvements are obtained without the level of analytical input required, for example, by an approach using Edgeworth expansion corrections. As noted in Beran (1988), the “possibility of direct nonanalytical implementation is a great practical merit” of bootstrap tests.

It should, however, be recognized that predictions about errors in rejection probabilities derived from asymptotic expansions might not be of great value for sample sizes of a magnitude of interest to empirical workers. It is possible that two test procedures have the same order of magnitude of the error of rejection probability but are found to exhibit different finite sample behaviour in simulation experiments. The form of the test statistic and, in particular, the quality of the covariance matrix estimate used in its construction may be important; see, for example, Davidson and MacKinnon (1992). There is even evidence that, when estimating impulse response coefficients in vector time series models, it might be better to bootstrap non-asymptotically pivotal quantities, rather than asymptotic pivots, because of problems with estimating variances; see Berkowitz and Kilian (2000) and the associated comments in Davidson (2000) for details.

In order to provide and evaluate evidence on the actual performance of asymptotic and simulation-based tests, much of the rest of this book is devoted to discussions of results from simulation experiments that are intended to be relevant to a number of situations of econometric interest. Chapters 3 and 4 contain results that, like those of this chapter, are relevant when the regression model has errors that are IID. However, applied workers may not wish to make the strong assumptions of independence and homoskedasticity. When either autocorrelation or heteroskedasticity is part of the model specification, the bootstrap methods described in this chapter are inappropriate because they are derived under the assumption that errors are IID with a finite variance. Chapter 5

contains descriptions of bootstrap techniques that have been proposed to accommodate departures from the IID assumption. Applications of these techniques are discussed in Chapter 6.

Although there are many results from analytical investigations and simulation experiments that support the use of bootstrap tests, there are also certain unusual cases in which the bootstrap is inconsistent; see Horowitz (2001, pp. 3167–3169). For example, the results in Athreya (1987) indicate that there would be problems if the errors were to have a heavy-tailed distribution with infinite variance. However, the assumption of finite variability of actual economic quantities seems reasonable. More generally, Samworth suggests that focussing on the consistency of a bootstrap method “can mask the finite-sample behaviour, and that inconsistent bootstrap estimators may in fact perform better than their consistent counterparts”; see Samworth (2003, p. 985). When discussing applications of bootstrap tests in this book, references will be supplied to provide potential users not only with information about theory-based results that are relevant to asymptotic validity, but also with a summary and comments on evidence derived from simulation experiments that throws light on finite sample properties.

3

Simulation-based Tests for Regression Models with IID Errors: Some Standard Cases

3.1. Introduction

Many applied studies involve the estimation and analysis of a linear regression model with IID errors. Several tests for this model were described in Chapter 1 and comments were made about the possible dangers of using asymptotic theory as the foundation for inference in empirical investigations. The purpose of this chapter is to show how simulation methods, which were discussed in Chapter 2, can be used to provide an improved basis for testing. The general structure of this chapter follows that of Chapter 2 in so far as exact Monte Carlo techniques are illustrated before the more generally applicable (but only asymptotically valid) nonparametric bootstrap methods are considered. When appropriate, the specific examples are accompanied by some comments on general issues relevant to applied work.

All of the examples discussed in this chapter involve test statistics that are either exactly pivotal or asymptotically pivotal. In the former case, the finite sample distribution of the test statistic, under the null hypothesis, does not depend upon any unknown parameters. In the latter case, it is only the asymptotic distribution of the test statistic, under the null hypothesis, that is independent of unknown parameters. In practice, few of the test statistics used in econometrics are exactly pivotal but many of them have the weaker property of being asymptotic pivots. Attention is restricted in this chapter to asymptotic pivots which, when the null hypothesis is true, have a known asymptotic distribution of a standard type, for example, $N(0, 1)$ or χ^2 . Test statistics that are either asymptotically pivotal with non-standard limit null distributions or not even asymptotically pivotal are discussed in the next chapter. In short,

this chapter deals with standard cases and the next chapter covers some non-standard cases.

As explained in Chapter 2, when a test statistic is exactly pivotal, it is possible to use the Monte Carlo approach to derive an exact test. For such a test, the error in rejection probability (ERP) is zero. When a test statistic is only asymptotically pivotal, critical values are often taken from a known limit null distribution. For tests based upon asymptotic critical values, the ERP term will tend to zero as the sample size goes to infinity, but can be large in practical situations. The results contained in Beran (1988) indicate that, if an asymptotically pivotal test statistic is combined with the bootstrap approach, the resulting ERP will tend to zero faster than that associated with the use of asymptotic critical values. In the terminology of Chapter 2, the bootstrap provides an asymptotic refinement.

There are many published results on the asymptotic refinements associated with bootstrap tests. This literature is technical and sometimes involves relatively complex asymptotic analysis. However, it is not always the case that such asymptotic analysis seems to provide a good explanation of what is observed in finite samples. The following remarks, which are taken from Davidson (2007), are pertinent:

A technique that has been used a good deal in work on asymptotic refinements for the bootstrap is Edgeworth expansion of distributions, usually distributions that become standard normal in the limit of infinite sample size. The standard reference to this line of work is Hall (1992), although there is no shortage of more recent work based on Edgeworth expansions. Whereas the technique can lead to useful theoretical insights, it is unfortunately not very useful as a quantitative explanation of the properties of bootstrap tests. In concrete cases, the true finite-sample distribution of a bootstrap P value, as estimated by simulation, can easily be further removed from an Edgeworth approximation to its distribution than from the asymptotic limiting distribution.

In this chapter, the emphasis will be on discussing evidence about finite sample behaviour of bootstrap tests which is derived from simulation experiments and interpreted in the context of available asymptotic theory.

The plan of this chapter is as follows. In Section 3.2, the use of a Monte Carlo test that gives exact control of finite sample significance levels is explained. The test is of the null hypothesis of Normality. The derivation

of an exact test for this null hypothesis is of interest because asymptotic critical values are sometimes very inaccurate in finite samples; see Section 1.5.2 above.

Monte Carlo tests, such as the one discussed in Section 3.2, require that the form of the error distribution be specified, either in the null hypothesis to be tested or in a set of untested auxiliary assumptions. However, applied regression analysis is often undertaken without any precise information about the shape of the error distribution. In such cases, it is natural to be concerned about the robustness of Monte Carlo tests and to examine the use of the nonparametric bootstrap as a better source of control of significance levels than asymptotic theory.

Section 3.3 contains theory-based and simulation results about the robustness of Monte Carlo tests for heteroskedasticity to incorrect specification of the error distribution, as well as evidence about the improvements associated with using nonparametric bootstrap checks, rather than standard asymptotic tests. Well-known and widely-used tests for heteroskedasticity are used to illustrate general issues.

In Section 3.4, bootstrap techniques are applied to the problem of testing linear restrictions on regression coefficients, which was covered at length in Chapter 1. Attention is drawn to the issue of whether to use restricted or unrestricted estimates from the actual data to define the bootstrap DGP. A small-scale simulation experiment is reported in which tests derived using these two types of bootstrap model are compared with each other and with procedures that use asymptotically valid critical values from standard distributions.

The choice between restricted and unrestricted estimates for use as the bootstrap world parameters is also examined in Section 3.5, in which the widely-used *serial correlation test* of Breusch (1978) and Godfrey (1978) is adopted as the procedure for study. On the basis of evidence from simulation experiments, it is argued that, when testing for serial correlation in dynamic regression models, it is better to use the restricted (null hypothesis) estimates for the parameters of the bootstrap process.

A summary of results and some concluding comments are contained in Section 3.6.

3.2. A Monte Carlo test of the assumption of Normality

This section is based on results that are contained in Dufour et al. (1998). Dufour et al. discuss the problem of testing the assumption that the IID errors have a Normal distribution, that is, the errors are $NID(0, \sigma^2)$.

The linear regression model is written as

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{u}, \quad (3.1)$$

in which the usual notation is used; see Section 1.2. The regressors of (3.1) are assumed to be either fixed in repeated sampling or strictly exogenous. The null hypothesis is

$$H_0 : \mathbf{u} \sim N(\mathbf{0}_n, \sigma^2 \mathbf{I}_n),$$

and this claim is to be tested after OLS estimation of (3.1). Let the OLS residuals be the elements of $\hat{\mathbf{u}}' = (\hat{u}_1, \dots, \hat{u}_n)$. These residuals can be used to obtain

$$\hat{\sigma}^2 = n^{-1} \sum_{i=1}^n \hat{u}_i^2, \quad (3.2)$$

which is consistent for σ^2 , under standard conditions. Let $\hat{\sigma}$ denote the positive square root of $\hat{\sigma}^2$.

The most popular test of H_0 in econometrics is probably the Jarque-Bera LM test (Jarque and Bera, 1980, 1987); see Dufour et al. (1998) for a discussion of this procedure and other tests. The Jarque-Bera statistic, denoted by JB , is defined as follows:

$$JB = n \left[(\sqrt{b_1})^2/6 + (b_2 - 3)^2/24 \right], \quad (3.3)$$

in which

$$\sqrt{b_1} = n^{-1} \sum_{i=1}^n \left(\frac{\hat{u}_i}{\hat{\sigma}} \right)^3 = \frac{n^{1/2} \sum_{i=1}^n \hat{u}_i^3}{\left[\sqrt{\sum_{i=1}^n \hat{u}_i^2} \right]^3}, \quad (3.4)$$

and

$$b_2 = n^{-1} \sum_{i=1}^n \left(\frac{\hat{u}_i}{\hat{\sigma}} \right)^4 = \frac{n \sum_{i=1}^n \hat{u}_i^4}{\left[\sum_{i=1}^n \hat{u}_i^2 \right]^2}. \quad (3.5)$$

Under H_0 , JB is asymptotically distributed as $\chi^2(2)$, but, as pointed out in Dufour et al. (1998), it does not have a tractable finite sample distribution. Unfortunately, as reported in Section 1.5.2, asymptotic theory cannot be relied upon to provide an accurate approximation.

It is clear from (3.3)–(3.5) that JB is simply a function of the OLS residuals and moreover that the value of JB would not be affected if every OLS residual \hat{u}_i were multiplied by a positive constant c , in other words, there is a function $g_{JB}(\cdot)$ such that

$$JB = g_{JB}(\hat{\mathbf{u}}) = g_{JB}(c\hat{\mathbf{u}}), \text{ for all } c > 0. \quad (3.6)$$

But, from (1.11) of Chapter 1, $\hat{\mathbf{u}} = (\mathbf{I}_n - \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}')\mathbf{u} = \mathbf{M}\mathbf{u}$; so that, setting $c = \sigma^{-1}$ in (3.6),

$$JB = g_{JB}(\mathbf{M}\mathbf{u}) = g_{JB}(\sigma^{-1}\mathbf{M}\mathbf{u}) = g_{JB}(\mathbf{M}\mathbf{z}), \quad (3.7)$$

in which $\mathbf{z} = \sigma^{-1}\mathbf{u}$. Under H_0 , $\mathbf{z}' = (z_1, \dots, z_n)$ is a vector with elements that are independent $N(0, 1)$ variables.

With access to a random number generator for the standard Normal distribution, a set of B independent n -dimensional vectors $\mathbf{z}_b^\dagger \sim N(\mathbf{0}_n, \mathbf{I}_n)$, $b = 1, \dots, B$, can be obtained. (All that is required is to use the random number generator nB times, with each call yielding a term denoted by z_{bi}^\dagger , $b = 1, \dots, B$ and $i = 1, \dots, n$.) Given the value of \mathbf{X} and hence of \mathbf{M} , these artificially generated vectors can then be used to calculate the artificial statistics $JB_b^\dagger = g_{JB}(\mathbf{M}\mathbf{z}_b^\dagger)$, $b = 1, \dots, B$. Conditionally upon the regressors, the terms JB_b^\dagger , $b = 1, \dots, B$, have the same finite sample distribution as JB , when H_0 is true. It is, therefore, possible to construct a Monte Carlo test of H_0 , as explained in Section 2.2.1, with a finite sample significance level which is equal to the desired significance level, the latter being denoted by α_d . As discussed in Section 2.2.1, it is important to choose B so that $(B + 1)\alpha_d$ is an integer. The value $B = 99$ is suitable for conventional values of α_d and no evidence that power is improved in important ways by using larger values is found in Dufour et al. (1998).

The process of implementing a Monte Carlo version of the Jarque-Bera statistic after OLS estimation of a regression model can be described in more detail and in more familiar terms as consisting of the following steps.

Monte Carlo Jarque-Bera test - Step 1

Use the actual data of $\mathbf{S} = (\mathbf{y}, \mathbf{X})$ to estimate (3.1) by OLS and to obtain the sample values of associated residuals, which are the elements of $\hat{\mathbf{u}} = (\hat{u}_1, \dots, \hat{u}_n)'$.

Monte Carlo Jarque-Bera test - Step 2

The residuals from step 1 are used to calculate the sample value of the Jarque-Bera statistic, denoted by \widehat{JB} ; see (3.3)–(3.5) above for the relevant formulae.

If, instead of relying upon the asymptotically valid $\chi^2(2)$ distribution for critical values, an exact Monte Carlo approach is adopted for assessing the statistical significance of \widehat{JB} , it is necessary to apply repeatedly steps like 1 and 2 to simulated samples of data. These samples must yield test statistics that have the same finite sample distribution, given \mathbf{X} , as \widehat{JB} when the null hypothesis is true, that is, the errors of (3.1) are $NID(0, \sigma^2)$. Let these statistics be denoted by $JB_b^\dagger, b = 1, \dots, B$.

The calculation of JB_b^\dagger is based on a corresponding simulated sample $\mathbf{S}_b^\dagger = (\mathbf{y}_b^\dagger, \mathbf{X})$, for $b = 1, \dots, B$. Each of these simulated samples comes from the same artificial population, which can be written as

$$\mathbf{y}^\dagger = \mathbf{X}\boldsymbol{\beta}^\dagger + \mathbf{u}^\dagger, \quad (3.8)$$

in which the errors in \mathbf{u}^\dagger are IID and come from a Normal distribution with zero mean and finite positive variance. Thus simulated data are drawn from a population which reflects the null hypothesis that is to be tested using the actual data. Clearly the researcher must have access to a subroutine intended to give random numbers that can be taken to be independent drawings from a Normal distribution with specified mean and variance. The mean of the Normal distribution should be set equal to zero. The value used for the variance is irrelevant since changes of scale leave the value of the Jarque-Bera statistic unaltered; see (3.6) and the associated comments. Drawings from the standard Normal distribution can, therefore, be used without any loss of generality. Consequently, if the error term for the i th observation of the b th simulated sample is denoted by u_{bi}^\dagger , $B \times n$ independent drawings from the $N(0, 1)$ distribution can be used as values for the errors $u_{bi}^\dagger, b = 1, \dots, B$ and $i = 1, \dots, n$.

Given error terms, simulated data on the dependent variable can be calculated from (3.8), provided that the value of $\boldsymbol{\beta}^\dagger$ is specified. As noted above, the Jarque-Bera statistic is just a function of OLS residuals and so, when calculated from Monte Carlo world data, is independent of the value of $\boldsymbol{\beta}^\dagger$ since

$$\begin{aligned} \hat{\mathbf{u}}^\dagger &= (\mathbf{I}_n - \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}')\mathbf{y}^\dagger \\ &= (\mathbf{I}_n - \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}')(\mathbf{X}\boldsymbol{\beta}^\dagger + \mathbf{u}^\dagger) = (\mathbf{I}_n - \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}')\mathbf{u}^\dagger, \end{aligned}$$

using $(\mathbf{I}_n - \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}')\mathbf{X}\boldsymbol{\beta}^\dagger = \mathbf{0}_n$. For the purpose of generating values of the dependent variable, the choice $\boldsymbol{\beta}^\dagger = \mathbf{0}_k$ is convenient and implies no loss of generality. Consequently, simulated data can be generated using $\mathbf{y}^\dagger = \mathbf{u}^\dagger$, with $\mathbf{u}^\dagger \sim N(\mathbf{0}_n, \mathbf{I}_n)$.

Steps 3, 4 and 5 for the b th of the B repetitions can be described as follows, with steps 4 and 5 being the counterparts in the simulation world of steps 1 and 2 for actual data.

Monte Carlo Jarque-Bera test - Step 3

Use a random number generator for the $N(0, 1)$ distribution to get the n values required for a realization of $\mathbf{u}_b^\dagger = (u_{b1}^\dagger, u_{b2}^\dagger, \dots, u_{bn}^\dagger)'$. Set $\mathbf{y}_b^\dagger = \mathbf{u}_b^\dagger$, that is, $y_{bi}^\dagger = u_{bi}^\dagger$ for $i = 1, \dots, n$. The data for the b th artificial sample $\mathbf{S}_b^\dagger = (\mathbf{y}_b^\dagger, \mathbf{X})$ are now available.

Monte Carlo Jarque-Bera test - Step 4

Use an appropriate computer routine to regress \mathbf{y}_b^\dagger on \mathbf{X} in order to obtain the associated OLS residual vector $\hat{\mathbf{u}}_b^\dagger = (\hat{u}_{b1}^\dagger, \hat{u}_{b2}^\dagger, \dots, \hat{u}_{bn}^\dagger)'$.

Monte Carlo Jarque-Bera test - Step 5

Use the OLS residuals from step 4 to calculate a test statistic JB_b^\dagger ; see (3.3).

Monte Carlo Jarque-Bera test - Step 6

Having carried out the full set of B repetitions of steps 3, 4 and 5, the final step is to calculate the Monte Carlo p -value of \hat{JB} , that is,

$$MCPV_{JB} = \frac{\sum_{b=1}^B \mathbf{1}(JB_b^\dagger \geq \hat{JB}) + 1}{B + 1}, \quad (3.9)$$

in which $\mathbf{1}(A)$ is the indicator variable that is equal to 1 if the event A is true and is otherwise equal to zero. The rejection rule for the Monte Carlo Jarque-Bera test is that the null hypothesis of Normal errors should be rejected as data-inconsistent if $MCPV_{JB} \leq \alpha_d$.

For the sake of exposition, the application of the Monte Carlo version of the Jarque-Bera test has been described in steps 1 to 6 as requiring $(1 + B)$ OLS regressions to be fitted. However, the matrix \mathbf{X} is fixed from the first and so the residual vectors $\hat{\mathbf{u}}_b^\dagger$ can be computed by pre-multiplying \mathbf{u}_b^\dagger by the constant matrix $\mathbf{M} = (\mathbf{I}_n - \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}')$ for $b = 1, \dots, B$. Consequently it is only necessary to invert $(\mathbf{X}'\mathbf{X})$ once (for step 1) so that the value of the fixed matrix \mathbf{M} can be stored for use in the B repetitions of the calculation of $\hat{\mathbf{u}}_b^\dagger = \mathbf{M}\mathbf{u}_b^\dagger$. It would, therefore, not be

computationally efficient to translate steps 1 to 6 directly into computer code. However, computers are now so powerful that the cost of inefficiency in terms of waiting time for a single application would probably be small. Considerations of computational efficiency are more important when finite sample properties under null and alternative hypotheses are investigated using simulation studies. In such studies, steps 1 to 6 would be repeated many thousands of times; see, for example, Dufour et al. (1998, section 4) for a description of a simulation study in which each finite sample rejection probability of interest is estimated using 10,000 replications.

The use of exact Monte Carlo tests for regression models is by no means restricted to the problem of testing the assumption of Normality. It is possible to apply arguments similar to those in Dufour et al. (1998) to other types of test statistic, provided the test statistic is exactly pivotal. For example, MacKinnon describes how the Monte Carlo approach can be used to derive an exact form of the widely-used Durbin-Watson test, under the maintained assumption that each error is $N(0, \sigma^2)$; see MacKinnon (2002, pp. 618–619). The scope for exact Monte Carlo tests is, however, limited by the restriction that the IID error terms of the linear regression model must have a distribution which is specified, up to knowledge of σ^2 , either by the null hypothesis or by an untested maintained hypothesis. It is not clear that applied econometricians will be able or willing to make such strong assumptions about error distributions and, therefore, the use of nonparametric bootstrap tests may be of greater interest. In the next section, which draws on the work reported in Godfrey et al. (2006), evidence is provided on the usefulness of nonparametric bootstrap techniques and the lack of robustness of Monte Carlo tests in the context of checks for heteroskedasticity.

3.3. Simulation-based tests for heteroskedasticity

There is an extensive literature on the construction, implementation and interpretation of tests for heteroskedasticity in the errors of linear regression models, which is usefully summarized in a recent paper by Dufour et al. (2004). As observed in that paper, most test procedures employ asymptotically valid critical values. Many researchers have carried out simulation experiments in order to learn about the finite sample behaviour of these asymptotic tests. The evidence that has been reported indicates that standard asymptotic distributions can be an unreliable basis for inference and that bootstrapping produces useful improvements in finite sample behaviour; see, for example, Cribari-Neto and Zarkos

(1999), Godfrey and Orme (1999) and Jeong and Lee (1999). However, checks for heteroskedasticity that are derived using a nonparametric bootstrap remain only asymptotically valid. In contrast, the results of Dufour et al. (2004) show that, when the distribution of the error of the regression model is known, or forms part of the null hypothesis, the test criteria for many homoskedasticity tests are exactly pivotal under the null. In such cases, it is possible to use Monte Carlo techniques to eliminate completely the discrepancy between actual and desired significance levels. The practical question that must be faced when considering Monte Carlo tests for heteroskedasticity is how robust they are to incorrect specification of the error distribution.

As in the previous section, the data for the dependent variable are assumed to be generated by the linear regression model (3.1). When constructing tests for heteroskedasticity, the errors of (3.1) are often written as

$$u_i = \sigma_i \varepsilon_i, 0 < \sigma_i < \infty, \quad (3.10)$$

in which the terms ε_i are IID, with CDF denoted by \mathcal{F}_ε , having zero mean and variance equal to one, $i = 1, \dots, n$. The null hypothesis for such tests is then

$$H_0^h : \sigma_i = \sigma, \text{ for all } i.$$

Monte Carlo tests for heteroskedasticity could be motivated by arguing that they should be regarded as general (omnibus) tests of a joint null hypothesis that comprises not only homoskedasticity but also correctness of mean function and the general form of the CDF of IID errors. However, this interpretation is not the conventional view of the standard tests for heteroskedasticity and there is little reason to believe that these tests would be useful for departures from the joint null which involve either incorrect mean functions or misspecified error distributions; see Davidson and MacKinnon (1985b) and Godfrey and Orme (1994, 1996).

In this section, therefore, Monte Carlo tests for heteroskedasticity are taken to be tests of H_0^h alone and to be based upon an untested supporting assumption that specifies the CDF of the errors of the null model. Such supporting assumptions are used in some of the best known tests for heteroskedasticity. Some of these tests use test statistics that, under an unspecified error CDF, are not asymptotically pivotal. Given that such test statistics are frequently discussed in textbooks and quite popular in empirical work, they will be included in this Section even though they are not asymptotically pivotal.

The LM test proposed in Breusch and Pagan (1979) is one of the most widely cited checks for heteroskedasticity. It is based upon the following untested assumptions: first, the conditional mean function is given by

$$E(\mathbf{y}|\mathbf{X}) = \mathbf{X}\boldsymbol{\beta}; \quad (3.11)$$

second, the regressors in (3.11) are strictly exogenous; third, the errors in $\mathbf{y} - \mathbf{X}\boldsymbol{\beta}$ are independent Normal variables, with common mean zero, and may be heteroskedastic; and fourth, the error variances are determined by

$$\text{Var}(u_i|z_{i1}, \dots, z_{iq}) = \sigma_i^2 = h(\gamma_0 + \sum_{j=1}^q z_{ij}\gamma_j), i = 1, \dots, n, \quad (3.12)$$

in which $h(\cdot)$ is a function with $h'(\gamma_0) \neq 0$ and the terms z_{ij} are observations on strictly exogenous variables that satisfy the regularity conditions of Breusch and Pagan (1979). The null hypothesis H_0^h is equivalent to the q restrictions of $\gamma_1 = \dots = \gamma_q = 0$, which imply $\sigma_i^2 = h(\gamma_0) = \sigma^2, i = 1, \dots, n$.

The Breusch-Pagan test statistic, denoted here by *BP*, is one-half of the explained sum of squares from the OLS estimation of the artificial regression

$$\frac{\hat{u}_i^2}{\hat{\sigma}^2} = \gamma_0 + \sum_{j=1}^q z_{ij}\gamma_j + \text{residual}, \quad (3.13)$$

and, if the null hypothesis is true, it is asymptotically distributed as $\chi^2(q)$, with the rejection region being in the right-hand side of this reference distribution. It is very important to note that the assumption that the errors have a Normal distribution is used in Breusch and Pagan (1979) to establish the asymptotic validity of this test. Normality implies that $E(u_i^4) = 3\sigma^4$, under H_0^h , which is a necessary condition for the Breusch-Pagan test to be asymptotically valid.

A second test that is often discussed in textbooks is the procedure given in Glejser (1969). The *Glejser test* is based upon the assumption that it is asymptotically valid to apply a conventional test of $\gamma_1 = \dots = \gamma_q = 0$ after OLS estimation of the artificial regression

$$|\hat{u}_i| = \gamma_0 + \sum_{j=1}^q z_{ij}\gamma_j + \text{residual}, \quad (3.14)$$

in which the regressors of (3.14) satisfy the same regularity conditions as those of (3.13). However, in order for such a test to be asymptotically valid, the error CDF \mathcal{F}_ε must be such that, under H_0^h , $\Pr(\varepsilon_i \geq 0) = 0.5$; see Godfrey (1996). The symmetry of the error distribution would, therefore, be sufficient for asymptotic validity of Glejser's test. The test statistic for Glejser's procedure is denoted by G .

The third and final test to be considered is *White's direct test* for heteroskedasticity; see White (1980). White shows how to construct an asymptotically valid test that is designed to detect any form of heteroskedasticity that invalidates conventional (homoskedasticity-valid) OLS-based inference. Under weak conditions that do not require specification of the general shape of the error CDF and do not require that $\Pr(\varepsilon_i \geq 0) = 0.5$, White proves that an asymptotically valid test can be derived after OLS estimation of the artificial regression model

$$\hat{u}_i^2 = \gamma_0 + \sum_{j=1}^r w_{ij}\gamma_j + \text{residual}, \quad (3.15)$$

in which the terms w_{ij} are the nonredundant variables from the squared values and cross-products of the regressors of (3.1), and r is the number of such variables (so that $r \leq k(k+1)/2$). If the R^2 -statistic for (3.15) is denoted by R_W^2 , the test statistic $W = nR_W^2$ is asymptotically distributed as $\chi^2(r)$, under the null hypothesis; see White (1980, section 3). Thus, of the three tests, only White's procedure is based upon a test statistic that is asymptotically pivotal. The asymptotic null distributions of BP and G , under H_0^h , depend (in different ways) upon the error CDF \mathcal{F}_ε , which, following Beran (1988), is regarded as part of the parameter vector.

3.3.1. Monte Carlo tests for heteroskedasticity

It is clear from (3.13), (3.14) and (3.15) that the test statistics BP_Z , G_Z and W_W depend upon the data of \mathbf{y} only through the OLS residuals $\hat{\mathbf{u}} = \mathbf{M}\mathbf{y} = (\mathbf{I}_n - \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}')\mathbf{y}$. Under the auxiliary assumption that the mean function (3.11) is specified correctly, $\hat{\mathbf{u}} = \mathbf{M}\mathbf{u}$; so that each of the test statistics is independent of $\boldsymbol{\beta}$. Moreover, as pointed out in Dufour et al. (2004), these statistics are also independent of the value of σ , when H_0^h is true. Consequently, corresponding to (3.7) which defines the Jarque-Bera statistic, the checks for heteroskedasticity can be written as:

$$BP = g_{BP}(\sigma^{-1}\mathbf{M}\mathbf{u}) = g_{BP}(\mathbf{M}\boldsymbol{\varepsilon}); \quad (3.16)$$

$$G = g_G(\sigma^{-1}\mathbf{M}\mathbf{u}) = g_G(\mathbf{M}\boldsymbol{\varepsilon}); \quad (3.17)$$

and

$$W = g_W(\sigma^{-1}\mathbf{Mu}) = g_W(\mathbf{M}\boldsymbol{\varepsilon}), \quad (3.18)$$

in which $\boldsymbol{\varepsilon} = (\varepsilon_1, \dots, \varepsilon_n)'$.

In the Monte Carlo test approach, it is assumed that the true error distribution is specified; see Dufour et al. (2004, eq. 3). Let the CDF for $\boldsymbol{\varepsilon}$ used for the Monte Carlo test be denoted by \mathcal{G}_ε . For any given statistic, derived from (3.1), which has a distribution that is independent of $\boldsymbol{\beta}$ and σ^2 , the Monte Carlo approach generates test statistics which, conditional upon the values of exogenous variables, possess the same finite sample distribution, when H_0^h is true and $\mathcal{G}_\varepsilon = \mathcal{F}_\varepsilon$.

More precisely, since, under H_0^h , there is no loss of generality implied by setting regression coefficients equal to zero and the error variance equal to one, B samples of simulation data can be generated using $\mathbf{y}_b^\dagger = \boldsymbol{\varepsilon}_b^\dagger$, with the n elements of

$$\boldsymbol{\varepsilon}_b^\dagger = (\varepsilon_{b1}^\dagger, \dots, \varepsilon_{bn}^\dagger)',$$

being obtained using a random number generator that mimics the process of drawing IID terms with CDF \mathcal{G}_ε , $b = 1, \dots, B$.

Simulation test statistics can then be calculated using the generated simulation data for the dependent variable and the actual values of the exogenous variables. The values of these test statistics are given by

$$BP_b^\dagger = g_{BP}(\mathbf{M}\boldsymbol{\varepsilon}_b^\dagger); \quad (3.19)$$

$$G_b^\dagger = g_G(\mathbf{M}\boldsymbol{\varepsilon}_b^\dagger); \quad (3.20)$$

and

$$W_b^\dagger = g_W(\mathbf{M}\boldsymbol{\varepsilon}_b^\dagger), \quad (3.21)$$

for $b = 1, \dots, B$. If $\mathcal{G}_\varepsilon = \mathcal{F}_\varepsilon$ and H_0^h is true, the B simulation values of any test statistic can be combined with the test statistic from the actual data to form a simple random sample of size $B + 1$. The rejection rules for Monte Carlo tests for BP , G and W are based upon the Monte Carlo p -values defined by

$$MCPV_{BP} = \frac{\sum_{b=1}^B \mathbf{1}(BP_b^\dagger \geq BP) + 1}{B + 1},$$

$$MCPV_G = \frac{\sum_{b=1}^B \mathbf{1}(G_b^\dagger \geq G) + 1}{B + 1},$$

and

$$MCPV_W = \frac{\sum_{b=1}^B \mathbf{1}(W_b^\dagger \geq W) + 1}{B + 1},$$

respectively. For each test statistic, the null hypothesis of homoskedasticity is rejected if the Monte Carlo p -value is less than or equal to the desired significance level, denoted by α_d . Under regularity conditions provided in Dufour et al. (2004), which include the requirement that the specified CDF \mathcal{G}_ε is correct, this rule provides an exact test of H_0^h when $\alpha_d(B + 1)$ is an integer.

Godfrey et al. consider the robustness of Monte Carlo tests for heteroskedasticity when an incorrect CDF has been assumed, that is, $\mathcal{G}_\varepsilon \neq \mathcal{F}_\varepsilon$; see Godfrey et al. (2006, section 2.3). They base their investigation on the general approach to asymptotic analysis which is used in Beran (1988) and come to the following conclusions about Monte Carlo tests of H_0^h :

1. if the test statistic is not asymptotically pivotal, the Monte Carlo test derived using the wrong CDF for the error distribution has an ERP which is $O(1)$ and therefore delivers asymptotically invalid inferences; and
2. if the test statistic is asymptotically pivotal, the Monte Carlo test derived using the wrong CDF for the error distribution is asymptotically valid and has an ERP which is of the same order in n as the test using asymptotically valid critical values.

Thus, in the case of White's direct test statistic, which is an asymptotic pivot, a Monte Carlo test with $\mathcal{G}_\varepsilon \neq \mathcal{F}_\varepsilon$ has the correct asymptotic significance level, but enjoys no refinement relative to the asymptotic test that uses critical values from the $\chi^2(r)$ distribution. The asymptotic significance levels of Monte Carlo versions of Breusch-Pagan and Glejser tests are not, in general, equal to the desired value when $\mathcal{G}_\varepsilon \neq \mathcal{F}_\varepsilon$. Godfrey et al. use asymptotic theory to illustrate this lack of robustness, using the example of the Breusch-Pagan test; see Godfrey et al. (2006, pp. 83–84). They show that, if the kurtosis implied by \mathcal{F}_ε is smaller (respectively, larger) than that implied by \mathcal{G}_ε , then, under homoskedasticity, the Monte Carlo version of the Breusch-Pagan test procedure will yield

asymptotic rejection probabilities which are smaller (respectively, larger) than the desired significance level α_d .

3.3.2. Bootstrap tests for heteroskedasticity

As an alternative to the Monte Carlo approach, simulation-based tests for heteroskedasticity can be conducted using a nonparametric bootstrap procedure. The results in Beran (1988) indicate that, corresponding to conclusions 1 and 2 of the previous subsection on Monte Carlo tests, bootstrap methods have the following characteristics:

1. if the test statistic is not asymptotically pivotal, the nonparametric bootstrap test of H_0^h is asymptotically valid and has an ERP which is of the same order in n as that of the procedure that uses critical values from the limit null distribution; and
2. if the test statistic is asymptotically pivotal, the nonparametric bootstrap test of H_0^h is asymptotically valid and has an ERP which is of smaller order in n than that of the procedure that uses critical values from the limit null distribution.

Comparisons of the properties of Monte Carlo and nonparametric bootstrap tests when $\mathcal{G}_\varepsilon \neq \mathcal{F}_\varepsilon$ indicates that the former are inferior to the latter in terms of asymptotic properties. It is, therefore, important not to use Monte Carlo tests, rather than bootstrap tests, unless there is very precise information about the error distribution or this distribution is specified as part of the hypothesis under test.

Nonparametric bootstrap tests for heteroskedasticity can be carried out using OLS results. Bootstrap conditional mean values are set equal to the OLS predicted values from actual data and bootstrap world errors are obtained by resampling either the actual OLS residuals or some transformations of these residuals; see (2.27), (2.30), (2.31) and (2.32) in Chapter 2. Thus the B bootstrap samples of size n can be generated from

$$y_{bi}^* = \mathbf{x}_i' \hat{\boldsymbol{\beta}} + u_{bi}^*, \quad i = 1, \dots, n,$$

where $u_{b1}^*, u_{b2}^*, \dots, u_{bn}^*$ is a random sample drawn, with replacement, from an asymptotically valid OLS residual-based EDF. If (3.1) does not have an intercept term, the OLS residuals must be recentred before they are used in a resampling scheme.

Bootstrap test statistics BP_b^* , G_b^* and W_b^* , $b = 1, \dots, B$, can then be calculated and the p -values of the corresponding actual test statistics can be

estimated by

$$BSPV_{BP} = \frac{\sum_{b=1}^B \mathbf{1}(BP_b^* \geq BP)}{B},$$

$$BSPV_G = \frac{\sum_{b=1}^B \mathbf{1}(G_b^* \geq G)}{B},$$

and

$$BSPV_W = \frac{\sum_{b=1}^B \mathbf{1}(W_b^* \geq W)}{B},$$

respectively. The null hypothesis of homoskedasticity is then rejected when $BSPV_\tau \leq \alpha_d$, where α_d is the desired significance level, $\tau = BP, G, W$. If this rule leads to bootstrap tests that have finite sample significance levels that are close to desired values, there is little incentive to risk using the Monte Carlo approach.

3.3.3. Simulation experiments and tests for heteroskedasticity

Simulation evidence on the relative merits of asymptotic, bootstrap and Monte Carlo tests for heteroskedasticity in finite samples can be obtained using the designs of the experiments reported in Dufour et al. (2004). In these experiments, the simulation data generation process, under H_0^h , can be written as

$$y_i = \sum_{j=1}^6 x_{ij} \beta_j + u_i, \quad u_i \text{ IID}(0, \sigma^2), \quad i = 1, \dots, n, \quad (3.22)$$

in which: $x_{i1} = 1$ for all i , so that β_1 is an intercept term; without loss of generality, $\beta_j = 1$ for all j ; and $n = 50, 100$. The regressor values x_{i2}, \dots, x_{i6} are independent drawings from the uniform distribution $U(0, 10)$ for $i = 1, \dots, n$.

The specification of regression model coefficients, together with the regressor values, allows the calculation of conditional mean values $E(y_i | \mathbf{x}_i)$, $i = 1, \dots, n$. The addition of a pseudo-random error to this value of $E(y_i | \mathbf{x}_i)$ gives an artificial observation. The random number generators used to obtain errors correspond to the following distributions: Normal; Student $t(5)$; and $\chi^2(2)$. Monte Carlo tests are derived using each of these three error distributions and also the Uniform and Lognormal distributions.

Having combined errors and means to obtain an artificial sample of n observations, (3.22) can be estimated by OLS and tests can be carried out. When Monte Carlo test techniques are employed, 399 Monte Carlo samples are generated for each of the five possible error distributions. Thus, for any given correct choice of the error distribution, there are also four incorrect models being used. As an alternative to using a parametric approach, the nonparametric bootstrap is implemented, as in Godfrey and Orme (1999), with 400 bootstrap samples. When implementing the nonparametric bootstrap, bootstrap errors are drawn as random samples, with replacement, from

$$\hat{\mathcal{F}}_{BS} : \text{probability } \frac{1}{n} \text{ on } \left(\frac{\hat{u}_i}{\sqrt{(1-h_{ii})}} - \frac{1}{n} \sum_{j=1}^n \frac{\hat{u}_j}{\sqrt{(1-h_{jj})}} \right), i = 1, \dots, n,$$

which corresponds to (2.32) of Chapter 2; see Davison and Hinkley (1997, page 275).

As is clear from (3.13) and (3.14), researchers wishing to use either Breusch-Pagan or Glejser statistics must choose a set of exogenous test variables z_{ij} . The version of the Breusch-Pagan statistic which is used in the experiments is computed using x_{i2}, \dots, x_{i6} from (3.22) for this purpose. This statistic is denoted by BP_x . When the error terms ε_t are IID and Normally distributed, BP_x is asymptotically distributed as $\chi^2(5)$. When the error terms ε_t are IID, but not Normally distributed, BP_x is not, in general, asymptotically distributed as $\chi^2(5)$; see Godfrey and Orme (1999, p. 174).

Following Dufour et al. (2004), the statistic for the Glejser test used here is the conventional F -statistic for testing that all slope coefficients equal zero in the artificial regression of $|\hat{u}_t|$ on the regressors of (3.22). The test statistic is denoted by G_x . Since it has a limit null distribution that depends upon characteristics of the error distribution, G_x is not asymptotically pivotal. The results in Godfrey (1996) imply that, if the null hypothesis is true and the errors have a symmetric distribution, G_x is asymptotically distributed as $\chi^2(5)/5$; so that critical values from the $F(5, n-6)$ distribution are asymptotically valid.

Given the form of (3.22), application of White's direct test would involve appealing to asymptotic theory to justify using critical values from a $\chi^2(20)$ distribution; see White (1980, section 3). In practical situations, it seems reasonable to try to ensure that orders of magnitude in the relevant asymptotic theory have some connection with the actual values implied by the data. In White's asymptotic analysis, r is fixed and

so $r/n \rightarrow 0$ as $n \rightarrow \infty$. Consequently, r/n should be small if appeal is to be made to this analysis. In the experiments, $r = 20$ and the sample sizes are $n = 50$ and $n = 100$. It seems useful to reduce the number of restrictions being tested. A modification of White's test is, therefore, examined. The modified version of White's test uses only the nonredundant levels and squares of regressors in (3.1) as the regressors of the artificial model (3.15). The asymptotic critical values for the modified test are, therefore, taken from the $\chi^2(10)$ distribution when data are generated by (3.22). The test statistic for this modification of the direct test is denoted by W_m .

Tables 3.1 to 3.3 contain representative samples of the results that are reported in Godfrey et al. (2006). In these tables, the desired significance level for all of the tests is set to 0.05, in other words, 5 per cent, as in the experiments in Dufour et al. (2004). Estimates of rejection probabilities are derived from 25,000 replications. The tests consist of "standard" versions, as well as bootstrap and Monte Carlo procedures. In "standard" tests, critical values for BP_x , G_x and W_m are taken from the $\chi^2(5)$, $F(5, n - 6)$ and $\chi^2(10)$ distributions, respectively, as would be suggested by a conventional textbook treatment. The bootstrap tests, denoted by $BSPV$, use the bootstrap p -value calculated using 400 bootstrap samples. The Monte Carlo tests combine the actual value of each test statistic with 399 corresponding artificial values so that $0.05 \times (399 + 1)$ is an integer, as required for an exact test. The notation for Monte Carlo tests is that $MC|Normal$ denotes the test with an assumed CDF \mathcal{G}_ε derived by standardizing a Normal distribution, with the other tests being defined using the same general notation. In each of the three tables of results, the estimates for the Monte Carlo test based upon the correct assumed error CDF are given in bold font.

In Table 3.1, the true error distribution is Normal and is, therefore, symmetric. It follows that all standard tests are asymptotically valid for the cases of Table 3.1. Asymptotic critical values give quite good control of finite sample significance levels for BP_x and G_x , but provide a poorer approximation for W_m . This feature of the results may reflect the fact that the modified White check W_m tests 10 restrictions, whereas the other two procedures only test 5 restrictions. The corresponding bootstrap tests work quite well, although the estimates for the bootstrap Breusch-Pagan test suggest that it is a little undersized. The first row of estimates for a Monte Carlo test corresponds to the correct choice of CDF and, in this case, the Monte Carlo test is exact, with estimates that are close to the desired level. However, the remaining results for Monte Carlo tests show the consequences of using the wrong error CDF.

The test statistics BP_x and G_x are not asymptotically pivotal and so Monte Carlo tests that use an incorrect CDF are not, in general, asymptotically valid; see Godfrey et al. (2006). The last four rows of results in Table 3.1 indicate how rejection rates may be far too low or far too high when an inappropriate Monte Carlo approach is applied to BP_x , which is not asymptotically distributed as χ^2 under any of the false Monte Carlo error distributions. The discussion in Godfrey et al. (2006, pp. 83–84) concerning the properties of inappropriate Monte Carlo tests of the Breusch-Pagan statistic is pertinent. Results for Monte Carlo tests using the Glejser statistic G_x and an incorrect distribution depend upon whether or not the false distribution is symmetric. If the true distribution is symmetric, as it is for Table 3.1, picking a different symmetric error distribution has asymptotically negligible effects on the rejection probability for G_x ; see Godfrey (1996). However, using an asymmetric distribution to carry out a Monte Carlo test will produce rejection rates that do not converge to desired levels. These predictions are borne out by the estimates in Table 3.1: $t(5)$ and uniform errors yield estimates close to 5 per cent, but $\chi^2(2)$ and lognormal distributions in Monte Carlo schemes lead to substantial under-rejection. The estimates of Table 3.1 also show that, with the asymptotically pivotal statistic W_m , the Monte Carlo approach is much more robust to incorrect choice of error CDF, which is consistent with the asymptotic analysis of Godfrey et al. (2006). However, the inappropriate Monte Carlo tests based upon W_m do not, in general, match the performance of the corresponding bootstrap test.

The symmetric standardized $t(5)$ distribution is used to produce the errors in the experiments that yield the estimates of Table 3.2. In contrast

Table 3.1 Estimates of rejection probabilities of standard, Monte Carlo and nonparametric bootstrap tests for heteroskedasticity: true error CDF is from standardized Normal and $\alpha_d = 5$ per cent

<i>Test</i>	<i>n</i> = 50			<i>n</i> = 100		
	BP_x	G_x	W_m	BP_x	G_x	W_m
<i>Standard</i>	4.77	5.30	3.48	5.02	5.28	4.22
<i>BSPV</i>	3.79	4.63	5.03	4.83	5.04	5.40
<i>MC Normal</i>	5.12	4.88	5.07	5.02	5.07	5.13
<i>MC t(5)</i>	0.09	4.78	4.81	0.00	4.95	5.19
<i>MC uniform</i>	27.15	4.81	4.85	36.43	4.97	4.87
<i>MC $\chi^2(2)$</i>	0.00	0.59	3.36	0.00	0.32	4.17
<i>MC lognormal</i>	0.00	0.29	3.24	0.00	0.17	4.15

to Table 3.1, the standard form of the Breusch-Pagan test is no longer asymptotically valid and the estimates indicate that the true rejection probability is much greater than the desired value when critical values are taken from the $\chi^2(5)$ distribution. The Glejser and White test based upon G_x and W_m are both asymptotically valid under the $t(5)$ distribution. Asymptotic critical values, however, do not seem to give very accurate approximations. The estimates for the standard test using G_x are a little too high, while those for the standard version based upon W_m are too low. As anticipated, the performance of the asymptotically valid tests is better for $n = 100$ than for $n = 50$.

Using the bootstrap, rather than invalid asymptotic critical values, with BP_x gives much better agreement between estimates and the desired level; but, since BP_x is not asymptotically pivotal, the bootstrap enjoys no refinement relative to the correct asymptotic test, see Beran (1988). The bootstrap works better with G_x and W_m , giving good agreement between estimated and desired significance levels. The estimated rejection rates for bootstrap versions of G_x and W_m are all in the range $\alpha_d \pm 0.1\alpha_d$.

The evidence in Table 3.2 concerning the behaviour of correct and incorrect Monte Carlo tests is similar to that provided by Table 3.1 and is consistent with the predictions of the asymptotic analysis in Godfrey et al. (2006). First, inappropriate Monte Carlo tests of BP_x produce estimates in the range 0.13 per cent to 68.58 per cent. Second, the consequences of using the wrong symmetric distribution with the Glejser test are much less serious than those of employing an asymmetric distribution to generate data for the Monte Carlo samples. Third, Monte Carlo tests of W_m are asymptotically valid, whatever the choice of CDF, but overall their finite sample behaviour is not as good as that of the bootstrap check.

Table 3.2 Estimates of rejection probabilities of standard, Monte Carlo and nonparametric bootstrap tests for heteroskedasticity: true error CDF is from standardized $t(5)$ and $\alpha_d = 5$ per cent

Test	$n = 50$			$n = 100$		
	BP_x	G_x	W_m	BP_x	G_x	W_m
Standard	22.94	5.77	3.74	31.84	5.46	4.27
BSPV	7.11	4.84	5.18	6.24	4.84	5.13
MC Normal	23.66	5.45	5.35	31.92	5.23	5.14
MC $t(5)$	5.04	5.18	5.16	5.34	5.16	5.03
MC uniform	52.13	5.14	5.25	68.58	5.07	4.94
MC $\chi^2(2)$	2.38	0.60	3.58	2.83	0.38	4.24
MC lognormal	0.16	0.35	3.48	0.13	0.14	4.23

Table 3.3 contains the estimates when errors are linear transformations of IID $\chi^2(2)$ variables and so have a heavily skewed distribution. The only standard test that is asymptotically valid under the error distribution for Table 3.3 is W_m . The standard versions of BP_x and G_x appear to have excessively high rejection probabilities. The nonparametric bootstrap is only moderately successful in eliminating the problem for these well-known tests, which use statistics that are not asymptotically pivotal.

The estimates for $MC|\chi^2(2)$ reflect the fact that it is exactly valid for all three statistics in the cases of Table 3.3. As in Tables 3.1 and 3.2, the use of the wrong error distribution to derive the Monte Carlo p -value of BP_x leads to estimates that are far from the desired value of 5 per cent: the relevant ranges of estimates in Table 3.3 are 0.15 per cent to 71.82 per cent for $n = 50$ and 0.02 per cent to 84.06 per cent for $n = 100$. When used with G_x in models with asymmetric errors, Monte Carlo tests that are based upon the wrong assumption about the form of the distribution are not, in general, asymptotically valid. The estimates for G_x in Table 3.3 provide very clear examples of the lack of robustness. As in Tables 3.1 and 3.2, inappropriate Monte Carlo schemes provide much closer agreement with the desired level when used with the asymptotic pivot W_m .

The results of Tables 3.1 to 3.3 provide information about the behaviour of Monte Carlo and bootstrap methods, both of which can be used to replace asymptotic critical values when testing for heteroskedasticity. The relative merits of these two simulation-based approaches depend upon the properties of the test statistic that is being used to check

Table 3.3 Estimates of rejection probabilities of standard, Monte Carlo and nonparametric bootstrap tests for heteroskedasticity: true error CDF is from standardized $\chi^2(2)$ and $\alpha_d = 5$ per cent

Test	$n = 50$			$n = 100$		
	BP_x	G_x	W_m	BP_x	G_x	W_m
Standard	41.13	21.11	5.02	53.32	24.10	5.02
BSPV	11.62	10.27	5.90	9.02	8.44	5.17
MC Normal	41.85	20.19	7.06	53.28	23.63	5.92
MC t(5)	10.70	19.49	6.74	10.60	23.24	5.81
MC uniform	71.82	19.40	6.76	84.06	23.32	5.69
MC $\chi^2(2)$	4.78	4.63	4.97	4.94	4.86	4.86
MC lognormal	0.15	3.04	4.84	0.02	2.88	5.08

for heteroskedasticity and the validity of the fixed error distribution that underpins the generation of Monte Carlo samples.

If the test statistic is exactly pivotal, the Monte Carlo approach offers the opportunity to have perfect control of finite sample significance levels. However, if the Monte Carlo data generation process is based upon the wrong assumption about the error distribution, the finite sample significance level will not equal the desired level and it will only tend to the desired value, as $n \rightarrow \infty$, if the test statistic is an asymptotic pivot. But, if the test statistic is an asymptotic pivot, the nonparametric bootstrap test is not only asymptotically valid but also enjoys an asymptotic refinement relative to an asymptotically valid Monte Carlo test that uses the wrong error CDF. Moreover, unlike a Monte Carlo test derived with the wrong error CDF, the nonparametric bootstrap test remains asymptotically valid when the test statistic is not an asymptotic pivot.

The evidence from the simulation experiments gives cause for real concern about the robustness of Monte Carlo methods. Given the uncertainty about the error distribution that is probably typical of applied work, it could be argued that Monte Carlo tests should only be used when the form of the error CDF is specified by the null hypothesis. The inclusion of the specification of the distribution in the null hypothesis is not common and seems to run counter to the persuasive arguments in Hansen (1999) about the role of distributional assumptions in modern econometrics. The remaining sections will, therefore, be restricted to discussions of nonparametric bootstrap tests and their implementation. An issue of interest in the implementation of such tests is the choice between restricted and unrestricted estimation when obtaining values for bootstrap world parameters from actual sample estimates. This issue will be considered first in the familiar context of the textbook F-test of linear coefficient restrictions.

3.4. Bootstrapping F tests of linear coefficient restrictions

3.4.1. Regression models with strictly exogenous regressors

It is very often the case that the unrestricted model of the alternative hypothesis can be written as (3.1), with the null hypothesis to be tested being that the regression coefficients satisfy a set of exact linear restrictions. In such a case, the null hypothesis is of the general form

$$H_0 : \mathbf{R}\boldsymbol{\beta} = \mathbf{r}, \quad (3.23)$$

in which \mathbf{R} and \mathbf{r} have elements that are known constants, with \mathbf{R} being $q \times k$ and \mathbf{r} being $q \times 1$, $q \leq k$. There are no redundant restrictions in the null hypothesis, so that \mathbf{R} has rank equal to q . The unrestricted OLS estimator for the alternative model (3.1) is denoted by $\hat{\boldsymbol{\beta}}$. The restricted least squares estimator that satisfies H_0 is denoted by $\tilde{\boldsymbol{\beta}}$. The n -dimensional restricted and unrestricted residual vectors are denoted by $\tilde{\mathbf{u}}$ and $\hat{\mathbf{u}}$, respectively. These residual vectors are used to obtain the residual sum of squares functions that define the F statistic for testing H_0 ; see equation (1.20).

Now it is well-known that

$$\hat{\boldsymbol{\beta}} = \boldsymbol{\beta} + (\mathbf{X}'\mathbf{X})^{-1} \mathbf{X}'\mathbf{u}, \quad (3.24)$$

and, from results contained in Greene (2008, section 5.3.2),

$$\tilde{\boldsymbol{\beta}} = \boldsymbol{\beta} + \left(\mathbf{I}_k - (\mathbf{X}'\mathbf{X})^{-1} \mathbf{R}' \left[\mathbf{R} (\mathbf{X}'\mathbf{X})^{-1} \mathbf{R}' \right]^{-1} \mathbf{R} \right) (\mathbf{X}'\mathbf{X})^{-1} \mathbf{X}'\mathbf{u}, \quad (3.25)$$

when H_0 is true. It follows that, given the value of \mathbf{X} , both estimators differ from the true value by a vector of linear combinations of the errors if (3.23) is valid. Consequently, when the errors of \mathbf{u} are IID with an unknown distribution and H_0 is true, the finite sample distributions of $\hat{\boldsymbol{\beta}}$ and $\tilde{\boldsymbol{\beta}}$, conditional upon \mathbf{X} , are unknown. Results for inference using standard distributions must, therefore, be derived using asymptotic theory.

It will be assumed in what follows that F tests of null hypotheses of the form (3.23) are asymptotically valid. General conditions for the asymptotic Normality of least squares estimators and the asymptotic validity of F tests of hypotheses like (3.23) have been provided in the statistics literature; see, for example, Arnold (1980) and Lai and Wei (1982). It is also assumed that bootstrapping the F -statistic yields an asymptotically valid test. Results on the convergence of bootstrap distributions to the required limits are available; see, for example, Freedman (1981) and Mammen (1992).

The general form of the F statistic for testing (3.23) is given by (1.20) of Chapter 1,

$$F = \frac{RSS(H_0) - RSS(H_1)}{RSS(H_1)} \cdot \frac{df(H_1)}{q},$$

in which RSS denotes a sum of squared residuals and $df(H_1)$ is the number of degrees of freedom, given here by $n - k$. It will be useful to note that

$$RSS(H_0) - RSS(H_1) = (\mathbf{R}\hat{\boldsymbol{\beta}} - \mathbf{r})' [\mathbf{R}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{R}']^{-1} (\mathbf{R}\hat{\boldsymbol{\beta}} - \mathbf{r});$$

so that the F statistic can be rewritten as

$$F = \frac{(\mathbf{R}\hat{\boldsymbol{\beta}} - \mathbf{r})' [\mathbf{R}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{R}']^{-1} (\mathbf{R}\hat{\boldsymbol{\beta}} - \mathbf{r})}{RSS(H_1)} \cdot \frac{df(H_1)}{q}. \quad (3.26)$$

The limit null distribution of F is $\chi^2(q)/q$, but many applied workers would probably use the asymptotically valid method of taking critical values from the $F(q, df(H_1))$ distribution. The application of nonparametric bootstrap techniques to the problem of implementing the F test is now considered.

In the nonparametric bootstrap approach, an artificial counterpart of the assumed actual DGP is used to study the sampling distribution of the test statistic. The assumed statistical model for the actual data has a parameter vector $\boldsymbol{\theta}' = (\boldsymbol{\beta}', \mathcal{F})$, where \mathcal{F} is the CDF for the error distribution and has mean equal to zero and variance equal to σ^2 , $0 < \sigma^2 < \infty$. The DGP for the bootstrap world is obtained, conditional upon the observed sample, by replacing $\boldsymbol{\theta}$ by a consistent estimator $\ddot{\boldsymbol{\theta}}$. The vector $\ddot{\boldsymbol{\theta}}$ is derived by combining an estimator of $\boldsymbol{\beta}$ and an estimator of \mathcal{F} . As explained in Chapter 2, the latter estimator is obtained from the empirical distribution function (EDF) of residuals. Thus, conditional upon $\mathbf{S} = (\mathbf{y}, \mathbf{X})$, the bootstrap DGP can be written as

$$\mathbf{y}^* = \mathbf{X}\ddot{\boldsymbol{\beta}} + \mathbf{u}^*, \quad (3.27)$$

in which \mathbf{u}^* contains IID errors that have a common CDF $\ddot{\mathcal{F}}$ given by the EDF for a set of residuals \ddot{u}_i , that is,

$$\ddot{\mathcal{F}} : \text{probability } \frac{1}{n} \text{ on } \ddot{u}_i, i = 1, \dots, n.$$

This specification of $\ddot{\mathcal{F}}$ is based upon the assumption that the residuals \ddot{u}_i sum to zero and, if this were not true, mean-adjustment would be required.

When H_0 is true, $\ddot{\boldsymbol{\beta}}$ and $\hat{\boldsymbol{\beta}}$ are both \sqrt{n} -consistent for $\boldsymbol{\beta}$ and so either could be used for $\boldsymbol{\beta}$. Similarly, under H_0 , $\ddot{\mathcal{F}}$ can be the EDF for either restricted or unrestricted residuals. Let $\tilde{\mathcal{F}}$ and $\hat{\mathcal{F}}$ denote the EDFs of restricted and unrestricted residuals, respectively. The use of one of these

EDFs for $\tilde{\mathcal{F}}$ implies that the corresponding residuals will be resampled randomly, with replacement, to serve as bootstrap world errors. As mentioned above, it is required that the errors of (3.27), like the errors of (3.1), have a population mean equal to zero. For simplicity of exposition, it is, therefore, assumed that (3.1) contains an intercept term which is not restricted by H_0 . It then follows that the sample means of restricted and unrestricted residuals both equal zero. (Least squares residuals that did not sum to zero over the n observations would have to be recentred by subtracting their sample average before being used to generate bootstrap errors.)

The bootstrap world parameter vector $\check{\theta}' = (\check{\beta}', \check{\mathcal{F}})$ can be defined in various ways. Four obvious combinations are: $\check{\theta}'_{(1)} = (\check{\beta}', \check{\mathcal{F}})$; $\check{\theta}'_{(2)} = (\hat{\beta}', \hat{\mathcal{F}})$; $\check{\theta}'_{(3)} = (\check{\beta}', \hat{\mathcal{F}})$; and $\check{\theta}'_{(4)} = (\hat{\beta}', \check{\mathcal{F}})$. The first choice uses only results from restricted estimation of (3.1) and provides the *restricted bootstrap test*. The second choice relies upon the results of unrestricted estimation of (3.1) and gives the *unrestricted bootstrap test*. The remaining two choices, viz. $\check{\theta}'_{(3)}$ and $\check{\theta}'_{(4)}$, combine results from both types of estimation and so give hybrid bootstrap tests. Results that are relevant to understanding the impact of the choice of $\check{\theta}$ on bootstrap tests are given in van Giersbergen and Kiviet (2002). These results will now be summarized.

Let the bootstrap counterparts of $\hat{\beta}$ and $\tilde{\beta}$ be denoted by $\hat{\beta}^*$ and $\tilde{\beta}^*$, respectively. Corresponding to (3.24), unrestricted estimators obtained using bootstrap data \mathbf{y}^* generated with $\check{\theta}' = (\check{\beta}', \check{\mathcal{F}})$ satisfy

$$\hat{\beta}^* = \check{\beta} + (\mathbf{X}'\mathbf{X})^{-1} \mathbf{X}'\mathbf{u}^*, \quad (3.28)$$

for both $\check{\beta} = \tilde{\beta}$ and $\check{\beta} = \hat{\beta}$. However, the form of the bootstrap counterpart of (3.25) depends on the choice for $\check{\beta}$. If $\check{\beta} = \tilde{\beta}$, so that $\mathbf{R}\check{\beta} = \mathbf{r}$,

$$\tilde{\beta}^* = \check{\beta} + \left(\mathbf{I}_k - (\mathbf{X}'\mathbf{X})^{-1} \mathbf{R}' \left[\mathbf{R} (\mathbf{X}'\mathbf{X})^{-1} \mathbf{R}' \right]^{-1} \mathbf{R} \right) (\mathbf{X}'\mathbf{X})^{-1} \mathbf{X}'\mathbf{u}^*. \quad (3.29)$$

However, if $\check{\beta} = \hat{\beta}$, $\mathbf{R}\check{\beta} \neq \mathbf{r}$ and the null hypothesis that is being tested using actual data will not be true in the artificial bootstrap world. When $\check{\beta} = \hat{\beta}$, (3.29) must be replaced by

$$\tilde{\beta}^* = \check{\beta} + \check{\delta} + \left(\mathbf{I}_k - (\mathbf{X}'\mathbf{X})^{-1} \mathbf{R}' \left[\mathbf{R} (\mathbf{X}'\mathbf{X})^{-1} \mathbf{R}' \right]^{-1} \mathbf{R} \right) (\mathbf{X}'\mathbf{X})^{-1} \mathbf{X}'\mathbf{u}^*, \quad (3.30)$$

in which

$$\ddot{\delta} = -(\mathbf{X}'\mathbf{X})^{-1} \mathbf{R}' \left[\mathbf{R} (\mathbf{X}'\mathbf{X})^{-1} \mathbf{R}' \right]^{-1} (\mathbf{R}\hat{\boldsymbol{\beta}} - \mathbf{r}). \tag{3.31}$$

Consequently, when the coefficients of (3.1) satisfy H_0 and $\hat{\boldsymbol{\beta}}$, the unrestricted estimated estimator of $\boldsymbol{\beta}$, is used to define $\tilde{\boldsymbol{\theta}}$, as in $\tilde{\boldsymbol{\theta}}_{(2)}$ and $\tilde{\boldsymbol{\theta}}_{(4)}$, bootstrap F test statistics of (3.23) do not have the same asymptotic distribution as the F -statistic given by (3.26). Following the suggestion for reflecting the null hypothesis that is made in Hall and Wilson (1991), asymptotically valid bootstrap inference with the choice $\tilde{\boldsymbol{\beta}} = \hat{\boldsymbol{\beta}}$ can be obtained by using the bootstrap data to test $H_0^u : \mathbf{R}\boldsymbol{\beta} = \mathbf{R}\hat{\boldsymbol{\beta}}$, rather than $H_0 : \mathbf{R}\boldsymbol{\beta} = \mathbf{r}$. With this change of null hypothesis, (3.28) and (3.29) will both be valid when $\tilde{\boldsymbol{\beta}} = \hat{\boldsymbol{\beta}}$; see van Giersbergen and Kiviet (2002) for a detailed discussion.

In fact, there is an equivalence between the F -statistic for testing $H_0 : \mathbf{R}\boldsymbol{\beta} = \mathbf{r}$ in bootstrap worlds defined with $\tilde{\boldsymbol{\beta}} = \tilde{\boldsymbol{\beta}}$ and the F -statistic for testing $H_0^u : \mathbf{R}\boldsymbol{\beta} = \mathbf{R}\hat{\boldsymbol{\beta}}$ in bootstrap worlds defined with $\tilde{\boldsymbol{\beta}} = \hat{\boldsymbol{\beta}}$. Let $\hat{\boldsymbol{\beta}}^*$ denote the unrestricted estimator for bootstrap data and $RSS^*(H_1)$ denote the corresponding sum of squared residuals. When $\tilde{\boldsymbol{\beta}} = \hat{\boldsymbol{\beta}}$ and the bootstrap-world null hypothesis is $H_0 : \mathbf{R}\boldsymbol{\beta} = \mathbf{r}$, the form of the bootstrap statistic F^* to be used to approximate the behaviour of the actual criterion F of (3.26) is

$$F^* = \frac{(\mathbf{R}\hat{\boldsymbol{\beta}}^* - \mathbf{r})' \left[\mathbf{R} (\mathbf{X}'\mathbf{X})^{-1} \mathbf{R}' \right]^{-1} (\mathbf{R}\hat{\boldsymbol{\beta}}^* - \mathbf{r})}{RSS^*(H_1)} \cdot \frac{df(H_1)}{q}. \tag{3.32}$$

When $\tilde{\boldsymbol{\beta}} = \hat{\boldsymbol{\beta}}$ and the bootstrap-world null hypothesis is $H_0^u : \mathbf{R}\boldsymbol{\beta} = \mathbf{R}\hat{\boldsymbol{\beta}}$, the corresponding bootstrap F -statistic is

$$F^* = \frac{(\mathbf{R}\hat{\boldsymbol{\beta}}^* - \mathbf{R}\hat{\boldsymbol{\beta}})' \left[\mathbf{R} (\mathbf{X}'\mathbf{X})^{-1} \mathbf{R}' \right]^{-1} (\mathbf{R}\hat{\boldsymbol{\beta}}^* - \mathbf{R}\hat{\boldsymbol{\beta}})}{RSS^*(H_1)} \cdot \frac{df(H_1)}{q}. \tag{3.33}$$

It is shown in van Giersbergen and Kiviet (2002, section 3) that these two expressions are equivalent and both can be rewritten as

$$F^* = \frac{(\mathbf{R} (\mathbf{X}'\mathbf{X})^{-1} \mathbf{X}'\mathbf{u}^*)' \left[\mathbf{R} (\mathbf{X}'\mathbf{X})^{-1} \mathbf{R}' \right]^{-1} (\mathbf{R} (\mathbf{X}'\mathbf{X})^{-1} \mathbf{X}'\mathbf{u}^*)}{\mathbf{u}^{*'} (\mathbf{I}_n - \mathbf{X} (\mathbf{X}'\mathbf{X})^{-1} \mathbf{X}') \mathbf{u}^*} \cdot \frac{df(H_1)}{q}, \tag{3.34}$$

which implies that, conditional upon \mathbf{X} , the bootstrap distribution of F^* depends upon that of \mathbf{u}^* and various constants. Given the same realization \mathbf{u}^* , (3.32) and (3.33) yield the same value of the test statistic. Hence, with strictly exogenous regressors, the importance of the choice between restricted and unrestricted estimations lies in the effects of the differences between the residual EDFs $\hat{\mathcal{F}}$ and $\tilde{\mathcal{F}}$ from which bootstrap errors are drawn.

The importance of the choice between $\hat{\mathcal{F}}$ and $\tilde{\mathcal{F}}$ has been the subject of some debate. The discussion has usually been based upon a consideration of the asymptotic properties of the EDF functions viewed as estimators of the true CDF \mathcal{F} . When testing a set of linear restrictions (3.23), $\hat{\mathcal{F}}$ will, under regularity conditions, be consistent and converge to \mathcal{F} , whether or not the null hypothesis is true. In contrast, the asymptotic behaviour of $\tilde{\mathcal{F}}$ does depend upon the validity of the null hypothesis. When the null is true, $\tilde{\mathcal{F}}$ will, like $\hat{\mathcal{F}}$, be consistent. When the null is false, $\tilde{\mathcal{F}}$, unlike $\hat{\mathcal{F}}$, is inconsistent for \mathcal{F} because the restricted residuals are derived using an inconsistent estimator of β .

These results on the asymptotic properties of $\hat{\mathcal{F}}$ and $\tilde{\mathcal{F}}$ as estimators of \mathcal{F} have led some researchers to argue for the use of unrestricted residuals to define the bootstrap error CDF on the grounds that this choice is likely to lead to higher power of bootstrap tests; see van Giersbergen and Kiviet (2002). However, others have argued against this conjecture, at least for asymptotically pivotal statistics like F of (3.26); see MacKinnon (2002). Evidence from simulation experiments on the relative merits of bootstrap tests derived from $\hat{\mathcal{F}}$ and $\tilde{\mathcal{F}}$ is discussed below. As will be seen, there seems little incentive not to use the restricted estimation version $\tilde{\mathcal{F}}$. It is also convenient to use $\tilde{\beta}$ for $\hat{\beta}$ since it is then valid to use the same null hypothesis for both actual data and artificial bootstrap data.

Consequently, in what follows, the parameter vector used to define the bootstrap world will usually be taken to be the restricted estimation vector $\tilde{\theta}'_{(1)} = (\tilde{\beta}', \tilde{\mathcal{F}})$. More precisely, the basic form of the bootstrap DGP is

$$\mathbf{y}^* = \mathbf{X}\tilde{\beta} + \mathbf{u}^*, \quad (3.35)$$

where $\mathbf{u}^{*'} = (u_1^*, \dots, u_n^*)$ is obtained by random sampling, with replacement, from

$$\tilde{\mathcal{F}} : \text{probability } \frac{1}{n} \text{ on } \tilde{u}_i, i = 1, \dots, n. \quad (3.36)$$

If the restricted residuals did not sum to zero, they would have to be recentered before being used in (3.36). Also (3.36) could be replaced by

some asymptotically valid variant obtained by modifying the restricted residuals using either the degrees-of-freedom adjustment or the leverage adjustment discussed in Chapter 2; see (2.31) and (2.32) for the corresponding expressions for unrestricted residuals.

The data process defined by (3.35) and (3.36) can be used to generate B samples of data, denoted by \mathbf{y}_b^* , $b = 1, \dots, B$. The value of the statistic F^* for testing (3.23) can be calculated from each of these bootstrap samples. These calculations provide a reference set (F_1^*, \dots, F_B^*) , with which the statistical significance of the actual-data statistic F can be assessed. Thus, using (2.12) of Chapter 2, the bootstrap p -value is computed as

$$BSPV_F = \frac{\#(F_b^* \geq F)}{B}, \quad (3.37)$$

and the restrictions of (3.23) are judged to be data-inconsistent when $BSPV_F \leq \alpha_d$, where α_d denotes the desired significance level.

An example: resampling restricted and unrestricted residuals

It is possible to illustrate the above results by considering a special case in which $q = k$ and the null hypothesis to be tested using actual data is $G_0 : \boldsymbol{\beta} = \mathbf{0}_k$. For this special case, $\hat{\boldsymbol{\beta}}$, as before, satisfies (3.24) and $\tilde{\boldsymbol{\beta}} = \mathbf{0}_k$. These two vectors are alternative choices for $\tilde{\boldsymbol{\beta}}$ when defining the parameter vector for the bootstrap data process.

If it is decided to set $\tilde{\boldsymbol{\beta}} = \hat{\boldsymbol{\beta}}$, bootstrap data are generated using

$$\mathbf{y}^* = \mathbf{X}\hat{\boldsymbol{\beta}} + \mathbf{u}^*,$$

and the bootstrap counterpart of (3.24) is

$$\hat{\boldsymbol{\beta}}^* = \hat{\boldsymbol{\beta}} + (\mathbf{X}'\mathbf{X})^{-1} \mathbf{X}'\mathbf{u}^*. \quad (3.38)$$

Given that the unrestricted (alternative hypothesis) estimator of $\boldsymbol{\beta}$ is used to define the bootstrap DGP, the null hypothesis in the bootstrap world is $G_0^u : \boldsymbol{\beta} = \hat{\boldsymbol{\beta}}$, not $G_0 : \boldsymbol{\beta} = \mathbf{0}_k$. With G_0^u under test, the bootstrap F statistic checks the joint significance of the elements of $\hat{\boldsymbol{\beta}}^* - \hat{\boldsymbol{\beta}}$, where $\hat{\boldsymbol{\beta}}^*$ is defined by (3.38). This F statistic is given by

$$\begin{aligned} F^* &= \frac{(\hat{\boldsymbol{\beta}}^* - \hat{\boldsymbol{\beta}})' (\mathbf{X}'\mathbf{X}) (\hat{\boldsymbol{\beta}}^* - \hat{\boldsymbol{\beta}})}{\mathbf{u}^{*'} (\mathbf{I}_n - \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1} \mathbf{X}') \mathbf{u}^*} \cdot \frac{n-k}{k} \\ &= \frac{(\mathbf{X}'\mathbf{u}^*)' (\mathbf{X}'\mathbf{X})^{-1} \mathbf{X}'\mathbf{u}^*}{\mathbf{u}^{*'} (\mathbf{I}_n - \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1} \mathbf{X}') \mathbf{u}^*} \cdot \frac{n-k}{k}; \end{aligned} \quad (3.39)$$

from (3.34).

Suppose next that the restricted estimator of β is employed to define the bootstrap DGP, so the artificial data are generated by

$$\mathbf{y}^* = \mathbf{u}^*,$$

using $\mathbf{X}\tilde{\beta} = \mathbf{X}(\mathbf{0}_k) = \mathbf{0}_n$. With this bootstrap DGP, the bootstrap counterpart of (3.24) is

$$\hat{\beta}^* = (\mathbf{X}'\mathbf{X})^{-1} \mathbf{X}'\mathbf{u}^*. \quad (3.40)$$

With the choice $\ddot{\beta} = \tilde{\beta} = \mathbf{0}_k$ for the bootstrap data process parameter vector, the null hypothesis to be tested is left in its original form, that is, $G_0 : \beta = \mathbf{0}_k$. The test of the joint significance of the elements of $\hat{\beta}^*$, given in (3.40), is based upon the F statistic

$$\begin{aligned} F^* &= \frac{\hat{\beta}^{*'} (\mathbf{X}'\mathbf{X}) \hat{\beta}^*}{\mathbf{u}^{*'} (\mathbf{I}_n - \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1} \mathbf{X}') \mathbf{u}^*} \cdot \frac{n-k}{k} \\ &= \frac{(\mathbf{X}'\mathbf{u}^*)' (\mathbf{X}'\mathbf{X})^{-1} \mathbf{X}'\mathbf{u}^*}{\mathbf{u}^{*'} (\mathbf{I}_n - \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1} \mathbf{X}') \mathbf{u}^*} \cdot \frac{n-k}{k}. \end{aligned} \quad (3.41)$$

It is clear from (3.39) and (3.41) that, given the same set of bootstrap errors \mathbf{u}^* , the two approaches will produce the same test statistic.

Turning to the choice of the EDF from which the elements of \mathbf{u}^* are to be drawn, $\hat{\mathcal{F}}$ and $\tilde{\mathcal{F}}$ are defined in this example by

$$\hat{\mathcal{F}} : \text{probability } \frac{1}{n} \text{ on } \hat{u}_i, i = 1, \dots, n,$$

in which \hat{u}_i is a typical element of the unrestricted residual vector,

$$\hat{\mathbf{u}} = (\mathbf{I}_n - \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1} \mathbf{X}') \mathbf{y} = (\mathbf{I}_n - \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1} \mathbf{X}') \mathbf{u},$$

whatever the value of β , and

$$\tilde{\mathcal{F}} : \text{probability } \frac{1}{n} \text{ on } y_i - \bar{y}, i = 1, \dots, n,$$

since $\tilde{\beta} = \mathbf{0}_k$ implies that the restricted residual vector is $\tilde{\mathbf{u}} = \mathbf{y}$, which must be centred before bootstrap errors are obtained. When $G_0 : \beta = \mathbf{0}_k$ is true, $\hat{\mathcal{F}}$ and $\tilde{\mathcal{F}}$ are derived from terms that differ from the corresponding errors by asymptotically negligible terms. If, however, $G_0 : \beta = \mathbf{0}_k$ is false, $\hat{\mathcal{F}}$ will tend to the true CDF \mathcal{F} , given the consistency of $\hat{\beta}$, but $\tilde{\mathcal{F}}$ will not because it is based upon an incorrect fixed value of β when $\beta \neq \mathbf{0}_k$.

3.4.2. Stable dynamic regression models

In addition to examining static regression models with strictly exogenous regressors, van Giersbergen and Kiviet also discuss stable dynamic regression models of the type for which bootstrap schemes were discussed in Section 2.3.3; see van Giersbergen and Kiviet (2002, section 4). Using the subscript t to denote a typical time series observation, a stable dynamic regression equation can be written as in (2.35) of Chapter 2,

$$y_t = \mathbf{y}'_{t(p)}\boldsymbol{\alpha} + \mathbf{x}'_t\boldsymbol{\beta} + u_t = \mathbf{w}'_t\boldsymbol{\gamma} + u_t, t = 1, \dots, n, \quad (3.42)$$

in which $\mathbf{y}'_{t(p)} = (y_{t-1}, \dots, y_{t-p})$, $p \geq 1$, $\mathbf{w}'_t = (\mathbf{y}'_{t(p)}, \mathbf{x}'_t)$, $\boldsymbol{\gamma}' = (\boldsymbol{\alpha}', \boldsymbol{\beta}')$ and the errors u_t are IID with zero mean, finite positive variance and CDF \mathcal{F} . Suppose that the null hypothesis to be tested in model (3.42) consists of the q_γ linear restrictions

$$H_\gamma : \mathbf{R}_\gamma\boldsymbol{\gamma} = \mathbf{r}_\gamma, \quad (3.43)$$

in which \mathbf{R}_γ is a $q_\gamma \times (p + k)$ matrix with elements that are known constants and \mathbf{r}_γ is a q_γ -dimensional vector of known constants.

Even if the errors of (3.42) were $\text{NID}(0, \sigma^2)$, the usual F -test of H_γ carried out after restricted and unrestricted estimation would only be asymptotically valid, given the inclusion of lagged dependent variables in the regressor set. The results in Freedman (1984) indicate that, under regularity conditions, bootstrap methods are asymptotically valid. (Freedman considers bootstrapping the two-stage least squares estimator in dynamic linear models, but this estimator includes OLS as a special case.)

The implementation of a bootstrap approach to testing H_γ , as an alternative to using asymptotic critical values, is straightforward. Given estimators $\hat{\boldsymbol{\alpha}}$ and $\hat{\boldsymbol{\beta}}$ that are consistent when H_γ is true, an autoregressive (recursive) simulation scheme, as defined by (2.36) in Chapter 2, can be used to generate B artificial samples of size n . With the strictly exogenous terms \mathbf{x}_t held constant, the bootstrap observations for the dependent variable are generated by

$$y_{bt}^* = \mathbf{y}'_{bt(p)}\hat{\boldsymbol{\alpha}} + \mathbf{x}'_t\hat{\boldsymbol{\beta}} + u_{bt}^*, t = 1, \dots, n, \quad (3.44)$$

in which actual data from (y_0, \dots, y_{1-p}) are used as the required start-up values and the error terms $(u_{bt}^*; t = 1, \dots, n)$ are obtained by random sampling, with replacement, from the EDF of the residuals associated

with $\ddot{\alpha}$ and $\ddot{\beta}$, in other words, from

$$\ddot{F} : \text{probability } \frac{1}{n} \text{ on } \ddot{u}_t, t = 1, \dots, n. \quad (3.45)$$

For each of the generated bootstrap samples, the value of the relevant F -statistic, denoted by F_b^* , $b = 1, \dots, B$, can be calculated and the full set of such statistics is then used, as in (3.37), to assess the strength of the evidence against H_γ that is provided by the value of F computed from the actual data.

Asymptotically valid bootstrap tests can be derived by using either the restricted estimate, or the unrestricted estimate, of γ from (3.42) to define the bootstrap DGP of (3.44) and (3.45), provided that the bootstrap world null hypothesis is redefined to be

$$H_\gamma^u : \mathbf{R}_\gamma \gamma = \mathbf{R}_\gamma \hat{\gamma}$$

when the unrestricted estimate $\hat{\gamma} = (\hat{\alpha}', \hat{\beta}')'$ is used. However, the equivalence established for models in which all regressors are strictly exogenous does not hold for dynamic models and the finite sample behaviour of tests can be strongly influenced by the choice between restricted and unrestricted estimation; see van Giersbergen and Kiviet (2002). On the basis of their results from simulation experiments, van Giersbergen and Kiviet conclude that the use of restricted estimates is to be preferred in dynamic models because unrestricted estimates suffer from problems of variability. Additional results on the importance of the choice between restricted and unrestricted estimation when defining bootstrap schemes can be obtained using experiments described in Section 1.5.1, in which the finite sample behaviour of asymptotically valid F tests was examined.

3.4.3. Some simulation evidence concerning asymptotic and bootstrap F tests

First, consider behaviour of tests under the null hypothesis. Whether the regression model is static or dynamic, the F statistic for testing a set of linear restrictions on regression coefficients has a known asymptotic distribution when those restrictions are valid, that is, the F statistic is asymptotically pivotal, under standard regularity conditions. Consequently the results in Beran (1988) indicate that a valid bootstrap test will have a smaller order of ERP than the asymptotic test. However, the evidence that was reported in Section 1.5.1 suggests that the asymptotic test, using critical values from the appropriate F distribution, can have estimated significance levels that are quite close to the desired levels.

Hence there may not be much scope for improvement when bootstrap versions of the F test are employed.

Second, suppose that experiments are conducted in which the null hypothesis is false. If asymptotic and bootstrap tests all have estimated significance levels that are close to the desired level, it will be possible to make empirically relevant comparisons of rejection frequencies obtained under the alternative hypothesis; see the comments in Horowitz and Savin (2000). In addition to the comparison of asymptotic and bootstrap tests, there is interest in two questions related to the implementation of bootstrap tests when the null hypothesis is untrue:

- What evidence is there about the link between the number of bootstrap samples and the power of bootstrap tests?
- What evidence is there to guide the choice between restricted and unrestricted residuals when defining the CDF of the error distribution for the bootstrap DGP?

Simulation results that are pertinent to these two questions and to the issue of the relative merits of asymptotic and various bootstrap tests can be obtained using the experimental designs described in Section 1.5.1. Consider the problem of testing the log-log estimating equation for a Cobb-Douglas model against the translog model. The unrestricted (alternative) model is

$$y_i = \beta_1 + \beta_2 x_{i2} + \beta_3 x_{i3} + \beta_4 x_{i2}^2 + \beta_5 x_{i2} x_{i3} + \beta_6 x_{i3}^2 + u_i, \quad (3.46)$$

and the null hypothesis is $H_0 : \beta_4 = \beta_5 = \beta_6 = 0$, which implies that the restricted model is

$$y_i = \beta_1 + \beta_2 x_{i2} + \beta_3 x_{i3} + u_i, \quad (3.47)$$

with the errors being IID with CDF \mathcal{F} and the variables of both models being as defined in Section 1.5.1. The real world data for the dependent variable and regressors in Greene (2008) are used to obtain $\hat{\mathbf{b}} = (\hat{b}_1, \hat{b}_2, \hat{b}_3, \hat{b}_4, \hat{b}_5, \hat{b}_6)'$, the unrestricted OLS estimate for the model (3.46), and $\bar{\mathbf{b}} = (\bar{b}_1, \bar{b}_2, \bar{b}_3, 0, 0, 0)'$, the corresponding restricted estimate. (The sub-vector $(\bar{b}_1, \bar{b}_2, \bar{b}_3)$, therefore, contains the estimates derived by applying OLS to the null model (3.47), using the actual data.)

Six simulation experiments are used, each of which has, like the original data series in Greene (2008), $n = 27$ as the sample size. Data generation processes are regression models with the same general form as either (3.46) or (3.47). The actual (real world) data in Greene (2008) are

used to supply the fixed regressor matrices of unrestricted and restricted models. Given these regressor matrices, it only remains to specify the parameter vector $\theta'_{[e]} = (\beta'_{[e]}, \mathcal{F}_{[e]})$ for experiment e , $e = 1, \dots, 6$. The null hypothesis is true in three of the six designs for simulation experiments, which have parameter vectors:

$$\theta'_{[1]} = (\tilde{\mathbf{b}}', \mathcal{F}_{[1]}), \theta'_{[2]} = (\tilde{\mathbf{b}}', \mathcal{F}_{[2]}) \text{ and } \theta'_{[3]} = (\tilde{\mathbf{b}}', \mathcal{F}_{[3]}),$$

in which $\mathcal{F}_{[1]}$, $\mathcal{F}_{[2]}$ and $\mathcal{F}_{[3]}$ are the CDFs implied by using standardized drawings from Normal, $t(5)$ and $\chi^2(2)$ distributions, respectively. As explained in Section 1.5.1, there is no loss of generality implied either by setting the error variance $\sigma_{[e]}^2$ equal to unity or by using the restricted estimate $\tilde{\mathbf{b}}$ as $\beta_{[e]}$, when the data process satisfies the null hypothesis.

The three simulation experiments in which the null hypothesis is false are defined by the parameter vectors:

$$\theta'_{[4]} = (\hat{\mathbf{b}}', \mathcal{F}_{[4]}), \theta'_{[5]} = (\hat{\mathbf{b}}', \mathcal{F}_{[5]}) \text{ and } \theta'_{[6]} = (\hat{\mathbf{b}}', \mathcal{F}_{[6]}),$$

in which $\mathcal{F}_{[4]}$, $\mathcal{F}_{[5]}$ and $\mathcal{F}_{[6]}$ are the CDFs implied by applying selected linear transformations to drawings from Normal, $t(5)$ and $\chi^2(2)$ distributions, respectively. The linear transformations are designed to yield errors with a zero mean and a variance such that power estimates are in an interesting range: neither too close to zero nor too close to one. These linear transformations are selected by trial and error in the simulation experiments.

Thus the data generation processes for the simulation experiments can be written as

$$\mathbf{y}_{[e]} = \mathbf{X}\beta_{[e]} + \mathbf{u}_{[e]}, \quad (3.48)$$

in which the 27 elements of $\mathbf{u}_{[e]}$ are IID with CDF $\mathcal{F}_{[e]}$, $e = 1, \dots, 6$. In the standard jargon of simulation studies, the number of replications used to estimate rejection probabilities of the tests is $R = 25,000$. Hence, for each of the simulation data generation schemes considered, 25,000 samples of data, each with 27 observations, are generated according to

$$\mathbf{y}_{[e,r]} = \mathbf{X}\beta_{[e]} + \mathbf{u}_{[e,r]}, \quad (3.49)$$

with $\mathbf{u}_{[e,r]}$ being a vector containing a simple random sample of 27 drawings from $\mathcal{F}_{[e]}$, for $r = 1, \dots, 25,000$.

Each realization from (3.49) can be used to calculate restricted and unrestricted estimates of $\beta_{[e]}$, denoted by $\tilde{\beta}_{[e,r]}$ and $\hat{\beta}_{[e,r]}$, with associated residual vectors $\tilde{\mathbf{u}}_{[e,r]}$ and $\hat{\mathbf{u}}_{[e,r]}$. The implied F -statistic for testing

$H_0 : \beta_4 = \beta_5 = \beta_6 = 0$ is denoted by $F_{[e,r]}$. An asymptotic test can be obtained by comparing $F_{[e,r]}$ with desired critical values of the $F(3, 21)$ distribution.

In the bootstrap-based alternative to the asymptotic test, the results obtained by estimating (3.49) are now employed to define a bootstrap data generation process from which B samples of bootstrap data are generated. Consequently the bootstrap counterpart of (3.49) is

$$\mathbf{y}_{[e,r,b]}^* = \mathbf{X}\boldsymbol{\beta}_{[e,r]}^* + \mathbf{u}_{[e,r,b]}^* \quad (3.50)$$

for $b = 1, \dots, B$. As explained below, the 27 error terms of $\mathbf{u}_{[e,r,b]}^*$ are obtained by resampling (possibly modified) residuals from estimation of (3.49). There are two values of B , $B = 400$ and $B = 1,000$, both of which are combined with six parameter vectors, denoted by $\boldsymbol{\theta}_{[e,r,j]}^{*'} = (\boldsymbol{\beta}_{[e,r]}^{*'}, \mathcal{F}_{[e,r,j]}^*)$, $j = 1, \dots, 6$, with the terms $\mathcal{F}_{[e,r,j]}^*$ each denoting an EDF derived from residuals calculated by estimating (3.49).

When setting up the bootstrap worlds, the null hypothesis is reflected by using $\boldsymbol{\beta}_{[e,r]}^* = \tilde{\boldsymbol{\beta}}_{[e,r]}$ in all vectors $\boldsymbol{\theta}_{[e,r,j]}^{*'}$, that is, the results of restricted estimation are used for the regression parameters. With this choice, the six parameter vectors for bootstrap data generation processes only differ in the EDF term and they can be written as:

$$\boldsymbol{\theta}_{[e,r,j]}^{*'} = \left(\tilde{\boldsymbol{\beta}}'_{[e,r]}, \tilde{\mathcal{F}}_{[e,r,j]} \right), j = 1, 2, 3,$$

and

$$\boldsymbol{\theta}_{[e,r,j+3]}^{*'} = \left(\tilde{\boldsymbol{\beta}}'_{[e,r]}, \hat{\mathcal{F}}_{[e,r,j]} \right), j = 1, 2, 3,$$

in which $\tilde{\mathcal{F}}_{[e,r,j]}$ and $\hat{\mathcal{F}}_{[e,r,j]}$ are derived from restricted and unrestricted residuals, respectively. As both types of residual sum to zero, it is not necessary to recentre them and $j = 1$ denotes an EDF for unadjusted residuals, $j = 2$ denotes an EDF for residuals adjusted by a degrees-of-freedom correction, and $j = 3$ denotes an EDF for the recentred versions of residuals adjusted by terms that take into account leverage values. Thus, for example, $\hat{\mathcal{F}}_{[e,r,j]}$, $j = 1, 2, 3$, correspond to (2.27), (2.30) and (2.31) above.

These combinations of the number of bootstraps and the bootstrap world parameter vector are intended to throw light on: (i) the effects of varying the number of bootstraps; (ii) the importance of the choice between restricted and unrestricted residuals when specifying the CDF for the bootstrap world; and (iii) the usefulness of adjusting residuals before using them to derive the bootstrap CDF.

Given a typical bootstrap sample from (3.50), the null hypothesis $H_0 : \beta_4 = \beta_5 = \beta_6 = 0$ can be tested. Let the resulting F -statistic be denoted by $F_{[e,r,b]}^*$. A p -value for $F_{[e,r]}$ is computed after completing the process of generating and analysing the full set of B bootstrap samples in replication r for simulation experiment e . Using $\mathbf{1}(A)$ to denote the usual indicator function for event A , this p -value estimate is calculated as

$$\hat{p}_{[e,r]} = \frac{\sum_{b=1}^B \mathbf{1}(F_{[e,r,b]}^* \geq F_{[e,r]})}{B}, \quad (3.51)$$

and H_0 is rejected if $\hat{p}_{[e,r]} \leq \alpha_d$, where α_d is the desired significance level. There are six estimates of the form (3.51); one for each of the choices from the set of bootstrap world parameters $\{\theta_{[e,r,j]}^*, j = 1, \dots, 6\}$.

Once the full set of $R = 25,000$ replications is completed for a given value of $\theta'_{[e]} = (\beta'_{[e]}, \mathcal{F}_{[e]})$, the finite sample rejection probability of a bootstrap test, under the design of simulation experiment e , can be estimated by the proportion of replications in which $\hat{p}_{[e,r]} \leq \alpha_d$, that is, by

$$\hat{p}_{BS[e]} = \frac{\sum_{r=1}^R \mathbf{1}(\hat{p}_{[e,r]} \leq \alpha_d)}{R}. \quad (3.52)$$

As with $\hat{p}_{[e,r]}$ of (3.51), there are six values for $\hat{p}_{BS[e]}$, each of which corresponds to one of the six approaches to specifying the bootstrap world parameter vector.

The rejection probability of the asymptotic test is also estimated using the full set of replications. The estimate is defined by

$$\hat{p}_{F[e]} = \frac{\sum_{r=1}^R \mathbf{1}(A_{[e,r]})}{R}, \quad (3.53)$$

in which $A_{[e,r]}$ is the event that $F_{[e,r]}$ is not less than the nominal critical value from the relevant F -distribution, which is $F(3, 21)$ in these experiments.

The desired significance level has three values in the experiments, with $\alpha_d = 0.10, 0.05, 0.01$. The precision of estimators for each of these desired levels can be assessed using the well-known formula for the standard error of the sample proportion. With $R = 25,000$, the values of $\sqrt{[\alpha_d(1 - \alpha_d)/R]}$ for $\alpha_d = 0.10, 0.05, 0.01$ are (approximately) 0.0019, 0.0014 and 0.0005, respectively. When levels of power are being estimated, the associated standard errors of estimators can be rather larger. For example, if a false null hypothesis were rejected in three-quarters of

the replications, the implied standard error would be calculated to be $\sqrt{(0.75)(0.25)/25,000}$, which is approximately 0.0027.

The results for significance levels under the three error distributions $\mathcal{F}_{[j]}$, $j = 1, 2, 3$, are given in Tables 3.4, 3.5 and 3.6. In each table, the first row of results corresponds to rejection rates for the asymptotic test that uses the $F(3, n - 6)$ distribution for critical values. (With $n = 27$ in all experiments, the reference distribution for the asymptotic test is always the $F(3, 21)$ distribution.) The remaining six rows of estimates correspond to bootstrap tests that are implemented by generating bootstrap data with parameter vectors of the general form $\theta_{[e,r,j]}^{*'} = (\beta_{[e,r]}^{*'}, \mathcal{F}_{[e,r,j]}^*)$, $j = 1, \dots, 6$, as defined above.

In Table 3.4, the errors have a Normal distribution and the use of the $F(3, 21)$ distribution for critical values is exactly valid. It is, therefore, not surprising that the estimates for this test are close to the corresponding desired levels. The bootstrap tests, being only asymptotically valid, could not do better, but do not do much worse. There is little to choose between the six bootstrap schemes on the basis of the evidence of Table 3.4, but the use of $B = 1,000$ appears to give better control of finite sample significance levels than $B = 400$.

The estimates in Table 3.5 are for the simulation world with symmetric $t(5)$ errors. The asymptotic test is no longer exactly valid and there is clear evidence of nonzero ERP terms, with all estimates being too large, relative to desired levels. Turning to the bootstrap tests, it really only with restricted residuals being used for the EDF and $B = 1,000$ that (minor)

Table 3.4 Estimates of significance levels of asymptotic and bootstrap (BS) F-tests of (3.47) against (3.46), with errors derived from $N(0, 1)$ distribution

α_d is equal to	1.0	5.0	10.0	1.0	5.0	10.0
$F(3, 21)$	1.0	4.9	10.0	1.0	4.9	10.0
BS test uses	$B = 400$			$B = 1,000$		
$(\tilde{\beta}'_{[e,r]}, \tilde{\mathcal{F}}_{[e,r,1]})$	1.2	5.2	10.3	1.1	5.2	10.2
$(\tilde{\beta}'_{[e,r]}, \tilde{\mathcal{F}}_{[e,r,2]})$	1.2	5.2	10.3	1.1	5.2	10.2
$(\tilde{\beta}'_{[e,r]}, \tilde{\mathcal{F}}_{[e,r,3]})$	1.2	5.2	10.3	1.0	5.2	10.2
$(\hat{\beta}'_{[e,r]}, \hat{\mathcal{F}}_{[e,r,1]})$	1.2	5.2	10.2	1.0	5.0	10.0
$(\hat{\beta}'_{[e,r]}, \hat{\mathcal{F}}_{[e,r,2]})$	1.2	5.2	10.2	1.0	5.0	10.0
$(\hat{\beta}'_{[e,r]}, \hat{\mathcal{F}}_{[e,r,3]})$	1.3	5.2	10.3	1.1	5.2	10.0

Notes: All estimates are for $n = 27$, are derived from 25,000 replications and are reported as percentages, rounded to one decimal place.

Table 3.5 Estimates of significance levels of asymptotic and bootstrap (BS) F-tests of (3.47) against (3.46), with errors derived from $t(5)$ distribution

α_d is equal to	1.0	5.0	10.0	1.0	5.0	10.0
$F(3, 21)$	1.4	5.7	10.8	1.4	5.7	10.8
BS test uses	$B = 400$			$B = 1,000$		
$(\tilde{\beta}'_{[e,r]}, \tilde{\mathcal{F}}_{[e,r,1]})$	1.5	5.7	10.9	1.2	5.5	10.6
$(\tilde{\beta}'_{[e,r]}, \tilde{\mathcal{F}}_{[e,r,2]})$	1.5	5.7	10.9	1.2	5.5	10.6
$(\tilde{\beta}'_{[e,r]}, \tilde{\mathcal{F}}_{[e,r,3]})$	1.5	5.7	10.9	1.2	5.5	10.6
$(\tilde{\beta}'_{[e,r]}, \hat{\mathcal{F}}_{[e,r,1]})$	1.6	5.8	10.9	1.4	5.7	10.6
$(\tilde{\beta}'_{[e,r]}, \hat{\mathcal{F}}_{[e,r,2]})$	1.6	5.8	10.9	1.4	5.7	10.6
$(\tilde{\beta}'_{[e,r]}, \hat{\mathcal{F}}_{[e,r,3]})$	1.6	5.9	11.0	1.4	5.7	10.6

Notes: All estimates are for $n = 27$, are derived from 25,000 replications and are reported as percentages, rounded to one decimal place.

Table 3.6 Estimates of significance levels of asymptotic and bootstrap (BS) F-tests of (3.47) against (3.46), with errors derived from $\chi^2(2)$ distribution

α_d is equal to	1.0	5.0	10.0	1.0	5.0	10.0
$F(3, 21)$	2.0	6.2	10.0	2.0	6.2	10.0
BS test uses	$B = 400$			$B = 1,000$		
$(\tilde{\beta}'_{[e,r]}, \tilde{\mathcal{F}}_{[e,r,1]})$	1.7	5.9	10.3	1.7	5.9	10.8
$(\tilde{\beta}'_{[e,r]}, \tilde{\mathcal{F}}_{[e,r,2]})$	1.7	5.9	10.3	1.7	5.9	10.8
$(\tilde{\beta}'_{[e,r]}, \tilde{\mathcal{F}}_{[e,r,3]})$	1.6	5.8	10.3	1.6	5.9	10.7
$(\tilde{\beta}'_{[e,r]}, \hat{\mathcal{F}}_{[e,r,1]})$	2.0	6.1	10.2	2.1	6.3	10.8
$(\tilde{\beta}'_{[e,r]}, \hat{\mathcal{F}}_{[e,r,2]})$	2.0	6.1	10.2	2.1	6.3	10.8
$(\tilde{\beta}'_{[e,r]}, \hat{\mathcal{F}}_{[e,r,3]})$	2.0	6.1	10.3	2.1	6.2	10.8

Notes: All estimates are for $n = 27$, are derived from 25,000 replications and are reported as percentages, rounded to one decimal place.

improvements are achieved. Modifications of restricted residuals do not seem important in these cases.

The use of the heavily-skewed $\chi^2(2)$ distribution for the error distribution produces the estimates in Table 3.6. The asymptotic test appears to reject too frequently for the 1 per cent and 5 per cent levels, but the estimated ERP is not large (about 1 percentage point). Bootstrap tests that use unrestricted residuals for resampling do not give systematic improvements. The use of restricted residuals to obtain the EDF terms $\tilde{\mathcal{F}}_{[e,r,j]}$, $j = 1, 2, 3$, combined with $B = 1,000$, yields such improvements, with there being

Table 3.7 Estimates of rejection probabilities of asymptotic and bootstrap (BS) F-tests of (3.47) against (3.46), with errors derived from Normal distribution

α_d is equal to	1.0	5.0	10.0	1.0	5.0	10.0
$F(3, 21)$	48.6	75.9	86.0	48.6	75.9	86.0
BS test uses	$B = 400$			$B = 1,000$		
$(\tilde{\beta}_{[e,r]}, \tilde{\mathcal{F}}_{[e,r,1]})$	51.0	76.5	86.1	49.6	76.1	85.5
$(\tilde{\beta}_{[e,r]}, \tilde{\mathcal{F}}_{[e,r,2]})$	51.0	76.5	86.1	49.6	76.1	85.5
$(\tilde{\beta}_{[e,r]}, \tilde{\mathcal{F}}_{[e,r,3]})$	50.9	76.6	86.2	49.5	76.0	85.5
$(\tilde{\beta}_{[e,r]}, \tilde{\mathcal{F}}_{[e,r,1]})$	50.7	76.4	86.0	49.0	75.8	85.4
$(\tilde{\beta}_{[e,r]}, \tilde{\mathcal{F}}_{[e,r,2]})$	50.7	76.4	86.0	49.0	75.8	85.4
$(\tilde{\beta}_{[e,r]}, \tilde{\mathcal{F}}_{[e,r,3]})$	51.1	76.6	86.0	49.6	76.1	85.5

Notes: All estimates are for $n = 27$, are derived from 25,000 replications and are reported as percentages, rounded to one decimal place.

only small differences between the results for the corresponding three bootstrap schemes.

Turning to evidence obtained when $H_0 : \beta_4 = \beta_5 = \beta_6 = 0$ is not satisfied, the general features of results are not sensitive to the choice of distribution for the errors of (3.48). Only the results for Normal errors are presented here and, as is clear from Table 3.4, estimated significance levels for this choice are sufficiently similar to support comparisons of rejection frequencies under the alternative. These rejection frequencies are given in Table 3.7.

The estimates reported in Table 3.7 do not reveal substantial differences in the ability of the tests to detect the departure from the null hypothesis. In relation to the two issues mentioned above, the rejection frequencies suggest that power is not enhanced either by increasing B from 400 to 1,000 or by using unrestricted, rather than restricted, residuals as the source of bootstrap errors. The former result is of limited importance, given the speed of modern computers, and, even though computing time is positive, the additional waiting time for the larger number of bootstrap samples is likely to be tiny. The latter result is consistent with those for asymptotically pivotal significance tests for regression models that are provided in Paparoditis and Politis (2005). It provides support for the use of a bootstrap scheme consisting of (3.35) and (3.36), that is, with both the bootstrap regression coefficient vector and the bootstrap error CDF being derived from the results of applying the restricted estimator to the actual data. These findings concerning bootstrap tests of (3.47) against (3.46) are corroborated by the estimates obtained from

a set of experiments involving autoregressive distributed lag relationships with non-Normal errors, in which (1.53) is tested against (1.54); see Section 1.5.1 for details of these models. The same main conclusions emerge under both null and alternative hypotheses.

As anticipated in the remarks at the start of this section, the application of bootstrap methods has not produced substantial improvements in finite sample behaviour of F -tests of simplifying linear restrictions. However, marked improvements from bootstrapping have been observed when different types of test have been studied. For example, Horowitz discusses the adequacy of asymptotic critical values for the well-known *information matrix* (IM) test proposed in White (1982a) and comments that “experiments carried out by many investigators have shown that with asymptotic critical values and sample sizes in the range found in applications, the true and nominal probabilities of rejecting a correct model can differ by a factor of 10 or more”; see Horowitz (2003, p. 213). Horowitz also comments that the estimates of ERP terms for IM tests are small when bootstrap methods are used. In fact, the results in Hall (1987) imply that the IM test for the linear regression model with exogenous regressors and NID errors can be implemented as a Monte Carlo procedure which is exactly valid.

White’s IM test differs from the F -tests discussed in this Section because it provides a check for misspecification. When testing for misspecifications, there is often uncertainty about the specification of the alternative hypothesis. (If there were very clear ideas about how a model was wrong, these ideas would presumably have been incorporated from the start of the empirical analysis.) Moreover, any test for misspecification can be linked to a family of *locally equivalent alternatives*, rather than to a single more general model; see Godfrey (1981) and Gourieroux and Monfort (1990, pp. 334–335). Consequently the issue of choosing between restricted and unrestricted residuals for defining bootstrap schemes is more complicated with misspecification tests than it is with the F -tests of this section. This choice will now be discussed in the context of testing for error serial correlation when the regression model has lagged values of the dependent variable as regressors, in other words, it is dynamic.

3.5. Bootstrapping LM tests for serial correlation in dynamic regression models

The importance of testing for serial correlation in the error terms of a linear regression model has been recognized for many years. In general,

the presence of serial correlation invalidates t and F tests, and leads to the inconsistency of OLS estimators if the regressors include lagged values of the dependent variable. Thus it is important for applied workers to have access to a reliable test for serial correlation when the regression model is dynamic. It is now standard practice to use the Lagrange Multiplier (LM) tests of Breusch (1978) and Godfrey (1978). The LM tests are flexible and use only OLS results. However, the LM tests suffer from the drawback that they are only asymptotically valid and the asymptotic χ^2 critical values have sometimes been found to give inadequate control of finite sample significance levels; see, for example, Davidson and MacKinnon (2007, section 8). The purpose of this section, which is based on Godfrey (2007b), is to give some results on bootstrapping LM tests in dynamic models. As in the previous section, the choice between restricted and unrestricted estimates for use as parameters in resampling schemes is considered in some detail.

3.5.1. Restricted or unrestricted estimates as parameters of bootstrap worlds

Suppose that data are generated by the stable dynamic linear model (3.42). The null hypothesis is that the errors u_t are IID with CDF \mathcal{F} , having mean zero and variance σ^2 . As in Breusch (1978), Durbin (1970) and Godfrey (1978), it is assumed that regularity conditions are satisfied, so that OLS estimators for (3.42) are asymptotically Normally distributed when this null hypothesis is true, with $(\hat{\boldsymbol{y}} - \boldsymbol{y})$ being $O_p(n^{-1/2})$.

All tests of the null hypothesis that the errors u_t are serially uncorrelated are constructed using the results of OLS estimation of (3.42). Let $\hat{\boldsymbol{y}}' = (\hat{\boldsymbol{\alpha}}', \hat{\boldsymbol{\beta}}')$ denote the OLS coefficient estimator for (3.42) and the terms $\hat{u}_t = y_t - \mathbf{w}'_t \hat{\boldsymbol{y}}$ be the corresponding residuals, $t = 1, \dots, n$. Whether the alternative is a G th-order autoregression, denoted by AR(G) and written as

$$u_t = \sum_{j=1}^G \phi_j u_{t-j} + \epsilon_t, \epsilon_t \text{ IID}(0, \sigma_\epsilon^2), \quad (3.54)$$

or a G th-order moving average, denoted by MA(G) and written as

$$u_t = \sum_{j=1}^G \theta_j \epsilon_{t-j} + \epsilon_t, \epsilon_t \text{ IID}(0, \sigma_\epsilon^2), \quad (3.55)$$

a suitable LM test of the null hypothesis can be computed as test of $\lambda = (\lambda_1, \dots, \lambda_G)' = \mathbf{0}$ in the augmented model

$$y_t = \mathbf{y}'_{t(p)} \boldsymbol{\alpha} + \mathbf{x}'_t \boldsymbol{\beta} + \hat{\mathbf{u}}'_{t(G)} \boldsymbol{\lambda} + u_t = \mathbf{w}'_t \boldsymbol{\gamma} + \hat{\mathbf{u}}'_{t(G)} \boldsymbol{\lambda} + u_t, \quad (3.56)$$

in which $\hat{\mathbf{u}}'_{t(G)} = (\hat{u}_{t-1}, \dots, \hat{u}_{t-G})$ and \hat{u}_{t-g} is set equal to zero for $t \leq g$. Let the OLS estimators for (3.56) be $\hat{\boldsymbol{\gamma}} = (\hat{\boldsymbol{\alpha}}', \hat{\boldsymbol{\beta}}')'$ and $\hat{\boldsymbol{\lambda}}$. The LM test is then a check of the joint significance of the elements of $\hat{\boldsymbol{\lambda}}$. The usual F -test is asymptotically valid and the F -statistic for testing (3.42) against (3.56) is denoted by LM_F .

Since the limit null distribution of LM_F is χ_G^2/G , this F -statistic is asymptotically pivotal, that is, its asymptotic distribution is independent of the nuisance parameters, which are taken to include the error distribution function \mathcal{F} . The results of Beran (1988), therefore, indicate that bootstrap tests may yield more accurate inferences than the asymptotic checks. A general recursive bootstrap scheme can be written as

$$y_t^* = \sum_{j=1}^p y_{t-j}^* \ddot{\alpha}_j + \mathbf{x}'_t \ddot{\boldsymbol{\beta}} + u_t^*, \quad t = 1, \dots, n, \quad (3.57)$$

in which: (i) presample values of y^* are set equal to those of y ; (ii) $\ddot{\boldsymbol{\alpha}} = (\ddot{\alpha}_1, \dots, \ddot{\alpha}_p)'$ and $\ddot{\boldsymbol{\beta}}$ are both consistent under the null hypothesis; and (iii) the distribution function of the bootstrap errors u_t^* converges to that of the true errors u_t when the null hypothesis is true. The choice of consistent estimator for the regression coefficients and the choice of scheme used to obtain u_t^* may both have small sample effects that cannot be neglected. The approaches adopted in the literature can be summarized as follows.

First, the restricted (null hypothesis) results can be used to mimic the assumed data process. In this approach, the OLS estimators $\hat{\boldsymbol{\alpha}}$ and $\hat{\boldsymbol{\beta}}$ from (3.42) are used for $\ddot{\boldsymbol{\alpha}}$ and $\ddot{\boldsymbol{\beta}}$, respectively, and the bootstrap errors u_t^* are obtained by simple random sampling, with replacement, from the empirical distribution function

$$\hat{\mathcal{F}} : \text{probability } \frac{1}{n} \text{ on } \hat{u}_t, \quad t = 1, \dots, n. \quad (3.58)$$

This combination gives the restricted (null hypothesis) bootstrap model

$$y_t^* = \sum_{j=1}^p y_{t-j}^* \hat{\alpha}_j + \mathbf{x}'_t \hat{\boldsymbol{\beta}} + u_t^*, \quad t = 1, \dots, n, \quad (3.59)$$

with u_t^* being derived using (3.58).

If (3.42) does not contain an intercept term, the sample mean of the OLS residuals should be subtracted from each \hat{u}_t before it is used in (3.58). Mean-adjustment is also required if the OLS residuals \hat{u}_t are modified by being divided by the square root of $(1 - h_{tt})$, where h_{tt} is the leverage value, $t = 1, \dots, n$.

Second, suppose that the bootstrap scheme is to be specified using a parameter estimator $\check{\theta}$ that is consistent under the alternative, as well as the null, in other words, an unrestricted-type estimator is used. Instrumental variable (IV) estimators are used in Rayner (1993) for $\check{\alpha}$ and $\check{\beta}$, with the instruments consisting of current and lagged exogenous variables. With this choice of instruments, the estimators are consistent under a fixed alternative hypothesis, that is, when at least one of the coefficients of the alternative model is a nonzero constant. In general, the statistic LM_F is $O_p(n)$ under such an alternative; so that the asymptotic rejection probability tends to unity for any finite critical value.

Let the IV parameter estimators and residuals for (3.42) be denoted by $\check{\alpha}$, $\check{\beta}$ and \check{u}_t , $t = 1, \dots, n$, respectively. In order to derive an unrestricted bootstrap scheme, it remains to specify how to obtain u_t^* . An AR(1) alternative is assumed in Rayner (1993) and a generalization for the AR(G) case involves applying OLS to

$$\check{u}_t = \phi_1 \check{u}_{t-1} + \dots + \phi_G \check{u}_{t-G} + e_t. \quad (3.60)$$

Let the residuals derived from OLS estimation of (3.60) be denoted by $\check{\check{e}}_t$. The unrestricted (fixed alternative hypothesis) bootstrap model is

$$y_t^* = \sum_{j=1}^p y_{t-j}^* \check{\alpha}_j + \mathbf{x}'_t \check{\beta} + u_t^*, \quad (3.61)$$

with u_t^* being obtained by simple random sampling, with replacement, from the empirical distribution function

$$\check{\mathcal{F}} : \text{probability } \frac{1}{n} \text{ on } \check{\check{e}}_t^c, t = 1, \dots, n, \quad (3.62)$$

where $\check{\check{e}}_t^c$ is the recentred version of $\check{\check{e}}_t$. The bootstrap test LM_F^* is then calculated using these artificial observations y_t^* and the bootstrap counterparts of (3.42) and (3.56).

In the unrestricted bootstrap model approach of Rayner (1993), the parameters of the conditional bootstrap law are estimators from observed data that are consistent in the specified unconditional fixed alternative model. This feature might be thought to yield residuals that give a better

approximation to the distribution of the error terms under the alternative. However, it is well-known that there is more than one alternative that leads to LM_F ; see, for example, Godfrey (1988, section 4.4.1). Consequently the use of the fitted autoregression (3.60) may be invalid under fixed alternatives. If the u_t were generated by a MA(G) process, the residuals \hat{e}_t^c used in (3.62) would be inappropriate; see Schwert (1987) for comments on the dangers of relying on pure autoregressions. Ramsey's criticism of the use of specific alternative models seems pertinent in the context of serial correlation tests; see Ramsey (1983, pp. 243–244). Thus there must be doubts about the general usefulness of the unrestricted bootstrap of Rayner (1993).

A different type of unrestricted bootstrap is considered in Mantalos (2003) but only for the case of an AR(1) alternative. A generalization that allows for the G th-order alternative consists of the following steps.

1. Estimate (3.42) by OLS to obtain $\hat{\alpha}, \hat{\beta}$ and the residuals \hat{u}_t .
2. Estimate (3.56) by OLS to obtain $\hat{\alpha}, \hat{\beta}, \hat{\lambda}$ and the residuals \hat{u}_t .
3. Draw e_1^*, \dots, e_n^* by simple random sampling with replacement from the empirical distribution function

$$\hat{\mathcal{F}} : \text{probability } \frac{1}{n} \text{ on } \hat{u}_t^c, t = 1, \dots, n, \quad (3.63)$$

where \hat{u}_t^c is the recentred version of \hat{u}_t ; asymptotically negligible modifications of the latter residuals (as described, for example, by MacKinnon, 2002, p. 620) are used in Mantalos (2003).

4. Given e_1^*, \dots, e_n^* from step 3, generate the bootstrap errors u_t^* recursively using

$$u_t^* = \sum_{j=1}^G \hat{\lambda}_j u_{t-j}^* + e_t^*, \quad (3.64)$$

with required starting values set equal to zero.

5. Generate bootstrap data using

$$y_t^* = \sum_{j=1}^G y_{t-j}^* \hat{\alpha}_j + \mathbf{x}_t' \hat{\beta} + u_t^*, t = 1, \dots, n, \quad (3.65)$$

in which the bootstrap errors are given by (3.64).

6. The bootstrap value of the Breusch-Godfrey LM_F statistic, denoted by LM_F^* , is then obtained by testing $H_0^u : \lambda = \hat{\lambda}$, not $H_0 : \lambda = \mathbf{0}_G$, in

the bootstrap counterpart of (3.56); see Section 3.3.1 above and van Giersbergen and Kiviet (2002, section 1.2). The OLS estimators of the coefficients of the bootstrap version of the artificial alternative (3.56) are denoted by $\hat{\gamma}^*$ and $\hat{\lambda}^*$.

In the version of an unrestricted bootstrap given by steps 1 to 6, there is no need to employ IV, as well as OLS, estimation. This saving reflects the fact that (3.56) is being used as an approximation to the specified alternative. The approximation is asymptotically valid under local alternatives in which parameters that are under test are $O(n^{-1/2})$, rather than $O(1)$; see Godfrey (1981). Thus the coefficients that determine the pattern of error autocorrelation are given by a Pitman-type drift, rather than being fixed constants.

Under an artificial sequence of alternatives that are drifting towards the null model at the specified rate in the unconditional (real world) law, the OLS estimators for (3.56) are consistent and, in particular, $\hat{\lambda}$ tends to a null vector so that (3.64) represents a local alternative, relative to $H_0 : \lambda = \mathbf{0}_G$, in the conditional (bootstrap) world. However, the local asymptotic theory of Pitman drifts may provide a poor approximation to actual behaviour in finite samples when observed serial correlation is not weak; see Eastwood and Godfrey (1992, section 4.2). It is argued in Godfrey (2007b) that, if asymptotic local theory fails to provide a good approximation in the bootstrap world, the unrestricted procedure based on steps 1–6 may lead to the true null hypothesis being rejected less frequently than desired in finite samples. The relevance of the asymptotic theory that underpins all three bootstrap approaches to the implementation of the Breusch-Godfrey test can be examined using simulation experiments.

3.5.2. Some simulation evidence on the choice between restricted and unrestricted estimates

The simulation experiments are based upon the dynamic model

$$y_t = \alpha_1 y_{t-1} + \alpha_2 y_{t-2} + \beta_1 + \beta_2 x_t + u_t, t = 1, \dots, n, \quad (3.66)$$

in which n is either 40 or 80. Under the null, the u_t are IID(0, σ^2). This process has been used in Dezhbakhsh (1990), Dezhbakhsh and Thursby (1995) and Godfrey and Tremayne (2005); parameter values in (3.66) are specified as in these earlier papers. The values of (α_1, α_2) are (0.5, 0.3), (0.7, -0.2), (1.0, -0.2), (1.3, -0.5), (0.9, -0.3) and (0.6, 0.2), which all

satisfy the conditions for dynamic stability. The value of (β_1, β_2) is $(1, 1)$ in all cases. The values of σ^2 are 1, 10 and 100. The OLS estimates of the parameters of (3.66) are denoted by $\hat{\alpha}_1, \hat{\alpha}_2, \hat{\beta}_1$ and $\hat{\beta}_2$.

The null hypothesis of serially uncorrelated errors is tested with $G = 4$; so that the test model corresponding to (3.56) is

$$y_t = \alpha_1 y_{t-1} + \alpha_2 y_{t-2} + \beta_1 + \beta_2 x_t + \sum_{j=1}^4 \lambda_j \hat{u}_{t-j} + \text{error}, \quad (3.67)$$

in which the terms \hat{u}_{t-j} are the lagged residuals from the OLS estimation of (3.66). The OLS estimate of (α_1, α_2) from (3.67) is denoted by $(\hat{\alpha}_1, \hat{\alpha}_2)$. A test of (3.66) against (3.67) corresponds to the kind of serial correlation check that might be used when working with quarterly data.

As in several other studies, x_t is generated by a stable AR(1) process,

$$x_t = \rho_x x_{t-1} + \zeta_t, |\rho_x| < 1, \zeta_t \text{ NID}(0, \sigma_\zeta^2). \quad (3.68)$$

The results discussed below are obtained using $\rho_x = 0.7$ and σ_ζ^2 selected so that $\text{Var}(x_t) = 1$. In order to obtain IV estimators for (3.66) that are consistent under a serial correlation alternative, x_t, x_{t-1}, x_{t-2} and an intercept term are used as instruments. The corresponding IV estimate of (α_1, α_2) from (3.66) is denoted by $(\check{\alpha}_1, \check{\alpha}_2)$.

The error terms u_t of (3.66) are generated by special cases of the mixed autoregressive-moving average ARMA(5, 5) model

$$u_t = \sum_{j=1}^5 \phi_j u_{t-j} + \epsilon_t + \sum_{j=1}^5 \theta_j \epsilon_{t-j}, \quad (3.69)$$

in which the ϵ_t are IID with zero mean and variance σ_ϵ^2 . All pre-sample values for (3.69) are set equal to zero. Similarly starting values for y and x are set equal to their respective unconditional means. The effects of this standard computational device are reduced by generating $n + 50$ observations and then using the last n of them.

Model (3.69) is sufficiently general to provide evidence about several aspects of the performance of the asymptotic and bootstrap tests derived from (3.66) and (3.67). By appropriate choices of the coefficients of (3.69), rejection rates can be estimated in the following cases: under the null hypothesis; under an AR(4) scheme, as is used in the unrestricted bootstrap tests; and under serial correlation models that are not AR(4).

The IID error term ϵ_t of (3.69) is drawn from three distributions. Following the previous studies of restricted and unrestricted bootstrap serial correlation tests, the Normal distribution is used. The other two choices give symmetric and asymmetric non-Normal distributions. The former involves drawing IID terms from a $t(5)$ distribution and then transforming them to have the required population mean and variance. For the latter, the $\chi^2(8)$ distribution is the source of the drawings that are transformed.

Tests are implemented in p -value form. The estimated p -values for the restricted, Mantalos-type unrestricted and Rayner-type unrestricted bootstrap tests are denoted by RES , MUR and RUR , respectively. These bootstrap p -values are calculated using 1,000 bootstrap samples. Rejection rates are obtained by comparing calculated p -values with nominal significance levels of 5 per cent and 10 per cent. Estimates of rejection probabilities are based upon a maximum of 25,000 replications but some replications are not suitable for computing bootstrap tests. The problem is that estimates of (α_1, α_2) , denoted by $(\check{\alpha}_1, \check{\alpha}_2)$, may fail to satisfy the conditions for dynamic stability and so cannot be used to define covariance stationary bootstrap data generation processes. The conditions that actual estimates must satisfy for stationarity in the bootstrap world are

$$-1 < \check{\alpha}_2 < 1, \check{\alpha}_1 + \check{\alpha}_2 < 1 \text{ and } \check{\alpha}_2 - \check{\alpha}_1 < 1.$$

In the discussion of the results from simulation experiments, it is useful to have codes for combinations of (α_1, α_2) and the error distribution. These codes are given in Table 3.8. For example, a case with code 2B has $(\alpha_1, \alpha_2) = (0.7, -0.2)$ and errors derived from the $t(5)$ distribution.

Table 3.8 Codes for combination of (α_1, α_2) and error distribution

1	$(\alpha_1, \alpha_2) = (0.5, 0.3)$
2	$(\alpha_1, \alpha_2) = (0.7, -0.2)$
3	$(\alpha_1, \alpha_2) = (1.0, -0.2)$
4	$(\alpha_1, \alpha_2) = (1.3, -0.5)$
5	$(\alpha_1, \alpha_2) = (0.9, -0.3)$
6	$(\alpha_1, \alpha_2) = (0.6, 0.2)$
A	error distribution from Normal
B	error distribution from $t(5)$
C	error distribution from $\chi^2(8)$

Before examining estimates of significance levels, consider first the results that shed light on how frequently bootstrap tests are applicable. The proportion of replications in which the estimates of (α_1, α_2) required for defining the bootstrap population are consistent with stationarity is an obvious index of applicability. Such proportions, measured in percentage terms, are referred to as applicability ratios. Perusal of the results indicates that applicability ratios are not very sensitive to the choice of error distribution from A, B and C of Table 3.8. As a representative sample, results on applicability for cases with Normal errors are reported in Table 3.9. Table 3.9 has 18 groups of results, each corresponding to a combination of bootstrap test (3 types), n (2 values) and σ^2 (3 values). There are 6 applicability ratios in each group, corresponding to the combinations of (α_1, α_2) in Table 3.8.

As can be seen from Table 3.9, there are important differences between the applicability ratios for the three different bootstrap tests. The restricted bootstrap check *RES* is almost always available whatever the combination of n and σ^2 . The unrestricted bootstrap tests, however, fail to match this level of performance. Instead the value of σ^2 is associated with substantial effects. As σ^2 increases so does the relative frequency with which fixed and local alternative hypothesis estimates of (α_1, α_2)

Table 3.9 Applicability ratios (percentages) for Normal errors

$n = 40$ with	$\sigma^2 = 1$	$\sigma^2 = 10$	$\sigma^2 = 100$
<i>RES</i>	100.0, 100.0, 100.0,	99.9, 100.0, 100.0,	99.8, 100.0, 100.0,
<i>test</i>	100.0, 100.0, 100.0	100.0, 100.0, 99.9	100.0, 100.0, 99.9
<i>RUR</i>	90.9, 96.6, 92.1,	49.7, 58.7, 53.7,	33.4, 38.2, 35.5,
<i>test</i>	86.9, 95.0, 92.4	49.7, 57.7, 51.4	36.3, 38.0, 33.6
<i>MUR</i>	97.8, 99.8, 99.4,	72.3, 79.7, 79.8,	58.3, 63.7, 64.5,
<i>test</i>	99.7, 99.8, 98.4	86.0, 82.8, 73.7	75.3, 68.5, 58.1
$n = 80$ with	$\sigma^2 = 1$	$\sigma^2 = 10$	$\sigma^2 = 100$
<i>RES</i>	100.0, 100.0, 100.0,	100.0, 100.0, 100.0,	100.0, 100.0, 100.0,
<i>test</i>	100.0, 100.0, 100.0	100.0, 100.0, 100.0	100.0, 100.0, 100.0
<i>RUR</i>	98.7, 99.6, 99.1,	63.2, 74.1, 67.1,	36.5, 42.3, 39.0,
<i>test</i>	96.0, 99.2, 99.0	62.0, 71.3, 65.4	38.7, 41.8, 36.2
<i>MUR</i>	99.9, 100.0, 100.0,	85.7, 91.1, 92.1,	65.2, 67.6, 72.1,
<i>test</i>	100.0, 100.0, 100.0	95.8, 92.8, 87.6	84.1, 74.9, 65.3

imply dynamically unstable bootstrap data processes. The problems with the IV-based procedure *RUR* derived from Rayner (1993) are so marked that it is difficult to recommend it as a tool for general use in applied work. The applicability of the Mantalos-type test *MUR* is not so severely impaired by error variance increases, but the effects of such increases are not negligible.

The sensitivity of unrestricted bootstrap tests to variations in σ^2 can be discussed after rewriting (3.66) as

$$y_t = \Psi(\mathcal{L})s_t + \sigma\Psi(\mathcal{L})a_t, \quad (3.70)$$

in which: \mathcal{L} is the lag operator, with $\mathcal{L}^j y_t = y_{t-j}$; $\Psi(\mathcal{L}) = 1 + \psi_1\mathcal{L} + \psi_2\mathcal{L}^2 + \dots = (1 - \alpha_1\mathcal{L} - \alpha_2\mathcal{L}^2)^{-1}$; $s_t = \beta_1 + \beta_2 x_t$; and $a_t = \sigma^{-1}u_t$. In this representation of the data process, s_t and a_t are uncorrelated with $\text{Var}(s_t) = \text{Var}(a_t) = 1$ in all experiments. It follows that, as σ increases, the importance of the exogenous component $\Psi(\mathcal{L})s_t$ decreases relative to that of $\sigma\Psi(\mathcal{L})a_t$.

In the experiments, Rayner's version of the unrestricted bootstrap test uses x_{t-1} and x_{t-2} as instruments for y_{t-1} and y_{t-2} . From (3.70), these instruments are only correlated with the exogenous component $\Psi(\mathcal{L})s_{t-j}$ of y_{t-j} , $j = 1, 2$. As σ increases, the exogenous components become less and less important, so that a type of weak instruments problem is approached and it is not surprising that IV estimates are not close to the corresponding true parameter values.

A different explanation is required for the sensitivity of the Mantalos-type check because it does not use IV estimation. The unrestricted bootstrap test that is proposed by Mantalos uses OLS estimators of (3.67) to define the bootstrap process. Under the null hypothesis, the estimators from (3.67) are inefficient relative to those for (3.66); the latter provide the parameter values for the restricted bootstrap. The degree of asymptotic variance inflation depends upon the extent to which y_{t-1} and y_{t-2} are "explained" in linear regressions by u_{t-1} , u_{t-2} , u_{t-3} and u_{t-4} . Consequently (3.70) implies that the effects of asymptotic variance inflation increase as σ increases. These effects may be reflected in finite samples by the greater frequency with which the estimates in $(\hat{\alpha}_1, \hat{\alpha}_2)$ from (3.67) imply nonstationary AR(2) regression models.

The fact that the test *MUR* is available more often than the test *RUR* is not sufficient to imply that the former is either well-behaved or superior to the latter. It is important to investigate the differences between actual and desired null rejection probabilities. When the data process of a

Table 3.10 Estimated null rejection probabilities for *RES*, *RUR* and *MUR* tests, with nominal significance level of 5 per cent

Case	<i>n</i> = 40			<i>n</i> = 80		
	<i>RES</i>	<i>RUR</i>	<i>MUR</i>	<i>RES</i>	<i>RUR</i>	<i>MUR</i>
1A	4.2	4.3	2.3	5.1	5.1	2.9
1B	4.2	4.2	2.3	4.9	4.9	2.7
1C	4.3	4.3	2.3	4.8	4.8	2.6
2A	4.6	4.6	2.0	5.0	5.0	2.7
2B	4.6	4.5	1.9	4.9	4.9	2.7
2C	4.4	4.3	1.9	5.0	4.9	2.6
3A	4.8	4.7	2.8	4.8	4.7	3.2
3B	4.3	4.1	2.5	4.8	4.7	3.3
3C	4.4	4.3	2.5	4.8	4.8	3.2
4A	4.3	4.3	3.3	4.9	4.9	4.4
4B	4.3	4.2	3.4	4.7	4.8	4.1
4C	4.3	4.3	3.4	4.9	4.8	4.1
5A	4.7	4.6	2.2	4.7	4.8	2.9
5B	4.3	4.2	2.2	5.0	5.0	3.2
5C	4.6	4.3	2.0	4.7	4.6	2.8
6A	4.9	4.8	2.6	4.8	4.8	2.7
6B	4.4	4.2	2.4	5.0	4.9	2.8
6C	4.5	4.5	2.3	5.0	4.9	2.6

Notes: The case codes given in the first column are derived from the codes of Table 3.8. The error variance is $\sigma^2 = 1$.

simulation experiment is such that all tests are usually available, the estimates for *MUR* are persistently below the desired values and the estimates for *RUR* suggest much closer agreement. The possibility of low rejection rates for *MUR* is discussed in Godfrey (2007b) and it appears that, in these experiments, AR error processes in the bootstrap world are not adequately approximated by the artificial alternative derived by adding lagged residuals to the original regression model. Table 3.10 contains results that illustrate these findings.

The results in Table 3.10 are for the data processes defined by combining the cases of Table 3.8 with $\sigma^2 = 1$. The use of $\sigma^2 = 1$ implies that all tests are available with quite high frequency. Only replications in which all tests are available are used to obtain the results of Table 3.10. Consequently the number of replications varies with case and sample size. At worst, there are over 21,000 replications, so that estimation should be sufficiently precise for practical purposes.

Table 3.11 Estimated null rejection probabilities of *RES* test, with $n = 40$ and a nominal significance level of 5 per cent.

Case	$\sigma^2 = 1$	$\sigma^2 = 10$	$\sigma^2 = 100$
1A	5.2	5.0	4.6
1B	4.8	4.9	4.7
1C	5.2	5.0	4.6
2A	5.2	5.3	5.1
2B	5.0	5.4	5.2
2C	4.8	5.0	5.1
3A	5.2	5.2	5.1
3B	4.9	4.9	4.9
3C	5.0	4.7	4.8
4A	5.4	5.2	5.0
4B	4.9	4.9	4.8
4C	4.8	5.1	5.1
5A	5.2	5.2	5.0
5B	5.1	5.3	5.0
5C	4.8	5.0	5.1
6A	5.2	5.1	5.0
6B	5.1	4.8	4.7
6C	5.0	5.0	4.9

Notes: The case codes given in the first column are derived from the codes of Table 3.8; and $n = 40$.

Rejection rates in Table 3.10 are derived with $\sigma^2 = 1$ so that all the tests can be compared. However, these results cannot be assumed to be representative of those for more general situations in which unrestricted bootstrap tests are not free of applicability problems. Attention is therefore also given to estimates for *RES* derived using all three values of σ^2 . Table 3.11 contains estimates for *RES* for all values of σ^2 , with $n = 40$. As indicated by Table 3.9, *RES* is almost always available and no estimate for this test in Table 3.11 is based upon fewer than 24,948 replications (many are based upon the full set of 25,000 replications). It is clear that *RES* performs well in the experiments. There is no indication of it being either persistently undersized or persistently oversized and fluctuations about the nominal size of 5 per cent are small. Every estimate for *RES* is in the range 0.9×5 per cent = 4.5 per cent to 1.1×5 per cent = 5.5 per cent; so that all satisfy the stringent criterion of robustness given in Serlin (2000).

Repeating the experiments of Table 3.11 with $n = 80$ does not lead to important changes. The same general patterns emerge, although the restricted bootstrap test has a slightly better performance when $n = 80$, with the largest difference between an estimate and the target value of 5 per cent being 0.3 per cent, compared with 0.4 per cent when $n = 40$.

To sum up, consideration of the results for experiments in which the null hypothesis is imposed leads to two conclusions. First, the restricted bootstrap leads to good control of finite sample null rejection rates. Second, doubt is cast upon the general usefulness of the two unrestricted bootstrap tests either because of the possibility of being frequently inapplicable or because of excessively low rejection rates.

Findings concerning estimates obtained under the null hypothesis have implications for comparisons of estimates derived under alternative hypotheses. In order to make sensible comparisons of power, there should not be important differences in estimates of null rejection probabilities. There is strong evidence that, when the null hypothesis is true, *MUR* rejects less frequently than *RES* and *RUR*, both of which have estimates that are closer to desired levels than those for *MUR*; see Table 3.10. It is therefore not surprising that *MUR* fails to detect serial correlation as frequently as the other tests when nonzero coefficients are used in (3.69). Given the arguments of Horowitz and Savin (2000), it was decided to exclude *MUR* from power comparisons, rather than to attempt to “size-correct” this test. Estimates for the remaining tests, viz., *RES* and *RUR*, are reported in Table 3.12. These estimates are representative of the full set derived with various serial correlation models that are special cases of (3.69).

Table 3.12 contains results for regression models with errors generated by

$$(1 - 0.7\mathcal{L} + 0.17\mathcal{L}^2 - 0.017\mathcal{L}^3 + 0.0006\mathcal{L}^4)u_t = \epsilon_t, \epsilon_t \text{ IID}(0, 1), \quad (3.71)$$

which has the same AR(4) structure as the version of (3.60) used to generate residuals for implementing the unrestricted bootstrap test *RUR*. The polynomial in \mathcal{L} used in (3.71) can be factorized as

$$(1 - 0.3\mathcal{L})(1 - 0.2\mathcal{L})(1 - 0.1\mathcal{L})^2.$$

It is clear from Table 3.12 that, whatever the combination of (α_1, α_2) and the error distribution, differences between power estimates are small and do not reveal a consistent ranking of *RES* and *RUR*. Consequently the results do not suggest that the unrestricted bootstrap test *RUR* has better power than the restricted bootstrap test *RES*. The findings derived

Table 3.12 Estimated alternative rejection probabilities for the autocorrelation model (3.71) and nominal significance level of 10 per cent

Case	RES	RUR
1A	55.6	55.2
1B	58.2	58.0
1C	56.8	56.6
2A	69.2	69.2
2B	70.3	70.3
2C	69.7	69.8
3A	73.7	73.7
3B	73.8	73.9
3C	73.8	73.9
4A	93.8	94.0
4B	94.3	94.4
4C	93.8	93.8
5A	79.0	79.0
5B	79.5	79.6
5C	79.5	79.6
6A	52.9	52.4
6B	53.9	53.7
6C	53.2	53.1

Notes: The case codes given in the first column are derived from the codes of Table 3.8; and $n = 80$.

from Table 3.12 are corroborated by the estimates from all other experiments with serially correlated errors; see Godfrey (2007b) for more details. This evidence, combined with the results on behaviour under the null hypothesis, provides support for the use of the restricted bootstrap scheme of (3.58) and (3.59) when testing for serial correlation after the OLS estimation of dynamic regression models.

Restricted bootstrap tests for serial correlation in dynamic regression models are also examined in Davidson and MacKinnon (2007). Davidson and MacKinnon use experiments in which the alternative is that the errors are generated by an AR(1) model and the regressors include only one lagged value of the dependent variable. They focus on the sample size $n = 20$ in order to highlight the improvements associated with the fast double bootstrap (FDB) method described in Section 2.5. The results from the simulation experiments conducted by Davidson and MacKinnon indicate that asymptotic critical values can be unreliable

and that, in general, the FDB tests work better than the single bootstrap versions. However, the correlations between the p -values from FDB and single bootstrap approaches are high, with all values being greater than 0.975.

3.6. Summary and concluding remarks

Two general approaches to using simulation to carry out tests in regression analysis have been discussed. First, the Monte Carlo approach has been described. This method can be used when the test statistic is exactly pivotal, that is, it has a distribution that does not depend upon any unknown parameters when the null hypothesis is true. Since the parameter vector of a regression model includes the CDF of the standardized errors, as well as the error variance and regression coefficients, a Monte Carlo test can only be applied if the CDF of the standardized errors is assumed to be known or is specified by the null hypothesis, for example, as in the test of Jarque and Bera (1980). It has been argued that these restrictions limit the usefulness of Monte Carlo tests in applied econometrics.

Second, the nonparametric bootstrap, which is more widely applicable, has been discussed. In contrast to the Monte Carlo technique, the nonparametric bootstrap does not use a prespecified (non-data based) CDF for the simulation-world errors, but instead uses an EDF derived from the residuals calculated from the actual sample data. Under general conditions, this approach yields a consistent estimator of an unspecified error CDF and the associated bootstrap test is asymptotically valid, but has a nonzero ERP in finite samples.

The behaviour of Monte Carlo tests that use a wrong assumption about the error distribution's CDF has been examined. Provided the test statistic is asymptotically pivotal, the incorrect Monte Carlo test is, like the bootstrap test, asymptotically valid. However, in such situations, the evidence indicates that finite sample significance levels of inappropriate Monte Carlo tests are not as well behaved as those of nonparametric bootstrap procedures. The predictions of the asymptotic analysis in Godfrey et al. (2006) are corroborated by results from simulation experiments that are discussed in Section 3.3.

It has, therefore, been argued that the nonparametric bootstrap is a more widely applicable and robust procedure for carrying out tests in regression models. The implementation of the bootstrap requires that a parameter vector for the bootstrap world data generation process be obtained from the actual data. The bootstrap parameter vector can be

constructed from either restricted (null hypothesis) estimates or unrestricted (alternative hypothesis) estimates. It is sometimes recommended that unrestricted estimates be used because they remain consistent under the alternative and this property might enhance the power of bootstrap tests. However, the evidence from simulation experiments does not provide strong support for this recommendation. Moreover, in many situations, there is uncertainty about the alternative, for example, when carrying out checks for misspecification. The results concerning standard F -tests in Section 3.4 indicate that restricted bootstrap tests are not outperformed by unrestricted bootstrap tests. Further evidence on this issue is provided in Section 3.5, in which the Breusch-Godfrey test for serial correlation is considered. This evidence indicates that unrestricted bootstrap implementation of serial correlation tests can lead to serious difficulties, whether the alternative is taken to be fixed or local. The evidence in Section 3.5 also reveals that the restricted bootstrap works well under the null and does not appear to be inferior in terms of power when compared with unrestricted bootstrap tests.

The focus of this chapter has been on test statistics that are asymptotically pivotal with a standard limiting distribution under the null hypothesis. Overall the recommendation that emerges is that a nonparametric bootstrap scheme, defined using restricted estimates, be employed to assess the statistical significance of such test statistics. The next chapter is devoted to discussion of results and evidence for tests that are either not asymptotically pivotal or are asymptotically pivotal with non-standard distributions. As will be seen, there are several important tests that are in these two categories.

4

Simulation-based Tests for Regression Models with IID Errors: Some Non-standard Cases

4.1. Introduction

There are important situations in which applied workers cannot use standard statistical tables to obtain asymptotic critical values when carrying out tests in regression analysis. In such cases, the test that is being applied will be called a non-standard asymptotic test. The purpose of this chapter is to provide discussions of some non-standard asymptotic tests of relevance to empirical econometrics. In the absence of convenient tabulated reference distributions, simulation methods offer the possibility of making asymptotically valid inferences. The form of the error distribution for the regression model is assumed to be unspecified and nonparametric bootstrap methods will be taken as the source of asymptotic tests and, in some cases, asymptotic refinements.

The test statistics that are discussed in this chapter are either asymptotically pivotal with non-standard limit null distributions or not even asymptotically pivotal. Applications of a single bootstrap cannot provide asymptotic refinements for the latter type of test statistic and so double bootstrap methods may be useful in applied work to achieve better control of finite sample significance levels; see Beran (1988). The computational costs of the conventional double bootstrap algorithm can be reduced by adopting the Fast Double Bootstrap (FDB) of Davidson and MacKinnon (2007); see Section 2.5 for discussions of double bootstrap techniques. The application of single bootstraps, double bootstraps and fast double bootstraps to non-standard asymptotic tests for regression models will be considered below. The three examples that will be used are those discussed in Section 1.6.

First, in Section 4.2, the implementation of the predictive test given in Chow (1960) is considered for cases in which the errors are IID with an

unspecified distribution. The standard form of the predictive test is based upon the very strong assumption that the errors are IID with a common Normal distribution. It is argued that, under more general assumptions about the error distribution, the test statistic is asymptotically nonpivotal. The results in Beran (1988) suggest that a single bootstrap will yield a test with the same order of ERP as the appropriate asymptotic test, but that a second stage of bootstrapping will be required to obtain asymptotic refinements, that is, an ERP of smaller order in the sample size. The implementation of double bootstrap predictive tests is discussed in some detail.

Second, Section 4.3 contains a discussion of the problem of controlling the overall significance level when the specification of a regression model is checked using a battery of test statistics. This situation is very commonly encountered in applied work because most modern programs will provide a number of checks for misspecification of the model, as well as point estimates of the regression coefficients and standard errors. When several separate asymptotic tests are carried out, conventional asymptotic theory can, in general, only provide bounds for the overall asymptotic significance level, even when all test statistics are asymptotically pivotal with standard limit null distributions. The use of single and double bootstrap schemes is again explored, with results from simulation experiments being used to throw light on the finite sample behaviour of the tests that are derived.

Third, the problem of detecting a *structural break* in a regression model when the *breakpoint* is unknown is examined in Section 4.4. Several authors have argued that it is rarely the case that the standard assumption of a known breakpoint is appropriate and that tests for structural breaks used in applied econometrics should be of the type proposed in Andrews (1993); see, for example, Hansen (1999) and Stock and Watson (2007). An asymptotic null distribution is derived by Andrews, who also provides tables of asymptotic critical values; see Andrews (1993, 2003a). As noted, for example, in Diebold and Chen (1996), the asymptotic null distribution is basically the distribution of the supremum of a collection of test statistics, each of which is individually asymptotically distributed as χ^2 . Diebold and Chen use simulation experiments in order to study the quality of the approximation provided by the results in Andrews (1993) and find evidence that the “bootstrap approximation to the finite-sample distribution appears consistently accurate, in contrast to the asymptotic approximation”; see Diebold and Chen (1996). Results on the relative merits of asymptotic and bootstrap tests for a structural break with an unknown breakpoint are discussed in Section 4.4, as is the usefulness of the FDB approach.

Finally, a summary of results concerning bootstrap versions of non-standard asymptotic tests and some concluding remarks are given in Section 4.5.

4.2. Bootstrapping predictive tests

Predictive tests, as described in Chow (1960), are widely used in applied work when estimating regression models by OLS. In such tests, the null hypothesis is that the same model applies to all the available data, that is, the regression parameters and error CDF are the same for all observations. The alternative hypothesis is that there are some observations that do not come from the same population as the rest. Such tests can be applied to estimated relationships as diagnostic checks when new data become available. They are also of interest when there is a priori information suggesting that some subset of the available data is not generated by the same process as other observations.

Godfrey and Orme draw the attention of applied workers to the following: (i) the strong auxiliary assumption of Normality that underpins the popular Chow predictive test; (ii) the impact of departures from Normality on this test; and (iii) the potential for using bootstrap methods to derive more robust inferences; see Godfrey and Orme (2000, 2002a). Single bootstrap methods are used with predictive tests in Godfrey and Orme (2000) and results on the gains associated with a double bootstrap technique are reported in Godfrey and Orme (2002a). In this section, relevant asymptotic results are summarized and evidence from simulation experiments is used to evaluate the practical importance of the asymptotic analysis and the usefulness of bootstrap techniques.

4.2.1. Asymptotic analysis for predictive test statistics

Following the notation used in Section 1.6, suppose that the full set of n observations is regarded as consisting of an estimation sample of $n_1 > k$ observations and a prediction sample of $n_2 = n - n_1$ observations. The null hypothesis is that the estimation sample and the prediction sample both come from the same population, with

$$y_i = \mathbf{x}_i' \boldsymbol{\beta} + u_i, i = 1, \dots, n, \quad (4.1)$$

in which the errors are IID, having CDF denoted by \mathcal{F} . The alternative hypothesis is that (4.1) is only valid for the estimation sample.

For convenience of exposition, let the estimation sample consist of the first n_1 observations. As explained in Section 1.6, predictions for the last n_2 observations $y_i, i = n_1 + 1, \dots, n$, are calculated using the OLS estimator of β derived from the estimation sample. This estimator can be written as

$$\check{\beta} = \left(\sum_{i=1}^{n_1} \mathbf{x}_i \mathbf{x}_i' \right)^{-1} \left(\sum_{i=1}^{n_1} \mathbf{x}_i y_i \right). \quad (4.2)$$

The predicted values, based upon the assumption of the same model applying to the second subsample, are then given by

$$\check{y}_{j+n_1} = \mathbf{x}'_{j+n_1} \check{\beta}, j = 1, \dots, n_2, \quad (4.3)$$

and the associated prediction residuals are

$$\check{e}_j = y_{j+n_1} - \check{y}_{j+n_1}, j = 1, \dots, n_2. \quad (4.4)$$

The null hypothesis implies that $E(y_{j+n_1}) = \mathbf{x}'_{j+n_1} \beta$, so

$$\begin{aligned} E(\check{e}_j) &= E(y_{j+n_1}) - E(\check{y}_{j+n_1}) \\ &= \mathbf{x}'_{j+n_1} \beta - \mathbf{x}'_{j+n_1} E(\check{\beta}) \\ &= 0, j = 1, \dots, n_2, \end{aligned}$$

since $E(u_{j+n_1}) = 0, j = 1, \dots, n_2$, and $E(\check{\beta}) = \beta$. A test of the joint significance of the prediction residuals is used to assess the data consistency of the null hypothesis.

Under classical assumptions, a simple extension of the basic regression model, using suitable dummy variables, allows calculation of the prediction residuals \check{e}_j and the required test of their joint significance; see Salkever (1976) for details. The dummy variables for the predictive test can be defined as follows. Let $\mathbf{d}_i, i = 1, \dots, n$, be a sequence of n_2 -dimensional vectors, with a typical element being $d_{ij} = \mathbf{1}(i = n_1 + j)$ for $j = 1, \dots, n_2$, where $\mathbf{1}(A)$ is the indicator variable that equals unity when A is true and is zero otherwise. The predictive test can then be implemented as a test of $\gamma = \mathbf{0}$ in the augmented model

$$y_i = \mathbf{x}'_i \beta + \mathbf{d}'_i \gamma + u_i, i = 1, \dots, n, \quad (4.5)$$

in which a typical element of $\boldsymbol{\gamma}$ is $\gamma_j = E(y_{j+n_1} | \mathbf{x}_{j+n_1}) - \mathbf{x}'_{j+n_1} \boldsymbol{\beta}$, $j = 1, \dots, n_2$. Salkever shows that the OLS estimator of $\boldsymbol{\gamma}$ in (4.5) equals the n_2 -dimensional vector of prediction residuals $(\check{\epsilon}_1, \check{\epsilon}_2, \dots, \check{\epsilon}_{n_2})'$, in which $\check{\epsilon}_j$ is defined in (4.4), $j = 1, \dots, n_2$; see Salkever (1976). Hence the predictive test can be implemented using the F -test of $\boldsymbol{\gamma} = \mathbf{0}$ in (4.5).

The F -statistic for testing $\boldsymbol{\gamma} = \mathbf{0}$ in (4.5) is simply the well-known Chow statistic of (1.60). If the standard assumptions concerning (4.1) are valid, this F -statistic is distributed as $F(n_2, n_1 - k)$. These standard assumptions are that (i) the regressors of \mathbf{x}_i are non-random or strictly exogenous and (ii) the errors u_i are independent $N(0, \sigma^2)$ variables, for $i = 1, \dots, n$; see Chow (1960), Hendry (1980) and Hendry and Santos (2005). In many situations, when carrying out an F -test, the overly restrictive assumption that the common distribution of the IID errors is $N(0, \sigma^2)$ can be relaxed at the cost of relying upon asymptotic theory, rather than finite sample results. Thus it is often the case that appeal is made to some form of Central Limit Theorem in order to claim that an F -test is asymptotically valid in the presence of unspecified non-Normality of the errors. However, the predictive test, as discussed in Chow (1960), cannot be regarded as being asymptotically robust to non-Normality.

It is assumed in Chow's discussion of predictive tests that the prediction sample is such that $n_2 \leq k$, with the estimation sample being such that $n_1 > k$. In conventional asymptotic theory for OLS estimators of regression models, k is taken as fixed with the number of observations used for estimation tending to infinity. Asymptotic analysis for the predictive test is, therefore, based upon allowing $n_1 \rightarrow \infty$, with k and n_2 fixed, perhaps with $n_2 \leq k$.

Since the OLS estimators of the coefficients γ_j in (4.5) equal the corresponding residuals $\check{\epsilon}_j$ of (4.4), it is appropriate to investigate the asymptotic robustness of the test of $\boldsymbol{\gamma} = \mathbf{0}$ in (4.5) by considering the limiting behaviour of these prediction residuals. The prediction residuals are given by

$$\check{\epsilon}_j = y_{j+n_1} - \mathbf{x}'_{j+n_1} \check{\boldsymbol{\beta}},$$

and so, under the assumption that the same model applies to all n observations,

$$\check{\epsilon}_j = u_{j+n_1} - \mathbf{x}'_{j+n_1} (\check{\boldsymbol{\beta}} - \boldsymbol{\beta}),$$

for $j = 1, \dots, n_2$. Now, as $n_1 \rightarrow \infty$, $p \lim (\check{\boldsymbol{\beta}} - \boldsymbol{\beta}) = \mathbf{0}$. Hence, under the null hypothesis, $\check{\epsilon}_j$ differs from u_{j+n_1} by a term that can be ignored as

$n_1 \rightarrow \infty$, for $j = 1, \dots, n_2$. It follows that the OLS estimator of $\boldsymbol{\gamma}$ in (4.5) tends to $(u_{n_1+1}, u_{n_1+2}, \dots, u_n)'$ when the null hypothesis is true.

Clearly it is not possible to appeal to a Central Limit Theorem to argue that the OLS estimator of $\boldsymbol{\gamma}$ in (4.5) is asymptotically Normal and consequently that the F -test of $\boldsymbol{\gamma} = \mathbf{0}$ is asymptotically valid, unless the individual errors of $(u_{n_1+1}, u_{n_1+2}, \dots, u_n)'$ are actually $N(0, \sigma^2)$. In general, the limit null distribution of the F -statistic for testing $\boldsymbol{\gamma} = \mathbf{0}$ in (4.5), as defined in (1.60), depends upon the unknown CDF of the errors, that is, \mathcal{F} . (Similar remarks apply to the asymptotic predictive test given in Hendry (1980). The $\chi^2(n_2)$ distribution used for asymptotic critical values in Hendry's test is only appropriate, as $n_1 \rightarrow \infty$, if errors are Normally distributed, as well as being IID.)

In the terminology of Beran (1988), the predictive test statistics proposed in Chow (1960) and Hendry (1980) are not asymptotically pivotal because of the dependence of their limit null distributions on the error CDF \mathcal{F} . The standard textbook versions, which assume Normality, are consequently asymptotically invalid when the common error distribution is not $N(0, \sigma^2)$. Beran's analysis indicates that the application of a single bootstrap provides a test that is asymptotically valid, with an ERP of the same order in magnitude in the sample size as the correct asymptotic test. The results in Beran (1988) also suggest that, in order to gain asymptotic refinements relative to the correct asymptotic test, a second stage of bootstrapping must be carried out.

4.2.2. Single and double bootstraps for predictive tests

Whether a single or double bootstrap approach is adopted, nonparametric methods can be used and there is no need to make spuriously precise assumptions about the form of the CDF \mathcal{F} . Indeed, the use of a parametric bootstrap approach in which an assumed error CDF, denoted by \mathcal{G} , is used for generating bootstrap errors would be ill-advised in practical situations. The dependence of the limit null distribution of the test statistic for the predictive check upon the true error CDF \mathcal{F} implies that a parametric bootstrap test is asymptotically invalid when $\mathcal{F} \neq \mathcal{G}$. The use of a consistent estimator of \mathcal{F} in a nonparametric bootstrap approach seems a much safer foundation for simulation-based inference.

It was recommended in Chapter 3 that tests be bootstrapped using restricted estimates from the actual data to define the parameter vector for the bootstrap scheme. In the context of predictive tests, the restricted estimator of $\boldsymbol{\beta}$ is the OLS estimator obtained when the null hypothesis that all observations are generated by the same model is imposed. This

restricted estimator is

$$\hat{\beta} = \left(\sum_{i=1}^n \mathbf{x}_i \mathbf{x}_i' \right)^{-1} \left(\sum_{i=1}^n \mathbf{x}_i y_i \right), \quad (4.6)$$

and the associated restricted residuals are defined by

$$\hat{u}_i = y_i - \mathbf{x}_i' \hat{\beta}, i = 1, \dots, n. \quad (4.7)$$

An asymptotically valid CDF for the bootstrap world, given the usual assumption that the regression model has an intercept term, is the basic restricted residual EDF

$$\hat{\mathcal{F}} : \text{probability } \frac{1}{n} \text{ on } \hat{u}_i, i = 1, \dots, n. \quad (4.8)$$

(Since the bootstrap data generation process must mimic the model under the null hypothesis, the bootstrap error must have an expected value equal to zero. If the regression model does not include an intercept term, the OLS residuals in $\hat{\mathcal{F}}$ should be recentered according to their sample mean.)

Remarks made in Section 3.4 concerning restricted and unrestricted estimates as bootstrap model parameters are pertinent. As is clear from (1.60), the statistic for the predictive test only depends upon residual sums of squares and known constants. There is, therefore, no dependence upon the choice of the regression coefficient vector when the regressors are strictly exogenous. For example, given a common bootstrap error vector from $\hat{\mathcal{F}}$, $\hat{\theta} = (\hat{\beta}', \hat{\mathcal{F}})'$ and $\check{\theta} = (\mathbf{0}'_k, \hat{\mathcal{F}})'$ lead to the same bootstrap Chow statistic; see the discussion of the result implied by (3.34) in Section 3.4.1.

Assuming that regressors are strictly exogenous, a single bootstrap method for carrying out Chow-type predictive tests in the presence of non-Normality can then be implemented using the following steps.

Predictive test: single bootstrap - Step 1

Estimate under the null hypothesis to obtain the sample values of the OLS estimate $\hat{\beta}$, defined in (4.6), and the associated residuals \hat{u}_i , $i = 1, \dots, n$. Also use the actual data to calculate the value of the standard Chow statistic P , which is given by (1.60).

Predictive test: single bootstrap - Step 2

Generate B bootstrap samples of size n , using the scheme

$$y_{bi}^* = \mathbf{x}_i' \hat{\boldsymbol{\beta}} + u_{bi}^*, i = 1, \dots, n, \quad (4.9)$$

in which the bootstrap errors $u_{bi}^*, i = 1, \dots, n$, are obtained by simple random sampling, with replacement, from $\hat{\mathcal{F}}$ of (4.8), $b = 1, \dots, B$. Application of the methods of Step 1 to each of the bootstrap samples provides values of the following items: Chow's statistic, denoted by P_b^* , the OLS parameter vector estimate, denoted by $\hat{\boldsymbol{\beta}}_b^*$; and the OLS residuals, which are denoted by $\hat{u}_{b1}^*, \dots, \hat{u}_{bn}^*$, for $b = 1, \dots, B$.

Predictive test: single bootstrap - Step 3

The B values of P_b^* could be ordered to obtain an estimate of the critical value that corresponds to the desired significance level α_d ; see, for example, Horowitz (1994). However, a more flexible approach is to estimate the p -value for the observed test statistic P calculated in Step 1 by

$$\hat{p}(P) = \frac{\sum_{b=1}^B \mathbf{1}(P_b^* \geq P)}{B}, \quad (4.10)$$

where $\mathbf{1}(\cdot)$ is the usual indicator function. The null hypothesis that the same regression model applies to all data is rejected if $\hat{p}(P) \leq \alpha_d$.

The single bootstrap procedure defined by these three steps leads to an asymptotically valid test when the strong auxiliary assumption of Normality is relaxed and \mathcal{F} is unspecified. In other words, the ERP of the single bootstrap predictive test is $o(1)$, under non-Normality, whereas that of the textbook Chow test, which uses critical values from the $F(n_2, n_1 - k)$ distribution, is $O(1)$ and does not vanish asymptotically. However, the results in Beran (1988) indicate that, under standard regularity conditions, it may be possible to improve upon the finite sample performance of the single bootstrap predictive test by using additional simulations. More precisely, a double bootstrap will have the benefit of an asymptotic refinement and its ERP will tend to zero faster than that of the single bootstrap variant.

The double bootstrap approach can be motivated by viewing the first-stage bootstrap p -value $\hat{p}(P)$, rather than P , as the test statistic because the former is asymptotically pivotal, having an asymptotic null distribution that is uniform between zero and unity; see Davison and Hinkley (1997, section 4.5). Consequently, following the description in Davison and

Hinkley (1997), a double bootstrap predictive test can be implemented by an “adjusted p -value” approach using the following steps.

Predictive test: double bootstrap - Step 1

This step is the same as the first step of the single bootstrap version of the predictive test.

Predictive test: double bootstrap - Step 2

This step is the same as the second step of the single bootstrap version of the predictive test.

Predictive test: double bootstrap - Step 3

Conditional upon each of the first-level bootstrap samples, generate C second-level bootstrap samples of size n according to

$$y_{bci}^{**} = \mathbf{x}_i' \hat{\boldsymbol{\beta}}_b^* + u_{bci}^{**}, \quad i = 1, \dots, n \text{ and } c = 1, \dots, C, \quad (4.11)$$

where the second-level bootstrap errors u_{bci}^{**} are obtained by simple random sampling, with replacement, from

$$\hat{\mathcal{F}}_b^* : \text{probability } \frac{1}{n} \text{ on } \hat{u}_{bi}^*, \quad i = 1, \dots, n. \quad (4.12)$$

(As with the second step of the single bootstrap method, the residuals \hat{u}_{bi}^* must be centred to have zero mean for the purpose of defining $\hat{\mathcal{F}}_b^*$ if $\boldsymbol{\beta}$ does not contain an intercept.)

Predictive test: double bootstrap - Step 4

Let the Chow statistics derived from the second-level bootstrap data y_{bci}^{**} , $i = 1, \dots, n$, be denoted by P_{bc}^{**} , $b = 1, \dots, B$ and $c = 1, \dots, C$. For each value of b , use the values of P_{bc}^{**} , $c = 1, \dots, C$, to estimate the p -value of the first-level test statistic P_b^* by

$$\hat{p}_b^*(P_b^*) = \frac{\sum_{c=1}^C \mathbf{1}(P_{bc}^{**} \geq P_b^*)}{C}, \quad b = 1, \dots, B.$$

The terms $\hat{p}_b^*(P_b^*)$ provide an empirical reference distribution for $\hat{p}(P)$ and can be used to gain an asymptotic refinement over the single bootstrap rejection rule “Reject null hypothesis if $\hat{p}(P) \leq \alpha_d$ ”. An adjusted p -value is calculated as

$$p_{adj}(P) = \frac{\sum_{b=1}^B \mathbf{1}(\hat{p}_b^*(P_b^*) \leq \hat{p}(P))}{B}, \quad (4.13)$$

and the null hypothesis is rejected if $p_{adj}(P) \leq \alpha_d$.

At first sight, the computational cost of the double bootstrap (adjusted p -value) approach might seem very large. The usual form of the statistic for Chow's predictive test involves the residual sums of squares from two regressions, one using the estimation sample and the other based upon the full (pooled) sample. (An equivalent method for the calculation of Chow's statistic is to estimate (4.1) and (4.5) by OLS and to carry out the F -test of the former against the latter.) Thus, the double bootstrap version of the predictive test might be thought to require $2(1 + B(1 + C))$ OLS estimations to generate the necessary values of the test statistic. However, in order to obtain both levels of bootstrap samples and all the associated values of the Chow statistic, the only regressions that need to be estimated are the two involving the genuine data, that is, those of potential economic interest. Consequently it is possible to reduce the cost of a double bootstrap predictive test to only a few seconds of waiting time on a modern computer. The savings, relative to repeated application of an OLS estimation algorithm, can be explained as follows.

A restricted bootstrap approach is to be used and the null hypothesis for actual data is to be imposed. When the null hypothesis is true, the Chow statistic from actual data is

$$P = \frac{[\mathbf{u}'\mathbf{M}\mathbf{u} - \mathbf{u}'_1\mathbf{M}_1\mathbf{u}_1]/n_2}{[\mathbf{u}'_1\mathbf{M}_1\mathbf{u}_1]/(n_1 - k)} = g_P(\mathbf{u}; \mathbf{M}, \mathbf{M}_1), \tag{4.14}$$

the values from first-level bootstrap data are given by

$$P_b^* = \frac{[\mathbf{u}_b^{*'}\mathbf{M}\mathbf{u}_b^* - \mathbf{u}_{b1}^{*'}\mathbf{M}_1\mathbf{u}_{b1}^*] / n_2}{[\mathbf{u}_{b1}^{*'}\mathbf{M}_1\mathbf{u}_{b1}^*] / (n_1 - k)} = g_P(\mathbf{u}_b^*; \mathbf{M}, \mathbf{M}_1), b = 1, \dots, B, \tag{4.15}$$

and, for the second-level bootstrap data, the corresponding expression is

$$P_{bc}^{**} = \frac{[\mathbf{u}_{bc}^{**'}\mathbf{M}\mathbf{u}_{bc}^{**} - \mathbf{u}_{bc1}^{**'}\mathbf{M}_1\mathbf{u}_{bc1}^{**}] / n_2}{[\mathbf{u}_{bc1}^{**'}\mathbf{M}_1\mathbf{u}_{bc1}^{**}] / (n_1 - k)} = g_P(\mathbf{u}_{bc}^{**}; \mathbf{M}, \mathbf{M}_1), c = 1, \dots, C, \tag{4.16}$$

in which $\mathbf{u} = (u_1, \dots, u_n)'$, $\mathbf{u}_1 = (u_1, \dots, u_{n_1})'$, $\mathbf{u}_b^* = (u_{b1}^*, \dots, u_{bn}^*)'$, $\mathbf{u}_{b1}^* = (u_{b1}^*, \dots, u_{bn_1}^*)'$, $\mathbf{u}_{bc}^{**} = (u_{bc1}^{**}, \dots, u_{bcn}^{**})'$, $\mathbf{u}_{bc1}^{**} = (u_{bc1}^{**}, \dots, u_{bcn_1}^{**})'$, $\mathbf{M} = \mathbf{I}_n - \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'$, \mathbf{X} is the $n \times k$ regressor matrix for the pooled sample, $\mathbf{M}_1 = \mathbf{I}_{n_1} - \mathbf{X}_1(\mathbf{X}'_1\mathbf{X}_1)^{-1}\mathbf{X}'_1$ and \mathbf{X}_1 is the $n_1 \times k$ regressor matrix for the estimation sample. The projection matrices \mathbf{M} and \mathbf{M}_1 in (4.15) and (4.16) are fixed for all bootstrap samples and need only be calculated

once. The terms that vary over artificial samples in (4.15) and (4.16) are \mathbf{u}_b^* and \mathbf{u}_{bc}^{**} . Given realizations of these vectors, there is no need to invert terms like $(\mathbf{X}'\mathbf{X})$ and $(\mathbf{X}'_1\mathbf{X}_1)$ for the $(B(1+C))$ bootstrap samples. Consider then the derivation of \mathbf{u}_b^* and \mathbf{u}_{bc}^{**} .

The elements of any realization of \mathbf{u}_b^* (and hence of its sub-vector \mathbf{u}_{b1}^*) are obtained by resampling the residuals from the OLS regression for actual data. Hence the generation of \mathbf{u}_{bc}^{**} (and hence of its sub-vector \mathbf{u}_{bc1}^{**}), as specified in (4.12), might be thought to require that an OLS regression be carried out, using first-level bootstrap data, to obtain residuals \hat{u}_{bi}^* , $i = 1, \dots, n$, which are then resampled, with replacement. However, it is more efficient to calculate the residuals \hat{u}_{bi}^* as the elements of the product of the fixed projection matrix \mathbf{M} and the realization of \mathbf{u}_b^* ; so that unnecessary calculations are avoided.

A similar device for reducing the computational costs of double bootstraps for linear regression models is described in McCullough and Vinod (1998, p. 93). McCullough and Vinod comment on important savings gained by calculating OLS parameter estimates for bootstrap data as the product of the fixed matrix $(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'$ and the bootstrap data vector for the dependent variable. They report that this approach produces a saving of about 80 per cent of the computing time, relative to repeated OLS estimations.

If the first-level computations required for the single bootstrap version of the predictive test are carried out and the results saved before second-level bootstrapping takes place, reductions of computation time can be obtained by using stopping rules to avoid carrying out unnecessary second-level bootstraps; see Horowitz et al. (2006, p. 861) for a description of three stopping rules. Horowitz et al. report very substantial savings of computing time as a result of employing stopping rules. However, for the applied worker, the short-run obstacle to using double bootstrap predictive tests in regression models is more likely to be the absence of suitable programs, than the time required to implement the procedures.

4.2.3. Simulation experiments and results

The above discussion of bootstrapping predictive tests has stressed the differences, under non-Normal errors, between the asymptotic orders of magnitude of ERP terms of standard textbook, single bootstrap and double bootstrap variants. It is, of course, important to have evidence about the finite sample relevance of asymptotic analysis relating to the relative magnitudes of these ERP terms. Results from some simulation experiments will, therefore, be provided. The details of the steps of the

computations required for the experiments are similar to those described in Section 3.4.3 on bootstrapping F -tests.

As in Godfrey and Orme (2000), the regression model employed in the experiments is

$$y_i = \sum_{j=1}^6 x_{ij} \beta_j + u_i, \quad (4.17)$$

in which: $x_{i1} = 1$; x_{i2} is drawn from a uniform distribution with parameters 1 and 31; x_{i3} is drawn from a Log-Normal distribution with $\ln(x_{i3}) \sim N(3, 1)$. Unlike x_{i2} and x_{i3} , the remaining regressors are serially correlated with

$$x_{i4} = 0.9x_{i-1,4} + v_{i4},$$

$$x_{i5} = 0.6x_{i-1,5} + v_{i5},$$

$$x_{i6} = 0.3x_{i-1,6} + v_{i6},$$

with v_{is} being independently Normally distributed, such that $E[x_{is}] = 0$ and $\text{var}[x_{is}] = 1$, for $s = 4, 5, 6$. The errors of (4.17) are IID with zero mean and variance σ^2 .

In the experiments based upon (4.17), (i) all regression coefficients β_j are set equal to zero and (ii) the error terms u_i have variance equal to one. Invariance results imply that neither (i) nor (ii) implies any loss of generality; see Breusch (1980). Since sensitivity to non-Normality is an important issue when discussing the behaviour of predictive tests, the errors u_i are obtained by standardizing pseudo-random variables drawn from several distributions. These error distributions are: Normal; Student's t with 5 degrees of freedom, $t(5)$; uniform over the unit interval; chi-square with 2 degrees of freedom, $\chi^2(2)$; and Log-Normal. This collection of distributions should provide a reasonable guide to how poor (oversized or undersized) any procedure might be in a practical situation.

The experiments involve three combinations of (n, n_2) with $n = (n_1 + n_2) = 30, 50, 80$ and $n_2 = 6$ in all cases. The reason for this choice of n_2 is that the Chow predictive test is usually applied when $n_2 \leq k$ and it is found in Godfrey and Orme (2000) that, given n , the performance of the single bootstrap procedure deteriorates as n_2 increases. Thus, in order to give a stringent check for predictive tests based upon (4.17), attention is restricted to the case of $n_2 = k = 6$. For the standard Chow predictive test, critical values are obtained from a $F(6, n_1 - 6)$ distribution,

$n_1 = 24, 44, 74$. In the notation of the previous subsection, single and double bootstrap predictive tests are implemented using $B = 500$ and $C = 100$.

Results are obtained for the three desired significance levels of 1 per cent, 5 per cent and 10 per cent. All three conventional significance levels are used because an applied worker may employ differing levels of significance according to the purpose of the test. For example, a 10 per cent level might be appropriate when testing for a genuine regime break, whilst 1 per cent might be appropriate if the test is just one of a number of diagnostics being employed. Estimates of the corresponding actual finite sample rejection probabilities are calculated using 25,000 replications.

In Tables 4.1 to 4.3, estimated rejection probabilities under the null hypothesis are given as percentages, rounded to one decimal place. For true rejection probabilities of 1 per cent, 5 per cent and 10 per cent, the values of twice the standard error of the corresponding estimators are (approximately) 0.1 per cent, 0.3 per cent and 0.4 per cent. The notation used in these tables is as follows. The standard Normality-valid test, using critical values from the assumed F -distribution, is denoted by SNV. The test based upon the single bootstrap p -value of the Chow statistic is denoted by SBS. Finally, DBS denotes the double bootstrap (adjusted p -value) test.

Table 4.1 Estimated rejection probabilities for SNV, SBS and DBS versions of Chow's predictive test, $n_1 = 24$ and $n_2 = 6$, with desired significance levels of 1 per cent, 5 per cent and 10 per cent.

Error distribution	Normal	$t(5)$	Uniform	$\chi^2(2)$	Log-Normal
a. Desired significance level of 1 per cent					
SNV	1.0	3.0	0.3	5.4	9.9
SBS	1.1	1.8	0.7	2.6	4.7
DBS	0.8	1.0	0.8	1.1	1.0
b. Desired significance level of 5 per cent					
SNV	5.2	8.3	2.4	11.1	15.1
SBS	5.5	7.4	3.9	8.8	11.6
DBS	5.1	6.0	4.3	6.6	7.9
c. Desired significance level of 10 per cent					
SNV	10.1	13.5	6.2	15.9	18.8
SBS	10.8	12.8	8.7	14.3	16.6
DBS	10.5	11.7	9.3	12.2	13.8

Notes: The $F(6, n_1 - 6)$ distribution provides critical values for the SNV test. The tests denoted by SBS and DBS are derived from single bootstrap and double bootstrap methods, respectively.

Table 4.2 Estimated rejection probabilities for SNV, SBS and DBS versions of Chow's predictive test, $n_1 = 44$ and $n_2 = 6$, with desired significance levels of 1 per cent, 5 per cent and 10 per cent.

Error distribution	<i>Normal</i>	<i>t(5)</i>	<i>Uniform</i>	$\chi^2(2)$	<i>Log-Normal</i>
a. Desired significance level of 1 per cent					
SNV	1.0	3.9	0.1	6.1	9.5
SBS	1.0	1.6	0.6	1.6	2.1
DBS	0.8	0.9	0.7	0.8	0.5
b. Desired significance level of 5 per cent					
SNV	4.8	8.8	1.4	11.3	13.4
SBS	5.2	7.0	4.0	7.7	8.7
DBS	4.8	5.6	4.6	5.7	5.8
c. Desired significance level of 10 per cent					
SNV	10.0	13.3	4.7	15.1	16.1
SBS	10.6	11.8	8.8	12.5	12.4
DBS	10.1	10.9	9.7	11.1	11.0

Notes: The $F(6, n_1 - 6)$ distribution provides critical values for the SNV test. The tests denoted by SBS and DBS are derived from single bootstrap and double bootstrap methods, respectively.

Table 4.3 Estimated rejection probabilities for SNV, SBS and DBS versions of Chow's predictive test, $n_1 = 74$ and $n_2 = 6$, with desired significance levels of 1 per cent, 5 per cent and 10 per cent.

Error distribution	<i>Normal</i>	<i>t(5)</i>	<i>Uniform</i>	$\chi^2(2)$	<i>Log-Normal</i>
a. Desired significance level of 1 per cent					
SNV	0.9	4.2	0.0	6.2	9.3
SBS	1.1	1.5	0.7	1.5	1.7
DBS	0.8	0.8	0.9	0.8	0.6
b. Desired significance level of 5 per cent					
SNV	4.9	8.8	0.9	11.1	12.9
SBS	5.2	6.5	4.1	6.8	7.2
DBS	4.7	5.4	4.6	5.5	5.9
c. Desired significance level of 10 per cent					
SNV	9.8	12.8	3.7	14.8	15.2
SBS	10.3	10.7	8.8	11.0	10.8
DBS	10.0	9.9	9.6	10.0	9.6

Notes: The $F(6, n_1 - 6)$ distribution provides critical values for the SNV test. The tests denoted by SBS and DBS are derived from single bootstrap and double bootstrap methods, respectively.

The estimates for the SNV test in Tables 4.1 to 4.3 show the importance of Normality for all combinations of n_1 and n_2 . This test is exactly valid when the errors are independent $N(0, \sigma^2)$ variables. Given the number of replications, it is, therefore, anticipated that estimates for SNV will be close to desired significance levels when the error distribution is Normal. However, there are important differences between estimates and desired significance levels under the non-Normal error distributions. The results suggest that, under the uniform error distribution, the actual null rejection probability is smaller than desired. The other non-Normal distributions lead to estimates that are rather greater than the desired levels, with this tendency being especially marked under the heavily skewed $\chi^2(2)$ and Log-Normal distributions. There is very clear evidence that the standard Chow predictive test is not robust to non-Normality.

The single bootstrap test SBS is only asymptotically valid, whether or not the error distribution is Normal. Tables 4.1 to 4.3 reveal the expected tendency of the behaviour of SBS to improve as n_1 increases. The estimates in these tables also show that SBS provides improved agreement with the desired significance levels, relative to the asymptotically invalid SNV test, when the error distribution is not Normal. However, the SBS test is still sensitive to the substantial skewness of $\chi^2(2)$ and Log-Normal distributions, with its estimated rejection probabilities being greater than the corresponding desired significance levels.

As expected from the analysis in Beran (1988), the rejection rates contained in Tables 4.1 to 4.3 strongly suggest that the double bootstrap test DBS provides better control of finite sample significance levels than SBS. Indeed, with the exception of cases that combine the smallest estimation sample size $n_1 = 24$ with the $\chi^2(2)$ and Log-Normal error distributions, the asymptotically valid DBS procedure has estimates that are quite close to the desired values. Overall, the conclusion that emerges is that, when carrying out predictive tests, the double bootstrap test DBS is a much more reliable basis for inference than either the single bootstrap check SBS or the Normality-valid SNV test described in Chow (1960). Indeed, the evidence from the simulation experiments shows that the textbook Chow test SNV appears to be so sensitive to non-Normality that it cannot be recommended for routine use in modern applied econometrics.

4.2.4. Dynamic regression models

The discussions of bootstrapping predictive tests and simulation experiments above have been based upon the assumption that the regressors

are strictly exogenous. The analysis can be generalized to allow for the inclusion of lagged values of the dependent variable in the regressor set. Recursive bootstrap methods, as described in Section 2.3.3, can be implemented when the regression model is dynamic; see Godfrey and Orme (2000) for details and evidence from simulation experiments. Further extensions of the basic framework are available and several important results, covering nonlinear models and estimation by instrumental variables, are available in Andrews (2003b).

4.3. Using bootstrap methods with a battery of OLS diagnostic tests

A common feature of modern computer programs for the estimation of a regression equation is that they offer a variety of OLS-based diagnostics designed to check the adequacy of the assumed model. There is some variation between programs, but a basic set of checks often includes:

- (i) a version of the RESET test of Ramsey (1969);
- (ii) procedures to detect serial correlation; and
- (iii) tests for heteroskedasticity.

As well as covering different sorts of violations of the assumptions of a linear regression model, the diagnostic checks of a program may allow the calculation of several different tests for a given type of misspecification. For example, the user may be allowed to choose the number of test variables for RESET of (i). For checks of type (ii), programs routinely permit users to choose the order of the alternative hypothesis for the Breusch-Godfrey LM tests and sometimes also give the statistic d discussed in Durbin and Watson (1950, 1951) and/or portmanteau checks; see Greene (2008, section 19.7) for a useful summary of tests for serial correlation. When obtaining checks of type (iii) for heteroskedasticity, the user may be able to define the regressor set for artificial regressions like (3.13) and (3.14), which are used to produce the test statistic. This breadth of coverage of diagnostic checks is a useful feature of programs. By its very nature, testing for misspecification has to be carried out without precise information since any very clear ideas about specification will presumably be used in the initial modelling.

Applied workers have taken advantage of tools for model checking and now often report a sizeable battery of diagnostics to accompany the standard OLS estimation results. However, an important problem arises when several diagnostics are used. As explained in Section 1.6,

the applied researcher cannot control the overall significance level using standard results. There are two sources of difficulty that impede this control. The first obstacle to precise control is that there may be important differences between actual and desired significance levels in finite samples when, as is often the case, individual tests are only asymptotically valid. The second difficulty would be present even if the asymptotic and finite sample significance levels of each check were equal. Unless all the individual tests are, under the null hypothesis of correct specification, independent, the overall significance level is unknown. Equivalently, it is not possible, in general, to determine the probability that a correctly specified model will survive all the checks to which it is subjected.

Standard asymptotic theory provides very limited guidance about the overall significance level of a collection of separate tests. Only bounds are available for the general case of asymptotically valid and dependent tests; see Darroch and Silvey (1963, section 2). The Bonferroni inequality implies that the overall significance level lies between the maximum of the significance levels of the individual tests and the sum of these individual significance levels. Thus, for example, with 6 diagnostic checks, each with an asymptotic significance level of 5 per cent, the overall asymptotic significance level is between 5 per cent and 30 per cent. The use of asymptotically valid bounds clearly cannot deliver precise control of the overall significance level of a battery of tests, especially when asymptotic theory does not give an accurate approximation to the unknown finite sample behaviour of individual diagnostic checks. The purpose of this section, which is based upon Godfrey (2005), is to examine how bootstrap methods might be used when no standard asymptotic procedure can be applied to control the overall significance level of a group of OLS-based checks for misspecification.

Before giving details of the bootstrap algorithms, it may be useful to comment on the purpose of the well-established practice of inspecting a collection of separate diagnostic checks after OLS estimation. The validity of conventional OLS analysis of the model that is under scrutiny requires that all of the null hypotheses of, for example, (i), (ii) and (iii) are true. Thus, the applied worker looking at such a battery of OLS diagnostic tests is interested in whether or not the separate checks of types (i)–(iii) above provide strong evidence against the *intersection null hypothesis* of correct mean function specification, no serial correlation and homoskedasticity. This emphasis on the joint validity of the separate null hypotheses is not typical of more general problems in multiple hypothesis testing; see Benjamini and Yekutieli (2001). It could be argued that a joint test approach would be more appropriate for econometric

diagnostic checks, but the reality is that it is standard practice to use a group of separate tests; see Godfrey and Veall (2000) for a discussion of joint and separate tests for misspecification.

4.3.1. Regression models and diagnostic tests

Suppose that a time series regression is under test. (If the regression model were to be estimated using cross-section data, there would be no interest in serial correlation tests, unless some sort of, for example, spatial correlation were suspected.) Using the subscript t to denote a typical time series observation, a stable dynamic regression equation can be written as in (2.35) of Chapter 2,

$$y_t = \mathbf{y}'_{t(p)}\boldsymbol{\alpha} + \mathbf{x}'_t\boldsymbol{\beta} + u_t = \mathbf{w}'_t\boldsymbol{\gamma} + u_t, t = 1, \dots, n, \quad (4.18)$$

in which: $\mathbf{y}'_{t(p)} = (y_{t-1}, \dots, y_{t-p})$, $p \geq 1$; \mathbf{x}_t contains a typical observation on each of k strictly exogenous variables; $\mathbf{w}'_t = (\mathbf{y}'_{t(p)}, \mathbf{x}'_t)$; $\boldsymbol{\gamma}' = (\boldsymbol{\alpha}', \boldsymbol{\beta}')$; and the errors u_t are assumed to be IID with zero mean, finite positive variance and unknown CDF \mathcal{F} . The model parameter vector is $\boldsymbol{\theta} = (\boldsymbol{\alpha}', \boldsymbol{\beta}', \mathcal{F})'$.

Diagnostic tests, which are designed to detect errors in the specification of (4.18), are assumed to be calculated after OLS estimation. The OLS estimator of $\boldsymbol{\gamma}' = (\boldsymbol{\alpha}', \boldsymbol{\beta}')$ is denoted by $\hat{\boldsymbol{\gamma}}' = (\hat{\boldsymbol{\alpha}}', \hat{\boldsymbol{\beta}}')$ and the associated OLS predicted values are $\hat{y}_t = \mathbf{w}'_t\hat{\boldsymbol{\gamma}}'$, $t = 1, \dots, n$. The corresponding OLS residuals, which are often used as proxies for the unobservable errors, are $\hat{u}_t = y_t - \hat{y}_t$, $t = 1, \dots, n$. Under the assumption that (4.18) contains an intercept term, the EDF of the OLS residuals can be used to generate bootstrap errors. This EDF is given in (4.8). Asymptotically valid alternatives involving modifications of OLS residuals, such as those given in (2.31) and (2.32), could be used, but are more difficult to justify in the context of dynamic regression models. (If no intercept is present in (4.18), the OLS residuals must be recentred to sum to zero before being used for resampling.)

It is worth stressing that it is assumed that the general form of the errors' CDF \mathcal{F} is unknown. This treatment of \mathcal{F} seems reasonable, given that it is unlikely that much information about the error distribution will be available in practical situations. Some researchers do assume that the error distribution is Normal, but this is a very strong assumption for which there is little justification. Moreover, many OLS procedures must be based upon asymptotic theory in the context of the dynamic model (4.18) and such theory usually applies with general non-Normal error distributions. Tests for the error distribution of the type discussed,

for example, in Jarque and Bera (1980) and Neumeyer et al. (2004) are, therefore, assumed not to be included in the battery of diagnostics. An implication of not imposing Normality is that, for example, the Breusch-Pagan (1979) check for heteroskedasticity and the predictive test statistic given in Chow (1960) are not asymptotically pivotal.

The diagnostic tests to be applied after OLS estimation of (4.18) are denoted by $\tau_j, j = 1, \dots, J$. It is convenient, for the discussion of p -values and the description of bootstrap methods below, to assume all diagnostic checks have rejection regions in the right-hand tail of the (asymptotic) null distribution of the test statistic. Tests of a single restriction may require a minor adjustment to fit into this framework. For example, consider Durbin's (1970) h -statistic, which is asymptotically $N(0, 1)$ when its null hypothesis of serial independence is true. If the alternative for this test is two-sided, the adjustment is to use $\tau = h^2$, with the asymptotic reference distribution being $\chi^2(1)$. If a one-sided alternative of negative autocorrelation is thought to be appropriate, the adjustment is to use $\tau = -h$. If a one-sided alternative of positive autocorrelation is thought to be appropriate, no adjustment is needed and $\tau = h$.

Let the number of restrictions tested by τ_j be denoted by $d_j, j = 1, \dots, J$. If all the diagnostics tested the same number of restrictions, an overall test might be derived by using the maximum of the test statistics as the criterion, but equality of the degrees-of-freedom parameters d_j is unlikely. It is more useful to examine the general case with $d_j \neq d_l$, for some j and l . In this general situation, it is not the magnitudes of individual diagnostics, but the corresponding p -values that reflect strength of evidence. In particular, for an indication of the strongest piece of evidence, the minimum of these p -values can be examined. The strength of the evidence against the assumed regression model that is provided by the minimum p -value is assessed using two levels of bootstrapping.

4.3.2. Bootstrapping the minimum p -value of several diagnostic test statistics

As noted above, the inclusion of the error distribution CDF \mathcal{F} in the model parameter vector $\theta = (\alpha', \beta', \mathcal{F})'$ implies that some well-known diagnostic checks are not asymptotically pivotal. For diagnostic test statistics that are not asymptotically pivotal, a single bootstrap can be used to obtain asymptotically valid inferences; see Beran (1988). In fact, it is proposed in Godfrey (2005) that all regression diagnostics being applied to the model under scrutiny should be implemented using bootstrap methods, even if they are asymptotically pivotal. (For any asymptotically pivotal tests being used, a bootstrap produces an ERP of

smaller order in n than that associated with the use of asymptotic critical values; see Beran, 1988.) Having used a single stage of bootstrapping to obtain estimated p -values for all of the individual checks, the minimum of these estimates can be derived. This minimum is not asymptotically pivotal and, in order to assess its statistical significance at a desired level α_d , it is necessary to use another stage of bootstrapping, again making appeal to the results in Beran (1988). The details of the bootstrap test of the minimum estimated p -value are as follows.

Minimum p -value test: Step 1

In the first step, the actual data are analysed. The null model (4.18) is estimated by OLS, using the genuine data set of n observations, and the sample values of the J diagnostic checks are calculated. These observed values are denoted by $(\tau_1^0, \dots, \tau_J^0)$.

It is recommended in Godfrey (2005) that the p -values of the observed test statistics τ_j^0 are estimated using a first-stage bootstrap, whether or not standard asymptotic distributions are available. In this first level of bootstrapping, B artificial samples, each of size n , are obtained. The next two steps are, therefore, repeated B times. Step 2 is used to generate bootstrap data and Step 3 involves applying to these bootstrap data the statistical techniques used in Step 1 with the actual data.

Minimum p -value test: Step 2

The data for bootstrap sample b are obtained using the recursive scheme

$$y_{bt}^* = \hat{\alpha}_1 y_{b,t-1}^* + \dots + \hat{\alpha}_p y_{b,t-p}^* + \mathbf{x}'_t \hat{\boldsymbol{\beta}} + u_{bt}^*, \quad t = 1, \dots, n, \quad (4.19)$$

where bootstrap sample starting values are set equal to actual estimation sample starting values (see Li and Maddala, 1996, section 2.3) and $\mathbf{u}_b^* = (u_{b1}^*, \dots, u_{bn}^*)'$ is derived by simple random sampling, with replacement, from the EDF in (4.8).

Minimum p -value test: Step 3

The regression model is then estimated using the bootstrap data from Step 2 to obtain OLS coefficient estimates, residuals and diagnostic test statistics, denoted by $\hat{\boldsymbol{\gamma}}_b^* = (\hat{\alpha}_{b1}^*, \dots, \hat{\alpha}_{bp}^*; \hat{\boldsymbol{\beta}}_b^{*'})'$, $\hat{\mathbf{u}}_b^* = (\hat{u}_{b1}^*, \dots, \hat{u}_{bn}^*)'$, and $(\tau_{b1}^*, \dots, \tau_{bJ}^*)$.

Minimum p -value test: Step 4

When Step 2 and Step 3 have been carried out B times, that is, for $b = 1, \dots, B$, the p -values of the observed values of the diagnostics τ_j^0 are

estimated in the fourth step by

$$\hat{p}_j^* = \frac{\sum_{b=1}^B \mathbf{1}(\tau_{bj}^* \geq \tau_j^0)}{B}, j = 1, \dots, J. \quad (4.20)$$

The minimum of these estimated bootstrap p -values is

$$\widehat{mp}^* = \min_j(\hat{p}_j^*). \quad (4.21)$$

An assessment of the statistical significance of \widehat{mp}^* of (4.21) is derived via a second-stage bootstrap. In the second-stage bootstrap, C artificial samples of size n are generated from each of the B first-stage bootstrap data sets. Consequently Step 5 and Step 6 must be repeated C times for each $b, b = 1, \dots, B$.

Minimum p -value test: Step 5

The generation of second-level bootstrap data is carried out in the fifth step. For each b , a typical second-level bootstrap sample of n observations is obtained using

$$y_{bct}^{**} = \hat{\alpha}_{b1}^* y_{bc,t-1}^{**} + \dots + \hat{\alpha}_{bp}^* y_{bc,t-p}^{**} + \mathbf{x}_t' \hat{\boldsymbol{\beta}}_b^* + u_{bct}^{**}, t = 1, \dots, n, \quad (4.22)$$

in which bootstrap sample starting values are set equal to actual estimation sample starting values and the n elements of $\mathbf{u}_{bc}^{**} = (u_{bc1}^{**}, \dots, u_{bcn}^{**})'$ are selected by simple random sampling, with replacement, from the first-level bootstrap residual EDF

$$\hat{\mathcal{F}}_b^* : \text{probability } 1/n \text{ on } \hat{u}_{bt}^*, t = 1, \dots, n. \quad (4.23)$$

Minimum p -value test: Step 6

In Step 6, the OLS procedures applied to the actual data in Step 1 are applied to the second-level bootstrap data from Step 5. Let the test statistics calculated from a typical second-stage bootstrap sample be denoted by $\tau_{bcj}^{**}, j = 1, \dots, J$.

Minimum p -value test: Step 7

Step 7 is the bootstrap-world counterpart of Step 4. After repeating Step 5 and Step 6 C times, the p -value for the first-level bootstrap statistic τ_{bj}^* , for any pair (b, j) , can be estimated by

$$\hat{p}_{bj}^{**} = \frac{\sum_{c=1}^C \mathbf{1}(\tau_{bcj}^{**} \geq \tau_{bj}^*)}{C}, j = 1, \dots, J \text{ and } b = 1, \dots, B. \quad (4.24)$$

Next, for a given value of b , the minimum of the estimated p -values \hat{p}_{bj}^{**} over the J tests can be found. This minimum p -value is denoted by

$$\widehat{mp}_b^{**} = \min_j(\hat{p}_{bj}^{**}). \quad (4.25)$$

Minimum p -value test: Step 8

The final step of the minimum p -value test is straightforward. After the first and second stages of bootstrapping have been completed, the B values of \widehat{mp}_b^{**} in (4.25) can be used to approximate the sampling distribution of \widehat{mp}^* of (4.21). Unusually small values of \widehat{mp}^* signal evidence of model inadequacy and the left-hand tail p -value of \widehat{mp}^* of (4.21) is estimated by

$$\hat{p}_{mp}^* = \frac{\sum_{b=1}^B \mathbf{1}(\widehat{mp}_b^{**} \leq \widehat{mp}^*)}{B}. \quad (4.26)$$

The rejection rule for the desired (nominal) overall significance level α_d is to reject the null model after application of the battery of checks if $\hat{p}_{mp}^* \leq \alpha_d$. Under regularity conditions, the test using this rule is asymptotically valid, but enjoys no refinement relative to the asymptotic theory test, despite involving two levels of bootstrapping.

4.3.3. Simulation experiments and results

Two experiments are employed to examine the reliability of the two-stage bootstrap method in finite samples. In the first experiment, hereafter Experiment A, the situation considered is one in which several checks are used to detect the same general type of misspecification. For the second experiment, hereafter Experiment B, the diagnostic checks consist of separate tests for invalid mean function, autocorrelation and heteroskedasticity, which correspond to (i), (ii) and (iii) above.

Experiments A and B have the following features in common. The bootstrap test based upon the minimum p -value of the diagnostic test statistics is implemented using $B = 500$ first-stage bootstrap samples and $C = 100$ second-stage bootstrap samples. The desired significance levels studied are 5 per cent and 10 per cent. Estimated rejection rates are calculated using $R = 10,000$ replications.

Experiment A is based upon a design described in Section 1.5.1. Consider the problem of testing the log-log estimating equation of a Cobb-Douglas model for omitted variables and/or incorrect functional

form. The restricted (null) model is

$$y_i = \beta_1 + \beta_2 x_{i2} + \beta_3 x_{i3} + u_i, E(u_i | x_{i2}, x_{i3}) = 0, \quad (4.27)$$

with the errors being IID with CDF \mathcal{F} and the variables of (4.27) being as defined in Section 1.5.1. The real world data for the dependent variable and regressors in Greene (2008) are used to obtain the corresponding OLS estimate $\hat{\mathbf{b}} = (\hat{b}_1, \hat{b}_2, \hat{b}_3)'$ and the error variance estimate s^2 . In order to investigate the effects of increasing sample size, an artificial data set of 54 observations on regressors is generated by setting

$$x_{i+27,2} = x_{i2} \text{ and } x_{i+27,3} = x_{i3}, i = 1, \dots, 27.$$

In the simulations for Experiment A, data on the dependent variable are generated using

$$y_i = \hat{b}_1 + \hat{b}_2 x_{i2} + \hat{b}_3 x_{i3} + u_i, i = 1, \dots, n, \quad (4.28)$$

with the errors u_i being IID drawings from Normal, $t(5)$ and $\chi^2(2)$ distributions, which are transformed to have zero population mean and population variance equal to the error variance estimate s^2 from OLS estimation of (4.27) using the actual data set of 27 observations. The sample size n is either 27 or 54.

The simulated data of Experiment A are used to investigate the usefulness of the bootstrap minimum p -value approach when three general checks of the mean function are carried out. These checks are as follows: a RESET test that uses only \hat{y}_i^2 as a test variable; a RESET test that uses the three variables of $(\hat{y}_i^2, \hat{y}_i^3, \hat{y}_i^4)$; and a procedure proposed in Thursby and Schmidt (1977) that involves testing (4.27) against

$$y_i = \beta_1 + \beta_2 x_{i2} + \beta_3 x_{i3} + \delta_1 x_{i2}^2 + \delta_2 x_{i3}^2 + \delta_3 x_{i2}^3 + \delta_4 x_{i3}^3 \\ + \delta_5 x_{i2}^4 + \delta_6 x_{i3}^4 + u_i.$$

The numbers of restrictions tested by these three checks are 1, 3 and 6, respectively.

The data generation process for Experiment B is based upon a simple version of an Okun's law-type relationship for time series data. The dependent variable y_t is to be interpreted as the change in the unemployment rate and the regressor x_t is the growth rate of output. Quarterly data for 1950.1 to 1983.4 on US output levels are used to obtain the values of

x_t ; see Godfrey (2005, section IV) for details. The data generation process for simulated samples is

$$y_t = \beta_1 + \beta_2 x_t + \beta_3 x_{t-1} + u_t, \quad (4.29)$$

in which $(\beta_1, \beta_2, \beta_3) = (1.5, -0.5, 0.0)$, as suggested by published OLS estimates. The errors for (4.29) are drawn in the same way as in Experiment A, except that the error variance is selected to give an average R^2 of about 0.5, corresponding to typical values observed in empirical work. The sample size is again either 27 or 54.

The model of (4.29) is subjected to three tests, each for a different general type of misspecification. First, as in Experiment A, a RESET test using only the squared value of the OLS predicted value is employed to detect omitted variables or incorrect functional form, so $d_1 = 1$. Second, the LM test of $\phi_1 = \dots = \phi_4 = 0$ in the expanded model

$$y_t = \beta_1 + \beta_2 x_t + \beta_3 x_{t-1} + \sum_{j=1}^4 \phi_j \hat{u}_{t-j} + u_t,$$

provides a check against general fourth order serial correlation, with $d_2 = 4$, as might be useful in a genuine empirical study based upon quarterly data. Third, a test of the type proposed by Koenker (1981) is applied to detect heteroskedasticity. It is implemented by calculating n times the coefficient of determination from the OLS estimation of the artificial regression of \hat{u}_t^2 on \hat{y}_t^2 and an intercept term, so $d_3 = 1$.

In both experiments, the bootstrap procedures described in the previous subsection are applied to each of the 10,000 samples of data simulated, given a choice of n and the error distribution. The proportion of replications in which comparison of \hat{p}_{mp}^* in (4.26) with α_d leads to rejection of the true intersection null hypothesis gives the estimate of the finite sample significance level that corresponds to α_d , for $\alpha_d = 5$ per cent or 10 per cent. The results are summarized in Table 4.4, which contains estimates as percentages, rounded to one decimal place.

Table 4.4 shows that, while the estimates are often slightly smaller than the desired level, there a reasonably good degree of control of the overall significance level in the examples used for Experiments A and B. Given that the number of replications is 10,000, estimators of significance levels can reasonably be treated as being approximately Normal and the test described in Godfrey and Orme (2000, p. 75) can be applied to assess the usefulness of the bootstrap procedure. According to this test every estimate in Table 4.4 for a case with $\alpha_d = 5$ per cent is consistent

Table 4.4 Estimated significance levels for bootstrap test of minimum p -value for battery of diagnostics

Error distribution	Normal	$t(5)$	$\chi^2(2)$
a. Desired significance level of 5 per cent			
<i>Experiment A</i> , $n = 27$	4.8	4.6	4.7
<i>Experiment A</i> , $n = 54$	4.8	4.5	4.4
<i>Experiment B</i> , $n = 27$	4.7	4.6	4.2
<i>Experiment B</i> , $n = 54$	4.7	4.5	4.1
b. Desired significance level of 10 per cent			
<i>Experiment A</i> , $n = 27$	9.6	9.2	9.6
<i>Experiment A</i> , $n = 54$	9.4	9.5	9.4
<i>Experiment B</i> , $n = 27$	10.5	10.2	8.8
<i>Experiment B</i> , $n = 54$	10.2	9.9	9.9

with the claim the true rejection probability is between 4.4 per cent and 5 per cent. Similarly, in Panel b of Table 4.4, with $\alpha_d = 10$ per cent, the test indicates every estimate is consistent with the claim that the true rejection probability is between 9 per cent and 10.1 per cent. This evidence about the degree of control with sample sizes as small as 27 and 54 is encouraging, especially under the extremely skewed errors derived from the $\chi^2(2)$ distribution.

It is reasonable to expect that even better results would be obtained if a double, rather than single, bootstrap were used with the asymptotically nonpivotal test statistic \widehat{mp}^* of (4.21). However, three levels of bootstrapping would be required to achieve such improvements and the computational costs might not be justified. A straightforward double-bootstrap test is available in the special case in which every test statistic is asymptotically pivotal and has a standard limit null distribution. In such a situation, there is no absolute necessity to obtain p -values of individual statistics by simulation. Instead suitable computer routines can provide p -values from each of the relevant standard asymptotic distributions.

Let the CDF of the asymptotic null distribution of τ_j be denoted by \mathcal{G}_j , $j = 1, \dots, J$. Given the assumption about the nature of the rejection region, the asymptotic p -value of the observed value τ_j^o is $p_j^a = 1 - \mathcal{G}_j(\tau_j^o)$, $j = 1, \dots, J$, and the minimum of these values is

$$mp^a = \min_j (1 - \mathcal{G}_j(\tau_j^o)), \quad (4.30)$$

which is the asymptotic theory counterpart of the bootstrap-based term \widehat{mp}^* of (4.21). The double bootstrap described in Section 2.5 can be applied to mp^a of (4.30) and is expected, on the basis of results in Beran (1988), to have finite sample significance levels that are closer to the desired levels than the single bootstrap test of \widehat{mp}^* of (4.21).

Godfrey provides simulation evidence on the relative sizes of ERP terms for the single bootstrap \widehat{mp}^* -test and the double bootstrap mp^a -test; see Godfrey (2005, section V). The experimental design corresponds to Experiment A above, except that 50,000 replications are used to obtain accurate estimates of differences in null rejection probabilities. These estimates are summarized in Table 4.5. The contents of Table 4.5 are in line with what is expected from the asymptotic analysis in Beran (1988). The estimated ERP term of the double bootstrap mp^a -test is, in every case, smaller than the corresponding value for the single bootstrap \widehat{mp}^* -test. Indeed the rejection rates of the former test are very close to the desired level.

While the mp^a -test appears to be superior to the \widehat{mp}^* -test, the former, unlike the latter, is only available when all test statistics being calculated are asymptotically pivotal. It would, therefore, seem sensible to avoid the unnecessary use of test statistics that are not asymptotically pivotal in applied work. For example, rather than using the LM statistic given in Breusch and Pagan (1979) to test for heteroskedasticity, the Studentized version given in Koenker (1981) can be employed. However, as explained in the previous section, it is not possible to adjust the statistic for Chow’s predictive test to obtain an asymptotically pivotal version. Consequently, the mp^a -test cannot be used to control the overall

Table 4.5 Estimated significance levels obtained from experiment A for the single bootstrap \widehat{mp}^* -test and the double bootstrap mp^a -test

		<i>n</i> = 27		<i>n</i> = 54	
		\widehat{mp}^* -test	mp^a -test	\widehat{mp}^* -test	mp^a -test
a. Desired significance level is $\alpha_d = 5$ per cent					
Error distribution is	<i>Normal</i>	4.4	4.7	4.6	4.8
Error distribution is	<i>t</i> (5)	4.5	4.8	4.3	4.6
Error distribution is	χ^2 (2)	4.6	5.0	4.5	4.7
b. Desired significance level is $\alpha_d = 10$ per cent					
Error distribution is	<i>Normal</i>	9.2	9.9	9.3	10.1
Error distribution is	<i>t</i> (5)	9.5	10.3	9.3	10.1
Error distribution is	χ^2 (2)	9.3	10.2	9.1	10.0

significance level when the predictive test is one of the battery of diagnostic checks. Godfrey obtains results on the performance of the $\widehat{m\hat{p}}^*$ -test when a battery of three diagnostic checks consists of a predictive test with $n_1 = 1$, combined with the autocorrelation and heteroskedasticity checks used in Experiment B; see Godfrey (2005, p. 276, Table 2). Godfrey finds that all estimates for the $\widehat{m\hat{p}}^*$ -test are consistent with Serlin's stringent condition that the actual finite sample significance level should be between $0.9\alpha_d$ and $1.1\alpha_d$; see Serlin (2000).

4.4. Bootstrapping tests for structural breaks

One of the best known tests for regression models is the F -test for structural breaks described in Chow (1960); see, for example, Hill et al. (2008, pp. 179–181) and Verbeek (2004, p. 64). The null hypothesis of the original Chow test is that the regression model (4.1), with NID errors, applies to all n observations to be used. In its simplest form, the alternative hypothesis is that there are two population models, which differ only in the value of the regression coefficient vector, and that, on the basis of nonsample information, it is known, under this alternative, which observations belong to each of the two populations. The corresponding subsample sizes are n_1 and n_2 , with $n_1 + n_2 = n$. For Chow's F -test to be available, it is required that $n_1 > k$ and $n_2 > k$, so that separate subsample estimation of all regression coefficients is possible.

Under the basic version of the alternative hypothesis, the unrestricted regression model in Chow (1960) can be written as

$$y_i = \mathbf{x}'_i \boldsymbol{\beta} + \mathbf{x}'_i (d_i \boldsymbol{\gamma}) + u_i, i = 1, \dots, n, \quad (4.31)$$

in which the errors u_i are $NID(0, \sigma^2)$ and d_i is a dummy variable, taking the values 0 or 1, with its value known for every observation. For the n_1 observations with $d_i = 0$, the regression model is

$$y_i = \mathbf{x}'_i \boldsymbol{\beta} + u_i, u_i NID(0, \sigma^2),$$

and for the n_2 observations with $d_i = 1$, it is

$$y_i = \mathbf{x}'_i (\boldsymbol{\beta} + \boldsymbol{\gamma}) + u_i, u_i NID(0, \sigma^2),$$

so that Normality and homoskedasticity are used as supporting assumptions for the F -test of $\boldsymbol{\gamma} = \mathbf{0}_k$ in (4.31). Provided that standard regularity conditions are satisfied, this F -test is asymptotically valid, under

unspecified forms of non-Normality, when k , the number of regression coefficients, is fixed and both subsample sizes tend to infinity, with $n_1 = O(n)$ and $n_2 = O(n)$. In keeping with previous discussions of tests for regression models, it will henceforth be assumed that the errors of (4.31) are simply IID with a CDF of unspecified general form, which is denoted by \mathcal{F} .

Extensions of the alternative hypothesis that relax the assumption of IID errors by allowing for variance parameter changes will be discussed in Chapter 6. Other types of modifications of the alternative hypothesis are possible, for example, (i) there might be more than two different regimes that are identified by the prior information, or (ii) it might be known that structural breaks, if present, only affect some of the elements of the regression coefficient vector. Such extensions can be handled using standard asymptotic theory for regression tests. However, as noted in Section 1.6, there is an extension of the alternative hypothesis that has great practical relevance and does not lead to a test using a standard asymptotically valid F -distribution for critical values. This extension is based upon recognition of the fact that there is often uncertainty about how the sample should be divided into subsamples.

In terms of the extended regression model of (4.31) with IID errors, there may well be uncertainty about the value of d_i for some values of i . If the data are ordered, for example, as in a time series study, and the two subsamples correspond to the first n_1 observations and the last n_2 observations, the problem can be referred to as being that the breakpoint under the alternative is unknown. Stock and Watson stress that applied workers who use the data in some way (either formally or informally) to determine the possible breakpoint cannot appeal to the standard textbook results in Chow (1960), which are based upon the assumption that the breakpoint is known on the basis of nonsample (a priori) information; see Stock and Watson (2007, pp. 569–570).

The problem of the absence of genuine nonsample information about the location of the structural break could be tackled by randomly selecting a value of n_1 that satisfies $k < n_1 < n - k$. This artificial device avoids using the data but may be of limited practical value. Hansen points out that this sort of arbitrary selection can lead to a true breakpoint being missed; see Hansen (2001, p. 118). A more systematic data-based method is required, but, as will be seen, such extensions of the original Chow test lead to statistics which have non-standard asymptotic distributions. It is certainly not valid to continue to rely upon the null distributions in Chow (1960), which are obtained under the assumption that the breakpoint for the alternative is known.

The purpose of this section is to outline asymptotic and bootstrap tests when there is a single unknown breakpoint under the alternative hypothesis. A very useful discussion of this case and other more general aspects of testing for structural breaks is provided in Perron (2006, section 8.4).

4.4.1. Testing constant coefficients against an alternative with an unknown breakpoint

Suppose then that, under the alternative that is being entertained, there is a single breakpoint for the regression coefficient vector, which puts the first n_1 observations in one regime and the last n_2 observations in a different regime. The alternative model can be written as

$$y_i = \sum_{j=1}^k x_{ij}\beta_j + \sum_{j=1}^k \mathbf{1}(i > n_1)x_{ij}\gamma_j + u_i, \quad (4.32)$$

in which $\mathbf{1}(\cdot)$ is the indicator function and the errors are IID with CDF \mathcal{F} . The value of n_1 is treated as an unknown constant in (4.32). If n_1 were instead known, it would, under general conditions, be asymptotically valid to compare the standard F -statistic for testing $\gamma_1 = \gamma_2 = \dots = \gamma_k = 0$ in (4.32) with critical values from the $F(k, n-2k)$ distribution. However, with the breakpoint being unknown, problems arise when constructing tests. Andrews summarizes some of the statistical literature on these problems in a very influential article; see Andrews (1993).

There are two points to note about the exposition in Andrews (1993). First, there are minor differences in terminology. Andrews refers to “structural change”, rather than “structural break” and to the “change point”, rather than the “breakpoint”. Second, Andrews finds it convenient to redefine a relevant parameter of interest and, rather than refer to n_1 as the breakpoint, he introduces $\pi = n_1/n$, which is treated as the unknown term to be estimated. Clearly π only appears as a parameter under the alternative hypothesis and it is this non-standard feature of the test problem that leads to complications.

The testing of null hypotheses when a nuisance parameter is present only under the alternative hypothesis is discussed in Davies (1977). Davies considers the application of the likelihood ratio principle and suggests that the maximized test statistic over all possible values of the parameter that vanishes under the null be used as a criterion for testing. The use of the same general idea in the specific context of linear (simple) regression models with two different regimes is suggested in two early articles by Quandt; see Quandt (1958, 1960).

Under the assumption of NID errors, the log-likelihood for (4.32) can be maximized over all possible values of n_1 : recall the inequalities $n_1 > k$ and $n_2 > k$ must both be satisfied. The maximum of these maxima can then be compared with the maximized log-likelihood for the null model (4.1) to arrive at a test statistic. This strategy leads to the test statistic

$$\sup_{\pi \in \Pi} LR_n(\pi), \quad (4.33)$$

in which: $LR_n(\pi)$ denotes the LR statistic for testing $\gamma_1 = \gamma_2 = \dots = \gamma_k = 0$ in (4.32), calculated using the breakpoint $\lfloor n\pi \rfloor$, with $\lfloor \cdot \rfloor$ being the integer part operator for the nonnegative term $n\pi$; and Π consists of values that satisfy

$$0 < \pi_1 \leq \pi \leq \pi_2 < 1,$$

for specified values of the lower bound π_1 and the upper bound π_2 . For notational convenience, the statistic in (4.33) will henceforth be written as *SupLR*. Quandt focusses on the likelihood ratio (LR) approach but the LM and Wald methods are also available. The statistics derived by using the general method in Davies (1977) with LM and Wald principles can be written as *SupLM* and *SupW*, respectively.

Quandt obtains his *SupLR* statistic under the assumption that the errors of the regression model are NID; so that OLS estimators are MLE. In keeping with the more recent contributions, this assumption is not made here. It is instead assumed that the errors are IID, with an unspecified common distribution, and that conditions for the consistency and asymptotic Normality of OLS estimators of regression coefficients are satisfied. It follows that the use of "*SupLR*", "*SupW*" and "*SupLM*" is, strictly speaking, incorrect. When the errors are IID, it would be more accurate to refer to OLS as a quasi-MLE and to modify the notation for test statistics to reflect the use of quasi-likelihood functions, for example, *SupQLR* could be used in place of *SupLR*. However, the use of a more accurate, but more cumbersome, notation would conflict with common usage in the literature and so *SupLR*, *SupLM* and *SupW* will be used, even though the OLS methods used to derive them are not the maximizers of the true likelihood function.

Standard results on the testing of linear coefficient restrictions in regression models with IID errors imply that, given a value of π and hence of n_1 , the statistics $LR_n(\pi)$, $LM_n(\pi)$ and $W_n(\pi)$ are all (different) functions of the F -statistic proposed in Chow (1960); see Godfrey (1988,

p. 51). This F -statistic is

$$F_C = \frac{RSS_0 - (RSS_1 + RSS_2)}{(RSS_1 + RSS_2)} \cdot \frac{n - 2k}{k}, \quad (4.34)$$

in which RSS_0 , RSS_1 and RSS_2 are the sums of squared residuals from the OLS regression of y_i on (x_{i1}, \dots, x_{ik}) for the full sample of n observations, the first n_1 observations and the last n_2 observations, respectively. (The RSS function from the OLS estimation of (4.32) equals $RSS_1 + RSS_2$.) When the null hypothesis is true, F_C does not depend upon the parameters $(\beta_1, \dots, \beta_k, \sigma^2)$. It follows that, given a value of π , $LR_n(\pi)$, $LM_n(\pi)$ and $W_n(\pi)$ all have distributions that are also independent of these parameters, under the null model.

It is also clear that, given a value of π , $LR_n(\pi)$, $LM_n(\pi)$ and $W_n(\pi)$ are all asymptotically distributed as $\chi^2(k)$ when $\gamma_1 = \gamma_2 = \dots = \gamma_k = 0$ in (4.32), whatever the form of the error CDF \mathcal{F} , provided that weak regularity conditions are satisfied. Thus the *Sup*-type statistics proposed in Andrews (1993) are asymptotically distributed as the supremum of a set of $\chi^2(k)$ variables, whatever the value of the parameter vector $\theta = (\beta_1, \dots, \beta_k, \mathcal{F})'$. The *Sup*-type statistics are, therefore, asymptotically pivotal. It follows that asymptotic critical values for general use may be available.

In order to obtain, for example, tables of critical values, it is necessary to determine the asymptotic null distributions of *SupLR*, *SupLM* and *SupW*. Andrews shows that the three statistics have the same non-standard limit null distribution. This equivalence under the null hypothesis corresponds to the classical result for standard test situations. It has already been established that the limit null distribution of the Andrews-type statistics does not depend upon $\theta = (\beta_1, \dots, \beta_k, \mathcal{F})'$, but it clearly depends upon k since each of the statistics is the supremum of a set of $\chi^2(k)$ variables. Andrews shows that the asymptotic null distribution of his *Sup*-statistics also depends upon

$$\lambda = \frac{\pi_2(1 - \pi_1)}{\pi_1(1 - \pi_2)}. \quad (4.35)$$

If it is decided to use symmetric trimming of the total sample, with $\pi_1 = \pi_0$ and $\pi_2 = (1 - \pi_0)$, λ of (4.35) can be replaced by

$$\lambda_s = \frac{(1 - \pi_0)^2}{\pi_0^2}. \quad (4.36)$$

Andrews suggests using symmetric trimming with $\pi_1 = 0.15$ and $\pi_2 = (1 - \pi_1) = 0.85$; see Andrews (1993, p. 826). He stresses that it is not possible to avoid the problem of selecting values for π_1 and π_2 by using the full range of values for π . With π unrestricted, the test statistics (and their critical value for any prespecified significance level) diverge to infinity, under the null hypothesis; see Andrews (1993, pp. 838–839).

A table of asymptotic critical values for the *Sup*-tests is provided in Andrews (1993). These critical values are derived using simulation; see Section 5.3 of Andrews (1993) for details. However, the critical values for $k = 8$ in Andrews (1993) are not correct. A corrected table, which is based upon more accurate procedures than the first version, appears in Andrews (2003a, Table 1, pp. 396–397). The table in Andrews (2003a) covers the following combinations: desired significance levels of $\alpha_d = 0.01, 0.05, 0.10$; $k = 1, \dots, 20$; and $\Pi = [\pi_0, 1 - \pi_0]$ for values of π_0 between 0.05 and 0.50. The form of Π implies symmetric trimming and, for each value of π_0 , Andrews gives the corresponding value of λ_s of (4.36).

Researchers wishing to use asymmetric trimming with $\pi_2 \neq (1 - \pi_1)$ can simply calculate the implied value of λ in (4.35) and then find, or approximate, the critical value by looking for similar values of λ_s of (4.36) in Table 1 of Andrews (2003a). For example, with $\pi_1 = 0.2$ and $\pi_2 = 0.6$,

$$\lambda = \frac{0.6(1 - 0.2)}{0.2(1 - 0.6)} = 6,$$

and interpolation uses the critical values for $\lambda_s = 5.44$ and $\lambda_s = 9.00$, which appear in Table 1 of Andrews (2003a).

Although the asymptotic critical values in Andrews (2003a) enable applied workers to judge the statistical significance of *Sup*-test statistics at conventional levels, some researchers may prefer to examine asymptotic *p*-values. However, the calculation of asymptotic *p*-values is not trivial because the relevant limit null distribution is non-standard. Hansen considers computationally convenient methods for approximating asymptotic *p*-values and provides empirical examples to illustrate the usefulness of his techniques; see Hansen (1997).

The analysis provided in Andrews (1993) has had a great impact and *Sup*-tests for structural breaks are often recommended for application in empirical econometrics. There are, however, two questions that need to be considered. First, do the asymptotic critical values provided by Andrews give useful approximations in finite samples? Second, given that the test statistics are asymptotically pivotal, can the application of

bootstrap methods produce useful improvements relative to the asymptotic theory tests? These questions have been addressed in a number of simulation studies.

4.4.2. Simulation evidence for asymptotic and bootstrap tests

The evidence which is discussed in this subsection is collected from published simulation studies. There are three general issues about which this evidence is informative: (i) the consequences of invalidly using the standard Chow-type critical values when the breakpoint is selected using data-based analysis of some sort (either formal or informal); (ii) the quality of the finite sample approximation provided by the asymptotic critical values in Andrews (2003a); and (iii) the merits of a bootstrap approach to assessing the statistical significance of a *Sup*-type statistic relative to the asymptotic theory method in Andrews (1993, 2003a).

Some important results are reported in Diebold and Chen (1996). Diebold and Chen carry out simulation experiments based upon the simple dynamic model in which, under the null hypothesis of no structural breaks, the data are generated by

$$y_t = \rho y_{t-1} + u_t, u_t \text{ NID}(0, 1), t = 1, \dots, n, \quad (4.37)$$

and

$$y_0 \sim N\left(0, \frac{1}{1 - \rho^2}\right), |\rho| < 1. \quad (4.38)$$

The alternative hypothesis is taken to consist of

$$y_t = \rho y_{t-1} + u_t, u_t \text{ NID}(0, 1), t = 1, \dots, n_1, \quad (4.39)$$

and

$$y_t = \rho_2 y_{t-1} + u_t, u_t \text{ NID}(0, 1), t = n_1 + 1, \dots, n, \quad (4.40)$$

with n_1 being unknown.

If the value of n_1 were known, a conventional asymptotic likelihood-based test of the single restriction $\rho = \rho_2$ would be possible, with reference being made to the $\chi^2(1)$ distribution for a critical value or asymptotic p -value. However, this approach is not appropriate when n_1 is unknown and common practice in modern econometrics is to adopt the approach discussed in Andrews (1993, 2003a).

As in Andrews (1993, 2003a), the unknown breakpoint parameter is defined to be $\pi = n_1/n$ and the values used in Diebold and Chen (1996) to implement the check for a structural break satisfy $\pi \in [0.15, 0.85]$. The associated *Sup*-type criteria are given by

$$SupLR = \max_{\pi} n \log \left[\frac{RSS_0}{(RSS_1 + RSS_2)} \right], \quad (4.41)$$

$$SupW = \max_{\pi} n \left[\frac{RSS_0 - (RSS_1 + RSS_2)}{(RSS_1 + RSS_2)} \right], \quad (4.42)$$

and

$$SupLM = \max_{\pi} n \left[\frac{RSS_0 - (RSS_1 + RSS_2)}{RSS_0} \right], \quad (4.43)$$

in which RSS_0 , RSS_1 and RSS_2 are the sums of squared residuals from the OLS estimation of (4.37), (4.39) and (4.40), respectively.

The problem that faces the applied worker who is using one of these *Sup*-statistics is to make accurate judgements about its statistical significance in a finite sample situation. In order to obtain results that assist the applied worker, data are simulated by Diebold and Chen, using (4.37) and (4.38) with a large number of combinations of ρ , n and desired significance level α_d ; see Diebold and Chen (1996, p. 225) for details. Sample sizes of $n = 10, 50, 100, 500$ and $1,000$ are examined. The parameter ρ has 18 values in the range $0.01 \leq \rho \leq 0.99$, so that there is considerable variation in the degree of autocorrelation of the dependent variable. Given the values of n and ρ , the data y_0, y_1, \dots, y_n are derived for each of 1,000 replications, using (4.37), (4.38) and a random number generator for the $N(0, 1)$ distribution. These replications are used to estimate finite sample null hypothesis rejection probabilities, which are compared with desired levels $\alpha_d = 1$ per cent, 2.5 per cent, 5 per cent and 10 per cent.

The estimates published in Diebold and Chen (1996) are derived under the assumption that the IID errors have a Normal distribution. However, Diebold and Chen report that their results are robust to departures from this assumption about the error distribution; see Diebold and Chen (1996, p. 223, footnote 2).

The following notation is used by Diebold and Chen when they discuss the results from their simulation experiments. Corresponding to (4.41), (4.42) and (4.43), the asymptotically invalid tests that use the $\chi^2(1)$ distribution for critical values are denoted by *ChiSupLR*,

ChiSupW and *ChiSupLM*. When the asymptotic critical values in Andrews (1993) are used with the *Sup*-statistics, the resulting tests are referred to as *AsySupLR*, *AsySupW* and *AsySupLM*. Finally the bootstrap variants of the *Sup*-test for a structural break with unknown breakpoint are denoted by *BootSupLR*, *BootSupW* and *BootSupLM*. These bootstrap tests are carried out using 1,000 bootstrap samples and their implementation for a typical replication of a simulation experiment can be described as follows.

Diebold and Chen (1996): Step 1

The first step, given the values of n and ρ , is to use (4.37), (4.38) and a random number generator for the Standard Normal distribution to obtain the replication sample y_1, \dots, y_n . These replication-level data are the counterparts in the simulation experiment of a sample of actual observations in a genuine application.

Diebold and Chen (1996): Step 2

The second step corresponds to restricted (null hypothesis) estimation. The data from Step 1 are used to estimate the first-order autoregression of (4.37) by OLS. Let the OLS estimate of ρ be denoted by $\hat{\rho}$ and the associated residuals by $\hat{u}_t, t = 1, \dots, n$. Note that (4.37) does not include an intercept and so the OLS residuals are not constrained to have a sample mean equal to zero, which is the population mean of an error term. For the purpose of obtaining bootstrap errors by resampling residuals, it is necessary to recentre the latter to have a zero mean, that is, the terms

$$\hat{u}_t^c = \hat{u}_t - \frac{1}{n} \sum_{s=1}^n \hat{u}_s, t = 1, \dots, n,$$

must be calculated.

Estimation of the alternative model for each value of n_1 which satisfies $[0.15n] \leq n_1 \leq [0.85n]$ is also carried out. The test statistics *SupLR*, *SupW* and *SupLM* are then calculated using the relevant results from estimation of null and/or alternative models.

Given the values of $\hat{\rho}$ and $\hat{u}_t^c, t = 1, \dots, n$, from the replication-level data, $B = 1,000$ bootstrap samples, each of size n , are generated. It is, therefore, necessary to repeat Step 3 and Step 4 B times. The former step provides bootstrap data and the latter step uses these artificial observations to derive bootstrap test statistics.

Diebold and Chen (1996): Step 3

For bootstrap sample b , the initial value y_{b0}^* is selected by random sampling from the replication-level data y_1, \dots, y_n and then the n observations $y_{b1}^*, \dots, y_{bn}^*$ are generated using

$$y_{bt}^* = \hat{\rho} y_{bt-1}^* + u_{bt}^*, t = 1, \dots, n, \quad (4.44)$$

in which the bootstrap errors u_{bt}^* are obtained by simple random sampling, with replacement, from the EDF of centred residuals defined by

$$\hat{\mathcal{F}} : \text{probability } \frac{1}{n} \text{ on } \hat{u}_t^c, t = 1, \dots, n. \quad (4.45)$$

The use of (4.45) implies that the bootstrap is not parametric and is appropriate when the CDF of the errors u_t is treated as unknown.

Diebold and Chen (1996): Step 4

The OLS-based estimation and testing procedures applied to replication-level data in Step 2 are now applied to the bootstrap sample. For bootstrap sample b , let the calculated values of the *Sup*-statistics be denoted by $SupLR_b^*$, $SupW_b^*$ and $SupLM_b^*$.

Diebold and Chen (1996): Step 5

The bootstrap p -values of the replication-level statistics $SupLR$, $SupW$ and $SupLM$ from Step 2 can be calculated when Step 3 and Step 4 have been carried out the required B times. However, as pointed out in Diebold and Chen (1996, p. 233), the three bootstrap tests are equivalent; this equivalence reflects the fact that each of LR , W and LM is a monotone increasing function of the F -statistic in (4.34). Consequently, if the common value of the bootstrap p -values for the three test statistics is denoted by \hat{p}_{SUP}^* ,

$$\begin{aligned} \hat{p}_{SUP}^* &= \frac{\sum_{b=1}^B \mathbf{1}(SupLR_b^* \geq SupLR)}{B} \\ &= \frac{\sum_{b=1}^B \mathbf{1}(SupW_b^* \geq SupW)}{B} \\ &= \frac{\sum_{b=1}^B \mathbf{1}(SupLM_b^* \geq SupLM)}{B}, \end{aligned}$$

with $B = 1,000$ in the simulation experiments in Diebold and Chen (1996). The asymptotically valid rejection rule for the bootstrap test with the desired significance level α_d is to reject the null hypothesis of no structural break if $\hat{p}_{SUP}^* \leq \alpha_d$.

When the above steps have been carried out for a complete set of R replications, it is possible to estimate the finite sample null rejection probability of the bootstrap test. The estimate is just the proportion of replications in which $\hat{p}_{SUP}^* \leq \alpha_d$. Diebold and Chen use $R = 1,000$, which is rather smaller than the numbers of replications used to examine the bootstrap tests discussed above, for example, in Chapter 3. The problem is that each replication is relatively costly. In general, with B bootstrap samples used to examine the *Sup*-tests of Andrews (1993), each replication involves the OLS estimation of $(B + 1)[1 + 2A]$ linear regressions, in which A is the number of different breakpoints examined, as implied by the values of the trimming parameters π_1 and π_2 , and the sample size n .

Having examined the estimates from their experiments, Diebold and Chen come to the conclusions that: actual rejection rates are much greater than desired significance levels if the consequences of the data-based estimation of the breakpoint are ignored and the invalid $\chi^2(1)$ distribution is used for critical values; the asymptotically valid critical values given in Andrews (1993) cannot be relied upon to give good control of finite sample significance levels, with this failing being especially marked for the *AsySupLM* test; and the performance of the bootstrap test is very good and consistently better than the corresponding asymptotic theory test.

In order to illustrate the results reported in Diebold and Chen (1996), Table 4.6 contains estimates with $n = 50$ for $\alpha_d = 5$ per cent and 10 per cent. This table is constructed from Tables 2, 3 and 4 in Diebold and Chen (1996), which cover several other cases. The results of Table 4.6 are representative of the full set reported in Diebold and Chen (1996). If the true null rejection probability is ε per cent, the standard error of its estimator, based upon 1,000 replications, is

$$h(\varepsilon) = \left[\sqrt{\frac{\varepsilon(100 - \varepsilon)}{1,000}} \right] \text{ per cent,}$$

with $h(5) \approx 0.7$ per cent and $h(10) \approx 0.9$ per cent. These approximate measures of precision are useful when comparing the estimates of Table 4.6 with the corresponding desired significance levels α_d or a range of values such as $\alpha_d \pm 0.1\alpha_d$; see Serlin (2000).

Table 4.6 Estimated null rejection probabilities of tests for structural break with unknown breakpoint: desired significance levels of 5 per cent and 10 per cent, with sample size = 50

Test	$\rho = 0.01$		$\rho = 0.50$		$\rho = 0.80$		$\rho = 0.99$	
<i>ChiSupW</i>	26.9	46.6	29.0	46.9	33.7	51.8	41.8	61.1
<i>ChiSupLR</i>	25.0	44.9	27.4	44.6	31.4	50.3	39.5	60.4
<i>ChiSupLM</i>	22.9	42.9	25.6	43.1	28.5	48.6	36.8	58.8
<i>AsySupW</i>	1.8	4.3	3.0	5.6	4.1	8.1	7.4	12.4
<i>AsySupLR</i>	1.0	3.1	1.9	4.5	3.3	6.3	5.2	10.3
<i>AsySupLM</i>	0.4	2.3	0.9	3.5	2.2	4.7	3.9	8.0
<i>BootSup</i>	6.6	11.4	5.2	10.3	4.2	10.3	6.1	12.0

Notes: All estimates are in percentage form. The tests *BootSupLR*, *BootSupW* and *BootSupLM* are equivalent and, in order to avoid duplication, are included under the single heading of *BootSup*. Estimates with $\alpha_d = 5$ per cent are given in normal font. Estimates with $\alpha_d = 10$ per cent are given in bold font.

Before examining the simulation results, it is useful to consider the relevant predictions of asymptotic theory in order to have a framework for the discussion. The estimates for *ChiSupW*, *ChiSupLR* and *ChiSupLM* are not expected to be close to α_d because these tests are asymptotically invalid, that is, each has an ERP term that is $O(1)$ and does not tend to zero as $n \rightarrow \infty$. Asymptotic theory leads to the results that *AsySupW*, *AsySupLR* and *AsySupLM* are asymptotically equivalent and have ERP terms that are $o(1)$ so rejection probabilities converge to desired levels (see Andrews, 1993). Thus, if asymptotic theory provided a good approximation, rejection rates would be close to desired significance levels and be similar for *AsySupW*, *AsySupLR* and *AsySupLM*. Finally, since the *Sup*-test statistics are asymptotically pivotal, the results in Beran (1988) suggest that the common bootstrap variant will behave better in finite samples than the asymptotic theory tests.

The general conclusions that can be drawn from the contents of Table 4.6 are the same for both values of the desired significance level. First, Table 4.6 provides strong evidence that using the critical value for a single $\chi^2(1)$ variable when assessing the statistical significance of the maximum of a set of $\chi^2(1)$ variables produces overrejection. Rejection rates for *ChiSupW*, *ChiSupLR* and *ChiSupLM* are all much greater than required, with departures from the desired level depending upon both ρ and the form of the test statistic. The costs of failing to take account of the data-based search for the unknown breakpoint in terms of misleading inferences are very clear.

Second, in contrast to the invalid tests using $\chi^2(1)$ critical values, the asymptotic test using the critical values in Andrews (1993) produces estimates less than the desired value, except when $\rho = 0.99$ and either *AsySupW* or *AsySupLR* is used. The variant *AsySupLM* suffers from the worst departures from the asymptotic significance level. As explained in Diebold and Chen (1996, p. 231), the pattern of inequalities between the rejection frequencies for *AsySupW*, *AsySupLR* and *AsySupLM* is in line with the well-known inequality relationship between the classical statistics for testing a set of linear restrictions; see Godfrey (1988, pp. 57–59) for comments on this relationship in the context of dynamic regression models.

Third, the bootstrap test gives better approximations than the asymptotic theory test and is less sensitive to variations in ρ . Application of the procedure described in Godfrey and Orme (2000a, p. 75) reveals that all estimates for *BootSup* in Table 4.6 are consistent with the claim that the actual null rejection probability is in the range $\alpha_d \pm 0.1\alpha_d$. The quality of this performance is not specific to the cases covered in Table 4.6. On the basis of their much larger complete set of results, Diebold and Chen remark that

the finite-sample size distortion associated with the use of bootstrapped critical values is minimal, regardless of the value of the nuisance parameter ρ (Diebold and Chen, 1996, p. 236).

Other authors have examined the usefulness of bootstrap tests for structural breaks with unknown breakpoints. For example, Christiano recommends a bootstrap approach and shows the importance of the problems caused by ignoring the consequences of pretest use of the data when selecting the breakpoint of the alternative hypothesis; see Christiano (1992). Several other studies are mentioned in Perron (2006, section 8.4), including the important contribution in Hansen (2000).

Hansen points out that the asymptotic theory in Andrews (1993) that underpins the *AsySup*-tests involves the assumption that the regressors are stationary. He generalizes previous analyses by allowing for structural breaks in the marginal distribution of the regressors; so that it is possible to focus on testing for structural breaks in the conditional model for the dependent variable. A “fixed-regressor” bootstrap is proposed in order to make asymptotically valid inferences in this more general framework; see Hansen (2000, section 5). In this version of the bootstrap, the regressors are treated as fixed even when they include lagged values of the dependent variable. Hansen shows that

this bootstrap produces an asymptotically valid test and obtains encouraging results from simulation experiments; see Hansen (2000, section 5.3). The fixed bootstrap test is more reliable than the asymptotic theory test, but “does not completely solve the inference problem”; see Hansen (2000, p. 110).

It might be thought useful to reduce the magnitude of the ERP term by using a double bootstrap test. However, as remarked above, the computational cost of bootstrap tests of *Sup*-statistics is relatively high even for a single bootstrap approach. A conventional double bootstrap of the type discussed in the previous sections of this chapter would have a very high computational cost. The possibility of applying the “Fast Double Bootstrap” (FDB) procedure, which is described in Section 2.5, is explored in Lamarche (2004). Lamarche finds that the estimates of ERP terms for bootstrap tests are small and that the faster bootstrap methods work very well. In contrast, the asymptotic theory test is badly behaved in finite samples. For example, in an experiment in which $\alpha_d = 5$ per cent, the sample size is $n = 20$ and the data are generated by a simple regression with an exogenous regressor, Lamarche obtains rejection frequencies for bootstrap and asymptotic tests equal to 5.2 per cent and 16.4 per cent, respectively. Lamarche also provides an example to demonstrate the feasibility of FDB tests of a *Sup*-statistic in applied work, given existing computer resources; see Lamarche (2004, section 3).

4.5. Summary and conclusions

This chapter has been concerned with situations in which asymptotic theory is at best intractable and, in any case, does not lead to one of the standard (tabulated) distributions as the source of critical values or *p*-values. It has been argued that the bootstrap can be especially useful when the standard asymptotic results for test statistics are not applicable. If theorists have not derived the relevant non-standard asymptotic null distribution, the bootstrap offers a way to carry out an asymptotically valid test when no theory-based method is available. When the non-standard asymptotic distribution has been derived, asymptotic critical values are usually approximated using simulation methods. However, such simulated asymptotic critical values may give poorer finite sample approximations than a bootstrap test. Moreover, applied workers may be interested in calculating estimates of *p*-values, rather than just using one of the three conventional significance levels. Published tables of asymptotic critical values are of limited value when the aim is to approximate *p*-values, but the bootstrap allows straightforward estimation.

There are many examples of non-standard tests that are important in applied econometric analysis. Three examples have been used in this chapter to illustrate the usefulness of bootstrap methods when test statistics have non-standard asymptotic null distributions. In the first example, the predictive test given in Chow (1960) was examined. This test is widely used, with critical values being taken from an F -distribution as proposed in Chow (1960) for the case of NID errors. However, if the errors are simply assumed to be IID, the predictive test statistic is not asymptotically pivotal. Instead its asymptotic null distribution depends upon the distribution of the IID errors, even as the sample size tends to infinity.

A single bootstrap version of the predictive test gives an asymptotically valid procedure and a double bootstrap provides an asymptotic refinement, as defined in Beran (1988). The predictions of the asymptotic analysis in Beran (1988) are supported by the results of simulation experiments, which are discussed in Section 4.2.3. Reliance on the F -distribution for critical values under non-Normality leads to rejection rates that are far from the desired significance levels. The use of a single bootstrap with the predictive test statistic gives quite a good approximation and a double bootstrap provides even better control of finite sample significance levels.

In the second example, the problem of controlling the overall significance level of a group of separate diagnostic checks was discussed. This is an important problem which is faced by any researcher who uses a standard program for the OLS estimation of a regression model. The limitations of conventional asymptotic theory were explained and applications of bootstrap methods were discussed. The basic idea is to use bootstrap techniques to judge the statistical significance of the minimum of the estimated p -values of the various individual diagnostic test statistics. Evidence from simulation experiments in Godfrey (2005) indicates the potential practical value of two asymptotically valid bootstrap approaches. The first of these procedures uses bootstrap estimates of individual p -values in order to derive their minimum. In the second approach, the minimum p -value is calculated from individual p -values that are obtained from known asymptotic distributions. Clearly the former method is more widely applicable than the latter because some test statistics are not asymptotically pivotal, for example, the predictive test of the first example. The former method may also be more useful than the latter when asymptotic distributions are available for all the tests being used, but there is evidence that suggests that such distributions can provide inaccurate approximations for some of these checks in finite samples.

The third and final example used in this chapter, like the first, is based upon a generalization of a test proposed in Chow (1960). The relevant procedure is Chow's test of the null hypothesis that the regression coefficient vector is the same for all observations, with an untested auxiliary assumption being that the errors are $NID(0, \sigma^2)$. (The auxiliary assumption of Normality can be relaxed if Chow's F -test is treated as an approximation based upon conventional asymptotic theory.) The alternative hypothesis in Chow's test can be viewed as consisting of two parts: (i) there are two different values of the regression coefficient vector; and (ii) it is known which of these two values applies to each observation of the complete set.

Many researchers have argued that part (ii) of Chow's alternative hypothesis is implausible and not representative of actual applications in which data are often used to determine the partitioning of the total sample that provides the strongest evidence against the null hypothesis. Unfortunately standard asymptotic theory cannot be applied when the data have been used in this way. Non-standard asymptotic theory is, however, available in the results of Andrews (1993). Given the relevant non-standard asymptotic null distribution, simulation can be used, as in Andrews (2003a), to approximate critical values for conventional levels of significance. The problem for applied work is that such asymptotic critical values may not be accurate in finite sample situations.

Simulation evidence was summarized. This evidence indicates that asymptotic critical values from the non-standard distributions in Andrews (1993) can be unreliable, with rejection frequencies that are not close to desired significance levels. The corresponding single bootstrap test does better in terms of the results from simulation experiments and is found to perform well. Results in Lamarche (2004) indicate that, despite the relatively high computational costs implied by the nature of the search-based test, a "Fast Double Bootstrap" test is feasible and is likely to give even more accurate results.

It is clear that the same general findings emerge from each of the three examples discussed in this chapter. Bootstrap tests can be useful and reliable when asymptotic theory fails to provide a standard distribution as an approximation to finite sample behaviour. In all three examples, simulation evidence supports the claim that bootstrap tests perform well for sample sizes of interest to applied workers. In contrast, the non-standard asymptotic critical values are either unavailable or have not been found to be generally reliable in simulation experiments.

When the findings of this chapter are combined with those of the previous chapter, it is evident that there is strong support for the routine use

of the bootstrap, whether or not standard asymptotic theory is available for the tests of interest. The generality of these encouraging results on bootstrap tests is, of course, limited by the assumption that the errors of the regression model are IID. Modern econometrics texts often contain the recommendation that inference should be based upon procedures that are asymptotically robust to non-Normality and heteroskedasticity. Indeed, in some cases, it is possible to derive tests that are asymptotically valid in the presence of unspecified forms of autocorrelation, heteroskedasticity and non-Normality. However, when either autocorrelation or heteroskedasticity of the errors is permitted, the generation of bootstrap errors by simple random sampling, with replacement, from the OLS residuals is inappropriate. Alternative bootstrap schemes are required in order to mimic the features of the assumed actual data generation process and so to lead to correct asymptotic distributions of test statistics. Some of these more general bootstrap schemes, which allow for autocorrelation and/or heteroskedasticity of the regression errors, are discussed in the next chapter.

5

Bootstrap Methods for Regression Models with Non-IID Errors

5.1. Introduction

After examining evidence from asymptotic analyses and simulation experiments, MacKinnon comments that he

would be very surprised to encounter a bootstrap test that did not work well in the context of a single-equation regression model . . . , provided the regressors are exogenous or predetermined and the underlying error terms are homoskedastic and serially uncorrelated (MacKinnon, 2002, p. 625).

Since the publication of MacKinnon's remarks, other researchers have come to the conclusion that bootstrap tests, derived by resampling residuals, should be used when the null hypothesis model is a linear regression with IID errors, whether or not standard asymptotic tests are available. However, the assumption that the errors are IID is restrictive and it is often thought necessary to allow for heteroskedasticity and/or autocorrelation in applied regression analysis.

It has been emphasized in the previous chapters that the artificial model used to obtain repeated samples of bootstrap data should mimic the model assumed to generate the actual data. Consequently, if the actual data are assumed to be produced by a regression model with non-IID errors, a bootstrap model that imposes IID errors will not be an accurate approximation and cannot be expected to lead to appropriate tests, even asymptotically. There is, therefore, a need for more general versions of bootstrap error processes, which permit the required departures from the framework of IID disturbances.

The purpose of this chapter is to outline important results for bootstrap schemes that are relevant to regression models with non-IID errors. Whether heteroskedasticity or autocorrelation of the errors is to be permitted, there is a choice between two general types of bootstrap test. The first type consists of *model-based bootstrap* procedures that are derived by assuming a specific error model to replace the assumption that the errors are IID, for example, a simple AR(1) scheme might be used to allow for autocorrelation. The specification of an error model is not required for the second type of bootstrap test. Instead tests are derived using bootstrap schemes that are designed to be asymptotically valid in the presence of unspecified forms of departure from either homoskedasticity or independence.

It seems reasonable to conjecture that the first type of test will be more efficient than the second type of procedure when the former is based upon the correct parametric form for the non-IID error model. However, the asymptotic validity of the first type of test will depend upon the adequacy of the assumed error model from which it is derived. Consequently the first type of test is less robust than the second type. The choice between the two approaches to deriving bootstrap tests, therefore, involves consideration of the trade-off between efficiency and robustness; see Liu and Singh (1992) for an analysis of efficiency and robustness in resampling. Both general types of bootstrap test are discussed below.

The contents of this chapter are as follows. In Section 5.2, bootstrap procedures that allow for heteroskedasticity of independently distributed errors are examined. Bootstraps for stationary autocorrelated processes are discussed in Section 5.3. The implementation of bootstrap schemes for generating pseudo-data that mimic both heteroskedasticity and autocorrelation is covered in Section 5.4. Finally, Section 5.5 contains a summary and some concluding remarks.

5.2. Bootstrap methods for independent heteroskedastic errors

The primary aim of regression analysis is to gain information about the conditional mean function for a dependent variable, given the values of regressors. As stressed, for example, in Hansen (1999), correct specification of the conditional mean function does not automatically imply that the conditional variances are all equal. Consequently it may often be reasonable to allow for heteroskedasticity. In order to simplify the exposition of bootstrap methods, discussion of dynamic models is postponed until the main ideas relating to *heteroskedasticity-valid bootstraps*

have been covered. It is, therefore, assumed initially that all regressors are either nonrandom or strictly exogenous.

The heteroskedastic regression model is written as

$$y_i = \sum_{j=1}^k x_{ij}\beta_j + u_i, i = 1, \dots, n, \tag{5.1}$$

in which, conditional upon the values of the regressors, the error terms are independently distributed with common zero mean but different variances. The conditional variances are denoted by $(\sigma_1^2, \dots, \sigma_n^2)$ and heteroskedasticity is present, so $\sigma_i^2 \neq \sigma_j^2$ for at least one pair of values i and j . Standard asymptotic analysis for heteroskedastic regression models uses the assumption that u_i^2 , like u_i , has a finite positive variance for all i ; so that the first four moments of the conditional distribution of each error are assumed to be finite. It is also assumed that the regressors satisfy assumptions required for the consistency and asymptotic Normality of the OLS estimators for (5.1); see, for example, White (1980).

As usual, the OLS estimators of the regression coefficients in (5.1) are denoted by $\hat{\beta}_j, j = 1, \dots, k$. The consistency of these OLS estimators implies that the former can play the role of the latter when constructing a bootstrap counterpart of the regression model which is assumed to generate the actual data. Hence, for a typical observation y_i^* of a bootstrap sample, the conditional mean could be

$$E^*(y_i^* | x_{i1}, \dots, x_{ik}) = \sum_{j=1}^k x_{ij}\hat{\beta}_j,$$

with the addition of a bootstrap error term u_i^* yielding

$$y_i^* = E^*(y_i^* | x_{i1}, \dots, x_{ik}) + u_i^* = \sum_{j=1}^k x_{ij}\hat{\beta}_j + u_i^*.$$

The classical residual resampling bootstrap proposed in previous chapters as a source for the error term u_i^* is no longer appropriate and does not lead to asymptotically valid tests in the presence of heteroskedasticity.

The failings of the IID-valid residual resampling method can be illustrated by considering a simple regression model with heteroskedastic errors. Suppose that

$$y_i = \alpha + \beta x_i + u_i, i = 1, \dots, n, \tag{5.2}$$

in which regressor values are not random and the heteroskedastic errors are independently distributed with

$$E(u_i) = 0 \text{ and } \text{Var}(u_i) = \sigma_i^2, i = 1, \dots, n.$$

Let the OLS estimators for (5.2) be denoted by $\hat{\alpha}$ and $\hat{\beta}$, with the associated residuals being $\hat{u}_i, i = 1, \dots, n$. As is well known, the OLS estimator of the slope coefficient satisfies

$$\hat{\beta} = \beta + \frac{\sum_i (x_i - \bar{x}) u_i}{\sum_i (x_i - \bar{x})^2}, \quad (5.3)$$

and, in the presence of heteroskedasticity, the variance of the $O_p(1)$ term $\sqrt{n}(\hat{\beta} - \beta)$ is

$$\text{Var}(\sqrt{n}(\hat{\beta} - \beta)) = \frac{n \sum_i (x_i - \bar{x})^2 \sigma_i^2}{[\sum_i (x_i - \bar{x})^2]^2}. \quad (5.4)$$

If, as part of an attempt to improve on approximations based upon asymptotic theory, repeated artificial samples were obtained using classical residual resampling, the bootstrap data would be generated by

$$y_i^* = \hat{\alpha} + \hat{\beta} x_i + u_i^*, i = 1, \dots, n, \quad (5.5)$$

with the bootstrap errors being drawn randomly, with replacement, from the EDF in

$$\hat{\mathcal{F}} : \text{probability } \frac{1}{n} \text{ on } \hat{u}_i, i = 1, \dots, n. \quad (5.6)$$

The implied bootstrap world counterpart of (5.3) is

$$\hat{\beta}^* = \hat{\beta} + \frac{\sum_i (x_i - \bar{x}) u_i^*}{\sum_i (x_i - \bar{x})^2}. \quad (5.7)$$

Now, under the classical resampling approach, the terms u_i^* are IID with common variance given by

$$\begin{aligned} \text{Var}^*(u^*) &= E^*[(u^*)^2] \\ &= \sum_{i=1}^n \Pr(u^* = \hat{u}_i) \hat{u}_i^2 \\ &= \frac{1}{n} \sum_{i=1}^n \hat{u}_i^2. \end{aligned}$$

It follows that, under the conditional (on observed data) bootstrap probability law based upon the false assumption of IID errors, the counterpart of (5.4) is

$$\text{Var} \left(\sqrt{n} (\hat{\beta}^* - \hat{\beta}) \right) = \frac{\sum_{i=1}^n \hat{u}_i^2}{\sum_i (x_i - \bar{x})^2},$$

which is asymptotically equivalent to

$$\frac{\sum_{i=1}^n \sigma_i^2}{\sum_i (x_i - \bar{x})^2}. \tag{5.8}$$

As $n \rightarrow \infty$, the term in (5.8), which is appropriate for IID bootstrap errors, does not have the same limit as the variance in (5.4), which is derived allowing for heteroskedasticity of actual errors. Consequently the use of an IID-based bootstrap implies that $\sqrt{n} (\hat{\beta}^* - \hat{\beta})$ and $\sqrt{n} (\hat{\beta} - \beta)$ do not have the same asymptotic variance when the errors u_i , $i = 1, \dots, n$, are heteroskedastic.

More generally, the classical (IID-valid) residual resampling scheme based upon (5.6), or some asymptotically equivalent scheme, does not lead to a consistent estimator of the covariance matrix of $\hat{\beta} = (\hat{\beta}_1, \dots, \hat{\beta}_k)'$ when (5.1) is estimated by OLS and so will not lead to valid large sample tests. Four different approaches to bootstrapping that do have the potential to yield valid asymptotic tests in the presence of heteroskedasticity will be now be examined.

5.2.1. Model-based bootstraps

Suppose first that a researcher is prepared to specify a model that determines variances of the dependent variable, as well as one for its mean values, given exogenous variables. In such a situation, the regression model (5.1) is combined with a *skedastic function*. The skedastic function often takes the form of a transformation of a linear combination of exogenous variables, which may or may not be related to the regressors of (5.1). Let

$$\mathbf{x}_i = (x_{i1}, \dots, x_{ik})' \text{ and } \mathbf{z}_i = (1, z_{i1}, \dots, z_{iq})'$$

denote typical observation vectors on exogenous variables taken to be relevant to mean and variance functions, respectively, with associated coefficient vectors

$$\boldsymbol{\beta} = (\beta_1, \dots, \beta_k)' \text{ and } \boldsymbol{\gamma} = (\gamma_0, \gamma_1, \dots, \gamma_q)'.$$

The assumed data generation process (DGP) can then be written as

$$y_i = \mathbf{x}'_i \boldsymbol{\beta} + u_i, i = 1, \dots, n, \quad (5.9)$$

in which the errors are independent, have zero mean and variances given by the skedastic function

$$\sigma_i^2 = v(\mathbf{z}'_i \boldsymbol{\gamma}), i = 1, \dots, n, \quad (5.10)$$

for some specified function $v(\cdot)$.

The literature on heteroskedastic regression models includes discussion of various forms of $v(\cdot)$. The multiplicative model is obtained by using $v(\cdot) = \exp(\cdot)$, that is,

$$\sigma_i^2 = \exp(\mathbf{z}'_i \boldsymbol{\gamma}), i = 1, \dots, n, \quad (5.11)$$

which is discussed in detail in Harvey (1976). The additive model

$$\sigma_i^2 = \mathbf{z}'_i \boldsymbol{\gamma}, i = 1, \dots, n, \quad (5.12)$$

is considered in Amemiya (1977). For both specifications, it is possible to obtain a consistent estimator of $\boldsymbol{\gamma}$ by applying OLS to an artificial regression in which the observations on the dependent variable are derived from the OLS residuals $\hat{u}_i = y_i - \mathbf{x}'_i \hat{\boldsymbol{\beta}}$, $i = 1, \dots, n$. For the multiplicative scheme (5.11), the artificial regression is

$$\log(\hat{u}_i^2) = \mathbf{z}'_i \boldsymbol{\gamma} + \text{error};$$

see Harvey (1976, p. 462). The corresponding artificial regression when the skedastic model is (5.12) can be written as

$$\hat{u}_i^2 = \mathbf{z}'_i \boldsymbol{\gamma} + \text{error};$$

see Amemiya (1977, pp. 366–367).

Given a consistent estimator of $\boldsymbol{\gamma}$, derived from OLS estimation of a suitable artificial regression, it is possible to use a feasible GLS method to re-estimate $\boldsymbol{\beta}$. Alternatively, some researchers have assumed Normality in order to make use of maximum likelihood techniques to estimate $\boldsymbol{\beta}$ and $\boldsymbol{\gamma}$; see Amemiya (1977, p. 368) and Harvey (1976, section 3).

For the purpose of outlining model-based bootstrap schemes for heteroskedastic regression models, let the estimators of $\boldsymbol{\beta}$ and $\boldsymbol{\gamma}$ be denoted

by $\ddot{\beta}$ and $\ddot{\gamma}$, respectively. Under standard assumptions, the methods mentioned above yield consistent estimators for the parameter vectors of (5.9) and (5.10). These consistent estimators can be used to define a bootstrap DGP. Let a sequence of scaled residuals be defined as follows:

$$\ddot{e}_i = (y_i - \mathbf{x}'_i \ddot{\beta}) / \sqrt{v(\mathbf{z}'_i \ddot{\gamma})}, i = 1, \dots, n. \tag{5.13}$$

A model-based bootstrap that reflects heteroskedasticity can be implemented using

$$y_i^* = \mathbf{x}'_i \ddot{\beta} + \left(\sqrt{v(\mathbf{z}'_i \ddot{\gamma})} \right) e_i^*, i = 1, \dots, n, \tag{5.14}$$

in which the terms e_1^*, \dots, e_n^* are obtained by random sampling, with replacement, from

$$\ddot{F}_e : \text{probability } \frac{1}{n} \text{ on } \left(\ddot{e}_i - \frac{1}{n} \sum_{j=1}^n \ddot{e}_j \right), i = 1, \dots, n. \tag{5.15}$$

There are, however, good reasons to question the practical value of bootstrapping using schemes of the type defined by (5.13) to (5.15). These schemes require the correct specification of the skedastic model (5.10) for their consistency. In applied work, there is not likely to be much information to guide the specification of (5.10). If the wrong functional form is selected for $v(\cdot)$, or \mathbf{z}_i does not include all relevant variables, the true pattern of heteroskedasticity will not be approached in the bootstrap world as $n \rightarrow \infty$. Consequently the model-based bootstrap approach must probably be judged to provide tools that are too fragile for general use. Its lack of robustness to misspecification of the variance model could lead to very misleading inferences; see Belsley (2002) for an analysis of the effects of using incorrect skedastic functions. Bootstraps that do not require the specification of variance models in order to supply heteroskedasticity-robust asymptotic procedures are, therefore, of interest.

5.2.2. Pairs bootstraps

The usual regression equation framework is not used in the second approach to bootstrapping in the presence of heteroskedasticity. Instead the bootstrap samples are obtained by resampling the *pairs* of observations y_i and \mathbf{x}_i . More formally, the bootstrap sample $\mathbf{S}^* = \{(y_i^*, \mathbf{x}_i^*), i = 1, \dots, n\}$ is derived by simple random sampling, with replacement, from

the joint EDF

$$\check{G} : \text{probability } \frac{1}{n} \text{ on } (y_i, \mathbf{x}_i), i = 1, \dots, n, \quad (5.16)$$

which is equivalent to

$$\check{G} : \text{probability } \frac{1}{n} \text{ on } (\mathbf{x}_i' \hat{\beta} + \hat{u}_i, \mathbf{x}_i), i = 1, \dots, n. \quad (5.17)$$

This pairs bootstrap is useful when the vectors (y_i, \mathbf{x}_i') are viewed as IID drawings from some joint distribution. As pointed out in Davidson and MacKinnon (2006, p. 822), this interpretation does not rule out heteroskedasticity in the conditional distributions of y_i , given \mathbf{x}_i , for $i = 1, \dots, n$.

Having obtained a pairs bootstrap sample $\mathbf{S}^* = \{(y_i^*, \mathbf{x}_i^*), i = 1, \dots, n\}$, the associated bootstrap OLS estimator, denoted by $\hat{\beta}^*$, can be calculated. In order for the pairs bootstrap to be a source of asymptotically valid inferences, $\sqrt{n}(\hat{\beta} - \beta)$ and $\sqrt{n}(\hat{\beta}^* - \hat{\beta})$ should, under their respective probability laws, have the same limiting distribution. Asymptotic analyses that deal with the consistency of the pairs bootstrap are contained in Freedman (1981) and Mammen (1993). However, while the consistency of the pairs bootstrap indicates its potential usefulness when the errors of a regression model are heteroskedastic, several authors have drawn attention to drawbacks.

First, there may be an interest in inference conditional upon the observed values of the exogenous regressors and the pairs bootstrap uses random sets of observations on regressors which may not stay close to the actual set; see Hinkley (1988, p. 331). Second, as pointed out in Horowitz (2001, p. 3215), randomly resampling the pairs (y_i, \mathbf{x}_i) does not impose a condition of the form $E(u_i | \mathbf{x}_i) = 0$. (Freedman remarks that “perhaps surprisingly, the condition . . . does not seem to be needed”; see Freedman (1981, p. 1220). It is shown in Mammen (1993, p. 256) that $E(u_i | \mathbf{x}_i) = \mathbf{0}_k$ is implied by the standard assumptions.) Third, modification of either the null hypothesis or the pairs bootstrap is required when statistics for testing linear restrictions are to be bootstrapped; see Davidson and MacKinnon (2006, p. 822). Given that the purpose of this book is to describe tests for regression models, it is worth commenting on the third point in a little more detail.

If the original pairs bootstrap is used, the null hypothesis to be tested using bootstrap samples must be modified to reflect what is true for the actual data. For example, rather than test the actual null hypothesis of

$\mathbf{R}\boldsymbol{\beta} = \mathbf{r}$, the modification is to examine $\mathbf{R}\boldsymbol{\beta} = \mathbf{R}\hat{\boldsymbol{\beta}}$ in the pairs bootstrap world. An alternative modification, which does not require adjustment of the null hypothesis, is to modify the pairs bootstrap. If the restricted estimator for the actual data is denoted by $\hat{\boldsymbol{\beta}}$, so that $\mathbf{R}\hat{\boldsymbol{\beta}} = \mathbf{r}$, the modified pairs bootstrap is to derive the sample $\mathbf{S}^* = \{(y_i^*, \mathbf{x}_i^*), i = 1, \dots, n\}$ by random sampling, with replacement, from

$$\check{G}_{\text{mod}} : \text{probability } \frac{1}{n} \text{ on } (\mathbf{x}'_i \tilde{\boldsymbol{\beta}} + \hat{u}_i, \mathbf{x}_i), i = 1, \dots, n; \tag{5.18}$$

see Flachaire (1999) and Mammen (1993, section 4). Flachaire reports evidence from small-scale simulation experiments that indicates that replacing (5.17) by (5.18) leads to more reliable tests. Mammen establishes that, if (5.18) is used in a modified pairs bootstrap, a bootstrap test of the standard F -statistic is asymptotically valid, despite the fact that the presence of unspecified forms of heteroskedasticity implies that this statistic is not asymptotically pivotal.

Given that the pairs bootstrap suffers from the three drawbacks described above, it is not surprising that there has been interest in alternative ways of bootstrapping regression statistics when unspecified types of heteroskedasticity are permitted. One popular approach, known as the *wild bootstrap*, will now be discussed.

5.2.3. Wild bootstraps

Wild bootstraps are more closely linked to the textbook discussions of regression models than the pairs bootstraps. A wild bootstrap involves adding together an estimated predicted part, which serves as a bootstrap world conditional mean, and a bootstrap error term. The key thing is that the bootstrap error should be obtained in a way that allows for heteroskedasticity of unknown form. A typical observation for a wild bootstrap scheme can be written as

$$y_i^* = \mathbf{x}'_i \hat{\boldsymbol{\beta}} + \ddot{u}_i \epsilon_i, \tag{5.19}$$

in which $\hat{\boldsymbol{\beta}}$ is an estimator of $\boldsymbol{\beta}$, \ddot{u}_i is a residual associated with a (possibly) different estimator $\check{\boldsymbol{\beta}}$, that is, $\ddot{u}_i = y_i - \mathbf{x}'_i \check{\boldsymbol{\beta}}$, and ϵ_i is a drawing from a *pick distribution*. (As usual, the residual \ddot{u}_i in (5.19) can be multiplied by an adjustment term that tends to 1, as $n \rightarrow \infty$, and reflects consideration of either leverage values or degrees-of-freedom relative to the sample size.) Conditional upon the observed data, the terms \mathbf{x}_i , $\hat{\boldsymbol{\beta}}$ and \ddot{u}_i in (5.19) are constants and the randomness of y_i^* about its conditional mean depends

upon the specification of the pick distribution. There is, therefore, no need to centre the actual residuals to have zero mean in order for the bootstrap-world expectation $E^*(\ddot{u}_i \epsilon_i) = \ddot{u}_i E^*(\epsilon_i)$ to be zero.

The general theory of the wild bootstrap is discussed in the statistics literature; see Liu (1988), Mammen (1993) and Wu (1986). In order to serve as basis for asymptotically valid (and hopefully reliable finite sample) tests, (5.19) should use estimators $\hat{\beta}$ and $\check{\beta}$ that are consistent under the relevant null hypothesis, with $\check{\beta}$ being a restricted estimator that satisfies the null hypothesis. The estimator $\hat{\beta}$, which provides the residual term \ddot{u}_i , can come from either restricted or unrestricted estimation. The pick distribution of (5.19) must also satisfy conditions for the consistency of the wild bootstrap. At a minimum, the terms ϵ_i should be IID random variables with mean $E^*(\epsilon_i) = 0$ and variance $E^*(\epsilon_i^2) = 1$, $i = 1, \dots, n$; see Wu (1986, section 7). A further restriction is sometimes imposed on the pick distribution. The distribution of a single linear combination of the elements of the OLS estimator $\hat{\beta}$ is considered in Liu (1988). Liu shows that, if $E^*(\epsilon_i^3) = 0$, the wild bootstrap enjoys second-order properties, with the first three moments of the relevant test statistic being estimated correctly to $O(n^{-1})$ by the wild bootstrap. Many different pick distributions are available and clearly evidence on their small sample properties is of interest to applied workers.

One obvious device for constructing a pick distribution from sample information is to calculate standardized versions \ddot{u}_i^Z of the residuals \ddot{u}_i , with

$$\ddot{u}_i^Z = a + b\ddot{u}_i, i = 1, \dots, n,$$

the constants a and b being chosen so that

$$\frac{1}{n} \sum_{i=1}^n \ddot{u}_i^Z = 0 \text{ and } \frac{1}{n} \sum_{i=1}^n (\ddot{u}_i^Z)^2 = 1,$$

and then to define the pick distribution by

$$\mathcal{D}_{\epsilon,1} : \Pr(\epsilon = \ddot{u}_i^Z) = \frac{1}{n}, i = 1, \dots, n. \quad (5.20)$$

When the terms ϵ_i are drawn randomly, with replacement, from $\mathcal{D}_{\epsilon,1}$, the two conditions given by Wu are satisfied but the additional condition discussed by Liu is not.

A simple alternative to the data-based approach that leads to $\mathcal{D}_{\epsilon,1}$ is to assume that the terms $\epsilon_i, i = 1, \dots, n$, are $NID(0, 1)$. This assumption gives

$$\mathcal{D}_{\epsilon,2} : \epsilon \sim N(0, 1). \tag{5.21}$$

As with $\mathcal{D}_{\epsilon,1}$, the first two moments for ϵ_i from $\mathcal{D}_{\epsilon,2}$ are $E^*(\epsilon_i) = 0$ and $E^*(\epsilon_i^2) = 1$. The symmetry of the standard Normal distribution implies that $E^*(\epsilon_i^3) = 0$; so that Liu's additional moment condition is not fulfilled by $\mathcal{D}_{\epsilon,2}$.

A pick distribution that satisfies all three moment conditions on the distribution of ϵ_i is proposed in Liu (1988). For this pick distribution, denoted by $\mathcal{D}_{\epsilon,3}$, ϵ is defined to be the mean-adjusted product of two Normal variables. More precisely, Liu's distribution $\mathcal{D}_{\epsilon,3}$ is given by

$$\mathcal{D}_{\epsilon,3} : \epsilon = df - E(d)E(f), \tag{5.22}$$

in which d and f are independent Normal variables with

$$d \sim N\left(\frac{1}{2}\left(\sqrt{\frac{17}{6}} + \sqrt{\frac{1}{6}}\right), \frac{1}{2}\right),$$

and

$$f \sim N\left(\frac{1}{2}\left(\sqrt{\frac{17}{6}} - \sqrt{\frac{1}{6}}\right), \frac{1}{2}\right).$$

An alternative to $\mathcal{D}_{\epsilon,3}$, which has the same first three moments, is given in Mammen (1993). This alternative pick distribution is denoted by $\mathcal{D}_{\epsilon,4}$ and is probably the most widely-used of the available specifications. It is based upon a discrete pick distribution with only two possible values, in which

$$\begin{aligned} \mathcal{D}_{\epsilon,4} : \Pr\left(\epsilon = \frac{-(\sqrt{5}-1)}{2}\right) &= \frac{(\sqrt{5}+1)}{2\sqrt{5}} \text{ and} \\ \Pr\left(\epsilon = \frac{(\sqrt{5}+1)}{2}\right) &= 1 - \frac{(\sqrt{5}+1)}{2\sqrt{5}}. \end{aligned} \tag{5.23}$$

Mammen also proposes a quadratic function of a standard Normal variable that gives the required values for the first three moments of ϵ_i in

(5.19). More precisely, Mammen defines

$$\mathcal{D}_{\epsilon,5} : \epsilon = \frac{z}{\sqrt{2}} - \frac{(z^2 - 1)}{2}, \quad (5.24)$$

in which $z \sim NID(0, 1)$.

A pick distribution which is much simpler than any of those already mentioned has been recommended in Davidson and Flachaire (2001, 2008). The scheme supported by Davidson and Flachaire is the *Rademacher distribution* given by

$$\mathcal{D}_{\epsilon,6} : \Pr(\epsilon = 1) = \Pr(\epsilon = -1) = \frac{1}{2}. \quad (5.25)$$

For this distribution, $E^*(\epsilon_i) = 0$, $E^*(\epsilon_i^2) = 1$ and $E^*(\epsilon_i^3) = 0$. At first glance, this pick distribution might seem to have very little to recommend it. It generates bootstrap errors from heteroskedastic two-point distributions, with

$$\Pr(u_i^* = \ddot{u}_i) = \Pr(u_i^* = -\ddot{u}_i) = 0.5,$$

so that $\text{Var}^*(u_i^*) = \ddot{u}_i^2$, $i = 1, \dots, n$. Few researchers would suggest such a simple specification for approximating the distributions of the actual regression errors. However, despite the simplicity of the Rademacher distribution, it appears to perform very well in many applications and some examples that illustrate its usefulness will be provided in the next chapter. Results from simulation experiments that are discussed in Flachaire (2005) suggest that Rademacher pick distribution $\mathcal{D}_{\epsilon,6}$ gives better performance than either pairs bootstraps or the widely-used wild bootstrap derived from Mammen's distribution $\mathcal{D}_{\epsilon,4}$.

5.2.4. Estimating function bootstraps

The problem of bootstrapping the OLS estimator of the coefficients of a linear regression model when the errors are heteroskedastic is examined in Hu and Zidek (1995). Hu and Zidek put forward an alternative to pairs and wild bootstraps. Their method is obtained from the estimating functions (often called the "normal equations") for the unrestricted OLS estimator $\hat{\beta}$. These estimating functions can be written as

$$\sum_{i=1}^n \mathbf{x}_i \hat{u}_i = \mathbf{0}_k,$$

and imply that the OLS estimator is given by

$$\hat{\beta} = \beta + \left(\sum_{i=1}^n \mathbf{x}_i \mathbf{x}_i' \right)^{-1} \left(\sum_{i=1}^n \mathbf{x}_i u_i \right).$$

Hu and Zidek suggest that $\mathbf{r}_i = \mathbf{x}_i \hat{u}_i$ be regarded as an estimate of $\rho_i = \mathbf{x}_i u_i, i = 1, \dots, n$. The *estimating function bootstrap* is then carried out by calculating

$$\hat{\beta}^* = \hat{\beta} + \left(\sum_{i=1}^n \mathbf{x}_i \mathbf{x}_i' \right)^{-1} \left(\sum_{i=1}^n \mathbf{r}_i^* \right), \tag{5.26}$$

in which $\{\mathbf{r}_1^*, \dots, \mathbf{r}_n^*\}$ is a bootstrap sample obtained by random sampling, with replacement, from

$$\mathcal{F}_T : \Pr(\mathbf{r}^* = \mathbf{x}_i \hat{u}_i) = \frac{1}{n}, i = 1, \dots, n.$$

Asymptotic analysis to justify the use of the estimating function bootstrap is provided in Hu and Zidek (1995). It is shown that, under regularity conditions, this bootstrap technique yields the correct asymptotic distribution. Hu and Zidek report encouraging results about small sample performance from simulation experiments. Hu and Zidek also compare the estimating function bootstrap with pairs and wild bootstraps; see Hu and Zidek (1995, section 3). Since the matrix inverse on the right-hand-side of (5.26) need only be computed once, the estimating function bootstrap has a smaller computational cost than the pairs bootstrap method, which involves OLS estimation for each of the regressor observations sets selected. Also, in contrast to the wild bootstrap, there is no need to make an assumption about the distribution of an auxiliary random variable such as ϵ_i in (5.19). While the choice of pick distribution does not affect the asymptotic validity of wild bootstrap tests, provided that the mean is 0 and the variance is 1, finite sample behaviour may be sensitive. Consequently there may be a lack of robustness of wild bootstrap tests to variations in the choice of pick distribution for sample sizes of relevance in applied work. Simulation results on this issue are provided in the next chapter.

It should be noted that the usefulness of the estimating function bootstrap is not confined to the linear regression model with heteroskedastic errors. Hu and Kalbfleisch provide a wide ranging discussion of this technique in which several different models are covered and argue that it is

often superior, in terms of accuracy and computational cost, to standard bootstrap methods; see Hu and Kalbfleisch (2000). There is, however, an important restriction on the availability of the estimating function bootstrap; the data must be independently distributed.

5.2.5. Bootstrapping dynamic regression models

It has been assumed so far that the regressors are all exogenous. This assumption may be appropriate when the regression analysis is based upon cross-section data and many textbooks include the remark that heteroskedasticity is often present when such data are employed; see, for example, Verbeek (2004, pp. 82–83). However, heteroskedasticity can also arise in time-series applications. When time-series regression models are under consideration, the assumption that all regressors are strictly exogenous must sometimes be relaxed because lagged values of the dependent variable are included in the regressor set. There is, therefore, a need for a discussion of bootstrap methods that can be used for dynamic regression models with heteroskedastic errors. It is convenient to use the subscript t , rather than i , for observations, given that it is being assumed that time series data are being used for the regression analysis.

In order to focus on the consequences of the regression equation being dynamic, suppose first that the DGP is the AR(p) model

$$y_t = \alpha_1 y_{t-1} + \cdots + \alpha_p y_{t-p} + u_t, t = 1, \dots, n, \quad (5.27)$$

in which: the integer p is finite, known and positive; and the elements of $\alpha = (\alpha_1, \dots, \alpha_p)'$ are such that $\alpha_p \neq 0$, with the equation

$$\alpha(\lambda) = \lambda^p - \alpha_1 \lambda^{p-1} - \cdots - \alpha_p = 0,$$

having all of its roots inside the unit circle. The OLS estimator for (5.27) is denoted by $\hat{\alpha} = (\hat{\alpha}_1, \dots, \hat{\alpha}_p)'$. In the standard asymptotic analysis of the behaviour of $\hat{\alpha}$, it is assumed that the errors u_t in (5.27) are IID, with zero mean, variance σ^2 and finite fourth moment; see, for example, Hamilton (1994, pp. 215–217). The generalization of these assumptions to allow for conditional heteroskedasticity is discussed in Gonçalves and Kilian (2004).

The weaker assumptions that are made by Gonçalves and Kilian include the requirement that the conditional mean of u_t , given the history of past errors, is zero, so that there is no serial correlation; see Gonçalves and Kilian (2004, p. 94) for detailed statements of the regularity conditions needed for the technical analysis. These assumptions

permit the presence of conditional heteroskedasticity schemes that are of interest in empirical research, for example, ARCH and GARCH processes. Consequently the serially uncorrelated errors can be statistically dependent, as in Horowitz et al. (2006).

Gonçaves and Kilian show that, under their assumptions, $\sqrt{n}(\hat{\alpha} - \alpha)$ is asymptotically Normal with zero mean vector and a finite, non-singular covariance matrix denoted by C ; see Gonçaves and Kilian (2004, Theorem 3.1). Not surprisingly, given the conditional heteroskedasticity of the errors, C is of the “sandwich” type, for example as in (1.38). Given a consistent estimator of C , it is possible to derive standard asymptotic χ^2 tests of hypotheses that impose linear restrictions on α . However, there is evidence that appropriate bootstrap tests might perform better in finite samples than such asymptotic theory tests; see Gonçaves and Kilian (2004).

Gonçaves and Kilian establish the first-order asymptotic validity of three bootstrap methods for autoregressions with conditionally heteroskedastic errors. First, a modification of the recursive bootstrap for the case of IID errors is examined. This modified dynamic bootstrap scheme is called the *recursive-design wild bootstrap* and can be written as

$$y_t^* = \hat{\alpha}_1 y_{t-1}^* + \cdots + \hat{\alpha}_p y_{t-p}^* + u_t^*, t = 1, \dots, n, \tag{5.28}$$

in which a wild bootstrap approach is used to obtain the bootstrap errors u_t^* , according to

$$u_t^* = \hat{u}_t \epsilon_t, t = 1, \dots, n, \tag{5.29}$$

with \hat{u}_t being a typical residual from the OLS estimation of (5.27) and the variables ϵ_t being IID drawings from a pick distribution with zero mean, variance one and finite fourth moment. The start-up values required for (5.28), that is, y_s^* with $s \leq 0$, are set equal to zero in Gonçaves and Kilian (2004), which reflects the absence from (5.27) of an intercept or other deterministic terms.

The second bootstrap considered by Gonçaves and Kilian is a *fixed-design wild bootstrap* in which the lagged dependent variables used as regressors in (5.27) are treated as if they were exogenous. In this approach, bootstrap data are generated by

$$y_t^* = \hat{\alpha}_1 y_{t-1} + \cdots + \hat{\alpha}_p y_{t-p} + u_t^*, t = 1, \dots, n, \tag{5.30}$$

in which the errors u_t^* are obtained using a scheme like (5.29).

The third bootstrap also treats the lagged dependent variables in the regressor set of (5.27) as if they were fixed in repeated sampling and just applies the pairs bootstrap approach; so that (5.16) is replaced by

$$\check{G}_{dyn} : \text{probability } \frac{1}{n} \text{ on } (y_t; y_{t-1}, \dots, y_{t-p}), t = 1, \dots, n.$$

Given suitable assumptions, Gonçalves and Kilian show that all three bootstrap methods yield the correct asymptotic distribution for the OLS estimator; see Gonçalves and Kilian (2004, Theorems 3.2–3.4). Thus it is possible to use either wild bootstraps or pairs bootstraps in autoregressive models with serially uncorrelated errors.

In their simulation experiments, which are designed to provide evidence about the finite sample performance of their bootstrap tests, Gonçalves and Kilian use a standard Normal distribution as the pick distribution, as in the wild “fixed regressor bootstrap” of Hansen (2000), but remark that their results are robust to alternative choices. They conclude that the recursive-design wild bootstrap of (5.28) and (5.29) “seems best suited for applications in empirical macroeconomics” and that classical IID-based residual resampling can lead to serious inaccuracies when there is conditional heteroskedasticity; see Gonçalves and Kilian (2004, p. 106).

The simulation experiments conducted by Gonçalves and Kilian allow for a GARCH(1, 1) process of the form

$$\sigma_t^2 = \omega_1 + \omega_2 u_{t-1}^2 + \omega_3 \sigma_{t-1}^2.$$

If a researcher were confident about the adequacy of a GARCH model, its estimated counterpart could be used in the bootstrap DGP; see Davidson and MacKinnon (2006, pp. 824–825). However, in the likely absence of such confidence, the recursive-design wild bootstrap is a more robust procedure.

The pure autoregression of (5.27) is, of course, a special case of the dynamic regression model

$$y_t = \mathbf{y}'_{t(p)} \boldsymbol{\alpha} + \mathbf{x}'_t \boldsymbol{\beta} + u_t = \mathbf{w}'_t \boldsymbol{\gamma} + u_t, t = 1, \dots, n, \quad (5.31)$$

in which $\mathbf{y}'_{t(p)} = (y_{t-1}, \dots, y_{t-p})$, $p \geq 1$, \mathbf{x}_t is exogenous, $\mathbf{w}'_t = (\mathbf{y}'_{t(p)}, \mathbf{x}'_t)$, $\boldsymbol{\gamma}' = (\boldsymbol{\alpha}', \boldsymbol{\beta}')$ and the errors u_t satisfy the assumptions of Gonçalves and Kilian (2004). Models which, like (5.31), contain both exogenous and lagged dependent variables are used frequently in applied time series regressions. Godfrey and Tremayne, relying upon the theoretical analysis

in Gonçalves and Kilian (2004), apply a recursive-design wild bootstrap, in which (5.28) is replaced by

$$y_t^* = \hat{\alpha}_1 y_{t-1}^* + \cdots + \hat{\alpha}_p y_{t-p}^* + \mathbf{x}_t' \hat{\boldsymbol{\beta}} + u_t^*, t = 1, \dots, n,$$

and the errors u_t^* are generated by schemes like (5.29) for various choices of the pick distribution; see Godfrey and Tremayne (2005). After carrying out a number of simulation experiments, Godfrey and Tremayne recommend that the Rademacher distribution $\mathcal{D}_{\epsilon,6}$ be used in preference to the other pick distributions described above. As will be seen in the next chapter, the Rademacher distribution, which is examined in detail in Davidson and Flachaire (2001, 2008), often emerges as the source of reliable inferences when carrying out wild bootstrap tests in regression models with heteroskedastic errors; also see, for example, Davidson et al. (2007).

5.3. Bootstrap methods for homoskedastic autocorrelated errors

Standard textbook discussions include reference to the possibility that the errors will be autocorrelated when time series data are employed in applied econometric work. The message is sometimes reinforced by providing plots of OLS residuals and/or summary statistics derived from these residuals. However, several authors have commented on the fact that marked autocorrelation of OLS residuals could reflect the effects of misspecification of the systematic part of the regression model, rather than the genuine autocorrelation of errors about a well-specified regression mean function; see Greene (2008, pp. 626–627) and Mizon (1995). The bootstrap techniques that are discussed in this section are only intended for the case of genuine error autocorrelation. They are not being recommended for application in the context of regression conditional mean functions that are subject to some unknown form of misspecification, for example, incorrect functional form. In addition, the discussion of bootstrap methods below is based upon the assumption that the error process is stationary; see Tremayne (2006, section 6.2) for useful explanations of stationarity and other relevant time series properties such as invertibility.

The format of this section is similar to that of the previous section, which dealt with heteroskedastic regression models. Model-based bootstrap methods are described first and then procedures designed to provide asymptotically valid inference in the presence of unspecified forms of

autocorrelation are considered. However, there is an important difference between the effects of heteroskedasticity and autocorrelation on OLS estimators. Provided standard regularity conditions are satisfied, the OLS estimator of the regression coefficient vector is consistent in the presence of unspecified forms of heteroskedasticity, whether or not lagged values of the dependent variable are included in the regressor set. In contrast, when there is error autocorrelation of unknown form, the consistency of the OLS estimator requires that lagged dependent variables are not used as regressors. Consequently, restricting the regressors to be exogenous is not just a convenient simplification for exposition that can be relaxed at an appropriate point and allowing for dynamic models would require the use of, for example, instrumental variable estimators in place of the OLS method.

5.3.1. Model-based bootstraps

Computer programs sometimes allow the estimation of regression models with errors that are generated by some special case of the general autoregressive-moving average (ARMA) model. For example, if it is assumed that the errors are from an ARMA(p, q) process, the error model can be written as

$$u_t = \sum_{j=1}^p \phi_j u_{t-j} + \epsilon_t + \sum_{j=1}^q \theta_j \epsilon_{t-j}. \quad (5.32)$$

The random variables ϵ_t in (5.32) are IID with zero mean and finite variance σ_ϵ^2 . The coefficients that appear in (5.32) must satisfy certain conditions for identifiability, invertibility and stationarity for the purposes of conventional asymptotic analysis. Detailed discussion of the properties of stationary ARMA processes can be found elsewhere, for example, Hamilton (1994, chapter 3) or Tremayne (2006). It is simply noted here that the following coefficient restrictions must be satisfied: the equations

$$\phi(\lambda) = \lambda^p - \sum_{j=1}^p \phi_j \lambda^{p-j} = 0,$$

and

$$\theta(\lambda) = \lambda^q + \sum_{j=1}^q \theta_j \lambda^{q-j} = 0,$$

have all roots inside the unit circle; $\phi(\lambda)$ and $\theta(\lambda)$ have no common roots; and $\phi_p \neq 0$ or $\theta_q \neq 0$.

It is often the case that emphasis is placed on the simpler class of pure AR(p) models when discussing autocorrelation models for regression errors. In particular, the stationary AR(1) model, which has the form

$$u_t = \phi_1 u_{t-1} + \epsilon_t, |\phi_1| < 1, \tag{5.33}$$

is frequently examined in some detail. This simple scheme will be used as the starting point for discussing model-based bootstraps for regression equations with autocorrelated errors.

Several articles have been published in which bootstrap methods are discussed in the context of regression models with exogenous regressors and stationary AR(1) errors; see Li and Maddala (1996, section 3.3) for a review. A clear explanation of a bootstrap procedure for AR(1) error models is given in Rayner (1991). After fitting a regression model of the form

$$y_t = \mathbf{x}'_t \boldsymbol{\beta} + u_t, t = 1, \dots, n, \tag{5.34}$$

Rayner uses the OLS residuals \hat{u}_t to estimate the coefficient of (5.33) by

$$\hat{\phi}_1 = \frac{\sum_{t=2}^n \hat{u}_t \hat{u}_{t-1}}{\sum_{t=1}^n \hat{u}_t^2}.$$

This estimate can be used to derive a sequence of residual counterparts of the IID terms ϵ_t , according to

$$\hat{\epsilon}_t = \hat{u}_t - \hat{\phi}_1 \hat{u}_{t-1}, \text{ for } t = 2, \dots, n.$$

Let $\hat{\epsilon}_t^c$ denote a typical centred residual, that is,

$$\hat{\epsilon}_t^c = \hat{\epsilon}_t - \frac{1}{n-1} \sum_{s=2}^n \hat{\epsilon}_s.$$

The n bootstrap errors $\{u_t^*; t = 1, \dots, n\}$ are obtained in Rayner (1991) as follows: (i) select one term randomly, with replacement, from the sequence $\{\hat{\epsilon}_t^c; t = 2, \dots, n\}$ and divide this term by $\sqrt{1 - \hat{\phi}_1^2}$ in order to construct u_1^* ; and (ii) given $\hat{\phi}_1$ and u_1^* from (i), generate

$$u_t^* = \hat{\phi}_1 u_{t-1}^* + \epsilon_t^*, t = 2, \dots, n,$$

in which the terms ϵ_t^* are obtained by random sampling, with replacement, from

$$\hat{\mathcal{F}}_\epsilon : \text{probability of } \frac{1}{n-1} \text{ on } \hat{\epsilon}_t^c \text{ for } t = 2, \dots, n.$$

Step (i) represents an attempt to reflect, in the bootstrap world, the fact that u_1 , like any term u_t that is generated by (5.33), has mean zero and variance $\sigma_\epsilon^2/(1 - \phi_1^2)$. Given the bootstrap errors from steps (i) and (ii), bootstrap data on the dependent variable can then be derived in the usual way, that is, with

$$y_t^* = \mathbf{x}_t' \hat{\boldsymbol{\beta}} + u_t^*, t = 1, \dots, n, \quad (5.35)$$

in which $\hat{\boldsymbol{\beta}}$ is the OLS estimator. However, the OLS estimator $\hat{\boldsymbol{\beta}}$ might be replaced by some feasible GLS estimator in a practical situation in which it had been assumed that the errors were generated by (5.33).

Steps like (i) and (ii) which are used by Rayner for the AR(1) error model must be modified when more complex types of autocorrelation are required. The asymptotic validity of appropriate modifications of the bootstrap scheme for the case of ARMA(p, q) autocorrelation models is suggested by results in Kreiss and Franke (1992). Kreiss and Franke prove asymptotic validity of a bootstrap based upon \sqrt{n} -consistent estimators of the parameters corresponding to those of (5.32) when an ARMA(p, q) model is fitted to observed data. The application of their results to regression modelling can be summarized as follows.

Suppose that the full model consists of (5.32) and (5.34), with the exogenous regressors of the latter satisfying conditions for the consistency of the OLS estimator. Let $\hat{\cdot}$ denote a \sqrt{n} -consistent estimator; so that

$$\sqrt{n}(\hat{\beta}_j - \beta_j) = O_p(1), j = 1, \dots, k,$$

$$\sqrt{n}(\hat{\phi}_j - \phi_j) = O_p(1), j = 1, \dots, p,$$

and

$$\sqrt{n}(\hat{\theta}_j - \theta_j) = O_p(1), j = 1, \dots, q,$$

when the true DGP consists of (5.32) and (5.34). The fitted residuals, which correspond to the IID terms ϵ_t , are denoted by $\hat{\epsilon}_t, t = 1, \dots, n$.

Following the suggestion for the treatment of observed data given in Kreiss and Franke (1992), these residuals can be derived by setting starting values ($\hat{\epsilon}_r, r \leq 0; \hat{u}_s, s \leq 0$) equal to zero in the recursive scheme

$$\hat{\epsilon}_t = \hat{u}_t - \sum_{j=1}^p \hat{\phi}_j \hat{u}_{t-j} - \sum_{j=1}^q \hat{\theta}_j \hat{\epsilon}_{t-j}, t = 1, \dots, n,$$

rather than adopting Rayner's strategy of trying to mimic the stationary distribution. Kreiss and Franke report that this approach works well in their bootstrap. Centred residuals $\hat{\epsilon}_t^C$ are then calculated by using

$$\hat{\epsilon}_t^C = \hat{\epsilon}_t - \frac{1}{n} \sum_{s=1}^n \hat{\epsilon}_s, t = 1, \dots, n.$$

The bootstrap errors u_t^* for the regression model are given by

$$u_t^* = \sum_{j=1}^p \hat{\phi}_j u_{t-j}^* + \epsilon_t^* + \sum_{j=1}^q \hat{\theta}_j \epsilon_{t-j}^*, t = 1, \dots, n, \tag{5.36}$$

in which starting values ($\epsilon_r^*, r \leq 0; u_s^*, s \leq 0$) are set equal to zero and the terms $\epsilon_t^*, t = 1, \dots, n$, are obtained by random sampling, with replacement, from the EDF

$$\hat{F}_\epsilon^t : \text{probability of } \frac{1}{n} \text{ on } \hat{\epsilon}_t^C \text{ for } t = 1, \dots, n.$$

The bootstrap errors are clearly not stationary because of the way in which the recursion (5.36) is started. Some researchers prefer to take action to reduce the effects of fixing starting values to be zero. A common device is to generate $n + m$ errors u_t^* , with m being moderately large, for example, $m = 100$, and then to discard the first m terms; see Davison and Hinkley (1997, p. 391). Given a sequence $\{u_t^*, t = 1, \dots, n\}$ which is to be used, bootstrap data can be derived using (5.35).

It might be thought that it is inappropriate to appeal to the results in Kreiss and Franke (1992) because they are derived under the assumption that observations are available from the ARMA(p, q) process, whereas regression errors are unobservable. However, given standard regularity conditions, the analysis in Pierce (1971) implies that the effects of replacing errors by OLS residuals, which is equivalent to replacing β by $\hat{\beta}$, are asymptotically negligible for regression models with only exogenous regressors.

Model-based bootstraps for autocorrelated errors are appealing when there is a high degree of confidence about the specification of the error model. If the true error process is the assumed $\text{ARMA}(p, q)$ model, the bootstrap using (5.35) and (5.36) will be asymptotically valid, under regularity conditions. Unfortunately, as in the discussion of model-based bootstraps for heteroskedastic errors, it seems reasonable to acknowledge that the applied worker will rarely have precise information about the error process that leads to the departure from the assumption of IID errors. Consequently, there is the real possibility that the assumed autocorrelation model will be wrong.

Suppose that the assumed error model is the $\text{ARMA}(p, q)$ scheme of (5.32) and the true model is $\text{ARMA}(P, Q)$. If $P > p$ and/or $Q > q$, the assumed model is underspecified and there will be important consequences. With all the regressors being strictly exogenous, the regression coefficients are estimated consistently but the use of an underspecified error model will lead to covariance matrices of estimators being estimated incorrectly. If there are lagged values of the dependent variable in the regressor set and the assumed autocorrelation model is an underspecified version of the true error process, the estimators of regression coefficients are, in general, inconsistent and the bootstrap may lead to very misleading inferences. More robust bootstrap procedures are, therefore, likely to be of greater practical value than the model-based bootstrap.

The techniques that are discussed in the rest of this section are intended for use when there is uncertainty about the autocorrelation model and the regressors can be assumed to be exogenous. The origins of these techniques are in the analysis of bootstraps for stationary time series variables with autocorrelation structures that are not assumed to be produced by a finite-dimensional parametric model. The relevant time series literature is sometimes very technical. Useful surveys are provided in Bühlmann (2002), Härdle et al. (2003), Mammen and Nandi (2004, section 2.4) and Politis (2003). Reviews written for an audience with interests in econometrics have also been published; see Berkowitz and Kilian (2000) and Li and Maddala (1996).

5.3.2. Block bootstraps

Suppose then that the aim is to devise a bootstrap scheme that is appropriate when the assumed model is (5.34), with the regressors being strictly exogenous and the errors being autocorrelated and strictly stationary. It is assumed that $E(u_t | \mathbf{x}_t) = 0$, which is equivalent to the assumption that $E(y_t | \mathbf{x}_t) = \mathbf{x}_t' \boldsymbol{\beta}$. The regression parameter $\boldsymbol{\beta}$ can be estimated

consistently by the OLS estimator $\hat{\beta}$, provided weak assumptions are satisfied, and so $\hat{y}_t = \mathbf{x}'_t \hat{\beta}$ can be used as the conditional mean $E^*(y_t^* | \mathbf{x}_t)$ in the bootstrap world. In order to implement (5.35), the remaining task is to obtain bootstrap errors u_t^* by using the OLS residuals \hat{u}_t to mimic the dependent errors $u_t, t = 1, \dots, n$. However, no attempt is to be made to specify a finite-dimensional model of the dependence of the errors. Instead resampling is based upon subsets of consecutive values of residuals in what is called a *block bootstrap*.

A simple form of the block bootstrap, which allows for unspecified forms of error autocorrelation, can be based upon an approach discussed in Carlstein (1986). This approach involves partitioning the set of n OLS residuals $(\hat{u}_1, \hat{u}_2, \dots, \hat{u}_n)$ into non-overlapping subsets, called *blocks*, each of which contains ℓ terms, that is, the *block-length* is ℓ . For simplicity of exposition, it is assumed that $n = b \times \ell$ for some integer value of b . Thus the number of *non-overlapping blocks* is b . The first block is $\hat{U}_1 = (\hat{u}_1, \hat{u}_2, \dots, \hat{u}_\ell)$, the second block is $\hat{U}_2 = (\hat{u}_{\ell+1}, \hat{u}_{\ell+2}, \dots, \hat{u}_{2\ell})$ and so on, with the last block being $\hat{U}_b = (\hat{u}_{n-\ell+1}, \hat{u}_{n-\ell+2}, \dots, \hat{u}_n)$. These blocks are used to define the probability model for a bootstrap error block of length ℓ , denoted by U^* , with

$$\Pr(U^* = \hat{U}_j) = \frac{1}{b} \text{ for } j = 1, \dots, b.$$

In the block bootstrap, a sequence of $n = b \times \ell$ bootstrap errors is obtained by pasting together the blocks $(U_1^*, U_2^*, \dots, U_b^*)$, which are obtained by random sampling, with replacement, from the EDF

$$\hat{\mathcal{F}}_U : \text{probability of } \frac{1}{b} \text{ on } \hat{U}_j \text{ for } j = 1, \dots, b. \tag{5.37}$$

Given that sampling is with replacement, an application of the block bootstrap can lead to some blocks of residuals being used more than once and others not being used at all in the derivation of bootstrap errors. For example, if $n = 50$ residuals are split into 10 blocks, each containing 5 values,

$$(\hat{u}_1, \hat{u}_2, \dots, \hat{u}_{50}) = (\hat{u}_1 : \hat{u}_2 : \hat{u}_3 : \hat{u}_4 : \hat{u}_5 : \hat{u}_6 : \hat{u}_7 : \hat{u}_8 : \hat{u}_9 : \hat{u}_{10}),$$

random sampling, with replacement, might generate the 50 bootstrap errors of

$$(u_1^*, u_2^*, \dots, u_{50}^*) = (\hat{u}_4 : \hat{u}_5 : \hat{u}_{10} : \hat{u}_4 : \hat{u}_7 : \hat{u}_7 : \hat{u}_1 : \hat{u}_6 : \hat{u}_8 : \hat{u}_6),$$

in which $\widehat{U}_j = (\widehat{u}_{5j-4}, \widehat{u}_{5j-3}, \widehat{u}_{5j-2}, \widehat{u}_{5j-1}, \widehat{u}_{5j})$ for $j = 1, 2, \dots, 10$. The blocks \widehat{U}_4 , \widehat{U}_6 and \widehat{U}_7 all appear twice, whereas \widehat{U}_2 , \widehat{U}_3 and \widehat{U}_9 are not used for this drawing of the bootstrap error vector.

The assumption that there is an integer value of b such that $n = b \times \ell$ can be relaxed without causing any important problems. Suppose that n/ℓ is not an integer and that $\lceil n/\ell \rceil$ is the smallest integer not less than n/ℓ . More than the required number of bootstrap errors can be obtained by resampling $\lceil n/\ell \rceil$ blocks, using (5.37). The last $[(\ell \times \lceil n/\ell \rceil) - n]$ bootstrap errors can then be deleted to define the n -dimensional bootstrap error vector; see Politis and Romano (1994, p. 1304).

When applying a block bootstrap, the hope is that, if the blocks of residuals are long enough, the autocorrelation structure of the true errors will be accurately reflected by the bootstrap errors and so repeated samples of bootstrap data derived from (5.35) will provide useful approximations for the distributions of statistics that are calculated from the actual data. The use of too small a value for ℓ will adversely affect the performance of the procedure and the inappropriate IID bootstrap is produced in the extreme case of using $\ell = 1$. There are, however, costs, as well as benefits, associated with high values of ℓ . Clearly $b = n/\ell$ decreases as ℓ increases, for a fixed sample size n , and so fewer unique bootstrap samples are available; see Davison and Hinkley (1997, pp. 396–397) for a discussion of the choice of the block-length. Also some authors point out that resampled terms may be too similar to the actual set if too high a value of ℓ is used; see Davidson and MacKinnon (2006, p. 830).

It is also possible to derive a *moving block bootstrap* by using overlapping blocks of residuals, as suggested in Künsch (1989). With ℓ again representing a fixed and common block length, the first block is $\widehat{U}_1 = (\widehat{u}_1, \widehat{u}_2, \dots, \widehat{u}_\ell)$, as before, but the second block is now $\widehat{U}_2 = (\widehat{u}_2, \widehat{u}_3, \dots, \widehat{u}_{\ell+1})$, and so on. With overlapping blocks of length ℓ and a sample size of n , there are $n - \ell + 1$ blocks, rather than n/ℓ blocks for the non-overlapping blocks method. For example, with $\ell = 5$ and $n = 50$, an overlapping blocks scheme yields 46 blocks and a non-overlapping blocks scheme has 10 blocks. The relative merits of overlapping and non-overlapping block bootstraps have been discussed in the statistics literature; see Horowitz (2001, section 4.1.1) for references. Horowitz remarks that the differences between the two methods are likely to be very small in applied work.

Whichever of the two forms of the block bootstrap is adopted, the choice of the block length can be very important. Consideration of the asymptotic theory of block bootstraps leads to the conclusion that, as

$n \rightarrow \infty$, ℓ should also tend to infinity but at a slower rate, so $\ell/n \rightarrow 0$. It is shown in Hall et al. (1995) that the *optimal block length* depends upon the context in which the block bootstrap is being used. More precisely, the optimal values are such that: $\ell = O(n^{1/3})$ for variance or bias estimation; $\ell = O(n^{1/4})$ for estimation of a one-sided distribution function; and $\ell = O(n^{1/5})$ for estimation of a two-sided distribution function. Unfortunately rules about the optimal asymptotic order of magnitude of ℓ may not be of great help to the applied researcher who is faced with the problem of analyzing a finite number of observations. Hall et al. discuss an empirical method for choosing the block length.

Whatever the choice of the fixed value of ℓ , the way in which the selected blocks $(\mathcal{U}_1^*, \mathcal{U}_2^*, \dots, \mathcal{U}_b^*)$ are pasted together implies that the bootstrap distribution (conditional upon actual data) is nonstationary. If, instead of being fixed, the lengths of blocks are assumed to be IID random variables, each having a geometric distribution, so that

$$\Pr(\ell = m) = \eta(1 - \eta)^{m-1}, 0 < \eta < 1, m = 1, 2, \dots,$$

it is possible to obtain a stationary bootstrap error series; see Politis and Romano (1994). With the block length ℓ being random, resampling continues until n bootstrap errors have been generated; see Politis and Romano (1994) for further discussion and numerical examples.

It is clearly useful that the block bootstrap methods do not require specific assumptions to be made about the form of the error autocorrelation. However, there are doubts about the practical value of such methods and open questions about their application in empirical work. Davidson and MacKinnon remark that the “biggest problem with block bootstrap methods is that they often do not work very well” (Davidson and MacKinnon, 2006, p. 831). Results can be very sensitive to the choice of the block lag length ℓ ; see, for example, Léger et al. (1992, section 3.3.1). Even with well-motivated choices for ℓ , asymptotic analyses and results from simulation experiments suggest that only small improvements may be gained by using the block bootstrap, rather than simply relying upon first-order asymptotic theory; see Härdle et al. (2003) and Horowitz (2003).

5.3.3. Sieve bootstraps

As an alternative to the use of the nonparametric block bootstrap approach, several authors have considered the possibility of approximating the error process by an autoregression. The use of an autoregression

leads to similarities with the model-based approach outlined above: a model is fitted; residuals are calculated; and bootstrap samples are generated by combining the estimated model with resampled sets of the (recentred) residuals. However, in contrast to the model-based technique, the order of the autoregression is not assumed to be finite and instead is assumed to tend to infinity as the sample size grows. This sort of approximation has been used in econometrics and time series analysis for many years. For example, it is used in the context of feasible generalized least squares estimation by Amemiya who provides an asymptotic analysis and comments on empirical application; see Amemiya (1973). In recent years, the use of autoregressions to approximate unspecified forms of autocorrelation has become widely known in the context of the Augmented Dickey Fuller (ADF) test for a unit root; see Chang and Park (2002) and Haldrup and Jansson (2006, section 7.2.1).

In order to make use of the time series results, for example, as reviewed in Bühlmann (2002), Mammen and Nandi (2004) or Politis (2003), the error term of (5.34) is assumed to be generated by an invertible and stationary process. More precisely, the errors are a linear time series with the $AR(\infty)$ representation

$$u_t = \sum_{j=1}^{\infty} \phi_j u_{t-j} + \epsilon_t, \quad (5.38)$$

in which $\sum_{j=1}^{\infty} \phi_j^2 < \infty$ and the *innovations* ϵ_t are IID with zero mean and finite variance. The autoregressive (AR) *sieve bootstrap*, which serves as an approximation to (5.38), is written as

$$u_t = \sum_{j=1}^{p(n)} \phi_j u_{t-j} + \epsilon_t, \quad (5.39)$$

in which $p(n) \rightarrow \infty$ with $p(n) = o(n)$, so that $p(n)/n \rightarrow 0$, as $n \rightarrow \infty$.

Asymptotic analysis indicates that, when it is applicable, the sieve bootstrap has properties that are superior to the block bootstrap; see, for example, the results given in Härdle et al. (2003, section 4.1). Indeed, Choi and Hall argue that

for linear time series the sieve bootstrap has substantial advantages over blocking methods, to such an extent that block-based methods are not really competitive (Choi and Hall, 2000).

However, for the time series results about the sieve bootstrap to have relevance in regression modelling, they must be applicable to the OLS

residuals that have to be used as proxies for the unobservable errors generated by (5.38). Consequently, there must be no asymptotically relevant effects associated with using residuals in place of errors. It is, therefore, assumed that the regressors of (5.34) are strictly exogenous and satisfy standard conditions for the consistency and asymptotic Normality of OLS estimators. Under the assumption that the use of the AR-sieve bootstrap is asymptotically justified when applied to OLS residuals, the bootstrap data for the regression model can be generated as follows.

First, estimate the model (5.34) by OLS and obtain the OLS residuals $\hat{u}_t, t = 1, \dots, n$. The next step is to specify a counterpart of (5.39) by choosing a value for the lag length $p(n)$. Some researchers have suggested that Akaike's information criterion (AIC) be used for this step; see Shibata (1976). Others put forward the idea that the lag length should be selected so that it leads to estimated residuals for the AR scheme (that is, estimates of innovations) that appear to be close to independent; see Amemiya (1973, p. 731). Given a selected value of $\hat{p}(n)$, the estimated AR coefficients $\hat{\phi}_j, j = 1, \dots, \hat{p}(n)$, are calculated in the third step. These estimates can be derived from OLS estimation of

$$\hat{u}_t = \sum_{j=1}^{\hat{p}(n)} \phi_j \hat{u}_{t-j} + \text{error}, t = \hat{p}(n) + 1, \dots, n, \tag{5.40}$$

or by using the Yule-Walker method, as described in Tremayne (2006, section 6.4.1). Let the residuals (estimated innovations) associated with $\hat{\phi}_j, j = 1, \dots, \hat{p}(n)$, be denoted by

$$r_t = \hat{u}_t - \sum_{j=1}^{\hat{p}(n)} \hat{\phi}_j \hat{u}_{t-j}, t = \hat{p}(n) + 1, \dots, n, \tag{5.41}$$

and their mean value be denoted by \bar{r} . The centred residuals ($r_t - \bar{r}$) are resampled to generate pseudo-innovations for the fourth step in which the bootstrap errors u_t^* are simulated.

It is worth explaining the fourth step in detail. The sieve bootstrap scheme for the generation of error terms u_t^* for the bootstrap world regression model can be represented by

$$u_t^* = \sum_{j=1}^{\hat{p}(n)} \hat{\phi}_j u_{t-j}^* + \epsilon_t^*, \tag{5.42}$$

in which the IID terms ϵ_t^* are drawn from the distribution defined by

$$\hat{\mathcal{F}}_r : \text{probability } \frac{1}{(n - \hat{p}(n))} \text{ on } r_t - \bar{r}, t = \hat{p}(n) + 1, \dots, n. \quad (5.43)$$

The scheme in (5.42) implies that, as in the case of the fixed-dimensional AR(p) model-based bootstrap, the treatment of starting values and issues of nonstationarity require attention in the context of the sieve bootstrap. A straightforward method is to set $u_t^* = 0$ for $t = 0, \dots, 1 - \hat{p}(n)$ and then to use (5.42) to generate $u_t^*, t = 1, \dots, n + m$, with m being judged to be large enough to make the effects of conditioning upon zero starting values very small for the last n terms. The required $n + m$ values of ϵ_t^* are obtained by random sampling, with replacement, from $\hat{\mathcal{F}}_r$ in (5.43). The first m simulated terms $\{u_t^* : t = 1, \dots, m\}$ are ignored and the required bootstrap data y_t^* are obtained from

$$y_t^* = \mathbf{x}'_t \hat{\boldsymbol{\beta}} + u_{t+m}^*, t = 1, \dots, n,$$

in order to obtain a good approximation to the stationary error distribution in (5.38).

However, the results from asymptotic analysis on the consistency and good convergence properties of the sieve bootstrap, as summarized in Härdle et al. (2003) and Horowitz (2001), do not imply that good approximations to every stationary and invertible error distribution will be obtained in finite sample situations. For example, investigations of the finite sample behaviour of ADF tests for unit roots indicate that the use of autoregressions to approximate pure moving average schemes can lead to substantial errors; see Schwert (1989). It may, therefore, be reasonable to be concerned about the importance of the choice of the order of the AR sieve bootstrap model (5.42).

The choice of the order of (5.42), that is, the value of $\hat{p}(n)$, is discussed in Choi and Hall (2000). After consideration of their results, Choi and Hall come to the conclusions that Akaike's information criterion is a useful way in which to select $\hat{p}(n)$ in empirical studies and the choice of $\hat{p}(n)$ in the AR-sieve bootstrap is not nearly as important as the choice of the block length ℓ in the block bootstrap. Results that indicate robustness of this type are reassuring to applied workers who wish to use the sieve bootstrap. Further reassurance is provided by Bühlmann who remarks that

Our empirical experience is that the choice of an approximating autoregressive order is quite *insensitive* with respect to the performance of the AR-sieve bootstrap, provided the chosen order is reasonable (Bühlmann, 2002, p. 58).

It is, of course, important that the sieve bootstrap model (5.42) be stationary, which implies restrictions on the estimates $\hat{\phi}_j$, $j = 1, \dots, \hat{p}(n)$. As noted above, either least squares or the Yule-Walker equations could be used for the estimation of ϕ_j , $j = 1, \dots, \hat{p}(n)$, in (5.40). The properties of these methods of estimation, as well as other techniques, have been the subject of study in the time series literature; see, for example, Paulsen and Tjøstheim (1985) and Tjøstheim and Paulsen (1983). The evidence appears to indicate that least squares estimation is superior to the Yule-Walker method, with the latter sometimes exhibiting severe bias, even with quite large samples. Unfortunately, the convenient and familiar least squares method can lead to problems in the context of the sieve bootstrap. Unrestricted OLS estimation of a model like (5.40) may produce estimates that do not satisfy the conditions for a stationary bootstrap world AR model (5.42). A third estimation procedure, known as the Burg method, should be used if appropriate software is available; see McCullough (1998) for a description of the Burg method and some empirical examples of its use.

5.3.4. Other methods

Davison and Hinkley propose a *post-blackening* scheme in which the block and sieve-type bootstraps are combined; see Davison and Hinkley (1997, p. 397). The initial step, which corresponds to the sieve approach, is to fit an AR model to *pre-whiten* the OLS residuals $\{\hat{u}_t; t = 1, \dots, n\}$. This fitted AR model is, however, only intended to eliminate much of the dependence of the OLS residuals \hat{u}_t and is not interpreted as an asymptotically valid representation of a true process for the errors u_t . Since some dependence remains, an IID-valid residual resampling scheme is inappropriate. Instead the residuals from the fitted AR model, in other words, terms that correspond to r_t in (5.41), are first recentered and then resampled, using the block bootstrap, to obtain a set of bootstrap innovations. Finally the generated innovations are post-blackened by combining them with the estimated coefficients of the AR approximation, as in (5.42) and (5.43), to derive the bootstrap errors u_t^* . In a practical situation, the order of the AR model for pre-whitening OLS residuals might be selected taking into account the nature

of the data, for example, an AR(4) scheme might be used with quarterly time series.

The bootstraps that have been discussed so far have used results developed from analysis in the *time domain*, that is, from analysis of terms ordered by a time subscript t . However, time series analysis can also be conducted in the *frequency domain*, with observations being regarded as weighted sums of periodic terms (sines and cosines); see Hamilton (1994, chapter 6). Bootstraps for regression models with autocorrelated errors have been derived using frequency domain techniques; see Christofferson (1997) and Hidalgo (2003). Time domain methods are, however, more familiar to economists and more widely used in empirical economics than frequency domain procedures.

There are alternatives to bootstrap methods for taking error autocorrelation into account. Politis and Romano provide a very clear explanation of an approach known as *subsampling* in their comments on Li and Maddala (1996); see Politis and Romano (1996). Politis and Romano propose that the sampling distribution of the statistic of interest be approximated using the values of this statistic that can be calculated from subsamples. For example, if the statistic to be studied is the OLS estimator defined by

$$\hat{\beta} = \left(\sum_{t=1}^n \mathbf{x}_t \mathbf{x}_t' \right)^{-1} \left(\sum_{t=1}^n \mathbf{x}_t y_t \right),$$

there are $(n - \ell + 1)$ subsample estimates of the form

$$\hat{\beta}_{(s)} = \left(\sum_{t=s}^{s+\ell-1} \mathbf{x}_t \mathbf{x}_t' \right)^{-1} \left(\sum_{t=s}^{s+\ell-1} \mathbf{x}_t y_t \right), s = 1, 2, \dots, (n - \ell + 1),$$

each of which is based upon a subsample of ℓ observations. Given stationarity and standard assumptions about asymptotic behaviour, $\sqrt{n}(\hat{\beta} - \beta)$ is $O_p(1)$ and has the same asymptotic distribution as $\sqrt{\ell}(\hat{\beta}_{(s)} - \hat{\beta})$, when $n \rightarrow \infty$, $\ell \rightarrow \infty$ and $\ell/n \rightarrow 0$, for $s = 1, 2, \dots, (n - \ell + 1)$. The EDF of the scaled terms $\sqrt{\ell}(\hat{\beta}_{(s)} - \hat{\beta})$ then provides an approximation to the distribution function of $\sqrt{n}(\hat{\beta} - \beta)$. Unfortunately, this simple technique has poor asymptotic properties in terms of the convergence of its approximation errors to zero; see Härdle et al. (2003), Horowitz (2001) and Mammen and Nandi (2004). Also, if, in a genuine application, the values of ℓ and n are to bear some relation to the asymptotic orders of magnitude assumed for the theory of subsampling procedures, ℓ must be

large and ℓ/n must be small; so that the sample size n may well have to be very large.

5.4. Bootstrap methods for heteroskedastic autocorrelated errors

Having discussed bootstraps designed for situations in which errors are either (i) autocorrelated and homoskedastic or (ii) heteroskedastic and serially uncorrelated, it remains to consider the more general case in which errors are both autocorrelated and (conditionally upon values of regressors) heteroskedastic. In such a case, tests that are designed to be asymptotically robust to the presence of unspecified forms of heteroskedasticity and autocorrelation are of interest. Procedures of this type are said to be heteroskedasticity and autocorrelation consistent (HAC) and are referred to below as *HAC tests*. It is useful to start by reviewing some theory-based results for asymptotic HAC tests before discussing bootstrap methods and their properties. These results indicate that, in contrast to previous cases, there are two alternative asymptotic theories from which tests can be derived and compared with bootstrap tests.

5.4.1. Asymptotic theory tests

The assumed model is the linear regression given in (5.34), that is,

$$y_t = \mathbf{x}'_t \boldsymbol{\beta} + u_t, t = 1, \dots, n,$$

in which: \mathbf{x}_t and $\boldsymbol{\beta}$ are k -dimensional vectors; and the errors u_t can now be autocorrelated and heteroskedastic, but must satisfy the moment condition that $E(u_t \mathbf{x}_t) = \mathbf{0}_k$. The moment condition is used in the technical analysis required to prove consistency and asymptotic Normality of the OLS estimator $\hat{\boldsymbol{\beta}}$; see, for example, the discussion in Fitzenberger (1998, section 2). As remarked above, there cannot be lagged values of y_t in \mathbf{x}_t if the moment condition is to hold when the errors u_t are generated by an autocorrelation scheme of unknown form. This is clearly an important restriction for those wishing to use time series regression models.

The traditional asymptotic theory for HAC tests will be examined first. Under the regularity conditions specified in Fitzenberger (1998), the standard asymptotic results apply and

$$\sqrt{n} (\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}) \sim_a N(\mathbf{0}_k, \mathbf{Q}^{-1} \Phi \mathbf{Q}^{-1}), \tag{5.44}$$

in which:

$$\mathbf{Q} = p \lim n^{-1} \sum_{t=1}^n \mathbf{x}_t \mathbf{x}_t';$$

and Φ is the asymptotic covariance matrix of $n^{-1/2} \sum_t \mathbf{x}_t u_t$, which, as in, for example, Kiefer and Vogelsang (2002), can be written as

$$\Phi = \Gamma(0) + \sum_{j=1}^{\infty} (\Gamma(j) + \Gamma(j)'), \quad (5.45)$$

where $\Gamma(j) = E(\mathbf{x}_t u_t u_{t-j} \mathbf{x}_{t-j}')$. The matrix \mathbf{Q} can be estimated consistently by

$$\hat{\mathbf{Q}} = n^{-1} \sum_{t=1}^n \mathbf{x}_t \mathbf{x}_t' \quad (5.46)$$

and so asymptotically valid inference about β requires a consistent estimator of Φ .

The discussion in Section 1.4.2 of the consistent estimation of Φ covered the well-known Newey-West estimator, defined in (1.47) to be

$$\hat{\mathbf{J}} = \hat{\Gamma}(0) + \sum_{j=1}^l \left(1 - \frac{j}{l+1}\right) (\hat{\Gamma}(j) + \hat{\Gamma}(j)'),$$

in which $\hat{\Gamma}(j) = n^{-1} \sum_{t=j+1}^n (\mathbf{x}_t \hat{u}_t \hat{u}_{t-j} \mathbf{x}_{t-j}')$, $j = 0, 1, \dots, l$, and $l \rightarrow \infty$ as $n \rightarrow \infty$, with $l = o(n)$. This estimator is one of a general family of variance estimators based upon *kernel functions*. A typical member of this family can be written as

$$\hat{\Phi} = k(0) \hat{\Gamma}(0) + \sum_{j=1}^{n-1} k\left(\frac{j}{M(n)}\right) (\hat{\Gamma}(j) + \hat{\Gamma}(j)'), \quad (5.47)$$

in which $k(w)$ is a kernel function such that $k(0) = 1$, $k(w)$ is continuous at $w = 0$, $|k(w)| \leq 1$ and $\int_{-\infty}^{\infty} k^2(w) dw < \infty$; see Andrews (1991, p. 821) for some examples of kernel functions. The term $M(n)$ that appears in the kernel function estimate in (5.47) is called the *bandwidth* and can be used to control truncation for kernel functions that satisfy $k(w) = 0$ if $w > 1$. A requirement for the bandwidth, which is sufficient for consistency of $\hat{\Phi}$ in standard cases, is that $M(n) \rightarrow \infty$ in such a way that $M(n)/n \rightarrow 0$,

as $n \rightarrow \infty$; see de Jong and Davidson (2000). It is this assumption about $M(n)$ that underpins the traditional asymptotic theory of HAC tests.

Given a consistent estimator $\hat{\Phi}$, with $M(n) = o(n)$ in (5.47), asymptotically valid inferences about β can be obtained in the traditional asymptotic theory by combining (5.44), (5.46) and (5.47). For example, if the null hypothesis consists of $q < k$ restrictions that can be written as $R\beta = r$, the Wald-type test statistic

$$W = (\mathbf{R}\hat{\beta} - \mathbf{r})' \left[\mathbf{R}\hat{Q}^{-1} \hat{\Phi} \hat{Q}^{-1} \mathbf{R}' \right]^{-1} (\mathbf{R}\hat{\beta} - \mathbf{r}), \tag{5.48}$$

is asymptotically distributed as $\chi^2(q)$ when the null is true. In the special case in which $q = 1$, a “*t*-ratio” procedure can be obtained, with the Standard Normal distribution being an asymptotically valid reference distribution for critical values. These results on asymptotic null distributions are the basis of traditional asymptotic HAC tests and are outlined in many textbooks. However, many authors have found that this asymptotic theory provides a poor approximation in finite samples, with evidence that actual significance levels are much higher than the desired levels; see, for example, Ligeralde and Brown (1995). Fortunately, there is an alternative approach to asymptotic analysis, which is proposed in Kiefer and Vogelsang (2005).

Kiefer and Vogelsang focus attention on the assumption that is made about the order of magnitude of the bandwidth term $M(n)$ in (5.47). It has been argued above that the assumptions that are made about asymptotic orders of magnitude should have two characteristics: first, they should lead to the test statistic having a non-degenerate asymptotic distribution when the null hypothesis is true; and second, they should bear some reasonable correspondence to the magnitude of terms in actual applications. Thus, traditional asymptotic analysis for HAC tests, in which $M(n) = o(n)$, may be useful when $M(n)/n$ is small and n is large. Kiefer and Vogelsang point out that $M(n)/n$ will be a positive fraction in any genuine application and derive an asymptotic theory for HAC tests under the assumption that $M(n)/n = c_{KV}, 0 < c_{KV} \leq 1$. (Kiefer and Vogelsang use a different notation, with b , not c_{KV} , denoting $M(n)/n$. However, b has been used above in this chapter for the number of blocks in a block bootstrap and will be used below for the same purpose.)

Clearly $M(n) = O(n)$, not $o(n)$, in the new asymptotic theory discussed in Kiefer and Vogelsang (2005). This change in the assumptions has important implications for asymptotic analysis. A conventional estimator of the asymptotic covariance matrix of $\sqrt{n}(\hat{\beta} - \beta)$ tends to a

matrix with elements that are random variables, rather than constants; see Kiefer and Vogelsang (2005, p. 1142). However, it is the limiting behaviour of test statistics, not of covariance matrix estimators, that is of primary interest. Kiefer and Vogelsang show that a statistic like \mathcal{W} in (5.48) is, under the assumptions of their new theory, asymptotically pivotal and has a non-standard asymptotic null distribution that depends upon the form of the kernel function in (5.47) and the value of c_{KV} . Critical values are simulated for the t -ratio that can be used when testing a single restriction and are provided via estimated coefficients for the critical value response surface

$$cv(c_{KV}) = a_0 + a_1 c_{KV} + a_2 c_{KV}^2 + a_3 c_{KV}^3,$$

for various combinations of the kernel function and the desired significance level; see Kiefer and Vogelsang (2005, p. 1146, Table 1). These new critical values provide an alternative to those from the $N(0, 1)$ distribution of the traditional asymptotic theory when approximating the unknown finite sample critical values. After comparing tests based upon their new theory with those derived from traditional asymptotic analysis, Kiefer and Vogelsang conclude that their “new approximation performs much better and gives insight into the choice of kernel and bandwidth” (Kiefer and Vogelsang, 2005, section 6).

5.4.2. Block bootstraps

In addition to two types of asymptotic test, there is, of course, the possibility of using a suitable bootstrap procedure. When deriving bootstrap HAC tests, researchers, for example, Fitzenberger (1998), have used a moving (overlapping) block bootstrap, as discussed in Künsch (1989). Since a test statistic like \mathcal{W} of (5.48) is asymptotically pivotal whichever of the two asymptotic theories is applied, it might be thought that, given the results in Beran (1988), the block bootstrap variant of the HAC test would enjoy an asymptotic refinement. However, as explained in Davison and Hall (1993), the use of the block bootstrap leads to complications.

Davison and Hall state that “the block bootstrap method damages the dependence structure of the data” (Davison and Hall, 1993, p. 216). This damage has an impact on the asymptotic covariance matrix of the OLS estimator in the bootstrap world and, unless adjustments are made, the block bootstrap is no more accurate than asymptotic theory; see Davison and Hall (1993), Götzte and Künsch (1996) and Horowitz (2001, p. 3191). If the statistics are calculated from bootstrap data using the same formula

as is employed for the actual data, with no adjustments to variance and covariance terms being made, the bootstrap test is said to be *naive*.

In order to describe the block bootstrap in the context of HAC tests, let ℓ denote the block length, which is assumed to be $o(n^{1/2})$ for the asymptotic validity of the bootstrap. As in Fitzenberger (1998), a typical block of observations on the variables of the regression model can be written as

$$\mathbf{B}_{s,\ell} = \{\mathbf{B}_{s,\ell}^y : \mathbf{B}_{s,\ell}^x\},$$

in which $\mathbf{B}_{s,\ell}^y = (y_s, \dots, y_{s+\ell-1})'$ and $\mathbf{B}_{s,\ell}^x$ is the $\ell \times k$ matrix with rows $\mathbf{x}'_t, t = s, \dots, s + \ell - 1$. There are $n - \ell + 1$ possible blocks $\mathbf{B}_{s,\ell}$, which are viewed as the possible values of the $\ell \times (1 + k)$ random matrix \mathbf{B} in a bootstrap probability model defined by

$$\mathcal{F}_B : \mathbf{B} = \mathbf{B}_{s,\ell} \text{ with probability } \frac{1}{n - \ell + 1}, s = 1, \dots, n - \ell + 1. \quad (5.49)$$

The desired bootstrap sample size is n^* , which is assumed to be $O(n)$, with $n^* = b\ell$ for some integer value of b . (If $n^* = n$, it is assumed, for simplicity of exposition, that $n = b\ell$ for some integer value of b .) The b required blocks can be obtained by random sampling, with replacement, from (5.49). Let the selected blocks be denoted by $\mathbf{B}_{s,\ell}^*, s = 1, \dots, b$. The bootstrap sample of n^* observations is then formed by joining together the b selected blocks in the usual way to obtain $\mathbf{S}^* = \{(y_t^*, \mathbf{x}_t^{*'}), t = 1, \dots, n^*\}$.

The bootstrap sample \mathbf{S}^* can be used to obtain the OLS estimator of the regression parameter vector, which is denoted by $\hat{\boldsymbol{\beta}}^*$. As in the simpler models considered in Freedman (1981), $\sqrt{n^*}(\hat{\boldsymbol{\beta}}^* - \hat{\boldsymbol{\beta}})$, under the conditional (on actual data) bootstrap probability model, and $\sqrt{n}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})$, under the assumed probability model for the actual data, have the same asymptotic distribution; see Fitzenberger (1998). This result about asymptotic distributions for estimators can be used to derive results for test statistics.

Suppose, as in the discussion of asymptotic HAC tests above, that the null hypothesis takes the form $\mathbf{R}\boldsymbol{\beta} = \mathbf{r}$. The results in Fitzenberger (1998) imply that the asymptotic distribution of $\sqrt{n}\mathbf{R}(\hat{\boldsymbol{\beta}}^* - \hat{\boldsymbol{\beta}})$ in the bootstrap world is the same as that of $\sqrt{n}\mathbf{R}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})$ under the model assumed to generate the real observations. Now

$$\sqrt{n}\mathbf{R}(\hat{\boldsymbol{\beta}}^* - \hat{\boldsymbol{\beta}}) = \sqrt{n}(\mathbf{R}\hat{\boldsymbol{\beta}}^* - \mathbf{R}\hat{\boldsymbol{\beta}}),$$

and

$$\sqrt{n}\mathbf{R}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}) = \sqrt{n}(\mathbf{R}\hat{\boldsymbol{\beta}} - \mathbf{R}\boldsymbol{\beta}) = \sqrt{n}(\mathbf{R}\hat{\mathbf{r}} - \mathbf{r}),$$

when the null hypothesis $\mathbf{R}\boldsymbol{\beta} = \mathbf{r}$ is true. Hence a bootstrap counterpart of \mathcal{W} in (5.48), which is naive in the sense of Davison and Hall (1993), is given by

$$\mathcal{W}^* = (\mathbf{R}\hat{\boldsymbol{\beta}}^* - \mathbf{R}\hat{\boldsymbol{\beta}})' \left[\mathbf{R}\hat{\mathbf{Q}}^{*-1} \hat{\boldsymbol{\Phi}}^* \hat{\mathbf{Q}}^{*-1} \mathbf{R}' \right]^{-1} (\mathbf{R}\hat{\boldsymbol{\beta}}^* - \mathbf{R}\hat{\boldsymbol{\beta}}), \quad (5.50)$$

in which an asterisk $*$ denotes a quantity defined in terms of bootstrap, rather than actual, data.

The statistic \mathcal{W}^* of (5.50) is a naive bootstrap statistic because it is defined using the same formula as \mathcal{W} of (5.48), apart from the replacement of \mathbf{r} by $\mathbf{R}\hat{\boldsymbol{\beta}}$; see, for example, Gonçalves and Vogelsang (2006). Gonçalves and Vogelsang remark that a well-established view is that the naive bootstrap test is no more accurate than the traditional asymptotic test and refer to the pertinent results in Davison and Hall (1993) and Götze and Künsch (1996). They show that the naive bootstrap is as accurate as the asymptotic test under the new asymptotic theory of HAC tests proposed in Kiefer and Vogelsang (2005); see Gonçalves and Vogelsang (2006, Theorem 4.1).

In addition to asymptotic analyses, several authors have reported results from simulation experiments in which the finite sample behaviour of HAC tests is investigated. Kiefer and Vogelsang find evidence that traditional asymptotic theory provides poorer approximations than their new asymptotic theory, with the latter providing finite sample performance which is comparable to that obtained with bootstrap methods; see Kiefer and Vogelsang (2005). The finite sample behaviour of HAC t -tests is also investigated in Gonçalves and Vogelsang (2006). The results of experiments carried out by Gonçalves and Vogelsang indicate that the naive bootstrap gives a much more accurate approximation than the traditional asymptotic test based upon the $N(0, 1)$ distribution. Gonçalves and Vogelsang also find that the simulations suggest that, when the block length is chosen appropriately, a naive block bootstrap can offer an asymptotic refinement relative to the non-traditional asymptotic test based upon the results in Kiefer and Vogelsang (2005). The importance of the block length is also considered in Fitzenberger (1998). Fitzenberger provides results from simulation experiments in which the block length is varied. He also gives programming code for

an example in which the moving block bootstrap is applied to a simple regression, with block length varying from 1 to 15; see Fitzenberger (1998, Appendix D). However, as noted by Fitzenberger, the optimal choice of block length in finite samples remains an open question.

5.4.3. Other methods

It has been assumed so far that the asymptotic covariance matrix of $\sqrt{n}(\hat{\beta} - \beta)$, denoted by C , is estimated using $\hat{C} = \hat{Q}^{-1} \hat{\Phi} \hat{Q}^{-1}$, where \hat{Q} and $\hat{\Phi}$ are defined in (5.46) and (5.47), respectively. Gonçalves and White establish the conditions under which the moving (overlapping) block bootstrap yields a consistent estimator of C ; see Gonçalves and White (2005). Given, say, A bootstrap samples, each of size n , the OLS estimates $\hat{\beta}_i^*$ can be calculated, $i = 1, \dots, A$. The bootstrap population mean vector can be estimated by

$$\bar{\beta}^* = \frac{1}{A} \sum_{i=1}^A \hat{\beta}_i^*, \tag{5.51}$$

and the bootstrap population covariance matrix of $\sqrt{n}(\hat{\beta}^* - \hat{\beta})$ can be estimated by

$$\hat{C}^* = \frac{n}{A} \sum_{i=1}^A \left(\hat{\beta}_i^* - \bar{\beta}^* \right) \left(\hat{\beta}_i^* - \bar{\beta}^* \right)'. \tag{5.52}$$

The use of estimates of the form (5.51) and (5.52) when applying the information matrix test to IID data is considered in Dhaene and Hoorelbeke (2004). Gonçalves and White conduct simulation experiments that provide evidence that bootstrap estimation of variances can lead to much more accurate inferences than the kernel-based method for estimating variances.

In the context of HAC tests, with moving block bootstrap samples, Gonçalves and White state that their theoretical analysis justifies a double bootstrap approach in which the first-level bootstrap delivers terms like $\mathbf{R}\hat{\beta}_i^* - \mathbf{R}\hat{\beta}$ and, for each first-level sample, a second-level set of B bootstrap samples is generated in order to estimate the asymptotic covariance matrix of $\mathbf{R}\hat{\beta}_i^* - \mathbf{R}\hat{\beta}$, $i = 1, \dots, A$. This double application of the moving block bootstrap allows the calculation of a set of A bootstrap counterparts of the statistic derived from the actual data and the bootstrap

variance-covariance matrix, that is,

$$\mathcal{W}_{BS} = n(\mathbf{R}\hat{\boldsymbol{\beta}} - \mathbf{r})' [\mathbf{R}\hat{\mathbf{C}}^* \mathbf{R}']^{-1} (\mathbf{R}\hat{\boldsymbol{\beta}} - \mathbf{r}),$$

and hence the derivation of a bootstrap p -value that can be compared with the desired significance level.

Not all researchers wishing to use bootstrap tests in the context of regression models with autocorrelated and heteroskedastic errors have used block bootstrap methods. Bisaglia and Procidano combine a sieve bootstrap (with its order selected by AIC) to proxy the autocorrelation and a wild bootstrap to reflect heteroskedasticity; see Bisaglia and Procidano (2002). The modification of the standard sieve bootstrap is to define ϵ_t^* in (5.42) to be the product of r_t in (5.41) and a drawing from one of the pick distributions described in Section 5.2.3.

The subsampling technique discussed in Politis and Romano (1996) can also be employed as an alternative to the block bootstrap approach. The use of subsampling in the context of the OLS estimation of a linear regression model with heteroskedastic data is examined in Politis et al. (1997). It is shown that, under regularity conditions, subsampling methods are asymptotically valid.

5.5. Summary and concluding remarks

Applied workers often carry out regression analyses without imposing the restrictive assumption that the errors are IID. The basic idea of the bootstrap approach is to set up an artificial data generation process that mimics the model assumed to generate the actual data. Consequently, the use of IID-valid bootstrap techniques, as discussed in Chapter 2, is inappropriate when the errors of the model for the actual data are allowed to be either heteroskedastic or autocorrelated. This chapter has been devoted to descriptions and discussions of bootstrap methods that, subject to standard conditions, can be employed in the presence of heteroskedasticity and/or autocorrelation. It is assumed throughout this chapter that heteroskedasticity and autocorrelation, if present, are genuinely properties of errors.

There is usually little reliable information about the nature of heteroskedasticity or autocorrelation; so that, in general, there are no firm foundations for the specification of a parametric model that determines the ways in which the assumption of IID errors is violated. Consequently, the emphasis has been on bootstrap methods that are not based upon parametric models for the errors and are instead asymptotically valid under unknown forms of heteroskedasticity and/or autocorrelation. This

emphasis reflects a view that it is safer for applied workers to use bootstrap methods that are robust under quite general conditions than to adopt model-based methods that are more efficient when the assumed error model is correct, but are invalid when it is misspecified.

The technique known as the wild bootstrap has been recommended for use in the context of heteroskedastic regression models. The wild bootstrap is asymptotically valid under weak conditions that are required for the consistency of the OLS estimators of the coefficients of the regression model. It can be used for both cross-section and time series regressions and, in the latter case, the regressors can include lagged values of the dependent variable, that is, the regression model can be dynamic. The implementation of the wild bootstrap requires the specification of an external “pick distribution” from which to draw standardized IID terms, each of which is multiplied by a corresponding estimated residual from the actual data in order to generate a bootstrap error. Several pick distributions have been described and a particularly simple version, called the Rademacher distribution, has been singled out as often being the source of the most reliable heteroskedasticity-robust bootstrap tests. Evidence on bootstrap tests for heteroskedastic regression models that supports the choice of the wild bootstrap approach with the Rademacher pick distribution will be provided in the next chapter.

It is more difficult to recommend a bootstrap procedure for general use when the errors are assumed to be homoskedastic and autocorrelated. If autocorrelation-robust inferences based upon OLS estimators are to be made in the presence of unknown types of autocorrelation, the regressors of the model must be strictly exogenous. However, lagged values of the dependent variable are often used as regressors in applied econometrics. The exclusion of lagged dependent variables from the regressor set is, therefore, an important restriction on the class of time series regression models that can be considered when obtaining autocorrelation-robust bootstrap tests after OLS estimation.

If attention is restricted to time series regressions with strictly exogenous regressors, consistent bootstrap methods are available and two of these have been discussed in detail, viz., the block bootstrap and the sieve bootstrap. As mentioned above, there is an extensive literature on the various forms of the block bootstrap. References to this literature and surveys of findings are provided in, for example, Härdle et al. (2003) and Li and Maddala (1996). In practical situations, however, the performance of block bootstrap tests can be sensitive to the choice of the block length and at best is often not much better than that of the corresponding asymptotic theory tests.

The results that have been obtained suggest that the sieve bootstrap, which is based upon an autoregressive approximation to the error autocorrelation model, outperforms the block bootstrap when both are available. Also the sieve bootstrap has been generalized by combining it with a wild bootstrap; so that, like the block bootstrap, it can be used with error terms that are both heteroskedastic and autocorrelated. However, there are many open questions about the application of bootstrap tests in the presence of unknown forms of autocorrelation; see Horowitz (2003). After summarizing relevant results from asymptotic analyses and simulation experiments, Davidson and MacKinnon conclude that

Neither the sieve bootstrap nor the best available block bootstrap methods can be relied upon to yield accurate inferences in samples of moderate size. Even for quite large samples, they may perform little better than asymptotic tests ... (Davidson and MacKinnon, 2006, p. 835).

Thus the overall picture that emerges is that, while there appears to be a fairly secure basis for obtaining heteroskedasticity-robust bootstrap tests for regression models with independent errors, there is much more uncertainty about how to tackle the problem of error autocorrelation. Moreover, if OLS is to be the assumed estimation method, asymptotically valid autocorrelation-robust bootstrap tests can only be derived for models in which all regressors are strictly exogenous. These remarks are clearly relevant to the situation in which errors are assumed to be both heteroskedastic and autocorrelated.

There is an alternative to viewing autocorrelation as a problem that can be fixed by adjusting the formula for the asymptotic covariance matrix of the OLS estimators. Some researchers have argued that, while it is certainly desirable that the form of the econometric model should be such that its errors are independent in different time periods, there is no reason why the independent errors should be associated with an autocorrelation model for the unobservable disturbance term of a regression model. Instead it might be more appropriate to use a dynamic specification of the regression model that is sufficiently general to justify the assumption of independent errors for the regression model itself; see, for example, Mizon (1995). Spanos refers to this approach to modelling as being in the LSE tradition; see Spanos (2006) for a stimulating discussion of the "error-fixing" strategy and alternatives.

If the restrictive assumptions of homoskedasticity and Normality contained in the model used in Spanos (2006, p. 33, equation 1.10) to

represent the LSE approach are relaxed, the starting point for regression analysis can be written as

$$y_t = \mathbf{y}'_{t(p)}\boldsymbol{\alpha} + \mathbf{x}'_t\boldsymbol{\beta} + u_t, t = 1, \dots, n,$$

in which: $\mathbf{y}'_{t(p)} = (y_{t-1}, \dots, y_{t-p})$, $p \geq 1$; \mathbf{x}_t contains current and lagged values of exogenous variables; and, conditionally on regressor values, the errors are independently distributed with common zero mean. Spanos remarks that, as part of the LSE approach, the data are used to select lag lengths in a “general to specific search”. The OLS-based significance tests used in such a search can be made asymptotically robust to heteroskedasticity but would be invalid if there were autocorrelation. Consequently, when the LSE modelling strategy is adopted, it is obviously important that the assumption of independent regression errors be tested. In particular, if, as is now widely recommended, the assumption of homoskedasticity is to be relaxed, checks for autocorrelation that are asymptotically robust to heteroskedasticity are required. As will be seen in the next chapter, the wild bootstrap approach mentioned above can be combined with standard autocorrelation tests to derive procedures that are asymptotically valid under heteroskedasticity and are likely to be useful to applied workers who wish to accept Mizon’s advice not to use potentially invalid “autocorrelation correction”.

6

Simulation-based Tests for Regression Models with Non-IID Errors

6.1. Introduction

The previous chapter contained descriptions of various bootstrap methods that are designed for application to regression models with non-IID errors. These bootstrap techniques can be used to implement OLS-based tests that are asymptotically valid in the presence of heteroskedasticity and/or autocorrelation. The purpose of this chapter is to discuss some important examples of such asymptotically robust tests and to examine evidence about which bootstrap scheme gives the best finite sample approximation when several are available to the applied researcher.

It is assumed below that the decision to relax the assumption of IID errors is made before empirical analysis is conducted. Applied workers should not turn to “robust inference” only in response to statistically significant values of checks for either heteroskedasticity or autocorrelation. Evidence against the use of screening tests when deciding whether or not to use heteroskedasticity-robust procedures is reported in Long and Ervin (2000, section 4.3). In addition to this evidence, there are other grounds for doubting the wisdom of using screening tests for heteroskedasticity. It is sometimes stated that significant values of such tests may reflect problems with the specification of the conditional mean function, rather than differences in variances of fluctuations about means. For example, Zietz argues that

evidence of heteroskedasticity should not be dismissed as unimportant and/or routinely treated with the application of White’s (1980) heteroskedasticity-consistent variance covariance matrix estimator. What is need instead is an examination of the underlying causes for heteroskedastic residuals (Zietz, 2001).

Similar remarks can be made in the context of testing for stationary autocorrelated errors by means of statistics calculated from the OLS residuals. Davidson and MacKinnon remark that

There is no universally effective way of avoiding misinterpreting misspecification of the regression function as the presence of serially correlated errors (Davidson and MacKinnon, 1993, p. 364).

In view of the above comments, it is recommended that checking the specification of the regression mean function, and testing other key assumptions required for the consistency of OLS estimators, should be an essential part of any empirical analysis in which it is claimed that inferences are robust to heteroskedasticity and/or autocorrelation. In the spirit of this recommendation, all of the examples discussed in this chapter are for robust tests of the null hypothesis that the mean vector of the conditional distribution of the errors, given regressor values, has every element equal to zero. This null hypothesis is the standard form of the moment condition that often plays a central role in asymptotic theory for OLS estimators and simulation methods.

The first example of a robust bootstrap-based test of the conditional mean assumption is given in Section 6.2. The long-established RESET test for omitted variables/incorrect functional form is used to illustrate the implementation and performance of bootstrap tests that are asymptotically valid under unspecified forms of heteroskedasticity and non-Normality. These robust tests are obtained using wild bootstrap methods. Section 6.2 includes results from simulation experiments. These results indicate the inadequacy of the finite sample approximation provided by asymptotic theory and also provide evidence about the relative merits of the pick distributions that were described in Section 5.2.3.

The heteroskedastic regression models in Section 6.2 are assumed to have regressors that are strictly exogenous. For such models, autocorrelation would not imply a nonzero conditional mean for an error term, given contemporaneous values of regressors. However, when lagged values of the dependent variable are included in the regressor set, autocorrelation will, in general, imply a nonzero conditional mean for the error term and the inconsistency of the OLS estimators of regression parameters. It is, therefore, important to have reliable tests for autocorrelation that are asymptotically robust to heteroskedasticity, when the regression model is dynamic.

Section 6.3 contains an examination of the usefulness of combining a recursive wild bootstrap with a heteroskedasticity-consistent version of

the widely-used Breusch-Godfrey test, which is appropriate for dynamic regression models. Results are obtained using simulation experiments. These results reinforce the evidence reported in Section 6.2 and indicate the usefulness of wild bootstrap methods in the presence of unspecified forms of heteroskedasticity.

The wild bootstrap tests that are discussed in Sections 6.2 and 6.3 share the characteristic of having asymptotically pivotal test statistics that possess standard limit distributions, under the null hypothesis; these standard distributions are either $N(0, 1)$ when a single restriction is under test or a χ^2 distribution in the more general case of testing several restrictions. It was, however, explained in Section 1.6 that some important test statistics have non-standard asymptotic distributions, when the null hypothesis is true.

An important example of a statistic with a non-standard asymptotic distribution is provided by the analysis in Andrews (1993). This analysis is concerned with the problem of testing for a break in parameter values when the alternative is that there is a single unknown breakpoint. Section 6.4 contains a summary of work in which heteroskedasticity-consistent versions of Andrews (1993)-type tests are studied. These robust bootstrap procedures are in keeping with the guidelines given in Hansen (1999) for a modern approach to regression analysis, which are that: (i) the model be subjected to a structural break test of the sort proposed in Andrews (1993); (ii) all test statistics should be asymptotically valid under heteroskedasticity; and (iii) bootstrap methods, rather than asymptotic theory, should be considered when making inferences.

The last example of a robust bootstrap test of the zero conditional error mean hypothesis is presented in Section 6.5. Hausman provides a general discussion of testing this hypothesis and, in the context of a discussion of detecting measurement errors, proposes an estimator contrast test; see Hausman (1978). The standard version of this test is derived under the assumption of IID errors. Section 6.5 contains discussions of versions of the Hausman test that are asymptotically valid under unspecified forms of stationary autocorrelation. Different forms of robust bootstrap tests are described, including one that makes use of the fast double bootstrap (FDB) mentioned in Section 2.5. As in other sections, simulation experiments are used to collect evidence about finite sample behaviour.

Finally, a summary and some concluding remarks are provided in Section 6.6.

6.2. Bootstrapping heteroskedasticity-robust regression specification error tests

The adequacy of the specification of the conditional mean function is a prerequisite for reliable inference in regression analysis. Many tests for specification error have been proposed. However, in several cases, the original form of the test is based upon the assumption of IID (or even NID) errors when the null hypothesis of no specification errors is correct. As argued by many authors in recent years, correct specification of the mean function does not automatically entail that the conditional variances are all equal. Consequently there is a need for heteroskedasticity-robust tests for specification errors. Appropriate forms of asymptotic and wild bootstrap tests are described in this section and simulation evidence is presented.

6.2.1. The forms of test statistics

Consider a linear regression model written, as in Chapter 1, in the form

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{u}, \quad (6.1)$$

in which: \mathbf{y} and \mathbf{u} are n -dimensional random vectors, with only the former being observable; \mathbf{X} is the $n \times k$ matrix of observations on regressor variables that, under correct specification, are assumed to be strictly exogenous; and $\boldsymbol{\beta}$ is the k -dimensional vector of unknown regression coefficients. The assumption that the mean function of (6.1) is specified correctly implies that the conditional means of errors satisfy

$$E(\mathbf{u}|\mathbf{X}) = \mathbf{0}_n, \quad (6.2)$$

which is called the *orthogonality assumption* in Hausman (1978). Equation (6.2) would be false if the conditional mean function of (6.1) were to suffer from certain specification errors, for example, omitted variables, incorrect functional form, endogenous regressors. The analysis given in Ramsey (1969) provides information about how the conditional mean function $E(\mathbf{u}|\mathbf{X})$ could be affected by such specification errors.

A conventional way in which to test the orthogonality assumption (6.2) is to test the null model (6.1) against the augmented equation

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{W}\boldsymbol{\gamma} + \mathbf{u}, \quad (6.3)$$

in which \mathbf{W} is an $n \times q$, $n > k + q$, matrix of test variables that are exogenous under the assumption that the mean function of the original

regression has no specification errors. A test of the q restrictions of $H_{\gamma} : \gamma = \mathbf{0}_q$ is interpreted as a check for specification errors; see, for example, Thursby and Schmidt (1977). It is worth noting that, in contrast to the model given in (6.1) which is under test, the model in (6.3) need not correspond to a genuine attempt to explain the behaviour of the dependent variable. Models like (6.3) are often better regarded as being artificial devices that permit convenient computation of test statistics.

For any given specification error that has been made when formulating (6.1), the power of the test associated with (6.3) will be affected by the choice of the test variables in \mathbf{W} . Ramsey stresses that specification error tests will be applied after the researcher has used the available information to decide upon how to specify (6.1) and that the test variables must, therefore, be selected with impoverished information; see Ramsey (1983, pp. 243–244). Ramsey's solution in his RESET test is to use second and higher order powers of the predicted values from OLS estimation of (6.1) to form \mathbf{W} ; see Ramsey (1969). The RESET test is discussed in many textbooks and included in many estimation programs. The original derivation of the RESET test is based upon the assumption that, under correct specification, the errors are $NID(0, \sigma^2)$ and, in order to indicate how it can be generalized to be asymptotically valid under much weaker conditions, it will be useful to introduce some notation.

Let the OLS estimator, predicted value and residual vectors for the null model (6.1) be denoted by $\hat{\beta} = (\hat{\beta}_1, \dots, \hat{\beta}_k)'$, $\hat{y} = \mathbf{X}\hat{\beta} = (\hat{y}_1, \dots, \hat{y}_n)'$ and $\hat{u} = (\hat{u}_1, \dots, \hat{u}_n)'$, respectively. Similarly let $(\tilde{\beta}', \tilde{\gamma}')'$ and \tilde{u} denote the OLS estimator vector and residual vector for the augmented model in (6.3). The check for specification errors is to be carried out by testing the joint significance of the elements of $\tilde{\gamma}$.

When $H_{\gamma} : \gamma = \mathbf{0}_q$ is true, standard results imply that

$$\begin{aligned} \tilde{\gamma} &= (\mathbf{W}'\mathbf{M}\mathbf{W})^{-1} \mathbf{W}'\mathbf{M}\mathbf{y} \\ &= (\mathbf{W}'\mathbf{M}\mathbf{W})^{-1} \mathbf{W}'\mathbf{M}\mathbf{u} \\ &= (\tilde{\mathbf{W}}'\tilde{\mathbf{W}})^{-1} \tilde{\mathbf{W}}'\mathbf{u} \\ &= (\tilde{\mathbf{W}}'\tilde{\mathbf{W}})^{-1} \tilde{\mathbf{W}}'\hat{\mathbf{u}}, \end{aligned} \tag{6.4}$$

in which $\mathbf{M} = \mathbf{I} - \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'$ and $\tilde{\mathbf{W}} = \mathbf{M}\mathbf{W}$ is the matrix of residuals from the OLS regression of \mathbf{W} on \mathbf{X} ; see, for example, Greene (2008, section 3.3). If the errors of $\mathbf{u} = (u_1, \dots, u_n)'$ are allowed to be

heteroskedastic, but assumed to be independent, their covariance matrix can be represented by a diagonal matrix $\Sigma = \text{diag}(\sigma_1^2, \dots, \sigma_n^2)$. Under standard conditions, for example, as set out in White (1980), the asymptotic null distribution of $n^{1/2}\tilde{\mathbf{y}}$ is multivariate Normal with zero mean vector and a covariance matrix which is given by

$$C_{\gamma\gamma} = p \lim n \left(\tilde{\mathbf{W}}' \tilde{\mathbf{W}} \right)^{-1} \tilde{\mathbf{W}}' \Sigma \tilde{\mathbf{W}} \left(\tilde{\mathbf{W}}' \tilde{\mathbf{W}} \right)^{-1}. \tag{6.5}$$

Hence it is asymptotically valid to test the null hypothesis by comparing sample values of a Wald-type statistic defined by

$$\mathcal{W} = n \tilde{\mathbf{y}}' \left[\ddot{C}_{\gamma\gamma} \right]^{-1} \tilde{\mathbf{y}}, \tag{6.6}$$

with critical values from the $\chi^2(q)$ distribution, provided that $\ddot{C}_{\gamma\gamma}$ is consistent for $C_{\gamma\gamma}$ of (6.5) when the null is true.

There are many asymptotically valid candidates for the heteroskedasticity-consistent covariance matrix estimator (HCCME) $\ddot{C}_{\gamma\gamma}$. In order to discuss some important versions of the HCCME, it is useful to write its general form as

$$\ddot{C}_{\gamma\gamma} = n \left(\tilde{\mathbf{W}}' \tilde{\mathbf{W}} \right)^{-1} \tilde{\mathbf{W}}' \mathbf{D} \tilde{\mathbf{W}} \left(\tilde{\mathbf{W}}' \tilde{\mathbf{W}} \right)^{-1}, \tag{6.7}$$

in which \mathbf{D} is an $n \times n$ diagonal matrix; so that $\mathbf{D} = \text{diag}(d_{11}, \dots, d_{nn})$. From (6.4), (6.6) and (6.7), the implied heteroskedasticity-robust Wald statistic is then

$$\mathcal{W} = \hat{\mathbf{u}}' \tilde{\mathbf{W}} \left[\tilde{\mathbf{W}}' \mathbf{D} \tilde{\mathbf{W}} \right]^{-1} \tilde{\mathbf{W}}' \hat{\mathbf{u}}. \tag{6.8}$$

Four well-known estimators of $C_{\gamma\gamma}$ are considered in MacKinnon and White (1985). These estimators are denoted by *HCO*, *HC1*, *HC2* and *HC3*. All are special cases of (6.7), with the diagonal elements of the matrix \mathbf{D} defined as follows:

$$d_{ii} = \tilde{u}_i^2, i = 1, \dots, n, \text{ for } HCO;$$

$$d_{ii} = \frac{n}{n - k - q} \tilde{u}_i^2, i = 1, \dots, n, \text{ for } HC1;$$

$$d_{ii} = (1 - h_{ii})^{-1} \tilde{u}_i^2, i = 1, \dots, n, \text{ for } HC2,$$

in which h_{ij} is a typical leverage value for the regressor matrix of (6.3); and

$$d_{ii} = (1 - h_{ii})^{-2} \tilde{u}_i^2, i = 1, \dots, n, \text{ for HC3.}$$

The *HC0* version is derived from the suggestion made in White (1980) and *HC1* incorporates a simple degrees of freedom adjustment. The versions given by *HC2* and *HC3* use leverage values to modify White's original expression; see Chesher and Jewitt (1987) for evidence on the importance of leverage values for the properties of a HCCME.

A fifth estimator, which is denoted by *HC4*, is provided in Cribari-Neto (2004). Cribari-Neto's HCCME uses the ratio of each leverage value h_{ii} to the sample average $\bar{h} = \frac{1}{n} \sum_j h_{jj} = (k+q)/n$ when modelling the effects of leverage values. More precisely, Cribari-Neto's estimator *HC4* is defined using the terms

$$d_{ii} = (1 - h_{ii})^{-\delta_i} \tilde{u}_i^2, \text{ where } \delta_i = \min\left(4, \frac{h_{ii}}{\bar{h}}\right), i = 1, \dots, n,$$

as the diagonal elements of **D**.

All five of these HCCME provide asymptotically valid inference in the presence of unspecified heteroskedasticity and also under homoskedasticity. However, it appears that the finite sample performance of test procedures can be greatly affected by the choice of HCCME. Several articles have been published in which the finite sample properties of heteroskedasticity-robust *t*-tests of a single restriction are studied by means of simulation experiments. In particular, the simulation results in Long and Ervin (2000) and MacKinnon and White (1985) are often cited. Both studies provide evidence that, when statistical significance is judged using asymptotic theory critical values, White's original proposal that *HC0* be used leads to *t*-tests that reject true null hypotheses too frequently and that *HC3* should be used in preference to *HC1* and *HC2*. The common conclusion is that *HC3*-based tests should be used routinely without screening tests for heteroskedasticity.

In subsequent work, Cribari-Neto compares his estimator *HC4* with *HC0* and *HC3*; see Cribari-Neto (2004). The focus is again on the finite sample behaviour of heteroskedasticity-consistent *t*-tests. Cribari-Neto obtains results from simulation experiments which indicate that *HC4* is superior to *HC3* and that the standard *HC0* form often produces estimated significance levels that are 2 or 3 times greater than the desired level.

Clearly this body of evidence on the relative merits of alternative versions of the HCCME is of interest to applied workers wishing to use

heteroskedasticity-robust tests. However, there is also an important contribution contained in Davidson and MacKinnon (1985a). Davidson and MacKinnon, like the previously mentioned authors, investigate the behaviour of quasi- t tests involving the use of asymptotic critical values. However, Davidson and MacKinnon extend the analysis to include examination of the importance of the choice between restricted and unrestricted estimation when constructing the HCCME. The expressions for $HC0$, $HC1$, $HC2$, $HC3$ and $HC4$ above use results from unrestricted estimation, that is, OLS for the regression model of (6.3). If these unrestricted results are replaced by those from the restricted estimation, that is, OLS for the model in (6.1), new versions of the HCCME are obtained. These versions are denoted by $HCR0$, $HCR1$ and so on. For example, the counterparts of the estimators studied in Long and Ervin (2000) and MacKinnon and White (1985) are derived by using (6.7), with the diagonal elements of the matrix D specified as:

$$d_{ii} = \hat{u}_i^2, i = 1, \dots, n, \text{ for } HCR0;$$

$$d_{ii} = \frac{n}{n-k} \hat{u}_i^2, i = 1, \dots, n, \text{ for } HCR1;$$

$$d_{ii} = (1 - h_{ii}^R)^{-1} \hat{u}_i^2, i = 1, \dots, n, \text{ for } HCR2,$$

in which h_{ii}^R is a typical leverage value for the regressor matrix of (6.1); and

$$d_{ii} = (1 - h_{ii}^R)^{-2} \hat{u}_i^2, i = 1, \dots, n, \text{ for } HCR3.$$

Davidson and MacKinnon carry out several simulation experiments from which the following findings emerge. When unrestricted estimation is used to define the HCCME, the ranking, from worst to best, based upon the behaviour of quasi- t tests, is $HC0$, $HC1$, $HC2$ and $HC3$, with $HC0$ sometimes having very substantial errors in rejection rates relative to desired significance levels. In contrast, the HCCME based upon restricted estimation leads to quasi- t tests that are well behaved in finite samples. Differences between tests based upon $HCR0$, $HCR1$, $HCR2$ and $HCR3$ are small, with $HCR0$ giving more reliable inferences than $HC3$, which is the best of the procedures based upon unrestricted estimation.

Further support for the use of restricted residuals when constructing t -tests that are heteroskedasticity-consistent is provided by simulation results in Flachaire (2005). Flachaire remarks that "tests based upon restricted residuals exhibit more power than those based upon unrestricted residuals" (Flachaire, 2005, p. 370). Flachaire also investigates

the relative merits of wild and pairs bootstraps in his simulation experiments. He finds that the Rademacher pick distribution of (5.25), which is recommended for general use in Davidson and Flachaire (2001, 2008), outperforms the pairs bootstrap and the other wild bootstraps that are considered in his study.

The results in Davidson and MacKinnon (1985a) on the advantages of using restricted residuals to obtain quasi- t tests are also reinforced by evidence from simulation experiments that are reported in Godfrey and Orme (2004). As well as examining heteroskedasticity-robust t -ratios, Godfrey and Orme generalize previous studies by considering heteroskedasticity-robust tests of several linear restrictions. They find that, even with restricted estimation being used to define the HCCME, asymptotic critical values (from the appropriate χ^2 distribution) do not give reliable tests, with the corresponding estimated finite sample significance levels being too small. After considering the results of simulation experiments, Godfrey and Orme propose that a wild bootstrap be used to control the significance level when the null hypothesis consists of several linear restrictions. Additional support for this proposal is found in Godfrey (2006).

Godfrey carries out experiments in which several restrictions are under test and errors are neither Normally distributed nor homoskedastic; see Godfrey (2006, section 3) for details. He finds that asymptotic critical values from a χ^2 distribution do not give accurate control of finite sample significance levels, with *HCR*-type versions of the HCCME tending to produce undersized tests. Godfrey then investigates the usefulness of wild bootstrap tests, using the Rademacher distribution $\mathcal{D}_{\epsilon,6}$ discussed in Section 5.2.3. After examining the results of a number of experiments, Godfrey concludes that this wild bootstrap approach provides much better behaved tests, with differences between the rejection rates associated with the various *HCR*-type estimates not being substantial. Overall, it appears that it would be reasonable to conjecture that using *HCR0* with a wild bootstrap will produce heteroskedasticity-consistent tests with reasonably good finite sample performance. The application of such a procedure to the problem of testing for specification errors in regression models will now be discussed.

6.2.2. Simulation experiments

The simulation experiments and results discussed in this subsection are taken from Godfrey and Orme (2002b). The Cobb-Douglas production function described in Section 1.5.1 is used to illustrate the implementation of heteroskedasticity-robust versions of the RESET test. Thus the

null model is

$$y_i = \beta_1 + \beta_2 x_{i2} + \beta_3 x_{i3} + u_i, i = 1, \dots, n, \quad (6.9)$$

in which x_{i2} and x_{i3} are the logs of data on labour and capital, respectively, taken from Greene (2008). The OLS predicted values, which are employed to obtain the RESET statistic, are denoted by

$$\hat{y}_i = \hat{\beta}_1 + \hat{\beta}_2 x_{i2} + \hat{\beta}_3 x_{i3}, i = 1, \dots, n.$$

As is quite common in applied work, the test variables are given by \hat{y}_i^2 , \hat{y}_i^3 and \hat{y}_i^4 ; so that the artificial alternative of the RESET test is a special case of (6.3) with $q = 3$, that is,

$$y_i = \beta_1 + \beta_2 x_{i2} + \beta_3 x_{i3} + \gamma_1 \hat{y}_i^2 + \gamma_2 \hat{y}_i^3 + \gamma_3 \hat{y}_i^4 + u_i, i = 1, \dots, n, \quad (6.10)$$

Heteroskedasticity is permitted and the errors of (6.9) are determined by

$$u_i = \sigma_i v_i, i = 1, \dots, n, \quad (6.11)$$

in which σ_i denotes a typical standard deviation and the terms v_i are IID random variables, with zero mean and variance equal to one, $i = 1, \dots, n$. The desired significance level of the heteroskedasticity-consistent RESET test is $\alpha_d = 5$ per cent.

In the experiments carried out in Godfrey and Orme (2002b), the regression coefficients of the mean function of (6.9) are set equal to the corresponding OLS estimates in Table 5.2 of Greene (2008, p. 91). The largest sample size used in the experiments is $n = 108$. There are only 27 observations in the original data set in Greene (2008) and so data for regressors in the mean function are reused, with

$$x_{ij} = x_{i+27,j} = x_{i+54,j} = x_{i+81,j}; i = 1, \dots, 27 \text{ and } j = 1, 2.$$

This strategy for obtaining regressor values is often adopted in simulation studies, see, for example, Cribari-Neto and Zarkos (1999) for references. Given simulation-world regression coefficients and data for the regressors, it only remains to specify how to obtain errors u_i and the value of the sample size n .

A skedastic function to determine values of σ_i and a distribution for v_i are both required for the error term u_i of (6.11), $i = 1, \dots, n$. Seven skedastic functions, which are described below, are used to study the behaviour of the heteroskedasticity-robust RESET test. Every skedastic

function is combined with three error distributions for v_i ; namely, standardized versions of Normal, $t(5)$ and $\chi^2(2)$ distributions. Each of these $7 \times 3 = 21$ error models is used with the mean function in (6.9) to form a data generation process (DGP). The 63 experiments in Godfrey and Orme (2002b) are then defined by specifying 3 values of the sample size for each DGP.

The seven skedastic models and associated sample sizes are as follows. First, the case of homoskedasticity is considered by using $\sigma_i = s, i = 1, \dots, n$, where s is the standard error of regression given in Table 5.2 of Greene (2008, p. 91) and $n = 27, 54, 108$. The next two schemes are special cases of the structural change model used in MacKinnon and White (1985), which requires n to be even. For $n = 26$ (not 27), 54 and 108, the standard deviations are given by

$$\sigma_i = s \text{ for } i = 1, \dots, n/2,$$

and

$$\sigma_i = c_{mw}s \text{ for } i = n/2 + 1, \dots, n,$$

with $c_{mw} = (2, 4)$; see Godfrey and Orme (2002b) for further discussion and comments on the choice of parameter values for these and other schemes. Next, the fourth and fifth types of variance model are given by

$$\sigma_i^2 = \lambda_1 + \lambda_2 x_{i2}^2 + \lambda_3 x_{i3}^2, i = 1, \dots, n,$$

in which $\lambda_1 = s^2$, $\lambda_2 = (0.0015, 0.0018)$ and $\lambda_3 = 0.5\lambda_2$, with $n = 27, 54, 108$. These two skedastic models can be thought of as arising from random coefficient models. The last two patterns of heteroskedasticity are generated by the multiplicative model

$$\sigma_i = s \exp(c_{mh}(\beta_1 + \beta_2 x_{i2} + \beta_3 x_{i3})), i = 1, \dots, n, \quad (6.12)$$

with $c_{mh} = (0.35, 0.55)$ and $n = 27, 54, 108$.

Given the specification of the DGP and the required sample size, Steps 1 to 6 are repeated for $R = 25,000$ replications in Godfrey and Orme (2002b).

Heteroskedasticity-robust RESET test: Step 1

Generate the "actual" data y_1, \dots, y_n , using (6.9) and (6.11) with the relevant skedastic function and error distribution.

Heteroskedasticity-robust RESET test: Step 2

Estimate the relationships that correspond to (6.9) and (6.10) by OLS. Let the OLS estimates, predicted values and residuals for (6.9) be denoted by $\hat{\beta}_j$, \hat{y}_i and \hat{u}_i , respectively; $i = 1, \dots, n$ and $j = 1, 2, 3$.

Heteroskedasticity-robust RESET test: Step 3

Use the OLS results from Step 2 to compute the heteroskedasticity-robust test of (6.9) against (6.10), with the covariance matrix estimator being *HCR0*. Let the test statistic be denoted by \mathcal{W}_R . Under the null hypothesis, \mathcal{W}_R is asymptotically distributed as $\chi^2(3)$.

Steps 4 and 5 are for the generation and analysis of wild bootstrap samples, respectively. In Godfrey and Orme (2002b), these steps are repeated $B = 399$ times before proceeding to Step 6.

Heteroskedasticity-robust RESET test: Step 4

Generate a bootstrap sample of n observations for each of the wild bootstrap schemes to be considered. The bootstrap data can be written as

$$y_{ij}^* = \hat{y}_i + \hat{u}_i \epsilon_{ij}, \quad (6.13)$$

in which \hat{y}_i and \hat{u}_i , $i = 1, \dots, n$, are from Step 2 and the terms $\epsilon_{1j}, \dots, \epsilon_{nj}$ are IID drawings from the pick distribution $\mathcal{D}_{\epsilon,j}$, $j = 1, \dots, 6$, defined by (5.20) to (5.25) in Chapter 5.

Heteroskedasticity-robust RESET test: Step 5

For each of the six pick distributions being examined, use the bootstrap data of (6.13) to estimate the regressions corresponding to (6.9) and (6.10). Apply the *HCR0* estimator to obtain the bootstrap counterpart of the test statistic \mathcal{W}_R calculated in Step 3. Let the bootstrap test statistic obtained using the pick distribution $\mathcal{D}_{\epsilon,j}$ be denoted by \mathcal{W}_{Rj}^* , $j = 1, \dots, 6$.

Once Steps 4 and 5 have been carried out B times, the p -value of the test statistic of Step 3 can be estimated, using the bootstrap values from Step 5, in the sixth step.

Heteroskedasticity-robust RESET test: Step 6

Having completed the process of obtaining wild bootstrap data and the associated test statistics, the bootstrap p -values of \mathcal{W}_R can be derived for each pick distribution. In the analysis in Godfrey and Orme (2002b), these p -values are calculated according to (2.11), rather than (2.12), of

Section 2.2.2, that is,

$$\tilde{p}_j = \frac{\left[\#(\mathcal{W}_{Rj}^* \geq \mathcal{W}_R) + 1 \right]}{B + 1}, j = 1, \dots, 6,$$

with $B = 399$. The decision rule is then that, for the pick distribution $\mathcal{D}_{\epsilon,j}$, the null hypothesis that (6.9) contains no specification errors is rejected if \tilde{p}_j is not greater than the desired significance level of 5 per cent, $j = 1, \dots, 6$.

Heteroskedasticity-robust RESET test: Step 7

Once Steps 1 to 6 have been repeated $R = 25,000$ times, the actual significance level of the heteroskedasticity-robust RESET test can be estimated for each pick distribution as the proportion of replications in which the null hypothesis is rejected in Step 6.

The results reported in Godfrey and Orme (2002b) concerning the estimated finite sample significance levels of the heteroskedasticity-robust version of Ramsey's RESET test can be summarized as follows. The use of the asymptotic critical value from the $\chi^2(3)$ distribution produces a tendency to underrejection relative to the required significance level of 5 per cent, but, as expected, the quality of the approximation tends to get better as n increases. The wild bootstrap tests enjoy varying degrees of success in controlling significance levels. There is, however, no reason to discuss each approach in detail because the Rademacher pick distribution $\mathcal{D}_{\epsilon,6}$ in (5.25) clearly gives the best overall performance and does extremely well with 59 of its 63 rejection frequencies being consistent with the claim that the actual significance level is between 4.5 per cent and 5.5 per cent; see Godfrey and Orme (2002b). The estimates derived using the multiplicative heteroskedasticity model (6.12) are given in Table 6.1 to illustrate the properties of the heteroskedasticity-robust RESET tests based upon the asymptotic critical value and the six pick distributions. Similar results are obtained with the other variance models adopted in Godfrey and Orme (2002b).

Overall the evidence from these experiments and others described in the literature indicates the possibility of using the Rademacher pick distribution of (5.25) in a wild bootstrap to obtain good control of the significance levels of heteroskedasticity-robust checks for specification errors when applying regression analysis. It is recommended that the HCCME used to obtain the robust form of the test statistic should be calculated using the residuals from restricted estimation, despite the implied inconvenience of recomputing the HCCME for each

Table 6.1 Estimated significance levels of asymptotic and wild bootstrap (WBS) versions of HCCME-based RESET test, using $HCR0$ and $\alpha_d = 5$ per cent, for the multiplicative heteroskedasticity model

error distribution: value of c_{MH} in (6.12):	$N(0, 1)$		$t(5)$		$\chi^2(2)$	
	0.35	0.55	0.35	0.55	0.35	0.55
a. $n = 27$						
asymptotic test	2.81	2.87	2.57	2.62	2.34	2.49
WBS test and $\mathcal{D}_{\epsilon,1}$	6.56	7.66	6.65	7.64	5.56	6.49
WBS test and $\mathcal{D}_{\epsilon,2}$	6.62	7.38	6.54	7.24	5.61	6.58
WBS test and $\mathcal{D}_{\epsilon,3}$	6.68	7.55	6.45	7.39	5.64	6.63
WBS test and $\mathcal{D}_{\epsilon,4}$	4.66	5.21	4.43	5.07	4.02	4.65
WBS test and $\mathcal{D}_{\epsilon,5}$	5.48	6.36	5.35	6.14	4.77	5.71
WBS test and $\mathcal{D}_{\epsilon,6}$	5.24	5.88	5.19	5.70	4.67	5.24
b. $n = 54$						
asymptotic test	3.57	4.23	3.01	3.40	2.74	3.36
WBS test and $\mathcal{D}_{\epsilon,1}$	6.34	8.18	6.42	7.77	5.05	6.95
WBS test and $\mathcal{D}_{\epsilon,2}$	6.23	7.78	6.13	7.10	5.33	6.60
WBS test and $\mathcal{D}_{\epsilon,3}$	6.40	8.14	6.12	7.32	5.40	6.87
WBS test and $\mathcal{D}_{\epsilon,4}$	4.70	6.09	4.50	5.47	4.17	5.02
WBS test and $\mathcal{D}_{\epsilon,5}$	5.59	7.39	5.22	6.54	4.67	6.19
WBS test and $\mathcal{D}_{\epsilon,6}$	4.91	5.74	4.85	5.28	4.31	5.09
c. $n = 108$						
asymptotic test	4.61	5.02	3.69	4.40	3.87	4.88
WBS test and $\mathcal{D}_{\epsilon,1}$	6.64	7.76	6.15	7.75	5.64	7.65
WBS test and $\mathcal{D}_{\epsilon,2}$	6.46	7.24	5.81	6.95	6.01	7.41
WBS test and $\mathcal{D}_{\epsilon,3}$	6.64	7.65	5.96	7.20	5.90	7.62
WBS test and $\mathcal{D}_{\epsilon,4}$	5.42	6.05	4.62	5.88	4.80	5.68
WBS test and $\mathcal{D}_{\epsilon,5}$	6.15	7.25	5.43	6.89	5.32	6.87
WBS test and $\mathcal{D}_{\epsilon,6}$	5.37	5.51	4.84	5.41	5.01	5.80

Notes: Each estimate is derived from 25,000 replications and **bold** font denotes that the estimate is consistent with level being between 4.5 per cent and 5.5 per cent.

null hypothesis under consideration. This recommendation is consistent with conclusions drawn by Davidson and Flachaire after carrying out simulation experiments for heteroskedasticity-robust t -tests; see Davidson and Flachaire (2008, p. 168).

6.3. Bootstrapping heteroskedasticity-robust autocorrelation tests for dynamic models

Conditional heteroskedasticity is a common feature of financial and macroeconomics time series data. When such heteroskedasticity is

present, standard asymptotic tests for autocorrelation in the errors of a dynamic regression model are inappropriate. In this section, a recursive wild bootstrap technique is used to derive a heteroskedasticity-robust version of the Breusch-Godfrey *autocorrelation test*. The finite sample performance of the robust bootstrap test is discussed in the light of simulation results, which are taken from Godfrey and Tremayne (2005).

6.3.1. The forms of test statistics

Suppose that time series data are to be used in an OLS-based analysis of a dynamic linear regression model. Using t as the subscript for the observations, this model is written as

$$y_t = \sum_{j=1}^p y_{t-j}\alpha_j + \sum_{j=1}^k x_{tj}\beta_j + u_t, \quad (6.14)$$

so that the regressor set consists of p lagged values of the dependent variable and k exogenous variables. It is assumed that n observations are available for the OLS estimation of the parameters of (6.14). An appropriate generalization of (6.1) is, therefore,

$$\mathbf{y} = \mathbf{Y}\boldsymbol{\alpha} + \mathbf{X}\boldsymbol{\beta} + \mathbf{u}, \quad (6.15)$$

in which: \mathbf{y} , \mathbf{X} , $\boldsymbol{\beta}$ and \mathbf{u} are as defined in (6.1); a typical element of the $n \times p$ matrix \mathbf{Y} is y_{t-j} , $t = 1, \dots, n$ and $j = 1, \dots, p$; and $\boldsymbol{\alpha}' = (\alpha_1, \dots, \alpha_p)$.

It is necessary to make some conventional assumptions in order to appeal to asymptotic theory for OLS-based analysis of (6.15). First, the model is assumed to have the property of dynamic stability with the coefficients in $\boldsymbol{\alpha}' = (\alpha_1, \dots, \alpha_p)$ being such that the roots of

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

are all strictly inside the unit circle. Second, the exogenous regressors are assumed to be asymptotically cooperative with

$$p \lim n^{-1} \mathbf{X}'\mathbf{X} = \boldsymbol{\Xi}_{xx},$$

being a finite positive-definite matrix. It remains to specify assumptions about the joint distribution of the errors of (6.15).

Under classical assumptions for OLS-based inference, the errors of (6.15) would be taken to be $NID(0, \sigma^2)$. Acceptance of the arguments in modern texts, for example, Stock and Watson (2007), leads to the

relaxation of both homoskedasticity and Normality, leaving only independence remaining as a restriction on the joint distribution of the errors. The general advice given in textbooks is that the OLS estimators of α and β are inconsistent if autocorrelation of the errors is permitted. There are, however, special cases in which OLS estimators remain consistent when the regressors include lagged values of the dependent variable and the errors are autocorrelated. For example, the dynamic model with MA(1) errors consisting of

$$y_t = \alpha_4 y_{t-4} + \beta_1 + \beta_2 x_{t2} + u_t, \quad |\alpha_4| < 1,$$

and

$$u_t = \epsilon_t + \theta_1 \epsilon_{t-1}, \quad |\theta_1| < 1, \quad \epsilon_t \text{ NID}(0, \sigma_\epsilon^2),$$

does not imply the inconsistency of OLS.

The key feature of such special cases is that the longest lag for which there is a non-zero error autocorrelation is less than the shortest lag on the lagged dependent variables that are included as regressors; see Phillips (1956) and Wise (1957). In general, there is no good reason to believe that such a restrictive condition will be satisfied and it seems reasonable to assume the inconsistency of OLS when estimating a dynamic regression model by OLS in the presence of unspecified forms of autocorrelation. Consequently the null hypothesis of independent errors remains of interest, even when the errors are allowed to exhibit heteroskedasticity and non-Normality of unknown forms. There is, therefore, a need for tests for autocorrelation that are asymptotically valid in the presence of unspecified heteroskedasticity.

A standard test, which is based upon the assumption of homoskedasticity, is the Lagrange Multiplier (LM) procedure given in Breusch (1978) and Godfrey (1978). Like the RESET test in the previous section, the LM test uses results from the OLS estimation of the null model to generate an artificial alternative model. Let the OLS estimators for α and β in (6.15) be denoted by $\hat{\alpha}$ and $\hat{\beta}$, respectively, and the residual vector be written as

$$\hat{\mathbf{u}} = \mathbf{y} - \mathbf{Y}\hat{\alpha} - \mathbf{X}\hat{\beta} = (\hat{u}_1, \dots, \hat{u}_n)'$$

The homoskedasticity-only version of the LM test, with the alternative hypothesis being either the AR(q) or the MA(q) scheme, can be implemented as an asymptotically valid F -test of $H_{LM} : \boldsymbol{\gamma} = \mathbf{0}_q$ in the augmented model

$$\mathbf{y} = \mathbf{Y}\alpha + \mathbf{X}\beta + \hat{\mathbf{U}}\boldsymbol{\gamma} + \mathbf{u}, \quad (6.16)$$

where $\hat{\mathbf{U}}$ is a $n \times q$ matrix of lagged OLS residuals with typical element \hat{u}_{t-j} , $t = 1, \dots, n$ and $j = 1, \dots, q$. (In order to have the same sample size for (6.15) and (6.16), it is assumed that $\hat{u}_{t-j} = 0$ when $t-j$ is not positive; this assumption has no impact on the asymptotic theory for the standard test.) The OLS estimators for the artificial alternative are denoted by $\hat{\alpha}$, $\hat{\beta}$ and $\hat{\gamma}$. The corresponding OLS residual vector is $\hat{\mathbf{u}}$. Following previous usage, the elements of $\hat{\mathbf{u}}$ and $\tilde{\mathbf{u}}$ are called the restricted and unrestricted residuals, respectively.

When the restrictive assumption of homoskedasticity is relaxed, the LM test must be based upon a heteroskedasticity-robust test of the joint significance of the elements of $\hat{\gamma}$. Under the null hypothesis $H_{LM} : \gamma = \mathbf{0}_q$,

$$\tilde{\gamma} = \left(\tilde{\mathbf{U}}' \tilde{\mathbf{U}} \right)^{-1} \tilde{\mathbf{U}}' \mathbf{u},$$

in which $\tilde{\mathbf{U}}$ is the matrix of residuals from the OLS regression of $\hat{\mathbf{U}}$ on \mathbf{Y} and \mathbf{X} . The general formulae of (6.6) and (6.7) can now be applied with $\tilde{\mathbf{W}}$ in the latter equation being replaced by $\tilde{\mathbf{U}}$ and the diagonal matrix \mathbf{D} being defined by one of the expressions in Section 6.2.1. This substitution leads to the general expression for the heteroskedasticity-consistent LM statistic as

$$LM_{HR} = \hat{\mathbf{u}}' \tilde{\mathbf{U}} \left(\tilde{\mathbf{U}}' \mathbf{D} \tilde{\mathbf{U}} \right)^{-1} \tilde{\mathbf{U}}' \hat{\mathbf{u}}, \quad (6.17)$$

which is asymptotically distributed as $\chi^2(q)$, when the null hypothesis is true.

The discussion in the previous section suggests that restricted residuals be used to calculate \mathbf{D} , that is, an *HCR*-type version of the HCCME should be employed, and also that a wild bootstrap should be considered as an alternative to using the limit null distribution for inference. Wild bootstrap samples of size n can be generated according to the recursive scheme

$$y_t^* = \sum_{j=1}^p y_{t-j}^* \hat{\alpha}_j + \sum_{j=1}^k x_{tj} \hat{\beta}_j + u_t^*, \quad t = 1, \dots, n, \quad (6.18)$$

in which the actual data are used to supply the initial values y_{t-j}^* for $t-j \leq 0$ and the bootstrap errors u_t^* are defined by

$$u_t^* = \hat{u}_t \epsilon_t, \quad t = 1, \dots, n, \quad (6.19)$$

in which the terms ϵ_t are IID drawings from a valid pick distribution. Evidence on the usefulness of the general approach of using a wild bootstrap to assess the statistical significance of LM_{HR} in (6.17) and on the relative merits of the various pick distributions that have been proposed in the literature can be sought using simulation experiments.

6.3.2. Simulation experiments

The experiments discussed in this subsection are taken from Godfrey and Tremayne (2005) in which simulation models proposed in Dezhbakhsh (1990) and Dezhbakhsh and Thursby (1995) are used. The dynamic regression model upon which experiments are based is

$$y_t = \alpha_1 y_{t-1} + \alpha_2 y_{t-2} + \beta_1 + \beta_2 x_t + u_t, t = 1, \dots, n, \quad (6.20)$$

in which x_t is a scalar variable and n equals 40 or 80. The finite sample significance levels of versions of heteroskedasticity-consistent tests for autocorrelation are estimated in Godfrey and Tremayne (2005), with the alternative being a fourth-order scheme, as might be appropriate when data are quarterly. Thus (6.20) is tested against

$$y_t = \alpha_1 y_{t-1} + \alpha_2 y_{t-2} + \beta_1 + \beta_2 x_t + \sum_{j=1}^4 \gamma_j \hat{u}_{t-j} + \text{error}, \quad (6.21)$$

with the test variables \hat{u}_{t-j} being lagged residuals from the OLS estimation of (6.20). The Wald-type statistic for testing $\gamma_1 = \dots = \gamma_4 = 0$ is calculated using the HCCME obtained by using squared restricted residuals for the diagonal elements of \mathbf{D} in (6.17), that is, $d_{tt} = \hat{u}_t^2, t = 1, \dots, n$.

Godfrey and Tremayne compare asymptotic tests with wild bootstrap tests. They also provide results for the comparison of two pick distributions. The first is the widely-used two-point distribution defined by

$$\begin{aligned} \epsilon &= -(\sqrt{5} - 1)/2 \text{ with probability } (\sqrt{5} + 1)/(2\sqrt{5}) \\ &= (\sqrt{5} + 1)/2, \text{ otherwise.} \end{aligned}$$

This distribution is denoted by $\mathcal{D}_{\epsilon,4}$ in (5.23) of Chapter 5. The second pick distribution is the Rademacher distribution $\mathcal{D}_{\epsilon,6}$ in (5.25) of Chapter 5, that is,

$$\begin{aligned} \epsilon &= 1 \text{ with probability } 0.5 \\ &= -1, \text{ otherwise.} \end{aligned}$$

Godfrey and Tremayne remark that other pick distributions, as described in Chapter 2, lead to results that are poorer than those derived from $\mathcal{D}_{\epsilon,4}$ and $\mathcal{D}_{\epsilon,6}$.

The choice of regression parameters in (6.20) follows Dezhbakhsh (1990) and Dezhbakhsh and Thursby (1995). The values of (α_1, α_2) are $(0.5, 0.3)$, $(0.7, -0.2)$, $(1.0, -0.2)$, $(1.3, -0.5)$, $(0.9, -0.3)$ and $(0.6, 0.2)$, which all satisfy the conditions for dynamic stability. These values for (α_1, α_2) are intended by Dezhbakhsh and Thursby to be typical of those observed in applied work. In all experiments, $\beta_1 = \beta_2 = 1$. Two methods are adopted to provide values on the exogenous regressor. First, the exogenous variable x_t is generated according to the first order autoregression

$$x_t = \psi_x x_{t-1} + v_t, \quad (6.22)$$

with v_t being $NID(0, \sigma_v^2)$, $\psi_x = 0.5$ or 0.9 , and σ_v^2 selected, given the value of ψ_x , so that $Var(x_t) = 1$. The starting value x_0 is a drawing from a standard Normal distribution. Second, Godfrey and Tremayne use a standardized version of the log of quarterly GDP in the UK as the exogenous regressor in (6.20).

With conditional heteroskedasticity permitted, the error terms u_t of (6.20) can be written as

$$u_t = \sqrt{\sigma_t^2} \zeta_t, \quad (6.23)$$

where σ_t^2 denotes a conditional variance and, under the null hypothesis, the terms ζ_t are IID with zero mean and variance equal to one. Various distributions for ζ_t are used. The Normal distribution serves as a benchmark and standardized forms of the $t(5)$ and $\chi^2(8)$ distributions are also employed. The $t(5)$ distribution is used, following Gonçalves and Kilian (2004), to investigate robustness of the wild bootstrap methods to symmetric non-Normal error distributions. The $\chi^2(8)$ distribution is used to provide evidence on the effects of marked skewness; see Godfrey and Tremayne (2005, p. 385).

The final component required to derive a typical error u_t , after drawing ζ_t , is the conditional standard deviation $\sqrt{\sigma_t^2}$. The following five specifications for variance schemes are used in Godfrey and Tremayne (2005). First, the errors are homoskedastic with

$$\sigma_t^2 = \sigma^2, t = 1, \dots, n, \quad (6.24)$$

σ^2 being set equal to 1 or 10. Second, the ARCH(1) process is used with

$$\sigma_t^2 = \phi_0 + \phi_1 u_{t-1}^2, \quad (6.25)$$

in which $\phi_0 = \sigma^2/(1 - \phi_1)$, $\phi_1 = 0.4$ or 0.8 , and σ^2 is defined as for (6.24). In the third model, the widely-used GARCH(1, 1) specification is used in the form

$$\sigma_t^2 = \psi_0 + \psi_1 u_{t-1}^2 + \psi_2 \sigma_{t-1}^2, \quad (6.26)$$

where $\psi_0 = 1$, $\psi_1 = 0.1$, and $\psi_2 = 0.8$; see Bollerslev (1986). The values of ψ_1 and ψ_2 are similar to those reported in empirical work.

Given the widespread use of quarterly data in applied work and the nature of the alternative hypothesis that underpins the LM test, it seems appropriate to examine seasonal schemes of heteroskedasticity. Consideration is, therefore, given to a fourth-order model taken from Engle (1982). As in (38) of Engle (1982, p. 1002), conditional variances σ_t^2 are written as

$$\sigma_t^2 = \phi_0 + \phi_1(0.4u_{t-1}^2 + 0.3u_{t-2}^2 + 0.2u_{t-3}^2 + 0.1u_{t-4}^2), \quad (6.27)$$

in which ϕ_0 and ϕ_1 are as defined for (6.25). The fifth and final model to be adopted has unconditional quarterly heteroskedasticity and can be written as

$$(\omega_1^2, \omega_2^2, \omega_3^2, \omega_4^2) = \left(\frac{2\sigma^2}{(1+c)}, \frac{2\sigma^2}{(1+c)}, \frac{2c\sigma^2}{(1+c)}, \frac{2c\sigma^2}{(1+c)} \right), \quad (6.28)$$

in which ω_j^2 denotes the variance in quarter j , the average of these terms is σ^2 , as specified in (6.24), and c equals 4 or 9. Model (6.28) is similar to that used in Burrige and Taylor (2001, p. 104).

All tests are carried out with a desired significance level of 5 per cent. The estimates for the asymptotic test are derived using the upper 5 per cent critical value from the $\chi^2(4)$ distribution. For the wild bootstrap tests, p -values are calculated using $B = 399$ bootstrap samples, according to (2.11), rather than (2.12), of Section 2.2.2. The number of replications for each experiment is $R = 25,000$. Consideration of standard errors and the approximate distribution of estimators of rejection probabilities implies that estimated significance levels that are either less than 4.28 or greater than 5.74 provide strong evidence against the claim that the true significance level is in the range 4.5 per cent to 5.5 per cent.

Given the specification of the simulation experiment DGP, the choice of a pick distribution (either $\mathcal{D}_{\epsilon,4}$ or $\mathcal{D}_{\epsilon,6}$) and the required sample size, there are $R = 25,000$ repetitions of Steps 1 to 6, which can be described as follows.

Heteroskedasticity-robust serial correlation test: Step 1

Use the selected version of (6.22) or the transformed UK GDP data to obtain n observations on the exogenous regressor. Generate the “actual” data y_1, \dots, y_n , using (6.20) and (6.23) with the selected combination of skedastic function and error distribution.

Heteroskedasticity-robust serial correlation test: Step 2

Estimate the relationships that correspond to (6.20) and (6.21) by OLS. Let the OLS estimated coefficients for (6.20) be denoted by $\hat{\alpha}_1$, $\hat{\alpha}_2$, $\hat{\beta}_1$ and $\hat{\beta}_2$, with associated predicted values and residuals denoted by \hat{y}_t and \hat{u}_t , respectively; $t = 1, \dots, n$.

Heteroskedasticity-robust serial correlation test: Step 3

Use the OLS results from Step 2 to compute the heteroskedasticity-robust Wald test of (6.20) against (6.21), with the covariance matrix estimator being $HCR0$. Let the test statistic be denoted by LM_{HR} . Under the null hypothesis, LM_{HR} is asymptotically distributed as $\chi^2(4)$.

Steps 4 and 5 are for the generation and analysis of wild bootstrap samples, respectively. These steps are repeated $B = 399$ times in Godfrey and Tremayne (2005) before proceeding to Step 6.

Heteroskedasticity-robust serial correlation test: Step 4

Generate a bootstrap sample of n observations using the pick distribution chosen for the experiment. The bootstrap data are obtained using a recursive wild bootstrap scheme, with

$$y_t^* = \hat{\alpha}_1 y_{t-1}^* + \hat{\alpha}_2 y_{t-2}^* + \hat{\beta}_1 + \hat{\beta}_2 x_t + u_t^*, t = 1, \dots, n, \quad (6.29)$$

and

$$u_t^* = \hat{u}_t \epsilon_t, t = 1, \dots, n, \quad (6.30)$$

in which \hat{u}_t , $t = 1, \dots, n$, are the restricted residuals from Step 2 and the terms $\epsilon_1, \dots, \epsilon_n$ are IID drawings from the selected pick distribution, that is, either $\mathcal{D}_{\epsilon,4}$ in (5.23) or $\mathcal{D}_{\epsilon,6}$ in (5.25).

Heteroskedasticity-robust serial correlation test: Step 5

For the pick distribution being examined, use the bootstrap data of (6.29) to estimate the regressions corresponding to (6.20) and (6.21). Apply the $HCR0$ estimator to derive the bootstrap counterpart of the test statistic LM_{HR} calculated in Step 3. Let the bootstrap test statistic obtained be denoted by LM_{HR}^* .

Heteroskedasticity-robust serial correlation test: Step 6

After repeating Steps 4 and 5 B times, the bootstrap p -value of LM_{HR} from Step 3 can be derived, following Godfrey and Tremayne (2005), as

$$\tilde{p}_{LM} = \frac{[\#(LM_{HR}^* \geq LM_{HR}) + 1]}{B + 1},$$

with $B = 399$. The decision rule is to reject the null hypothesis if $\tilde{p}_{LM} \leq 5$ per cent.

Heteroskedasticity-robust serial correlation test: Step 7

Once Steps 1 to 6 have been carried out for the complete set of $R = 25,000$ replications, the finite sample significance level of the heteroskedasticity-robust LM test can be estimated by the proportion of replications in which the null hypothesis is rejected in Step 6.

Steps 1 to 7 can be carried out for all the required combinations of simulation-world DGP, sample size and pick distribution for the wild bootstrap in order to build up a body of evidence on finite sample properties of robust checks for autocorrelation in heteroskedastic dynamic regression models.

Consideration of the results given in Godfrey and Tremayne (2005) leads to the following conclusions about the finite sample behaviour of heteroskedasticity-robust LM tests for autocorrelation.

First, when the HCCME-based statistic LM_{HR} is compared with asymptotic critical values, there is not good control of finite sample significance levels and instead there is clear evidence of underrejection relative to the desired level.

Second, the use of the Rademacher pick distribution, that is, $\mathcal{D}_{\epsilon,6}$ of (5.25), leads to a well behaved test and produces results that are usually in closer agreement with the desired significance level than those obtained with the pick distribution $\mathcal{D}_{\epsilon,4}$ of (5.23). This finding supports the recommendations made in Flachaire (2005) for the implementation of heteroskedasticity-robust tests in regression analysis.

A sample of results, which is representative of the full set obtained by Godfrey and Tremayne, is used to illustrate their key findings

Table 6.2 Estimated significance levels of asymptotic and wild bootstrap (WBS) versions of HCCME-based LM_{HR} test, using *HRCRO*, with desired significance level of 5 per cent and $n = 40$

a. Homoskedasticity with $\sigma^2 = 1$ in (6.24)			
<i>Error distribution</i>	$N(0, 1)$	$t(5)$	$\chi^2(8)$
asymptotic test	3.69	2.88	3.24
WBS test and $\mathcal{D}_{\epsilon,4}$	5.98	5.48	5.55
WBS test and $\mathcal{D}_{\epsilon,6}$	5.04	4.41	4.61
b. ARCH(1) with $\phi_1 = 0.8$ in (6.25)			
<i>Error distribution</i>	$N(0, 1)$	$t(5)$	$\chi^2(8)$
asymptotic test	3.41	2.90	3.18
WBS test and $\mathcal{D}_{\epsilon,4}$	5.99	5.82	6.01
WBS test and $\mathcal{D}_{\epsilon,6}$	5.08	4.78	4.92
c. GARCH(1,1) with $\psi_1 = 0.1$ and $\psi_2 = 0.8$ in (6.26)			
<i>Error distribution</i>	$N(0, 1)$	$t(5)$	$\chi^2(8)$
asymptotic test	3.34	2.98	3.09
WBS test and $\mathcal{D}_{\epsilon,4}$	5.86	5.92	5.75
WBS test and $\mathcal{D}_{\epsilon,6}$	5.07	5.14	4.94
d. ARCH(4) with $\phi_0 = 0.1$ and $\phi_1 = 0.8$ in (6.27)			
<i>Error distribution</i>	$N(0, 1)$	$t(5)$	$\chi^2(8)$
asymptotic test	3.52	3.04	3.20
WBS test and $\mathcal{D}_{\epsilon,4}$	6.26	5.78	5.88
WBS test and $\mathcal{D}_{\epsilon,6}$	5.20	4.87	4.96
e. Seasonal variances with $c = 9$ in (6.28)			
<i>Error distribution</i>	$N(0, 1)$	$t(5)$	$\chi^2(8)$
asymptotic test	3.36	2.62	2.99
WBS test and $\mathcal{D}_{\epsilon,4}$	6.62	6.08	6.40
WBS test and $\mathcal{D}_{\epsilon,6}$	5.91	5.47	5.76

Notes: Each estimate is derived from 25,000 replications and **bold** font denotes that the estimate is consistent with the true significance level being between 4.5 per cent and 5.5 per cent, as indicated by the test given in Godfrey and Orme (2000, p. 75).

Source: From Godfrey and Tremayne, 2005, p. 389, Table 3.

on the performance of asymptotic and wild bootstrap versions of heteroskedasticity-robust LM tests. This sample of results is presented in Table 6.2. The estimated significance levels given in Table 6.2 show the potential value of the very simple two-point pick distribution $\mathcal{D}_{\epsilon,6}$ of (5.25); see Davidson and Flachaire (2001, 2008) for a detailed analysis of this pick distribution. In most cases, the estimate for the wild bootstrap

test that uses LM_{HR} , computed with the restricted residual-based HCCME $HCRO$, as the test statistic and the Rademacher distribution $\mathcal{D}_{\epsilon,6}$ as the pick distribution is consistent with the claim that the true significance level is within 0.5 per cent of the desired value of 5 per cent.

6.4. Bootstrapping heteroskedasticity-robust structural break tests with an unknown breakpoint

Section 4.4 contained a discussion of the problem of testing the null hypothesis that the coefficients of a regression model are constant against the alternative that there is a single breakpoint, which is unknown. As part of this discussion, evidence on the usefulness of basing tests on an IID-valid bootstrap scheme was summarized; see (4.44) and (4.45) for an example of such a scheme. However, the use of an IID-valid bootstrap may produce very misleading inferences when the actual data are generated by a regression model with heteroskedastic errors. In this section, the possibility of using bootstrap methods that are asymptotically valid under unknown forms of heteroskedasticity is discussed. Attention is also paid to the treatment of dynamic specification when setting up bootstrap schemes.

Under the unknown breakpoint alternative, the full set of n observations consists of two samples, with the conditional mean functions of the corresponding two populations differentiated by their regression parameter vectors. The sample sizes for these two samples, denoted by n_1 and n_2 , are unknown, but it is assumed that both are large enough to allow OLS estimation of the corresponding model. As in Andrews (1993), the unknown ratio n_1/n is denoted by π .

With lagged values of the dependent variable included in the regressor set, the null (restricted) model can be written as

$$y_t = \sum_{j=1}^p y_{t-j} \alpha_j + \sum_{j=1}^k x_{tj} \beta_j + u_t, \quad (6.31)$$

and the alternative (unrestricted) model as

$$y_t = \sum_{j=1}^p y_{t-j} \alpha_j + \sum_{j=1}^p \left[\mathbf{1}\left(\frac{t}{n} > \pi\right) y_{t-j} \right] \delta_j + \sum_{j=1}^k x_{tj} \beta_j + \sum_{j=1}^k \left[\mathbf{1}\left(\frac{t}{n} > \pi\right) x_{tj} \right] \gamma_j + u_t, \quad (6.32)$$

for $t = 1, \dots, n$, in which: $1(\frac{t}{n} > \pi)$ is the indicator function which is zero when the event " $\frac{t}{n} > \pi$ " is false and is one when this event is true; and the heteroskedastic errors u_t are distributed independently with zero means. While the exact value of π is unknown, it is assumed that a parameter space Π can be defined by specified lower and upper bounds, with

$$0 < \pi_1 \leq \pi \leq \pi_2 < 1.$$

In the classic article by Chow, it is assumed that n_1 is known to the researcher; see Chow (1960). Hence, under Chow's assumptions, π is known and so the values of the regressors of (6.32) are known. The F -statistic for testing the $p + k$ restrictions of

$$H_C : \delta_1 = \dots = \delta_p = \gamma_1 = \dots = \gamma_k = 0, \quad (6.33)$$

can, therefore, be calculated after OLS estimation of (6.31) and (6.32). Under IID errors, critical values from the $F(p+k, n-2p-2k)$ distribution are asymptotically valid. However, in the situation that is now being discussed, π is unknown and the absence of a priori information about the breakpoint implies that Chow's original test is no longer available.

All that is known is Π and so (6.31) can be tested against each of the possible alternatives defined by having a breakpoint parameter in Π . For example, if $n = 100$ and it is assumed that $0.15 \leq \pi \leq 0.85$, with $p+k < 15$, there are 71 possible alternatives, each of which can be used to test H_C of (6.33) by means of an F -statistic. It is now common for inference to be based upon the maximum of these F -statistics and the value of π that corresponds to the maximum F -statistic provides the estimate of the breakpoint. With general n and Π , the F -test statistic for testing (6.31) against (6.32) can be denoted by $F_n(\pi)$, $\pi \in \Pi$, and the largest of the set of statistics obtained by varying π between π_1 and π_2 is

$$SupF = \sup_{\pi \in \Pi} F_n(\pi).$$

As discussed in Section 4.4, an asymptotic theory for a test based upon $SupF$ is provided in Andrews (1993). A bootstrap-based approach which permits weaker assumptions than those used by Andrews is proposed in Hansen (2000). Structural breaks of the distributions of regressors, lagged dependent variables in the regressor set and heteroskedasticity of the errors are all allowed in Hansen's analysis; see Hansen (2000). The presence of heteroskedasticity implies that the IID-valid bootstrap techniques for Sup -type statistics discussed in Section 4.4 are inappropriate.

The key features of the alternative to the IID bootstrap that is proposed in Hansen (2000) can be summarized as follows: (i) the regressors, including any lagged values of the dependent variable, are treated as fixed; and (ii) a wild bootstrap is used to mimic heteroskedasticity. Hansen shows that a *SupF* test based upon his “fixed regressor bootstrap” is asymptotically valid; see Hansen (2000, p. 107, Theorem 6).

Hansen not only establishes the asymptotic validity of his heteroskedastic fixed regressor bootstrap test, but also provides some evidence from simulation experiments about its finite sample properties. The regression model in these experiments has the form

$$y_t = \alpha_1 y_{t-1} + \alpha_2 y_{t-2} + \alpha_3 y_{t-3} + \beta_1 x_{t-1} + \beta_2 x_{t-2} + \beta_3 x_{t-3} + \beta_4 + u_t, \\ t = 1, \dots, n, \quad (6.34)$$

with $n = 50$, $\alpha_1 = 0.5$, $\beta_1 = 1$ and all other regression parameters set equal to zero; see Hansen (2000, section 5.3) for full details of the experimental designs. The bootstrap world counterpart of (6.34), given that all regressors are treated as fixed, is

$$y_t^* = \ddot{\alpha}_1 y_{t-1} + \ddot{\alpha}_2 y_{t-2} + \ddot{\alpha}_3 y_{t-3} + \ddot{\beta}_1 x_{t-1} + \ddot{\beta}_2 x_{t-2} + \ddot{\beta}_3 x_{t-3} + \ddot{\beta}_4 + u_t^*, \\ t = 1, \dots, n, \quad (6.35)$$

with the “double-dot” notation being used to denote a parameter of the bootstrap model and the bootstrap errors u_t^* being obtained using a wild bootstrap.

As a consequence of all regressors being treated as fixed in (6.35), standard invariance results imply that the choice of bootstrap world regression parameters $\ddot{\alpha}_j$ and $\ddot{\beta}_j$ is unimportant when estimating the significance levels of tests of H_C ; see Breusch (1980). Hansen simply sets the values of such parameters equal to zero in the bootstrap world. Thus, his heteroskedasticity-valid fixed regressor bootstrap can be written as

$$y_t^* = u_t^*, t = 1, \dots, n, \quad (6.36)$$

and Hansen obtains the bootstrap errors by using

$$u_t^* = \check{u}_t \epsilon_t, t = 1, \dots, n, \quad (6.37)$$

in which ϵ_t is a typical drawing from the $N(0, 1)$ distribution, that is, $\mathcal{D}_{\epsilon, 2}$ in (5.21), and \check{u}_t is a typical residual from the OLS estimation of the model (6.32) that yields the smallest residual sum of squares as π varies between π_1 and π_2 ; see Hansen (2000, p. 106).

Hansen uses (6.36) and (6.37) to generate $B = 1,000$ wild bootstrap samples in his simulation experiments. The estimates of finite sample significance levels for bootstrap and other tests are derived from $R = 5,000$ replications in these experiments, with the desired significance level being 10 per cent. Hansen finds that, in the presence of heteroskedasticity, there is evidence of the following: the asymptotic theory in Andrews (1993) provides a very misleading approximation; an IID-valid bootstrap leads to a badly behaved test; and his heteroskedasticity-valid bootstrap test works much better. However, while the heteroskedastic fixed regressor bootstrap provides a substantial improvement relative to the other two approaches, it sometimes suffers from over-rejection, relative to the desired significance level of 10 per cent.

The approach adopted in Hansen (2000) differs from that recommended in previous sections in three ways. First, the lagged dependent variables that appear as regressors in the model for actual data are treated as fixed in the model for bootstrap data, in other words, (6.35) is used, rather than a recursive scheme like (6.18). Second, when defining the wild bootstrap error in (6.37), the scaling factor is an unrestricted residual from the estimation of the best-fitting alternative, not the restricted residual as in, e.g., (6.30). Third, the pick distribution for Hansen's scheme is the $N(0, 1)$ distribution in (5.21), whereas the Rademacher distribution of (5.25) has been recommended several times above.

A wild bootstrap test for a single structural break at an unknown point discussed in Jouini (2008) is much closer to the general recommendation made above than Hansen's procedure. Like Hansen, Jouini allows for nonstationary regressors and provides simulation evidence about finite sample behaviour of tests when the regressors include lagged values of the dependent variable. In Jouini's simulation experiments, the null model, which corresponds to (6.34) in Hansen (2000), is

$$y_t = \alpha_1 y_{t-1} + \alpha_2 y_{t-2} + \beta_1 x_{t-1} + \beta_2 x_{t-2} + \beta_3 + u_t, t = 1, \dots, n, \quad (6.38)$$

with $n = 50, 100$. However, in contrast to Hansen's fixed bootstrap, Jouini generates bootstrap data using the recursive scheme

$$y_t^* = \hat{\alpha}_1 y_{t-1}^* + \hat{\alpha}_2 y_{t-2}^* + \hat{\beta}_1 x_{t-1} + \hat{\beta}_2 x_{t-2} + \hat{\beta}_3 + u_t^*, t = 1, \dots, n, \quad (6.39)$$

in which: pre-sample values are given by $y_0^* = y_0 = 0$ and $y_{-1}^* = y_{-1} = 0$; bootstrap model coefficients $\hat{\alpha}_j$ and $\hat{\beta}_j$ are obtained from the OLS

estimation of (6.38); and the wild bootstrap error is defined by

$$u_t^* = \left[\hat{u}_t (1 - h_{tt}^R)^{-1/2} \right] \epsilon_t, \quad (6.40)$$

in which \hat{u}_t is a typical residual from OLS estimation of (6.38), h_{tt}^R is a typical leverage value for a regressor matrix for (6.38) and ϵ_t is a drawing from the Rademacher pick distribution in (5.25). Thus the only difference between Jouini's scheme and the one proposed as a general approach above is that he includes an asymptotically negligible leverage-based adjustment in (6.40).

As in the results reported in Hansen (2000), Jouini finds that the *SupF* test derived from the asymptotic theory in Andrews (1993) provides misleading inferences in the presence of heteroskedasticity. The wild bootstrap test for *SupF* obtained using (6.39) and (6.40) is superior to the heteroskedastic fixed regressor bootstrap test given in Hansen (2000) and works well, with estimated significance levels that are close to the desired levels. These results are encouraging but it could be argued that they are of limited interest because, if heteroskedastic errors were suspected, the applied worker would use a heteroskedasticity-robust statistic to test (6.31) against (6.32), not the *F*-statistic which is derived under the assumption of homoskedasticity.

MacKinnon has generalized the classic Chow test to make it robust to heteroskedasticity of unknown form; see MacKinnon (1989). He also extends the analysis by allowing for a nonlinear regression function, but, for simplicity of exposition, attention is restricted here to linear models. The null (restricted) model for the full set of n observations is then given by (6.31). Under the assumption that the alternative hypothesis specifies the breakpoint π in (6.32), the heteroskedasticity-robust check for structural change can be obtained from OLS estimation of the unrestricted model (6.32) by using a Wald statistic based upon an appropriate HCCME; see (6.8) and (6.7), respectively, for the general forms of these terms. An asymptotic test is straightforward. Sample values of the statistic for the heteroskedasticity-consistent Wald test of H_C can be compared with critical values from the asymptotic null distribution, which is $\chi^2(p + k)$. However, as stressed in Stock and Watson (2007), applied workers usually will not be certain about the breakpoint for the alternative model and that it is, therefore, inappropriate to treat π as a known constant in (6.32).

Suppose that, in the absence of perfect information about the breakpoint, (6.32) is to be estimated for each value of π in the specified set Π and a HCCME-based Wald test of H_C is carried out for each of these

values. A typical HCCME-based test statistic, as derived in MacKinnon (1989), is denoted by $\mathcal{M}_{HR}(\pi)$, with notation making its dependence on the value of the breakpoint $\pi \in \Pi$ explicit. The obvious generalization of the well-known *SupF* criterion, based upon MacKinnon's robust method, is then given by

$$\text{Sup}\mathcal{M}_{HR} = \sup_{\pi \in \Pi} \mathcal{M}_{HR}(\pi),$$

which is, under H_C , asymptotically distributed as the maximum of a set of $\chi^2(p+k)$ variables, whether or not heteroskedasticity is present.

Some simulation results on the reliability of wild bootstrap tests of the significance of the statistic $\text{Sup}\mathcal{M}_{HR}$ are reported in Lamarche (2003). Lamarche uses the specification of simulation experiments given in Hansen (2000), that is, the basic DGP is (6.34). Lamarche also follows Hansen by adopting the fixed regressor bootstrap scheme (6.36) and by using an estimated unrestricted residual in (6.37). (The unrestricted residuals are those associated with the largest of the test statistics $\mathcal{M}_{HR}(\pi)$.) Lamarche considers two pick distributions, namely, the Rademacher distribution and the standard Normal distribution. He also examines the adjustments of residuals described in MacKinnon and White (1985) in the context of the construction of the HCCME. His simulation experiments yield estimates of errors in rejection probabilities for values of the desired significance level in the range zero to 30 per cent.

The new bootstrap $\text{Sup}\mathcal{M}_{HR}$ test appears to be better behaved than *SupF* and to be quite reliable in finite samples, provided that the bootstrap models are asymptotically valid in the presence of heteroskedasticity. Lamarche remarks that his "test has good finite sample properties under different resampling procedures and transformations of the residuals"; see Lamarche (2003). Given the results in Jouini (2008), it seems reasonable to conjecture that good control of finite sample significance levels of the $\text{Sup}\mathcal{M}_{HR}$ test might be achieved by using a recursive wild bootstrap of the type that consists of (6.18) and (6.19), with restricted, rather than unrestricted, residuals being employed to scale drawings from the Rademacher pick distribution. This sort of approach is used to obtain tests of parameter stability in O'Reilly and Whelan (2005).

The simulation results that are reported in O'Reilly and Whelan (2005) cover tests that are based upon the assumption of IID errors and more robust procedures that are derived using a heteroskedasticity-consistent covariance matrix estimator. After conducting simulation experiments, O'Reilly and Whelan find that the performance of the fixed regressor bootstrap in Hansen (2000) is not robust to variations in the values of

the coefficient α_1 in the stable AR(1) model

$$y_t = \alpha_1 y_{t-1} + \beta_1 + u_t, |\alpha_1| < 1, t = 1, \dots, n.$$

The agreement between estimated and desired significance levels gets worse as α_1 increases from the value of 0.5, which is employed in Hansen (2000), with the fixed regressor bootstrap leading to excessively large estimates. O'Reilly and Whelan also find that combining a recursive bootstrap with a wild bootstrap, using the Rademacher distribution for the pick distribution, works well whether or not there is heteroskedasticity. Additional evidence in favour of the recursive wild bootstrap method, compared with the heteroskedastic fixed regressor bootstrap, is derived from experiments with an intercept term, y_{t-1} and a single autocorrelated exogenous variable as regressors; see O'Reilly and Whelan (2005, section 4).

6.5. Bootstrapping autocorrelation-robust Hausman tests

The asymptotic validity of conventional OLS-based techniques requires that the regressors are neither endogenous nor measured with errors. Tests for endogeneity and errors-in-variables are, therefore, of great importance. A test that is proposed in Hausman (1978) has become very popular and is described in many textbooks. The standard form of Hausman's test is derived under the assumption of IID errors, with the test statistic being asymptotically pivotal, having a limit null distribution that is χ^2 . As anticipated from the results in Beran (1988), the use of an IID-valid bootstrap produces better behaviour of Hausman's test than is associated with the asymptotic critical values when the errors are IID; see Wong (1996). The purpose of this section is to examine the use of bootstrap methods that permit the asymptotically valid application of Hausman's test in the presence of error autocorrelation of unknown form.

6.5.1. The forms of test statistics

In almost every application of regression analysis, there will be regressors that can be assumed to be measured without error and to be exogenous, if only because of the inclusion of an intercept term with, say, $x_{tk} = 1$ for all t . It will, therefore, be convenient to write the regression model in the partitioned form

$$\mathbf{y} = \mathbf{X}_1 \boldsymbol{\beta}_1 + \mathbf{X}_2 \boldsymbol{\beta}_2 + \mathbf{u}, \quad (6.41)$$

in which the status of the variables in X_1 is open to question, but it is maintained that X_2 contains neither endogenous variables nor incorrectly measured regressors. The orthogonality condition that is under test is that $E(\mathbf{u}|X_1, X_2)$ has all elements equal to zero.

It is useful to introduce some additional notation and to specify some assumptions. The vectors \mathbf{y} and \mathbf{u} in (6.41) are both n -dimensional, with X_1 and X_2 being $n \times k_1$ and $n \times k_2$, respectively. The terms β_1 and β_2 are $k_1 \times 1$ and $k_2 \times 1$, respectively. When there is no need to differentiate between the two types of regressor, the partitioned form of the regression model can be replaced by

$$\mathbf{y} = \mathbf{X}\beta + \mathbf{u}, \quad (6.42)$$

in which $\mathbf{X} = (X_1, X_2)$ and $\beta' = (\beta_1', \beta_2')$. As usual, let k denote the number of regressors, that is, $k = k_1 + k_2$. The errors are assumed to be stationary, autocorrelated variables with common zero mean when the orthogonality condition is true. Thus, under the null hypothesis that the regressors of X_1 are correctly measured strictly exogenous variables,

$$E(\mathbf{u}|\mathbf{X}) = \mathbf{0}_n, \quad (6.43)$$

and

$$E(\mathbf{u}\mathbf{u}'|\mathbf{X}) = \sigma^2\mathbf{R}, \quad (6.44)$$

in which \mathbf{R} is a symmetric positive-definite matrix with a typical element being the autocorrelation coefficient $\text{Corr}(u_s, u_t) = \rho(|s - t|)$, say, $s, t = 1, \dots, n$. When the errors are assumed to be IID, \mathbf{R} is the $n \times n$ identity matrix.

Hausman obtains a test of the validity of (6.43) by considering the difference between OLS and Instrumental Variable (IV) estimators of β in (6.42). When the orthogonality condition is true, both estimators are consistent and the difference between them will tend to a vector with every element equal to zero. When the orthogonality condition is not satisfied, the OLS estimator is inconsistent and will not have the same probability limit as the IV estimator; so that the difference between these estimators does not tend to a zero vector.

In order to present alternative forms for the Hausman test, it is necessary to outline some of the relevant results; see Davidson and MacKinnon (2004, section 8.7) and Greene (2008, section 12.4) for more detailed discussions. As in previous analyses, let the OLS estimator for (6.42) be

denoted by

$$\hat{\beta} = (\hat{\beta}'_1, \hat{\beta}'_2)',$$

with an associated residual vector

$$\hat{\mathbf{u}} = \mathbf{y} - \mathbf{X}\hat{\beta}.$$

It is useful to note that the normal equations for OLS estimators can be written as

$$\mathbf{X}'_1 \hat{\mathbf{u}} = \mathbf{0}_{k_1}, \quad (6.45)$$

and

$$\mathbf{X}'_2 \hat{\mathbf{u}} = \mathbf{0}_{k_2}. \quad (6.46)$$

The variables of \mathbf{X}_2 are maintained to be measured without errors and to be strictly exogenous. These regressors are, therefore, valid instruments for the estimation of (6.42). It is assumed that, as seems sensible, these regressors are included in the instrument set. The $n \times m$ matrix of instruments can then be written in the form

$$\mathbf{Z} = (\mathbf{Z}_1, \mathbf{X}_2), \quad (6.47)$$

in which \mathbf{Z}_1 is $n \times m_1$, $m_1 \geq k_1$; so that $m = m_1 + k_2 \geq k$, that is, there are sufficient instruments for estimation. Let the projection matrix $\mathbf{Z}(\mathbf{Z}'\mathbf{Z})^{-1}\mathbf{Z}'$ be denoted by \mathbf{P}_Z . The IV estimator, with which $\hat{\beta}$ is to be compared, is then

$$\tilde{\beta} = (\mathbf{X}'\mathbf{P}_Z\mathbf{X})^{-1} \mathbf{X}'\mathbf{P}_Z\mathbf{y}. \quad (6.48)$$

The focus in Hausman (1978) is on the vector of *estimator contrasts* denoted by \mathbf{q} , that is,

$$\mathbf{q} = \tilde{\beta} - \hat{\beta}.$$

Hausman shows that, under the null hypothesis,

$$\sqrt{n}\mathbf{q} \sim_a N(\mathbf{0}_k, \mathbf{C}_{qq}). \quad (6.49)$$

It might be thought, given previous results on the construction of test statistics, that it would be possible to derive an asymptotically valid test of the significance of the joint significance of the elements of $\sqrt{n}\mathbf{q}$ in

which critical values were taken from the $\chi^2(k)$ distribution. However, this is not the case because the covariance matrix C_{qq} in (6.49) is singular and care must be taken to avoid invalid critical values; see, for example, Krämer and Sonnberger (1986).

In order to show the cause of the singularity of the covariance matrix C_{qq} , it is useful to substitute the OLS-based identity

$$\mathbf{y} = \mathbf{X}\hat{\boldsymbol{\beta}} + \hat{\mathbf{u}},$$

in (6.48). This substitution leads to

$$\begin{aligned} \mathbf{q} &= \tilde{\boldsymbol{\beta}} - \hat{\boldsymbol{\beta}} \\ &= (\mathbf{X}'\mathbf{P}_Z\mathbf{X})^{-1}\mathbf{X}'\mathbf{P}_Z(\mathbf{X}\hat{\boldsymbol{\beta}} + \hat{\mathbf{u}}) - \hat{\boldsymbol{\beta}} \\ &= (\mathbf{X}'\mathbf{P}_Z\mathbf{X})^{-1}\mathbf{X}'\mathbf{P}_Z\hat{\mathbf{u}}. \end{aligned}$$

Since $\mathbf{P}_Z\mathbf{X}$ is the matrix of predicted values from the OLS regression of \mathbf{X} on \mathbf{Z} and \mathbf{X}_2 is included in \mathbf{Z} , $\mathbf{X}'\mathbf{P}_Z\hat{\mathbf{u}}$ can be partitioned as

$$\begin{pmatrix} \mathbf{X}'_1\mathbf{P}_Z\hat{\mathbf{u}} \\ \mathbf{X}'_2\mathbf{P}_Z\hat{\mathbf{u}} \end{pmatrix} = \begin{pmatrix} \mathbf{X}'_1\mathbf{P}_Z\hat{\mathbf{u}} \\ \mathbf{X}'_2\hat{\mathbf{u}} \end{pmatrix} = \begin{pmatrix} \mathbf{X}'_1\mathbf{P}_Z\hat{\mathbf{u}} \\ \mathbf{0}_{k_2} \end{pmatrix},$$

in which use has been made of (6.46). Hence

$$\mathbf{q} = \begin{pmatrix} \mathbf{q}_1 \\ \mathbf{q}_2 \end{pmatrix} = (\mathbf{X}'\mathbf{P}_Z\mathbf{X})^{-1} \begin{pmatrix} \mathbf{X}'_1\mathbf{P}_Z\hat{\mathbf{u}} \\ \mathbf{0}_{k_2} \end{pmatrix},$$

and so the last k_2 elements of the estimator contrast vector \mathbf{q} are linear combinations of the first k_1 elements. It follows that, when \mathbf{X}_2 is included in \mathbf{Z} , the relevant reference distribution for Hausman's test is $\chi^2(k_1)$, not $\chi^2(k)$. There are two well-known forms of an appropriate asymptotic theory test statistic.

The first form is a direct Wald test of the significance of

$$\mathbf{q}_1 = \tilde{\boldsymbol{\beta}}_1 - \hat{\boldsymbol{\beta}}_1; \tag{6.50}$$

so that only the contrasts corresponding to suspect regressors are used. Under standard regularity conditions, the asymptotic null distribution of $\sqrt{n}\mathbf{q}_1$ is given by

$$\sqrt{n}\mathbf{q}_1 \sim_a N(\mathbf{0}_{k_1}, C_{q_1q_1}), \tag{6.51}$$

in which $C_{q_1q_1}$ is a non-singular matrix. Let $\ddot{C}_{q_1q_1}$ be an estimator of $C_{q_1q_1}$ that is consistent when the null hypothesis is true. The estimator contrast form of Hausman's test is then based upon the result that

$$\mathcal{H}_1 = n\mathbf{q}'_1 \left[\ddot{C}_{q_1q_1} \right]^{-1} \mathbf{q}_1 \sim_a \chi^2(k_1), \quad (6.52)$$

under the null hypothesis.

Hausman is able to derive a very simple expression for the covariance matrix estimator in (6.52) by imposing the auxiliary assumption of IID errors; see Hausman (1978, p. 1254). This expression is obtained by considering the covariance between $\sqrt{n}(\tilde{\beta}_1 - \beta_1)$ and $\sqrt{n}(\hat{\beta}_1 - \beta_1)$. As explained in Greene (2008, p. 208), "Hausman's essential result is that the covariance of an efficient estimator with its difference from an inefficient estimator is zero." This result allows the covariance matrix in (6.51) to be written as the difference between the covariance matrices of IV and OLS estimators; see Greene (2008, pp. 208–9). However, when the assumption of IID errors is relaxed, OLS is no longer asymptotically efficient and Hausman's simplifying result is not applicable. The form of the covariance matrix estimator used in (6.52) is more complicated when either autocorrelation or heteroskedasticity is present. An expression that is appropriate under general stationary autocorrelation is given in Li (2006) for the special case in which just enough instruments are used, that is, \mathbf{Z} in (6.47) is $n \times k$.

Given the complexity of the form (6.52) under non-IID errors, a second (indirect) approach is attractive. In this second approach, straightforward manipulations are used to establish that Hausman's test can be implemented by testing $\boldsymbol{\gamma} = \mathbf{0}_{k_1}$ in the augmented model

$$\mathbf{y} = \mathbf{X}_1\boldsymbol{\beta}_1 + \mathbf{X}_2\boldsymbol{\beta}_2 + [\mathbf{P}_Z\mathbf{X}_1]\boldsymbol{\gamma} + \mathbf{u}; \quad (6.53)$$

see, for example, Davidson and MacKinnon (2004, section 8.7). If the OLS estimator for $\boldsymbol{\gamma}$ in (6.53) is denoted by $\check{\boldsymbol{\gamma}}$, a robust version of Hausman's procedure can be obtained by checking the significance of $\sqrt{n}\check{\boldsymbol{\gamma}}$, using an appropriate robust covariance matrix estimator in a Wald statistic. According to the departures from the assumption of IID errors that are being allowed, the covariance matrix should be autocorrelation-consistent, or a HCCME, or a HAC estimate. If the relevant covariance matrix estimate is denoted by $\check{C}_{\boldsymbol{\gamma}\boldsymbol{\gamma}}$, then, under the null hypothesis,

$$\mathcal{H}_2 = n\check{\boldsymbol{\gamma}}' \left[\check{C}_{\boldsymbol{\gamma}\boldsymbol{\gamma}} \right]^{-1} \check{\boldsymbol{\gamma}} \sim_a \chi^2(k_1), \quad (6.54)$$

with significantly large values indicating the inconsistency of the null hypothesis with the data.

The two forms of Hausman's test that are provided in (6.52) and (6.54) both rely upon asymptotic theory. Given the results in Wong (1996) for the IID errors case that illustrate the improved performance of Hausman's test when bootstrap methods replace asymptotic critical values, the use of bootstrap techniques that are appropriate in the presence of autocorrelation of unknown form, as represented by (6.44), is clearly worth considering.

Perhaps the most obvious way in which to employ bootstrap methods is to apply the same formula for the test statistic to the actual sample and to each of a suitably large number of bootstrap samples. The bootstrap p -value of the statistic from the actual data can then be calculated and compared with the desired significance level. In order to be a basis for asymptotically valid inference under autocorrelation of unspecified form, the bootstrap samples would have to be generated using one of the methods discussed in Section 5.3, with the block and sieve bootstraps being obvious candidates.

It is possible that the direct application of a bootstrap as outlined in the previous paragraph would lead to difficulties and uncertainties related to the choice of method used to obtain the relevant autocorrelation-consistent covariance matrix estimate, either $\check{C}_{q_1 q_1}$ in (6.52) or $\check{C}_{\gamma\gamma}$ in (6.54). An alternative bootstrap-based strategy is to use the results in Gonçalves and White (2005) on the bootstrap estimation of covariance matrices.

Suppose, for example, that the estimator contrast \mathbf{q}_1 in (6.50) is to be checked for statistical significance. If B bootstrap samples are obtained, using, for example, a block bootstrap as described in Section 5.3.2, then B bootstrap counterparts of the contrast \mathbf{q}_1 can be derived. Let these bootstrap statistics be denoted by $\mathbf{q}_{1(b)}^*$, $b = 1, \dots, B$. A simple estimate of the covariance matrix of $\sqrt{n}\mathbf{q}_1$, which can be used in (6.52), is given by

$$\mathbf{C}_{q_1 q_1}^* = \frac{n}{B} \sum_{b=1}^B \left[(\mathbf{q}_{1(b)}^* - \bar{\mathbf{q}}_1^*) (\mathbf{q}_{1(b)}^* - \bar{\mathbf{q}}_1^*)' \right],$$

in which

$$\bar{\mathbf{q}}_1^* = \frac{1}{B} \sum_{b=1}^B \mathbf{q}_{1(b)}^*.$$

A statistic of the form (6.52) can then be computed as

$$\mathcal{H}_1^* = n\mathbf{q}'_1 \left[\mathbf{C}_{q_1 q_1}^* \right]^{-1} \mathbf{q}_1, \quad (6.55)$$

and its sample value could be compared with a right-hand-tail critical value from the $\chi^2(k_1)$ distribution.

As with the studies of heteroskedasticity-robust bootstrap tests discussed above, the special case of testing a single restriction has received attention in the literature on bootstrap versions of Hausman tests. If $k_1 = 1$, the statistic \mathcal{H}_2 in (6.54) can be calculated as the square of the autocorrelation-robust t -ratio for testing $\gamma = 0$ in

$$\mathbf{y} = \mathbf{x}_1\beta_1 + \mathbf{X}_2\beta_2 + [\mathbf{P}_Z\mathbf{x}_1]\gamma + \mathbf{u}, \quad (6.56)$$

with the asymptotic critical values taken from the $\chi^2(1)$ distribution. The relevant t -ratio can be written as $\check{\gamma}/ACSE(\check{\gamma})$, with $ACSE(\check{\gamma})$ denoting the autocorrelation-consistent standard error of the OLS estimator $\check{\gamma}$. If there is uncertainty about how best to calculate $ACSE(\check{\gamma})$, or there is concern that the estimate might be excessively variable for sample sizes of relevance to the applied worker, it may be reasonable to bootstrap the coefficient estimate $\check{\gamma}$, rather than its associated autocorrelation-consistent t -ratio; see the comments on Berkowitz and Kilian (2000) in Davidson (2000, p. 49).

If, for example, a block bootstrap were used to generate B bootstrap samples, then a p -value for the non-asymptotically pivotal coefficient estimate $\check{\gamma}$ could be calculated as

$$\hat{p}^*(\check{\gamma}) = \frac{\#\{|\check{\gamma}_{(b)}^*| \geq |\check{\gamma}|\}}{B}, \quad (6.57)$$

in which $\check{\gamma}_{(b)}^*$ is the bootstrap counterpart of $\check{\gamma}$, $b = 1, \dots, B$. However, given that the point estimate is not asymptotically pivotal, there is no asymptotic refinement derived from the bootstrap. A double bootstrap can be used in an attempt to improve the control of finite sample significance levels. In particular, a fast double bootstrap (FDB) of the type described in Section 2.5 may offer the benefits of good performance at relatively low computational cost; see Davidson and MacKinnon (2002b, 2007) for discussions of this approach.

Some simulation evidence that throws light on the relative merits of alternative bootstrap methods for providing autocorrelation-robust Hausman tests is discussed next. This evidence is obtained using the experimental designs given in Li (2006).

6.5.2. Simulation experiments

In the simulation experiments that are conducted in Li (2006), a block bootstrap is used to carry out an autocorrelation-consistent version of Hausman's test of an estimator contrast. The results from the experiments provide information about the finite sample properties of this bootstrap test and those of the corresponding test based upon the asymptotic null distribution. Li describes his block bootstrap test as consisting of the following three steps.

Li (2006): Step 1

Use the actual data to apply OLS to the regression model (6.42); the OLS residuals in $\hat{\mathbf{u}}$ and the point estimates in $\hat{\boldsymbol{\beta}}$ are saved for use in the next step. Before proceeding to the second step, (6.42) is estimated by IV and the statistic \mathcal{H}_1 , as defined in (6.52), is calculated, using an appropriate autocorrelation-consistent covariance matrix estimate. An invalid form of the test statistic based upon the false assumption of IID errors is also computed in Li (2006). This inappropriate statistic is denoted by \mathcal{H}_1^\times .

Li proposes that a block bootstrap be used to generate B samples, each of which is used to calculate a value of the autocorrelation-robust Hausman test statistic. Step 2 is, therefore, repeated B times.

Li (2006): Step 2

The vector of OLS residuals $\hat{\mathbf{u}} = (\hat{u}_1, \hat{u}_2, \dots, \hat{u}_n)'$ from Step 1 is divided into non-overlapping blocks of equal length. Let the common block length be denoted by ℓ and the number of blocks be $r = n/\ell$. The r blocks of residuals can be written as

$$\widehat{\mathcal{U}}_j = (\hat{u}_{1+(j-1)\ell}, \hat{u}_{2+(j-1)\ell}, \dots, \hat{u}_{j\ell}), j = 1, 2, \dots, r.$$

These blocks are used to define the probability model for a random bootstrap error block of length ℓ , denoted by \mathcal{U}^* , with

$$\Pr(\mathcal{U}^* = \widehat{\mathcal{U}}_j) = \frac{1}{r} \text{ for } j = 1, \dots, r. \quad (6.58)$$

For repetition b of Step 2, a sequence of n bootstrap errors $\mathbf{u}_{(b)}^* = (u_{1(b)}^*, u_{2(b)}^*, \dots, u_{n(b)}^*)'$ is then generated by randomly sampling, with replacement, r blocks from the population defined by (6.58) and then joining the blocks together. The vector $\mathbf{u}_{(b)}^*$ is combined with the OLS predicted value of Step 1 to obtain the b th sample of bootstrap data for

the dependent variable, according to

$$\mathbf{y}_{(b)}^* = \mathbf{X}\hat{\boldsymbol{\beta}} + \mathbf{u}_{(b)}^*. \tag{6.59}$$

These bootstrap data are used, like the actual data in Step 1, to calculate the estimator contrast version of Hausman’s test, as given in (6.52). The bootstrap counterpart of \mathcal{H}_1 from Step 1 is denoted by $\mathcal{H}_{1(b)}^*$.

Li (2006): Step 3

Conventional asymptotic theory can be used to judge whether or not the sample value of \mathcal{H}_1 from Step 1 is statistically significant. However, after Step 2 has been carried out for $b = 1, \dots, B$, a bootstrap p -value test can be used. More precisely, the null hypothesis is rejected if

$$\frac{\#(\mathcal{H}_{1(b)}^* \geq \mathcal{H}_1)}{B} \leq \alpha_d,$$

in which α_d is the desired significance level. In Li’s analysis, the inappropriate statistic \mathcal{H}_1^\times is compared with critical values from the asymptotic null distribution of \mathcal{H}_1 in order to illustrate the consequences of failing to take into account the presence of autocorrelation.

The experiments that are used in Li (2006) to investigate the usefulness of bootstrap tests provided by Steps 1, 2 and 3 can be described as follows. The regression model in Li (2006) is

$$y_t = x_{t1}\beta_1 + \beta_2 + u_t, \tag{6.60}$$

with u_t generated according to either the MA(2) scheme

$$u_t = e_t + \theta_1 e_{t-1} + \theta_2 e_{t-2}, \quad (\theta_2 = 0.5),$$

or the AR(1) model

$$u_t = \phi u_{t-1} + e_t, \quad |\phi| < 1,$$

and

$$\begin{pmatrix} e_t \\ x_{t1} \\ z_{t1} \end{pmatrix} \sim NID \left(\begin{pmatrix} 0 \\ 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 1 & \rho_{xe} & 0 \\ \rho_{xe} & 1 & \rho_{xz} \\ 0 & \rho_{xz} & 1 \end{pmatrix} \right), \tag{6.61}$$

for $t = 1, \dots, n$. The non-singularity of the covariance matrix in (6.61) requires that $1 - \rho_{xz}^2 - \rho_{xe}^2 > 0$. Observation vectors for regressors and

instruments are defined as $\mathbf{x}_t = (x_{t1}, 1)'$ and $\mathbf{z}_t = (z_{t1}, 1)'$, respectively, $t = 1, \dots, n$. The null hypothesis corresponds to $\rho_{xe} = 0$ and small values of ρ_{xz} provides insights into the issue of weak instruments, as discussed in Staiger and Stock (1997).

All experiments in Li (2006) have $\beta_1 = \beta_2 = 1$. The sample size is $n = 100$ and the block length is $\ell = 4$; so that the number of blocks is $r = 25$. The number of block bootstrap samples is $B = 1,000$. The desired significance levels are 5 per cent and 10 per cent and, for each experiment, $R = 2,000$ replications are used to estimate the corresponding finite sample significance levels. Li summarizes his results as follows. As might be expected, the invalid test \mathcal{H}_1^\times has estimates that indicate substantial departures from the intended significance levels (estimates are lower than desired and sometimes very low, for example, less than 0.5 per cent). The estimates for the correct asymptotic theory test of \mathcal{H}_1 have better agreement with the desired values, but satisfactory performance is not always observed with some estimates being outside of the range $\alpha_d \pm 0.2\alpha_d$. The bootstrap test of the significance of \mathcal{H}_1 appears to work well, with evidence of good control of significance levels in almost every case. Li interprets his results as showing that, after correcting the Hausman statistic to be asymptotically robust to autocorrelation, a block bootstrap variant of the test outperforms the version derived using the standard first-order asymptotic theory. However, these results are subject to some limitations.

First, when carrying out experiments with MA(2) errors, Li uses a covariance matrix estimator based upon knowledge of the correct autocorrelation model, that is, it is assumed that autocorrelations of third and higher order equal zero. The non-zero autocorrelation terms are calculated using residuals from OLS estimation of (6.60); see Li (2006, p. 80). Li's use of "null hypothesis" residuals when estimating autocorrelation matrices is supported by results to be found in Ligeralde and Brown (1995) on restricted versus unrestricted residuals. However, the use of the correct MA(2) autocorrelation model could be viewed as tending to lead to estimates that overstate the accuracy of both asymptotic and block bootstrap tests.

Second, for the experiments with AR(1) errors, Li uses a Bartlett kernel with bandwidth equal to 5; this choice leads to a covariance matrix for the OLS estimator defined by appropriate special cases of (1.47) and (1.48) with truncation lag $l = 4$. However, as demonstrated in Kiefer and Vogelsang (2005), the choice of bandwidth and kernel may be important and there are many alternatives to the combination used in Li (2006).

In view of these remarks, it seems worthwhile to examine results obtained by repeating the experiments in Li (2006) without using information about the correct error model and without imposing his choices for the bandwidth and kernel. Three procedures are to be considered and none of them requires an applied researcher to select a bandwidth and kernel combination in order to obtain an autocorrelation-consistent covariance matrix estimate. All are based upon the added variable approach to computing the Hausman statistic. Since the status of only one regressor is under test in the experiments of Li (2006), a special case of (6.56) can be used to calculate test statistics of $\gamma = 0$.

In the first test, denoted by \mathcal{T}_1 , block bootstrap samples are used to estimate the autocorrelation-consistent standard error required for a quasi- t ratio, which is compared with asymptotic critical values from the $N(0, 1)$ distribution for a two-sided alternative; see Gonçalves and White (2005) on the bootstrap estimation of standard errors in linear regression models. For the second and third tests, no standard error is required and instead it is the point estimate of the coefficient of the test variable in (6.56) that is the object of interest in block bootstrap samples. The block bootstrap is used in the second procedure, denoted by \mathcal{T}_2 , to estimate a p -value, as given in (6.57), which is compared with desired significance levels. For the third test, denoted by \mathcal{T}_3 , the p -value calculated for \mathcal{T}_2 is used as the test statistic in a FDB approach. Several descriptions of standard single level bootstrap tests have been provided above and so only the application of the FDB test will be described here. The FDB version of the autocorrelation-robust Hausman test with a single regressor under scrutiny can, like Li's procedure, be described as a three-step method.

FDB autocorrelation-robust Hausman test: Step 1

The actual data are used to obtain OLS results for the regression equations that correspond to (6.42) and (6.56). The OLS residual vector $\hat{\mathbf{u}}$ and parameter estimate $\hat{\beta}$ for the former model are saved to be used for the generation of block bootstrap data. The point estimate of γ in the model corresponding to (6.56) is the object used as the statistic of interest when performing the FDB test of the null hypothesis that the regressor under investigation is exogenous and measured without error. This point estimate is denoted by $\check{\gamma}$.

As in the approach described in Li (2006), the second step is to be repeated B times before proceeding to Step 3.

FDB autocorrelation-robust Hausman test: Step 2

It is convenient to break the second step, which is repeated until B block bootstrap samples have been generated and analysed, into two parts.

(a) As in Li's method, the vector of OLS residuals $\hat{\mathbf{u}} = (\hat{u}_1, \hat{u}_2, \dots, \hat{u}_n)'$ is divided into non-overlapping blocks of equal length ℓ . The $r = n/\ell$ blocks of OLS residuals can, as before, be written as

$$\hat{u}_j = (\hat{u}_{1+(j-1)\ell}, \hat{u}_{2+(j-1)\ell}, \dots, \hat{u}_{j\ell}), j = 1, 2, \dots, r.$$

For repetition b of Step 2, a sequence of n bootstrap errors $\mathbf{u}_{(b)}^* = (u_{1(b)}^*, u_{2(b)}^*, \dots, u_{n(b)}^*)'$ is generated by randomly sampling, with replacement, r blocks from the population defined by (6.58) and then joining the blocks together. The vector $\mathbf{u}_{(b)}^*$ is combined with the OLS predicted values of Step 1 to obtain the b th sample of bootstrap data for the dependent variable, according to (6.59). The implied vector of n bootstrap observations on the dependent variable is $\mathbf{y}_{(b)}^*$, which is now subject to the same OLS analyses as was the actual value \mathbf{y} in Step 1. In particular, the OLS estimation of the first-level bootstrap counterpart of (6.42) yields point estimates of coefficients and estimated residuals as the elements of the vectors $\hat{\boldsymbol{\beta}}_{(b)}^*$ and $\hat{\mathbf{u}}_{(b)}^* = (\hat{u}_{1(b)}^*, \hat{u}_{2(b)}^*, \dots, \hat{u}_{n(b)}^*)'$, respectively. Also the OLS estimation of the first-level bootstrap counterpart of (6.56) yields an estimate for the coefficient of the test variable, which is denoted by $\check{\gamma}_{(b)}^*$.

(b) In order to implement the FDB test, the block bootstrap is then applied to $\hat{\mathbf{u}}_{(b)}^*$ in the same way that it was applied to $\hat{\mathbf{u}}$ in part (a), but only once. The sampling and joining together of blocks derived from $\hat{\mathbf{u}}_{(b)}^*$ gives a single second-level bootstrap vector $\mathbf{u}_{1(b)}^{**}$. The corresponding single vector of second-level observations on the dependent variable is then

$$\mathbf{y}_{1(b)}^{**} = \mathbf{X}\hat{\boldsymbol{\beta}}_{(b)}^* + \mathbf{u}_{1(b)}^{**}.$$

Now the second-level block bootstrap counterpart of the artificial model (6.56) can be estimated by OLS, with the statistic of interest being the estimated coefficient of the test variable. This estimate is denoted by $\check{\gamma}_{1(b)}^{**}$.

FDB autocorrelation-robust Hausman test: Step 3

After Step 2 has been carried out B times, the estimated p -value of the actual estimate $\check{\gamma}$ can be calculated using (6.57). In the FDB test, this quantity, denoted by $\hat{p}^*(\check{\gamma})$, is now viewed as the test statistic and, in order to derive a reference distribution, the second-level results are used

to estimate p -values for each of the first-level terms $\check{\gamma}_{(b)}^*$. Thus a collection of B estimated p -values is calculated according to

$$\hat{p}_{(b)}^{**}(\check{\gamma}_{(b)}^*) = \frac{1}{B} \left[\mathbf{1}(|\check{\gamma}_{1(1)}^{**}| \geq |\check{\gamma}_{(b)}^*|) + \mathbf{1}(|\check{\gamma}_{1(2)}^{**}| \geq |\check{\gamma}_{(b)}^*|) + \dots + \mathbf{1}(|\check{\gamma}_{1(B)}^{**}| \geq |\check{\gamma}_{(b)}^*|) \right], b = 1, 2, \dots, B,$$

in which $\mathbf{1}(\cdot)$ is the usual indicator function. The FDB adjusted p -value is obtained by comparing $\hat{p}^*(\check{\gamma})$ with $\hat{p}_{(b)}^{**}(\check{\gamma}_{(b)}^*)$, $b = 1, 2, \dots, B$, and is defined by

$$\hat{p}_F(\check{\gamma}) = \frac{1}{B} \sum_{b=1}^B \mathbf{1}(\hat{p}_{(b)}^{**}(\check{\gamma}_{(b)}^*) \leq \hat{p}^*(\check{\gamma})).$$

The null hypothesis is rejected if $\hat{p}_F(\check{\gamma}) \leq \alpha_d$.

Given that \mathcal{T}_1 is an asymptotic theory test, with the bootstrap used to calculate the standard error, not a critical value, it might be anticipated that its small sample performance would be similar to that of \mathcal{T}_2 , which combines a statistic that is not asymptotically pivotal with a bootstrap. It is also to be expected that \mathcal{T}_3 would outperform \mathcal{T}_2 as a result of the use of the FDB. Evidence on the behaviour of these three tests in finite samples is obtained using the experiments employed in Li (2006). In these simulation experiments, the null (restricted) model is (6.60) and artificial alternative (unrestricted) model is

$$y_t = x_{t1}\beta_1 + \beta_2 + \hat{x}_{t1}\gamma + u_t, \tag{6.62}$$

in which \hat{x}_{t1} is a typical predicted value from the OLS regression of x_{t1} on $\mathbf{z}_t = (z_{t1}, 1)'$, $t = 1, \dots, n$. Simulation results for \mathcal{T}_1 , \mathcal{T}_2 and \mathcal{T}_3 are derived using the same experiments as are specified in Li (2006), except that the number of replications is increased to $R = 25,000$.

The results from the simulation experiments concerning the finite sample significance levels of the three block bootstrap-based tests \mathcal{T}_1 , \mathcal{T}_2 and \mathcal{T}_3 are summarized in Table 6.3. These results correspond to those presented in Table 1 of Li (2006). Consequently the two values used for the desired significance level in Table 6.3 are given by $\alpha_d = 5$ per cent, 10 per cent. The contents of Table 6.3 indicate that \mathcal{T}_1 is, on the whole, slightly inferior to \mathcal{T}_2 in terms of the agreement between the estimated and desired significance levels. Both tests reject more frequently than desired, but the problem is not very serious, with the claim that the actual significance level is between α_d and $1.2\alpha_d$ being consistent

Table 6.3 Estimated significance levels of bootstrap-based autocorrelation-robust Hausman tests, using block bootstrap samples, $n = 100$ and $\ell = 4$

a. MA(2) errors							
Parameters		\mathcal{T}_1 test		\mathcal{T}_2 test		\mathcal{T}_3 test	
ρ_{xz}	θ_1	(i)	(ii)	(i)	(ii)	(i)	(ii)
0.1	0.3	5.6	10.9	5.5	10.6	5.2	10.2
0.1	0.7	5.6	11.0	5.5	10.8	5.3	10.3
0.4	0.3	5.5	10.9	5.5	10.6	5.2	10.1
0.4	0.7	5.7	11.0	5.5	10.7	5.3	10.3
0.7	0.3	5.4	10.9	5.3	10.6	5.0	10.1
0.7	0.7	5.5	10.9	5.4	10.7	5.2	10.2

b. AR(1) errors							
Parameters		\mathcal{T}_1 test		\mathcal{T}_2 test		\mathcal{T}_3 test	
ρ_{xz}	ϕ	(i)	(ii)	(i)	(ii)	(i)	(ii)
0.1	0.3	5.5	11.0	5.3	10.5	5.0	10.2
0.1	0.8	5.8	11.1	5.9	11.0	5.6	10.5
0.4	0.3	5.7	11.2	5.5	10.9	5.3	10.5
0.4	0.8	5.8	11.3	5.8	11.2	5.5	10.6
0.7	0.3	5.8	11.2	5.7	10.7	5.5	10.3
0.7	0.8	5.9	11.2	5.9	11.2	5.6	10.6

Notes: Estimates that are given under (i) and (ii) correspond to desired significance levels of 5 per cent and 10 per cent, respectively. Each estimate is derived from 25,000 replications and **bold** font denotes that the estimate is consistent with the claim that the true significance level is between α_d and $1.1\alpha_d$, $\alpha_d = 5$ per cent, 10 per cent, as indicated by the test in Godfrey and Orme (2000, p. 75).

with all estimates for \mathcal{T}_1 and \mathcal{T}_2 , for $\alpha_d = 5$ per cent, 10 per cent. As expected, the evidence in Table 6.3 indicates that the FDB test \mathcal{T}_3 outperforms both \mathcal{T}_1 and \mathcal{T}_2 . The tests in Godfrey and Orme (2000, p. 75) yield the outcome that none of the estimates for \mathcal{T}_3 would lead to rejection of the claim that the actual significance levels of this test are in range α_d to $1.1\alpha_d$, $\alpha_d = 5$ per cent, 10 per cent. The FDB test, therefore, has well-behaved finite sample significance levels and, in contrast to the block bootstrap test proposed in Li (2006), does not require the user to specify a choice for the bandwidth and kernel required to obtain an autocorrelation-consistent standard error.

The ability of a test to detect departures from the null hypothesis is, of course, of considerable importance when such departures imply the inconsistency of the OLS estimators of the regression model. Given that the FDB test \mathcal{T}_3 and Li's block bootstrap test based upon \mathcal{H}_1 in (6.52) both seem to be well-behaved under the null hypothesis, it is reasonable to compare their rejection frequencies under the alternative hypothesis, that is, with $\rho_{xe} \neq 0$. Li denotes his bootstrap version of the autocorrelation-consistent estimator contrast test by \mathcal{W}^b and presents results on its empirical power; see Li (2006, Table 2). Several of these results are for situations in which rejection rates are close to the desired significance level. The comparisons of \mathcal{T}_3 and \mathcal{W}^b given here are for the more interesting cases with higher estimates. More precisely, cases are selected so that, with a desired significance level of 10 per cent, rejection rates are approximately 20 per cent, 30 per cent, ..., 90 per cent and 100 per cent. The estimates for such cases are contained in Table 6.4. The differences between estimates are sometimes small, but, when they are more substantial, it is \mathcal{T}_3 that outperforms \mathcal{W}^b .

It might be thought that, despite the evidence in Tables 6.3 and 6.4, the FDB Hausman test \mathcal{T}_3 is unattractive because it has a high computational cost. However, there are simple devices that remove the need for repeated use of OLS estimation programs in the three steps for the calculation of \mathcal{T}_3 that are given above. The key to savings is that the observations in \mathbf{x}_1 , \mathbf{X}_2 and \mathbf{Z}_1 are fixed over first and second levels of block bootstrap samples. In order to explain the savings, use is made of the n -dimensional vector \mathbf{v} , which is given by

$$\mathbf{v} = (\mathbf{x}'_1 \mathbf{P}_Z \mathbf{M} \mathbf{P}_Z \mathbf{x}_1)^{-1} \mathbf{M} \mathbf{P}_Z \mathbf{x}_1,$$

in which \mathbf{M} and \mathbf{P}_Z are as defined above.

Given the definition of \mathbf{v} , it follows that $\check{\gamma} = \mathbf{v}'\mathbf{y} = \mathbf{v}'\mathbf{u}$, when $\gamma = 0$ in (6.56). In Step 2 of the scheme for implementing the FDB test, all bootstrap data are generated under the null hypothesis, that is, with $\gamma = 0$ in (6.56). It is, therefore, only necessary to compute \mathbf{v} from actual data in Step 1 and its value can be stored for use in the B block bootstrap samples of Step 2 to obtain $\check{\gamma}_{(b)}^*$ and $\check{\gamma}_{1(b)}^{**}$, $b = 1, \dots, B$. In part (a) of Step 2, having obtained the first-level block bootstrap errors $\mathbf{u}_{(b)}^*$ from $\hat{\mathbf{u}}$, the results that $\hat{\mathbf{u}}_{(b)}^* = \mathbf{M}\mathbf{u}_{(b)}^*$ and $\check{\gamma}_{(b)}^* = \mathbf{v}'\mathbf{u}_{(b)}^*$ can be used to obtain the required residuals and point estimate without use of OLS estimation for first-level bootstrap data. In part (b) of Step 2, having derived the corresponding second-level block bootstrap error vector $\mathbf{u}_{1(b)}^{**}$ from

Table 6.4 Estimated power of bootstrap-based autocorrelation-robust Hausman tests, using block bootstrap samples, $n = 100$ and $\ell = 4$, with desired significance levels of 5 per cent and 10 per cent.

a. MA(2) errors						
<i>Parameters</i>			\mathcal{T}_3 test		\mathcal{W}^b test	
ρ_{xz}	ρ_{xe}	θ_1	(i)	(ii)	(i)	(ii)
0.4	0.3	0.5	19.0	28.7	17.6	27.6
0.4	0.5	0.5	49.2	61.7	46.4	60.4
0.7	0.3	0.5	69.9	80.1	59.5	74.6
0.7	0.5	0.5	99.8	99.9	98.6	99.7

b. AR(1) errors						
<i>Parameters</i>			\mathcal{T}_3 test		\mathcal{W}^b test	
ρ_{xz}	ρ_{xe}	ϕ	(i)	(ii)	(i)	(ii)
0.4	0.3	0.8	13.0	20.8	12.9	21.4
0.4	0.5	0.8	27.8	39.2	26.6	39.4
0.7	0.3	0.8	43.6	56.4	36.1	50.6
0.7	0.5	0.8	89.9	94.7	80.2	92.0

Notes: Estimates that are given under (i) and (ii) correspond to desired significance levels of 5 per cent and 10 per cent, respectively. Estimates for \mathcal{T}_3 are derived from 25,000 replications. Estimates for Li's test \mathcal{W}^b are taken from Table 2 in Li (2006, p. 80).

$\hat{\mathbf{u}}_{(b)}^*$, $\check{\gamma}_{1(b)}^{***} = \mathbf{v}'\mathbf{u}_{1(b)}^{**}$ can be used to obtain the required coefficient estimate, rather than direct estimation of the counterpart of (6.56). The only time that OLS estimation is required is when the actual sample data are analysed.

6.6. Summary and conclusions

In recent years, there has been increasing emphasis on the need to use robust methods of inference in regression analysis and to abandon methods derived under the restrictive assumption of IID errors. The robust significance tests that are now often recommended for general use are based upon estimators of the sampling covariances and variances of OLS coefficient estimators that are consistent when the errors are not IID. The former set of estimators are the elements of the covariance matrix

estimator, which can be constructed to be consistent in the presence of autocorrelation and/or heteroskedasticity of the errors. Important early examples of such covariance matrix estimators are to be found in Newey and West (1987) and White (1980).

In practical situations, attempts to use robust tests often rely upon standard first-order asymptotic theory. However, the finite sample distributions of test statistics may be poorly approximated by such asymptotically valid results and the use of appropriate bootstrap methods may be of value in applied work. Since the bootstrap must mimic the data generation process assumed to apply to the actual observations, the bootstrap scheme must reflect the permitted departures from the assumption of IID errors.

The standard results for asymptotic and bootstrap “robust” tests require that the OLS estimators of the null model are consistent and asymptotically Normally distributed. Consequently it is vital that there are no specification errors that imply the inconsistency of OLS estimators. An essential part of regression analysis allowing for non-IID errors should, therefore, be the calculation and examination of checks for specification errors, which, in the terminology of Verbeek, correspond to “cases where the OLS estimator cannot be saved”; see Verbeek (2004, section 5.2). The tests for these important specification errors should not be based upon the assumption of IID errors; arguments about robustness apply to such checks as well as to other significance tests.

Several examples of robust tests for the causes of OLS inconsistency have been discussed in this chapter, with both asymptotic and bootstrap versions being considered. Three of the examples illustrate the application of heteroskedasticity-robust tests. These examples are: (i) the RESET test for incorrect functional form and omitted regressors; (ii) the Breusch-Godfrey test for autocorrelation in dynamic regression models; and (iii) testing for a structural break with an unknown breakpoint.

One conclusion that emerges from the discussion of the heteroskedasticity-robust tests is that asymptotic critical values do not always provide satisfactory control of finite sample significance levels. Evidence of much better and more reliable control is observed for a wild bootstrap test in which the pick distribution is the Rademacher distribution discussed in Davidson and Flachaire (2001, 2008). This wild bootstrap scheme has been found to be useful in several quite different applications and is recommended for general use. When the regression model is dynamic, it is additionally recommended that a recursive version of this wild bootstrap be used, rather than relying upon a “fixed regressor” wild bootstrap approach.

It seems likely that heteroskedasticity-robust tests will become very commonly used and it is hoped that standard estimation programs will allow the use of the Rademacher-based wild bootstrap. As discussed above and illustrated by the first three examples, there is now a considerable body of evidence to support the general use of the wild bootstrap recommended in Davidson and Flachaire (2001, 2008), whether or not first-order asymptotic theory leads to a standard distribution for critical values. However, not every popular test in regression analysis can be made asymptotically valid in the presence of heteroskedasticity of unknown form. In particular, it is argued in Godfrey (2008) that the predictive tests discussed in Sections 1.6 and 4.2 cannot be modified to be heteroskedasticity-robust using, for example, the results in White (1980).

As might be anticipated from the remarks in Section 5.5, there is less evidence available for bootstrap tests that allow for autocorrelated errors than there is for heteroskedasticity-robust bootstrap tests. When stationary autocorrelation of unspecified form is permitted, no lagged values of the dependent variable can be included in the regressor set because the OLS estimators of regression coefficients are required to be consistent and asymptotically Normally distributed. The relevant literature has been focussed on the use of the block bootstrap as a tool for approximating the actual dependence of the errors when the regressors are strictly exogenous. The block bootstrap can be used to obtain a p -value for a test statistic computed using a selected combination of bandwidth and kernel to derive an autocorrelation-consistent covariance matrix estimate, or it can be used to estimate the covariance matrix in order to calculate a Wald statistic that is then compared with an asymptotic critical value.

The final example in this chapter illustrates the application of such bootstrap techniques and others to a widely-used test proposed in Hausman (1978), which is intended to be powerful against endogeneity and errors-in-variables. The block bootstrap p -value approach has been found to outperform the asymptotic critical value version of an autocorrelation-robust form of Hausman's test; see Li (2006). A fast double bootstrap test has been used for the special case of a single suspect regressor in Section 6.5 above and gives very close agreement with desired significance levels. The simulation evidence obtained indicates that this method works relatively well under null and alternative hypotheses, does not require the estimation of an autocorrelation-robust standard deviation and need not have a heavy computational cost.

The sieve bootstrap could be used as an alternative to the block bootstrap to derive autocorrelation-robust procedures. There are several

applications of the former method to situations in which nonstationary processes are under consideration; see, for example, Chang (2004), Chang et al. (2006) and Fuertes (2008). However, the use of sieve bootstraps with OLS-based tests would, like the block bootstrap, require that no lagged dependent variables be used as regressors.

It has been stressed above that the selected bootstrap scheme must match the assumptions made about the errors of the model that has been proposed as the data generation process. For example, the wild bootstrap tests described in this chapter involve sampling independently distributed artificial errors and so would not be appropriate if the researcher were to need bootstrap tests that are asymptotically valid in the presence of both heteroskedasticity and autocorrelation. When all regressors are strictly exogenous, the moving blocks bootstrap discussed in Fitzenberger (1998) gives asymptotically valid HAC tests, with blocks of data on the regressand and regressors, rather than blocks of residuals, being resampled; see Section 5.4.2 for details of Fitzenberger's bootstrap approach. The resampling of blocks of data on regressors implies that the bootstrap observations in the $n \times k$ regressor matrix X^* are not fixed over bootstrap samples and so the computational savings outlined in, for example, Section 6.5, are not available.

7

Simulation-based Tests for Non-nested Regression Models

7.1. Introduction

It has been argued in previous chapters that the proper use of bootstrap methods can often produce better control of finite sample significance levels than can be obtained from asymptotic theory. Moreover, an appropriate bootstrap approach can sometimes be used to derive a valid large sample test when none is available from standard asymptotic theory. However, all of the previous discussions and recommendations have been based upon the assumption that the model of the null hypothesis is a special case of the model of the alternative hypothesis; it has, therefore, been possible to refer to the former as *restricted* and the latter as *unrestricted*.

When the null hypothesis model can be obtained by restricting the coefficients of the alternative specification against which it is being tested, the two hypotheses are said to be *nested*. More precisely, the null hypothesis model is nested in the alternative model. In very many tests of nested hypotheses, a set of linear coefficient restrictions is imposed on the alternative model to obtain the null model. In other applications of nested hypothesis tests, one or more coefficients of the alternative are assumed to tend to some limit in order to derive the null specification to be tested. While it is certainly the case that tests of nested hypotheses are very common in applied econometrics, there are important situations in which nesting is not possible, that is, the hypotheses involved in the test are *non-nested*. This chapter contains discussions of asymptotic and simulation-based tests of non-nested hypotheses in regression analysis. As in the chapters on tests for nested hypotheses, the relevant asymptotic theory and the available evidence from studies of finite sample properties will be outlined.

The primary purposes of this chapter are to explain the application of bootstrap techniques to tests of non-nested models and to summarize evidence about the usefulness of such techniques. No attempt is made to give a detailed discussion of the theoretical results for tests of non-nested models that have been obtained. Readers who are interested in learning more about the theory that underpins tests of non-nested hypotheses should consult the very useful surveys that have been published; see, for example, Gourieroux and Monfort (1994), McAleer and Pesaran (1986), MacKinnon (1983), Pesaran and Dupleich Ulloa (2008), Pesaran and Weeks (2001) and Szroeter (1999). An assessment of the impact of the theoretical literature on the practice of applied workers is provided in McAleer (1995). Although theoretical results and empirical applications are available for non-nested hypotheses in a wide variety of situations, only the case of competing linear regression models is discussed in this chapter.

It is convenient for the purposes of exposition to restrict attention to cases that can be discussed in the familiar framework of a linear regression model with IID errors, which is estimated by OLS. Thus the competing models are assumed to have the same dependent variable, but to have different sets of regressors that are not nested. This assumption is made in standard textbook discussions of non-nested regression models; see, for example, Greene (2008, section 7.3). Greene gives, as an example of such a situation, the two competing non-nested models

$$y_t = \beta_1 + \beta_2 x_t + \beta_3 x_{t-1} + u_{t0},$$

and

$$y_t = \gamma_1 + \gamma_2 x_t + \gamma_3 y_{t-1} + u_{t1},$$

in which y_t and x_t denote the levels of consumption and income in period t , respectively. Clearly, there is the potential for non-nested models to occur whenever economists disagree about the choice of regressors. The case of non-nested regressor sets has received a great deal of attention, with a number of asymptotic and bootstrap tests being proposed and examined in the literature.

The asymptotic tests proposed for regression models that have non-nested regressors are outlined in Section 7.2. As will be seen, several issues arise. Classical likelihood-based approaches to testing lead to results that are not like those derived in the context of nested hypotheses. Also, some researchers have used the statistics proposed for testing model validity

in the presence of non-nested alternatives as tools for model selection. Remarks on both types of use are provided in Section 7.2.

Bootstrap versions of tests for non-nested regression models are discussed in Section 7.3. A review of results from simulation experiments that provide information about the finite sample properties of asymptotic and bootstrap tests is given. These results are relevant to assessing the usefulness of asymptotic critical values, single bootstraps and fast double bootstraps in both standard and non-standard situations.

Attention is restricted in Section 7.3 to consideration of bootstrapping test statistics that, under the null hypothesis, have proper asymptotic distributions (which may or may not have standard forms). When analysis moves beyond the simple framework of linear regression models, the derivation of such statistics is sometimes difficult. In one approach to overcoming problems of analytical intractability, the bootstrap is simply applied to an easily computed index of relative fit. The simplicity of this approach is, at first sight, attractive, but, as explained in Section 7.4, it does not, in general, provide an appropriate comparison basis for estimating p -values.

Finally, Section 7.5 contains a summary and some concluding remarks.

7.2. Asymptotic tests for models with non-nested regressors

Many econometricians tackling the problems of testing when regressor sets are non-nested have used results taken from two influential articles by Cox in which he proposes that a starting point in a general approach should be the *log-likelihood ratio* (LLR) statistic; see Cox (1961, 1962). (Cox refers to non-nested hypotheses as *separate families of hypotheses*.) However, direct application of the ideas put forward by Cox does not always lead to convenient tests and various alternative approaches have been proposed.

In this section, important tests based upon different approaches are outlined. As a starting point for exposition, it is assumed that there are just two non-nested models to be considered. Regularity conditions and the problems associated with near-orthogonality of the two non-nested sets of regressors are discussed. The initial assumption that the validity of one model is to be tested, given a single non-nested alternative, is then relaxed and testing in the presence of several non-nested alternatives is discussed. Next, as an alternative to checking the assumption of validity, the use of tests for model selection is considered. Finally some evidence about finite sample properties of asymptotic tests for

non-nested regression models that has been obtained from simulation experiments is summarized.

7.2.1. Cox-type LLR tests

In order to provide details of the implementation of Cox's suggestion in the context of non-nested linear regression models, Pesaran assumes that there are two competing models, which can be written as:

$$H_0 : \mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{u}_0, \mathbf{u}_0 \sim N(\mathbf{0}_n, \sigma_0^2 \mathbf{I}_n), \quad (7.1)$$

and

$$H_1 : \mathbf{y} = \mathbf{Z}\boldsymbol{\gamma} + \mathbf{u}_1, \mathbf{u}_1 \sim N(\mathbf{0}_n, \sigma_1^2 \mathbf{I}_n), \quad (7.2)$$

in which \mathbf{X} and \mathbf{Z} are $n \times k_0$ and $n \times k_1$ matrices of observations on strictly exogenous regressors for H_0 and H_1 , respectively; see Pesaran (1974). The two sets of regressors are non-nested; so that not every variable in \mathbf{X} is a linear combination of those in \mathbf{Z} and not every variable in \mathbf{Z} is a linear combination of those in \mathbf{X} .

In contrast to the classical case of nested hypotheses, H_0 and H_1 are not reserved to denote the null and alternative hypotheses, respectively. Applied workers often wish to test H_1 against H_0 , as well as H_0 against H_1 . In such situations, there are four possible outcomes: (i) both models can be rejected; (ii) both models can be accepted; (iii) H_0 can be accepted and H_1 can be rejected; and (iv) H_0 can be rejected and H_1 can be accepted; see, for example, Pesaran and Dupleich Ulloa (2008, pp. 107–108) for comments.

The regression models of (7.1) and (7.2) can be written in terms of conditional distributions as

$$H_0 : \mathbf{y} | \mathbf{X}, \mathbf{Z} \sim N(\mathbf{X}\boldsymbol{\beta}, \sigma_0^2 \mathbf{I}_n),$$

and

$$H_1 : \mathbf{y} | \mathbf{X}, \mathbf{Z} \sim N(\mathbf{Z}\boldsymbol{\gamma}, \sigma_1^2 \mathbf{I}_n),$$

respectively. Note that the same conditioning is used for both specifications; see Gourieroux and Monfort (1994, p. 2592).

For the purpose of explaining the *Cox-type test*, suppose that the validity of H_0 is to be tested, using H_1 as the alternative. Given the models (7.1) and (7.2), maximum likelihood estimates, and hence the likelihood

ratio statistic recommended by Cox, can be obtained by Ordinary Least Squares (OLS) estimation. In order to provide details of the test, it is useful to introduce some additional notation. Let OLS regression coefficient estimates, predicted values and residuals for the two models be the elements of the vectors

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y},$$

$$\hat{\mathbf{y}} = (\mathbf{Z}'\mathbf{Z})^{-1}\mathbf{Z}'\mathbf{y},$$

$$\hat{\mathbf{y}}_0 = \mathbf{X}\hat{\boldsymbol{\beta}} = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y},$$

$$\hat{\mathbf{y}}_1 = \mathbf{Z}\hat{\mathbf{y}} = \mathbf{Z}(\mathbf{Z}'\mathbf{Z})^{-1}\mathbf{Z}'\mathbf{y},$$

$$\hat{\mathbf{u}}_0 = (\hat{u}_{10}, \dots, \hat{u}_{n0})' = \mathbf{M}_X\mathbf{y},$$

and

$$\hat{\mathbf{u}}_1 = (\hat{u}_{11}, \dots, \hat{u}_{n1})' = \mathbf{M}_Z\mathbf{y},$$

where, in the definitions of residual vectors,

$$\mathbf{M}_X = \mathbf{I}_n - \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}',$$

and

$$\mathbf{M}_Z = \mathbf{I}_n - \mathbf{Z}(\mathbf{Z}'\mathbf{Z})^{-1}\mathbf{Z}'.$$

The maximum likelihood estimates of the error variances for (7.1) and (7.2) are denoted by $\hat{\sigma}_0^2 = n^{-1}\mathbf{y}'\mathbf{M}_X\mathbf{y}$ and $\hat{\sigma}_1^2 = n^{-1}\mathbf{y}'\mathbf{M}_Z\mathbf{y}$, respectively.

The LLR statistic is $\hat{L}_{01} = \hat{L}_0 - \hat{L}_1$, in which \hat{L}_0 is the maximized log-likelihood for (7.1) and \hat{L}_1 is the maximized log-likelihood for (7.2). The two maximized log-likelihood functions are given by

$$\hat{L}_i = \sum_{t=1}^n \hat{l}_{ti}, \quad (7.3)$$

and

$$\hat{l}_{ti} = -\frac{1}{2} \ln(2\pi) - \frac{1}{2} \ln(\hat{\sigma}_i^2) - \frac{\hat{u}_{ti}^2}{2\hat{\sigma}_i^2}, \quad (7.4)$$

that is, \hat{l}_{ti} is the contribution to \hat{L}_i , $i = 0, 1$, that is associated with observation t , $t = 1, \dots, n$. It follows that

$$\hat{L}_{01} = \hat{L}_0 - \hat{L}_1 = \frac{n}{2} \ln \left(\frac{\hat{\sigma}_1^2}{\hat{\sigma}_0^2} \right); \quad (7.5)$$

see, for example, Pesaran (1974, pp. 156–157) for the derivation of \hat{L}_{01} and other details.

If the models were nested, with $\mathbf{M}_Z\mathbf{X}$ being an $n \times k_0$ matrix with every element equal to zero, $-2\hat{L}_{01}$ would be asymptotically distributed as $\chi^2(k_1 - k_0)$, under regularity conditions, when (7.1) is valid. However, this result no longer applies when (7.1) and (7.2) are non-nested and a different asymptotic theory is required to determine an appropriate basis for making inferences. In his pioneering articles, Cox concentrates on the general ideas of his approach to testing non-nested hypotheses, rather than on details of regularity conditions. General regularity conditions for Cox's test of non-nested hypotheses are considered in White (1982b). The details of technical assumptions that imply the asymptotic validity of the Cox-type approach in the context of linear regression models are given in White (1982b, section 3). It is important to note that one of these assumptions is that the regressors of the two models are not asymptotically orthogonal, that is, $\text{plim } n^{-1}\mathbf{X}'\mathbf{Z}$ must not have every element equal to zero.

Now, under the regularity conditions for non-nested regression models, \hat{L}_{01} is $O_p(n)$ when (7.1) is the true model, not $O_p(1)$ as it would be if (7.1) were nested in (7.2). The scaled LLR statistic $n^{-1}\hat{L}_{01}$ has probability limit given by

$$\mu_{01} = \text{plim}_0 \ n^{-1}\hat{L}_{01} = \frac{1}{2} \ln \left(\frac{\sigma_0^2 + \boldsymbol{\beta}'\boldsymbol{\Lambda}\boldsymbol{\beta}}{\sigma_0^2} \right), \quad (7.6)$$

in which plim_0 denotes that the probability limit is taken assuming that (7.1) is the true data generation process (DGP) and

$$\boldsymbol{\Lambda} = \text{plim } n^{-1}\mathbf{X}'\mathbf{M}_Z\mathbf{X}.$$

Under the null hypothesis model (7.1), the term μ_{01} defined in (7.6) can be estimated consistently by

$$\hat{\mu}_{01} = \frac{1}{2} \ln \left(\frac{\hat{\sigma}_0^2 + \hat{\boldsymbol{\beta}}'\hat{\boldsymbol{\Lambda}}\hat{\boldsymbol{\beta}}}{\hat{\sigma}_0^2} \right), \quad (7.7)$$

in which

$$\hat{\Lambda} = n^{-1} X' M_Z X. \quad (7.8)$$

Consequently, the *centred* LLR statistic

$$T_{01} = n^{-1} \hat{L}_{01} - \hat{\mu}_{01}, \quad (7.9)$$

converges in probability to zero when (7.1) is true.

As discussed in Cox (1961, 1962), provided regularity conditions are satisfied, $\sqrt{n}T_{01}$ is asymptotically distributed as a Normal variable with zero mean and finite variance V_{01} , under the null hypothesis H_0 . Moreover, as shown in Pesaran (1974, p. 158), the estimator

$$\hat{V}_{01} = \frac{\hat{\sigma}_0^2 \left(\hat{\beta}' \left[n^{-1} X' M_Z M_X M_Z X \right] \hat{\beta} \right)}{(\hat{\sigma}_0^2 + \hat{\beta}' \hat{\Lambda} \hat{\beta})^2}, \quad (7.10)$$

is consistent for V_{01} when (7.1) is true. Hence the standardized criterion

$$N_{01} = \frac{T_{01}}{\sqrt{\hat{V}_{01}}}, \quad (7.11)$$

is asymptotically distributed as $N(0, 1)$ under the null hypothesis that (7.1) is valid.

The statistic N_{01} in (7.11) can be compared with critical values from the $N(0, 1)$ distribution to obtain a valid asymptotic test. The literature includes discussions of both one-sided and two-sided tests. If the model of (7.2) is the only alternative to which consideration is to be given, then a one-sided test can be used since the probability limit of N_{01} is negative under H_1 . (Note that, when a one-sided test is employed, a statistically significant negative value of N_{01} does not imply that H_1 is correct; see MacKinnon (1983, pp. 91–92).) On the other hand, if the applied worker wishes to allow for the possibility of some third unspecified model being correct, a two-sided test of the significance of N_{01} is appropriate. In the latter case, the Cox-type test can be regarded as a check for general misspecification in (7.1), rather than a specific test against (7.2).

The basic form of the Cox statistic N_{01} is adjusted in Godfrey and Pesaran (1983) in attempts to obtain procedures with better finite sample properties; the simulation evidence on small sample behaviour is discussed at the end of this section. Godfrey and Pesaran make use of

the assumptions of exogenous regressors and NID errors to derive mean and variance adjustments. Their adjusted test statistic is denoted by \tilde{N}_{01} ; see Godfrey and Pesaran (1983, section 2). Godfrey and Pesaran also use an asymptotically valid linearization to obtain a second adjusted statistic, which can be denoted by W_{01} . The adjustments used by Godfrey and Pesaran are asymptotically negligible, under weak conditions, and their asymptotic tests \tilde{N} and W would be valid if the errors were simply assumed to be IID. However, their variance adjustments rely upon Normality for their justification.

The centred LLR test of Cox (1961, 1962) and the adjusted variants proposed in Godfrey and Pesaran (1983) have not been widely used; see McAleer (1995, Table 4). This lack of popularity may reflect the absence of a clear motivation for considering the LLR statistic when the models are non-nested. Alternatively it may be that the variance estimates, for example, as given in (7.10) for N_{01} , do not suggest intuitively appealing test statistics.

7.2.2. Artificial regression tests

Davidson and MacKinnon provide an alternative to the Cox-type test that is simpler to implement and much easier to motivate; see Davidson and MacKinnon (1981). The procedure described by Davidson and MacKinnon, which is called the *J test*, is probably the most widely-used method for testing model specification in the presence of nonnested alternative models. In the context of (7.1) and (7.2), Davidson and MacKinnon propose that the former is tested using information about the latter by carrying out an asymptotically valid *t*-test of $\delta = 0$ in the artificial regression model

$$\mathbf{y} = \delta \hat{\mathbf{y}}_1 + \mathbf{X}\boldsymbol{\beta} + \mathbf{u}_0, \mathbf{u}_0 \sim N(\mathbf{0}_n, \sigma_0^2 \mathbf{I}_n). \quad (7.12)$$

If the relevant *t*-ratio is denoted by J_{01} , Davidson and MacKinnon show that, as $n \rightarrow \infty$, J_{01} tends to $-N_{01}$, when H_0 is true; see Davidson and MacKinnon (1981, pp. 789–790). However, the *J test* method has the advantages of being simpler to implement than the Cox-type procedure and has a clearer motivation in terms of assessing the relevance of the predicted value from the competing specification.

The artificial regression (7.12) is not the only way in which the null model (7.1) can be nested in an equation for the purposes of deriving a test statistic. An adjusted *J*-statistic, denoted by *JA*, is proposed in Fisher and McAleer (1981). The *JA*-type method for checking the validity of (7.1) uses, as its test variable, the predicted value from the OLS regression of

$\hat{\mathbf{y}}_0$ on \mathbf{Z} . Thus the statistic JA_{01} is the t -ratio for testing $\delta = 0$ in the artificial regression

$$\mathbf{y} = \delta [\mathbf{P}_Z \hat{\mathbf{y}}_0] + \mathbf{X}\boldsymbol{\beta} + \mathbf{u}_0, \mathbf{u}_0 \sim N(\mathbf{0}_n, \sigma_0^2 \mathbf{I}_n), \quad (7.13)$$

in which $\mathbf{P}_Z = \mathbf{Z}(\mathbf{Z}'\mathbf{Z})^{-1}\mathbf{Z}'$. The J test and JA test are asymptotically equivalent under the null hypothesis. However, in contrast to the J test, which is only asymptotically valid, the JA test is exactly valid when (7.1) is true.

7.2.3. Comprehensive model F -test

As an alternative to the one-degree-of-freedom J and JA tests, it is possible to form a nesting model by adding the regressors specific to (7.2) to the regressors of (7.1). This nesting, or *comprehensive*, model can be written as

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}_1\boldsymbol{\gamma}_1 + \mathbf{u}_0, \mathbf{u}_0 \sim N(\mathbf{0}_n, \sigma_0^2 \mathbf{I}_n), \quad (7.14)$$

in which: $\mathbf{Z} = (\mathbf{Z}_0, \mathbf{Z}_1)$; $\mathbf{Z}'_0\mathbf{M}_X\mathbf{Z}_0$ is a matrix with every element equal to zero; and $\mathbf{Z}'_1\mathbf{M}_X\mathbf{Z}_1$ is positive definite. If \mathbf{Z}_1 is $n \times k_{11}$, the F -statistic for testing (7.1) against (7.14), which is denoted by F_{01} , is distributed as $F(k_{11}, n - k_0 - k_{11})$ when the former is valid. If there is only one regressor which is specific to (7.2), that is, $k_{11} = 1$, this F -statistic is the square of the value of both the J_{01} and JA_{01} statistics. Hence the J test is exactly valid in the special case with $k_{11} = 1$; see Godfrey (1984, p. 75).

7.2.4. Regularity conditions and orthogonal regressors

It is possible to motivate tests of (7.1) against the artificial regression models (7.12), (7.13) and (7.14) without explicit reference to the centred LLR statistic which is the focus of Cox's analysis. Consequently there is no compelling reason to derive such tests under the restrictive assumption of Normality, which is used in Pesaran (1974) to derive the log-likelihood functions. Asymptotic validity under weaker assumptions that permit the inclusion of lagged dependent variables in the regressor set and non-Normality of the errors is established in MacKinnon et al. (1983). From now on, it is simply assumed that the errors of the true model are IID and that regularity conditions corresponding to those given in MacKinnon et al. (1983) are satisfied.

It has been emphasized in the literature that the conditions for the asymptotic validity of procedures such as the Cox-type and J tests require that the regressors of the competing models are not orthogonal; see, for example, MacKinnon (1983, p. 96). This condition is appropriate when the models have no regressors in common. However, in most cases, the

models have one or more regressors in common, for example, there will usually be an intercept term in both models. As indicated in Michelis (1999, p. 371), the required absence-of-orthogonality condition is a little more complicated when some regressors are in both models, or more generally when some (but not all) regressors of one are linear combinations of the regressors of the other.

The absence-of-orthogonality condition used in Godfrey and Pesaran (1983), which covers the case of non-nested models with regressors in common, is that

$$\phi^* = \text{plim}_0 n^{-1} \beta' X' P_Z M_X P_Z X \beta, \quad (7.15)$$

should exist and be a finite positive quantity. This condition implies that the test variables in (7.12) and (7.13) are asymptotically cooperative; see Schmidt (1976, section 2.7) for discussion of asymptotically uncooperative regressors. The simpler condition that $n^{-1} X' Z$ should not tend to a null matrix is, however, not sufficient to ensure that $\phi^* > 0$; see Godfrey and Pesaran (1983, appendix A).

Michelis provides an asymptotic analysis that throws light on the consequences of weakly correlated regressors in nonnested models; see Michelis (1999). Michelis allows for a local form of orthogonality, which he terms *near population orthogonality* (NPO). In the NPO framework, $\beta' X' P_Z M_X P_Z X \beta$ in (7.15) is $O_p(1)$, not $O_p(n)$, and so $\phi^* = 0$, implying that the condition given in Godfrey and Pesaran (1983) is not satisfied.

Michelis proves that, under the NPO assumption, the J -statistic does not tend to a $N(0, 1)$ variable when the null hypothesis is true, but instead tends to a random function that depends upon a nuisance parameter, a $N(0, 1)$ variable and a χ^2 variable. Consequently, when the NPO assumption provides a good approximation to the behaviour of a J -statistic, the use of critical values from the conventional $N(0, 1)$ distribution will not give the desired significance level in large samples. Michelis carries out simulation experiments to assess the impact of weak correlations between the non-nested regressors and finds evidence that actual rejection probabilities can be much greater than the desired level when the assumed asymptotic reference distribution for the J -statistic is $N(0, 1)$; see Michelis (1999, section 5).

7.2.5. Testing with multiple alternatives

It is not only the failure of the absence-of-orthogonality assumption that can lead to the standard asymptotic results being an inadequate basis for inference when testing models with non-nested regressors. In particular,

problems can arise when the null model is to be tested in the presence of several non-nested alternative models. Suppose that, as before, the model to be tested is (7.1), except that the errors are assumed to be IID, rather than NID. However, it is now assumed that, rather than just having one non-nested alternative, the researcher has m , $m > 1$, non-nested regression models that can be used as a source of evidence against the null model.

The vectors of OLS predicted values from these alternatives are denoted by $\hat{\mathbf{y}}_j$, $j = 1, \dots, m$. The corresponding m binary J tests, in each of which (7.1) is tested against just one of the alternatives, are the tests of $\delta_j = 0$ in

$$\mathbf{y} = \delta_j \hat{\mathbf{y}}_j + \mathbf{X}\boldsymbol{\beta} + \mathbf{u}_0, \mathbf{u}_0 \sim IID(\mathbf{0}_n, \sigma_0^2 \mathbf{I}_n), j = 1, \dots, m, \quad (7.16)$$

where the IID errors have an unspecified CDF, denoted by \mathcal{F}_0 . The t -statistics for testing $\delta_j = 0$ in models like (7.16) are denoted by J_{0j} , $j = 1, \dots, m$.

A joint test could be implemented by using an asymptotically valid F -test of the m restrictions of $\delta_1 = \delta_2 = \dots = \delta_m = 0$ in the artificial model

$$\mathbf{y} = \sum_{j=1}^m \delta_j \hat{\mathbf{y}}_j + \mathbf{X}\boldsymbol{\beta} + \mathbf{u}_0, \mathbf{u}_0 \sim IID(\mathbf{0}_n, \sigma_0^2 \mathbf{I}_n), \quad (7.17)$$

provided $n > k_0 + m$. If the sample size were sufficiently large, it would also be possible to carry out an asymptotic F -test of the validity of (7.1) against a comprehensive model that includes the null model and all m alternative models as special cases. The comprehensive model for this F -test can be written as

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}^\dagger \boldsymbol{\gamma}^\dagger + \mathbf{u}_0, \mathbf{u}_0 \sim IID(\mathbf{0}_n, \sigma_0^2 \mathbf{I}_n), \quad (7.18)$$

in which \mathbf{Z}^\dagger is obtained from $(\mathbf{Z}_1, \dots, \mathbf{Z}_m)$ by first removing any variable that is a linear combination of the regressors in (7.1) and then deleting any redundant variables; so that $(\mathbf{X}, \mathbf{Z}^\dagger)$ has full column rank for sufficiently large n .

McAleer draws attention to the fact that, when faced by multiple non-nested alternatives, applied workers have used collections of separate binary tests more frequently than they employed a joint test; see McAleer (1995, p. 162, Table 6). This practice leads to a non-standard overall asymptotic test. Suppose that two-sided binary J tests are being used; so that the squared t -ratio can serve as the test statistic for each

of the m binary checks. Each of these squared t -ratios $J_{0j}^2, j = 1, \dots, m$, is asymptotically distributed as $\chi^2(1)$ when the null model is valid and regularity conditions are satisfied. If all m binary statistics are insignificant at an individual desired significance level of α_d , then $J_{0j}^2 < c$, for $j = 1, \dots, m$, where $\text{Prob}(\chi^2(1) \geq c) = \alpha_d$. It follows that, in such a situation, $\text{Sup}J_0^2 = \max(J_{01}^2, \dots, J_{0m}^2)$ is less than c . The limit null distribution of $\text{Sup}J_0^2$ is, however, not $\chi^2(1)$ and the Bonferroni inequality simply implies that the asymptotic significance level associated with the rule "Reject H_0 if $\text{Sup}J_0^2 \geq c$ " is between α_d and $m\alpha_d$. Consequently, if (7.1) is only accepted when all of the m separate binary J tests yield insignificant outcomes, the overall significance level is unknown, even asymptotically. The use of the bootstrap in this non-standard case is discussed in the next section.

In all that has been covered so far, the purpose of testing has been to check the validity of one model in the light of evidence that is provided by one or more other non-nested models. In particular, following Cox (1961, 1962), the LLR statistic \hat{L}_{01} in (7.5) has been used to obtain a test of the validity of (7.1) with (7.2) being the non-nested alternative; the roles of the models can, of course, be reversed if the validity of each in turn is to be tested. However, there is another way in which the LLR statistic can be used in a test. As explained by Lien and Vuong, \hat{L}_{01} in (7.5) can be employed when the researcher is interested "in discriminating between the competing models by testing the hypothesis that the models are 'equivalent' under some appropriate definition"; see Lien and Vuong (1987).

7.2.6. Tests for model selection

A general discussion of testing for the purpose of model selection, as opposed to model validation, is provided in Vuong (1989). The null hypothesis and regularity conditions used in Vuong imply that, when the claim that the non-nested models are equivalent is true, $\text{plim}_{\mathcal{T}} n^{-1}\hat{L}_{01} = 0$, where $\text{plim}_{\mathcal{T}}$ denotes a probability limit taken under the unknown true DGP. Vuong shows that, in general, when the models are equivalent according to his definition,

$$n^{-1/2}\hat{L}_{01} \sim_a N(0, \omega^2);$$

see Vuong (1989, sections 5 and 6). Given a consistent estimator of ω^2 , denoted by $\hat{\omega}^2$, a test procedure using critical values from the $N(0, 1)$ distribution is straightforward to apply. Suppose that the asymptotic

significance level for the model selection test is 5 per cent. There are three possible outcomes. First, if the sample value of $n^{-1/2}\hat{L}_{01}/\sqrt{\hat{\omega}^2}$ is smaller than -1.96 , the data are interpreted as suggesting that (7.2) is better than (7.1). Second, if the sample value of $n^{-1/2}\hat{L}_{01}/\sqrt{\hat{\omega}^2}$ is between -1.96 and 1.96 , the data are judged to be consistent with the claim that the models fit equally well. Third, if the sample value of $n^{-1/2}\hat{L}_{01}/\sqrt{\hat{\omega}^2}$ is greater than 1.96 , the data are taken to indicate that (7.1) is better than (7.2). Further discussion of the implementation of Vuong's test and an example of its application are provided in Greene (2008, pp. 140–142).

As has been noted in the literature, the statistic used in Vuong's test can be adjusted so that it uses well-known *model selection criteria*. More precisely,

$$\begin{aligned} n^{-1/2}\hat{L}_{01} &= n^{-1/2}(\hat{L}_0 - \hat{L}_1) \\ &= n^{-1/2}\left[\left(\hat{L}_0 - s(n, k_0)\right) - \left(\hat{L}_1 - s(n, k_1)\right)\right] + o_p(1), \\ &= n^{-1/2}[IC_0 - IC_1] + o_p(1), \end{aligned}$$

in which IC_0 and IC_1 are information criteria derived by applying penalty functions $s(n, k_0)$ and $s(n, k_1)$ that adjust for the dimension of the model and are $o(n^{1/2})$. Many computer programs calculate, as part of the results for OLS estimation, the values of the *Akaike Information Criterion* (AIC) and *Schwarz Bayesian Information Criterion* (BIC). These criteria are defined using penalty functions of the type $s(\cdot, \cdot)$, with $s(n, k_i) = k_i$ for AIC and $s(n, k_i) = k_i \ln(n)/2$ for BIC, $i = 0, 1$. Since $\ln(n) > 2$ for sample sizes of relevance to applied work, the BIC measure clearly gives a greater penalty per regressor than AIC and is recommended in Hansen (1999) for model selection.

Clarke, like Vuong, has examined the problem of testing for model selection, rather than for model validity; see Clarke (2003, 2007). He refers to the former approach as leading to tests of *relative discrimination* and to the latter approach as yielding tests of *absolute discrimination*. Clarke provides a simple alternative to Vuong's procedure. The maximized log-likelihoods for the models (7.1) and (7.2) are both regarded as the sum of the n per observation contributions; so that

$$\hat{L}_{01} = (\hat{L}_0 - \hat{L}_1) = \sum_{t=1}^n \hat{l}_{t0} - \sum_{t=1}^n \hat{l}_{t1} = \sum_{t=1}^n (\hat{l}_{t0} - \hat{l}_{t1}),$$

using the notation in (7.3) and (7.4). Clarke considers the sequence of differences $\hat{l}_{t0} - \hat{l}_{t1}$, $t = 1, \dots, n$, and computes

$$\hat{p}_C = n^{-1} \sum_{t=1}^n \mathbf{1}(\hat{l}_{t0} - \hat{l}_{t1} > 0),$$

that is, the proportion of times in which a positive value of $\hat{l}_{t0} - \hat{l}_{t1}$ is observed. He argues that, if the models are equivalent, then, for large samples, \hat{p}_C should be close to 0.5. A standard test for a binomial proportion, with hypothesized value equal to 0.5, is used to assess the evidence provided by the sample value of \hat{p}_C . As with the test in Vuong (1989), the procedure given in Clarke (2003, 2007) can be modified to be based upon model selection criteria like AIC and BIC, in place of the unadjusted maximized log-likelihood functions.

Evidence from simulation experiments is used by Clarke to argue that the simple test based upon the signs of the terms $\hat{l}_{t0} - \hat{l}_{t1}$, $t = 1, \dots, n$, is superior to the Vuong test; see Clarke (2007). However, neither Clarke's test nor Vuong's test is designed to detect misspecification of a model. The purpose of model selection tests is to test the hypothesis that two models are equally good (or equally bad) according to a specified criterion, not to throw light on whether either of them is consistent with the sample data. In the rest of this chapter, discussion will be restricted to tests for model validity.

7.2.7. Evidence from simulation experiments

The finite sample properties of asymptotic tests for model validity, when regressor sets are non-nested, are investigated in a number of studies that use simulation experiments for the estimation of rejection probabilities. The evidence from these experiments is summarized in surveys, for example, McAleer and Pesaran (1986, section 5), Pesaran and Weeks (2001, section 5.5) and Szroeter (1999, section 3.1.2). More detailed discussions of the results are provided in Davidson and MacKinnon (1982, 2002b), Ericsson (1986), Godfrey and Pesaran (1983) and Michelis (1999). The various researchers have found very similar general features in their results.

First, the use of asymptotic critical values with unadjusted Cox-type statistics, such as N_{01} in (7.11), leads to excessively high estimates of significance levels; an attempt to explain the observed discrepancies is given in Ericsson (1986). Second, the comparison of J -statistics like J_{01} with critical values from the asymptotically valid $N(0, 1)$ distribution also produces rejection rates that are too high relative to the desired

significance level α_d . For example, with $\alpha_d = 5$ per cent, estimated significance levels for the J test in the range 15 per cent to 25 per cent are obtained in the experiments in Godfrey and Pesaran (1983). The results given in Michelis (1999) are consistent with those in Godfrey and Pesaran (1983) and indicate that the unadjusted Cox-type test is even more badly behaved than the J test when judged by finite sample rejection frequencies of true models. Third, the asymptotic tests based the JA -statistic proposed in Fisher and McAleer (1981) and the F -test derived from a comprehensive model like (7.14) have finite sample significance levels that seem to be reasonably close to desired values, even in the presence of lagged dependent variables in the regressor sets and non-Normal error distributions; see, for example, Godfrey and Pesaran (1983).

However, as argued in MacKinnon (1983), the relatively favourable findings about the behaviour of the JA test and comprehensive model F test, under the null hypothesis, are not sufficient to imply that these procedures can be recommended for routine use in applied econometrics because the ability of tests to detect false models is also important. Unfortunately, there is evidence that the JA test and F test can both lack power in situations of practical relevance; see Godfrey and Pesaran (1983, section 4), McAleer and Pesaran (1986, section 5) and MacKinnon (1983, section 3).

The adjusted Cox-type tests based upon the \tilde{N}_{01} and W_{01} statistics proposed in Godfrey and Pesaran (1983) have been found to have estimates of finite sample significance levels that are quite close to the desired values, even when the errors do not have a Normal distribution and the regressors include the lagged value of the dependent variable; see Godfrey and Pesaran (1983, section 4). Moreover, in power comparisons, Godfrey and Pesaran observe that the \tilde{N} and W tests are superior to both the JA test and the F test. Unfortunately, while applied workers can compute tests based upon artificial regressions like (7.12), (7.13) and (7.14) quite easily, the implementation of the \tilde{N} and W tests is much less convenient unless special routines are included in the estimation program.

It seems likely that the pattern of usage reflected by Table 4 in McAleer (1995) will continue in the future with the J test being the most popular tool for testing regression models with non-nested regressors. However, the asymptotic theory does not appear to provide an adequate foundation for empirical analysis using this test (or some of the other procedures). It is, therefore, not surprising that econometricians have turned from the standard asymptotic theory for tests of non-nested regression models and looked to bootstrap techniques in an attempt

to derive procedures that have acceptable properties whether the null model is true or false.

7.3. Bootstrapping tests for models with non-nested regressors

There are several simulation studies in which results are reported for bootstrap tests of a null model against a single non-nested alternative model. These results are summarized in Section 7.3.1 and indicate that bootstrapping gives better control of finite sample significance levels than asymptotic theory. Serious problems of size distortion that are observed when asymptotic critical values are employed are solved effectively, in general, by the application of a simple bootstrap method. However, this is not the only advantage of the bootstrap. As explained in Section 7.2.5, it is not uncommon for applied workers to be faced by several non-nested alternatives. When the individual tests of the null against each of the non-nested alternatives are used in a battery of checks, asymptotic theory fails to provide a standard reference distribution for the induced overall test; see Darroch and Silvey (1963). The bootstrap, however, delivers an easily implemented and asymptotically valid procedure. Evidence relevant to situations in which there are multiple alternatives is discussed in Section 7.3.2. Many of the results are taken from Godfrey (1998).

7.3.1. One non-nested alternative regression model: significance levels

Results that illustrate the potential inadequacy of asymptotic critical values and the usefulness of bootstrap tests, with a single alternative, are provided by simulation experiments reported in Fan and Li (1995). In these experiments, Fan and Li use, as the null model,

$$H_0^{FL} : y_t = x_{t1}\beta_1 + x_{t2}\beta_2 + \beta_3 + u_{t0}, t = 1, \dots, n,$$

and their alternative model has the form

$$H_1^{FL} : y_t = z_{t1}\gamma_1 + z_{t2}\gamma_2 + \gamma_3 + u_{t1}, t = 1, \dots, n,$$

so that, in the notation of the previous section, $k_0 = k_1 = 3$.

The data for simulation experiments are generated using a DGP of a type that has often been used in the literature; see, for example, Delgado and Stengos (1994), Godfrey (1998), Godfrey and Pesaran (1983) and

Pesaran (1982). As a special case of this DGP, Fan and Li specify the following: $\beta_1 = \beta_2 = \beta_3 = 1$;

$$\begin{pmatrix} x_{t1} \\ x_{t2} \\ u_{t0} \end{pmatrix} \sim NID \left(\begin{pmatrix} 0 \\ 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 5 \end{pmatrix} \right), t = 1, \dots, n,$$

which implies an asymptotic R^2 index equal to $2/7$; and $n = 25$. The non-nested regressors in Fan and Li (1995) are obtained using

$$z_{ti} = \lambda x_{ti} + w_{ti}, i = 1, 2,$$

in which the terms w_{ti} are $NID(0, 1)$ and λ is selected to control the population correlation coefficient between x_{ti} and z_{ti} . More precisely, if this population correlation is denoted by ρ , then $\lambda = \rho/\sqrt{(1 - \rho^2)}$. Fan and Li use $\rho = 0.1$ and $\rho = 0.7$, with the former intended to capture near-orthogonality of the type discussed in Michelis (1999).

Fan and Li examine the application of the J test. (They also consider the JA test, but this procedure is exactly valid when the reference distribution is $t(25 - 3 - 1)$, given that the DGP in Fan and Li (1995) has exogenous regressors and NID errors.) The J test is applied as a one-sided test, with large positive values indicating strong evidence against the null model. The asymptotic critical values are taken from the $N(0, 1)$ distribution, as suggested in Davidson and MacKinnon (1981). (Critical values are sometimes taken from the $t(n - k_0 - 1)$ distribution, but there is no formal result to justify this practice, in general.)

The implementation of the bootstrap version of the J test of (7.1) against (7.2), as described in Fan and Li (1995), consists of the following steps.

Fan and Li (1995): Step 1

Use the original data $\mathbf{S} = (\mathbf{y}, \mathbf{X}, \mathbf{Z})$ to compute the OLS estimates of the two non-nested models and to calculate the value of the J -statistic for testing (7.1) against (7.2). The predicted values and residuals from the OLS estimation of (7.1) are denoted by \hat{y}_{t0} and \hat{u}_{t0} , $t = 1, \dots, n$, respectively. The sample value of the J -statistic is denoted by J_{01} .

Steps 2 to 4 involve the generation and analysis of bootstrap samples; this set of steps is repeated B times.

Fan and Li (1995): Step 2

Draw a sequence of bootstrap errors u_{t0}^* , $t = 1, \dots, n$, by random sampling, with replacement, from the EDF of the recentred OLS residuals,

that is, from

$$\hat{F}_0^{FL} : \Pr \left(u_0^* = \hat{u}_{t0} - \frac{1}{n} \sum_{s=1}^n \hat{u}_{s0} \right) = \frac{1}{n}, t = 1, \dots, n. \quad (7.19)$$

Fan and Li (1995): Step 3

Use the bootstrap errors from Step 2 with the OLS predicted values for the null model obtained in Step 1 to derive the bootstrap sample data, with typical observation given by

$$y_t^* = \hat{y}_{t0} + u_{t0}^*.$$

Let $\mathbf{y}^* = (y_1^*, \dots, y_n^*)'$.

Fan and Li (1995): Step 4

Apply the OLS procedures used on the actual data $\mathbf{S} = (\mathbf{y}, \mathbf{X}, \mathbf{Z})$ in Step 1 to the bootstrap data $\mathbf{S}^* = (\mathbf{y}^*, \mathbf{X}, \mathbf{Z})$. Let the bootstrap counterpart of the actual test statistic J_{01} be denoted by J_{01}^* .

Fan and Li (1995): Step 5

After repeating Steps 2–4 B times, Fan and Li use a one-sided bootstrap version of the J test, with the p -value of J_{01} from Step 1 being estimated as

$$\hat{p}_J^{FL} = \frac{\#(J_{01}^* \geq J_{01})}{B}. \quad (7.20)$$

The null model is rejected if $\hat{p}_J^{FL} \leq \alpha_d$, where α_d is the desired significance level.

The finite sample significance levels associated with asymptotic and bootstrap variants of the J test are estimated by Fan and Li, with the former using critical values from the $N(0, 1)$ distribution and the latter being based upon $B = 1,000$ bootstrap samples. Estimates that correspond to desired levels of $\alpha_d = 1$ per cent, 5 per cent and 10 per cent are calculated using $R = 1,000$ replications. With $R = 1,000$ replications, the standard error measures $\sqrt{\alpha_d(100 - \alpha_d)/R}$ are approximately equal to 0.31 per cent, 0.67 per cent and 0.95 per cent for $\alpha_d = 1$ per cent, 5 per cent and 10 per cent, respectively. The results from Fan and Li (1995) are given in Table 7.1.

Table 7.1 Estimated significance levels for asymptotic and bootstrap versions of J test for $n = 25$; see Tables 1 and 2 in Fan and Li (1995, pp. 110–111)

α_d	<i>Asymptotic version</i>		<i>Bootstrap version</i>	
	$\rho = 0.1$	$\rho = 0.7$	$\rho = 0.1$	$\rho = 0.7$
1 per cent	11.7	5.4	1.5	1.2
5 per cent	29.3	14.8	5.9	5.3
10 per cent	44.2	22.8	10.9	10.7

Notes: Each estimate is derived from 1,000 replications and 1,000 bootstrap samples are used to carry out the bootstrap versions of J .

The estimates in Table 7.1 indicate that the asymptotic critical values from the standard Normal distribution provide very poor control of the finite sample significance levels, especially when the non-nested regressors are nearly orthogonal. In contrast, the bootstrap J test performs much better, with quite close agreement with desired values, even when the regressors are only weakly correlated.

The results for bootstrapped J tests that come from the simulation experiments in Fan and Li (1995) are encouraging, but their generality is especially open to question. The errors are always assumed to be $NID(0, 5)$ and, in terms of the numbers of regressors in the competing models, $k_0 = k_1 = 3$ in all experiments. It is argued in Godfrey and Pesaran (1983, p. 144) that it is useful to examine sensitivity to non-Normality and that the finite sample significance levels of the asymptotic J test may deviate from desired values when one or more of the following features are present: (i) a poor fit of the true model (so that variations of the error variance in the model design are of interest); (ii) weak correlations between the regressors of the non-nested models; and (iii) the false alternative model has more regressors than does the true null model, that is, $k_0 < k_1$. Consequently, additional evidence is required before a strong case can be made for bootstrapping the J test (and other procedures) for assessing the validity of regression equations when their regressors are non-nested. The findings from a larger set of experiments are reported in Godfrey (1998).

When estimating significance levels, Godfrey uses the following regression equations as null and alternative models, respectively:

$$H_0^G : y_t = \sum_{i=1}^{k_0} x_{ti}\beta_i + u_{t0}, \quad u_{t0} \text{ IID}(0, \sigma_0^2), \quad (7.21)$$

and

$$H_1^G : y_t = \sum_{i=1}^{k_1} z_{ti} \gamma_i + u_{t1}, \quad u_{t1} \text{ IID}(0, \sigma_1^2), \quad (7.22)$$

for $t = 1, \dots, n$. As in Fan and Li (1995), the regressors x_{ti} of (7.21) are $N(0, 1)$ variables that are independent over both t and i . The regressors z_{ti} for (7.22) are generated using

$$z_{ti} = \lambda x_{ti} + w_{ti}, \quad i = 1, 2, \dots, \min(k_0, k_1), \quad (7.23)$$

and, if $k_0 < k_1$,

$$z_{ti} = w_{ti}, \quad i = k_0 + 1, k_0 + 2, \dots, k_1, \quad (7.24)$$

with the terms w_{ti} being $N(0, 1)$ variables that are independent over both t and i . Following the conventional approach to designing experiments, λ is selected in order to obtain a required value of the population correlation coefficient between x_{ti} and z_{ti} , which is denoted by ρ .

The regression coefficients in (7.21) are all set equal to unity, so $\beta_i = 1, i = 1, \dots, k_0$. The errors of (7.21) are $\text{IID}(0, \sigma_0^2)$, being derived by taking appropriate linear transformations of drawings from one of the following distributions: Normal; LogNormal; and $\chi^2(2)$. The value of σ_0^2 is set to control the population R^2 -statistic, denoted by R_0^2 , by means of the relationship

$$\sigma_0^2 = k_0(1 - R_0^2)/R_0^2.$$

The desired significance level is 5 per cent and finite sample values are estimated using experiments with values of design parameters taken from

$$n = 40, 60,$$

$$\rho = 0.3, 0.6, 0.9,$$

$$R_0^2 = 0.3, 0.6, 0.9,$$

$$(k_0, k_1) = (2, 2), (2, 4), (4, 2), (4, 4).$$

The tests of (7.21) that Godfrey considers are: the F -test against the artificial comprehensive model that corresponds to (7.14); the unadjusted Cox test N ; the J test; the JA test; and the adjusted Cox-type procedures \tilde{N} and W ; see Godfrey and Pesaran (1983) and Godfrey (1998) for details. The first of these tests, denoted by F_{01} , is a standard test

of joint significance of a subset of regressors, with critical values taken from the right-hand tail of the relevant $F(k_1, n - k_0 - k_1)$ distribution. As the tests of the validity of models like (7.21) cannot usually be based upon the assumption that (7.22) is the only possible alternative model, the remaining five tests are all implemented as two-sided procedures, but with some variation in the choice of reference distribution for the critical values. These critical values come from the $t(n - k_0 - 1)$ distribution for the procedures that use an artificial regression, viz. J and JA . The adjusted and unadjusted Cox-type tests (N , \tilde{N} and W) use the $N(0, 1)$ distribution. Estimates of the finite sample significance levels associated with asymptotically valid critical values are derived using $R = 10,000$ replications.

Table 7.2 contains results obtained in the experiments in Godfrey (1998). These results indicate that the F_{01} and JA_{01} tests, both of which are calculated using an artificial model that nests (7.21), are well-behaved in all cases; these tests are, of course, exactly valid under Normal errors. However, the J_{01} test, which also comes from an artificial nesting model, does not have rejection frequencies that are close to the desired level of 5 per cent, with some estimates being between two and three times the desired value. The failings of the asymptotic distribution to serve as a

Table 7.2 Estimated significance levels using asymptotic critical values for cases with (7.21) as the null model, (7.22) as the alternative model, $R_0^2 = 0.6$, $\rho = 0.3$ and $n = 40$

Design parameters		Test					
(k_0, k_1)	Error dbn.	F_{01}	J_{01}	JA_{01}	W_{01}	\tilde{N}_{01}	N_{01}
(2, 2)	Normal	5.3	7.0	5.2	4.6	5.2	12.7
(2, 2)	LogNormal	5.2	6.8	5.3	4.7	5.4	11.8
(2, 2)	$\chi^2(2)$	4.7	6.2	5.0	4.4	4.9	11.5
(2, 4)	Normal	5.3	15.1	5.0	4.5	5.1	17.4
(2, 4)	LogNormal	5.2	13.4	5.1	4.2	4.8	16.0
(2, 4)	$\chi^2(2)$	4.8	14.2	5.0	4.1	4.9	17.1
(4, 2)	Normal	5.6	7.7	5.1	4.2	4.8	14.5
(4, 2)	LogNormal	5.2	7.3	5.0	4.4	4.9	13.8
(4, 2)	$\chi^2(2)$	5.2	7.3	5.1	4.3	4.8	14.3
(4, 4)	Normal	5.2	13.4	5.0	3.7	4.5	16.4
(4, 4)	LogNormal	5.2	12.5	5.2	3.8	4.6	15.1
(4, 4)	$\chi^2(2)$	5.3	13.7	5.2	4.2	4.8	16.7

Notes: Estimates are based upon 10,000 replications and are rounded to one decimal place. The desired significance level is 5 per cent in all cases. Source: Godfrey (1998, p. 67, Table 2).

useful approximation to the finite sample distribution of the J test is a serious problem, given the widespread use of this test.

The unadjusted Cox-statistic N_{01} is even more badly behaved in terms of the too frequent rejection of the true null model. In contrast, the adjusted variants W_{01} and \tilde{N}_{01} do not suffer from over-rejection and, if anything, the former test is a little under-sized. Although the W_{01} and \tilde{N}_{01} tests are quite well behaved for all distributions used in the experiments, their theoretical justification rests upon an assumption of Normality.

The estimated significance levels obtained in Fan and Li (1995) and in Godfrey (1998) show that asymptotic theory does not provide a generally reliable basis for inference when either the unadjusted Cox-type N -test or the more popular J test is used to test models with non-nested regressors. Godfrey, like Fan and Li, examines the possibility that bootstrapping these tests might produce better agreement between finite sample and desired significance levels. His approach is similar to the five-step method used in Fan and Li (1995), which is summarized above. There are the following differences: (i) Godfrey applies the bootstrap to the artificial comprehensive model test F_{01} , as well as to the five tests that have their origins in Cox (1961, 1962); (ii) when bootstrapping the latter group of tests, a two-sided alternative is used, so that the p -value in (7.20) is replaced by

$$\hat{p}_J^G = \frac{\#(|J_{01}^*| \geq |J_{01}|)}{B} = \frac{\#((J_{01}^*)^2 \geq (J_{01})^2)}{B}, \quad (7.25)$$

with the estimates for JA , N , \tilde{N} and W being defined in a similar way; and (iii) as well as using the residual resampling scheme in Step 2 of the Fan-Li procedure, Godfrey employs degrees-of-freedom and leverage adjustments, which are described in Godfrey (1998, p. 68).

It is found in Godfrey (1998) that the alternative resampling methods of (iii) do not produce results that differ in important ways from the simple scheme (7.19). However, some authors prefer the use of adjusted residuals when implementing an IID bootstrap. The results reported in Godfrey (1998) for bootstrap tests are based upon the following bootstrap world CDF:

$$\hat{F}_0^G : \Pr(u_0^* = a + b\hat{u}_{t0}^a) = \frac{1}{n}, \quad t = 1, \dots, n, \quad (7.26)$$

in which: \hat{u}_{t0}^a is a leverage-adjusted residual, as used in (2.32), defined by

$$\hat{u}_{t0}^a = \frac{\hat{u}_{t0}}{\sqrt{(1 - h_{tt})}},$$

where $1 - h_{tt}$ is a typical diagonal element of \mathbf{M}_X ; and the constants a and b are chosen so that

$$E^*(u_0^*) = \frac{1}{n} \sum_{t=1}^n (a + b\hat{u}_{t0}^a) = 0,$$

and

$$\text{Var}^*(u_0^*) = \frac{1}{n} \sum_{t=1}^n (a + b\hat{u}_{t0}^a)^2 = \frac{1}{n - k_0} \sum_{t=1}^n \hat{u}_{t0}^2 = s_0^2, \text{ say.}$$

Following Horowitz (1994), Godfrey bases the bootstrap tests on the outcomes of $B = 100$ artificial samples for each of the $R = 10,000$ replications of a given experiment. However, in view of the advances in low-cost computing since the publication of Godfrey (1998), it would involve very little waiting time to use $B = 1,000$ in actual applied work. Table 7.3 shows what happens in the cases covered by Table 7.2 when asymptotic tests are replaced by bootstrap tests.

Table 7.3 Estimated significance levels using bootstrap p -values for cases with (7.21) as the null model, (7.22) as the alternative model, $R_0^2 = 0.6$, $\rho = 0.3$ and $n = 40$

<i>Design parameters</i>		<i>Test</i>					
(k_0, k_1)	<i>Error dbn.</i>	F_{01}	J_{01}	JA_{01}	W_{01}	\tilde{N}_{01}	N_{01}
(2, 2)	Normal	4.6	4.9	4.9	5.0	5.0	4.9
(2, 2)	LogNormal	4.9	5.0	5.1	5.0	5.1	5.0
(2, 2)	$\chi^2(2)$	4.8	4.9	4.8	4.7	4.6	4.6
(2, 4)	Normal	4.8	5.2	4.9	5.0	5.1	5.0
(2, 4)	LogNormal	5.0	5.0	4.9	4.9	5.0	4.8
(2, 4)	$\chi^2(2)$	4.5	5.0	5.3	4.9	5.0	4.8
(4, 2)	Normal	5.2	5.2	5.2	5.1	5.2	4.9
(4, 2)	LogNormal	5.5	5.1	5.4	5.0	5.1	5.1
(4, 2)	$\chi^2(2)$	4.8	5.1	5.2	5.3	5.4	5.1
(4, 4)	Normal	4.7	5.0	4.7	4.7	4.8	4.9
(4, 4)	LogNormal	5.0	4.9	4.8	4.6	4.6	4.8
(4, 4)	$\chi^2(2)$	5.4	5.3	5.2	5.2	5.3	5.4

Notes: Estimates are based upon 10,000 replications and 100 bootstrap samples. They are rounded to one decimal place. The desired significance level is 5 per cent in all cases. *Source:* Godfrey (1998, p. 70, Table 3).

Comparison of the estimates in Table 7.3 with those in Table 7.2 provides clear evidence of the way in which bootstrapping can produce agreement between actual and desired significance levels that is superior to that obtained using asymptotic critical values. The size distortions of the N and J tests are essentially eliminated. All of the estimates in Table 7.3 are in the range 4.5 per cent to 5.5 per cent and so they indicate close agreement with the desired significance level of 5 per cent. Since the estimated significance levels of the bootstrap versions of the tests are close to the desired value, the differences between them are not so large as to cast doubt upon power comparisons; see Horowitz and Savin (2000) for a discussion of when empirically relevant power comparisons can be made.

7.3.2. One non-nested alternative regression model: power

Table 7.4 contains power estimates for the bootstrap tests. These estimates are obtained by testing (7.21) when data are generated by (7.22), not the other way around; see Godfrey (1984, pp. 76–77) for comments relevant to this choice of experimental design. When the data are generated using (7.22), the coefficients are selected as follows: $\gamma_i = 1$, for

Table 7.4 Power estimates using bootstrap tests for cases with (7.21) as the false null model, (7.22) as the true alternative model, $R_1^2 = 0.3$, and $n = 40$

Design parameters		Test					
(k_0, k_1)	ρ	F_{01}	J_{01}	JA_{01}	W_{01}	\tilde{N}_{01}	N_{01}
(2,2)	0.3	88.4	90.0	66.1	91.8	91.9	83.4
(2,2)	0.6	76.5	83.3	72.7	86.4	86.4	84.6
(2,2)	0.9	29.5	40.0	36.2	42.0	44.5	46.1
(2,4)	0.3	82.4	84.8	41.0	85.4	85.5	70.8
(2,4)	0.6	72.8	80.4	43.8	81.4	81.8	70.0
(2,4)	0.9	33.6	47.5	26.4	42.8	45.4	41.3
(4,2)	0.3	86.5	88.5	70.6	91.0	91.1	86.4
(4,2)	0.6	73.9	80.8	71.4	83.9	84.5	82.4
(4,2)	0.9	27.7	37.7	34.0	40.2	43.0	44.1
(4,4)	0.3	77.8	82.4	48.2	84.0	84.2	83.2
(4,4)	0.6	60.6	74.7	56.6	75.4	76.5	78.2
(4,4)	0.9	21.1	40.1	29.8	35.8	38.9	42.8

Notes: Estimates are based upon 2,500 replications and 100 bootstrap samples. They are rounded to one decimal place. Source: Godfrey (1998, p. 71, Table 4).

$i = 1, \dots, k_1$; and the error variance σ_1^2 is selected by choosing a value for the population coefficient of determination, according to

$$\begin{aligned}\sigma_1^2 &= k_1(1 + \lambda^2)(1 - R_1^2)/R_1^2 \text{ if } k_0 \geq k_1, \\ &= (k_0\lambda^2 + k_1)(1 - R_1^2)/R_1^2 \text{ if } k_0 < k_1,\end{aligned}$$

with $R_1^2 = (0.3, 0.6, 0.9)$. The power estimates reported in Table 7.4 are derived using $R = 2,500$ replications; so that the maximum standard error is 1 per cent, corresponding to a true power of 50 per cent. The desired significance level of all tests of the false model (7.21) equals 5 per cent.

The cases used to obtain the results in Table 7.4 correspond to power estimates in an interesting range, that is, not close to either 5 per cent or 100 per cent. As in studies of asymptotic tests, the bootstrap forms of the comprehensive model test F_{01} and the Fisher-McAleer test JA_{01} are sometimes observed to be less powerful than the bootstrap variants of the tests proposed in Pesaran (1974), Davidson and MacKinnon (1981) and Godfrey and Pesaran (1983), viz. N_{01} , J_{01} , \tilde{N}_{01} and W_{01} . The relative performance of the bootstrapped version of N_{01} is a little variable and this test does not seem to offer any advantage over using one of J_{01} , \tilde{N}_{01} and W_{01} . The differences between estimates for J_{01} , \tilde{N}_{01} and W_{01} in Table 7.4 are small, but they show that \tilde{N}_{01} is slightly better than W_{01} and that, in most cases, J_{01} is outperformed by \tilde{N}_{01} and W_{01} . However, the estimates do not suggest a serious degree of inferiority of the bootstrap J test, which has the advantages of being very much easier to implement and motivate than the modified Cox-tests \tilde{N}_{01} and W_{01} . The J test also has the advantage that it is more amenable to theory-based analysis of its finite sample behaviour.

7.3.3. One non-nested alternative regression model: extreme cases

Theoretical analysis can sometimes be used to obtain results that provide a framework for interpreting simulation-based evidence and can be used to design experiments that provide stringent checks of the usefulness of bootstrap tests. In Godfrey and Pesaran (1983), the classical assumptions of exogenous regressors and *NID* errors are used to obtain mean and variance adjustments for the basic form of the Cox-type statistic N_{01} , when testing (7.1) against (7.2), and to identify situations in which asymptotic tests might behave badly. These assumptions are employed in Davidson and MacKinnon (2002a) to derive a much fuller analysis of the finite

sample properties of the J test. The theoretical analysis that Davidson and MacKinnon provide serves two purposes: first, it explains when and why the asymptotic distribution of the test statistic provides a poor approximation; and second, it helps to identify certain unusual cases in which bootstrapping will not yield very good control of significance levels.

Davidson and MacKinnon find that a key factor in determining the finite sample distribution of J_{01} , when (7.1) is true, is the scalar

$$\Delta_0^2 = \frac{\beta'X'P_ZM_XP_ZX\beta}{\sigma_0^2}; \quad (7.27)$$

see Davidson and MacKinnon (2002a, Theorem 1). The larger the value of Δ_0^2 , the closer the finite sample distribution of J_{01} is to $N(0, 1)$. While this finding is obtained under the assumptions of exogenous regressors and *NID* errors, simulation evidence reported in Davidson and MacKinnon (2002a) indicates that there is a useful degree of robustness to the inclusion of a lagged dependent variable in the regressors and non-Normality of the errors.

The quantity Δ_0^2 in (7.27) has a simple interpretation. When H_0 in (7.1) is true, the OLS estimator of δ in the artificial regression model

$$\mathbf{y} = \delta [\mathbf{P}_Z\mathbf{X}\boldsymbol{\beta}] + \mathbf{X}\boldsymbol{\beta} + \mathbf{u}_0, \quad \mathbf{u}_0 \sim N(\mathbf{0}_n, \sigma_0^2\mathbf{I}_n), \quad (7.28)$$

is distributed as $N(0, 1/\Delta_0^2)$; so that Δ_0^2 is a natural measure of the precision of this estimator. The vector $\mathbf{P}_Z\mathbf{X}\boldsymbol{\beta}$ is obviously unobservable, but it is easy to verify that it is the expected value, under the null model, of both $\hat{\mathbf{y}}_1$ and $\mathbf{P}_Z\hat{\mathbf{y}}_0$, which links the test variable in (7.28) to those in (7.12) and (7.13).

Davidson and MacKinnon use their theoretical analysis to design simulation experiments that represent extreme cases; see Davidson and MacKinnon (2002b). In these extreme cases, the widely-used asymptotic J test often rejects a true null model more than 50 per cent of the time and the single bootstrap does not fully correct this problem. Consequently, as well as studying the performance of the single bootstrap version of the J test, Davidson and MacKinnon investigate the usefulness of their Fast Double bootstrap (FDB) in these extreme cases.

Davidson and MacKinnon obtain B (first-level) bootstrap samples, according to

$$\mathbf{y}_{(b)}^* = \hat{\mathbf{y}}_0 + \mathbf{u}_{(b)}^*,$$

in which the n elements of the bootstrap error vector $\mathbf{u}_{(b)}^*$ are derived by random sampling, with replacement, from

$$\hat{F}_0^{DM} : \Pr \left(u_0^* = \sqrt{\frac{n}{n-k_0}} (\hat{u}_{t0} - \frac{1}{n} \sum_{s=1}^n \hat{u}_{s0}) \right) = \frac{1}{n}, t = 1, \dots, n,$$

for $b = 1, \dots, B$. The degrees-of-freedom adjustment employed in \hat{F}_0^{DM} is useful in experiments in Davidson and MacKinnon (2002b) for which k_0/n is not small. Davidson and MacKinnon do not find any gain associated with more complex adjustments of the OLS residuals, for example, using leverage values. Given the values of the bootstrap statistics J_{01}^* from the B generated samples, the p -value of the actual test statistic J_{01} can be calculated using either (7.20) or (7.25), depending upon whether a one-sided or two-sided test is required.

The FDB procedure is implemented by generating a single second-level bootstrap sample from each of the first-level bootstrap samples $\mathbf{S}_{(b)}^* = (\mathbf{y}_{(b)}^*, \mathbf{X}, \mathbf{Z})$, $b = 1, \dots, B$. These second-level samples are obtained by treating the first-level bootstrap data as if they were actual data; see Davidson and MacKinnon (2002b, section 3) for details. An adjusted p -value can then be computed, as explained in Section 2.5 above. Davidson and MacKinnon explain why the FDB approach is likely to be more accurate than the single bootstrap approach, while being much cheaper to carry out than the standard double bootstrap. The results from the experiments used in Davidson and MacKinnon (2002b) show that the FDB version of the J test works remarkably well.

There are assumptions that permit perfect control of the significance levels of the J test. Luger sets out conditions under which exact permutation tests can be derived; see Luger (2006) for specific details for tests of non-nested models and Kennedy (1995) for a more general discussion. For the case of linear regression models, Luger's assumptions can be explained as follows. Let the models under consideration be written as

$$H_0^L : y_t = \mathbf{x}'_{t0} \boldsymbol{\beta}_0 + \mathbf{w}'_t \boldsymbol{\psi}_0 + u_{t0}, t = 1, \dots, n,$$

and

$$H_1^L : y_t = \mathbf{z}'_{t1} \boldsymbol{\gamma}_1 + \mathbf{w}'_t \boldsymbol{\psi}_1 + u_{t1}, t = 1, \dots, n,$$

in which \mathbf{x}_{t0} and \mathbf{z}_{t1} contain observations on the (non-overlapping) regressors specific to H_0^L and H_1^L , respectively, and \mathbf{w}_t is a vector of observations on (overlapping) variables that are common to the two

models. An exact J test of H_0^L against H_1^L is available if the following two assumptions are satisfied:

1. Either $\{y_t, \mathbf{x}_{t0}, \mathbf{w}_t; t = 1, \dots, n\}$ or $\{z_{t1}; t = 1, \dots, n\}$ is a collection of exchangeable random variables; and
2. The vectors $\{y_t, \mathbf{x}_{t0}, \mathbf{w}_t; t = 1, \dots, n\}$ are independent of $\{z_{t1}; t = 1, \dots, n\}$.

The notion of “exchangeable random variables” used in the first assumption is that, for such variables, the joint probability density function is not altered when the variables are permuted, that is, under shuffles of the observations. Thus the assumption that the variables are IID is sufficient (but not necessary) for them to be exchangeable. The single bootstraps and FDB methods proposed in Fan and Li (1995), Godfrey (1998) and Davidson and MacKinnon (2002b) do not require these two strong assumptions about $\{y_t, \mathbf{x}_{t0}, \mathbf{w}_t, z_{t1}; t = 1, \dots, n\}$, but only enjoy asymptotic, not exact, validity. These assumptions are not made below, but the implied loss of control associated with relying upon bootstrap techniques is not likely to be important. On the basis of the results from their simulation experiments, Davidson and MacKinnon conclude that “In practice, we would expect the single bootstrap to work extremely well, and our FDB procedure to work nearly perfectly, in virtually every case that an econometrician would be likely to encounter” (Davidson and MacKinnon, 2002b, p. 428).

7.3.4. Two non-nested alternative regression models: significance levels

The bootstrap methods described above are recommended for application when there are two competing models with non-nested regressor sets. As noted in McAleer (1995), applied workers are sometimes in a situation in which there are three or more non-nested models under consideration. In such a situation, if (7.1) is taken to be the null model to be tested, the $m > 1$ alternatives can be written as

$$\mathbf{y} = \mathbf{Z}_j \boldsymbol{\gamma}_j + \mathbf{u}_j, \mathbf{u}_j \sim IID(\mathbf{0}_n, \sigma_j^2 \mathbf{I}_n), j = 1, \dots, m. \quad (7.29)$$

Simulation experiments that provide evidence for the special case of $m = 2$ non-nested alternatives are described in Godfrey (1998, section 3.2).

In his experiments for the estimation of significance levels, Godfrey uses (7.21) as the null model and the two alternative models have $k_1 = k_2 = 2$ regressors. The first alternative has $\{z_{t1}, z_{t3}; t = 1, \dots, n\}$ as regressor

values and the second has $\{z_{t2}, z_{t4}; t = 1, \dots, n\}$, with the data on these regressors being obtained using (7.23) and (7.24). The test statistics that Godfrey examines are as follows.

First, the comprehensive F test statistic, denoted by F_0^C , is derived using as the alternative

$$y_t = \sum_{i=1}^{k_0} x_{ti}\beta_i + \sum_{i=1}^4 z_{ti}\gamma_i + u_{t0}, \quad u_{t0} \text{ IID}(0, \sigma_0^2), \quad t = 1, \dots, n,$$

with the asymptotic test being based upon the $F(4, n - k_0 - 4)$, $k_0 = 2, 4$, distributions. Second, in order to examine the joint version of the J test, derived from an artificial model corresponding to (7.17), an asymptotically valid F test of (7.21) against

$$y_t = \sum_{i=1}^{k_0} x_{ti}\beta_i + \delta_1 \hat{y}_{t1} + \delta_2 \hat{y}_{t2} + u_{t0}, \quad u_{t0} \text{ IID}(0, \sigma_0^2), \quad t = 1, \dots, n,$$

is carried out, with the test statistic being written as F_0^J . The asymptotic critical values for F_0^J are taken from the $F(2, n - k_0 - 2)$, $k_0 = 2, 4$, distributions.

Joint tests for the adjusted Cox-type tests \tilde{N} and W , which are recommended for a single alternative in Godfrey and Pesaran (1983), cannot be implemented so easily. McAleer observes that it is common practice for applied workers to calculate a binary test of the null against each of the alternatives and to reject the null unless all of the separate binary tests are statistically insignificant. This practice is equivalent to using the most extreme of the m separate binary test statistics, or equivalently the minimum of the p -values for these test statistics, as the overall test criterion. The statistics $Sup\tilde{N}_0^2 = \max(\tilde{N}_{01}^2, \tilde{N}_{02}^2)$ and $SupW_0^2 = \max(W_{01}^2, W_{02}^2)$, therefore, merit consideration and are included in the discussion below.

Conventional tables of critical values are not available when either $Sup\tilde{N}_0^2$ or $SupW_0^2$ is used. As explained in the previous section, the maximum of the test statistics does not have a standard asymptotic null distribution. However, as pointed out in MacKinnon (2007), "One of the big advantages of bootstrap testing is that ... it can easily be used to assign a P value to the maximum of a possibly large number of test statistics." Consequently, it is of interest to investigate the application of the bootstrap to $Sup\tilde{N}_0^2$ and $SupW_0^2$ in the experiments with two alternatives in Godfrey (1998).

Estimated significance levels for valid asymptotic tests, invalid asymptotic tests and bootstrap tests are presented in Table 7.5. As in Tables 7.2

Table 7.5 Estimates of significance levels with two alternative models for cases with Normal errors, $k_0 = 2$, and $n = 40$

Design coefficients		Test							
		F_0^C		F_0^J		$Sup\tilde{N}_0^2$		$SupW_0^2$	
R_0^2	ρ	(a)	(d)	(b)	(d)	(c)	(d)	(c)	(d)
0.3	0.3	4.8	4.5	12.6	4.4	10.0	4.9	8.7	4.9
0.6	0.3	5.2	5.1	8.9	4.9	9.7	4.9	8.7	4.9
0.9	0.3	5.1	5.0	5.4	4.4	8.5	4.5	8.3	4.6
0.3	0.6	5.2	5.1	9.6	5.0	9.8	5.3	8.3	5.3
0.6	0.6	4.6	4.6	6.4	4.8	9.1	4.8	7.9	4.7
0.9	0.6	5.1	5.1	5.4	5.3	9.4	5.3	9.1	5.1
0.3	0.9	5.0	4.8	8.2	4.9	8.2	4.9	7.3	5.0
0.6	0.9	4.8	4.9	6.7	4.6	8.7	4.9	8.3	4.9
0.9	0.9	5.4	5.5	5.5	5.2	9.7	5.2	9.4	5.0

Notes: Estimates are based upon 10,000 replications and 100 bootstrap samples. They are rounded to one decimal place. The critical values are as follows: (a) $F(4, 34)$; (b) $F(2, 36)$; (c) $\chi^2(1)$; and (d) bootstrap. Source: Godfrey (1998, p. 74, Table 5).

and 7.3, the estimates are calculated using 10,000 replications and correspond to a desired significance level of 5 per cent. All cases in Table 7.5 are for true null models with *NID* errors. The comprehensive model check of F_0^C is, therefore, exactly valid and the estimates for F_0^C in Table 7.5 are not surprisingly close to 5 per cent. The asymptotic joint *J* test, based upon F_0^J , behaves quite well when $R_0^2 = 0.9$. However, the quality of the asymptotic approximation deteriorates as R_0^2 decreases, with $R_0^2 = 0.3$ producing evidence of substantial size distortions. The asymptotically invalid tests in which $Sup\tilde{N}_0^2$ and $SupW_0^2$ are compared with the 5 per cent critical value of the $\chi^2(1)$ distribution lead to estimates that fall within the Bonferroni bounds of 5 per cent and 10 per cent, but are not close to the former, which is the desired value. Application of the single bootstrap gives good control of significance levels for all procedures, even the non-asymptotically pivotal tests based upon $Sup\tilde{N}_0^2$ and $SupW_0^2$. The use of the bootstrap is, therefore, recommended when there are multiple non-nested alternatives, whichever of the approaches to testing is used.

7.3.5. Two non-nested alternative regression models: power

The finding that the bootstrap can be used to fix the finite sample significance level when checking the maximum of a set of binary test statistics

does not imply that this method gives better results than a joint test. In particular, it cannot be guaranteed that a bootstrap *Sup*-test is more powerful than a joint test. In any case, it should not be assumed that, in the event of the null model being rejected, the alternative that yields the most extreme test statistic is the correct specification.

In order to derive some evidence concerning power properties, the estimation of rejection probabilities when the null model is false is based upon data generated using

$$H_1^m : y_t = z_{t1} + z_{t3} + u_{t1}, t = 1, \dots, n,$$

in which the errors u_{t1} are $NID(0, \sigma_1^2)$ and the parameter σ_1^2 is controlled via the population R^2 for H_1^m , according to

$$\sigma_1^2 = (2 + \lambda^2)(1 - R_1^2)/R_1^2,$$

when $k_0 = 2$, and

$$\sigma_1^2 = 2(1 + \lambda^2)(1 - R_1^2)/R_1^2,$$

when $k_0 = 4$, with $R_1^2 = (0.3, 0.6, 0.9)$. Results obtained using 2,500 replications are given in Table 7.6.

The estimates in Table 7.6 come from a very small set of experiments and it is hoped that more extensive experiments will be carried out in future research. However, these estimates do indicate that the bootstrapped joint J test, denoted by F_0^J , outperforms the bootstrapped

Table 7.6 Estimates of power of bootstrap tests with two alternative models for cases with Normal errors, $R_1^2 = 0.3$ and $n = 40$

<i>Design parameters</i>		<i>Test</i>			
ρ	k_0	F_0^C	F_0^J	$Sup\tilde{N}_0^2$	$SupW_0^2$
0.3	2	81.4	82.8	88.0	87.9
0.6	2	72.3	78.0	82.6	82.5
0.9	2	33.9	43.8	44.4	42.3
0.3	4	76.0	80.1	85.3	85.1
0.6	4	61.4	69.6	77.7	77.2
0.9	4	19.6	29.0	36.8	34.2

Notes: Estimates are based upon 2,500 replications and 100 bootstrap samples. They are rounded to one decimal place.

comprehensive model test F_0^C . Both of these easily implemented F tests are estimated to be less powerful than the bootstrapped $Sup\tilde{N}_0^2$ and $SupW_0^2$ procedures. The results in Tables 7.5 and 7.6, therefore, suggest that bootstrap methods are useful when testing a null model in the presence of several non-nested alternatives and that joint tests that could use standard asymptotic distributions may be inferior to bootstrapped Sup -type criteria.

7.4. Bootstrapping the LLR statistic with non-nested models

An essential part of the famous procedure described in Cox (1961, 1962) is the derivation of the mean and variance of the LLR statistic in order to obtain a standardized variate that, under the null hypothesis, is asymptotically distributed as $N(0, 1)$. Unfortunately, in many cases, it is very difficult to obtain analytical expressions for the mean and variance. The absence of such difficulties in the analyses of previous sections reflects the fact that the non-nested regression models share the same dependent variable and IID error model. In practice, it may be necessary to deal with other types of non-nested regressions.

First, it is possible that the dependent variable of one model is a known one-to-one transformation of the dependent variable of another model; see Yeo (2005) for a more general discussion of transformations in regression. For example, there is a substantial literature on the problem of testing linear and log-linear specifications of the general forms

$$H_0^L : y_t = \sum_{i=1}^{k_0} x_{ti}\beta_i + u_{t0}, \quad u_{t0} \text{ IID}(0, \sigma_0^2),$$

and

$$H_0^{LL} : \ln(y_t) = \sum_{i=1}^{k_1} z_{ti}\gamma_i + u_{t1}, \quad u_{t1} \text{ IID}(0, \sigma_1^2);$$

see the references in Godfrey and Santos Silva (2004).

Second, the models may have the same dependent variable and regressors, but have non-nested error models. For example, the first-order autoregressive and moving schemes

$$u_{t0} = \phi u_{t-1,0} + \epsilon_{t0}, \quad |\phi| < 1, \quad \epsilon_{t0} \text{ IID}(0, \zeta_0^2),$$

and

$$u_{t1} = \epsilon_{t1} + \theta\epsilon_{t-1,1}, \quad |\theta| < 1, \quad \epsilon_{t1} \text{ IID}(0, \zeta_1^2),$$

might be competing non-nested specifications of the error model; see Walker (1967, 1970) for a detailed analysis of tests for non-nested time series models.

The purpose of this section is to examine a bootstrap technique that has been suggested for application outside the simple framework in which models are linear regressions that are only non-nested in their regressor sets. Various methods for overcoming the problem of analytical intractability have been proposed in the literature. Some authors have recommended the use of simulation-based estimates; see, for example, Monfardini (2003), Pesaran and Pesaran (1993, 1995) and Yeo (2005). Other researchers have moved some distance from the original ideas in Cox (1961, 1962) in order to obtain tests that at least have the virtue of being convenient to implement in empirical work; see Baltagi and Li (1995) and MacKinnon et al. (1983, p. 56). A third approach is simply to bootstrap the raw LLR statistic; see, for example, Coulibaly and Brorsen (1999) and Kim et al. (1998, section 5.2). It is this third approach that is discussed in this section.

Suppose that, as discussed in previous applications of the bootstrap, B bootstrap samples have been generated, using a bootstrap DGP with parameter vector equal to an estimated parameter vector for the null model. As in Section 7.2, let the LLR statistic for the actual data be denoted by \hat{L}_{01} . The corresponding bootstrap statistics are denoted by $\hat{L}_{01(b)}^*$, $b = 1, \dots, B$. The proportion of bootstrap values $\hat{L}_{01(b)}^*$ less than or equal to the actual value \hat{L}_{01} is then

$$prop = \frac{\#(\hat{L}_{01(b)}^* \leq \hat{L}_{01})}{B}. \quad (7.30)$$

A small sample adjustment in which one is added to both the numerator and denominator of (7.30) is recommended in Coulibaly and Brorsen (1999), but this adjustment is unimportant for reasonable values of B . The important issue is whether or not (7.30) delivers an asymptotically valid p -value.

The application of the bootstrap to the LLR statistic can be illustrated by using the case of non-nested linear regression models with NID errors, which was discussed in Section 7.2. In this special case, when (7.1) is to

be tested against (7.2), the actual LLR statistic is

$$\hat{L}_{01} = \frac{n}{2} \ln \left(\frac{\mathbf{y}' \mathbf{M}_Z \mathbf{y}}{\mathbf{y}' \mathbf{M}_X \mathbf{y}} \right),$$

and, given suitable bootstrap samples $\mathbf{y}^*_{(b)}$ from

$$\mathbf{y}^* \sim N(\mathbf{X}\hat{\boldsymbol{\beta}}, \hat{\sigma}_0^2 \mathbf{I}_n),$$

the corresponding bootstrap LLR statistics are

$$\hat{L}^*_{01(b)} = \frac{n}{2} \ln \left(\frac{\mathbf{y}^{*'}_{(b)} \mathbf{M}_Z \mathbf{y}^*_{(b)}}{\mathbf{y}^{*'}_{(b)} \mathbf{M}_X \mathbf{y}^*_{(b)}} \right),$$

for $b = 1, \dots, B$.

All previous attempts to use the bootstrap have been concerned with test statistics that had proper asymptotic distributions, at least after appropriate centering and scaling. In contrast, \hat{L}_{01} is such that, under the null, $n^{-1}\hat{L}_{01}$ converges in probability to a constant, that is, $n^{-1}\hat{L}_{01}$ has a degenerate asymptotic null distribution. In order to transform to obtain variates with proper asymptotic distributions, it is useful to note that the numerator of (7.30) can be written as

$$\#(\hat{L}^*_{01(b)} \leq \hat{L}_{01}) = \#(\sqrt{n} [n^{-1}\hat{L}^*_{01(b)} - \hat{\mu}_{01}] \leq \sqrt{n} [n^{-1}\hat{L}_{01} - \hat{\mu}_{01}]),$$

where $\hat{\mu}_{01}$ is as given in (7.7), that is,

$$\hat{\mu}_{01} = \frac{1}{2} \ln \left(\frac{\hat{\sigma}_0^2 + \hat{\boldsymbol{\beta}}' \hat{\boldsymbol{\Lambda}} \hat{\boldsymbol{\beta}}}{\hat{\sigma}_0^2} \right).$$

Under regularity conditions, $T_{01} = \sqrt{n} [n^{-1}\hat{L}_{01} - \hat{\mu}_{01}]$ has an asymptotic null distribution of the form $N(0, V_{01})$, $0 < V_{01} < \infty$. The consistency of the bootstrapped LLR test, therefore, requires that, in the bootstrap world, $\sqrt{n} [n^{-1}\hat{L}^*_{01(b)} - \hat{\mu}_{01}]$ has the same asymptotic distribution. However, in the bootstrap world, $\hat{\boldsymbol{\beta}}$ and $\hat{\sigma}_0^2$, which appear in the expression for $\hat{\mu}_{01}$, are the counterparts of $\boldsymbol{\beta}$ and σ_0^2 , respectively, in the model assumed to generate the actual data. The terms

$$\sqrt{n} \left[n^{-1}\hat{L}_{01} - \frac{1}{2} \ln \left(\frac{\hat{\sigma}_0^2 + \hat{\boldsymbol{\beta}}' \hat{\boldsymbol{\Lambda}} \hat{\boldsymbol{\beta}}}{\hat{\sigma}_0^2} \right) \right],$$

and

$$\sqrt{n} \left[n^{-1} \hat{L}_{01} - \frac{1}{2} \ln \left(\frac{\sigma_0^2 + \beta' \hat{\Lambda} \beta}{\sigma_0^2} \right) \right],$$

do not have the same asymptotic null distribution because

$$\sqrt{n} \left[\frac{1}{2} \ln \left(\frac{\hat{\sigma}_0^2 + \hat{\beta}' \hat{\Lambda} \hat{\beta}}{\hat{\sigma}_0^2} \right) - \frac{1}{2} \ln \left(\frac{\sigma_0^2 + \beta' \hat{\Lambda} \beta}{\sigma_0^2} \right) \right]$$

is not asymptotically negligible. It follows that (7.30) does not yield an asymptotically valid p -value when two non-nested linear regression models with NID errors are under scrutiny.

Results on the asymptotic properties of the bootstrapped LLR procedure in a more general framework are provided in Godfrey (2007a). Godfrey argues that $\sqrt{n} \left[n^{-1} \hat{L}_{01(b)}^* - \hat{\mu}_{01} \right]$ has a bootstrap world asymptotic null distribution that is Normal with the correct (zero) mean, but the wrong variance relative to that of the asymptotic distribution of $\sqrt{n} \left[n^{-1} \hat{L}_{01} - \hat{\mu}_{01} \right]$ under the null model; see Godfrey (2007a, p. 411). Simulation evidence on the consequences of using the bootstrapped LLR method in the context of testing linear and log-linear functional forms in regression analysis is given in Godfrey and Santos Silva (2004). The results reported in Godfrey and Santos Silva (2004) include examples of important departures from the desired significance level.

7.5. Summary and concluding remarks

In a review of applications of bootstrap techniques in econometrics, Jeong and Maddala suggest that there are two main uses of the bootstrap that have firm theoretical foundations and support from empirical studies or simulation experiments; see Jeong and Maddala (1993, p. 575). First, the bootstrap often gives better approximations than asymptotic theory when the latter is tractable, but fails to give an acceptable level of accuracy for sample sizes of a magnitude of interest to applied workers. Second, when asymptotic theory is not tractable, the bootstrap can sometimes offer asymptotically valid procedures that are reasonably well-behaved in finite samples.

In this chapter, tests of non-nested regression models have been discussed and examples of both of the uses identified by Jeong and Maddala have been provided. The testing of a null model against a single non-nested alternative yields evidence of the value of the bootstrap when

asymptotic theory is tractable and provides tests that are easy to carry out, but is of very doubtful relevance for sample sizes of interest. The example of the widely-used J test, proposed in Davidson and MacKinnon (1981), is especially clear. This procedure is easily motivated, requires only a t -test after OLS estimation of an artificial regression model and uses asymptotic critical values from the familiar $N(0, 1)$ distribution. However, these asymptotic critical values can lead to extremely misleading inferences. MacKinnon reports that, when carried out with a desired significance level of 5 per cent, the J test can reject a true model more than 80 per cent of the time, even with a sample size of 50; see MacKinnon (2002, p. 617).

The results discussed in Section 7.3 suggest that the bootstrap is remarkably effective in removing the excess rejection frequency of the J test and achieves the same outcome for the unadjusted Cox-type test derived in Pesaran (1974), which, if anything, can be even more badly behaved than the J test when asymptotic critical values are used. Davidson and MacKinnon have used theoretical analysis to design simulation experiments that produce extreme cases of the bad behaviour of the asymptotic J test. In such extreme cases, asymptotic critical values are very poor approximations to finite sample values, a single bootstrap gives much better control of significance levels and the Fast Double Bootstrap yields excellent results; see Davidson and MacKinnon (2002a, 2002b).

The second main use of the bootstrap that Jeong and Maddala describe has also been illustrated. No standard asymptotic distribution is available when a null model is tested separately against each of two or more non-nested alternatives and an induced test based upon the minimum p -value of the individual test statistics is used. The evidence in Section 7.3.4 indicates that the application of a single bootstrap gives good control of finite sample significance levels. The results of the simulation experiments described in Section 7.3.5 suggest that the bootstrap test based upon the minimum p -value can be more powerful than a joint J test for which asymptotic theory is tractable.

It is, of course, important that the bootstrap be applied in a correct way. Some researchers have tried to simplify the testing of non-nested models by simply bootstrapping the log-likelihood ratio statistic, or some other measure of relative goodness of fit. It cannot be guaranteed that such an approach will yield asymptotically valid inferences. If the probability limit of the average log-likelihood ratio statistic, that is, $n^{-1}\hat{L}_{01}$ in the notation of Section 7.2, depends upon the unknown parameters of the null model, the implied bootstrap asymptotic distribution will be inappropriate, having the wrong variance. The discussion of this point in Section 7.4 is reminiscent of the findings reported in Durbin (1970)

concerning naive tests in which estimates are treated as if they were the true parameters.

There are many topics that deserve further investigation. A more detailed study of the behaviour of bootstrap tests when there are multiple non-nested alternatives would provide useful information. In such a study of overall assessment based upon the minimum p -value associated with a collection of separate binary tests, the single bootstrap used in Godfrey (1998) might be compared with the two-level bootstrap approach proposed in Godfrey (2005).

A second important topic for future research is the relaxation of the assumption of IID errors that has been made in this chapter. In particular, it has been suggested several times in previous chapters that heteroskedasticity-robust tests should be used whenever possible. Tests for non-nested hypotheses that are calculated by applying OLS to artificial regression models could easily be modified to be based upon heteroskedasticity-consistent covariance matrix estimators, combined with a wild bootstrap of the type described in Section 5.2.3. Tests that are amenable to such modification include the procedures F_{01} , J_{01} and JA_{01} , derived for a single alternative, which are described in Section 7.3.1, and also the joint tests for multiple alternatives F_0^C and F_0^J , which are considered in Section 7.3.4. Given the results in Chapter 6, the wild bootstrap would appear to offer the promise of reliable tests for non-nested regression equations in the presence of heteroskedasticity of unknown form.

8

Epilogue

I hope that the previous chapters have made the following clear: first, the approximations to the finite sample distributions of test statistics that are provided by conventional (first-order) asymptotic theory are sometimes inadequate for practical purposes; second, the use of an appropriate bootstrap method can lead to improved control of finite sample significance levels, relative to critical values from asymptotic theory, with some improvements being very substantial; and third, there are important problems in testing for which a bootstrapping approach can provide a tractable solution but asymptotic theory cannot.

In order to simplify the exposition, I have focussed on tests that are applied after the OLS estimation of a linear regression model in which all regressors are at least predetermined and have the property that sample means and sample variances calculated from regressor values tend to population values, which are finite, as the sample size increases. However, the bootstrap can be applied in the context of more general models than the linear regression equation, with more complicated estimators and with nonstationary data processes.

As well as being studied extensively in the setting of linear regression, the bootstrap approach has been used with, for example, nonlinear regressions, systems of simultaneous equations, panel-data regressions, logit, probit and Tobit models. In terms of extending the coverage of bootstrap procedures beyond OLS results, researchers have devised bootstrap methods for application after estimation by generalized least squares, quantile regression methods, instrumental variables and Generalized Method of Moments techniques. Nonstationarity of the variables in the regression model can be permitted and there is a substantial and expanding literature on bootstrap tests for investigating the presence of unit roots and cointegration.

Readers wishing to learn more about bootstrap methods that are outside the scope of this book can consult the several excellent surveys that have been published; for example, see Berkowitz and Kilian (2000), Davidson and MacKinnon (2006), Horowitz (2001, 2003) and Li and Maddala (1996). However, looking at these surveys and the contributions to which they contain references will only be a first step. The study of bootstrap methods in econometrics attracts many energetic and talented researchers; so that web-based searches are strongly advised in order to learn about what results are available for the problem of interest.

At the time of writing, it is almost 30 years since Efron's article on the bootstrap appeared; see Efron (1979). It is difficult to describe just how important Efron's work has proved to be in applied and theoretical research. It offers an alternative mind-set for statistical inference. Instead of using formal analysis and theorems about asymptotic behaviour, it is possible to substitute capital for labour by employing a personal computer with a suitable program to generate artificial samples after specifying a bootstrap world, given the actual data. The statistics calculated from the bootstrap samples then provide a convenient reference set for judging the statistical significance of the real-world value of the test statistic. However, the importance of the bootstrap approach to those wishing to learn and use econometrics is not limited to hypothesis testing.

As I confessed in the Preface, my research obsession is working on tests for econometricians, but I am also very interested in teaching econometrics and statistics, especially at the introductory level. My experience from teaching is that students sometimes find understanding concepts related to sampling distributions difficult. As persuasively argued in Kennedy (2001), the bootstrap may prove to be a very valuable teaching tool and help with motivation and understanding even in introductory modules. I hope that the authors of first-level econometric textbooks will adopt the ideas in Kennedy (2001) and that standard econometric estimation packages will be written to include code that allows the implementation of the bootstrap techniques that have been described in this book.

Bibliography

- Ali, M. M. and Sharma, S. C. (1996). Robustness to nonnormality of regression F-tests. *Journal of Econometrics*, **71**, 175–205.
- Amemiya, T. (1973). Generalized least squares with an estimated autocovariance matrix. *Econometrica*, **41**, 723–732.
- Amemiya, T. (1977). A note on a heteroscedastic model. *Journal of Econometrics*, **6**, 365–370.
- Andrews, D. W. K. (1991). Heteroskedasticity and autocorrelation consistent covariance matrix estimation. *Econometrica*, **59**, 817–858.
- Andrews, D. W. K. (1993). Tests for parameter instability and structural change with unknown change point. *Econometrica*, **61**, 821–856.
- Andrews, D. W. K. (2003a). Tests for parameter instability and structural change with unknown change point: a corrigendum. *Econometrica*, **71**, 395–397.
- Andrews, D. W. K. (2003b). End-of-sample instability tests. *Econometrica*, **71**, 1661–1694.
- Andrews, D. W. K. and Buchinsky, M. (2002). On the number of bootstrap repetitions for BC_a confidence intervals. *Econometric Theory*, **18**, 962–984.
- Andrews, D. W. K. and Monahan, J. C. (1992). An improved heteroskedasticity and autocorrelation consistent covariance matrix estimator. *Econometrica*, **60**, 953–966.
- Anselin, L. (2006). Spatial econometrics. In Mills, T. C. and Patterson, K. (eds.), *Palgrave Handbook of Econometrics, Volume 1*, pp. 901–969. New York: Palgrave Macmillan.
- Arnold, S. F. (1980). Asymptotic validity of F tests for the ordinary linear model and the multiple correlation model. *Journal of the American Statistical Society*, **75**, 890–894.
- Athreya, K. (1987). Bootstrap of the mean in the infinite variance case. *Annals of Statistics*, **15**, 724–731.
- Baltagi, B. H. and Li, Q. (1995). Testing AR(1) against MA(1) disturbances in an error component model. *Journal of Econometrics*, **68**, 133–151.
- Barnard, G. A. (1963). Discussion on The spectral analysis of point processes (by M. S. Bartlett). *Journal of the Royal Statistical Society, Series B*, **25**, 294.
- Belsley, D. A. (2002). An investigation of an unbiased correction for heteroskedasticity and the effects of misspecifying the skedastic function. *Journal of Economic Dynamics and Control*, **26**, 1379–1396.
- Belsley, D. A., Kuh, E. and Welsch, R. E. (1980). *Regression Diagnostics: Identifying Influential Data and Sources of Collinearity*. New York: Wiley.
- Benjamini, Y. and Yekutieli, D. (2001). The control of the false discovery rate in multiple testing under uncertainty. *Annals of Statistics*, **29**, 1165–1188.
- Beran, R. (1988). Prepivoting test statistics: a bootstrap view of asymptotic refinements. *Journal of the American Statistical Association*, **83**, 687–697.
- Berkowitz, J. and Kilian, L. (2000). Recent developments in bootstrapping time series. *Econometric Reviews*, **19**, 1–48.

- Bisaglia, L. and Procidano, I. (2002). On the power of the augmented Dickey-Fuller test against fractional alternatives using bootstrap. *Economics Letters*, **77**, 343–347.
- Bollerslev, T. (1986). Generalized autoregressive conditional heteroskedasticity. *Journal of Econometrics*, **31**, 307–327.
- Booth, J. G. and Hall, P. (1994). Monte Carlo approximation and the iterated bootstrap. *Biometrika*, **81**, 331–340.
- Bowman, K. O. and Shenton, L. R. (1975). Omnibus contours for departures from normality based upon $\sqrt{b_1}$ and b_2 . *Biometrika*, **62**, 243–250.
- Bradley, J. V. (1978). Robustness? *British Journal of Mathematical and Statistical Psychology*, **31**, 144–152.
- Breusch, T. S. (1978). Testing for autocorrelation in dynamic linear models. *Australian Economic Papers*, **17**, 334–355.
- Breusch, T. S. (1980). Useful invariance results for generalized regression models. *Journal of Econometrics*, **13**, 327–340.
- Breusch, T. S. and Pagan, A. R. (1979). A simple test for heteroscedasticity and random coefficient variation. *Econometrica*, **47**, 1287–1294.
- Broman, K. W. and Caffo, B. S. (2003). Simulation-based P values: response to North et al. *American Journal of Human Genetics*, **72**, 496.
- Bühlmann, P. (2002). Bootstraps for time series. *Statistical Science*, **17**, 57–72.
- Burridge, P. and Taylor, A. M. R. (2001). On regression-based tests for seasonal unit roots in the presence of periodic heteroscedasticity. *Journal of Econometrics*, **104**, 91–118.
- Carlstein, E. (1986). The use of subseries values for estimating the variance of a general statistic from a stationary sequence. *Annals of Statistics*, **14**, 1171–1179.
- Chang, Y. (2004). Bootstrap unit root tests in panels with cross-sectional dependency. *Journal of Econometrics*, **120**, 263–293.
- Chang, Y. and Park, J. Y. (2002). On the asymptotics of ADF tests for unit roots. *Econometric Reviews*, **21**, 431–447.
- Chang, Y., Park, J. Y. and Song, K. (2006). Bootstrapping cointegrated regressions. *Journal of Econometrics*, **133**, 703–739.
- Chesher, A. and Jewitt, I. (1987). The bias of a heteroskedasticity consistent covariance matrix estimator. *Econometrica*, **55**, 1217–1222.
- Choi, E. and Hall, P. (2000). Bootstrap confidence regions computed from autoregressions of arbitrary order. *Biometrika*, **62**, 461–477.
- Chow, G. C. (1960). Tests of equality between sets of coefficients in two linear regressions. *Econometrica*, **28**, 591–605.
- Christiano, L. J. (1992). Searching for a break in GNP. *Journal of Business and Economic Statistics*, **10**, 237–250.
- Christofferson, J. (1997). A resampling method for regression models with serially correlated errors. *Computational Statistics and Data Analysis*, **25**, 43–53.
- Clarke, K. A. (2003). Nonparametric model discrimination in international relations. *Journal of Conflict Resolution*, **47**, 72–93.
- Clarke, K. A. (2007). A simple distribution-free test for nonnested model selection. *Political Analysis*, **15**, 347–363.
- Coulibaly, N. and Brorsen, B. W. (1999). Monte Carlo sampling approach to testing nonnested hypotheses: Monte Carlo results. *Econometric Reviews*, **18**, 195–209.

- Cox, D. R. (1961). Tests of separate families of hypotheses. *Proceedings of the Fourth Berkeley Symposium on Mathematical Statistics and Probability, Volume 1*, pp. 105–123. Berkeley: University of California Press.
- Cox, D. R. (1962). Further results on tests of separate families of hypotheses. *Journal of the Royal Statistical Society, Series B*, **24**, 406–424.
- Cox, D. R. and Hinkley, D. V. (1974). *Theoretical Statistics*. London: Chapman and Hall.
- Cribari-Neto, F. (2004). Asymptotic inference under heteroskedasticity of unknown form. *Computational Statistics and Data Analysis*, **45**, 215–233.
- Cribari-Neto, F. and Zarkos, S. G. (1999). Bootstrap methods for heteroskedastic regression models: evidence on estimation and testing. *Econometric Reviews*, **18**, 211–228.
- D'Agostino, R. B. (1986). Tests for the normal distribution. In D'Agostino, R. B. and Stephens, M. A. (eds.), *Goodness of Fit Techniques*, pp. 367–419. New York: Marcel Dekker.
- Darroch, J. N. and Silvey, S. D. (1963). On testing more than one hypothesis. *Annals of Mathematical Statistics*, **34**, 555–567.
- Davidson, J. (1994). *Stochastic Limit Theory*. Oxford: Oxford University Press.
- Davidson, J. (2000). *Econometric Theory*. Oxford: Blackwell.
- Davidson, J., Monticini, A. and Peel, D. (2007). Implementing the wild bootstrap using a two-point distribution. *Economics Letters*, **96**, 309–315.
- Davidson, R. (2000). Comment on “Recent developments in bootstrapping time series”. *Econometric Reviews*, **19**, 49–54.
- Davidson, R. (2007). Bootstrapping econometric models. Working Paper, GREQAM.
- Davidson, R. and Flachaire, E. (2001). The wild bootstrap, tamed at last. Queen's Economics Department Working Paper No. 1000, Queen's University, Kingston, Canada.
- Davidson, R. and Flachaire, E. (2008). The wild bootstrap, tamed at last. *Journal of Econometrics*, **146**, 162–169.
- Davidson, R. and MacKinnon, J. G. (1981). Several tests for model specification in the presence of alternative hypotheses. *Econometrica*, **49**, 781–793.
- Davidson, R. and MacKinnon, J. G. (1982). Some non-nested hypothesis tests and the relations among them. *Review of Economic Studies*, **49**, 551–565.
- Davidson, R. and MacKinnon, J. G. (1985a). Heteroskedasticity-robust tests in regression directions. *Annales de l'INSEE*, **59/60**, 183–218.
- Davidson, R. and MacKinnon, J. G. (1985b). The interpretation of test statistics. *Canadian Journal of Economics*, **18**, 38–57.
- Davidson, R. and MacKinnon, J. G. (1992). A new form of the information matrix test. *Econometrica*, **60**, 145–157.
- Davidson, R. and MacKinnon, J. G. (1993). *Estimation and Inference in Econometrics*. Oxford: Oxford University Press.
- Davidson, R. and MacKinnon, J. G. (1999). The size distortion of bootstrap tests. *Econometric Theory*, **15**, 361–376.
- Davidson, R. and MacKinnon, J. G. (2000). Bootstrap tests: how many bootstraps? *Econometric Reviews*, **19**, 55–68.
- Davidson, R. and MacKinnon, J. G. (2002a). Bootstrap J tests of nonnested linear regression models. *Journal of Econometrics*, **109**, 167–193.

- Davidson, R. and MacKinnon, J. G. (2002b). Fast double bootstrap tests of nonnested linear regression models. *Econometric Reviews*, **21**, 419–429.
- Davidson, R. and MacKinnon, J. G. (2004). *Econometric Theory and Methods*. Oxford: Oxford University Press.
- Davidson, R. and MacKinnon, J. G. (2006). Bootstrap methods in econometrics. In Mills, T. C. and Patterson, K. (eds.), *Palgrave Handbook of Econometrics, Volume 1*. Basingstoke: Palgrave Macmillan.
- Davidson, R. and MacKinnon, J. G. (2007). Improving the reliability of bootstrap tests with the fast double bootstrap. *Computational Statistics and Data Analysis*, **51**, 3259–3281.
- Davies, R. B. (1977). Hypothesis testing when a nuisance parameter is present only under the alternative. *Biometrika*, **64**, 247–254.
- Davidson, A. C. and Hall, P. (1993). On Studentizing and blocking methods for implementing the bootstrap with dependent data. *Australian Journal of Statistics*, **35**, 215–224.
- Davidson, A. C. and Hinkley, D. V. (1997). *Bootstrap Methods and their Application*. Cambridge: Cambridge University Press.
- De Jong, R. M. and Davidson, J. (2000). Consistency of kernel estimators of heteroscedastic and autocorrelated covariance matrices. *Econometrica*, **68**, 407–423.
- Deb, P. and Sefton, M. (1996). The distribution of a Lagrange multiplier test of normality. *Economics Letters*, **51**, 123–130.
- Delgado, M. A. and Stengos, T. (1994). Semiparametric specification testing of non-nested econometric models. *Review of Economic Studies*, **61**, 291–303.
- Den Hann, W. J. and Levin, A. (1997). A practitioner's guide to robust covariance estimation. In Maddala, G. S. and Rao, C. R. (eds.), *Handbook of Statistics: Robust Inference*. New York: Elsevier.
- Dezhbakhsh, H. (1990). The inappropriate use of serial correlation tests in dynamic linear models. *Review of Economics and Statistics*, **72**, 126–132.
- Dezhbakhsh, H. and Thursby, J. G. (1995). A Monte Carlo comparison of tests based upon the Durbin-Watson statistic with other autocorrelation tests in dynamic models. *Econometric Reviews*, **14**, 347–366.
- Dhaene, G. and Hoorelbeke, D. (2004). The information matrix test with bootstrap-based covariance matrix estimation. *Economics Letters*, **82**, 341–347.
- Diebold, F. X. and Chen, C. (1996). Testing structural stability with endogenous breakpoint: a size comparison of analytic and bootstrap procedures. *Journal of Econometrics*, **70**, 221–241.
- Dufour, J.-M., and Khalaf, L. (2001). Monte Carlo test methods in econometrics. In Baltagi, B. (ed.), *A Companion to Econometric Theory*, pp. 494–519. Oxford: Blackwell.
- Dufour, J.-M., Farhat, A., Gardiol, L., and Khalaf, L. (1998). Simulation-based finite sample normality tests in linear regressions. *Econometrics Journal*, **1**, 154–173.
- Dufour, J.-M., Khalaf, L., Bernard, J.-T. and Genest, I. (2004). Simulation-based finite-sample tests for heteroskedasticity and ARCH effects. *Journal of Econometrics*, **122**, 317–347.
- Durbin, J. (1970). Testing for serial correlation in least squares regression when some of the regressors are lagged dependent variables. *Econometrica*, **38**, 410–421.

- Durbin, J. and Watson, G. S. (1950). Testing for serial correlation in least squares regression I. *Biometrika*, **37**, 409–428.
- Durbin, J. and Watson, G. S. (1951). Testing for serial correlation in least squares regression II. *Biometrika*, **38**, 159–178.
- Dwass, M. (1957). Modified randomization tests for nonparametric hypotheses. *Annals of Mathematical Statistics*, **28**, 181–187.
- Eastwood, A. and Godfrey, L. G. (1992). The properties and constructive use of misspecification tests for multiple regression models. In Godfrey, L. G. (ed.), *The Implementation and Constructive Use of Misspecification Tests in Econometrics*, pp. 109–175. Manchester: Manchester University Press.
- Efron, B. (1979). Bootstrap methods: another look at the jackknife. *Annals of Statistics*, **7**, 1–26.
- Efron, B. and Gong, G. (1983). A leisurely look at the bootstrap, the jackknife and cross-validation. *American Statistician*, **37**, 36–48.
- Efron, B. and Tibshirani, R. J. (1993). *An Introduction to the Bootstrap*. London: Chapman and Hall.
- Eicker, F. (1967). Limit theorems for regression with unequal and dependent errors. In LeCam, L. and Neyman, J. (eds.), *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability*, pp.59–82. Berkeley: University of California Press.
- Engle, R. F. (1982). Autoregressive conditional heteroscedasticity with estimates of the variance of U.K. inflation. *Econometrica*, **50**, 987–1008.
- Ericsson, N. R. (1986). Post-simulation analysis of Monte Carlo experiments: interpreting Pesaran's (1974) study of non-nested test statistics. *Review of Economic Studies*, **53**, 691–707.
- Fan, Y. and Li, Q. (1995). Bootstrapping J-type tests for non-nested regression models. *Economics Letters*, **48**, 107–112.
- Fisher, G. R. and McAleer, M. (1981). Alternative procedures and associated tests of significance for non-nested hypotheses. *Journal of Econometrics*, **16**, 103–119.
- Fitzenberger, B. (1998). The moving blocks bootstrap and robust inference for linear least squares and quantile regressions. *Journal of Econometrics*, **82**, 235–287.
- Flachaire, E. (1999). A better way to bootstrap pairs. *Economics Letters*, **64**, 257–262.
- Flachaire, E. (2005). Bootstrapping heteroskedastic regression models: wild bootstrap vs. pairs bootstrap. *Computational Statistics and Data Analysis*, **49**, 361–376.
- Franke, J., Kreiss, J.-P. and Mammen, E. (2002). Bootstrap of kernel smoothing in nonlinear time series. *Bernoulli*, **8**, 1–37.
- Freedman, D. A. (1981). Bootstrapping regression models. *Annals of Statistics*, **9**, 1218–1228.
- Freedman, D. A. (1984). On bootstrapping two-stage least-squares estimates in stationary linear models. *Annals of Statistics*, **12**, 827–842.
- Fuertes, A.-M. (2008). Sieve bootstrap *t*-tests on long-run average parameters. *Computational Statistics and Data Analysis*, **52**, 3354–3370.
- van Giersbergen, N. P. A. and Kiviet, J. F. (2002). How to implement the bootstrap in static or stable dynamic regression models: test statistics versus confidence region approach. *Journal of Econometrics*, **108**, 133–156.
- Glejser, H. (1969). A new test of heteroskedasticity. *Journal of the American Statistical Association*, **64**, 316–323.

- Godfrey, L. G. (1978). Testing against general autoregressive and moving average error models when the regressors include lagged dependent variables. *Econometrica*, **46**, 1293–1301.
- Godfrey, L. G. (1981). On the invariance of the Lagrange multiplier test with respect to certain changes in the alternative hypothesis. *Econometrica*, **49**, 1443–1455.
- Godfrey, L. G. (1984). On the uses of misspecification checks and tests of non-nested hypotheses in empirical econometrics. *Economic Journal*, **94**, 69–81.
- Godfrey, L. G. (1987). Discriminating between autocorrelation and misspecification in regression analysis: an alternative strategy. *Review of Economics and Statistics*, **69**, 128–134.
- Godfrey, L. G. (1988). *Misspecification Tests in Econometrics*. Cambridge: Cambridge University Press.
- Godfrey, L. G. (1996). Some results on the Glejser and Koenker tests for heteroskedasticity. *Journal of Econometrics*, **72**, 275–299.
- Godfrey, L. G. (1998). Tests for non-nested regression models: some results on small sample behaviour and the bootstrap. *Journal of Econometrics*, **84**, 59–74.
- Godfrey, L. G. (2005). Controlling the overall significance level of a battery of least squares diagnostic tests. *Oxford Bulletin of Economics and Statistics*, **67**, 263–279.
- Godfrey, L. G. (2006). Tests for regression models with heteroskedasticity of unknown form. *Computational Statistics and Data Analysis*, **50**, 2715–2733.
- Godfrey, L. G. (2007a). On the asymptotic validity of a bootstrap method for testing nonnested hypotheses, *Economics Letters*, **94**, 408–413.
- Godfrey, L. G. (2007b). Alternative approaches to implementing Lagrange multiplier tests for serial correlation in dynamic regression models. *Computational Statistics and Data Analysis*, **51**, 3282–3295.
- Godfrey, L. G. (2008). Testing for heteroskedasticity and predictive failure in linear regression models. *Oxford Bulletin of Economics and Statistics*, **70**, 415–429.
- Godfrey, L. G. and Orme, C. D. (1994). The sensitivity of some general checks to omitted variables in linear models. *International Economic Review*, **35**, 489–506.
- Godfrey, L. G. and Orme, C. D. (1996). On the behaviour of conditional moment tests in the presence of unconsidered local alternatives. *International Economic Review*, **37**, 263–281.
- Godfrey, L. G. and Orme, C. D. (1999). The robustness, reliability and power of heteroskedasticity tests. *Econometric Reviews*, **18**, 169–194.
- Godfrey, L. G. and Orme, C. D. (2000). Controlling the significance levels of prediction error tests for linear regression models. *Econometrics Journal*, **3**, 66–83.
- Godfrey, L. G. and Orme, C. D. (2002a). Using bootstrap methods to obtain nonnormality robust Chow prediction tests. *Economics Letters*, **76**, 429–436.
- Godfrey, L. G. and Orme, C. D. (2002b). Significance levels of heteroskedasticity-robust tests for specification and misspecification: some results on the use of wild bootstraps. Unpublished paper, University of York.
- Godfrey, L. G. and Orme, C. D. (2004). Controlling the finite sample significance levels of heteroskedasticity-robust tests of several linear restrictions on regression coefficients. *Economics Letters*, **82**, 281–287.
- Godfrey, L. G. and Pesaran, M. H. (1983). Tests of non-nested regression models: small sample adjustments and Monte Carlo evidence. *Journal of Econometrics*, **21**, 133–154.

- Godfrey, L. G. and Santos Silva, J. M. C. (2004). Bootstrap tests of nonnested hypotheses: some further results, *Econometric Reviews*, **23**, 325–340.
- Godfrey, L. G. and Tremayne, A. R. (1988). Misspecification tests for univariate time series models and their applications in econometrics. *Econometric Reviews*, **7**, 1–42.
- Godfrey, L. G. and Tremayne, A. R. (2005). The wild bootstrap and heteroskedasticity robust tests for serial correlation in dynamic regression models. *Computational Statistics and Data Analysis*, **49**, 377–395.
- Godfrey, L. G. and Veall, M. R. (2000). Alternative approaches to testing by variable addition. *Econometric Reviews*, **19**, 241–261.
- Godfrey, L. G., Orme, C. D. and Santos Silva, J. M. C. (2006). Simulation-based tests for heteroskedasticity in linear regression models: some further results. *Econometrics Journal*, **9**, 76–97.
- Gonçalves, S. and Kilian, L. (2004). Bootstrapping autoregressions with conditional heteroskedasticity of unknown form. *Journal of Econometrics*, **123**, 89–120.
- Gonçalves, S. and Vogelsang, T. J. (2006). Block bootstrap HAC robust tests: the sophistication of the naive bootstrap. Working Paper, Department of Economics, Cornell University.
- Gonçalves, S. and White, H. (2005). Bootstrap standard error estimates for linear regression. *Journal of the American Statistical Association*, **100**, 970–979.
- Götze, F. and Künsch, H. R. (1996). Second-order correctness of the blockwise bootstrap for stationary observations. *Annals of Statistics*, **24**, 1914–1933.
- Gourieroux, C. and Monfort, A. (1990). *Time Series and Dynamic Models*. Cambridge: Cambridge University Press.
- Gourieroux, C. and Monfort, A. (1994). Testing nonnested hypotheses. In Engle, R. F. and McFadden, D. L. (eds.), *Handbook of Econometrics, Volume 4*. Amsterdam: North-Holland.
- Greene, W. H. (2008). *Econometric Analysis*, 6th Edition. New Jersey: Prentice-Hall.
- Gujarati, D. N. (2003). *Basic Econometrics*, 4th Edition. New York: McGraw-Hill.
- Haldrup, N. and Jansson, M. (2006). Improving size and power in unit root testing. In Mills, T. C. and Patterson, K. (eds.), *Palgrave Handbook of Econometrics, Volume 1*. Basingstoke: Palgrave Macmillan.
- Hall, A. (1987). The information matrix test for the linear model. *Review of Economic Studies*, **54**, 257–263.
- Hall, P. (1992). *The Bootstrap and Edgeworth Expansion*. New York: Springer-Verlag.
- Hall, P. and Titterton, D. M. (1989). The effect of simulation order on level accuracy and power of Monte Carlo tests. *Journal of the Royal Statistical Society B*, **51**, 459–467.
- Hall, P. and Wilson, S. R. (1991). Two guidelines for bootstrap hypothesis testing. *Biometrics*, **47**, 757–762.
- Hall, P. and Wilson, S. R. (1992). Bootstrap hypothesis testing. *Biometrics*, **48**, 970.
- Hall, P., Horowitz, J. L. and Jing, B.-Y. (1995). On blocking rules for the bootstrap for dependent data. *Biometrika*, **82**, 561–574.
- Hamilton, J. D. (1994). *Time Series Analysis*. Princeton: Princeton University Press.
- Hansen, B. E. (1997). Approximate asymptotic p values for structural-change tests. *Journal of Business and Economic Statistics*, **15**, 60–67.
- Hansen, B. E. (1999). Discussion of “Data mining reconsidered”. *Econometrics Journal*, **2**, 192–201.

- Hansen, B. E. (2000). Testing for structural change in conditional models. *Journal of Econometrics*, **97**, 93–115.
- Hansen, B. E. (2001). The new econometrics of structural change: dating breaks in U.S. labor productivity. *Journal of Economic Perspectives*, **15**, 117–128.
- Härdle, W., Horowitz, J. and Kreiss, J.-P. (2003). Bootstrap methods for time series. *International Statistical Review*, **71**, 435–459.
- Harvey, A. C. (1976). Estimating regression models with multiplicative heteroskedasticity. *Econometrica*, **44**, 461–465.
- Hausman, J. (1978). Specification tests in econometrics. *Econometrica*, **46**, 1251–1271.
- Hedges, S. B. (1992). The number of replications needed for accurate estimation of the bootstrap p value in phylogenetic studies. *Molecular Biology and Evolution*, **9**, 366–369.
- Hendry, D. F. (1980). Predictive failure and econometric modelling in macro-econometrics: the transaction demand for money. In Omerod, P. (ed.), *Modelling the UK Economy*. Heinemann: London.
- Hendry, D. F. and Santos, C. (2005). Regression models with data-based indicator variables. *Oxford Bulletin of Economics and Statistics*, **67**, 571–595.
- Hidalgo, J. (2003). An alternative bootstrap to moving blocks for time series regression models. *Journal of Econometrics*, **117**, 369–399.
- Hill, R. C., Griffiths, W. E. and Lim, G. C. (2008). *Principles of Econometrics*, 3rd Edition. New York: Wiley.
- Hinkley, D. V. (1988). Bootstrap methods. *Journal of the Royal Statistical Society, Series B*, **50**, 321–337.
- Horowitz, J. L. (1994). Bootstrap-based critical values for the information-matrix test. *Journal of Econometrics*, **61**, 395–411.
- Horowitz, J. L. (1997). Bootstrap methods in econometrics. In Kreps, D. M. and Wallis, K. F. (eds.), *Advances in Economics and Econometrics: Theory and Applications, Seventh World Congress, Volume 3*. Cambridge: Cambridge University Press.
- Horowitz, J. L. (2001). The bootstrap. In Heckman, J. J. and Leamer, E. E. (eds.), *Handbook of Econometrics, Volume 5*, pp. 3159–3228. Amsterdam: North Holland.
- Horowitz, J. L. (2003). The bootstrap in econometrics. *Statistical Science*, **18**, 211–218.
- Horowitz, J. L. and Savin, N. E. (2000). Empirically relevant critical values for hypothesis tests: a bootstrap approach. *Journal of Econometrics*, **95**, 375–389.
- Horowitz, J. L., Lobato, I. N., Nankervis, J. C. and Savin, N. E. (2006). Bootstrapping the Box-Pierce Q test: a robust test of uncorrelatedness. *Journal of Econometrics*, **133**, 841–862.
- Hsieh, D. A. (1983). A heteroscedasticity-consistent covariance matrix estimator for time series regressions. *Journal of Econometrics*, **22**, 281–290.
- Hu, F. and Kalbfleisch, J. D. (2000). The estimating function bootstrap. *Canadian Journal of Statistics*, **28**, 449–499.
- Hu, F. and Zidek, J. V. (1995). A bootstrap based on the estimating equations of the linear model. *Biometrika*, **82**, 263–275.
- Jarque, C. M. and Bera, A. K. (1980). Efficient tests for normality, homoscedasticity and serial independence of regression residuals. *Economics Letters*, **6**, 255–259.
- Jarque, C. M. and Bera, A. K. (1987). A test for normality of observations and regression residuals. *International Statistical Review*, **55**, 163–172.

- Jeong, J. and Lee, K. (1999). Bootstrapped White's test for heteroskedasticity in regression models. *Economics Letters*, **63**, 261–267.
- Jeong, J. and Maddala, G. S. (1993). A perspective on application of bootstrap methods in econometrics. In Maddala G. S., Rao, C. R. and Vinod, H. D. (eds.), *Handbook of Statistics, Volume 11*, pp. 573–610. Amsterdam: Elsevier.
- Jöckel, K.-H. (1986). Finite sample properties and asymptotic efficiency of Monte Carlo tests. *Annals of Statistics*, **14**, 336–347.
- Jouini, J. (2008). Bootstrap methods for single structural change tests: power versus corrected size and empirical illustration. *Statistical Papers* [online version] 6 March, www.springerlink.com/content/x880v9496206454h/?p=bcb2c86a9ea64763956ad93b90a3179e&pi=0.
- Kennedy, P. E. (1995). Randomization tests in econometrics. *Journal of Business and Economic Statistics*, **13**, 85–94.
- Kennedy, P. E. (2001). Bootstrapping student understanding of what is going on in econometrics. *Journal of Economic Education*, **32**, 110–123.
- Kiefer, N. M. and Vogelsang, T. J. (2002). Heteroskedasticity-autocorrelation robust standard errors using the Bartlett kernel without truncation. *Econometrica*, **70**, 2093–2095.
- Kiefer, N. M. and Vogelsang, T. J. (2005). A new asymptotic theory for heteroskedasticity-autocorrelation robust tests. *Econometric Theory*, **21**, 1130–1164.
- Kiefer, N. M., Vogelsang, T. J. and Bunzel, H. (2000). Simple robust testing of regression hypotheses. *Econometrica*, **68**, 695–714.
- Kim, S., Shephard, N. and Chib, S. (1998). Stochastic volatility: likelihood inference and comparison with ARCH models. *Review of Economic Studies*, **65**, 361–393.
- Kiviet, J. F. (1986). On the rigour of some specification tests for modelling dynamic relationships. *Review of Economic Studies*, **53**, 241–262.
- Koenker, R. (1981). A note on Studentizing a test for heteroskedasticity. *Journal of Econometrics*, **17**, 107–112.
- Krämer, W. and Sonnberger, H. (1986). Computational pitfalls of the Hausman test. *Journal of Economic Dynamics and Control*, **10**, 163–165.
- Kreiss, J.-P. and Franke, J. (1992). Bootstrapping stationary autoregressive moving-average models. *Journal of Time Series Analysis*, **13**, 297–317.
- Künsch, H. R. (1989). The jackknife and the bootstrap for general stationary observations. *Annals of Statistics*, **17**, 1217–1241.
- Lai, T. L. and Wei, C. Z. (1982). Least squares estimation in stochastic regression models with applications to identification and control of dynamic systems. *Annals of Statistics*, **10**, 154–166.
- Lamarche, J.-F. (2003). A robust bootstrap test under heteroskedasticity. *Economics Letters*, **79**, 353–359.
- Lamarche, J.-F. (2004). The numerical performance of fast bootstrap procedures. *Computational Economics*, **23**, 379–389.
- Léger, C., Politis, D. N. and Romano, J. P. (1992). Bootstrap technology and applications. *Technometrics*, **34**, 378–398.
- Li, H. and Maddala, G. S. (1996). Bootstrapping time series models. *Econometric Reviews*, **15**, 115–158.
- Li, J. (2006). The block bootstrap test of Hausman's exogeneity in the presence of serial correlation. *Economics Letters*, **91**, 76–82.

- Lien, D. and Vuong, Q. H. (1987). Selecting the best linear regression model: a classical approach. *Journal of Econometrics*, **35**, 3–23.
- Ligeralde, A. V. and Brown, B. W. (1995). Band covariance matrix estimation using restricted residuals: a Monte Carlo analysis. *International Economic Review*, **36**, 751–767.
- Liu, R. Y. (1988). Bootstrap procedures under some non i.i.d. models. *Annals of Statistics*, **16**, 1696–1708.
- Liu, R. Y. and Singh, K. (1992). Efficiency and robustness in resampling. *Annals of Statistics*, **20**, 370–384.
- Long, J. S. and Ervin, L. H. (2000). Using heteroskedasticity-consistent standard errors in the linear regression model. *American Statistician*, **32**, 217–224.
- Luger, R. (2006). Exact permutation tests for non-nested non-linear regression models. *Journal of Econometrics*, **133**, 513–529.
- McAleer, M. (1995). The significance of testing empirical non-nested models. *Journal of Econometrics*, **67**, 149–171.
- McAleer, M. and Pesaran, M. H. (1986). Statistical inference in nonnested econometric models. *Applied Mathematics and Computation*, **20**, 271–311.
- McCabe, B. and Tremayne, A. R. (1993). *Elements of Modern Asymptotic Theory with Statistical Applications*. Manchester: Manchester University Press.
- McCullough, B. D. (1998). Algorithm choice for (partial) autocorrelation functions. *Journal of Economic and Social Measurement*, **24**, 265–278.
- McCullough, B. D. and Vinod, H. (1993). Implementing the single bootstrap: some computational considerations. *Computational Economics*, **6**, 1–15.
- McCullough, B. D. and Vinod, H. (1998). Implementing the double bootstrap. *Computational Economics*, **12**, 79–95.
- MacKinnon, J. G. (1983). Model specification tests against non-nested alternatives. *Econometric Reviews*, **2**, 85–110.
- MacKinnon, J. G. (1989). Heteroskedasticity-robust tests for structural change. *Empirical Economics*, **14**, 77–92.
- MacKinnon, J. G. (2002). Bootstrap inference in econometrics. *Canadian Journal of Economics*, **35**, 615–645.
- MacKinnon, J. G. (2007). Bootstrap hypothesis testing. Working Paper No. 1127, Queen's Economics Department, Kingston, Canada.
- MacKinnon, J. G. and White, H. (1985). Some heteroscedasticity consistent covariance matrix estimators with improved finite sample properties. *Journal of Econometrics*, **29**, 305–325.
- MacKinnon, J. G., White, H. and Davidson, R. (1983). Test for model specification in the presence of alternative hypotheses. *Journal of Econometrics*, **21**, 53–70.
- McQuarrie, A. D. R. and Tsai, C.-L. (1998). *Regression and the Time Series Model Selection*. Singapore: World Scientific Publishing.
- Mammen, E. (1992). *When Does Bootstrap Work?* Berlin: Springer-Verlag.
- Mammen, E. (1993). Bootstrap and wild bootstrap for high dimensional linear models. *Annals of Statistics*, **21**, 255–285.
- Mammen, E. and Nandi, S. (2004). Bootstrap and resampling. In Gentle, J. E., Härdle, W. and Mori, Y. (eds.), *Handbook of Computational Statistics*, pp. 467–496. Berlin: Springer-Verlag.
- Mann, H. and Wald, H. (1943). On stochastic limit and order relationships. *Annals of Statistics*, **14**, 217–226.

- Mantalos, P. (2003). Bootstrapping the Breusch-Godfrey autocorrelation test for a single equation dynamic model: bootstrapping the restricted versus unrestricted model. *Monte Carlo Methods and Applications*, **9**, 257–269.
- Marriott, F. H. C. (1979). Barnard's Monte Carlo tests: how many simulations? *Applied Statistics*, **28**, 75–77.
- Michelis, L. (1999). The distributions of the J and Cox non-nested tests in regression models with weakly correlated regressors. *Journal of Econometrics*, **93**, 369–401.
- Milliken, G. A. and Graybill, F. A. (1970). Extensions of the general linear hypothesis model. *Journal of the American Statistical Association*, **65**, 797–807.
- Mizon, G. E. (1995). A simple message to autocorrelation correctors: Don't. *Journal of Econometrics*, **69**, 267–288.
- Monfardini, C. (2003). An illustration of Cox's non-nested testing procedure for logit and probit models. *Computational Statistics and Data Analysis*, **42**, 425–444.
- Navidi, W. (1989). Edgeworth expansions for bootstrapping regression models. *Annals of Statistics*, **17**, 1472–1478.
- Neumeyer, N., Dette, H. and Nagel, E. - R. (2004). Bootstrap tests for the error distribution in linear and nonparametric regression models. Working Paper, Ruhr-Universität Bochum.
- Newey, W. K. and West, K. D. (1987). A simple, positive semi-definite, heteroskedasticity and autocorrelation consistent covariance matrix. *Econometrica*, **55**, 703–708.
- O'Reilly, G. and Whelan, K. (2005). Testing parameter stability: a wild bootstrap approach. Unpublished paper, Central Bank and Financial Services Authority of Ireland.
- Paparoditis, E. and Politis, D. N. (2005). Bootstrap hypothesis testing in regression models. *Statistics and Probability Letters*, **74**, 356–365.
- Paulsen, J. and Tjøstheim, D. (1985). On the estimation of residual variance and order in autoregressive time series. *Journal of the Royal Statistical Society B*, **47**, 216–228.
- Perron, P. (2006). Dealing with structural breaks. In Mills, T. C. and Patterson, K. (eds.), *Palgrave Handbook of Econometrics, Volume 1*. Basingstoke: Palgrave Macmillan.
- Pesaran, M. H. (1974). On the general problem of model selection. *Review of Economic Studies*, **41**, 153–171.
- Pesaran, M. H. (1982). Comparison of local power of alternative tests of non-nested regression models. *Econometrica*, **50**, 1287–1305.
- Pesaran, M. H. and Dupleich Ulloa, M. R. (2008). Non-nested hypotheses. In Durlauf, S. N. and Blume, L. E. (eds.), *The New Palgrave Dictionary of Economics*, second edition, pp. 107–114. Basingstoke: Palgrave Macmillan.
- Pesaran, M. H. and Pesaran, B. (1993). A simulation approach to the problem of computing Cox's statistic for testing nonnested models. *Journal of Econometrics*, **57**, 377–392.
- Pesaran, M. H. and Pesaran, B. (1995). A non-nested test of level differences versus log-differenced stationary models. *Econometric Reviews*, **14**, 213–227.
- Pesaran, M. H. and Weeks, M. (2001). Non-nested hypothesis testing: an overview. In Baltagi, B. H. (ed.), *Companion to Theoretical Econometrics*. Oxford: Basil Blackwell.

- Phillips, A. W. (1956). Some notes on the estimation of time-forms of reactions in interdependent dynamic systems. *Economica*, **23**, 99–113.
- Pierce, D. A. (1971). Least squares estimation in the regression model with autoregressive-moving average errors. *Biometrika*, **58**, 299–312.
- Politis, D. N. (2003). The impact of bootstrap methods on time series analysis. *Statistical Science*, **18**, 219–230.
- Politis, D. N. and Romano, J. P. (1994). The stationary bootstrap. *Journal of the American Statistical Association*, **89**, 1303–1313.
- Politis, D. N. and Romano, J. P. (1996). Subsampling for econometric models - comments on "Bootstrapping Time Series Models". *Econometric Reviews*, **15**, 169–176.
- Politis, D. N., Romano, J. P. and Wolf, M. (1997). Subsampling for heteroskedastic time series. *Journal of Econometrics*, **81**, 281–317.
- Quandt, R. E. (1958). The estimation of the parameters of a linear regression system obeying two separate regimes. *Journal of the American Statistical Association*, **53**, 873–880.
- Quandt, R. E. (1960). Tests of the hypothesis that a regression system obeys two separate regimes. *Journal of the American Statistical Association*, **55**, 324–330.
- Ramsey, J. B. (1969). Tests for specification errors in classical linear least squares regression analysis. *Journal of the Royal Statistical Society B*, **31**, 350–371.
- Ramsey, J. B. (1983). Diagnostic tests as residuals analysis: perspective and comment. *Econometric Reviews*, **2**, 241–248.
- Rayner, R. K. (1991). Resampling methods for tests in regression models with autocorrelated errors. *Economics Letters*, **36**, 281–284.
- Rayner, R. K. (1993). Testing for serial correlation in regression models with lagged dependent variables. *Review of Economics and Statistics*, **75**, 716–721.
- Rowley, J. C. R. and Wilton, D. A. (1973). Quarterly models of wage determination: some new efficient estimates. *American Economic Review*, **63**, 380–389.
- Salkever, D. S. (1976). The use of dummy variables to compute predictions, prediction errors and confidence intervals. *Journal of Econometrics*, **4**, 393–397.
- Samworth, R. (2003). A note on methods of restoring consistency to the bootstrap. *Biometrika*, **90**, 985–990.
- Schmidt, P. (1976). *Econometrics*. New York: Dekker.
- Schwert, G. W. (1987). Effects of model specification on tests for unit roots in macroeconomic data. *Journal of Monetary Economics*, **20**, 73–103.
- Schwert, G. W. (1989). Tests for unit roots: a Monte Carlo investigation. *Journal of Business and Economic Statistics*, **7**, 147–159.
- Serlin, R. C. (2000). Testing for robustness in Monte Carlo studies. *Psychological Methods*, **5**, 230–240.
- Shibata, R. (1976). Selection of the order of an autoregressive model by Akaike's information criterion. *Biometrika*, **63**, 117–126.
- Spanos, A. (2006). Econometrics in retrospect and prospect. In Mills, T. C. and Patterson, K. (eds.), *Palgrave Handbook of Econometrics, Volume 1*, pp. 3–60. New York: Palgrave Macmillan.
- Staiger, D. and Stock, J. H. (1997). Instrumental variables regression with weak instruments. *Econometrica*, **65**, 557–586.
- Stewart, K. G. (1997). Exact testing in multivariate regression. *Econometric Reviews*, **16**, 321–352.

- Stock, J. H. and Watson, M. W. (2007). *Introduction to Econometrics*, 2nd edition. Boston: Pearson.
- Szroeter, J. (1999). Testing non-nested econometric models. *The Current State of Economic Science*, **1**, 223–253.
- Thursby, J. G. and Schmidt, P. (1977). Some properties of tests for specification error in a linear regression model. *Journal of the American Statistical Association*, **72**, 635–641.
- Tibshirani, R. (1992). Bootstrap hypothesis testing. *Biometrics*, **48**, 969–970.
- Tjøstheim, D. and Paulsen, J. (1983). Bias of some commonly-used time series estimates. *Biometrika*, **70**, 389–399.
- Tremayne, A. R. (2006). Stationary linear univariate time series models. In Mills, T. C. and Patterson, K. (eds.), *Palgrave Handbook of Econometrics, Volume 1*. Basingstoke: Palgrave Macmillan.
- Urzúa, C. M. (1996). On the correct use of omnibus tests for normality. *Economics Letters*, **53**, 247–251.
- Verbeek, M. (2004). *A Guide to Modern Econometrics*, 2nd edition. New York: Wiley.
- Vuong, Q. H. (1989). Likelihood ratio tests for model selection and non-nested hypotheses. *Econometrica*, **57**, 307–333.
- Walker, A. M. (1967). Some tests of separate families of hypotheses in time series analysis. *Biometrika*, **54**, 39–68.
- Walker, A. M. (1970). Corrections: some tests of separate families of hypotheses in time series analysis. *Biometrika*, **57**, 226.
- White, H. (1980). A heteroscedasticity-consistent covariance matrix estimator and a direct test for heteroscedasticity. *Econometrica*, **50**, 483–500.
- White, H. (1982a). Maximum likelihood estimation of misspecified models. *Econometrica*, **50**, 1–25.
- White, H. (1982b). Regularity conditions for Cox's test of nonnested hypotheses. *Journal of Econometrics*, **19**, 301–318.
- White, H. (1984). *Asymptotic Theory for Econometricians*. Orlando: Academic Press.
- White, H. and MacDonald, G. M. (1980). Some large-sample tests for nonnormality in the linear regression model. *Journal of the American Statistical Association*, **75**, 16–28.
- Wise, J. (1957). The estimation of the time-response functions in complete economic systems. *Economica*, **24**, 67–70.
- Wong, K.-F. (1996). Bootstrapping Hausman's exogeneity test. *Economics Letters*, **53**, 139–143.
- Wooldridge, J. M. (2006). *Introductory Econometrics*, 2nd edition. Mason: Thomson, South-Western.
- Wu, C. F. J. (1986). Jackknife, bootstrap and other resampling methods in regression analysis (with discussion). *Annals of Statistics*, **14**, 1261–1295.
- Yeo, I.-K. (2005). Variable selection and transformation in linear regression models. *Statistics and Probability Letters*, **72**, 219–226.
- Zietz, J. (2001). Heteroskedasticity and neglected parameter heterogeneity. *Oxford Bulletin of Economics and Statistics*, **63**, 263–273.

This page intentionally left blank

Author Index

- Ali, M. M., 14, 28
Amemiya, T., 182, 202–3
Andrews, D. W. K., 21–2, 37, 69, 135,
149, 162, 164–8, 170–2, 175, 208,
220, 241–2, 244–5
Anselin, L., 15
Arnold, S. F., 102
Athreya, K., 79
- Baltagi, B. H., 298
Barnard, G. A., 47
Belsley, D. A., 16, 28, 183
Benjamini, Y., 150
Bera, A. K., 31–3, 40, 84, 132, 152
Beran, R., 55, 70–2, 78–9, 82, 91, 93–4,
99, 110, 120, 134–5, 139, 141,
148, 152–3, 159, 171, 174,
210, 247
Berkowitz, J., 79, 198, 253, 304
Bisaglia, L., 214
Bollerslev, T., 237
Booth, J. G., 74
Bradley, J. V., 26
Breusch, T. S., 8–9, 33, 56–7, 83, 90,
119, 145, 152, 159, 233, 243
Broman, K. W., 53
Brorsen, B. W., 298
Brown, B. W., 209, 256
Buchinsky, M., 69
Bühlmann, P., 198, 202, 205
Bunzel, H., 21, 23
Burridge, P., 237
- Caffo, B. S., 53
Carlstein, E., 199
Chang, Y., 202, 264
Chen, C., 37, 135, 166–70, 172
Chesher, A., 224
Choi, E., 202, 204
Chow, G. C., 36–7, 134, 136, 138–9,
148, 152, 160–1, 163, 174–5, 242
- Christiano, L. J., 172
Christoffersson, J., 206
Clarke, K. A., 278–9
Coulibaly, N., 298
Cox, D. R., 48, 268, 272–3, 277, 287,
297–8
Cribari-Neto, F., 88, 224, 227
- D’Agostino, R. B., 32
Darroch, J. N., 40, 150, 281
Davidson, J., 2, 11, 209
Davidson, R., 2, 8, 25–6, 28, 34, 54,
64, 68–71, 75, 77, 79, 82, 89, 119,
131, 134, 184, 188, 192–3, 200–1,
216, 219, 225–6, 231, 240, 248,
251, 253, 263, 273, 279, 282,
290–3, 301, 304
Davies, R. B., 162–3
Davison, A. C., 45, 51, 53, 64, 74,
96, 141, 197, 200, 205,
210, 212
Deb, P., 32
De Jong, R. M., 209
Delgado, M. A., 281
Den Hann, W. J., 23
Dezhbakhsh, H., 123, 235–6
Dhaene, G., 213
Diebold, F. X., 37, 135, 166–70, 172
Dufour, J.-M., 47–8, 53, 59, 78, 83–5,
88–9, 91–3, 95–7
Dupleich Ulloa, M. R., 267, 269
Durbin, J., 119, 149, 152, 301
Dwass, M., 47
- Eastwood, A., 123
Efron, B., 45, 51, 53, 61, 304
Eicker, F., 18
Engle, R. F., 237
Ericsson, N. R., 279
Ervin, L. H., 20, 218, 224–5

- Fan, Y., 281–5, 287, 293
 Fisher, G. R., 273, 280
 Fitzenberger, B., 207, 210–13, 265
 Flachaire, E., 185, 188, 193, 225–6,
 231, 239–40, 263
 Franke, J., 66, 196–7
 Freedman, D. A., 72, 102, 109,
 184, 211
 Fuertes, A.-M., 264
- van Giersbergen, N. P. A., 104–6,
 109–10, 123
 Glejser, H., 90
 Godfrey, L. G., 9, 13, 24, 33, 35, 39,
 49–50, 56, 75, 83, 88–9, 91, 93,
 96–9, 118–19, 122–3, 128, 131–2,
 136, 145, 149–53, 157, 159–60,
 163, 172, 174, 193, 226–30,
 232–3, 235–6, 238–40, 259, 263,
 272–5, 279–81, 284–90, 293–4,
 297, 300, 302
 Gong, G., 51
 Gonçalves, S., 190–3, 212–13, 236,
 252, 257
 Götze, F., 210, 212
 Gourieroux, C., 118, 267, 269
 Graybill, F. A., 10
 Greene, W. H., 1–2, 15, 17, 27, 37,
 102, 111, 149, 156, 193, 222,
 227–8, 248, 251, 267, 278
 Gujarati, D. N., 1
- Haldrup, N., 202
 Hall, A., 118
 Hall, P., 45, 54–5, 64, 67–8, 71, 74, 82,
 105, 201–2, 204, 210, 212
 Hamilton, J. D., 22, 190, 194, 206
 Hansen, B. E., 17, 42, 78, 101, 135,
 161, 165, 172–3, 178, 192, 220,
 242–7, 278
 Härdle, W., 198, 201–2, 204, 206, 215
 Harvey, A. C., 182
 Hausman, J., 24, 220–1, 247, 249,
 251, 264
 Hedges, S. B., 68
 Hendry, D. F., 37–8, 138–9
 Hidalgo, J., 206
- Hill, R. C., 160
 Hinkley, D. V., 45, 48, 51, 53, 64, 74,
 96, 141–2, 184, 197, 200, 205
 Hoorelbeke, D., 213
 Horowitz, J. L., 34, 67, 69, 71, 79, 111,
 118, 130, 141, 144, 184, 191,
 200–1, 204, 206, 210, 216,
 288–9, 304
 Hsieh, D. A., 21
 Hu, F., 188–90
- Jansson, M., 202
 Jarque, C. M., 31–2, 40, 84, 132, 152
 Jeong, J., 89, 300
 Jewitt, I., 224
 Jöckel, K.-H., 67
 Jouini, J., 244, 246
- Kalbfleisch, J. D., 190
 Kennedy, P. E., 292, 304
 Khalaf, L., 47, 78
 Kiefer, N. M., 21, 23, 208–10, 212, 256
 Kilian, L., 79, 190–3, 198, 236,
 253, 304
 Kim, S., 298
 Kiviet, J. F., 36, 38, 104–6, 109–10, 123
 Koenker, R., 157, 159
 Krämer, W., 24, 249
 Kreiss, J.-P., 196–7
 Künsch, H. R., 200, 210, 212
- Lai, T. L., 102
 Lamarche, J.-F., 173, 175, 246
 Lee, K., 41, 89
 Léger, C., 201
 Levin, A., 23
 Li, H., 153, 195, 198, 206, 215, 304
 Li, J., 251, 253–7, 259–60, 264
 Li, Q., 298
 Lien, D., 277
 Ligeralde, A. V., 209, 256
 Liu, R. Y., 178, 186–7
 Long, J. S., 20, 218, 224–5
 Luger, R., 292

- McAleer, M., 267, 273, 276,
279–80, 293
- McCabe, B., 11
- McCullough, B. D., 67, 144, 205
- MacDonald, G. M., 27, 31
- MacKinnon, J. G., 2, 8, 20, 25–6, 28,
34, 64, 67–71, 75, 77, 79, 88–9,
106, 119, 122, 131, 134, 177, 184,
192, 200–1, 216, 219, 223–6, 228,
245–6, 248, 251, 253, 267, 272–4,
279–80, 282, 290–4, 298, 301, 304
- McQuarrie, A. D. R., 64
- Maddala, G. S., 153, 195, 198, 206,
215, 300, 304
- Mammen, E., 45, 102, 184–7, 198,
202, 206
- Mann, H., 11
- Mantalos, P., 122
- Marriott, F. H. C., 49, 67
- Michelis, L., 275, 279–80, 282
- Milliken, G. A., 10
- Mizon, G. E., 193, 216
- Monahan, J. C., 21, 23
- Monfardini, C., 298
- Monfort, A., 118, 267, 269
- Nandi, S., 45, 198, 202, 206
- Navidi, W., 78
- Neumeyer, N., 59, 152
- Newey, W. K., 21, 23, 262
- O'Reilly, G., 246–7
- Orme, C. D., 33, 35, 39, 75, 89, 96,
136, 145, 149, 157, 172,
226–30, 259
- Pagan, A. R., 8, 33, 56–7, 90, 152, 159
- Paparoditis, E., 63, 117
- Park, J. Y., 202
- Paulsen, J., 205
- Perron, P., 162, 172
- Pesaran, B., 298
- Pesaran, M. H., 267, 269, 271–5,
279–82, 284–5, 290, 294, 298, 301
- Phillips, A. W., 233
- Pierce, D. A., 197
- Politis, D. N., 63, 117, 198, 200–2,
206, 214
- Procidano, I., 214
- Quandt, R. E., 37, 162
- Ramsey, J. B., 9, 24, 122, 149, 221–2
- Rayner, R. K., 121–2, 127, 195
- Romano, J. P., 200–1, 206, 214
- Rowley, J. C. R., 15
- Salkever, D. S., 137–8
- Samworth, R., 80
- Santos, C., 138
- Santos Silva, J. M. C., 297, 300
- Savin, N. E., 111, 130, 289
- Schmidt, P., 156, 222, 275
- Schwert, G. W., 122, 204
- Sefton, M., 32
- Serlin, R. C., 26, 68, 129, 160, 170
- Sharma, S. C., 14, 28
- Shibata, R., 203
- Silvey, S. D., 40, 150, 281
- Singh, K., 178
- Sonnberger, H., 24, 249
- Spanos, A., 216
- Staiger, D., 255
- Stengos, T., 281
- Stewart, K. G., 10
- Stock, J. H., 17, 37, 135, 161, 232,
245, 255
- Szroeter, J., 267, 279
- Taylor, A. M. R., 237
- Thursby, J. G., 123, 156, 222, 235–6
- Tibshirani, R. J., 45, 53–4, 61
- Titterton, D. M., 55, 67–8
- Tjøstheim, D., 205
- Tremayne, A. R., 11, 13, 123, 193–4,
203, 232, 235–6, 238–9
- Tsai, C.-L., 64
- Urzúa, C. M., 32

- Veall, M. R., 151
Verbeek, M., 160, 190, 263
Vinod, H., 67, 144
Vogelsang, T. J., 21, 23, 208–10,
212, 256
Vuong, Q. H., 277, 279
- Wald, H., 11
Walker, A. M., 298
Watson, G. S., 149
Watson, M. W., 17, 37, 135, 161,
232, 245
Weeks, M., 267, 279
Wei, C. Z., 102
West, K. D., 21, 23, 262
Whelan, K., 246–7
- White, H., 11, 18–20, 27, 31, 33, 40,
91, 96, 118, 179, 213, 218, 223–5,
228, 246, 252, 257, 262, 264, 271
Wilson, S. R., 54
Wilton, D. A., 15
Wise, J., 233
Wong, K.-F., 247, 251
Wooldridge, J. M., 1
Wu, C. F. J., 186
- Yekutieli, D., 150
Yeo, I.-K., 297–8
- Zarkos, S. G., 88, 227
Zidek, J. V., 188–9
Zietz, J., 218

Subject Index

- absolute discrimination tests, 278
- adjusted p -value, 74, 142, 258
 - see also* double bootstrap
- Akaike Information Criterion (AIC), 203, 204, 278
- Andrews *SupF* test, 242, 245
- asymptotic p -values, 41, 57, 158, 165
- asymptotic pivot, 49, 70, 72, 78, 93
- asymptotic refinements (associated with bootstrap), 70–1, 74, 76, 82, 101, 141, 210
- asymptotic theory, 2, 11–2, 93, 171
 - tests based on, 2–4, 19, 24–5, 69, 138, 150, 191, 201, 207–10, 215, 242, 245, 266, 271
- asymptotically cooperative, 232, 275
- asymptotically pivotal statistic, 49, 55, 70, 74, 98, 106
- asymptotically valid tests, 1, 17, 36, 179
- autoregressive (recursive) bootstrap, 65–6, 109, 120, 191–3, 234, 244, 263
- autoregressive distributed lag (ADL) model, 30
- autocorrelation test, 9, 231–2
 - see also* Breusch-Godfrey test
- autocovariance matrices, 22
- autoregressive (AR) models, 30, 119, 121–2, 124, 127, 130–1, 190, 195, 202, 204, 206, 233, 255
- autoregressive-moving average (ARMA) models, 124, 194
 - bootstrap for, 196–8
- battery of tests, 135, 149–50, 152, 155, 281
- bandwidth, 208–10, 256, 259
- blurring of the critical region (blurring effect), 49, 67
- blocks, 199
 - moving (overlapping), 200
 - non-overlapping, 199
- block bootstraps, 198–201
 - for HAC tests, 210–3
 - for autocorrelation-consistent Hausman test, 253–6
- block length, 199–201, 215
 - in HAC tests, 211
- bootstrap
 - asymptotic properties of, 69–72
 - asymptotic refinements, 70, 74, 76, 78, 82, 141, 210
 - autocorrelation-robust Hausman test, 247–53
 - autoregressive (recursive), 65–6, 109, 120, 191–3, 234, 244, 263
 - cumulative distribution function (CDF), 62, 103, 106, 111, 140, 287
 - data generation process (DGP), 50, 54–6, 60, 62–3, 66, 70, 75–7, 103, 106, 183, 192, 298
 - error distribution, 61
 - F-tests, 63, 101–18
 - fixed regressor, 66, 172
 - gains from, 69–70
 - golden rules, 54, 62
 - heteroskedasticity-robust
 - autocorrelation tests, 231–41
 - heteroskedasticity-robust regression specification error tests, 221–31
 - heteroskedasticity-robust structural break tests with unknown breakpoint, 241–7
 - inconsistency, 79–80
 - law, 52, 121
 - methods for heteroskedastic
 - autocorrelated errors, 207–14
 - methods for homoskedastic
 - autocorrelated errors, 193–207
 - methods for independent
 - heteroskedastic errors, 178–93

- bootstrap – *continued*
 - minimum p -value test, 152–5
 - model based, 178, 181–3, 194, 198
 - nonparametric, 55, 78, 94, 96, 103, 132–3
 - parametric, 55–6, 59, 139
 - p -value, 53–4, 68, 72, 107, 264
 - recursive (autoregressive), 65–6, 109, 120, 191–3, 234, 244, 263
 - restricted, 104, 131, 133, 143
 - sample mean estimator, 52
 - sample variance estimator, 52
 - samples, 50, 54, 67, 72, 74, 111, 213, 304
 - tests, 45, 50–5, 62, 64, 69–72, 78–9, 82, 177
 - tests for functional form, 300
 - tests for heteroskedasticity, 94–5
 - tests for linear coefficient restrictions, 101–8
 - tests for nonnested regression models, 281–9
 - tests for predictive failure, 136–49
 - tests for serial correlation, 118–9
 - tests for structural breaks, 160–73
 - unrestricted, 104, 121–3, 133
 - variance-covariance matrix, 213
 - wild, 185–8, 214–5, 226, 237, 239, 243
- breakpoint, 135, 161–3, 167, 172, 241–2
- Breusch-Godfrey test, 9, 83, 149
 - bootstrap version of, 118–32
 - heteroskedasticity-robust bootstrap version of, 231–41
- Breusch-Pagan (BP) statistic, 58, 90
- Breusch-Pagan test, 9, 33, 90, 232, 233
 - rejection rates, 33
- Burg method, 205

- Central Limit Theorem (CLT), 11, 14, 33, 50
- Chow's prediction error test
 - statistic, 39
- Chow's test, 36–9
 - for prediction errors, 37–9
 - for structural breaks, 36–7, 160
- classical linear regression model, 3–5
 - tests for, 3–10
 - assumptions about the error term, 3
- conditional heteroskedasticity, 231
- consistency
 - of bootstrap covariance matrix estimator, 213
 - of bootstrap for ARMA model, 196
 - of bootstrap for LLR test of covariance matrix estimator, 17, 19–20, 22–3, 208–9, 262
 - of estimator of error CDF, 60, 106, 139
 - of OLS error variance estimator, 6
 - of OLS estimator, 19, 194, 233, 264
 - of OLS estimator under autocorrelation, 194
 - of OLS estimator under Pitman drift, 123
 - of pairs bootstrap, 184
 - of sieve bootstrap, 204
 - of wild bootstrap, 186
- Cox-type LLR tests, 269–73, 294
- criteria for robustness, 26
- critical value (for F test), 7
- cross-section data, 20, 151, 190
- cumulative distribution function (CDF), 39, 49, 55, 59, 93, 132, 139, 151

- data generation process (DGP), 25, 66, 103, 182, 190, 271, 277
- double bootstrap, 72–7, 139, 143–4, 173, 213
 - fast, 75, 173, 175
 - see also* fast double bootstrap (FDB)
 - for minimum p -value test, 153–5
 - p -value, 74
 - stopping rules, 75, 144
 - predictive tests, 141–4
- decision rule
 - and hypothesis testing, 7
 - for F test, 7
 - for one-sided t test, 8
 - for two-sided t test, 8
- Durbin h -test, 152
- Durbin-Watson test, 88, 149

- dynamic models, 29, 64, 109–10, 232
dynamic stability, 65–6, 124–5, 232
- Edgeworth expansions, 71, 79, 82
empirical distribution function (EDF),
51, 57, 60, 62, 70, 76, 106, 206
endogeneity, 247, 264
error in rejection probability (ERP),
70–1, 75, 82, 93–4, 118, 139, 141,
171, 173
errors-in-variables, 247, 264
estimated standard error, 7
estimating function bootstrap, 188–90
estimation sample, 37–8, 136–8, 143
estimator contrasts, 249
exact validity
of F test, 7, 10–1
of Monte Carlo test, 47, 49, 59
of t test, 10
exchangeable random variables, 293
exogenous regressors, 10
expansion of CDF of test statistic,
70, 74
- F statistic, 7–8, 103, 242
the distribution of, 7, 10
under non-Normality, 28–9
see also Wald statistic
- F test, 9, 27, 42, 102, 118
asymptotic validity, 27
and non-Normality of errors, 14,
28–9
comprehensive model, 274
exact validity, 7, 10–1, 27,
for structural breaks, 160
for production function example, 27
relationship with t test, 7
- Fast Double Bootstrap (FDB), 75–7,
131, 134, 291–3
of autocorrelation robust Hausman
test, 257–61
- Feasible Generalised Least Squares
(FGLS), 16–7, 182
assumptions of, 16
versus OLS in finite samples, 17
frequency domain, 206
fixed alternative, 13
- Generalized Least Squares (GLS)
estimator, 15
conditional covariance of, 15
feasibility of, 16
- Generalized linear regression
model, 14
tests for, 14–25
- Glejser test, 90
- HAC tests, 21–5, 201–14
autocovariance matrices in HAC
estimation, 22
new asymptotic theory, 209–10
traditional asymptotic theory, 207–9
using instrumental variables, 23
- hat matrix, 5
- Hausman test 24, 220, 247–49
bootstrap for
autocorrelation-consistent
version, 247–53
- HCCME-based tests, 18–21
HCR version, 33, 226, 234
- Hendry's test for predictive failure,
36, 39
- heteroskedasticity and autocorrelation
consistent (HAC) estimator, 17,
23, 251
- heteroskedasticity consistent
covariance matrix estimator
(HCCME), 17, 19, 20, 223–6, 230
- heteroskedasticity-robust tests, 19
use of in significance testing, 33–5
- Heteroskedasticity test
Breusch and Pagan, 8
Glejser, 90
Koenker, 157
White, 91
- Heteroskedasticity-valid bootstraps,
178–81
- ideal bootstrap, 67
- IID valid bootstrap, 214, 241–2,
244, 247
- IID valid covariance matrix, 17
- impulse response coefficients, 79

- information matrix (IM) test, 118
instrumental variable estimation,
24, 121
intersection null hypothesis, 150
- J test, 273–5, 279–80, 292
adjusted J-test, 273
bootstrapped, 282–5
- Jarque-Bera test, 31–2, 40–1, 84–7
Monte Carlo version, 85–7
asymptotic distribution, 32
- kernel function, 208
Koenker test, 157
- Lagrange Multiplier (LM) tests, 9,
56–7, 84, 118–22, 157, 159, 233–5
heteroskedasticity-robust, 234
see also Breusch-Godfrey test *and*
Breusch-Pagan test
- Law of Large Numbers, 11
leverage values, 5, 14, 28
likelihood ratio (LR), 163–4
locally equivalent alternatives, 118
log-likelihood ratio (LLR) statistic, 268
bootstrapping with non-nested
models, 297–301
Cox-type, 269–73
centred, 272
LSE approach, 216–7
- minimum p -value test, 153–5
MLE estimator of variance, 6
model selection criteria, 278
model based bootstrap, 178, 181–3,
194–8
Monte Carlo
 p -value, 48
rejection rule, 48
robustness to misspecification of
distribution, 49, 93
tests, 47–50, 58, 88, 101
for heteroskedasticity, 89–94
for non-Normality, 83–8
test rejection rule, 48
- moving average (MA) models, 119,
122, 233, 255–6
moving block bootstrap, 200
in HAC tests, 213
Multiple testing with
diagnostic checks, 149–60
non-nested alternatives, 275–7
multivariate normal distribution, 4
- naive test, 211–2, 302
nested hypotheses, 266–7
Newey-West estimator, 23, 208
NID (normally and independently
distributed), 3
non-nested hypotheses, 266, 268
non-nested time series models, 298
non-Normality of errors
tests for, 31–3
see also Jarque-Bera test, 31
non-standard tests, 36
Nonlinear Least Squares (NLS), 16
nonparametric bootstraps, 55, 78, 94,
103, 133
nuisance parameters, 46
null distribution, 37, 53, 63, 70, 72,
81, 91, 94, 103, 134, 139, 141,
164, 209–10, 277, 299–300
null hypothesis, 6–7, 10–2, 19, 24, 27,
31, 36, 38, 46, 59, 84, 89–90, 101,
136, 160, 184–5, 209, 221, 233,
241, 248, 271, 277
number of artificial samples, 47, 52,
choice of, 67–9
- Okun's law, 156
optimal block length, 201
orders of magnitude for random
variables, 11
order of probability, 12
Ordinary Least Squares (OLS)
estimator, 4, 270
consistency and asymptotic
Normality, 12, 19, 22
degrees of freedom, 5–7, 14
inconsistency and specification
errors, 263

- Ordinary Least Squares (OLS)
 estimator – *continued*
 inconsistency in dynamic models
 with autocorrelated errors, 233
 predicted values, 4
 residual sum of squares (RSS) from, 5
 residuals, 5
 under the assumptions of the
 generalized regression model,
 14–8
 under the classical assumptions,
 3–10
 under weaker assumptions (random
 regressors and IID errors), 10–4
 orthogonality assumption, 221
 orthogonal regressors, 274
 absence-of-orthogonality
 condition, 275
 near population orthogonality
 (NPO), 275
 overall significance level, 40
- pairs bootstrap, 64, 183–5, 188–9,
 192, 226
 modified version, 185
 parametric bootstrap, 55–9
 permutation test, 292
 pick distribution, 185–9, 193, 226,
 230, 235, 244–7, 263
 Pitman drift, 13, 123
 pivotal statistic, 47, 49, 81, 88–9,
 101, 132
 post-blackening, 205
 pre-whitening, 205
 prediction sample, 37, 136, 138
 predictive failure, 37–8
 predictive tests, 37, 136–9
 bootstrapping, 139–44
 prepivoting, 72
see also double bootstrap
 production function example, 27, 226
 projection matrices, 75, 143–4, 249
 pseudo-samples, 45
p-value, 41
 for bootstrap test, 53–4
 for double bootstrap test, 74
 for fast double bootstrap test, 75–6
 for two-sided alternative, 48
- Quandt likelihood ratio (QLR)
 method, 37
 Quandt SupLR statistic, 163
 quasi-likelihood, 163
- Rademacher distribution, 188, 193,
 215, 226, 230, 235, 239, 240, 245
see also pick distribution
 recentred values, 52, 54, 61–2, 94,
 104, 151, 205
 recursive (autoregressive) bootstrap,
 65–6, 109, 120, 191–3, 234,
 244, 263
 regularity conditions, 19, 21, 39, 48,
 57, 66, 72, 93, 109, 119, 189–90,
 207, 214, 271–2, 274, 277
 relative discrimination tests, 278
 replications, 28, 125, 128, 170, 283
 resampling, 45, 54–5, 62–4, 66, 69,
 177–8
 restricted and unrestricted residuals,
 107–9
 using classical residual scheme
 under heteroskedasticity, 180–1
 RESET F-statistic, 9–10
 RESET test, 9, 24, 149, 222
 and HAC estimator of covariance
 matrix, 24
 heteroskedasticity-robust, 226–31
 restricted bootstrap test, 104, 130–1,
 133
 restricted estimator, 6, 62, 77, 139,
 185–6
 restricted (null) model, 8, 34, 156, 266
 restricted residuals, 6, 24–5, 35, 62–3,
 103–4, 106, 111, 118, 140, 225–6,
 234–5, 256
 use in HAC estimation, 24
 use in HCCME, 225, 230–1
 robustness of tests, criteria for, 26
- sandwich covariance matrix, 18,
 22, 191
 Schwarz Bayesian Information
 Criterion (BIC), 278

- separate family of hypotheses, 268
 - see also* non-nested hypotheses
- sequence of local alternatives, 13
 - see also* Pitman drift, 13
- sieve bootstrap, 201–5, 216, 264
 - with wild bootstrap, 214
- significance of a subset of regressors, 24
- simple regression, 63, 179, 213
- simulation-based tests
 - for heteroskedasticity, 88–101
- simulation experiments and results
 - for asymptotic and bootstrap F tests, 110–8
 - for autocorrelation-robust Hausman test, 253–62
 - for battery of OLS diagnostic tests, 155–60
 - for heteroskedasticity-robust test for autocorrelation, 235–41
 - for heteroskedasticity-robust test for specification errors, 226–31
 - for non-nested regression models (one alternative), 281–90
 - for non-nested regression models (two alternatives), 293–7
 - for tests for significance of regressors, 27–31
 - for tests for heteroskedasticity, 95–101
 - for tests for predictive failure, 144–8
 - for tests for serial correlation, 123–32
 - for tests for structural breaks, 166–73
- skedastic function, 181–3, 227
- standard errors, 7, 13
 - autocorrelation robust, 253, 257
 - heteroskedasticity and autocorrelation robust, 17
 - heteroskedasticity robust, 17, 35
- stationary variables, 15, 125, 172, 193–4, 198, 201–2, 205
- strictly exogenous regressors, 10
- structural break, 135
- Student *t* distribution, 7
- subsampling, 206, 214
- Sup* test, 37, 164–5, 173
 - for F statistics, 242–3, 245
 - for LM statistics, 163–4, 167
 - for LR statistics, 163–4, 167
 - for nonnested models, 277, 294, 297
 - for W statistics, 163–4, 167
- symmetric error distribution, 91, 98–100
- t-ratio, 7
- t-statistic, 12
- t-test, 8, 11, 273, 301
 - autocorrelation-robust, 253
 - heteroskedasticity-robust, 33–4, 231
 - relationship with F test, 7
- time domain, 206
- time series
 - data, 11, 21, 37, 156, 193, 231
 - linear, 202
 - models, 66, 79, 298
 - regressions, 20–1, 42, 192, 215
- trimming of sample, 164–5
- unit root, 21, 202, 204, 303
- unrestricted (alternative) model, 8, 24, 34, 266
- unrestricted bootstrap test, 104, 124, 133
- unrestricted estimator, 6, 62, 104–5
- unrestricted residuals, 6, 25, 35, 63, 103–4, 106, 111, 118, 225, 234, 246, 256
 - use in HAC estimation, 25
 - use in HCCME, 19–20
- Wald statistic
 - for autocorrelation-robust Hausman test, 250–1
 - for GLS, 15
 - HAC version for OLS, 209
 - heteroskedasticity-robust version for OLS, 19–20, 223, 235, 245

- White's heteroskedasticity-consistent covariance matrix estimator (HCCME), 17, 18, 34–5, 218, 223–6, 230, 245–6
- White's test, 40–1, 91, 93, 96–7
- Wild bootstrap, 185–8, 214–5, 226, 237, 239, 243
- fixed-design, 191
- recursive-design, 191
- with sieve bootstrap, 214
- Yule-Walker method, 205