

# A Companion to Theoretical Econometrics

*Edited by*

---

BADI H. BALTAGI  
*Texas A & M University*



---

A COMPANION TO THEORETICAL ECONOMETRICS

## **Blackwell Companions to Contemporary Economics**

The *Blackwell Companions to Contemporary Economics* are reference volumes accessible to serious students and yet also containing up-to-date material from recognized experts in their particular fields. They focus on basic, bread-and-butter issues in economics as well as popular contemporary topics often not covered in textbooks. Coverage avoids the overly technical, is concise, clear, and comprehensive. Each Companion features an introductory essay by the editor, bibliographical reference sections, and an index.

*A Companion to Theoretical Econometrics* edited by Badi H. Baltagi

*A Companion to Business Forecasting* edited by Michael P. Clements and David F. Hendry

### **Forthcoming:**

*A Companion to the History of Economic Thought* edited by Warren J. Samuels, Jeff E. Biddle, and John B. Davis

© 2001, 2003 by Blackwell Publishing Ltd

350 Main Street, Malden, MA 02148-5018, USA  
108 Cowley Road, Oxford OX4 1JF, UK  
550 Swanston Street, Carlton South, Melbourne, Victoria 3053, Australia  
Kurfürstendamm 57, 10707 Berlin, Germany

All rights reserved. No part of this publication may be reproduced, stored in a retrieval system, or transmitted, in any form or by any means, electronic, mechanical, photocopying, recording or otherwise, except as permitted by the UK Copyright, Designs, and Patents Act 1988, without the prior permission of the publisher.

First published 2001  
First published in paperback 2003 by Blackwell Publishing Ltd

*Library of Congress Cataloging-in-Publication Data*

A companion to theoretical econometrics / edited by Badi H. Baltagi.  
p. cm. — (Blackwell companions to contemporary economics)  
A collection of articles by an international group of scholars.  
Includes bibliographical references and index.  
ISBN 0-631-21254-X (hb : alk. paper) — ISBN 1-4051-0676-X (pb. : alk.  
paper)  
1. Econometrics. I. Title: Theoretical econometrics. II. Baltagi, Badi H.  
(Badi Hani) III. Series.

HB139.C643 2000  
330'.01'5195—dc21

00-025862

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

Set in 10 on 12pt Book Antique  
by Graphicraft Ltd., Hong Kong  
Printed and bound in the United Kingdom  
by T. J. International Ltd., Padstow, Cornwall

For further information on  
Blackwell Publishing, visit our website:  
<http://www.blackwellpublishing.com>

# Contents

<i>List of Figures</i>	viii
<i>List of Tables</i>	ix
<i>List of Contributors</i>	x
<i>Preface</i>	xii
<i>List of Abbreviations</i>	xiv
<i>Introduction</i>	1
<b>1 Artificial Regressions</b> <i>Russell Davidson and James G. MacKinnon</i>	16
<b>2 General Hypothesis Testing</b> <i>Anil K. Bera and Gamini Premaratne</i>	38
<b>3 Serial Correlation</b> <i>Maxwell L. King</i>	62
<b>4 Heteroskedasticity</b> <i>William E. Griffiths</i>	82
<b>5 Seemingly Unrelated Regression</b> <i>Denzil G. Fiebig</i>	101
<b>6 Simultaneous Equation Model Estimators: Statistical Properties and Practical Implications</b> <i>Roberto S. Mariano</i>	122
<b>7 Identification in Parametric Models</b> <i>Paul Bekker and Tom Wansbeek</i>	144

<b>8</b>	<b>Measurement Error and Latent Variables</b>	162
	<i>Tom Wansbeek and Erik Meijer</i>	
<b>9</b>	<b>Diagnostic Testing</b>	180
	<i>Jeffrey M. Wooldridge</i>	
<b>10</b>	<b>Basic Elements of Asymptotic Theory</b>	201
	<i>Benedikt M. Pötscher and Ingmar R. Prucha</i>	
<b>11</b>	<b>Generalized Method of Moments</b>	230
	<i>Alastair R. Hall</i>	
<b>12</b>	<b>Collinearity</b>	256
	<i>R. Carter Hill and Lee C. Adkins</i>	
<b>13</b>	<b>Nonnested Hypothesis Testing: An Overview</b>	279
	<i>M. Hashem Pesaran and Melvyn Weeks</i>	
<b>14</b>	<b>Spatial Econometrics</b>	310
	<i>Luc Anselin</i>	
<b>15</b>	<b>Essentials of Count Data Regression</b>	331
	<i>A. Colin Cameron and Pravin K. Trivedi</i>	
<b>16</b>	<b>Panel Data Models</b>	349
	<i>Cheng Hsiao</i>	
<b>17</b>	<b>Qualitative Response Models</b>	366
	<i>G.S. Maddala and A. Flores-Lagunes</i>	
<b>18</b>	<b>Self-Selection</b>	383
	<i>Lung-fei Lee</i>	
<b>19</b>	<b>Random Coefficient Models</b>	410
	<i>P.A.V.B. Swamy and George S. Tavlas</i>	
<b>20</b>	<b>Nonparametric Kernel Methods of Estimation and Hypothesis Testing</b>	429
	<i>Aman Ullah</i>	
<b>21</b>	<b>Durations</b>	444
	<i>Christian Gouriéroux and Joann Jasiak</i>	
<b>22</b>	<b>Simulation Based Inference for Dynamic Multinomial Choice Models</b>	466
	<i>John Geweke, Daniel Houser, and Michael Keane</i>	
<b>23</b>	<b>Monte Carlo Test Methods in Econometrics</b>	494
	<i>Jean-Marie Dufour and Lynda Khalaf</i>	
<b>24</b>	<b>Bayesian Analysis of Stochastic Frontier Models</b>	520
	<i>Gary Koop and Mark F.J. Steel</i>	
<b>25</b>	<b>Parametric and Nonparametric Tests of Limited Domain and Ordered Hypotheses in Economics</b>	538
	<i>Esfandiar Maasoumi</i>	

---

<b>26</b>	<b>Spurious Regressions in Econometrics</b>	557
	<i>Clive W.J. Granger</i>	
<b>27</b>	<b>Forecasting Economic Time Series</b>	562
	<i>James H. Stock</i>	
<b>28</b>	<b>Time Series and Dynamic Models</b>	585
	<i>Aris Spanos</i>	
<b>29</b>	<b>Unit Roots</b>	610
	<i>Herman J. Bierens</i>	
<b>30</b>	<b>Cointegration</b>	634
	<i>Juan J. Dolado, Jesús Gonzalo, and Francesc Marmol</i>	
<b>31</b>	<b>Seasonal Nonstationarity and Near-Nonstationarity</b>	655
	<i>Eric Ghysels, Denise R. Osborn, and Paulo M.M. Rodrigues</i>	
<b>32</b>	<b>Vector Autoregressions</b>	678
	<i>Helmut Lütkepohl</i>	
	<i>Index</i>	700

# Figures

<b>19.1</b>	Short-term interest rate elasticity for RCM1 (without concomitants)	425
<b>19.2</b>	Short-term interest rate elasticity for RCM2 (with concomitants)	425
<b>21.1</b>	Censoring scheme: unemployment spells	455
<b>21.2</b>	Truncation scheme	456
<b>21.3</b>	Hazard functions for accelerated hazard models	457
<b>21.4</b>	Hazard functions for proportional hazard models	458
<b>21.5</b>	(Under) Overdispersion of intertrade durations	463
<b>22.1</b>	Marginal posterior densities of first log-wage equation's parameters from data set 3-EMAX	485
<b>22.2</b>	EMAX and polynomial future components evaluated at mean values of state variables at each period	487
<b>27.1</b>	US unemployment rate, recursive AR(BIC)/unit root pretest forecast, and neural network forecast	568
<b>27.2</b>	Six-month US CPI inflation at an annual rate, recursive AR(BIC)/unit root pretest forecast, and neural network forecast	568
<b>27.3</b>	90-day Treasury bill at an annual rate, recursive AR(BIC)/unit root pretest forecast, and neural network forecast	569
<b>27.4</b>	Six-month growth of US industrial production at an annual rate, recursive AR(BIC)/unit root pretest forecast, and neural network forecast	569
<b>27.5</b>	Six-month growth of total real US manufacturing and trade inventories at an annual rate, recursive AR(BIC)/unit root pretest forecast, and neural network forecast	570
<b>28.1</b>	US industrial production index	586
<b>28.2</b>	De-trended industrial production index	587
<b>29.1</b>	Density of $\rho_0$	617
<b>29.2</b>	Density of $\tau_0$ compared with the standard normal density	618
<b>29.3</b>	Density of $\rho_1$	619
<b>29.4</b>	Density of $\tau_1$ compared with the standard normal density	620
<b>29.5</b>	Density of $\rho_2$	630
<b>29.6</b>	Density of $\tau_2$ compared with the standard normal density	631

# Tables

<b>12.1</b>	Matrix of variance proportions	261
<b>12.2</b>	Harmful collinearity decision matrix	267
<b>19.1</b>	Long-run elasticities	424
<b>19.2</b>	Long-run elasticities and direct effects from RCM2	424
<b>22.1</b>	Quality of polynomial approximation to the true future component	477
<b>22.2</b>	Choice distributions and mean accepted wages in the data generated with true and OLS polynomial future components	480–1
<b>22.3</b>	Descriptive statistics for posterior distributions of the model's structural parameters for several different data sets generated using polynomial future component	482
<b>22.4</b>	Log-wage equation estimates from OLS on observed wages generated under the polynomial future component	483
<b>22.5</b>	Descriptive statistics for posterior distributions of the model's structural parameters for several different data sets generated using true future component	484
<b>22.6</b>	Wealth loss when posterior polynomial approximation is used in place of true future component	486
<b>23.1</b>	IV-based Wald/Anderson–Rubin tests: empirical type I errors	499
<b>23.2</b>	Kolmogorov–Smirnov/Jarque–Bera residuals based tests: empirical type I errors	501
<b>23.3</b>	Empirical type I errors of multivariate tests: uniform linear hypotheses	503
<b>26.1</b>	Regression between independent AR(1) series	560
<b>27.1</b>	Comparison of simulated out-of-sample linear and nonlinear forecasts for five US macroeconomic time series	575
<b>27.2</b>	Root mean squared forecast forecast errors of VARs, relative to AR(4)	575
<b>32.1</b>	Models and LR type tests	687

# Contributors

**Lee C. Adkins**, Oklahoma State University  
**Luc Anselin**, University of Illinois  
**Paul Bekker**, University of Groningen  
**Anil K. Bera**, University of Illinois  
**Herman J. Bierens**, Pennsylvania State University  
**A. Colin Cameron**, University of California – Davis  
**Russell Davidson**, Queen's University, Ontario, and GREQAM, Marseilles  
**Juan Dolado**, Universidad Carlos III de Madrid  
**Jean-Marie Dufour**, University of Montreal  
**Denzil G. Fiebig**, University of Sydney  
**A. Flores-Lagunes**, University of Arizona  
**John Geweke**, University of Minnesota and University of Iowa  
**Eric Ghysels**, Pennsylvania State University  
**Jesús Gonzalo**, Universidad Carlos III de Madrid  
**Christian Gouriéroux**, CREST and CEPREMAP, Paris  
**Clive W.J. Granger**, University of California – San Diego  
**William E. Griffiths**, University of Melbourne  
**Alastair R. Hall**, North Carolina State University  
**R. Carter Hill**, Louisiana State University  
**Daniel Houser**, University of Arizona  
**Cheng Hsiao**, University of Southern California  
**Joann Jasiak**, York University  
**Michael Keane**, University of Minnesota and New York University  
**Lynda Khalaf**, University of Laval  
**Maxwell L. King**, Monash University  
**Gary Koop**, University of Glasgow  
**Lung-fei Lee**, Hong Kong University of Science & Technology  
**Helmut Lütkepohl**, Humboldt University  
**Esfandiar Maasoumi**, Southern Methodist University, Dallas  
**James MacKinnon**, Queen's University, Ontario

- G.S. Maddala, Ohio State University  
Roberto S. Mariano, University of Pennsylvania  
Francesc Marmol, Universidad Carlos III de Madrid  
Erik Meijer, University of Groningen  
Denise R. Osborn, University of Manchester  
M. Hashem Pesaran, Cambridge University  
Benedikt M. Pötscher, University of Vienna  
Gamini Premaratne, University of Illinois  
Ingmar R. Prucha, University of Maryland  
Paulo M.M. Rodrigues, University of Algarve  
Aris Spanos, Virginia Polytechnic Institute and State University  
Mark F.J. Steel, University of Kent  
James H. Stock, Harvard University  
P.A.V.B. Swamy, Department of the Treasury, Washington  
George S. Tavlas, Bank of Greece  
Pravin K. Trivedi, Indiana University  
Aman Ullah, University of California – Riverside  
Tom Wansbeek, University of Groningen  
Melvyn Weeks, Cambridge University  
Jeffrey M. Wooldridge, Michigan State University

# Preface

This companion in theoretical econometrics is the first in a series of companions in economics published by Blackwell. The emphasis is on graduate students of econometrics and professional researchers who need a guide, a friend, a companion to lead them through this exciting yet ever growing and expanding field. This is not a handbook of long chapters or exhaustive surveys on the subject. These are simple chapters, written by international experts who were asked to give a basic introduction to their subject. These chapters summarize some of the well known results as well as new developments in the field and direct the reader to supplementary reading. Clearly, one single volume cannot do justice to the wide variety of topics in theoretical econometrics. There are five handbooks of econometrics published by North-Holland and two handbooks of applied econometrics published by Blackwell, to mention a few. The 32 chapters in this companion give only a sample of the important topics in theoretical econometrics. We hope that students, teachers, and professionals find this companion useful. I would like to thank Al Bruckner who approached me with this idea and who entrusted me with the editorial job, the 50 authors who met deadlines and page limitations.

I would also like to thank the numerous reviewers who read these chapters and commented on them. These include Seung Ahn, Paul Bekker, Anil Bera, Herman Bierens, Erik Biorn, Siddahrtha Chib, James Davidson, Francis Diebold, Juan Dolado, Jean-Marie Dufour, Neil Ericsson, Denzil Fiebig, Philip Hans Franses, John Geweke, Eric Ghysels, David Giles, Jesús Gonzalo, Clive Granger, William Greene, William Griffith, Alastair Hall, Bruce Hansen, R. Carter Hill, Cheng Hsiao, Hae-Shin Hwang, Svend Hylleberg, Michael Keane, Lynda Khalaf, Gary Koop, Lung-fei Lee, Qi Li, Oliver Linton, Helmut Lütkepohl, Essie Maasoumi, James MacKinnon, G.S. Maddala, Masao Ogaki, Denise Osborn, Pierre Perron, Peter Phillips, Ingmar Prucha, Peter Schmidt, Mark Steel, James Stock, Pravin Trivedi, Aman Ullah, Marno Verbeek, Tom Wansbeek, Rainer Winkelmann, and Jeffrey Wooldridge.

---

On a sad note, G.S. Maddala, a contributing author to this volume, died before this book was published. He was a leading figure in econometrics and a prolific researcher whose writings touched students all over the world. He will be sorely missed.

Finally, I would like to acknowledge the support and help of Blackwell Publishers and the secretarial assistance of Teri Bush at various stages of the preparation of this companion.

BADI H. BALTAGI  
*Texas A&M University  
College Station, Texas*

# Abbreviations

2SLS	two-stage least squares
3SLS	three-stage least squares
a.s.	almost sure
ACD	Autoregressive Conditional Duration
ADF	Augmented Dickey–Fuller
AE	asymptotically equivalent
AIC	Aikake's information criteria
AIMA	asymptotically ideal model
AIMSE	average integrated mean square error
AR	autoregressive
AR(1)	first-order autoregressive
ARCH	autoregressive conditional heteroskedasticity
ARFIMA	autoregressive fractionally integrated moving average
ARIMA	autoregressive integrated moving average
ARMA	autoregressive moving average
BDS	Brock, Dechert, and Scheinkman
BIC	Bayesian information criteria
BKW	Belsley, Kuh, and Welsch
BLUE	best, linear unbiased estimator
BMC	bound Monte Carlo
CAPM	capital asset pricing model
CAPS	consistent adjusted least squares
CDF (or cdf)	cumulative distribution function
CES	constant elasticity of substitution
CFI	comparative fit index
CG matrix	matrix of contributions to the gradient
CI	confidence interval
CLT	central limit theorem
CM	conditional moment
CME	conditional mean encompassing

CMT	conditional moment test
CPI	consumer price index
CPS	current population survey
CUAN	consistent and uniformly asymptotic normal
DEA	data envelopment analysis
DF	Dickey–Fuller
DGLS	dynamic generalized least squares
DGM	data generating mechanism
DGP	data generating process
DHF	Dickey, Hasza, and Fuller
DLR	double-length artificial regression
DOLS	dynamic ordinary least squares
DW	Durbin–Watson
DWH	Durbin–Wu–Hausman
EBA	elimination-by-aspects
ECM	expectation conditional maximization
EM	expectation maximization
EPE	estimated prediction error
ESS	explained sums of squares
ESS <sub>R</sub>	restricted sum of squares
ESS <sub>U</sub>	unrestricted error sum of squares
EWMA	exponentially weighted moving average
FCLT	functional central limit theorem
FGLS	feasible generalized least squares
FIML	full information maximum likelihood
FIVE	full information instrumental variables efficient
FM-OLS	fully modified ordinary least squares estimator
FSD	first-order stochastic dominate
FWL	Frisch–Waugh–Lovell
GARCH	generalized autoregressive conditional heteroskedastic
GEV	generalized extreme value
GHK	Geweke, Hajivassiliou, and Keane
GHM	Gouriéoux, Holly, and Montfort
GIS	geographic information systems
GL	generalized Lorenz
GLM	generalized linear model
GLN	Ghysels, Lee, and Noh
GLS	generalized least squares
GML	generalized maximum likelihood
GMM	generalized method of moments
GNR	Gauss–Newton regression
GSUR	generalized seemingly unrelated regression
HAC	heteroskedasticity and autocorrelation consistent
HEBA	hierarchical elimination-by-aspects
HEGY	Hylleberg, Engle, Granger, and Yoo
H-K	Honoré and Kyriazidou

HRGNR	heteroskedasticity-robust Gauss–Newton regression
i.p.	in probability
IC	information criteria
ID	independently distributed
IIA	independence of irrelevant alternatives
IID	independently identically distributed
IIV	iterated instrumental variable
ILS	indirect least squares
IM	information matrix
IMSE	integrated mean square error
INAR	integer autoregressive
IP	industrial production
IV	instrumental variable
JB	Jarque–Bera
KLIC	Kullback–Leibler information criterion
KPSS	Kwiatkowski, Phillips, Schmidt, and Shin
KS	Kolmogorov–Smirnov
KT	Kuhn–Tucker
LBI	locally best invariant
LCLS	local constant least squares
LEF	linear exponential family
LI	limited information
LIML	limited information maximum likelihood
LIVE	limited information instrumental variables efficient
LL	local linear
LLLS	local linear least squares
LLN	law of large numbers
LLS	local least squares
LM	Lagrange multiplier
LMC	local Monte Carlo
LMP	locally most powerful
LMPU	locally most powerful unbiased
LPLS	local polynomial least squares
LR	likelihood ratio
LS	least squares
LSE	least squares estimation
LSTAR	logistic smooth transition autoregression
M2SLS	modified two-stage least squares
MA	moving average
MA(1)	first-order moving average
MC	Monte Carlo
MCMC	Markov Chain Monte Carlo
MD	martingale difference
MDML	multivariate dynamic linear regression
MIMIC	multiple indicators–multiple causes
ML	maximum likelihood

---

MLE	maximum likelihood estimation
MLR	multivariate linear regression
MM	method of moments
MMC	maximized Monte Carlo
MML	maximum marginal likelihood
MNL	multinomial logit
MNP	multinomial probit
MP	most powerful
MS	maximum score
MSE	mean square error
MSFE	mean squared forecast error
MSL	method of simulated likelihood
MSM	method of simulated moments
MSS	method of simulated scores
NB	negative binomial
NFI	normed fit index
NLS	nonlinear least squares
NMNL	nested multinomial logit
NN	neural network
N-P	Neyman–Pearson
NPRSS	nonparametric residual sum of squares
N-W	Nadaraya–Watson
NYSE	New York Stock Exchange
OLS	ordinary least squares
OPG	outer-product-of-the-gradient
PDF	probability distribution function
PLS	predictive least squares
PML	pseudo-ML
PP	Phillips–Perron
PR	probabilistic reduction
PRSS	parametric residual sum of squares
psd	positive semi-definite
PSP	partial sum process
QML	quasi-ML
QP	quadratic programming
QRM	qualitative response model
RCM	random coefficient models
RESET	regression error specification test
RIS	recursive importance sampling
RLS	restricted least squares
RMSE	root mean squared error
RMSFE	root mean squared forecast error
RNI	relative noncentrality index
RRR	reduced rank regression
RS	Rao’s score
RSS	residual sum of squares

<i>s/n</i>	signal-to-noise
SA	simulated annealing
SAR	spatial autoregressive
SD	stochastic dominance
SEM	simultaneous equations model
SET	score encompassing test
SMA	spatial moving average
SML	simulated maximum likelihood
SNP	semi-nonparametric
SP	semiparametric
SSD	second-order stochastic dominate
SSE	sum of square error
SSR	sum of squared residuals
STAR	smooth transition autoregression
SUR(E)	seemingly unrelated regression
SVD	Stochastic Volatility Duration
TAR	transition autoregression
TSD	third-order stochastic
UI	union intersection
UL	uniform linear
ULLN	uniform law of large numbers
UMP	uniformly most powerful
UMPI	uniformly most powerful invariant
UMPU	uniformly most powerful unbiased
VAR	vector autoregression
VECM	vector error correction model
VIF	variance-inflation factor
VNM	von Neumann–Morgenstern
W	Wald
WET	Wald encompassing test
wrt	with respect to

# Introduction

*Badi H. Baltagi*

This is the first companion in econometrics. It covers 32 chapters written by international experts in the field. The emphasis of this companion is on “keeping things simple” so as to give students of econometrics a guide through the maze of important topics in econometrics. These chapters are helpful for readers and users of econometrics who are not looking for exhaustive surveys on the subject. Instead, these chapters give the reader some of the basics and point to further readings on the subject. The topics covered vary from basic chapters on serial correlation and heteroskedasticity, which are found in standard econometrics texts, to specialized topics that are covered by econometric society monographs and advanced books on the subject like count data, panel data, and spatial correlation. The authors have done their best to keep things simple. Space and time limitations prevented the inclusion of other important topics, problems and exercises, empirical applications, and exhaustive references. However, we believe that this is a good start and that the 32 chapters contain an important selection of topics in this young but fast growing field.

Chapter 1 by Davidson and MacKinnon introduces the concept of an artificial regression and gives three conditions that an artificial regression must satisfy. The widely used Gauss–Newton regression (GNR) is used to show how artificial regressions can be used for minimizing criterion functions, computing one-step estimators, calculating covariance matrix estimates, and more importantly computing test statistics. This is illustrated for testing the null hypothesis that a subset of the parameters of a nonlinear regression model are zero. It is shown that the test statistic can be computed as an explained sum of squares of the GNR divided by a consistent estimate of the residual variance. Two ways of computing this statistic are: (i) the sample size times the uncentered  $R^2$  of the GNR, or (ii) an ordinary F-statistic testing the subset of the parameters are zero from the GNR. The two statistics are asymptotically equivalent under the null hypothesis. The GNR can be used for other types of specification tests, including serial correlation, nonnested hypothesis and obtaining Durbin–Wu–Hausman type tests. This chapter also shows how to make the GNR robust to heteroskedasticity of

unknown form. It also develops an artificial regression for the generalized method of moments (GMM) estimation. The outer-product-of-the-gradient (OPG) regression is also discussed. This is a simple artificial regression which can be used with most models estimated by maximum likelihood. It is shown that the OPG satisfies the three conditions of an artificial regression and has the usual uses of an artificial regression. It is appealing because it requires only first derivatives. However, it is demonstrated that the OPG regression yields relatively poor estimates of the covariance matrices and unreliable test statistics in small samples. In fact, test statistics based on the OPG regression tend to overreject, often very severely. Davidson and MacKinnon also discuss double-length or triple-length artificial regressions where each observation makes two or three contributions to the criterion function. In this case, the artificial regression has twice or three times the sample size. This artificial regression can be used for many purposes, including tests of models with different functional forms. Finally, this chapter extends the GNR to binary response models such as the logit and probit models. For further readings on this subject, see Davidson and MacKinnon (1993).

Chapter 2 by Bera and Premaratne gives a brief history of hypothesis testing in statistics. This journey takes the reader through the basic testing principles leading naturally to several tests used by econometricians. These tests are then linked back to the basic principles using several examples. This chapter goes through the Neyman–Pearson lemma and the likelihood ratio test. It explains what is meant by locally most powerful tests and it gives the origins of the Rao-score test. Next, locally most powerful unbiased tests and Neyman's smooth test are reviewed. The interrelationship among the holy trinity of test statistics, i.e., the Wald, likelihood ratio, and Rao-score tests is brought home to the reader by an amusing story. Neyman's  $C(\alpha)$  test is derived and motivated. This approach provides an attractive way of dealing with nuisance parameters. Next, this chapter goes through some of the application of testing principles in econometrics. Again a brief history of hypothesis testing in econometrics is given beginning with the work of Ragnar Frisch and Jan Tinbergen, going into details through the Durbin–Watson statistic linking its origins to the Neyman–Pearson lemma. The use of the popular Rao-score test in econometrics is reviewed next, emphasizing that several tests in econometrics old and new have been given a score test interpretation. In fact, Bera and Premaratne consider the conditional moment test developed by Newey (1985) and Tauchen (1985) and derive its score test interpretation. Applications of Neyman's  $C(\alpha)$  test in econometrics are cited and some of the tests in econometrics are given a smooth test interpretation. The chapter finishes with a double warning about testing: be careful how you interpret the test results. Be careful what action you take when the null is rejected.

Chapter 3 by King surveys the problem of serial correlation in econometrics. Ignoring serial correlation in the disturbances can lead to inefficient parameter estimates and a misleading inference. This chapter surveys the various ways of modeling serial correlation including Box–Jenkins time series models. Auto-regressive and moving average (ARMA) models are discussed, highlighting the contributions of Cochrane and Orcutt (1949) for the AR(1) model; Thomas and Wallis (1971) for the restricted AR(4) process and Nichols, Pagan, and Terrell

(1975) for the MA(1) process. King argues that modeling serial correlation involves taking care of the dynamic part of model specification. The simple version of a dynamic model includes the lagged value of the dependent variable among the regressors. The relationship between this simple dynamic model and the AR(1) model is explored. Estimation of the linear regression model with ARMA disturbances is considered next. Maximum likelihood estimation (MLE) under normality is derived and a number of practical computation issues are discussed. Marginal likelihood estimation methods are also discussed which work well for estimating the ARMA process parameters in the presence of nuisance parameters. In this case, the nuisance parameters include the regression parameters and the residual variance. Maximizing marginal likelihoods were shown to reduce the estimation bias of maximum likelihood methods. Given the nonexperimental nature of economic data and the high potential for serial correlation, King argues that it is important to test for serial correlation. The von Neuman as well as Durbin–Watson (DW) tests are reviewed and their statistical properties are discussed. For example, the power of the DW test tends to decline to zero when the AR(1) parameter  $\rho$  tends to one. In this case, King suggests a class of point optimal tests that provide a solution to this problem. LM, Wald and LR tests for serial correlation are mentioned, but King suggests constructing these tests using the marginal likelihood rather than the full likelihood function. Testing for AR(1) disturbances in the dynamic linear regression model is also studied, and the difficulty of finding a satisfactory test for this model is explained by showing that the model may suffer from a local identification problem. The last section of this chapter takes up the problem of deciding what lags should be used in the ARIMA model. Model selection criteria are recommended rather than a test of hypotheses and the Bayesian information criteria is favored because it is consistent. This means that as the sample size goes to infinity, this criteria selects the correct model from a finite number of models with probability one.

Chapter 4 by Griffiths gives a lucid treatment of the heteroskedasticity problem. The case of a known variance covariance term is treated first and generalized least squares is derived. Finite sample inference under normality as well as large sample inference without the normality assumption are summarized in the context of testing linear restrictions on the regression coefficients. In addition, inference for nonlinear restrictions on the regression coefficients is given and the consequences of heteroskedasticity on the least squares estimator are explained. Next, the case of the unknown variance covariance matrix is treated. In this case, several specifications of the form of heteroskedasticity are entertained and maximum likelihood estimation under normality is derived. Tests of linear restrictions on the regression coefficients are then formulated in terms of the ML estimates. Likelihood ratio, Wald and LM type tests of heteroskedasticity are given under normality of the disturbances. Other tests of heteroskedasticity as well as Monte Carlo experiments comparing these tests are cited. Adaptive estimators that assume no form of heteroskedasticity are briefly surveyed as well as several other miscellaneous extensions. Next, this chapter discusses Bayesian inference under heteroskedasticity. The joint posterior probability density function is specified assuming normality for the regression and uninformative priors

on heteroskedasticity. An algorithm for obtaining the marginal posterior probability density function is given.

Chapter 5 by Fiebig, surveys the most recent developments on seemingly unrelated regressions (SUR) including both applied and theoretical work on the specification, estimation and testing of SUR models. This chapter updates the survey by Srivastava and Dwivedi (1979) and the book by Srivastava and Giles (1987). A basic introduction of the SUR model introduced by Zellner (1962), is given along with extensions of the model to allow for more general stochastic specifications. These extensions are driven in part by diagnostic procedures as well as theoretical economic arguments presented for behavioral models of consumers and producers. Here the problem of testing linear restrictions which is important for testing demand systems or estimating say a cost function with share equations is studied. In addition, tests for the presence of contemporaneous correlation in SUR models as well as spatial autocorrelation are discussed. Next, SUR with missing observations and computational matters are reviewed. This leads naturally to a discussion of Bayesian methods for the SUR model and improved estimation methods for SUR which include several variants of the Stein-rule family and the hierarchical Bayes estimator. Finally, a brief discussion of misspecification, robust estimation issues, as well as extensions of the SUR model to time series modeling and count data Poisson regressions are given.

Chapter 6 by Mariano, considers the problem of estimation in the simultaneous equation model. Both limited as well as full information estimators are discussed. The inconsistency of ordinary least squares (OLS) is demonstrated. Limited information instrumental variable estimators are reviewed including two-stage least squares (2SLS), limited information instrumental variable efficient (LIVE), Theil's k-class, and limited information maximum likelihood (LIML). Full information methods including three-stage least squares (3SLS), full information instrumental variables efficient (FIVE) and full information maximum likelihood (FIML) are studied next. Large sample properties of these limited and full information estimators are summarized and conditions for their consistency and asymptotic efficiency are stated without proof. In addition, the finite sample properties of these estimators are reviewed and illustrated using the case of two included endogenous variables. Last, but not least, practical implications of these finite sample results are given. These are tied up to the recent literature on weak instruments.

Chapter 7 by Bekker and Wansbeek discusses the problem of identification in parametric models. Roughly speaking, a model is identified when meaningful estimates of its parameters can be obtained. Otherwise, the model is under-identified. In the latter case, different sets of parameter values agree well with the statistical evidence rendering scientific conclusions based on any estimates of this model void and dangerous. Bekker and Wansbeek define the basic concepts of observational equivalence of two parameter points and what is meant by local and global identification. They tie up the notion of identification to that of the existence of a consistent estimator, and provide a link between identification and the rank of the information matrix. The latter is made practically useful by presenting it in terms of the rank of a Jacobian matrix. Although the chapter

is limited to the problem of parametric identification based on sample information and exact restrictions on the parameters, extensions are discussed and the reader is referred to the book by Bekker, Merckens, and Wansbeek (1994) for further analysis.

Chapter 8 by Wansbeek and Meijer discusses the measurement error problem in econometrics. Many economic variables like permanent income, productivity of a worker, consumer satisfaction, financial health of a firm, etc. are latent variables that are only observed with error. This chapter studies the consequences of measurement error and latent variables in econometric models and possible solutions to these problems. First, the linear regression model with errors in variables is considered, the bias and inconsistency of OLS is demonstrated and the attenuation phenomenon is explained. Next, bounds on the parameters of the model are obtained by considering the reverse regression. Solutions to the errors in variables include restrictions on the parameters to identify the model and hence yield consistent estimators of the parameters. Alternatively, instrumental variables estimation procedures can be employed, or nonnormality of the errors may be exploited to obtain consistent estimates of these parameters. Repeated measurements like panel data on households, firms, regions, etc. can also allow the consistent estimation of the parameters of the model. The second part of this chapter gives an extensive discussion of latent variable models including factor analysis, the multiple indicators-multiple causes (MIMIC) model and a frequently used generalization of the MIMIC model known as the reduced rank regression model. In addition, general linear structural equation models estimated by LISREL are considered and maximum likelihood, generalized least squares, test statistics, and model fit are studied.

Chapter 9 by Wooldridge provides a comprehensive account of diagnostic testing in econometrics. First, Wooldridge explains how diagnostic testing differs from classical testing. The latter assumes a correctly specified parametric model and uses standard statistics to test restrictions on the parameters of this model, while the former tests the model for various misspecifications. This chapter considers diagnostic testing in cross section applications. It starts with diagnostic tests for the conditional mean in the linear regression model. Conditional mean diagnostics are computed using variable addition statistics or artificial regressions (see Chapter 1 by Davidson and MacKinnon). Tests for functional form are given as an example and it is shown that a key auxiliary assumption needed to obtain a usable limiting distribution for the usual  $nR^2$  (LM) test statistic is homoskedasticity. Without this assumption, the limiting distribution of the LM statistic is not  $\chi^2$  and the resulting test based on chi-squared critical values may be asymptotically undersized or oversized. This LM statistic is adjusted to allow for heteroskedasticity of unknown form under the null hypothesis. Next, testing for heteroskedasticity is considered. A joint test of the conditional mean and conditional variance is an example of an omnibus test. However, if this test rejects it is difficult to know where to look. A popular omnibus test is White's (1982) information matrix test. This test is explicitly a test for homoskedasticity, conditional symmetry and homokurtosis. If we reject, it may be for any of these reasons and it is not clear why one wants to toss out a model because of asymmetry, or

because its fourth and second moments do not satisfy the same relationship as that for a normal distribution. Extensions to nonlinear models are discussed next as well as diagnostic tests for completely specified parametric models like limited dependent variable models, probit, logit, tobit, and count data type models. The last section deals with diagnostic testing in time series models. In this case, one can no longer assume that the observations are independent of one another and the discussion of auxiliary assumptions under the null is more complicated. Wooldridge discusses different ways to make conditional mean diagnostics robust to serial correlation as well as heteroskedasticity. Testing for heteroskedasticity in time series contexts and omnibus tests on the errors in time series regressions, round up the chapter.

Chapter 10 by Pötscher and Prucha gives the basic elements of asymptotic theory. This chapter discusses the crucial concepts of convergence in probability and distribution, the convergence properties of transformed random variables, orders of magnitude of the limiting behavior of sequences of random variables, laws of large numbers both for independent and dependent processes, and central limit theorems. This is illustrated for regression analysis. Further readings are suggested including the recent book by Pötscher and Prucha (1997).

Chapter 11 by Hall provides a thorough treatment of the generalized method of moments (GMM) and its applications in econometrics. Hall explains that the main advantage of GMM for econometric theory is that it provides a general framework which encompasses many estimators of interest. For econometric applications, it provides a convenient method of estimating nonlinear dynamic models without complete knowledge of the distribution of the data. This chapter gives a basic definition of the GMM estimation principle and shows how it is predicated on the assumption that the population moment condition provides sufficient information to uniquely determine the unknown parameters of the correctly specified model. This need not be the case, and leads naturally to a discussion of the concepts of local and global identification. In case the model is overidentified, Hall shows how the estimation effects a decomposition on the population moment condition into identifying restrictions upon which the estimation is based, and overidentifying restrictions which are ignored in the estimation. This chapter describes how the estimated sample moment can be used to construct the overidentification restrictions test for the adequacy of the model specification, and derives the consistency and asymptotic distribution of the estimator. This chapter also characterizes the optimal choice of the weighting matrix and shows how the choice of the weight matrix impacts the GMM estimator via its asymptotic variance. MLE is shown to be a special case of GMM. However, MLE requires that we know the distribution of the data. GMM allows one to focus on the information used in the estimation and thereby determine the consequences of choosing the wrong distribution. Since the population moment condition is not known in practice, the researcher is faced with a large set of alternatives to choose from. Hall focuses on two extreme scenarios where the best and worst choices are made. The best choice is the population moment condition which leads to the estimator with the smallest asymptotic variance. The worst choice is when the population moment condition does not provide enough information to

identify the unknown parameters of our model. This leads to a discussion of nearly uninformative population moment conditions, their consequence and how they might be circumvented.

Chapter 12 by Hill and Adkins gives a lucid discussion of the collinearity problem in econometrics. They argue that collinearity takes three distinct forms. The first is where an explanatory variable exhibits little variability and therefore makes it difficult to estimate its effect in a linear regression model. The second is where two explanatory variables exhibit a large correlation leaving little independent variation to estimate their separate effects. The third is where there may be one or more nearly exact linear relationships among the explanatory variables, hence obscuring the effects of each independent variable on the dependent variable. Hill and Adkins examine the damage that multicollinearity does to estimation and review collinearity diagnostics such as the variance decomposition of Belsley, Kuh, and Welsch (1980), the variance-inflation factor and the determinant of  $X'X$ , the sum of squares and cross-product of the regressor matrix  $X$ . Once collinearity is detected, this chapter discusses whether collinearity is harmful by using Belsley's (1982) test for adequate signal to noise ratio in the regression model and data. Next, remedies to harmful collinearity are reviewed. Here the reader is warned that there are only two safe paths. The first is obtaining more and better data which is usually not an option for practitioners. The second is imposing additional restrictions from economic theory or previous empirical research. Hill and Adkins emphasize that although this is a feasible option, only good nonsample information should be used and it is never truly known whether the information introduced is good enough. Methods of introducing exact and inexact nonsample information including restricted least squares, Stein-rule estimators, inequality restricted least squares, Bayesian methods, the mixed estimation procedure of Theil and Goldberger (1961) and the maximum entropy procedure of Golan, Judge, and Miller (1996) are reviewed. In addition, two estimation methods designed specifically for collinear data are discussed if only to warn the readers about their use. These are ridge regression and principal components regression. Finally, this chapter extends the collinearity analysis to nonlinear models.

Chapter 13 by Pesaran and Weeks gives an overview of the problem of non-nested hypothesis testing in econometrics. This problem arises naturally when rival economic theories are used to explain the same phenomenon. For example, competing theories of inflation may suggest two different sets of regressors neither of which is a special case of the other. Pesaran and Weeks define non-nested models as belonging to separate families of distributions in the sense that none of the individual models may be obtained from the remaining either by imposition of parameter restrictions or through a limiting process. This chapter discusses the problem of model selection and how it relates to non-nested hypothesis testing. By utilizing the linear regression model as a convenient framework, Pesaran and Weeks examine three broad approaches to non-nested hypotheses testing: (i) the modified (centered) log-likelihood ratio procedure also known as the Cox test; (ii) the comprehensive models approach, whereby the non-nested models are tested against an artificially constructed general model that includes the

non-nested models as special cases; and (iii) the encompassing approach, where the ability of one model to explain particular features of an alternative model is tested directly. This chapter also focuses on the Kullback-Leibler divergence measure which has played a pivotal role in the development of a number of non-nested test statistics. In addition, the Vuong (1989) approach to model selection, viewed as a hypothesis testing problem is also discussed. Finally, practical problems involved in the implementation of the Cox procedure are considered. This involves finding an estimate of the Kullback-Leibler measure of closeness of the alternative to the null hypothesis which is not easy to compute. Two methods are discussed to circumvent this problem. The first examines the simulation approach and the second examines the parametric bootstrap approach.

Chapter 14 by Anselin provides an excellent review of spatial econometrics. These methods deal with the incorporation of spatial interaction and spatial structure into regression analysis. The field has seen a recent and rapid growth spurred both by theoretical concerns as well as the need to apply econometric models to emerging large geocoded databases. This chapter outlines the basic terminology and discusses in some detail the specification of spatial effects including the incorporation of spatial dependence in panel data models and models with qualitative variables. The estimation of spatial regression models including maximum likelihood estimation, spatial 2SLS, method of moments estimators and a number of other approaches are considered. In addition, specification tests for spatial effects as well as implementation issues are discussed.

Chapter 15 by Cameron and Trivedi gives a brief review of count data regressions. These are regressions that involve a dependent variable that is a count, such as the number of births in models of fertility, number of accidents in studies of airline safety, hospital or doctor visits in health demand studies, number of trips in models of recreational demand, number of patents in research and development studies or number of bids in auctions. In these examples, the sample is concentrated on a few discrete values like 0, 1 and 2. The data is skewed to the left and the data is intrinsically heteroskedastic with its variance increasing with the mean. Two methods of dealing with these models are considered. The first is a fully parametric approach which completely specifies the distribution of the data and restricts the dependent variable to nonnegative integer values. This includes the Poisson regression model which is studied in detail in this chapter including its extensions to truncated and censored data. Limitations of the Poisson model, notably the excess zeros problem and the overdispersion problem are explained and other parametric models, superior to the Poisson are presented. These include continuous mixture models, finite mixture models, modified count models and discrete choice models. The second method of dealing with count data is a partial parametric method which focuses on modeling the data via the conditional mean and variance. This includes quasi-maximum likelihood estimation, least squares estimation and semiparametric models. Extensions to other types of data notably time series, multivariate, and panel data are discussed and the chapter concludes with some practical recommendations. For further readings and diagnostic procedures, the reader is referred to the recent econometric society monograph on count data models by Cameron and Trivedi (1998).

Chapter 16 by Hsiao gives a selected survey of panel data models. First, the benefits from using panels are discussed. This includes more degrees of freedom, controlling for omitted variable bias, reducing the problem of multicollinearity and improving the accuracy of parameter estimates and predictions. A general encompassing linear panel data model is provided which includes as special cases the error components model, the random coefficients model and the mixed fixed and random coefficients model. These models assume that some variables are subject to stochastic constraints while others are subject to deterministic constraints. In practice, there is little knowledge about which variables are subject to stochastic constraints and which variables are subject to deterministic constraints. Hsiao recommends the Bayesian predictive density ratio method for selecting between two alternative formulations of the model. Dynamic panel data models are studied next and the importance of the initial observation with regards to the consistency and efficiency of the estimators is emphasized. Generalized method of moments estimators are proposed and the problem of too many orthogonality conditions is discussed. Hsiao suggests a transformed maximum likelihood estimator that is asymptotically more efficient than GMM. Hsiao also reports that the mean group estimator suggested by Pesaran and Smith (1995) does not perform well in finite samples. Alternatively, a hierarchical Bayesian approach performs well when  $T$  is small and the initial value is assumed to be a fixed constant. Next, the existence of individual specific effects in nonlinear models is discussed and the conditional MLE approach of Chamberlain (1980) is given. The problem becomes more complicated if lagged dependent variables are present.  $T \geq 4$  is needed for the identification of a logit model and this conditional method will not work with the presence of exogenous variables. In this case, a consistent and asymptotically normal estimator proposed by Honoré and Kyriazidou (1997) is suggested. An alternative semiparametric approach to estimating nonlinear panel models is the maximum score estimator proposed by Manski (1975). This applies some data transformation to eliminate the individual effects if the nonlinear model is of the form of a single index model with the index possessing a linear structure. This estimator is consistent but not root  $n$  consistent. A third approach proposed by Lancaster (1998) finds an orthogonal reparametrization of the fixed effects such that the new fixed effects are independent of the structural parameters in the information matrix sense. Hsiao discusses the limitations of all three methods, emphasizing that none of these approaches can claim general applicability and that the consistency of nonlinear panel data estimators must be established on a case by case basis. Finally, Hsiao treats missing observations in panels. If individuals are missing randomly, most estimators in the balanced panel case can be easily generalized to the unbalanced case. With sample selection, Hsiao emphasizes the dependence of the MLE and Heckman's (1979) two-step estimators on the exact specification of the joint error distribution. If this distribution is misspecified, then these estimators are inconsistent. Alternative semiparametric methods are discussed based on the work of Ahn and Powell (1993), Kyriazidou (1997), and Honoré and Kyriazidou (1998).

Chapter 17 by Maddala and Flores-Lagunes gives an update of the econometrics of qualitative response models. First, a brief introduction to the basic material on

the estimation of binary and multinomial logit and probit models is given and the reader is referred to Maddala (1983) for details. Next, this chapter reviews specification tests in qualitative response models and the reader is referred to the recent review by Maddala (1995). Panel data with qualitative variables and semiparametric estimation methods for qualitative response models are reviewed including Manski's maximum score, quasi-maximum likelihood, generalized maximum likelihood and the semi-nonparametric estimator. Maddala and Flores-Lagunes comment on the empirical usefulness and drawbacks of the different methods. Finally, simulation methods in qualitative response models are reviewed. The estimation methods discussed are the method of simulated moments, the method of simulated likelihood and the method of simulated scores. Some examples are given comparing these simulation methods.

Chapter 18 by Lee gives an extensive discussion of the problem of self-selection in econometrics. When the sample observed is distorted and is not representative of the population under study, sample selection bias occurs. This may be due to the way the sample was collected or it may be due to the self-selection decisions by the agents being studied. This sample may not represent the true population no matter how large. This chapter discusses some of the conventional sample selection models and counterfactual outcomes. A major part of this chapter concentrates on the specification, estimation, and testing of parametric models of sample selection. This includes Heckman's two-stage estimation procedure as well as maximum likelihood methods, polychotomous choice sample selection models, simulation estimation methods, and the estimation of simultaneous equation sample selection models. Another major part of this chapter focuses on semiparametric and nonparametric approaches. This includes semiparametric two-stage estimation, semiparametric efficiency bound and semiparametric maximum likelihood estimation. In addition, semiparametric instrumental variable estimation and conditional moments restrictions are reviewed, as well as sample selection models with a tobit selection rule. The chapter concludes with the identification and estimation of counterfactual outcomes.

Chapter 19 by Swamy and Tavlas describes the purpose, estimation, and use of random coefficient models. Swamy and Tavlas distinguish between first generation random coefficient models that sought to relax the constant coefficient assumption typically made by researchers in the classical tradition and second generation random coefficient models that relax the assumptions made regarding functional forms, excluded variables, and absence of measurement error. The authors argue that the latter are useful approximations to reality because they provide a reasonable approximation to the underlying "true" economic relationship. Several model validation criteria are provided. Throughout, a demand for money model is used as a backdrop to explain random coefficient models and an empirical application to United Kingdom data is given.

Chapter 20 by Ullah provides a systematic and unified treatment of estimation and test of hypotheses for nonparametric and semiparametric regression models. Parametric approaches to specifying functional form in econometrics may lead to misspecification. Nonparametric and semiparametric approaches provide alternative estimation procedures that are more robust to functional form

misspecification. This chapter studies the developments in the area of kernel estimation in econometrics. Nonparametric estimates of conditional means as well as higher order moments and function derivatives are reviewed. In addition, some new goodness of fit procedures for nonparametric regressions are presented and their application to determining the window width and variable selection are discussed. Next, a combination of parametric and nonparametric regressions that takes into account the tradeoff between good fit (less bias) and smoothness (low variance) is suggested to improve (in mean-squared error sense) the drawbacks of each approach used separately. Additive nonparametric regressions and semiparametric models are given as possible solutions to the curse of dimensionality. The chapter concludes with hypothesis testing in nonparametric and semiparametric models.

Chapter 21 by Gourieroux and Jasiak gives a review of duration models. These models have been used in labor economics to study the duration of individual unemployment spells and in health economics to study the length of hospital stays to mention just two examples. Gourieroux and Jasiak discuss the main duration distribution families including the exponential, gamma, Weibull, and lognormal distributions. They explain what is meant by survivor functions, hazard functions, and duration dependence. Maximum likelihood estimation for the exponential duration model without heterogeneity as well as the gamma distributed heterogeneity model are given. The latter leads to the Pareto regression model. Next, the effect of unobservable heterogeneity and its relationship with negative duration dependence as well as the problem of partial observability of duration variables due to truncation or censoring effects are considered. Semiparametric specifications of duration models are also studied. These distinguish a parametric scoring function and an unconstrained baseline distribution. Accelerated and proportional hazard models are introduced and the estimation methods for the finite dimensional functional parameters are given. Finally, dynamic models for the analysis of time series durations are given. These are especially useful for applications to financial transactions data. Some recent developments in this field are covered including the Autoregressive Conditional Duration (ACD) model and the Stochastic Volatility Duration (SVD) model.

Chapter 22 by Geweke, Houser, and Keane provides a detailed illustration of how to implement simulation methods to dynamic discrete choice models where one has available a panel of multinomial choice histories and partially observed payoffs. The advantages of this procedure is that it does not require the econometrician to solve the agents' dynamic optimization problem, or to make strong assumptions about the way individuals form expectations. The chapter focuses exclusively on simulation based Bayesian techniques. Monte Carlo results demonstrate that this method works well in relatively large state-space models with only partially-observed payoffs, where very high dimensional integrations are required.

Chapter 23 by Dufour and Khalaf reviews Monte Carlo test methods in econometrics. Dufour and Khalaf demonstrate that this technique can provide exact randomized tests for any statistic whose finite sample distribution may be intractable but can be simulated. They illustrate this technique using several

specification tests in linear regressions including tests for normality, tests for independence, heteroskedasticity, ARCH, and GARCH. Also, nonlinear hypotheses in univariate and SUR models, tests on structural parameters in instrumental variable regressions, tests on long-run multipliers in dynamic models, long-run identification constraints in VAR (vector autoregression) models and confidence intervals on ratios of coefficients in discrete choice models. Dufour and Khalaf demonstrate, using several econometric examples, that standard testing procedures that rely on asymptotic theory can produce questionable  $p$ -values and confidence intervals. They recommend Monte Carlo test methods over these standard testing procedures because the former procedures produce a valid inference in small samples.

Chapter 24 by Koop and Steel uses the stochastic frontier models to illustrate Bayesian methods in econometrics. This is contrasted with the classical econometric stochastic frontier methods and the strengths and weaknesses of both approaches are reviewed. In particular, the stochastic frontier model with cross-sectional data is introduced with a simple log-linear model (e.g., Cobb-Douglas or translog) and a simple Gibbs sampler is used to carry out Bayesian inference. This is extended to nonlinear production frontiers (e.g., constant elasticity of substitution or the asymptotically ideal model) where more complicated posterior simulation methods are necessary. Next, the stochastic frontier model with panel data is considered and the Bayesian fixed effects and random effects models are contrasted with their classical alternatives. Koop and Steel show that the Bayesian fixed effects model imposes strong and possibly unreasonable prior assumptions. In fact, only relative efficiencies and not absolute efficiencies can be computed by this model and it is shown to be quite sensitive to prior assumptions. In contrast, the Bayesian random effects model makes explicit distributional assumptions for the inefficiencies and allows robust inference on the absolute efficiencies.

Chapter 25 by Maasoumi summarizes some of the technical and conceptual advances in testing multivariate linear and nonlinear inequality hypotheses in econometrics. This is done in the context of substantive empirical settings in economics in which either the null, or the alternative, or both hypotheses define more limited domains than the two-sided alternatives typically tested in classical hypotheses testing. The desired goal is increased power. The impediments are a lack of familiarity with implementation procedures, and characterization problems of distributions under some composite hypotheses. Several empirically important cases are identified in which practical “one-sided” tests can be conducted by either the mixed  $\chi^2$  distribution, or the union intersection mechanisms based on the Gaussian variate, or the popular resampling/simulation techniques. Point optimal testing and its derivatives find a natural medium here whenever unique characterization of the null distribution for the “least favorable” cases is not possible. Applications in the parametric, semiparametric and nonparametric testing area are cited.

Chapter 26 by Granger gives a brief introduction to the problem of spurious regressions in econometrics. This problem is traced historically from its origins (see Yule 1926). A definition of spurious correlation is given and the problem is

illustrated with simulations. The theoretical findings on spurious regressions undertaken by Phillips (1986) are summarized and some extensions of this work are cited. This chapter emphasizes that this spurious regression problem can occur with stationary series also. Applied econometricians have been worrying about spurious regressions only with nonstationary series. Usually, testing for unit roots before entering a regression. Granger warns that one should also be worried about this problem with stationary series and recommends care in the proper specification of one's model using lags of all variables.

Chapter 27 by Stock provides an introduction to the main methods used for forecasting economic time series. Throughout the chapter, Stock emphasizes the responsible production and interpretation of economic forecasts which have entered into many aspects of economic life. This requires a clear understanding of the associated econometric tools, their limitations, and an awareness of common pitfalls in their application. Stock provides a theoretical framework for considering some of the tradeoffs in the construction of economic forecasts. In particular, he decomposes the forecast error into three terms and shows how two sources of this error entail a tradeoff. This tradeoff is between model approximation error and model estimation error. This leads naturally to considering information criteria that select among competing models. Stock also discusses prediction intervals and forecast comparison methods. Next, this chapter provides a glimpse at some of the relevant empirical features of five monthly macroeconomic time series data for the USA. These include the rate of inflation, output growth, the unemployment rate, a short-term interest rate and total real manufacturing and trade inventories. In addition, this chapter provides an overview of univariate forecasts which are made solely using past observations on the series. For linear models, simple forecasting methods like the exponential smoothing method and the mainstay of univariate forecasting, i.e., autoregressive moving average models are discussed. Two nonlinear models are considered, smooth transition autoregression and artificial neural networks. These models provide parametric and nonparametric approaches to nonlinear forecasting. Finally, this chapter tackles multivariate forecasting where forecasts are made using historical information on multiple time series. In particular, it considers vector autoregressions, (see also Chapter 32 by Lütkepohl), and forecasting with leading economic indicators.

Chapter 28 by Spanos examines time series and dynamic models in econometrics. After providing a brief historical perspective and a probabilistic framework, the chapter reviews AR models, MA models and ARMA models including VARs. Then it places these time series models in the perspective of econometric linear regression models. Throughout, the emphasis is on the statistical adequacy of the various models, and ways by which that adequacy may be evaluated.

Chapter 29 by Bierens explains the two most frequently applied unit root tests, namely the Augmented Dickey–Fuller tests and the Phillips–Perron tests. Bierens emphasizes three reasons why it is important to distinguish stationary processes from unit root processes. The first points to the fact that regressions involving unit roots may give spurious results (see Chapter 26 by Granger). The second reason is that for two or more unit root processes, there may exist linear

combinations which are stationary, and these linear combinations may be interpreted as long-run relationships (see Cointegration, Chapter 30, by Dolado, Gonzalo, and Marmol). The third reason emphasizes that tests for parameter restrictions in autoregressions involving unit root processes have in general different null distributions than in the case of stationary processes. Therefore, naive application of classical inference may give misleading results. Bierens considers in details the Gaussian AR(1) case without an intercept; the Gaussian AR(1) case with an intercept under the alternative of stationarity; the general AR( $p$ ) process with a unit root and the Augmented Dickey–Fuller test; the general ARIMA processes and the Phillips–Perron test; and the unit root with drift versus trend stationarity.

Chapter 30 by Dolado, Gonzalo, and Marmol discusses how the important concept of cointegration in econometrics bridged the gap between economic theorists who had much to say about equilibrium but relatively little to say about dynamics and the econometricians whose models concentrated on the short-run dynamics disregarding the long-run equilibrium. “In addition to allowing the data to determine short-run dynamics, cointegration suggests that models can be significantly improved by including long-run equilibrium conditions suggested by economic theory.” This chapter discusses the implications of cointegration and gives the basic estimation and testing procedures in a single equation framework, when variables have a single unit root. These include the Engle and Granger (1987) two-step procedure and the fully modified ordinary least squares procedure proposed by Phillips and Hansen (1990). Next, the analysis is extended to the multivariate setting where the estimation and testing of more general system based approaches to cointegration are considered. These include the Johansen (1995) maximum likelihood estimation procedure based on the reduced rank regression method, and the Stock and Watson (1988) methodology of testing for the dimension of “common trends.” The latter is based on the dual relationship between the number of cointegrating vectors and the number of common trends in the system. Several extensions of this literature are discussed including (i) higher order cointegrated systems; (ii) fractionally cointegrated systems; (iii) nearly cointegrated systems; (iv) nonlinear error correction models; and (v) structural breaks in cointegrated systems.

Modelling seasonality has progressed from the traditional view that seasonal patterns are a nuisance which need to be removed to the current view, see Ghysels (1994), that they are an informative feature of economic time series that should be modeled explicitly. Chapter 31 by Ghysels, Osborn, and Rodrigues discusses the properties of stochastic seasonal nonstationary processes. In particular, they consider the characteristics of a simple seasonal random walk model and then generalize the discussion to seasonally integrated autoregressive moving average processes. In addition, they discuss the implications of misdiagnosing nonstationary stochastic seasonality as deterministic. Finally, the asymptotic properties of several seasonal unit root tests are studied and the results are generalized to the case of a near seasonally integrated framework.

Chapter 32 by Lütkepohl gives a concise review of the vector autoregression (VAR) literature in econometrics. The poor performance of simultaneous equations

macroeconometric models resulted in a critical assessment by Sims (1980) who advocated the use of VAR models as alternatives. This chapter gives a brief introduction to these models, their specification and estimation. Special attention is given to cointegrated systems. Model specification and model checks studied include testing for the model order and exclusion restrictions, determining the autoregressive order by model selection criteria, specifying the cointegrating rank as well as various model checks that are based on the residuals of the final model. Among the uses of VAR models discussed is forecasting and economic analysis. The concept of Granger causality is introduced which is based on forecast performance, impulse responses are considered as instruments for analyzing causal relationships between variables. Finally, forecast error variance decompositions and policy analysis are discussed.

CHAPTER ONE

# Artificial Regressions

Russell Davidson and James G. MacKinnon

## 1 INTRODUCTION

All popular nonlinear estimation methods, including nonlinear least squares (NLS), maximum likelihood (ML), and the generalized method of moments (GMM), yield estimators which are asymptotically linear. Provided the sample size is large enough, the behavior of these nonlinear estimators in the neighborhood of the true parameter values closely resembles the behavior of the ordinary least squares (OLS) estimator. A particularly illuminating way to see the relationship between any nonlinear estimation method and OLS is to formulate the *artificial regression* that corresponds to the nonlinear estimator.

An artificial regression is a linear regression in which the regressand and regressors are constructed as functions of the data and parameters of the nonlinear model that is really of interest. In addition to helping us understand the asymptotic properties of nonlinear estimators, artificial regressions are often extremely useful as calculating devices. Among other things, they can be used to estimate covariance matrices, as key ingredients of nonlinear optimization methods, to compute one-step efficient estimators, and to calculate test statistics.

In the next section, we discuss the defining properties of an artificial regression. In the subsequent section, we introduce the Gauss–Newton regression, which is probably the most popular artificial regression. Then, in Section 4, we illustrate a number of uses of artificial regressions, using the Gauss–Newton regression as an example. In Section 5, we develop the most important use of artificial regressions, namely, hypothesis testing. We go beyond the Gauss–Newton regression in Sections 6 and 7, in which we introduce two quite generally applicable artificial regressions, one for models estimated by maximum likelihood, and one for models estimated by the generalized method of moments. Section 8 shows how artificial regressions may be modified to take account of the presence of heteroskedasticity of unknown form. Then, in Sections 9 and 10, we discuss double-length regressions and artificial regressions for binary response models, respectively.

## 2 THE CONCEPT OF AN ARTIFICIAL REGRESSION

Consider a fully parametric, nonlinear model that is characterized by a parameter vector  $\theta$  which belongs to a parameter space  $\Theta \subseteq \mathbb{R}^k$  and which can be estimated by minimizing a criterion function  $Q(\theta)$  using  $n$  observations. In the case of a nonlinear regression model estimated by nonlinear least squares,  $Q(\theta)$  would be one half the sum of squared residuals, and in the case of a model estimated by maximum likelihood,  $Q(\theta)$  would be minus the loglikelihood function.

If an artificial regression exists for such a model, it always involves two things: a regressand,  $r(\theta)$ , and a matrix of regressors,  $R(\theta)$ . The number of regressors for the artificial regression is equal to  $k$ , the number of parameters. The number of “observations” for the artificial regression is often equal to  $n$ , but it may also be equal to a small integer, such as 2 or 3, times  $n$ . We can write a generic artificial regression as

$$r(\theta) = R(\theta)b + \text{residuals}, \quad (1.1)$$

where  $b$  is a  $k$ -vector of coefficients. “Residuals” is used here as a neutral term to avoid any implication that (1) is a statistical model. The regressand and regressors in (1) can be evaluated at any point  $\theta \in \Theta$ , and the properties of the artificial regression will depend on the point at which they are evaluated. In many cases, we will want to evaluate (1) at a vector of estimates  $\hat{\theta}$  that is root- $n$  consistent. This means that, if the true parameter vector is  $\theta_0 \in \Theta$ , then  $\hat{\theta}$  approaches  $\theta_0$  at a rate proportional to  $n^{-1/2}$ . One such vector that is of particular interest is  $\hat{\theta}$ , the vector of estimates which minimizes the criterion function  $Q(\theta)$ .

For (1.1) to constitute an artificial regression, the vector  $r(\theta)$  and the matrix  $R(\theta)$  must satisfy certain defining properties. These may be stated in a variety of ways, which depend on the class of models to which the artificial regression is intended to apply. For the purposes of this chapter, we will say that (1.1) is an artificial regression if it satisfies the following three conditions:

1. The estimator  $\hat{\theta}$  is defined, uniquely in a neighborhood in  $\Theta$ , by the  $k$  equations  $R^\top(\hat{\theta})r(\hat{\theta}) = 0$ ;
2. for any root- $n$  consistent  $\hat{\theta}$ , a consistent estimate of  $\text{var}(\text{plim } n^{1/2}(\hat{\theta} - \theta_0))$  is given by the inverse of  $n^{-1}R^\top(\hat{\theta})R(\hat{\theta})$ . Formally,

$$\text{var}\left(\text{plim}_{n \rightarrow \infty} n^{1/2}(\hat{\theta} - \theta_0)\right) = \text{plim}_{n \rightarrow \infty} (n^{-1}R^\top(\hat{\theta})R(\hat{\theta}))^{-1};$$

3. if  $\tilde{b}$  denotes the vector of estimates from the artificial regression (1.1) with regressand and regressors evaluated at  $\hat{\theta}$ , then

$$\hat{\theta} + \tilde{b} = \hat{\theta} + o_p(n^{-1/2}).$$

Many artificial regressions actually satisfy a stronger version of condition (1):

$$g(\theta) = -R^\top(\theta)r(\theta), \quad (1.1')$$

where  $g(\theta)$  denotes the gradient of the criterion function  $Q(\theta)$ . Clearly, condition (1.1') implies condition (1), but not vice versa. The minus sign in (1.1') is due to the arbitrary choice that the estimator is defined by minimizing  $Q(\theta)$  rather than maximizing it.

Condition (2) has been written in a particularly simple form, and some non-standard artificial regressions do not actually satisfy it. However, as we will see, this does not prevent them from having essentially the same properties as artificial regressions that do satisfy it.

Condition (3), which is perhaps the most interesting of the three conditions, will be referred to as the *one-step property*. It says that, if we take one step from an initial consistent estimator  $\hat{\theta}$ , where the step is given by the coefficients  $\hat{b}$  from the artificial regression, we will obtain an estimator that is asymptotically equivalent to  $\hat{\theta}$ .

The implications of these three conditions will become clearer when we study specific artificial regressions in the remainder of this chapter. These conditions differ substantially from the conditions used to define an artificial regression in Davidson and MacKinnon (1990), because that paper was concerned solely with artificial regressions for models estimated by maximum likelihood.

### 3 THE GAUSS-NEWTON REGRESSION

Associated with every nonlinear regression model is a somewhat nonstandard artificial regression which is probably more widely used than any other. Consider the univariate, nonlinear regression model

$$y_t = x_t(\beta) + u_t, \quad u_t \sim \text{iid}(0, \sigma^2), \quad t = 1, \dots, n, \quad (1.2)$$

where  $y_t$  is the  $t$ th observation on the dependent variable, and  $\beta$  is a  $k$ -vector of parameters to be estimated. The scalar function  $x_t(\beta)$  is a nonlinear regression function. It determines the mean value of  $y_t$  as a function of unknown parameters  $\beta$  and, usually, of explanatory variables, which may include lagged dependent variables. The explanatory variables are not shown explicitly in (1.2), but the  $t$  subscript on  $x_t(\beta)$  reminds us that they are present. The model (1.2) may also be written as

$$\mathbf{y} = \mathbf{x}(\beta) + \mathbf{u}, \quad \mathbf{u} \sim \text{iid}(0, \sigma^2 \mathbf{I}), \quad (1.3)$$

where  $\mathbf{y}$  is an  $n$ -vector with typical element  $y_t$ , and  $\mathbf{x}(\beta)$  is an  $n$ -vector of which the  $t$ th element is  $x_t(\beta)$ .

The nonlinear least squares (NLS) estimator  $\hat{\beta}$  for model (1.3) minimizes the sum of squared residuals. It is convenient to use this sum divided by 2. Thus we define

$$Q(\beta) = \frac{1}{2}(\mathbf{y} - \mathbf{x}(\beta))^\top(\mathbf{y} - \mathbf{x}(\beta)). \quad (1.4)$$

The Gauss–Newton regression can be derived as an approximation to *Newton's Method* for the minimization of  $Q(\beta)$ . In this case, Newton's Method consists of the following iterative procedure. One starts from some suitably chosen starting value,  $\beta_{(0)}$ . At step  $m$  of the procedure,  $\beta_{(m)}$  is updated by the formula

$$\beta_{(m+1)} = \beta_{(m)} - H_{(m)}^{-1} g_{(m)},$$

where the  $k \times 1$  vector  $g_{(m)}$  and the  $k \times k$  matrix  $H_{(m)}$  are, respectively, the gradient and the Hessian of  $Q(\beta)$  with respect to  $\beta$ , evaluated at  $\beta_{(m)}$ . For general  $\beta$ , we have

$$g(\beta) = -X^\top(\beta)(y - x(\beta)),$$

where the matrix  $X(\beta)$  is an  $n \times k$  matrix with  $t$ th element the derivative of  $x_t(\beta)$  with respect to  $\beta_i$ , the  $i$ th component of  $\beta$ . A typical element of the Hessian  $H(\beta)$  is

$$H_{ij}(\beta) = -\sum_{t=1}^n \left( (y_t - x_t(\beta)) \frac{\partial X_{ti}(\beta)}{\partial \beta_j} - X_{ti}(\beta) X_{tj}(\beta) \right), \quad i, j = 1, \dots, k. \quad (1.5)$$

The Gauss–Newton procedure is one of the set of so-called **quasi-Newton** procedures, in which the exact Hessian is replaced by an approximation. Here, only the second term in (1.5) is used, so that the  $H(\beta)$  of Newton's method is replaced by the matrix  $X^\top(\beta)X(\beta)$ . Thus the Gauss–Newton updating formula is

$$\beta_{(m+1)} = \beta_{(m)} + (X_{(m)}^\top X_{(m)})^{-1} X_{(m)}^\top (y - x_{(m)}), \quad (1.6)$$

where we write  $X_{(m)} = X(\beta_{(m)})$  and  $x_{(m)} = x(\beta_{(m)})$ . The updating term on the right-hand side of (1.6) is the set of OLS parameter estimates from the *Gauss–Newton regression*, or *GNR*,

$$y - x(\beta) = X(\beta)b + \text{residuals}, \quad (1.7)$$

where the variables  $r(\beta) \equiv y - x(\beta)$  and  $R(\beta) \equiv X(\beta)$  are evaluated at  $\beta_{(m)}$ . Notice that there is no regressor in (1.7) corresponding to the parameter  $\sigma^2$ , because the criterion function  $Q(\beta)$  does not depend on  $\sigma^2$ . This is one of the features of the GNR that makes it a nonstandard artificial regression.

The GNR is clearly a linearization of the nonlinear regression model (1.3) around the point  $\beta$ . In the special case in which the original model is linear,  $x(\beta) = X\beta$ , where  $X$  is the matrix of independent variables. Since  $X(\beta)$  is equal to  $X$  for all  $\beta$  in this special case, the GNR will simply be a regression of the vector  $y - X\beta$  on the matrix  $X$ .

An example is provided by the nonlinear regression model

$$y_t = \beta_1 Z_{t1}^{\beta_2} Z_{t2}^{1-\beta_2} + u_t, \quad u_t \sim \text{iid}(0, \sigma^2), \quad (1.8)$$

where  $Z_{t1}$  and  $Z_{t2}$  are independent variables. The regression function here is nonlinear and has the form of a Cobb–Douglas production function. In many cases, of course, it would be reasonable to assume that the error term is multiplicative, and it would then be possible to take logarithms of both sides and use ordinary least squares. But if we wish to estimate (1.8) as it stands, we must use nonlinear least squares. The GNR that corresponds to (1.8) is

$$y_t - \beta_1 Z_{t1}^{\beta_2} Z_{t2}^{1-\beta_2} = b_1 Z_{t1}^{\beta_2} Z_{t2}^{1-\beta_2} + b_2 \beta_1 Z_{t2} \left( \frac{Z_{t1}}{Z_{t2}} \right)^{\beta_2} \log \left( \frac{Z_{t1}}{Z_{t2}} \right) + \text{residual}.$$

The regressand is  $y_t$  minus the regression function, the first regressor is the derivative of the regression function with respect to  $\beta_1$ , and the second regressor is the derivative of the regression function with respect to  $\beta_2$ .

Now consider the defining conditions of an artificial regression. We have

$$R^\top(\theta)r(\theta) = X^\top(\beta)(y - x(\beta)), \quad (1.9)$$

which is just minus the gradient of  $Q(\beta)$ . Thus condition (1.1') is satisfied.

Next, consider condition (3). Let  $\hat{\beta}$  denote a vector of initial estimates, which are assumed to be root- $n$  consistent. The GNR (1.7) evaluated at these estimates is

$$y - \hat{x} = \hat{X}\hat{b} + \text{residuals},$$

where  $\hat{x} \equiv x(\hat{\beta})$  and  $\hat{X} \equiv X(\hat{\beta})$ . The estimate of  $b$  from this regression is

$$\hat{b} = (\hat{X}^\top \hat{X})^{-1} \hat{X}^\top (y - \hat{x}). \quad (1.10)$$

The *one-step efficient estimator* is then defined to be

$$\hat{\beta} \equiv \hat{\beta} + \hat{b}. \quad (1.11)$$

By Taylor expanding the expression  $n^{-1/2} \hat{X}^\top (y - \hat{x})$  around  $\beta = \beta_0$ , where  $\beta_0$  is the true parameter vector, and using standard asymptotic arguments, it can be shown that, to leading order,

$$n^{-1/2} \hat{X}^\top (y - \hat{x}) = n^{-1/2} X_0^\top u - n^{-1} X_0^\top X_0 n^{1/2} (\hat{\beta} - \beta_0),$$

where  $X_0 \equiv X(\beta_0)$ . This relation can be solved to yield

$$n^{1/2} (\hat{\beta} - \beta_0) = (n^{-1} X_0^\top X_0)^{-1} (n^{-1/2} X_0^\top u - n^{-1/2} \hat{X}^\top (y - \hat{x})). \quad (1.12)$$

Now it is a standard result that, asymptotically,

$$n^{1/2} (\hat{\beta} - \beta_0) = (n^{-1} X_0^\top X_0)^{-1} (n^{-1/2} X_0^\top u); \quad (1.13)$$

see, for example, Davidson and MacKinnon (1993, section 5.4). By (1.10), the second term on the right-hand side of (1.12) is asymptotically equivalent to  $-n^{1/2}\hat{b}$ . Thus (1.12) implies that

$$n^{1/2}(\hat{\beta} - \beta_0) = n^{1/2}(\hat{\beta} - \beta_0) - n^{1/2}\hat{b}.$$

Rearranging this and using the definition (1.11), we see that, to leading order asymptotically,

$$n^{1/2}(\hat{\beta} - \beta) = n^{1/2}(\hat{\beta} + \hat{b} - \beta_0) = n^{1/2}(\hat{\beta} - \beta_0).$$

In other words, after both are centered and multiplied by  $n^{1/2}$ , the one-step estimator  $\hat{\beta}$  and the NLS estimator  $\hat{\beta}$  tend to the same random variable asymptotically. This is just another way of writing condition (3) for model (1.3).

Finally, consider condition (2). Since  $X(\beta)$  plays the role of  $R(\theta)$ , we see that

$$\frac{1}{n}R^\top(\theta)R(\theta) = \frac{1}{n}X^\top(\beta)X(\beta). \quad (1.14)$$

If the right-hand side of (1.14) is evaluated at any root- $n$  consistent estimator  $\hat{\beta}$ , it must tend to the same probability limit as  $n^{-1}X_0^\top X_0$ . It is a standard result, following straightforwardly from (1.13), that, if  $\hat{\beta}$  denotes the NLS estimator for the model (1.3), then

$$\lim_{n \rightarrow \infty} \text{var}(n^{1/2}(\hat{\beta} - \beta_0)) = \sigma_0^2 \text{plim}_{n \rightarrow \infty} (n^{-1}X_0^\top X_0)^{-1}, \quad (1.15)$$

where  $\sigma_0^2$  is the true variance of the error terms; see, for example, Davidson and MacKinnon (1993, ch. 5). Thus the GNR would satisfy condition (2) except that there is a factor of  $\sigma_0^2$  missing. However, this factor is automatically supplied by the regression package. The estimated covariance matrix will be

$$\widehat{\text{var}}(\hat{b}) = s^2(\hat{X}^\top \hat{X})^{-1}, \quad (1.16)$$

where  $s^2 = \text{SSR}/(n - k)$  is the estimate of  $\sigma^2$  from the artificial regression. It is not hard to show that  $s^2$  estimates  $\sigma_0^2$  consistently, and so it is clear from (1.15) that (1.16) provides a reasonable way to estimate the covariance matrix of  $\hat{\beta}$ .

It is easy to modify the GNR so that it actually satisfies condition (2). We just need to divide both the regressand and the regressors by  $s$ , the standard error from the original, nonlinear regression. When this is done, (1.14) becomes

$$\frac{1}{n}R^\top(\theta)R(\theta) = \frac{1}{ns^2}X^\top(\beta)X(\beta),$$

and condition (2) is seen to be satisfied. However, there is rarely any reason to do this in practice.

Although the GNR is the most commonly encountered artificial regression, it differs from most artificial regressions in one key respect: there is one parameter,  $\sigma^2$ , for which there is no regressor. This happens because the criterion function,  $Q(\beta)$ , depends only on  $\beta$ . The GNR therefore has only as many regressors as  $\beta$  has components. This feature of the GNR is responsible for the fact that it does not quite satisfy condition (2). The fact that  $Q(\beta)$  does not depend on  $\sigma^2$  also causes the asymptotic covariance matrix to be block diagonal between the  $k \times k$  block that corresponds to  $\beta$  and the  $1 \times 1$  block that corresponds to  $\sigma^2$ .

#### 4 USES OF THE GNR

The GNR, like other artificial regressions, has several uses, depending on the parameter values at which the regressand and regressors are evaluated. If we evaluate them at  $\hat{\beta}$ , the vector of NLS parameter estimates, regression (1.7) becomes

$$\mathbf{y} - \hat{\mathbf{x}} = \hat{\mathbf{X}}\hat{\beta} + \text{residuals}, \quad (1.17)$$

where  $\hat{\mathbf{x}} \equiv \mathbf{x}(\hat{\beta})$  and  $\hat{\mathbf{X}} \equiv \mathbf{X}(\hat{\beta})$ . By condition (1), which follows from the first-order conditions for NLS estimation, the OLS estimate  $\hat{\beta}$  from this regression is a zero vector. In consequence, the explained sum of squares, or ESS, from regression (1.17) will be 0, and the SSR will be equal to

$$\|\mathbf{y} - \hat{\mathbf{x}}\|^2 = (\mathbf{y} - \hat{\mathbf{x}})^\top(\mathbf{y} - \hat{\mathbf{x}}),$$

which is the SSR from the original nonlinear regression.

Although it may seem curious to run an artificial regression all the coefficients of which are known in advance to be zero, there can be two very good reasons for doing so. The first reason is to check that the vector  $\hat{\beta}$  reported by a program for NLS estimation really does satisfy the first-order conditions. Computer programs for calculating NLS estimates do not yield reliable answers in every case; see McCullough (1999). The GNR provides an easy way to see whether the first-order conditions are actually satisfied. If all the  $t$ -statistics for the GNR are not less than about  $10^{-4}$ , and the  $R^2$  is not less than about  $10^{-8}$ , then the value of  $\hat{\beta}$  reported by the program should be regarded with suspicion.

The second reason to run the GNR (1.17) is to calculate an estimate of  $\text{var}(\hat{\beta})$ , the covariance matrix of the NLS estimates. The usual OLS covariance matrix from regression (1.17) is

$$\widehat{\text{var}}(\hat{\beta}) = s^2(\hat{\mathbf{X}}^\top \hat{\mathbf{X}})^{-1}, \quad (1.18)$$

which is similar to (1.16) except that everything is now evaluated at  $\hat{\beta}$ . Thus running the GNR (1.17) provides an easy way to calculate what is arguably the best estimate of  $\text{var}(\hat{\beta})$ . Of course, for (1.18) to provide an asymptotically valid covariance matrix estimate, it is essential that the error terms in (1.2) be independent and identically distributed, as we have assumed so far. We will discuss ways to drop this assumption in Section 7.

Since the GNR satisfies the one-step property, it and other artificial regressions can evidently be used to obtain one-step efficient estimates. However, although one-step estimation is of considerable theoretical interest, it is generally of modest practical interest, for two reasons. First, we often do not have a root- $n$  consistent estimator to start from and, secondly, modern computers are so fast that the savings from stopping after just one step are rarely substantial.

What is often of great practical interest is the use of the GNR as part of a numerical minimization algorithm to find the NLS estimates  $\hat{\beta}$  themselves. In practice, the classical Gauss–Newton updating procedure (1.6) should generally be replaced by

$$\beta_{(m)} = \beta_{(m-1)} + \alpha_{(m)} b_{(m)},$$

where  $\alpha_{(m)}$  is a scalar that is chosen in various ways by different algorithms, but always in such a way that  $Q(\beta_{(m+1)}) < Q(\beta_{(m)})$ . Numerical optimization methods are discussed by Press *et al.* (1992), among many others. Artificial regressions other than the GNR allow these methods to be used more widely than just in the least squares context.

## 5 HYPOTHESIS TESTING WITH ARTIFICIAL REGRESSIONS

Artificial regressions like the GNR are probably employed most frequently for hypothesis testing. Suppose we wish to test a set of  $r$  equality restrictions on  $\theta$ . Without loss of generality, we can assume that these are zero restrictions. This allows us to partition  $\theta$  into two subvectors,  $\theta_1$  of length  $k - r$ , and  $\theta_2$  of length  $r$ , the restrictions being that  $\theta_2 = 0$ . If the estimator  $\hat{\theta}$  is not only root- $n$  consistent but also asymptotically normal, an appropriate statistic for testing these restrictions is

$$\hat{\theta}_2^\top (\widehat{\text{var}}(\hat{\theta}_2))^{-1} \hat{\theta}_2, \quad (1.19)$$

which will be asymptotically distributed as  $\chi^2(r)$  under the null if  $\widehat{\text{var}}(\hat{\theta}_2)$  is a suitable estimate of the covariance matrix of  $\hat{\theta}_2$ .

Suppose that  $r(\theta)$  and  $R(\theta)$  define an artificial regression for the estimator  $\hat{\theta}$ . Let  $\hat{\theta} \equiv [\hat{\theta}_1 \vdots 0]$  be a vector of root- $n$  consistent estimates under the null. Then, if the variables of the artificial regression are evaluated at  $\hat{\theta}$ , the regression can be expressed as

$$r(\hat{\theta}_1, 0) = R_1(\hat{\theta}_1, 0)b_1 + R_2(\hat{\theta}_2, 0)b_2 + \text{residuals}, \quad (1.20)$$

where the partitioning of  $R = [R_1 \ R_2]$  corresponds to the partitioning of  $\theta$  as  $[\theta_1 \vdots \theta_2]$ . Regression (1.20) will usually be written simply as

$$\bar{r} = \bar{R}_1 b_1 + \bar{R}_2 b_2 + \text{residuals},$$

although this notation hides the fact that  $\hat{\theta}$  satisfies the null hypothesis.

By the one-step property,  $\hat{b}_2$  from (1.20) is asymptotically equivalent under the null to the estimator  $\hat{\theta}_2$ , since under the null the true value of  $\theta_2$  is zero. This suggests that we may replace  $\hat{\theta}_2$  in (1.19) by  $\hat{b}_2$ . By property (2), the asymptotic covariance matrix of  $n^{1/2}(\hat{\theta} - \theta_0)$  is estimated by  $(n^{-1}\hat{R}^\top\hat{R})^{-1}$ . A suitable estimate of the covariance matrix of  $\hat{\theta}_2$  can be obtained from this by use of the Frisch–Waugh–Lovell (FWL) theorem: See Davidson and MacKinnon (1993, ch. 1) for a full treatment of the FWL theorem. The estimate is  $(\hat{R}_2^\top\hat{M}_1\hat{R}_2)^{-1}$ , where the orthogonal projection matrix  $\hat{M}_1$  is defined by

$$\hat{M}_1 = I - \hat{R}_1(\hat{R}_1^\top\hat{R}_1)^{-1}\hat{R}_1^\top. \quad (1.21)$$

By the same theorem, we have that

$$\hat{b}_2 = (\hat{R}_2^\top\hat{M}_1\hat{R}_2)^{-1}\hat{R}_2^\top\hat{M}_1\hat{r}. \quad (1.22)$$

Thus the artificial regression version of the test statistic (1.19) is

$$\hat{b}_2^\top\hat{R}_2^\top\hat{M}_1\hat{R}_2\hat{b}_2 = \hat{r}^\top\hat{M}_1\hat{R}_2(\hat{R}_2^\top\hat{M}_1\hat{R}_2)^{-1}\hat{R}_2^\top\hat{M}_1\hat{r}. \quad (1.23)$$

The following theorem demonstrates the asymptotic validity of (1.23).

**Theorem 1.** If the regressand  $r(\theta)$  and the regressor matrix  $R(\theta)$  define an artificial regression for the root- $n$  consistent, asymptotically normal, estimator  $\hat{\theta}$ , and if the partition  $R = [R_1 \ R_2]$  corresponds to the partition  $\theta = [\theta_1 : \theta_2]$ , then the statistic (1.23), computed at any root- $n$  consistent  $\hat{\theta} = [\hat{\theta}_1 : 0]$ , is asymptotically distributed as  $\chi^2(r)$  under the null hypothesis that  $\theta_2 = 0$ , and is asymptotically equivalent to the generic statistic (1.19).

**Proof.** To prove this theorem, we need to show two things. The first is that

$$n^{-1}\hat{R}_2^\top\hat{M}_1\hat{R}_2 = n^{-1}R_2(\theta_0)M_1(\theta_0)R_2(\theta_0) + o_p(1),$$

where  $\theta_0$  is the true parameter vector, and  $M_1(\theta_0)$  is defined analogously to (1.21). This result follows by standard asymptotic arguments based on the one-step property. The second is that the vector

$$n^{-1/2}\hat{R}_2^\top\hat{M}_1\hat{r} = n^{-1/2}R_2^\top(\theta_0)M_1(\theta_0)r(\theta_0) + o_p(1)$$

is asymptotically normally distributed. The equality here also follows by standard asymptotic arguments. The asymptotic normality of  $\hat{\theta}$  implies that  $\hat{b}$  is asymptotically normally distributed. Therefore, by (1.22),  $n^{-1/2}\hat{R}_2^\top\hat{M}_1\hat{r}$  must also be asymptotically normally distributed. These two results imply that, asymptotically under the null hypothesis, the test statistic (1.23) is a quadratic form in a normally distributed  $r$ -vector, the mean of which is zero, and the inverse of its covariance matrix. Such a quadratic form follows the  $\chi^2(r)$  distribution. ■

**Remarks.** The statistic (1.23) can be computed as the difference between the sums of squared residuals (SSR) from the regressions

$$\hat{r} = \hat{R}_1 b_1 + \text{residuals, and} \quad (1.24)$$

$$\check{r} = \check{R}_1 b_1 + \check{R}_2 b_2 + \text{residuals.} \quad (1.25)$$

Equivalently, it can be computed as the difference between the explained sums of squares (ESS), with the opposite sign, or as the ESS from the FWL regression corresponding to (1.25):

$$\hat{M}_1 \hat{r} = \hat{M}_1 \check{R}_2 b_2 + \text{residuals.}$$

If  $\text{plim } n^{-1} \hat{r}^\top \hat{r} = 1$  for all root- $n$  consistent  $\hat{\theta}$ , there are other convenient ways of computing (1.23), or statistics asymptotically equivalent to it. One is the ordinary F-statistic for  $b_2 = 0$  in regression (25):

$$F = \frac{\hat{r}^\top \hat{M}_1 \check{R}_2 (\check{R}_2^\top \hat{M}_1 \check{R}_2)^{-1} \check{R}_2^\top \hat{M}_1 \hat{r} / r}{\hat{r}^\top \hat{M}_1 \hat{r} / (n - k)}, \quad (1.26)$$

which works because the denominator tends to a probability limit of 1 as  $n \rightarrow \infty$ . This statistic is, of course, in  $F$  rather than  $\chi^2$  form.

Another frequently used test statistic is available if  $\hat{\theta}$  is actually the vector of restricted estimates, that is, the estimator that minimizes the criterion function when the restriction that  $\theta_2 = 0$  is imposed. In this case,  $n$  times the uncentered  $R^2$  from (1.25) is a valid test statistic. With this choice of  $\hat{\theta}$ , the ESS from (1.24) is zero, by property (1). Thus (1.23) is just the ESS from (1.25). Since  $nR^2 = \text{ESS}/(\text{TSS}/n)$ , where TSS denotes the total sum of squares, and since  $\text{TSS}/n \rightarrow 1$  as  $n \rightarrow \infty$ , it follows that this statistic is asymptotically equivalent to (1.23).

Even though the GNR does not satisfy condition (2) when it is expressed in its usual form with all variables not divided by the standard error  $s$ , the  $F$ -statistic (1.26) and the  $nR^2$  statistic are still valid test statistics, because they are both ratios. In fact, variants of the GNR are routinely used to perform many types of specification tests. These include tests for serial correlation similar to the ones proposed by Godfrey (1978), nonnested hypothesis tests where both models are parametric (Davidson and MacKinnon, 1981), and nonnested hypothesis tests where the alternative model is nonparametric (Delgado and Stengos, 1994). They also include several Durbin–Wu–Hausman, or DWH, tests, in which an efficient estimator is compared with an inefficient estimator that is consistent under weaker conditions; see Sections 7.9 and 11.4 of Davidson and MacKinnon (1993).

## 6 THE OPG REGRESSION

By no means all interesting econometric models are regression models. It is therefore useful to see if artificial regressions other than the GNR exist for wide classes of models. One of these is the *outer-product-of-the-gradient regression*, or *OPG*

regression, a particularly simple artificial regression that can be used with most models that are estimated by maximum likelihood. Suppose we are interested in a model of which the loglikelihood function can be written as

$$\ell(\theta) = \sum_{t=1}^n \ell_t(\theta), \quad (1.27)$$

where  $\ell_t(\cdot)$  denotes the contribution to the loglikelihood function associated with observation  $t$ . This is the log of the density of the dependent variable(s) for observation  $t$ , conditional on observations  $1, \dots, t-1$ . Thus lags of the dependent variable(s) are allowed. The key feature of (1.27) is that  $\ell(\theta)$  is a sum of contributions from each of the  $n$  observations.

Now let  $G(\theta)$  be the matrix with typical element

$$G_{it}(\theta) \equiv \frac{\partial \ell_t(\theta)}{\partial \theta_i}; \quad t = 1, \dots, n, i = 1, \dots, k.$$

The matrix  $G(\theta)$  is called the *matrix of contributions to the gradient*, or the *CG matrix*, because the derivative of the sample loglikelihood (1.27) with respect to  $\theta_i$ , the  $i$ th component of  $\theta$ , is the sum of the elements of column  $i$  of  $G(\theta)$ . The OPG regression associated with (1.27) can be written as

$$\mathbf{1} = G(\theta)\mathbf{b} + \text{residuals}, \quad (1.28)$$

where  $\mathbf{1}$  denotes an  $n$ -vector of 1s.

It is easy to see that the OPG regression (1.28) satisfies the conditions for it to be an artificial regression. Condition (1.1') is evidently satisfied, since  $R^\top(\theta)r(\theta) = G^\top(\theta)\mathbf{1}$ , the components of which are the derivatives of  $\ell(\theta)$  with respect to each of the  $\theta_i$ . Condition (2) is also satisfied, because, under standard regularity conditions, if  $\theta$  is the true parameter vector,

$$\operatorname{plim}_{n \rightarrow \infty} (n^{-1} R^\top(\theta) R(\theta)) = \operatorname{plim}_{n \rightarrow \infty} (n^{-1} G^\top(\theta) G(\theta)) = \mathcal{I}(\theta).$$

Here  $\mathcal{I}(\theta)$  denotes the information matrix, defined as

$$\mathcal{I}(\theta) = \lim_{n \rightarrow \infty} \frac{1}{n} \sum_{t=1}^n E(G_t^\top(\theta) G_t(\theta)),$$

where  $G_t(\cdot)$  is the  $t$ th row of  $G(\cdot)$ . Since, as is well known, the asymptotic covariance matrix of  $n^{1/2}(\hat{\theta} - \theta_0)$  is given by the inverse of the information matrix, condition (2) is satisfied under the further weak regularity condition that  $\mathcal{I}(\theta)$  should be continuous in  $\theta$ . Condition (3) is also satisfied, since it can be shown that one-step estimates from the OPG regression are asymptotically equivalent to maximum likelihood estimates. The proof is quite similar to the one for the GNR given in Section 3.

It is particularly easy to compute an LM test by using the OPG regression. Let  $\tilde{\theta}$  denote the constrained ML estimates obtained by imposing  $r$  restrictions when maximizing the loglikelihood. Then the ESS from the OPG regression

$$\mathbf{t} = \mathbf{G}(\tilde{\theta})\mathbf{b} + \text{residuals}, \quad (1.29)$$

which is equal to  $n$  times the uncentered  $R^2$ , is the OPG form of the LM statistic. Like the GNR, the OPG regression can be used for many purposes. The use of what is essentially the OPG regression for obtaining maximum likelihood estimates and computing covariance matrices was advocated by Berndt, Hall, Hall, and Hausman (1974). Using it to compute Lagrange Multiplier, or LM, tests was suggested by Godfrey and Wickens (1981), and using it to compute information matrix tests was proposed by Chesher (1983) and Lancaster (1984). The OPG regression is appealing for all these uses because it applies to a very wide variety of models and requires only first derivatives. In general, however, both estimated covariance matrices and test statistics based on the OPG regression are not very reliable in finite samples. In particular, a large number of papers, including Chesher and Spady (1991), Davidson and MacKinnon (1985a, 1992), and Godfrey, McAleer, and McKenzie (1988), have shown that, in finite samples, LM tests based on the OPG regression tend to overreject, often very severely.

Despite this drawback, the OPG regression provides a particularly convenient way to obtain various theoretical results. For example, suppose that we are interested in the variance of  $\hat{\theta}_2$ , the last element of  $\hat{\theta}$ . If  $\theta_1$  denotes a vector of the remaining  $k - 1$  elements, and  $\mathbf{G}(\theta)$  and  $\mathbf{b}$  are partitioned in the same way as  $\theta$ , the OPG regression becomes

$$\mathbf{t} = \mathbf{G}_1(\theta)\mathbf{b}_1 + \mathbf{G}_2(\theta)\mathbf{b}_2 + \text{residuals},$$

and the FWL regression derived from this by retaining only the last regressor is

$$\mathbf{M}_1\mathbf{t} = \mathbf{M}_1\mathbf{G}_2\mathbf{b}_2 + \text{residuals},$$

where  $\mathbf{M}_1 \equiv \mathbf{I} - \mathbf{G}_1(\mathbf{G}_1^\top \mathbf{G}_1)^{-1}\mathbf{G}_1^\top$ , and the dependence on  $\theta$  has been suppressed for notational convenience. The covariance matrix estimate from this is just

$$(\mathbf{G}_2^\top \mathbf{M}_1 \mathbf{G}_2)^{-1} = (\mathbf{G}_2^\top \mathbf{G}_2 - \mathbf{G}_2^\top \mathbf{G}_1 (\mathbf{G}_1^\top \mathbf{G}_1)^{-1} \mathbf{G}_1^\top \mathbf{G}_2)^{-1}. \quad (1.30)$$

If we divide each of the components of (1.30) by  $n$  and take their probability limits, we find that

$$\lim_{n \rightarrow \infty} \text{var}(n^{1/2}(\hat{\theta}_2 - \theta_{20})) = (\mathcal{J}_{22} - \mathcal{J}_{21}\mathcal{J}_{11}^{-1}\mathcal{J}_{12})^{-1},$$

where  $\theta_{20}$  is the true value of  $\theta_2$ . This is a very well-known result, but, since its relation to the FWL theorem is not obvious without appeal to the OPG regression, it is not usually obtained in such a convenient or illuminating way.

## 7 AN ARTIFICIAL REGRESSION FOR GMM ESTIMATION

Another useful artificial regression, much less well known than the OPG regression, is available for a class of models estimated by the generalized method of moments (GMM). Many such models can be formulated in terms of functions  $f_i(\theta)$  of the model parameters and the data, such that, when they are evaluated at the true  $\theta$ , their expectations conditional on corresponding information sets,  $\Omega_t$ , vanish. The  $\Omega_t$  usually contain all information available prior to the time of observation  $t$ , and so, as with the GNR and the OPG regression, lags of dependent variables are allowed.

Let the  $n \times l$  matrix  $W$  denote the instruments used to obtain the GMM estimates. The  $t$ th row of  $W$ , denoted  $W_t$ , must contain variables in  $\Omega_t$  only. The dimension of  $\theta$  is  $k$ , as before, and, for  $\theta$  to be identified, we need  $l \geq k$ . The GMM estimates with  $l \times l$  weighting matrix  $A$  are obtained by minimizing the criterion function

$$Q(\theta) = \frac{1}{2} f^\top(\theta) W A W^\top f(\theta) \quad (1.31)$$

with respect to  $\theta$ . Here  $f(\theta)$  is the  $n$ -vector with typical element  $f_i(\theta)$ . For the procedure known as efficient GMM, the weighting matrix  $A$  is chosen so as to be proportional, asymptotically at least, to the inverse of the covariance matrix of  $W^\top f(\theta)$ . In the simplest case, the  $f_i(\theta)$  are serially uncorrelated and homoskedastic with variance 1, and so an appropriate choice is  $A = (W^\top W)^{-1}$ . With this choice, the criterion function (1.31) becomes

$$Q(\theta) = \frac{1}{2} f^\top(\theta) P_W f(\theta), \quad (1.32)$$

where  $P_W$  is the orthogonal projection on to the columns of  $W$ .

Let  $J(\theta)$  be the negative of the  $n \times k$  Jacobian matrix of  $f(\theta)$ , so that the  $t$ th element of  $J(\theta)$  is  $-\partial f_t / \partial \theta_i(\theta)$ . The first-order conditions for minimizing (1.32) are

$$J^\top(\theta) P_W f(\theta) = 0. \quad (1.33)$$

By standard arguments, it can be seen that the vector  $\hat{\theta}$  that solves (1.33) is asymptotically normal and asymptotically satisfies the equation

$$n^{1/2}(\hat{\theta} - \theta_0) = (n^{-1} J_0^\top P_W J_0)^{-1} n^{-1/2} J_0^\top P_W f_0, \quad (1.34)$$

with  $J_0 = J(\theta_0)$  and  $f_0 = f(\theta_0)$ . See Davidson and MacKinnon (1993, ch. 17), for a full discussion of GMM estimation.

Now consider the artificial regression

$$f(\theta) = P_W J(\theta) b + \text{residuals}. \quad (1.35)$$

By the first-order conditions (1.33) for  $\theta$ , this equation clearly satisfies condition (1), and in fact it also satisfies condition (1') for the criterion function  $Q(\theta)$  of (1.32). Since the covariance matrix of  $f(\theta_0)$  is just the identity matrix, it follows from (1.34) that condition (2) is also satisfied. Arguments just like those presented in Section 3 for the GNR can be used to show that condition (3), the one-step property, is also satisfied by (1.35).

If the  $f_i(\theta_0)$  are homoskedastic but with unknown variance  $\sigma^2$ , regression (1.35) can be used in exactly the same way as the GNR. Either the regressand and regressors can be divided by a suitable consistent estimate of  $\sigma$ , or else all test statistics can be computed as ratios, in  $F$  or  $nR^2$  form, as appropriate.

An important special case of (1.35) is provided by the class of regression models, linear or nonlinear, estimated with instrumental variables (IV). Such a model can be written in the form (1.3), but it will be estimated by minimizing, not the criterion function (1.4) related to the sum of squared residuals, but rather

$$Q(\beta) \equiv \frac{1}{2}(y - x(\beta))^\top P_W(y - x(\beta)),$$

where  $W$  is an  $n \times 1$  matrix of instrumental variables. This criterion function has exactly the same form as (1.32), with  $\beta$  instead of  $\theta$ , and with  $f(\beta) = y - x(\beta)$ . In addition,  $J(\beta) = X(\beta)$ , where  $X(\beta)$  is defined, exactly as for the GNR, to have the  $t$ th element  $\partial x_t / \partial \beta_i(\beta)$ . The resulting artificial regression for the IV model, which takes the form

$$y - x(\beta) = P_W X(\beta) b + \text{residuals}, \quad (1.36)$$

is often referred to as a GNR, because, except for the projection matrix  $P_W$ , it is identical to (1.7): See Davidson and MacKinnon (1993, ch. 7).

## 8 ARTIFICIAL REGRESSIONS AND HETEROSKEDASTICITY

Covariance matrices and test statistics calculated via the GNR (1.7), or via artificial regressions such as (1.35) and (1.36), are not asymptotically valid when the assumption that the error terms are iid is violated. Consider a modified version of the nonlinear regression model (1.3), in which  $E(uu^\top) = \Omega$ , where  $\Omega$  is an  $n \times n$  diagonal matrix with  $t$ th diagonal element  $\omega_t^2$ . Let  $\hat{\Omega}$  denote an  $n \times n$  diagonal matrix with the squared residual  $\hat{u}_t^2$  as the  $t$ th diagonal element. It has been known since the work of White (1980) that the matrix

$$(\hat{X}^\top \hat{X})^{-1} \hat{X}^\top \hat{\Omega} \hat{X} (\hat{X}^\top \hat{X})^{-1} \quad (1.37)$$

provides an estimator of  $\text{var}(\hat{\beta})$ , which can be used in place of the usual estimator,  $s^2(\hat{X}^\top \hat{X})^{-1}$ . Like the latter, this *heteroskedasticity-consistent covariance matrix estimator*, or *HCCME*, can be computed by means of an artificial regression. We will refer to this regression as the *heteroskedasticity-robust Gauss–Newton regression*, or *HRGNR*.

In order to derive the HRGNR, it is convenient to begin with a linear regression model  $y = X\beta + u$ , and to consider the criterion function

$$Q(\beta) = \frac{1}{2}(y - X\beta)^T X(X^T \Omega X)^{-1} X^T (y - X\beta).$$

The negative of the gradient of this function with respect to  $\beta$  is

$$X^T X(X^T \Omega X)^{-1} X^T (y - X\beta), \quad (1.38)$$

and its Hessian is the matrix

$$X^T X(X^T \Omega X)^{-1} X^T X, \quad (1.39)$$

of which the inverse is the HCCME if we replace  $\Omega$  by  $\hat{\Omega}$ . Equating the gradient to zero just yields the OLS estimator, since  $X^T X$  and  $X^T \Omega X$  are  $k \times k$  nonsingular matrices.

Let  $V$  be an  $n \times n$  diagonal matrix with  $t$ th diagonal element equal to  $\omega_t$ ; thus  $V^2 = \Omega$ . Consider the  $n \times k$  regressor matrix  $R$  defined by

$$R = V X (X^T V^2 X)^{-1} X^T X = P_{VX} V^{-1} X, \quad (1.40)$$

where  $P_{VX}$  projects orthogonally on to the columns of  $VX$ . We have

$$R^T R = X^T X (X^T \Omega X)^{-1} X^T X, \quad (1.41)$$

which is just the Hessian (1.39). Let  $U(\beta)$  be a diagonal matrix with  $t$ th diagonal element equal to  $y_t - X_t \beta$ . Then, if we define  $R(\beta)$  as in (1.40) but with  $V$  replaced by  $U(\beta)$ , we find that  $\hat{R}^T \hat{R}$  is the HCCME (1.37).

In order to derive the regressand  $r(\beta)$ , note that, for condition (1') to be satisfied, we require

$$R^T(\beta)r(\beta) = X^T X (X^T U^2(\beta) X)^{-1} X^T (y - X\beta);$$

recall (1.38). Since the  $t$ th element of  $U(\beta)$  is  $y_t - X_t \beta$ , this implies that

$$r(\beta) = U^{-1}(\beta)(y - X\beta) = \iota.$$

In the general nonlinear case,  $X$  becomes  $X(\beta)$ , and the HRGNR has the form

$$\iota = P_{U(\beta)X(\beta)} U^{-1}(\beta) X(\beta) b + \text{residuals}, \quad (1.42)$$

where now the  $t$ th diagonal element of  $U(\beta)$  is  $y_t - x_t(\beta)$ . When  $\beta = \hat{\beta}$ , the vector of NLS estimates,

$$\begin{aligned} \hat{r}^T \hat{R} &= \iota^T P_{\hat{U}\hat{X}} \hat{U}^{-1} \hat{X} \\ &= \iota^T \hat{U} \hat{X} (\hat{X}^T \hat{U} \hat{U} \hat{X})^{-1} \hat{X}^T \hat{U} \hat{U}^{-1} \hat{X} \\ &= \hat{u}^T \hat{X} (\hat{X}^T \hat{\Omega} \hat{X})^{-1} \hat{X}^T \hat{X} = 0, \end{aligned} \quad (1.43)$$

because the NLS first-order conditions give  $\hat{X}^\top \hat{u} = 0$ . Thus condition (1) is satisfied for the nonlinear case. Condition (2) is satisfied by construction, as can be seen by putting hats on everything in (1.41).

For condition (3) to hold, regression (1.42) must satisfy the one-step property. We will only show that this property holds for linear models. Extending the argument to nonlinear models would be tedious but not difficult. In the linear case, evaluating (1.42) at an arbitrary  $\hat{\beta}$  gives

$$\hat{b} = (X^\top \tilde{U}^{-1} P_{\tilde{U}X} \tilde{U}^{-1} X)^{-1} X^\top \tilde{U}^{-1} P_{\tilde{U}X} l.$$

With a little algebra, it can be shown that this reduces to

$$\hat{b} = (X^\top X)^{-1} X^\top \tilde{u} = (X^\top X)^{-1} X^\top (y - X\hat{\beta}) = \hat{\beta} - \hat{\beta}, \quad (1.44)$$

where  $\hat{\beta}$  is the OLS estimator. It follows that the one-step estimator  $\hat{\beta} + \hat{b}$  is equal to  $\hat{\beta}$ , as we wished to show. In the nonlinear case, of course, we obtain an asymptotic equality rather than an exact equality.

As with the ordinary GNR, the HRGNR is particularly useful for hypothesis testing. If we partition  $\beta$  as  $[\beta_1 \vdots \beta_2]$  and wish to test the  $r$  zero restrictions  $\beta_2 = 0$ , we need to run two versions of the regression and compute the difference between the two SSRs or ESSs. The two regressions are:

$$l = P_{\tilde{U}X} \tilde{U}^{-1} \tilde{X}_1 b_1 + \text{residuals, and} \quad (1.45)$$

$$l = P_{\tilde{U}X} \tilde{U}^{-1} \tilde{X}_1 b_1 + P_{\tilde{U}X} \tilde{U}^{-1} \tilde{X}_2 b_2 + \text{residuals.} \quad (1.46)$$

It is important to note that the first regression is *not* the HRGNR for the restricted model, because it uses the matrix  $P_{\tilde{U}X}$  rather than the matrix  $P_{\tilde{U}\tilde{X}_1}$ . In consequence, the regressand in (1.45) will not be orthogonal to the regressors. This is why we need to run two artificial regressions. We could compute an ordinary  $F$ -statistic instead of the difference between the SSRs from (1.45) and (1.46), but there would be no advantage to doing so, since the  $F$ -form of the test merely divides by a stochastic quantity that tends to 1 asymptotically.

The HRGNR appears to be new. The trick of multiplying  $X(\beta)$  by  $U^{-1}(\beta)$  in order to obtain an HCCME by means of an OLS regression was used, in a different context, by Messer and White (1984). This trick does cause a problem in some cases. If any element on the diagonal of the matrix  $U(\beta)$  is equal to 0, the inverse of that element cannot be computed. Therefore, it is necessary to replace any such element by a small, positive number before computing  $U^{-1}(\beta)$ .

A different, and considerably more limited, type of heteroskedasticity-robust GNR, which is applicable only to hypothesis testing, was first proposed by Davidson and MacKinnon (1985b). It was later rediscovered by Wooldridge (1990, 1991) and extended to handle other cases, including regression models with error terms that have autocorrelation as well as heteroskedasticity of unknown form.

It is possible to construct a variety of artificial regressions that provide different covariance matrix estimators for regression models. From (1.43) and (1.44), it follows that any artificial regression with regressand

$$\mathbf{r}(\beta) = \mathbf{U}^{-1}(\beta)(\mathbf{y} - \mathbf{x}(\beta))$$

and regressors

$$\mathbf{R}(\beta) = \mathbf{P}_{\mathbf{U}(\beta)\mathbf{X}(\beta)}\mathbf{U}^{-1}(\beta)\mathbf{X}(\beta)$$

satisfies properties (1) and (3) for the least-squares estimator, for any nonsingular matrix  $\mathbf{U}(\beta)$ . Thus any sandwich covariance matrix estimator can be computed by choosing  $\mathbf{U}(\beta)$  appropriately; the estimator (1.37) is just one example. In fact, it is possible to develop artificial regressions that allow testing not only with a variety of different HCCMEs, but also with some sorts of heteroskedasticity and autocorrelation consistent (HAC) covariance matrix estimators. It is also a simple matter to use such estimators with modified versions of the artificial regression (1.35) used with models estimated by GMM.

## 9 DOUBLE-LENGTH REGRESSIONS

Up to this point, the number of observations for all the artificial regressions we have studied has been equal to  $n$ , the number of observations in the data. In some cases, however, artificial regressions may have  $2n$  or even  $3n$  observations. This can happen whenever each observation makes two or more contributions to the criterion function.

The first *double-length artificial regression*, or *DLR*, was proposed by Davidson and MacKinnon (1984a). We will refer to it as *the DLR*, even though it is no longer the only artificial regression with  $2n$  observations. The class of models to which the DLR applies is a subclass of the one used for GMM estimation. Such models may be written as

$$f_t(y_t, \theta) = \varepsilon_t, \quad t = 1, \dots, n, \quad \varepsilon_t \sim \text{NID}(0, 1), \quad (1.47)$$

where, as before, each  $f_t(\cdot)$  is a smooth function that depends on the data and on a  $k$ -vector of parameters  $\theta$ . Here, however, the  $f_t$  are assumed to be normally distributed conditional on the information sets  $\Omega_t$ , as well as being of mean zero, serially uncorrelated, and homoskedastic with variance 1. Further,  $f_t$  may depend only on a scalar dependent variable  $y_t$ , although lagged dependent variables are allowed as explanatory variables.

The class of models (1.47) is much less restrictive than it may at first appear to be. In particular, it is not essential that the error terms follow the normal distribution, although it is essential that they follow some specified, continuous distribution, which can be transformed into the standard normal distribution, so as to allow the model to be written in the form of (1.47). A great many models that involve transformations of the dependent variable can be put into the form of (1.47). For example, consider the Box–Cox regression model

$$\tau(y_t, \lambda) = \sum_{i=1}^k \beta_i \tau(X_{ti}, \lambda) + \sum_{j=1}^l \gamma_j Z_{tj} + u_t, \quad u_t \sim N(0, \sigma^2), \quad (1.48)$$

where  $\tau(x, \lambda) = (x^\lambda - 1)/\lambda$  is the Box–Cox transformation (Box and Cox, 1964),  $y_t$  is the dependent variable, the  $X_{ti}$  are independent variables that are always positive, and the  $Z_{tj}$  are additional independent variables. We can rewrite (1.48) in the form of (1.47) by making the definition

$$f_t(y_t, \theta) = \frac{1}{\sigma} \left( \tau(y_t, \lambda) - \sum_{i=1}^k \beta_i \tau(X_{ti}, \lambda) - \sum_{j=1}^l \gamma_j Z_{tj} \right).$$

For the model (1.47), the contribution of the  $t$ th observation to the loglikelihood function  $\ell(y, \theta)$  is

$$\ell_t(y_t, \theta) = -\frac{1}{2} \log(2\pi) - \frac{1}{2} f_t^2(y_t, \theta) + k_t(y_t, \theta),$$

where

$$k_t(y_t, \theta) \equiv \log \left| \frac{\partial f_t(y_t, \theta)}{\partial y_t} \right|$$

is a Jacobian term. Now let us make the definitions

$$F_{ti}(y_t, \theta) \equiv \frac{\partial f_t(y_t, \theta)}{\partial \theta_i} \quad \text{and} \quad K_{ti}(y_t, \theta) \equiv \frac{\partial k_t(y_t, \theta)}{\partial \theta_i}$$

and define  $F(y, \theta)$  and  $K(y, \theta)$  as the  $n \times k$  matrices with typical elements  $F_{ti}(y_t, \theta)$  and  $K_{ti}(y_t, \theta)$  and typical rows  $F_t(y, \theta)$  and  $K_t(y, \theta)$ . Similarly, let  $f(y, \theta)$  be the  $n$ -vector with typical element  $f_t(y_t, \theta)$ .

The DLR, which has  $2n$  artificial observations, may be written as

$$\begin{bmatrix} f(y, \theta) \\ \iota \end{bmatrix} = \begin{bmatrix} -F(y, \theta) \\ K(y, \theta) \end{bmatrix} b + \text{residuals.} \quad (1.49)$$

Since the gradient of  $\ell(y, \theta)$  is

$$g(y, \theta) = -F^\top(y, \theta)f(y, \theta) + K^\top(y, \theta)\iota, \quad (1.50)$$

we see that regression (1.49) satisfies condition (1'). It can also be shown that it satisfies conditions (2) and (3), and thus it has all the properties of an artificial regression.

The DLR can be used for many purposes, including nonnested hypothesis tests of models with different functional forms (Davidson and MacKinnon, 1984a), tests of functional form (MacKinnon and Magee, 1990), and tests of linear and loglinear regressions against Box–Cox alternatives like (1.48) (Davidson and MacKinnon, 1985a). The latter application has recently been extended to models with AR(1) errors by Baltagi (1999). An accessible discussion of the DLR may be

found in Davidson and MacKinnon (1988). When both the OPG regression and the DLR are available, the finite-sample performance of the latter always seems to be very much better than that of the former.

As we remarked earlier, the DLR is not the only artificial regression with  $2n$  artificial observations. In particular, Orme (1995) showed how to construct such a regression for the widely-used tobit model, and Davidson and MacKinnon (1999) provided evidence that Orme's regression generally works very well. It makes sense that a double-length regression should be needed in this case, because the tobit loglikelihood is the sum of two summations, which are quite different in form. One summation involves all the observations for which the dependent variable is equal to zero, and the other involves all the observations for which it takes on a positive value.

## 10 AN ARTIFICIAL REGRESSION FOR BINARY RESPONSE MODELS

For binary response models such as the logit and probit models, there exists a very simple artificial regression that can be derived as an extension of the Gauss–Newton regression. It was independently suggested by Engle (1984) and Davidson and MacKinnon (1984b).

The object of a binary response model is to predict the probability that the binary dependent variable,  $y_t$ , is equal to 1 conditional on some information set  $\Omega_t$ . A useful class of binary response models can be written as

$$E(y_t | \Omega_t) = \Pr(y_t = 1) = F(Z_t \beta). \quad (1.51)$$

Here  $Z_t$  is a row vector of explanatory variables that belong to  $\Omega_t$ ,  $\beta$  is the vector of parameters to be estimated, and  $F(x)$  is the differentiable cumulative distribution function (CDF) of some scalar probability distribution. For the probit model,  $F(x)$  is the standard normal CDF. For the logit model,  $F(x)$  is the logistic function

$$\frac{\exp(x)}{1 + \exp(x)} = (1 + \exp(-x))^{-1}.$$

The loglikelihood function for this class of binary response models is

$$\ell(\beta) = \sum_{t=1}^n ((1 - y_t) \log(1 - F(Z_t \beta)) + y_t \log(F(Z_t \beta))), \quad (1.52)$$

If  $f(x) = F'(x)$  is the density corresponding for the CDF  $F(x)$ , the first-order conditions for maximizing (1.52) are

$$\sum_{t=1}^n \frac{(y_t - \hat{F}_t) \hat{f}_t Z_{ti}}{\hat{F}_t(1 - \hat{F}_t)} = 0, \quad i = 1, \dots, k, \quad (1.53)$$

where  $Z_{ti}$  is the  $i$ th component of  $Z_t$ ,  $\hat{f}_t \equiv f(Z_t \hat{\beta})$  and  $\hat{F}_t \equiv F(Z_t \hat{\beta})$ .

There is more than one way to derive the artificial regression that corresponds to the model (1.51). The easiest is to rewrite it in the form of the nonlinear regression model

$$y_t = F(Z_t \beta) + u_t. \quad (1.54)$$

The error term  $u_t$  here is evidently nonnormal and heteroskedastic. Because  $y_t$  is like a Bernoulli trial with probability  $p$  given by  $F(Z_t \beta)$ , and the variance of a Bernoulli trial is  $p(1 - p)$ , the variance of  $u_t$  is

$$v_t(\beta) \equiv F(Z_t \beta)(1 - F(Z_t \beta)). \quad (1.55)$$

The ordinary GNR for (1.54) would be

$$y_t - F(Z_t \beta) = f(Z_t \beta)Z_t b + \text{residual},$$

but the ordinary GNR is not appropriate because of the heteroskedasticity of the  $u_t$ . Multiplying both sides by the square root of the inverse of (1.55) yields the artificial regression

$$v_t^{-1/2}(\beta)(y_t - F(Z_t \beta)) = v_t^{-1/2}(\beta)f(Z_t \beta)Z_t b + \text{residual}. \quad (1.56)$$

This regression has all the usual properties of artificial regressions. It can be seen from (1.53) that it satisfies condition (1'). Because a typical element of the information matrix corresponding to (1.52) is

$$\mathcal{I}_{ij}(\beta) = \text{plim}_{n \rightarrow \infty} \left( \frac{1}{n} \sum_{t=1}^n Z_{ti} Z_{tj} \frac{f(Z_t \beta)^2}{F(Z_t \beta)(1 - F(Z_t \beta))} \right),$$

it is not difficult to show that regression (1.56) satisfies condition (2). Finally, since (1.56) has the structure of a GNR, the arguments used in Section 3 show that it also satisfies condition (3), the one-step property.

As an artificial regression, (1.56) can be used for all the things that other artificial regressions can be used for. In particular, when it is evaluated at restricted estimates  $\tilde{\beta}$ , the explained sum of squares is an LM test statistic for testing the restrictions. The normalization of the regressand by its standard error means that other test statistics, such as  $nR^2$  and the ordinary F-statistic for the coefficients on the regressors that correspond to the restricted parameters to be zero, are also asymptotically valid. However, they seem to have slightly poorer finite-sample properties than the ESS (Davidson and MacKinnon, 1984b). It is, of course, possible to extend regression (1.56) in various ways. For example, it has been extended to tests of the functional form of  $F(x)$  by Thomas (1993) and to tests of ordered logit models by Murphy (1996).

## 11 CONCLUSION

In this chapter, we have introduced the concept of an artificial regression and discussed several examples. We have seen that artificial regressions can be useful for minimizing criterion functions, computing one-step estimates, calculating covariance matrix estimates, and computing test statistics. The last of these is probably the most common application. There is a close connection between the artificial regression for a given model and the asymptotic theory for that model. Therefore, as we saw in Section 6, artificial regressions can also be very useful for obtaining theoretical results.

Most of the artificial regressions we have discussed are quite well known. This is true of the Gauss–Newton regression discussed in Sections 3 and 4, the OPG regression discussed in Section 6, the double-length regression discussed in Section 9, and the regression for binary response models discussed in Section 10. However, the artificial regression for GMM estimation discussed in Section 7 does not appear to have been treated previously in published work, and we believe that the heteroskedasticity-robust GNR discussed in Section 8 is new.

## *References*

- Baltagi, B. (1999). Double length regressions for linear and log-linear regressions with AR(1) disturbances. *Statistical Papers* 4, 199–209.
- Berndt, E.R., B.H. Hall, R.E. Hall, and J.A. Hausman (1974). Estimation and inference in nonlinear structural models. *Annals of Economic and Social Measurement* 3, 653–65.
- Box, G.E.P., and D.R. Cox (1964). An analysis of transformations. *Journal of the Royal Statistical Society, Series B* 26, 211–52.
- Chesher, A. (1983). The information matrix test: simplified calculation via a score test interpretation. *Economics Letters* 13, 45–8.
- Chesher, A., and R. Spady (1991). Asymptotic expansions of the information matrix test statistic. *Econometrica* 59, 787–815.
- Davidson, R., and J.G. MacKinnon (1981). Several tests for model specification in the presence of alternative hypotheses. *Econometrica* 49, 781–93.
- Davidson, R., and J.G. MacKinnon (1984a). Model specification tests based on artificial linear regressions. *International Economic Review* 25, 485–502.
- Davidson, R., and J.G. MacKinnon (1984b). Convenient Specification Tests for Logit and Probit Models. *Journal of Econometrics* 25, 241–62.
- Davidson, R., and J.G. MacKinnon (1985a). Testing linear and loglinear regressions against Box–Cox alternatives. *Canadian Journal of Economics* 18, 499–517.
- Davidson, R., and J.G. MacKinnon (1985b). Heteroskedasticity-robust tests in regression directions. *Annales de l'INSEE* 59/60, 183–218.
- Davidson, R., and J.G. MacKinnon (1988). Double-length artificial regressions. *Oxford Bulletin of Economics and Statistics* 50, 203–17.
- Davidson, R., and J.G. MacKinnon (1990). Specification tests based on artificial regressions. *Journal of the American Statistical Association* 85, 220–7.
- Davidson, R., and J.G. MacKinnon (1992). A new form of the information matrix test. *Econometrica* 60, 145–57.
- Davidson, R., and J.G. MacKinnon (1993). *Estimation and Inference in Econometrics*. New York: Oxford University Press.

- Davidson, R., and J.G. MacKinnon (1999). Bootstrap testing in nonlinear models. *International Economic Review* 40, 487–508.
- Delgado, M.A., and T. Stengos (1994). Semiparametric specification testing of non-nested econometric models. *Review of Economic Studies* 61, 291–303.
- Engle, R.F. (1984). Wald, Likelihood Ratio and Lagrange Multiplier Tests in Econometrics. In Zvi Griliches and Michael D. Intriligator (eds.). *Handbook of Econometrics*, Vol. II, Amsterdam: North-Holland.
- Godfrey, L.G. (1978). Testing against general autoregressive and moving average error models when the regressors include lagged dependent variables. *Econometrica* 46, 1293–301.
- Godfrey, L.G., and M.R. Wickens (1981). Testing linear and log-linear regressions for functional form. *Review of Economic Studies* 48, 487–96.
- Godfrey, L.G., M. McAleer, and C.R. McKenzie (1988). Variable addition and Lagrange Multiplier tests for linear and logarithmic regression models. *Review of Economics and Statistics* 70, 492–503.
- Lancaster, T. (1984). The covariance matrix of the information matrix test. *Econometrica* 52, 1051–3.
- MacKinnon, J.G., and L. Magee (1990). Transforming the dependent variable in regression models. *International Economic Review* 31, 315–39.
- McCullough, B.D. (1999). Econometric software reliability: EViews, LIMDEP, SHAZAM, and TSP. *Journal of Applied Econometrics* 14, 191–202.
- Messer, K., and H. White (1984). A note on computing the heteroskedasticity consistent covariance matrix using instrumental variable techniques. *Oxford Bulletin of Economics and Statistics* 46, 181–4.
- Murphy, A. (1996). Simple LM tests of mis-specification for ordered logit models. *Economics Letters* 52, 137–41.
- Orme, C. (1995). On the use of artificial regressions in certain microeconomic models. *Econometric Theory* 11, 290–305.
- Press, W.H., S.A. Teukolsky, W.T. Vetterling, and B.P. Flannery (1992). *Numerical Recipes in C*, 2nd edn., Cambridge: Cambridge University Press.
- Thomas, J. (1993). On testing the logistic assumption in binary dependent variable models. *Empirical Economics* 18, 381–92.
- White, H. (1980). A heteroskedasticity-consistent covariance matrix estimator and a direct test for heteroskedasticity. *Econometrica* 48, 817–38.
- Wooldridge, J.M. (1990). A unified approach to robust, regression-based specification tests. *Econometric Theory* 6, 17–43.
- Wooldridge, J.M. (1991). On the application of robust, regression-based diagnostics to models of conditional means and conditional variances. *Journal of Econometrics* 47, 5–46.

CHAPTER TWO

# General Hypothesis Testing

*Anil K. Bera and Gamini Premaratne\**

## 1 INTRODUCTION

The history of statistical hypothesis testing is, indeed, very long. Neyman and Pearson (1933) traced its origin to Bayes (1763). However, systematic applications of hypothesis testing began only after the publication of Karl Pearson's (1900) goodness-of-fit test, which is regarded as one of the 20 most important scientific breakthroughs in this century. In terms of the development of statistical methods, Ronald Fisher took up where Pearson left off. Fisher (1922) can be regarded as the analytical beginning of statistical methods. In his paper Fisher advocated the use of maximum likelihood estimation and provided the general theory of parametric statistical inference. In order to develop various statistical techniques, Fisher (1922) also introduced such basic concepts as consistency, efficiency, and sufficiency that are now part of our day-to-day vocabulary. Fisher, however, was not particularly interested in testing *per se*, and he occupied himself mostly in solving problems of estimation and sampling distributions. Neyman and Pearson (1928) suggested the likelihood ratio (LR) test, but that was mostly based on intuitive arguments. The foundation of the theory of hypothesis testing was laid by Neyman and Pearson (1933), and for the first time the concept of "optimal test" was introduced through the analysis of "power function." The result was the celebrated Neyman-Pearson (N-P) lemma. This lemma provides a way to find the most powerful (MP) and uniformly most powerful (UMP) tests. Neyman and Pearson (1936) generalized the basic N-P lemma to restrict optimal tests to suitable subclasses since the UMP test rarely exists. On the basis of Neyman-Pearson's foundation in testing, several general test principles were gradually developed, such as Neyman's (1937) smooth test, Wald (1943), Rao's (1948) score test and Neyman's (1959)  $C(\alpha)$  test. During the last four decades no new fundamental test principle has emerged. However, econometricians have produced a

large number of general test procedures such as those in Hausman (1978), Newey (1985), Tauchen (1985), and White (1982). Also, simultaneously, econometricians applied the basic test principles, most notably Rao's score test, and developed model diagnostic and evaluation techniques for the basic assumptions such as serial independence, homoskedasticity, and normality for the regression models, and these procedures are now routinely used in applied econometrics.

The aim of this chapter is very modest. Our main purpose is to explain the basic test principles with examples in a simple way and discuss how they have been used by econometricians to develop procedures to suit their needs. In the next section, we review the general test procedures suggested in the statistics literature. In Section 3, we discuss some standard tests in econometrics and demonstrate how they are linked to some basic principles. The last section offers some concluding remarks. At the outset we should state that there is nothing original about the material covered in this chapter. Students of econometrics are sometimes not aware of the origin of many of the tests they use. Here, we try to provide intuitive descriptions of some test principles with simple examples and to show that various econometric model evaluations and diagnostic procedures have their origins in some of the basic statistical test principles. Much of what we cover can be found, though in a scattered fashion, in Lehmann (1986, 1999), Godfrey (1988), Bera and Ullah (1991), Davidson and MacKinnon (1993), Gouriéroux and Monfort (1995), Bera (2000), Bera and Billias (2000) and Rao (2000).

## 2 SOME TEST PRINCIPLES SUGGESTED IN THE STATISTICS LITERATURE

We start by introducing some notation and concepts. Suppose we have  $n$  independent observations  $y_1, y_2, \dots, y_n$  on a random variable  $Y$  with density function  $f(y; \theta)$ , where  $\theta$  is a  $p \times 1$  parameter vector with  $\theta \in \Theta \subset \Re^p$ . It is assumed that  $f(y; \theta)$  satisfies the regularity conditions stated in Rao (1973, p. 364) and Serfling (1980, p. 144). The likelihood function is given by

$$L(\theta, y) \equiv L(\theta) = \prod_{i=1}^n f(y_i; \theta), \quad (2.1)$$

where  $y = (y_1, y_2, \dots, y_n)'$  denotes the sample.

Suppose we are interested in testing a simple hypothesis  $H_0 : \theta = \theta_0$  against another simple hypothesis  $H_1 : \theta = \theta_1$ . Let  $S$  denote the sample space. In standard test procedures,  $S$  is partitioned into two regions,  $\omega$  and its compliment  $\omega^c$ . We reject the null hypothesis if the sample  $y \in \omega$ ; otherwise, we do not reject  $H_0$ . Let us define a test function  $\phi(y)$  as:  $\phi(y) = 1$  when we reject  $H_0$ , and  $\phi(y) = 0$  when we do not reject  $H_0$ . Then,

$$\begin{aligned} \phi(y) &= 1 \quad \text{if } y \in \omega \\ &= 0 \quad \text{if } y \in \omega^c. \end{aligned} \quad (2.2)$$

Therefore, the probability of rejecting  $H_0$  is given by

$$\gamma(\theta) = E_{\theta}[\phi(y)] = \int \phi(y)L(\theta)dy, \quad (2.3)$$

where  $E_{\theta}$  denotes expectation when  $f(y; \theta)$  is the probability density function. Type-I and type-II error probabilities are given by, respectively,

$$\Pr(\text{Reject } H_0 | H_0 \text{ is true}) = E_{\theta_0}[\phi(y)] = \gamma(\theta_0) \quad (2.4)$$

and

$$\Pr(\text{Accept } H_0 | H_1 \text{ is true}) = E_{\theta_1}[1 - \phi(y)] = 1 - \gamma(\theta_1). \quad (2.5)$$

Note that  $\gamma(\theta_1)$  is the probability of making a correct decision and it is called the power of the test. An ideal situation would be if we could simultaneously minimize  $\gamma(\theta_0)$  and maximize  $\gamma(\theta_1)$ . However, because of the inverse relationship between the type-I and type-II error probabilities and also because Neyman and Pearson (1933) wanted to avoid committing the error of first kind and did not want  $\gamma(\theta_0)$  to exceed a preassigned value, they suggested maximizing the power  $\gamma(\theta_1)$  after keeping  $\gamma(\theta_0)$  at a low value, say  $\alpha$ , that is, maximize  $E_{\theta_0}[\phi(y)]$  subject to  $E_{\theta_0}[\phi(y)] = \alpha$  [see also Neyman, 1980, pp. 4–5]. A test  $\phi^*(y)$  is called a most powerful (MP) test if  $E_{\theta_1}[\phi^*(y)] \geq E_{\theta_1}[\phi(y)]$  for any  $\phi(y)$  satisfying  $E_{\theta_1}[\phi(y)] = \alpha$ . If an MP test maximizes power uniformly in  $\theta \in \Theta_1 \subset \Theta$ , the test is called a uniformly most powerful (UMP) test. A UMP test, however, rarely exists, and therefore, it is necessary to restrict optimal tests to a suitable subclass by requiring the test to satisfy other criteria such as local optimality, unbiasedness, and invariance, etc. For the N-P Lemma, there is only one side condition, that is, the size ( $\alpha$ ) of the test. Once the test is restricted further, in addition to the size there will be more than one side condition, and one must use the generalized N-P lemma given in Neyman and Pearson (1936).

## 2.1 Neyman–Pearson generalized lemma and its applications

The lemma can be stated as follows:

Let  $g_1, g_2, \dots, g_m, g_{m+1}$  be integrable functions and  $\phi$  be a test function over  $S$  such that  $0 \leq \phi \leq 1$ , and

$$\int \phi g_i dy = c_i \quad i = 1, 2, \dots, m, \quad (2.6)$$

where  $c_1, c_2, \dots, c_m$  are given constants. Further, let there exist a  $\phi^*$  and constants  $k_1, k_2, \dots, k_m$  such that  $\phi^*$  satisfies (2.6), and

$$\begin{aligned}\phi^* &= 1 \quad \text{if} \quad g_{m+1} > \sum_{i=1}^m k_i g_i \\ &= 0 \quad \text{if} \quad g_{m+1} < \sum_{i=1}^m k_i g_i\end{aligned}\tag{2.7}$$

then,

$$\int \phi^* g_{m+1} dy \geq \int \phi g_{m+1} dy.\tag{2.8}$$

For a proof, see, for example, Rao (1973, pp. 446–8) and Lehmann (1986, pp. 96–101). Several results can be derived from the above lemma as discussed below.

### N-P LEMMA AND THE LIKELIHOOD RATIO TEST

To obtain the basic N-P lemma, we put  $m = 1$ ,  $g_1 = L(\theta_0)$ ,  $g_2 = L(\theta_1)$ ,  $c_1 = \alpha$  and  $k_1 = k$ . Then among all test functions  $\phi(y)$  having size  $\alpha$ , that is,

$$\int \phi(y) L(\theta_0) dy = \alpha,\tag{2.9}$$

the function  $\phi^*(y)$  defined as

$$\begin{aligned}\phi^*(y) &= 1 \quad \text{when} \quad L(\theta_1) > kL(\theta_0) \\ &= 0 \quad \text{when} \quad L(\theta_1) < kL(\theta_0),\end{aligned}$$

and also satisfying (2.9), we will have

$$\int \phi^*(y) L(\theta_1) dy \geq \int \phi(y) L(\theta_1) dy,\tag{2.10}$$

that is,  $\phi^*(y)$  will provide the MP test. Therefore, in terms of critical region,

$$\omega = \left\{ y \left| \frac{L(\theta_1)}{L(\theta_0)} > k \right. \right\},\tag{2.11}$$

where  $k$  is such that  $\Pr\{\omega | H_0\} = \alpha$ , is the MP critical region.

The N-P lemma also provides the logical basis for the LR test. To see this, consider a general form of null hypothesis,  $H_0 : h(\theta) = c$  where  $h(\theta)$  is an  $r \times 1$  vector function of  $\theta$  with  $r \leq p$  and  $c$  a known constant vector. It is assumed that  $H(\theta) = \frac{\partial h(\theta)}{\partial \theta}$  has full column rank, that is,  $\text{rank}[H(\theta)] = r$ . We denote the maximum likelihood estimator (MLE) of  $\theta$  by  $\hat{\theta}$ , and by  $\tilde{\theta}$ , the restricted MLE of  $\theta$ , that is,

$\hat{\theta}$  is obtained by maximizing the loglikelihood function  $l(\theta) = \ln L(\theta)$  subject to the restriction  $h(\theta) = c$ . Neyman and Pearson (1928) suggested their LR test as

$$\text{LR} = 2 \left[ \ln \frac{L(\hat{\theta})}{L(\tilde{\theta})} \right] = 2[l(\hat{\theta}) - l(\tilde{\theta})]. \quad (2.12)$$

Their suggestion did not result from any search procedure satisfying an optimality criterion, and it was purely based on intuitive grounds and Fisher's (1922) likelihood principle.<sup>1</sup> Comparing the MP critical region in (2.11) with (2.12) we can see the logical basis of the LR test.

### LOCALLY MP (LMP) AND RAO'S (1948) SCORE TESTS

Let us consider a simple case, say  $p = 1$  and test  $H_0 : \theta = \theta_0$ . Assuming that the power function  $\gamma(\theta)$  in (2.3) admits Taylor series expansion, we have

$$\gamma(\theta) = \gamma(\theta_0) + (\theta - \theta_0)\gamma'(\theta_0) + \frac{(\theta - \theta_0)^2}{2}\gamma''(\theta^*), \quad (2.13)$$

where  $\theta^*$  is a value in between  $\theta$  and  $\theta_0$ . If we consider *local* alternatives of the form  $\theta = \theta_0 + \delta/\sqrt{n}$ ,  $0 < \delta < \infty$ , the third term will be of order  $O(n^{-1})$ . To obtain highest power, we need to maximize,

$$\gamma'(\theta_0) = \left. \frac{\partial}{\partial \theta} \gamma(\theta) \right|_{\theta=\theta_0} = \int \phi(y) \frac{\partial}{\partial \theta} L(\theta_0) dy, \quad (2.14)$$

for  $\theta > \theta_0$ . Therefore, for an LMP test of size  $\alpha$  we should have

$$\int \phi(y) L(\theta_0) dy = \alpha,$$

and maximize  $\int \phi(y) \frac{\partial}{\partial \theta} L(\theta_0) dy$ . In the N-P generalized lemma, let us put  $m = 1$ ,  $g_1 = L(\theta_0)$ ,  $g_2 = \frac{\partial}{\partial \theta} L(\theta_0)$ ,  $c_1 = \alpha$  and  $k_1 = k$ . Then from (2.7) and (2.8), the LMP test will have critical region

$$\frac{\partial}{\partial \theta} L(\theta_0) > k L(\theta_0)$$

or

$$\frac{\partial}{\partial \theta} \ln L(\theta_0) = \left. \frac{\partial}{\partial \theta} l(\theta) \right|_{\theta=\theta_0} > k. \quad (2.15)$$

The quantity  $s(\theta) = \partial l(\theta)/\partial \theta$  is known as the score function. The above result was first discussed in Rao and Poti (1946), who stated that an LMP test for  $H_0 : \theta = \theta_0$  is given by

$$l_1 s(\theta_0) > l_2, \quad (2.16)$$

where  $l_2$  is so determined that the size of test is equal to a preassigned value  $\alpha$  with  $l_1$  as  $+1$  or  $-1$ , respectively, for alternative  $\theta > \theta_0$  and  $\theta < \theta_0$ . Test criterion (2.16) is a precursor to Rao's score (RS) or the Lagrange multiplier (LM) test that has been very useful to econometrics for developing various model diagnostic procedures, as we will discuss later.

The LMP test can also be obtained directly from the N-P lemma (2.11). By expanding  $L(\theta_1)$  around  $\theta_0$  as

$$L(\theta_1) = L(\theta_0) + (\theta_1 - \theta_0) \frac{\partial}{\partial \theta} L(\theta^*), \quad (2.17)$$

where  $\theta^*$  is in between  $\theta_0$  and  $\theta_1$ . Therefore, according to (2.11) we reject  $H_0$  if

$$1 + (\theta_1 - \theta_0) \frac{1}{L(\theta_0)} \cdot \frac{\partial}{\partial \theta} L(\theta^*) > k. \quad (2.18)$$

Now as  $\theta_1 \rightarrow \theta_0$ , it is clear that this critical region reduces to that of (2.15) [see Gouriéroux and Monfort, 1995, p. 32].

**Example 1.** As an example of an LMP test consider testing for the median of a Cauchy distribution with probability density

$$f(y; \theta) = \frac{1}{\pi} \frac{1}{1 + (y - \theta)^2} \quad -\infty < y < \infty. \quad (2.19)$$

We test  $H_0 : \theta = 0$  against  $H_1 : \theta > 0$ . For simplicity, take  $n = 1$ , and therefore, we reject  $H_0$  for large values of

$$\left. \frac{\partial \ln f(y; \theta)}{\partial \theta} \right|_{\theta=0} = \frac{2y}{1 + y^2}. \quad (2.20)$$

As constructed, this will provide an optimal test for  $\theta$  close to zero (*local* alternatives). Now suppose  $\theta \gg 0$ , and we can see that as  $\theta \rightarrow \infty$ ,  $2y/(1 + y^2) \rightarrow 0$ . Therefore, for distant alternatives the power of the test will be zero.

Therefore, what works for local alternatives may not work at all for not-so-local alternatives. The situation, however, is not so grim universally. Consider the following standard example.

**Example 2.** Let  $Y \sim N(\mu, 1)$  and test  $H_0 : \mu = 0$  against  $H_1 : \mu > 0$  based on a sample of size 1. We have

$$\left. \frac{\partial \ln f(y; \mu)}{\partial \mu} \right|_{\mu=0} = y. \quad (2.21)$$

Therefore, we reject  $H_0$  if  $y > k$ , where  $k = Z_\alpha$ , the upper  $\alpha$  percent cut-off point of standard normal. The power of this test is  $1 - \Phi(Z_\alpha - \mu)$ , where  $\Phi(\cdot)$  is the distribution function of the standard normal density. And as  $\mu \rightarrow \infty$ , the power of the test goes to 1. Therefore, the test  $y > Z_\alpha$  is not only LMP, it is also uniformly most powerful (UMP) for all  $\mu > 0$ .

Now let us consider what happens to the power of this test when  $\mu < 0$ . The power function  $\Pr(y > Z_\alpha | \mu < 0)$  still remains  $1 - \Phi(Z_\alpha - \mu)$ , but it is now less than  $\alpha$ , the size of the test. Therefore, the test is not MP for all  $\mu \neq 0$ . To get an MP test for *two-sided* alternatives, we need to add unbiasedness as an extra condition in our requirements.

### LOCALLY MOST POWERFUL UNBIASED (LMPU) TEST

A test  $\phi(y)$  of size  $\alpha$  is unbiased for  $H_0 : \theta \in \Omega_0$  against  $H_1 : \theta \in \Omega_1$  if  $E_\theta[\phi(y)] \leq \alpha$  for  $\theta \in \Omega_0$  and  $E_\theta[\phi(y)] \geq \alpha$  for  $\theta \in \Omega_1$ . Suppose we want to find an LMPU test for testing  $H_0 : \theta = \theta_0$  against  $H_1 : \theta \neq \theta_0$ . By expanding the power function  $\gamma(\theta)$  in (2.3) around  $\theta = \theta_0$  for local alternatives, we have

$$\begin{aligned}\gamma(\theta) &= \gamma(\theta_0) + (\theta - \theta_0)\gamma'(\theta_0) + \frac{(\theta - \theta_0)^2}{2}\gamma''(\theta_0) + o(n^{-1}) \\ &= \alpha + \frac{(\theta - \theta_0)^2}{2}\gamma''(\theta_0) + o(n^{-1}).\end{aligned}\quad (2.22)$$

Unbiasedness requires that the “power” should be minimum at  $\theta = \theta_0$ , and, hence,  $\gamma'(\theta_0) = 0$ . To maximize the local power, we, therefore, need to maximize  $\gamma''(\theta_0)$  for both  $\theta > \theta_0$  and  $\theta < \theta_0$ , and this leads to the LMPU test. Neyman and Pearson (1936, p. 9) called the corresponding critical region “type-A region,” and this requires maximization of  $\gamma''(\theta_0)$  subject to two side-conditions  $\gamma(\theta_0) = \alpha$  and  $\gamma'(\theta_0) = 0$ . In the N-P generalized lemma, let us put  $m = 2$ ,  $c_1 = 0$ ,  $c_2 = \alpha$ ,  $g_1 = \frac{\partial l(\theta_0)}{\partial \theta}$ ,  $g_2 = L(\theta_0)$  and  $g_3 = \frac{\partial^2 l(\theta_0)}{\partial \theta^2}$ , then from (2.7) and (2.8), the optimal test function  $\phi^* = 1$  if

$$\frac{\partial^2 L(\theta_0)}{\partial \theta^2} > k_1 \frac{\partial L(\theta_0)}{\partial \theta} + k_2 L(\theta_0) \quad (2.23)$$

and  $\phi^* = 0$ , otherwise. Critical region (2.23) can be expressed in terms of the derivatives of the loglikelihood function as

$$\frac{\partial^2 l(\theta_0)}{\partial \theta^2} + \left[ \frac{\partial l(\theta_0)}{\partial \theta} \right]^2 > k_1 \frac{\partial l(\theta_0)}{\partial \theta} + k_2. \quad (2.24)$$

In terms of the score function  $s(\theta) = \partial l(\theta)/\partial \theta$  and its derivative  $s'(\theta)$ , (2.24) can be written as

$$s'(\theta_0) + [s(\theta_0)]^2 > k_1 s(\theta_0) + k_2. \quad (2.25)$$

**Example 2.** (*continued*) For this example, consider now testing  $H_0 : \mu = 0$  against  $H_1 : \mu \neq 0$ . It is easy to see that  $s(\theta_0) = y$ ,  $s'(\theta_0) = -1$ . Therefore, a uniformly most powerful unbiased test (UMPU) will reject  $H_0$  if

$$y^2 + k'_1 y + k'_2 > 0$$

or

$$y < k''_1 \quad \text{and} \quad y > k''_2,$$

where  $k'_1$ ,  $k'_2$ ,  $k''_1$ , and  $k''_2$  are some constants determined from satisfying the size and unbiasedness conditions. After some simplification, the LMPU principle leads to a symmetric critical region of the form  $y < -Z_{\alpha/2}$  and  $y > Z_{\alpha/2}$ .

In many situations,  $s'(\theta)$  can be expressed as a linear function of the score  $s(\theta)$ . For those cases, LMPU tests will be based on the score function only, just like the LMP test in (2.15). Also for certain test problems  $s(\theta_0)$  vanishes, then from (2.25) we see that an LMPU test can be constructed using the second derivative of the loglikelihood function.

**Example 3.** (Godfrey, 1988, p. 92). Let  $y_i \sim N(0, (1 + \theta^2 z_i))$ ,  $i = 1, 2, \dots, n$ , where  $z_i$ s are given positive constants. We are interested in testing  $H_0 : \theta = 0$ , that is,  $y_i$  has constant variance. The loglikelihood function and the score function are, respectively, given by

$$l(\theta) = \text{const} - \frac{1}{2} \sum_{i=1}^n \ln(1 + \theta^2 z_i) - \frac{1}{2} \sum_{i=1}^n y_i^2 / (1 + \theta^2 z_i), \quad (2.26)$$

and

$$s(\theta) = \frac{\partial}{\partial \theta} l(\theta) = -\theta \sum_{i=1}^n \left[ \frac{z_i}{(1 + \theta^2 z_i)} - \frac{z_i y_i^2}{(1 + \theta^2 z_i)^2} \right]. \quad (2.27)$$

It is clear that  $s(\theta) = 0$  at  $H_0 : \theta = 0$ . However,

$$s'(\theta) \Big|_{\theta=0} = \frac{\partial s(\theta)}{\partial \theta} \Bigg|_{\theta=0} = \frac{1}{2} \sum_{i=1}^n z_i (y_i^2 - 1) \quad (2.28)$$

and from (2.25), the LMPU test could be based on the above quantity. In fact, it can be shown that (Godfrey, 1988, p. 92)

$$\frac{\sum_{i=1}^n z_i (y_i^2 - 1)}{\sqrt{2 \sum_{i=1}^n z_i^2}} \xrightarrow{d} N(0, 1). \quad (2.29)$$

where  $\xrightarrow{d}$  denotes convergence in distribution.

### NEYMAN'S SMOOTH TEST

Pearson (1900) suggested his goodness-of-fit test to see whether an assumed probability model adequately described the data at hand. Suppose we divide data into  $j$ -classes and the probability of the  $j$ th class is  $\theta_j$ ,  $j = 1, 2, \dots, p$ , and  $\sum_{j=1}^p \theta_j = 1$ . Suppose according to the assumed probability model  $\theta_j = \theta_{j0}$ ; therefore, our null hypothesis could be stated as  $H_0 : \theta_j = \theta_{j0}$ ,  $j = 1, 2, \dots, p$ . Let  $n_j$  denote the observed frequency of the  $j$ th class, with  $\sum_{j=1}^p n_j = n$ . Pearson (1900) suggested the goodness-of-fit statistic

$$P = \sum_{j=1}^p \frac{(n_j - n\theta_{j0})^2}{n\theta_{j0}} = \sum_{j=1}^p \frac{(O_j - E_j)^2}{E_j}, \quad (2.30)$$

where  $O_j$  and  $E_j$  denote, respectively, the observed and expected frequencies for the  $j$ th class.

Neyman's (1937) criticism to Pearson's test was that (2.30) does not depend on the order of positive and negative differences  $(O_j - E_j)$ . Neyman (1980) gives an extreme example represented by two cases. In the first, the signs of the consecutive differences  $(O_j - E_j)$  are not the same, and in the other, there is run of, say, a number of "negative" differences, followed by a sequence of "positive" differences. These two possibilities might lead to similar values of  $P$ , but Neyman (1937, 1980) argued that in the second case the goodness-of-fit should be more in doubt, even if the value of  $P$  happens to be small.

Suppose we want to test the null hypothesis ( $H_0$ ) that  $f(y; \theta)$  is the true density function for the random variable  $Y$ . The specification of  $f(y; \theta)$  will be *different* depending on the problem on hand. Let us denote the alternative hypothesis as  $H_1 : Y \sim g(y)$ . Neyman (1937) transformed any hypothesis-testing problem of this type to testing only *one kind of hypothesis*. Let  $z = F(y)$  denote the distribution function of  $Y$ , then the density of the random variable  $Z$  is given by

$$h(z) = g(y) \frac{dy}{dz} = \frac{g(y)}{f(y; \theta)}, \quad (2.31)$$

when  $H_0 : Y \sim f(y; \theta)$ , then

$$h(z) = 1 \quad 0 < z < 1. \quad (2.32)$$

Therefore, testing  $H_0$  is equivalent to testing whether  $Z$  has uniform distribution in the interval  $(0, 1)$ , irrespective of the specification of  $f(y; \theta)$ . As for the specific alternative to the uniform distribution, Neyman (1937) suggested a *smooth* class. By smooth alternatives Neyman meant those densities that have few intersections with the null density function and that are close to the null. He specified the alternative density as

$$h(z) = C(\delta) \exp \left[ \sum_{j=1}^r \delta_j \pi_j(z) \right], \quad (2.33)$$

where  $C(\delta)$  is the constant of integration that depends on the  $\delta_j$  values, and  $\pi_j(z)$  are orthogonal polynomials satisfying

$$\begin{aligned} \int_0^1 \pi_j(z) \pi_k(z) dy &= 1 \quad \text{for } j = k \\ &= 0 \quad \text{for } j \neq k. \end{aligned} \quad (2.34)$$

Under the hypothesis  $H_0 : \delta_1 = \delta_2 = \dots = \delta_r = 0$ ,  $C(\delta) = 1$  and  $h(z)$  in (2.33) reduces to the uniform density (2.32). Using the generalized N-P lemma, Neyman (1937) derived a locally most powerful symmetric unbiased test for  $H_0$ , and the test statistic is given by

$$\psi_r^2 = \sum_{j=1}^r \frac{1}{n} \left[ \sum_{i=1}^n \pi_j(z_i) \right]^2. \quad (2.35)$$

The test is symmetric in the sense that the asymptotic power of the test depends only on the "distance"  $\sum_{j=1}^r \delta_j^2$  between the null and alternative hypotheses.

## 2.2 Tests based on score function and Wald's test

We have already discussed Rao's (1948) score principle of testing as an LMP test in (2.15) for the scalar parameter  $\theta(p = 1)$ . For the  $p \geq 2$  case, there will be scores for each individual parameter, and the problem is to combine them in an "optimal" way. Let  $H_0 : \theta = \theta_0$ , where now  $\theta = (\theta_1, \theta_2, \dots, \theta_p)'$  and  $\theta_0 = (\theta_{10}, \theta_{20}, \dots, \theta_{p0})'$ , and the (local) alternative hypothesis be as  $H_1 : \theta = \theta_\delta$ , where  $\theta_\delta = (\theta_{10} + \delta_1, \theta_{20} + \delta_2, \dots, \theta_{p0} + \delta_p)'$ . The proportionate change in the loglikelihood function for moving from  $\theta_0$  to  $\theta_\delta$  is given by  $\delta's(\theta_0)$ , where  $\delta = (\delta_1, \delta_2, \dots, \delta_p)'$  and  $s(\theta_0)$  is the score function evaluated at  $\theta = \theta_0$ . Let us define the information matrix as

$$I(\theta) = -E \left[ \frac{\partial^2 l(\theta)}{\partial \theta \partial \theta'} \right]. \quad (2.36)$$

Then, the asymptotic variance of  $\delta's(\theta_0)$  is  $\delta'I(\theta_0)\delta$ ; and, if  $\delta'$  were known, a test could be based on

$$\frac{[\delta's(\theta_0)]^2}{\delta'I(\theta_0)\delta}, \quad (2.37)$$

which under  $H_0$  will be asymptotically distributed as  $\chi_1^2$ . To eliminate the  $\delta$ 's and to obtain a linear function that would yield maximum discrimination, Rao (1948) maximized (2.37) with respect to  $\delta$  and obtained<sup>2</sup>

$$\sup_{\delta} \frac{[\delta's(\theta_0)]^2}{\delta'I(\theta_0)\delta} = s(\theta_0)' I(\theta_0)^{-1} s(\theta_0) \quad (2.38)$$

with optimal value  $\delta = I(\theta_0)^{-1}s(\theta_0)$ . In a sense,  $\delta = I(\theta_0)^{-1}s(\theta_0)$  signals the *optimal* direction of the alternative hypothesis that we should consider. For example, when  $p = 1$ ,  $\delta = +1$  or  $-1$ , as we have seen in (2.16). Asymptotically, under the null, the statistic in (2.38) follows a  $\chi_p^2$  distribution in contrast to (2.37), which follows  $\chi_1^2$ . When the null hypothesis is composite, like  $H_0 : h(\theta) = c$  with  $r \leq p$  restrictions, the general form of Rao's score (RS) statistic is

$$\text{RS} = s(\tilde{\theta})' I(\tilde{\theta})^{-1} s(\tilde{\theta}), \quad (2.39)$$

where  $\tilde{\theta}$  is the restricted MLE of  $\theta$ . Under  $H_0 : \text{RS} \xrightarrow{d} \chi_r^2$ . Therefore, we observe *two* optimality principles behind the RS test; first, in terms of the LMP test as given in (2.15), and second, in deriving the "optimal" direction for the multi-parameter case.

Rao (1948) suggested the score test as an alternative to the Wald (1943) statistic, which for testing  $H_0 : h(\theta) = c$  is given by

$$W = (h(\hat{\theta}) - c)' [H(\hat{\theta})' I(\hat{\theta})^{-1} H(\hat{\theta})]^{-1} (h(\hat{\theta}) - c). \quad (2.40)$$

Rao (1948, p. 53) stated that his test "besides being simpler than Wald's has some theoretical advantages," such as invariance under transformation of parameters. Rao (2000) recollects the motivation and background behind the development of the score test.

The three statistics LR, W, and RS given, respectively in (2.12), (2.40), and (2.39) are referred to as the "holy trinity." We can look at these statistics in terms of different measures of distance between the null and alternative hypotheses. When the null hypothesis is true, we would expect the restricted and unrestricted MLEs of  $\theta$ ,  $\hat{\theta}$ , and  $\tilde{\theta}$  to be close, and likewise the loglikelihood functions. Therefore the LR statistic measures the distance through the loglikelihood function and is based on the difference  $I(\hat{\theta}) - I(\tilde{\theta})$ . To see the intuitive basis of the score test, note that  $s(\hat{\theta})$  is zero by construction, and we should expect  $s(\tilde{\theta})$  to be close to zero if  $H_0$  is true. And hence the RS test exploits the distance through the score function  $s(\theta)$  and can be viewed as being based on  $s(\tilde{\theta}) - s(\hat{\theta})$ . Lastly, the W test considers the distance directly in terms of  $h(\theta)$  and is based on  $[h(\hat{\theta}) - c] - [h(\tilde{\theta}) - c]$ , where by construction  $h(\hat{\theta}) = c$ . This reveals a duality between the Wald and score tests. At the unrestricted MLE  $\hat{\theta}$ ,  $s(\hat{\theta}) = 0$ , and the Wald test checks whether  $h(\hat{\theta})$  is away from  $c$ . On the other hand, at the restricted MLE  $\tilde{\theta}$ ,  $h(\tilde{\theta}) = c$  by construction, and the score test verifies whether  $s(\tilde{\theta})$  is far from a null vector.<sup>3</sup>

**Example 4.** Consider a multinomial distribution with  $p$  classes and let the probability of an observation belonging to the  $j$ th class be  $\theta_j$ , so that  $\sum_{j=1}^p \theta_j = 1$ . Denote the frequency of  $j$ th class by  $n_j$  with  $\sum_{j=1}^p n_j = n$ . We are interested in testing  $H_0 : \theta_j = \theta_{j0}, j = 1, 2, \dots, p$ , where  $\theta_{j0}$ s are known constants. It can be shown that for this problem the score statistic is given by

$$s(\theta_0)' I(\theta_0)^{-1} s(\theta_0) = \sum_{j=1}^p \frac{(n_j - n\theta_{j0})^2}{n\theta_{j0}}, \quad (2.41)$$

where  $\theta_0 = (\theta_{10}, \dots, \theta_{p0})'$ . Therefore, the RS statistic is the same as Pearson's  $P$  given in (2.30). It is quite a coincidence that Pearson (1900) suggested a score test mostly based on intuitive grounds almost 50 years before Rao (1948). For this problem, the other two test statistics LR and W are given by

$$\text{LR} = 2 \sum_{i=1}^p n_j \ln \left( \frac{n_j}{n\theta_{j0}} \right) = \sum_{j=1}^p O_j \ln \left( \frac{O_j}{E_j} \right) \quad (2.42)$$

and

$$W = \sum_{j=1}^p \frac{(n_j - n\theta_{j0})^2}{n_j} = \sum_{j=1}^p \frac{(O_j - E_j)^2}{O_j}. \quad (2.43)$$

The equivalence of the score and Pearson's tests and their local optimality has not been fully recognized in the statistics literature. Many researchers considered the LR statistic to be superior to  $P$ . Asymptotically, both statistics are locally optimal and equivalent, and, in terms of finite sample performance,  $P$  performs better [see for example Rayner and Best, 1989, pp. 26–7].

The three tests LR, W, and RS are based on the (efficient) maximum likelihood estimates. When consistent (rather than efficient) estimators are used there is another attractive way to construct a score-type test, which is due to Neyman (1954, 1959). In the literature this is known as the  $C(\alpha)$ , or effective score or Neyman–Rao test. To follow Neyman (1959), let us partition  $\theta$  as  $\theta = [\theta_1', \theta_2]'$ , where  $\theta_2$  is a scalar and test  $H_0 : \theta_2 = \theta_{20}$ . Therefore,  $\theta_1$  is the nuisance parameter with dimension  $(p-1) \times 1$ . Neyman's fundamental contribution is the *derivation* of an asymptotically optimal test using consistent estimators of the nuisance parameters. He achieved this in two steps. First he started with a class of function  $g(y; \theta_1, \theta_2)$  satisfying regularity condition of Cramér (1946, p. 500).<sup>4</sup>

For simplicity let us start with a normed Cramér function, that is,  $g(y; \theta_1, \theta_2)$  has zero mean and unit variance. We denote  $\sqrt{n}$ -consistent estimator of  $\theta$  under  $H_0$  by  $\theta^+ = (\theta_1^+, \theta_{20})'$ . Neyman asked the question what should be the property of  $g(\cdot)$  such that replacing  $\theta$  by  $\theta^+$  in the test statistic would not make any difference asymptotically, and his Theorem 1 proved that  $g(\cdot)$  must satisfy

$$\text{Cov}[g(y; \theta_1, \theta_{20}), s_{1j}(y; \theta_1, \theta_{20})] = 0, \quad (2.44)$$

where  $s_{1j} = \frac{\partial l(\theta)}{\partial \theta_{1j}}$ , i.e. the score for  $j$ th component of  $\theta_1$ ,  $j = 1, 2, \dots, p-1$ . In other words, the function  $g(y; \theta)$  should be orthogonal to  $s_1 = \frac{\partial l(\theta)}{\partial \theta_1}$ . Starting from a normed Cramér function let us construct

$$\bar{g}(y; \theta_1, \theta_{20}) = g(y; \theta_1, \theta_{20}) - \sum_{j=1}^{p-1} b_j s_{1j}(\theta_1, \theta_{20}), \quad (2.45)$$

where  $b_j$ ,  $j = 1, 2, \dots, p-1$ , are the regression coefficients of regressing  $g(y; \theta_1, \theta_{20})$  on  $s_{11}, s_{12}, \dots, s_{1p-1}$ . Denote by  $\sigma^2(\theta_1, \theta_{20})$  the minimum variance of  $\bar{g}(y; \theta_1, \theta_{20})$ , and define

$$g^*(y; \theta_1, \theta_{20}) = \frac{\bar{g}(y; \theta_1, \theta_{20})}{\sigma(\theta_1, \theta_{20})}. \quad (2.46)$$

Note that  $g^*(y; \theta_1, \theta_{20})$  is also a normed Cramér function, and the covariance between  $g^*(y; \theta_1, \theta_{20})$  and  $s_{1j}(\theta_1, \theta_{20})$  is also zero,  $j = 1, 2, \dots, p - 1$ . Therefore, a class of  $C(\alpha)$  test can be based on  $Z_n(\theta_1^+, \theta_{20}) = \frac{1}{\sqrt{n}} \sum_{i=1}^n g^*(y_i; \theta_1^+, \theta_{20})$ . Condition (2.44) ensures that  $Z_n(\theta_1, \theta_{20}) - Z_n(\theta_1^+, \theta_{20}) = o_p(1)$ . The second step of Neyman was to find the starting function  $g(y; \theta)$  itself. Theorem 2 of Neyman (1959) states that under the sequence of local alternatives  $H_{1n} : \theta_2 = \theta_{20} + \frac{\delta}{\sqrt{n}}$ ,  $0 < \delta < \infty$ ,  $Z_n(\theta_1^+, \theta_{20})$  is asymptotically distributed as normal with mean  $\delta \rho \sigma_2$  and variance unity, where

$$\rho = \text{Corr} \left[ \bar{g}(y; \theta_1, \theta_{20}), \frac{\partial l(\theta)}{\partial \theta_2} \right], \quad (2.47)$$

and

$$\sigma_2 = \text{Var} \left[ \frac{\partial l(\theta)}{\partial \theta_2} \right].$$

The asymptotic power of the test will be purely guided by  $\rho$ , and to maximize the power we should select the function  $g(y; \theta)$  so that  $\rho = 1$ , that is, the optimal choice should be  $g(y; \theta) = \frac{\partial l(\theta)}{\partial \theta_2} = s_2(\theta_1, \theta_{20})$  say, score for the testing parameter  $\theta_2$ . Therefore, from (2.45), an asymptotically and locally optimal test should be based on the part of the score for the parameter tested that is orthogonal to the score for the nuisance parameter, namely,

$$s_2(\theta_1^+, \theta_{20}) - \sum_{j=1}^{p-1} b_j s_{1j}(\theta_1^+, \theta_{20}). \quad (2.48)$$

In (2.48)  $b_j$ ,  $j = 1, 2, \dots, p - 1$  are now regression coefficients of regressing  $s_2(\theta_1^+, \theta_{20})$  on  $s_{11}, s_{12}, \dots, s_{1p-1}$ , and we can express (2.48) as

$$s_2(\theta^+) - I_{21}(\theta^+) I_{11}^{-1}(\theta^+) s_1(\theta^+) = s_2^*(\theta), \text{ say,} \quad (2.49)$$

where  $I_{ij}(\theta)$  are the appropriate blocks of the information matrix  $I(\theta)$  corresponding to  $\theta_1$  and  $\theta_2$ .  $s_2^*(\theta)$  is called the *effective* score for  $\theta_2$  and its variance,  $I_{22}^*(\theta) = I_{22}(\theta) - I_{21}(\theta) I_{11}^{-1}(\theta) I_{12}(\theta)$  is termed effective information. Note that, since  $s_2^*(\theta)$  is the residual score obtained from running a regression of  $s_2(\theta)$  on  $s_1(\theta)$ , it will be orthogonal to the score for  $\theta_1$ . The operational form of Neyman's  $C(\alpha)$  test is

$$C(\alpha) = s_2^*(\theta^+)' I_{22}^*(\theta^+)^{-1} s_2^*(\theta^+). \quad (2.50)$$

Bera and Billias (2000) derived this test using the Rao (1948) framework [see equations (2.38) and (2.39)]. If we replace the  $\sqrt{n}$ -consistent estimator  $\theta^+$  by the restricted MLE, then  $s_2^*(\theta^+)$  and  $I_{22}^*(\theta^+)$  reduce to  $s_2(\tilde{\theta})$  and  $I_{22}(\tilde{\theta})$ , respectively, and the  $C(\alpha)$  test becomes the standard RS test.

**Example 5.** (Neyman, 1959) Let us consider testing  $H_0 : \theta_2 = 0$  in the following Cauchy density

$$f(y; \theta_1, \theta_2) = \frac{\theta_1}{\pi} \cdot \frac{1}{\theta_1^2 + (y - \theta_2)^2} \quad -\infty < y < \infty, \quad (2.51)$$

where  $\theta_1 > 0$  and  $-\infty < \theta_2 < \infty$ . It is easy to see that

$$\left. \frac{\partial \ln f(y; \theta_1, \theta_2)}{\partial \theta_2} \right|_{\theta_2=0} = \frac{2y}{\theta_1^2 + y^2}, \quad (2.52)$$

$I_{12}(\theta) = 0$  and  $I_{22}(\theta) = \frac{1}{2\theta_1^2}$  under  $H_0$ . Therefore,  $s_2^*(\theta) = s_2(\theta)$  and  $I_{22}^*(\theta) = I_{22}(\theta)$ . Hence the  $C(\alpha)$  statistic (2.50) based on a sample  $\mathbf{y} = (y_1, y_2, \dots, y_n)'$  is given by

$$C(\alpha) = \left( \sum_{i=1}^n \frac{2y_i}{\theta_1^{+2} + y_i^2} \right)^2 \Bigg/ \left( \frac{n}{2\theta_1^{+2}} \right). \quad (2.53)$$

For  $\theta_1^+$ , we can use any  $\sqrt{n}$ -consistent estimator such as the difference between the third and first sample quartiles. Since  $I_{21}(\theta) = 0$  under  $H_0$ , the RS test will have the same algebraic form (2.53), but for  $\theta_1$ , we need to use the restricted MLE,  $\hat{\theta}_1$ .

Neyman's  $C(\alpha)$  approach provides an attractive way to take into account the nuisance parameter  $\theta_1$ . Bera and Yoon (1993) applied this approach to develop tests that are valid under a *locally* misspecified model. They showed that replacing  $\theta_1$ , even by a null vector in the final form of the test, would lead to a valid test procedure.

This ends our discussion of the test principles proposed in the statistics literature. We have covered only those tests that have some relevance to testing and evaluating econometric models. In the next section we discuss some of their applications.

### 3 APPLICATIONS OF TEST PRINCIPLES TO ECONOMETRICS

Compared with statistics, econometrics is a relatively new discipline. Work in econometrics began in the 1920s mainly due to the initiatives of Ragnar Frisch and Jan Tinbergen. One of the first macro-econometric models was that of the Dutch economy built by Tinbergen in the mid-1930s. Possibly the first formal application of statistical tests was carried out by Tinbergen in his book, *Statistical Testing of Business-Cycle Theories*.<sup>5</sup> Tinbergen was commissioned from the Central Statistical Bureau of the Netherlands by the League of Nations to study trade cycles, and the outcome was the book published in 1939. Tinbergen used three different approaches to statistical inference, namely, Fisher's classical method, Frisch's bunch-map analysis and Tjalling Koopmans' time series approach.

John Maynard Keynes was skeptical about applying statistical techniques to economic data, as can be seen in his review of Tinbergen's book. It was left to Haavelmo (1944) to successfully defend the application of statistical methodologies to economic data within the framework of the *joint* probability distribution of variables. Trygve Haavelmo was clearly influenced by Jerzy Neyman,<sup>6</sup> and Haavelmo (1944) contains a seven page account of the Neyman–Pearson theory. He clearly stated the limitation of the standard hypothesis testing approach and explicitly mentioned that a test is, in general, constructed on the basis of a given fixed set of possible alternatives that he called a priori admissible hypotheses. And whenever this priori admissible set deviates from the data generating process, the test loses its optimality [for more on this see (Bera and Yoon, 1993) and (Bera, 2000)]. Haavelmo, however, did not himself formally apply the Neyman–Pearson theory to econometric testing problems. That was left to Anderson (1948) and Durbin and Watson (1950).

### 3.1 The Neyman–Pearson lemma and the Durbin–Watson test

The *first* formal specification test in econometrics, the Durbin–Watson (DW) (1950) test for autocorrelation in the regression model has its foundation in the UMP test principle via a theorem of Anderson (1948). Most econometrics textbooks provide a detail discussion of the DW test but do not mention its origin. Let us consider the standard linear regression model with autocorrelated errors:

$$y_t = x'_t \beta + \varepsilon_t \quad (2.54)$$

$$\varepsilon_t = \rho \varepsilon_{t-1} + u_t, \quad (2.55)$$

where  $y_t$  is the  $t$ th observation on the dependent variable,  $x_t$  is the  $t$ th observation on  $k$  strictly exogenous variables,  $|\rho| < 1$  and  $u_t \sim \text{iidN}(0, \sigma^2)$ ,  $t = 1, 2, \dots, n$ . The problem is testing  $H_0 : \rho = 0$ . Using the N–P lemma, Anderson (1948) showed that UMP tests for serial correlation can be obtained against one-sided alternatives.<sup>7</sup> A special case of Anderson's lemma is as follows:

If the probability density of  $\varepsilon$  can be written in the form,

$$f(\varepsilon) = \text{const. exp} \left[ -\frac{1}{2\sigma^2} \{(1 + \rho^2)\varepsilon'\varepsilon - 2\rho\varepsilon'D\varepsilon\} \right], \quad (2.56)$$

where  $\varepsilon = (\varepsilon_1, \varepsilon_2, \dots, \varepsilon_n)'$  and columns of  $X = (x_1, x_2, \dots, x_n)$  are generated by  $k$  eigen-vectors of  $D$ , the UMP test of  $H_0 : \rho = 0$  against  $H_1 : \rho > 0$  is given by  $a > a_0$ , where

$$a = \frac{\hat{\varepsilon}'D\hat{\varepsilon}}{\hat{\varepsilon}'\hat{\varepsilon}}, \quad (2.57)$$

and  $a_0$  is such that  $\Pr[a > a_0 | \rho = 0] = \alpha$ , the size of the test. Here  $\hat{\epsilon} = (\hat{\epsilon}_1, \hat{\epsilon}_2, \dots, \hat{\epsilon}_n)'$  with  $\hat{\epsilon}_t = y_t - x_t \hat{\beta}$  and  $\hat{\beta}$  is the ordinary least squares (OLS) residual vector.

For the model given in (2.54) and (2.55), the probability distribution of  $\epsilon$ , that is, the likelihood function is given by<sup>8</sup>

$$\text{const. exp} \left[ -\frac{1}{2\sigma^2} \left\{ (1 + \rho^2) \epsilon' \epsilon - \rho^2 (\epsilon_1^2 + \epsilon_n^2) - 2\rho \sum_{t=2}^n \epsilon_t \epsilon_{t-1} \right\} \right]. \quad (2.58)$$

Comparing (2.56) and (2.58), we see that the latter cannot be written in the form of the former. Durbin and Watson (1950) approached the problem from the opposite direction and selected a form of  $D$  in such a way that (2.56) becomes "close" to (2.58). They chose  $D = I_n - \frac{1}{2}A$ , where

$$A = \begin{bmatrix} 1 & -1 & 0 & 0 & \cdots & 0 & 0 & 0 \\ -1 & 2 & -1 & 0 & \cdots & 0 & 0 & 0 \\ 0 & -1 & 2 & -1 & \cdots & 0 & 0 & 0 \\ \vdots & \vdots & \vdots & \vdots & \cdots & \vdots & \vdots & \vdots \\ 0 & 0 & 0 & 0 & \cdots & -1 & 2 & -1 \\ 0 & 0 & 0 & 0 & \cdots & 0 & -1 & 1 \end{bmatrix} \quad (2.59)$$

so that

$$\begin{aligned} \epsilon' D \epsilon &= \sum_{t=1}^n \epsilon_t^2 - \frac{1}{2} \sum_{t=2}^n (\epsilon_t - \epsilon_{t-1})^2 \\ &= \frac{1}{2} (\epsilon_1^2 + \epsilon_n^2) + \sum_{t=2}^n \epsilon_t \epsilon_{t-1}. \end{aligned} \quad (2.60)$$

Then, the density (2.56) reduces to

$$f(\epsilon) = \text{const. exp} \left[ -\frac{1}{2\sigma^2} \left\{ (1 + \rho^2) \epsilon' \epsilon - \rho (\epsilon_1^2 + \epsilon_n^2) - 2\rho \sum_{t=2}^n \epsilon_t \epsilon_{t-1} \right\} \right]. \quad (2.61)$$

Now the only difference between the likelihood function (2.58) and (2.61) is on the middle terms involving  $\rho^2$  and  $\rho$ , and the difference can be neglected. Anderson's theorem suggests that a UMP test should be based on

$$\begin{aligned} a &= \frac{\hat{\epsilon}' D \hat{\epsilon}}{\hat{\epsilon}' \hat{\epsilon}} = \frac{\hat{\epsilon}' \hat{\epsilon} - \frac{1}{2} \hat{\epsilon}' A \hat{\epsilon}}{\hat{\epsilon}' \hat{\epsilon}} \\ &= 1 - \frac{1}{2} \frac{\hat{\epsilon}' A \hat{\epsilon}}{\hat{\epsilon}' \hat{\epsilon}}. \end{aligned} \quad (2.62)$$

Durbin and Watson (1950) used a slight transformation of “ $a$ ” to form their test statistic, namely,

$$\begin{aligned} d &= 2(1 - a) \\ &= \frac{\hat{\epsilon}' A \hat{\epsilon}}{\hat{\epsilon}' \hat{\epsilon}} = \frac{\sum_{t=2}^n (\hat{\epsilon}_t - \hat{\epsilon}_{t-1})^2}{\sum_{t=1}^n \hat{\epsilon}_t^2}. \end{aligned} \quad (2.63)$$

Note that “ $a$ ” in (2.62) is approximately equal to the estimate  $\hat{p} = \sum_{t=2}^n \hat{\epsilon}_t \hat{\epsilon}_{t-1} / \sum_{t=1}^n \hat{\epsilon}_t^2$ , whereas  $d \approx 2(1 - \hat{p})$ . Most econometrics textbooks discuss in details about tables of bounds for the DW test, and we will not cover that here. Our main purpose is to trace the origin of the DW test to the N-P lemma. For a historical account of the DW test, see (King, 1987). In spatial econometrics, Moran’s (1950) test for spatial dependence has the similar form

$$I = \frac{\hat{\epsilon}' W \hat{\epsilon}}{\hat{\epsilon}' \hat{\epsilon}}, \quad (2.64)$$

where  $W$  is a spatial weight matrix that represents “degree of potential interaction” among neighboring locations. Using the above analysis it is easy to link  $I$  to the N-P lemma and demonstrate its optimality [for more on this, see (Anselin and Bera, 1998)].

### 3.2 Use of score test in econometrics

In the econometrics literature, Rao’s score test is known as the Lagrange multiplier test. This terminology came from Silvey (1959). Note that the restricted MLE of  $\theta$  under the restriction  $H_0 : h(\theta) = c$  can be obtained from the first order condition of the Lagrangian function

$$\mathcal{L} = l(\theta) - \lambda'[h(\theta) - c], \quad (2.65)$$

where  $\lambda$  is an  $r \times 1$  vector of Lagrange multipliers. The first order conditions are

$$s(\tilde{\theta}) - H(\tilde{\theta})\tilde{\lambda} = 0 \quad (2.66)$$

$$h(\tilde{\theta}) = c, \quad (2.67)$$

where  $H(\theta) = \frac{\partial h(\theta)}{\partial \theta}$ . Therefore, we have  $s(\tilde{\theta}) = H(\tilde{\theta})\tilde{\lambda}$ . Given that  $H(\tilde{\theta})$  has full rank,  $s(\tilde{\theta}) = 0$  is equivalent to  $\tilde{\lambda} = 0$ , that is, the Lagrange multipliers vanish. These multipliers can be interpreted as the implicit cost (shadow prices) of imposing the restrictions. It can be shown that

$$\tilde{\lambda} = \frac{\partial l(\tilde{\theta})}{\partial c}, \quad (2.68)$$

that is, the multipliers give the rate of change of the maximum attainable value with respect to the change in the constraints. If  $H_0 : h(\theta) = c$  is true and  $l(\tilde{\theta})$  gives the optimal value,  $\tilde{\lambda}$  should be close to zero. Given this "economic" interpretation in terms of multipliers, it is not surprising that econometricians prefer the term LM rather than RS. In terms of Lagrange multipliers, (2.39) can be expressed as

$$RS = LM = \tilde{\lambda}'H(\tilde{\theta})'I(\tilde{\theta})^{-1}H(\tilde{\theta})\tilde{\lambda}. \quad (2.69)$$

Byron (1968), probably the first to apply the RS test in econometrics, used the version (2.69) along with the LR statistic for testing homogeneity and symmetry restrictions in demand equations. It took another decade for econometricians to realize the potential of the RS test. In this regard, the work of Breusch and Pagan (1980) has been the most influential. They collected relevant research reported in the statistics literature, presented the RS test in a general framework in the context of evaluating econometric models, and discussed many applications. Since the late 1970s, econometricians have applied the score principle to a variety of econometric testing problems and studied the properties of the resulting tests. Now the RS tests are the most common items in the econometricians' kit of testing tools. We will make no attempt to provide a test of all applications of the RS test in econometrics for these are far too many. For example, consider the linear regression model (2.54). The OLS analysis of this model is based on four basic assumptions: correct linear functional form; the assumptions of disturbance normality; homoskedasticity; and serial independence. Violation of these affects both estimation and inference results. With the aid of the RS principle, many procedures have been proposed to test the above assumptions and these are now routinely reported in most of the standard econometric software packages. In most cases, the algebraic forms of the LR and W tests can hardly be simplified beyond their original formulae (2.12) and (2.40). On the other hand, in many cases the RS test statistics, apart from its computational ease, can be reduced to neat and elegant formulae enabling its easy incorporation into computer software. Breusch and Pagan (1980), Godfrey (1988), Bera and Ullah (1991), Davidson and MacKinnon (1993), and Bera and Billias (2000) discussed many applications of the score tests in econometrics and demonstrated that many of the old and new econometric tests could be given a score-test interpretation. For example, test procedures developed in Hausman (1978), Newey (1985), Tauchen (1985), and White (1982) could be put in the framework of the score test. To see this, let us consider the Newey (1985) and Tauchen (1985) moment test and write the moment restriction as

$$E_f[m(y; \theta)] = 0, \quad (2.70)$$

where  $E_f$  means that (2.70) is true only when  $f(y; \theta)$  is the correct p.d.f. A test for this hypothesis can be based on the estimate of the sample counterpart of (2.70), namely,

$$\frac{1}{n} \sum_{i=1}^n m(y_i; \theta). \quad (2.71)$$

Now consider an auxiliary p.d.f.

$$f^*(y; \theta, \gamma) = f(y; \theta) \exp[\gamma m(y; \theta) - \phi(\theta, \gamma)], \quad (2.72)$$

where  $\phi(\theta, \gamma) = \ln \int \exp[\gamma m(y; \theta)] f(y; \theta) dy$ . Note that if  $f(y; \theta)$  is the correct p.d.f., then  $\gamma = 0$  in (2.72). Therefore, a test for the correct specification of  $f(y; \theta)$  can be achieved by testing  $\gamma = 0$ . Writing the loglikelihood function under the alternative hypothesis as

$$l^*(\theta, \gamma) = \sum_{i=1}^n \ln f^*(y_i; \theta, \gamma),$$

we see that the score function for testing  $\gamma = 0$  in (2.72) is

$$\left. \frac{\partial l^*(\theta, \gamma)}{\partial \gamma} \right|_{\gamma=0} = \sum_{i=1}^n m(y_i; \theta), \quad (2.73)$$

and it gives the identical moment test. This interpretation of the moment test as a score test was first noted by White (1984). Recently, Chesher and Smith (1997) gave more general and rigorous treatments of this issue. There are uncountably many choices of the auxiliary p.d.f.  $f^*(y; \theta, \gamma)$ , and the score test is invariant with respect to these choices. The LR test, however, will be sensitive to the form of  $f^*(y; \theta, \gamma)$ .

Neyman's (1959)  $C(\alpha)$  formulation, which formally established that "every" locally optimal test should be based on the score function (see equation (2.47)), also has been found to be useful in econometrics. For testing complicated nonlinear restrictions the Wald test has a computational advantage; however, on this particular occasion the Wald test runs into a serious problem of non-invariance, as pointed out by Gregory and Veal (1985) and Vaeth (1985). In this situation the score tests are somewhat difficult to compute, but the  $C(\alpha)$  tests are the most convenient to use, as demonstrated by Dagenais and Dufour (1991) and Bera and Billias (2000).

Rayner and Best (1989, sections 4.2 and 6.1) showed that Neyman's smooth statistic  $\psi_r^2$  in (2.35) can be derived as a score test for testing  $H_0 : \delta_1 = \delta_2 = \dots = \delta_r = 0$  in (2.33). In fact, Neyman's smooth test can be viewed as a first formally derived Rao's score test from the Neyman–Pearson principle. We have seen no formal application of the smooth test in econometrics. However, Lawrence Klein gave a seminar on this topic at MIT during academic year 1942–3 to draw attention to Neyman (1937), since the paper was published in a rather recondite journal [see Klein, 1991]. Unfortunately, Klein's effort did not bring Neyman's test to econometrics. However, some of the tests in econometrics can be given a smooth test interpretation. The test for normality suggested in Jarque and Bera (1980) and Bera and Jarque (1981) can be viewed as a smooth test [see Rayner and Best, 1989, p. 90]. Orthogonal polynomial tests suggested by Smith (1989) and Cameron and Trivedi (1990) are also in the spirit of Neyman (1937). In a recent paper

Diebold, Gunther, and Tay (1998) suggested the use of the density of the probability integral transformation (2.31) for density forecast evaluation. They adopted a graphical approach; however, their procedure can be formalized by using a test of the form  $\psi_r^2$ .

In Example 3 we saw that the score function vanished under the null hypothesis. In economics this kind of situation is encountered often [see, for instance, Bera, Ra, and Sarkar, 1998]. Lee and Chesher (1986) have offered a comprehensive treatment of this problem, and one of the examples they considered is the stochastic production frontier model of Aigner, Lovell, and Schmidt (1977). In Lee and Chesher (1986) the score vanished when they tried to test the null hypothesis that all the firms are efficient. They suggested using the second-order derivatives of the loglikelihood function. From (2.25) we get the same test principle by putting  $s(\theta_0) = 0$ , namely, reject the null if

$$s'(\theta_0) = \frac{\partial^2 l(\theta)}{\partial \theta^2} > k_2.$$

Therefore, again using the Neyman–Pearson principle, we see that when the score vanishes, we cannot get a locally best test but we can obtain a locally best unbiased test based on the second derivative of the loglikelihood function.

## 4 EPILOGUE

In this chapter we have first explored the general test principles developed by statisticians with some simple examples and, then, briefly discussed how those principles have been used by econometricians to construct tests for econometric models. It seems we now have a large enough arsenal to use for model testing. We should, however, be careful particularly about two aspects of testing: first, interpreting the test results, and second, taking the appropriate action when the null hypothesis is rejected. When asked about his contribution to linear models, Geoff Watson mentioned the DW test but added [see Beran and Fisher, 1998, p. 91], “What do I do if I have a regression and find the errors don’t survive the Durbin–Watson test? What do I actually do? There is no robust method. You’d like to use a procedure that would be robust against errors no matter what the covariance matrix is. Most robustness talk is really about outliers, long-tail robustness. Dependence robustness is largely untouched.”<sup>9</sup> This problem can arise after applying any test. Use of large sample tests when we have very limited data and the issue of pretesting are other important concerns. As this century and the millennium rush to a close, more research to solve these problems will make econometric model evaluation and testing more relevant in empirical work.

## Notes

\* We would like to thank Badi Baltagi, Roger Koenker and two anonymous referees for many pertinent comments. We are also grateful to Yulia Kotlyarova who provided very competent research assistance during the summer of 1998 and offered many helpful

suggestions on an earlier draft of this chapter. We, however, retain the responsibility for any remaining errors. Financial support from the Research Board of the University of Illinois at Urbana-Champaign and the Office of Research, College of Commerce and Business Administration, University of Illinois at Urbana-Champaign are gratefully acknowledged.

- 1 Neyman (1980, p. 6) stated their intuition as, "The intuitive background of the likelihood ratio test was simply as follows: if among the contemplated admissible hypotheses there are some that ascribe to the facts observed probabilities much larger than that ascribed by the hypothesis tested, then it appears 'reasonable' to reject the hypothesis."
- 2 This result follows from the generalized Cauchy–Schwarz inequality (Rao, 1973, p. 54)

$$(u'v)^2 \leq (u'Au)(v'A^{-1}v),$$

where  $u$  and  $v$  are column vectors and  $A$  is a non-singular matrix. Equality holds when  $u = A^{-1}v$ .

- 3 The interrelationships among these three tests can be brought home to students through an amusing story. Once around 1946 Ronald Fisher invited Jerzy Neyman, Abraham Wald, and C.R. Rao to his lodge for afternoon tea. During their conversation, Fisher mentioned the problem of deciding whether his dog, who had been going to an "obedience school" for some time, was disciplined enough. Neyman quickly came up with an idea: leave the dog free for some time and then put him on his leash. If there is not much difference in his behavior, the dog can be thought of as having completed the course successfully. Wald, who lost his family in the concentration camps, was adverse to any restrictions and simply suggested leaving the dog free and seeing whether it behaved properly. Rao, who had observed the nuisances of stray dogs in Calcutta streets, did not like the idea of letting the dog roam freely and suggested keeping the dog on a leash at all times and observing how hard it pulls on the leash. If it pulled too much, it needed more training. That night when Rao was back in his Cambridge dormitory after tending Fisher's mice at the genetics laboratory, he suddenly realized the connection of Neyman and Wald's recommendations to the Neyman–Pearson LR and Wald tests. He got an idea and the rest is history.
- 4 In the  $C(\alpha)$  test the letter "C" refers to Cramér and " $\alpha$ " to the level of significance. Neyman (1959) was published in a Festschrift for Harald Cramér. Neyman frequently referred to this work as his "last performance." He was, however, disappointed that the paper did not attract as much attention as he had hoped for, and in later years, he regretted publishing it in a Festschrift as not many people read Festschriften.
- 5 In the introduction (p. 11), Tinbergen stated,

The purpose of this series of studies is to submit to statistical test some of the theories which have been put forward regarding the character and causes of cyclical fluctuation in business activity. Many of these theories, however, do not exist in a form immediately appropriate for statistical testing while most of them take account of the same body of economic phenomena – viz., the behavior of investment, consumption, incomes, prices, etc. Accordingly, the method of procedure here adopted is not to test the various theories one by one (a course which would involve much repetition), but to examine in succession, in the light of the various explanations which have been offered, the relation between certain groups of economic phenomena.

He, however, cautioned against relying too much on the test results, "for no statistical test can prove a theory to be correct" (p. 12). For more on this see (Duo, 1993, Chapter 5).

- 6 In his Nobel lecture he stated (Haavelmo, 1997),

"For my own part I was lucky enough to be able to visit the United States in 1939 on a scholarship. . . . I then had the privilege of studying with the world famous statistician Jerzy Neyman in California for a couple of months. At that time, young and naive, I thought I knew something about econometrics. I exposed some of my thinking on the subject to Professor Neyman. Instead of entering into a discussion with me, he gave me two or three exercises for me to work out. He said he would talk to me when I had done these exercises. When I met him for the second talk, I had lost most of my illusions regarding the understanding of how to do econometrics. But Professor Neyman also gave me hopes that there might be other more fruitful ways to approach the problem of econometric methods than those which had so far caused difficulties and disappointments."

- 7 Technically speaking, this model does not fit in our earlier framework due to the dependence structure. However, once a proper likelihood function is defined we can derive our earlier test statistics.
- 8 Instead of dealing with the joint distribution conditional on the explanatory variables in all time periods, a better approach would be to consider sequential conditional distribution under much weaker assumptions. Wooldridge (1994) discusses the merits of modeling sequential distributions.
- 9 We should, however, note that econometricians have developed a number of procedures to estimate a consistent variance–covariance matrix to take account of the unknown form of dependence; for a discussion of this and other robust procedures see (Bera, 2000) and (Wooldridge, 2000).

## References

- Aigner, D.J., C.A.K. Lovell, and P. Schmidt (1977). Formulation and estimation of stochastic frontier production function model. *Journal of Econometrics* 6, 21–37.
- Anderson, T.W. (1948). On the theory of testing serial correlation. *Skandinavisk Aktuarietidskrift* 31, 88–116.
- Anselin, L., and A.K. Bera (1998). Spatial dependence in linear regression models with an introduction to spatial econometrics. In A. Ullah and D.E.A. Giles (eds.), *Handbook of Applied Economic Statistics*. New York: Marcel Dekker, 237–89.
- Bayes, Rev. T. (1763). An essay toward solving a problem in the doctrine of chances. *Philosophical Transactions of the Royal Society* 53, 370–418.
- Bera, A.K. (2000). Hypothesis testing in the 20th century with a special reference to testing with misspecified models. In C.R. Rao and G. Szekely (eds.), *Statistics for the 21st Century*. New York: Marcel Dekker, 33–92.
- Bera, A.K., and Y. Billias (2000). Rao's score, Neyman's  $C(\alpha)$  and Silvey's LM test: An essay on historical developments and some new results. *Journal of Statistical Planning and Inference* Forthcoming.
- Bera, A.K., and C.M. Jarque (1981). An efficient large-sample test for normality of observations and regression residuals. Working Paper in Economics and Econometrics, Number 40, The Australian National University, Canberra.
- Bera, A.K., and A. Ullah (1991). Rao's score test in econometrics. *Journal of Quantitative Economics* 7, 189–220.
- Bera, A.K., and M.J. Yoon (1993). Specification testing with locally misspecified alternatives. *Econometric Theory* 9, 649–58.

- Bera, A.K., S.-S. Ra, and N. Sarkar (1998). Hypothesis testing for some nonregular cases in econometrics. In S. Chakravarty, D. Coondoo, and R. Mukherjee (eds.), *Econometrics: Theory and Practice*, New Delhi: Allied Publishers, 319–51.
- Beran, R.J., and N.I. Fisher (1998). A conversation with Geoff Watson. *Statistical Science* 13, 75–93.
- Breusch, T.S., and A.R. Pagan (1980). The Lagrange multiplier test and its applications to model specification in econometrics. *Review of Economic Studies* 47, 239–53.
- Byron, R.P. (1968). Methods for estimating demand equations using prior information: A series of experiments with Australian data. *Australian Economic Papers* 7, 227–48.
- Cameron, A.C., and P.K. Trivedi (1990). Conditional moment tests and orthogonal polynomials, Working Paper in Economics, Number 90–051, Indiana University.
- Chesher, A., and R. Smith (1997). Likelihood ratio specification tests. *Econometrica* 65, 627–46.
- Cramér, H. (1946). *Mathematical Methods of Statistics*. New Jersey: Princeton University Press.
- Dagenais, M.G., and J.-M. Dufour (1991). Invariance, nonlinear models, and asymptotic tests. *Econometrica* 59, 1601–15.
- Davidson, R., and J.G. MacKinnon (1993). *Estimation and Inference in Econometrics*. Oxford: Oxford University Press.
- Diebold, F.X., T.A. Gunther, and A.S. Tay (1998). Evaluating density forecasts with application to financial risk management. *International Economic Review*, 39, 863–905.
- Duo, Q. (1993). *The Foundation of Econometrics: A Historical Perspective*. Oxford: Clarendon Press.
- Durbin, J., and G.S. Watson (1950). Testing for serial correlation in least squares regression I. *Biometrika* 37, 409–28.
- Fisher, R.A. (1922). On the mathematical foundations of theoretical statistics. *Philosophical Transaction of the Royal Society A222*, 309–68.
- Godfrey, L.G. (1988). *Misspecification Tests in Econometrics, The Lagrange Multiplier Principle and Other Approaches*. Cambridge: Cambridge University Press.
- Gouriéroux, C., and A. Monfort (1995). *Statistics and Econometric Models* 2. Cambridge: Cambridge University Press.
- Gregory, A.W., and M.R. Veal (1985). Formulating Wald tests of nonlinear restrictions. *Econometrica* 53, 1465–8.
- Haavelmo, T. (1944). The probability approach in econometrics. *Supplements to Econometrica* 12.
- Haavelmo, T. (1997). Econometrics and the welfare state: Nobel lecture, December 1989. *American Economic Review* 87, 13–5.
- Hausman, J.J. (1978). Specification tests in econometrics. *Econometrica* 46, 1215–72.
- Jarque, C.M., and A.K. Bera (1980). Efficient tests for normality, homoskedasticity and serial independence of regression residuals. *Economics Letters* 6, 255–9.
- King, M.L. (1987). Testing for autocorrelation in linear regression models: A survey. In M.L. King and D.E.A. Giles (eds.), *Specification Analysis in the Linear Model*. London: Routledge and Kegan Paul, 19–73.
- Klein, L. (1991). The statistics seminar, MIT, 1942–1943. *Statistical Science* 6, 320–30.
- Lee, L.F., and A. Chesher (1986). Specification testing when score test statistics are individually zero. *Journal of Econometrics* 31, 121–49.
- Lehmann, E.L. (1986). *Testing Statistical Hypotheses*. New York: John Wiley & Sons.
- Lehmann, E.L. (1999). *Elements of Large Sample Theory*. New York: Springer-Verlag.
- Moran, P.A.P. (1950). A test for the serial independence of residuals. *Biometrika* 37, 178–81.
- Newey, W. (1985). Maximum likelihood specification testing and conditional moment tests. *Econometrica* 53, 1047–70.

- Neyman, J. (1937). "Smooth test" for goodness of fit. *Skandinavisk Akturarietidskrift* 20, 150–99.
- Neyman, J. (1954). Sur une famille de tests asymptotiques des hypothèses statistiques compasées. *Trabajos de Estadística* 5, 161–8.
- Neyman, J. (1959). Optimal asymptotic test of composite statistical hypothesis. In U. Grenander (ed.), *Probability and Statistics, the Harald Cramér Volume*. Uppsala: Almqvist and Wiksell, 213–34.
- Neyman, J. (1980). Some memorable incidents in probabilistic/statistical studies. In I.M. Chakravarti (ed.), *Asymptotic Theory of Statistical Tests and Estimation*, New York: Academic Press, 1–32.
- Neyman, J., and E.S. Pearson (1928). On the use and interpretation of certain test criteria for purpose of statistical inference. *Biometrika* 20, 175–240.
- Neyman, J., and E.S. Pearson (1933). On the problem of the most efficient tests of statistical hypothesis. *Philosophical Transactions of the Royal Society, Series A* 231, 289–337.
- Neyman, J., and E.S. Pearson (1936). Contribution to the theory of testing statistical hypothesis I: Unbiased critical regions of type A and type A<sub>1</sub>. *Statistical Research Memoirs* 1, 1–37.
- Pearson, K. (1900). On the criterion that a given system of deviations from the probable in the case of a correlated system of variables is such that it can be reasonably supposed to have arisen from random sampling. *Philosophical Magazine Series 5*, 50, 157–75.
- Rao, C.R. (1948). Large sample tests of statistical hypotheses concerning several parameters with applications to problems of estimation. *Proceedings of the Cambridge Philosophical Society* 44, 50–7.
- Rao, C.R. (1973). *Linear Statistical Inference and Its Applications*. New York: John Wiley and Sons.
- Rao, C.R. (2000). Two score and ten years of score tests. *Journal of Statistical Planning and Inference*.
- Rao, C.R., and S.J. Poti (1946). On locally most powerful tests when alternatives are one sided. *Sankhyā* 7, 439–40.
- Rayner, J.C.W., and D.J. Best (1989). *Smooth Tests of Goodness of Fit*. New York: Oxford University Press.
- Serfling, R.J. (1980). *Approximation Theorems of Mathematical Statistics*. New York: John Wiley and Sons.
- Silvey, S.D. (1959). The Lagrange multiplier test. *Annals of Mathematical Statistics* 30, 389–407.
- Smith, R. (1989). On the use of distributional mis-specification checks in limited dependent variable models. *Economic Journal* 99, 178–92.
- Tauchen, G. (1985). Diagnostic testing and evaluation of maximum likelihood models. *Journal of Econometrics* 30, 415–43.
- Vaeth, M. (1985). On the use of Wald's test in exponential families. *International Statistical Review* 53, 199–214.
- Wald, A. (1943). Tests of statistical hypothesis concerning several parameters when the number of observations is large. *Transactions of the American Mathematical Society* 54, 426–82.
- White, H. (1982). Maximum likelihood estimation of misspecified models. *Econometrica* 50, 1–25.
- White, H. (1984). Comment on "Tests of specification in econometrics." *Econometric Reviews* 3, 261–7.
- Wooldridge, J.M. (1994). Estimation and inference for dependent processes. In R.F. Engle and D.L. McFadden (eds.), *Handbook of Econometrics* Vol. 4. Amsterdam: North-Holland, 2639–738.
- Wooldridge, J.M. (2000). Diagnostic testing. Chapter 9 this volume.

CHAPTER THREE

# Serial Correlation

*Maxwell L. King*

## 1 INTRODUCTION

In its most general form, serial correlation involves the correlation of successive time series observations. It has been the subject of much research in econometrics over the last half century, particularly with respect to serial correlation in regression disturbances. The seminal work of Cochrane and Orcutt (1949) did much to alert econometricians to the difficulties of assuming independent regression errors in time series applications of the standard linear regression model. It is now well known that the neglect of disturbance correlation can lead to inefficient parameter estimates, misleading inferences from hypothesis tests and inefficient predictions. This led to a vast literature on testing for serial correlation in linear regression models and other models, see for example King (1987) for a review of this literature.

Early econometric models were often built using a static representation of the economic forces at work. Such static models were generally estimated using annual data and typically any problems resulting from the poor specification of the dynamics of the economic situation being modeled were swept into the error term. The availability of quarterly data in the 1970s brought the realization that better modeling of the dynamics was needed. This led to a greater interest by econometricians in a class of univariate time series models proposed by Box and Jenkins (1970). It also led to greater use of dynamic linear regression models in which the lagged dependent variable is included as a regressor. In summary, modeling serial correlation really involves taking care of the dynamic part of a model specification.

The class of Box–Jenkins models provides important building blocks for a wide range of model specifications that handle the dynamics of the process. In this chapter we provide a summary of these different types of models and consider related issues of estimation and testing. We constrain our attention purely to univariate models. These models can all be generalized to linear simultaneous equations and vector autoregressive models.

The plan of this chapter is as follows. A range of models are introduced in Section 2. These include Box–Jenkins time series models, regression disturbance models of serial correlation and dynamic linear regression models. Section 3 discusses the problem of estimation of the linear regression model with serial correlation in the disturbances. Hypothesis testing is the topic of Section 4. Particular emphasis is placed on the Durbin–Watson and related tests in the context of the standard linear regression model and the dynamic linear regression model. The chapter ends with a short discussion on model selection.

## 2 MODELS

### 2.1 The Box–Jenkins class of models

The simplest time series model of serial correlation is the first-order autoregressive (AR(1)) process, which for a time series,  $y_t$ ,  $t = 1, \dots, n$ , with mean zero can be written as

$$y_t = \rho y_{t-1} + \varepsilon_t, \quad |\rho| < 1, \quad \varepsilon_t \sim \text{iid}(0, \sigma^2). \quad (3.1)$$

$\varepsilon_t$  is the error or innovation of the process at time  $t$  and is assumed to be independently distributed with mean zero and constant variance  $\sigma^2$ . The requirement on the scalar parameter  $\rho$  that  $|\rho| < 1$ , is called the stationarity condition which stops  $y_t$  from being an explosive process and ensures its variance is a constant  $\sigma^2/(1 - \rho^2)$ . Model (3.1) is incomplete without a statement or assumption about how  $y_0$  is generated. A standard assumption is that the process has been going on forever and  $y_0$  has the same distribution as any other  $y_t$  (stationarity assumption). This is often expressed by  $y_0$  being distributed with mean zero and variance  $\sigma^2/(1 - \rho^2)$ . An alternative approach is to treat  $y_0$  as a constant – often zero. This results in a nonstationary process because its variance is not constant.

If we denote by  $y = (y_1, \dots, y_n)'$ , the vector of observations on the process, then  $y$  has a mean which is the  $n \times 1$  vector of zeros and has the covariance matrix in the stationary case of

$$\text{var}(y) = \frac{\sigma^2}{1 - \rho^2} \begin{bmatrix} 1 & \rho & \rho^2 & \cdots & \cdots & \cdots & \cdots & \rho^{n-1} \\ \rho & 1 & \rho & & & & & \rho^{n-2} \\ \rho^2 & \rho & 1 & & & & & \rho^{n-3} \\ \vdots & & & \ddots & & & & \vdots \\ \vdots & & & & \ddots & & & \vdots \\ \vdots & & & & & \ddots & & \vdots \\ \vdots & & & & & & 1 & \rho \\ \rho^{n-1} & \rho^{n-2} & \cdots & \cdots & \cdots & \cdots & \rho & 1 \end{bmatrix}.$$

The latter follows from

$$\text{cov}(y_t, y_{t-i}) = \rho^i \text{var}(y_t) = \rho^i \sigma^2 / (1 - \rho^2), \quad i = 1, \dots, n - 1.$$

This implies that the AR(1) process has an autocorrelation function of the form

$$\rho_{(i)} = \text{cov}(y_t, y_{t-i}) / (\text{var}(y_t) \text{ var}(y_{t-i}))^{1/2} = \rho^i$$

which declines at an exponential rate as  $i$  increases.

Another simple model is the first-order moving average (MA(1)) process which can be written as

$$y_t = \varepsilon_t + \gamma \varepsilon_{t-1}, \quad \varepsilon_t \sim \text{iid}(0, \sigma^2) \quad (3.2)$$

where  $\varepsilon_t$  is defined as in (3.1). The  $n \times 1$  vector  $y$  has mean zero and covariance matrix

$$\text{var}(y) = \sigma^2 \begin{bmatrix} 1 + \gamma^2 & \gamma & 0 & \dots & \dots & \dots & 0 \\ \gamma & 1 + \gamma^2 & \gamma & & & & 0 \\ 0 & \gamma & 1 + \gamma^2 & & & & 0 \\ \vdots & & & \ddots & & & \vdots \\ \vdots & & & & \ddots & & \vdots \\ \vdots & & & & & 1 + \gamma^2 & \gamma \\ 0 & 0 & \dots & \dots & \dots & \gamma & 1 + \gamma^2 \end{bmatrix}. \quad (3.3)$$

Note that  $\text{var}(y_t) = \sigma^2(1 + \gamma^2)$  and the autocorrelation function of an MA(1) process is

$$\begin{aligned} \rho_{(i)} &= \gamma / (1 + \gamma^2), \quad i = 1 \\ &= 0, \quad \quad \quad i > 1. \end{aligned}$$

We usually assume  $|\gamma| \leq 1$  although it is possible to have an MA(1) process with  $|\gamma| > 1$ , but for normally distributed errors, it is impossible to distinguish between the likelihood function for (3.2) with  $(\gamma^*, \sigma^{*2}) = (\gamma, \sigma^2)$  and  $(\gamma^*, \sigma^{*2}) = (1/\gamma, \sigma^2\gamma^2)$  because (3.3) takes the same value for these two sets of parameter values. The standard solution to this minor identification problem is to restrict  $\gamma$  to the interval  $|\gamma| \leq 1$ . This is known as the invertibility condition. An MA(1) process is stationary because  $y_t$  is a simple weighted sum of the innovations  $\varepsilon_t, \varepsilon_{t-1}$ , so no condition is required for stationarity.

If we combine (3.1) and (3.2), we get an autoregressive moving average (ARMA(1, 1)) process which can be written as

$$y_t = \rho y_{t-1} + \varepsilon_t + \gamma \varepsilon_{t-1}, \quad |\rho| < 1, \quad \varepsilon_t \sim \text{iid}(0, \sigma^2). \quad (3.4)$$

Note that

$$\text{var}(y_t) = \sigma^2 \{(1 + \gamma^2) + 2\gamma\rho\} / (1 - \rho^2)$$

$$\begin{aligned} \text{cov}(y_t, y_{t-i}) &= \rho \text{ var}(y_t) + \gamma \sigma^2, \quad \text{for } i = 1 \\ &= \rho^i \text{ var}(y_t), \quad \quad \quad \text{for } i \geq 2. \end{aligned}$$

It is also worth observing that if

$$y_t = \varepsilon_t, \quad (3.5)$$

i.e., we have a white noise model for  $y_t$  with no serial correlation, then by lagging (3.5) one period, multiplying it by  $\rho$  and subtracting from (3.5) we get

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

which is (3.4) with  $\gamma = -\rho$ . This is known as a model with a common factor. It appears to be an ARMA(1, 1) model but it is in fact a white noise model.

As in the AR(1) case, (3.4) is not complete until we have made an assumption about  $y_0$ . Again the usual assumption is stationarity, i.e. to assume that (3.4) is a process that has been going on forever.

The  $p$ th order autoregressive process (AR( $p$ )) is a generalization of (3.1) and can be written as

$$y_t = \rho_1 y_{t-1} + \rho_2 y_{t-2} + \dots + \rho_p y_{t-p} + \varepsilon_t, \quad \varepsilon_t \sim \text{iid}(0, \sigma^2). \quad (3.6)$$

The stationarity condition generalizes to the requirement that the roots of the polynomial equation

$$1 - \rho_1 z - \rho_2 z^2 - \dots - \rho_p z^p = 0 \quad (3.7)$$

lie outside the unit circle. In other words, all the roots of (3.7) must be larger than one in absolute value. For an AR(2) process, this requires

$$\rho_1 + \rho_2 < 1, \quad \rho_2 - \rho_1 < 1 \quad \text{and} \quad \rho_2 > -1.$$

In a similar way, the  $q$ th order moving average process (MA( $q$ )) is a generalization of (3.2) and has the form

$$y_t = \varepsilon_t + \gamma_1 \varepsilon_{t-1} + \dots + \gamma_q \varepsilon_{t-q}, \quad \varepsilon_t \sim \text{iid}(0, \sigma^2). \quad (3.8)$$

The invertibility condition now becomes such that the roots of the polynomial equation

$$1 + \gamma_1 z + \gamma_2 z^2 + \dots + \gamma_q z^q = 0$$

lie outside the unit circle. Again no further condition is required for stationarity.

Observe that

$$\text{var}(y_t) = \sigma^2(1 + \gamma_1^2 + \dots + \gamma_q^2)$$

$$\begin{aligned} \text{cov}(y_t y_{t-i}) &= \sigma^2(\gamma_i + \gamma_1 \gamma_{i+1} + \dots + \gamma_{q-i} \gamma_q), & \text{for } i = 1, \dots, q, \\ &= 0, & \text{for } i > q. \end{aligned}$$

In addition, (3.6) and (3.8) can be combined to produce an ARMA( $p, q$ ) process which has the form

$$y_t = \rho_1 y_{t-1} + \rho_2 y_{t-2} + \dots + \rho_p y_{t-p} + \varepsilon_t + \gamma_1 \varepsilon_{t-1} + \dots + \gamma_q \varepsilon_{t-q}, \quad \varepsilon_t \sim \text{iid}(0, \sigma^2). \quad (3.9)$$

This class of models is popular because it can often allow the serial correlation in a stationary time series to be modeled with a handful of parameters.

A further important generalization of the class of ARMA( $p, q$ ) models is the class of autoregressive integrated moving average (ARIMA) models. If  $L$  denotes the lag operator such that  $Ly_t = y_{t-1}$  then the ARIMA( $p, d, q$ ) model is just an ARMA( $p, q$ ) model applied to the transformed variable

$$(1 - L)^d y_t \quad (3.10)$$

rather than to  $y_t$ . When  $d = 1$ , then  $(1 - L)^d y_t = y_t - y_{t-1}$  which is the first difference of  $y_t$ . Often  $y_t$  is not a stationary series but its first (or higher) difference,  $y_t - y_{t-1}$ , is. Hence it may be sensible to fit an ARMA model to the first (or higher) difference of  $y_t$ .

A special ARIMA model is the ARIMA(0, 1, 0) model which is also known as the random walk model because it has the form

$$y_t = y_{t-1} + \varepsilon_t, \quad \varepsilon_t \sim \text{iid}(0, \sigma^2). \quad (3.11)$$

Observe that (3.11) can also be written as

$$y_t = y_0 + \varepsilon_1 + \dots + \varepsilon_t$$

which means  $\text{var}(y_t - y_0) = t\sigma^2$ , a quantity that goes to infinity as  $t$  increases.

Finally there is the related class of fractionally integrated models, denoted ARFIMA, in which the  $d$  in (3.10) is not restricted to taking only an integer value.

For further details on ARIMA models, see Box and Jenkins (1970, 1976). A good survey of fractionally integrated processes and their applications in econometrics is given by Baillie (1996).

## 2.2 Serial correlation in the disturbances of the linear regression model

The seminal work of Cochrane and Orcutt (1949) alerted the econometric profession to the difficulties of assuming uncorrelated disturbances in time series applications of the linear regression model. It soon became well known that neglecting serial correlation in regression disturbances can lead to inefficient parameter estimates, misleading hypothesis tests on these parameters and inefficient regression forecasts. The initial focus was on the linear regression model

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{u} \quad (3.12)$$

where  $y$  is  $n \times 1$ ,  $X$  is an  $n \times k$  nonstochastic matrix of rank  $k < n$ ,  $\beta$  is a  $k \times 1$  parameter vector and  $u$  is an  $n \times 1$  disturbance vector whose elements are assumed to follow a stationary AR(1) process

$$u_t = \rho u_{t-1} + \varepsilon_t, \quad |\rho| < 1, \quad \varepsilon_t \sim \text{iid}(0, \sigma^2). \quad (3.13)$$

It was recognized (see for example Johnston, 1972), that the usual reason for including the disturbance term in the regression model is to account for the effects of omitted or unobservable regressors, errors in the measurement of the dependent variable, arbitrary human behavior and functional approximations. Given that some of these effects are typically autocorrelated, it was thought that the AR(1) disturbance process (3.13) might be a good model for the regression disturbances in econometric applications.

From the 1970s onwards, there has been an increased understanding of other forms of serially correlated disturbance processes, no doubt helped by the work of Box and Jenkins (1970) and others. With the increased use of quarterly time series data, the simple AR(4) process

$$u_t = \rho_4 u_{t-4} + \varepsilon_t, \quad |\rho_4| < 1, \quad \varepsilon_t \sim \text{iid}(0, \sigma^2) \quad (3.14)$$

was made popular by the work of Thomas and Wallis (1971) and Wallis (1972). Combined with an AR(1) process, it also leads to the restricted AR(5) process

$$u_t = \rho_1 u_{t-1} + \rho_4 u_{t-4} - \rho_1 \rho_4 u_{t-5} + \varepsilon_t, \quad \varepsilon_t \sim \text{iid}(0, \sigma^2). \quad (3.15)$$

New numerical algorithms and improvements in computer hardware in the 1970s and 1980s have allowed MA disturbance models to be an alternative to the standard AR models – see for example, Nichols, Pagan, and Terrell (1975) for a review. The MA(1) disturbance model takes the form

$$u_t = \varepsilon_t + \gamma \varepsilon_{t-1}, \quad \varepsilon_t \sim \text{iid}(0, \sigma^2) \quad (3.16)$$

and the simple MA(4) process

$$u_t = \varepsilon_t + \gamma_4 \varepsilon_{t-4}, \quad \varepsilon_t \sim \text{iid}(0, \sigma^2) \quad (3.17)$$

is an alternative model to (3.14) for quarterly data. Of course there is really no reason not to consider a general ARMA( $p, q$ ) model for the disturbance process which is what many econometricians these days consider rather than just an AR(1) model.

### 2.3 The dynamic linear regression model

In many applications of the linear regression model, it is clear that the value of the dependent variable depends very much on some fraction of its value in the

previous period as well as on other independent variables. This leads naturally to the dynamic linear regression model, the simplest version of which is

$$y_t = \alpha_1 y_{t-1} + x'_t \beta + u_t, \quad t = 1, \dots, n, \quad (3.18)$$

where  $\alpha_1$  is a scalar,  $\beta$  is a  $k \times 1$  parameter vector,  $x_t$  is a  $k \times 1$  vector of exogenous variables and  $u_t$  is the disturbance term. A more general model is

$$y_t = \alpha_1 y_{t-1} + \dots + \alpha_p y_{t-p} + x'_t \beta + u_t, \quad t = 1, \dots, n. \quad (3.19)$$

For completion, we need to also provide an assumption on the generation of  $y_0, y_{-1}, \dots, y_{1-p}$ . One approach is to treat  $y_0, y_{-1}, \dots, y_{1-p}$  as constants. Another is to assume they each have a constant mean equal to  $E(y_1)$  and that  $v_t = y_t - E(y_t)$  follows the stationary AR( $p$ ) process

$$v_t = \alpha_1 v_{t-1} + \alpha_2 v_{t-2} + \dots + \alpha_p v_{t-p} + u_t$$

in which  $u_t$  is the error term in (3.19). The former assumption is appropriate if we wish to make inferences conditional on the values taken by  $y_0, y_{-1}, \dots, y_{1-p}$  while the latter assumption has been made by Tse (1982) and Inder (1985, 1986).

In some circumstances, there is not much difference between the linear regression (3.12) with AR(1) errors and the first-order dynamic regression model (3.18). To see this, suppose  $x_t$  is made up of a constant intercept regressor and the time trend, i.e.  $x_t = (1, t)'$ . Consider

$$y_t = x'_t \beta + u_t \quad (3.20)$$

in which  $u_t$  follows the AR(1) process (3.13). If we lag (3.20) one period, multiply it by  $\rho$  and subtract from (3.20), we get

$$y_t = \rho y_{t-1} + x'_t \beta - \rho x'_{t-1} \beta + \epsilon_t. \quad (3.21)$$

Observe that (3.21) is a regression with a well-behaved error term and five regressors. Because when  $x_t$  is lagged, it remains a constant regressor and a linear trend regressor, there is perfect multicollinearity between the regressors of  $x_t$  and  $x_{t-1}$ . This problem can be solved by dropping the  $\rho x'_{t-1} \beta$  term, in which case (3.21) becomes the simple dynamic linear model (3.18) with  $u_t \sim \text{iid}(0, \sigma^2)$ . Thus in the general case of a linear regression model of the form of (3.20), if  $x_t$  is "lag invariant" in the sense that all regressors in  $x_{t-1}$  can be written as linear combinations of the regressors purely from  $x_t$ , then (3.18) and (3.20) are equivalent. This simple analysis has ignored the first observation. Any difference between the two models could depend largely on what is assumed about the first observation in each model.

### 3 ESTIMATION

This section discusses the problem of estimation for the models of the previous section. Recall that subsection 2.1 considered models with mean zero. These can be readily generalized to models in which  $y_t$  has a constant but unknown mean, say  $\beta_1$ . Such models can be written as a special case of a regression model (3.12) in which  $X$  is the vector of ones. It is very rare that the mean of  $y_t$  is known to be zero, so models which allow for nonzero means are more realistic. Thus in this section, we shall restrict our attention to the estimation of the linear regression model (3.12) with various ARMA-type disturbance processes.

#### 3.1 Maximum likelihood estimation

An  $n \times 1$  vector  $y$  generated by the linear regression model (3.12) with an ARMA( $p, q$ ) disturbance process with normal errors, i.e.  $\varepsilon_t \sim IN(0, \sigma^2)$ , can be shown to be distributed as

$$y \sim N(X\beta, \sigma^2\Omega(\rho, \gamma)) \quad (3.22)$$

where  $\sigma^2\Omega(\rho, \gamma)$  is the covariance matrix of the ARMA( $p, q$ ) process in which  $\rho = (\rho_1, \dots, \rho_p)'$  and  $\gamma = (\gamma_1, \dots, \gamma_q)'$  are parameter vectors. The loglikelihood function of this model is

$$\begin{aligned} f(\beta, \sigma^2, \rho, \gamma | y) &= -\frac{n}{2} \log (2\pi\sigma^2) - \frac{1}{2} \log |\Omega(\rho, \gamma)| \\ &\quad - \frac{1}{2\sigma^2} (y - X\beta)' \Omega(\rho, \gamma)^{-1} (y - X\beta). \end{aligned} \quad (3.23)$$

For any given values of  $\rho$  and  $\gamma$ , the values of  $\beta$  and  $\sigma^2$  that maximize (3.23) are

$$\tilde{\beta}_{\rho, \gamma} = (X'\Omega(\rho, \gamma)^{-1}X)^{-1}X'\Omega(\rho, \gamma)^{-1}y \quad (3.24)$$

and

$$\tilde{\sigma}_{\rho, \gamma}^2 = (y - X\tilde{\beta}_{\rho, \gamma})' \Omega(\rho, \gamma)^{-1} (y - X\tilde{\beta}_{\rho, \gamma}) / n. \quad (3.25)$$

If these estimators of  $\beta$  and  $\sigma^2$  are substituted back into (3.23), we get the profile or concentrated loglikelihood:

$$f_p(\rho, \gamma | y) = -\frac{n}{2} \log (2\pi\tilde{\sigma}_{\rho, \gamma}^2) - \frac{1}{2} \log |\Omega(\rho, \gamma)| - \frac{n}{2}. \quad (3.26)$$

Full maximum likelihood estimates of the parameters in this model are therefore obtained by first finding the values of the  $\rho$  and  $\gamma$  vectors which maximize (3.26). These values of  $\rho$  and  $\gamma$ , denoted  $\tilde{\rho}$  and  $\tilde{\gamma}$ , are then used in (3.24) and (3.25) to find  $\tilde{\beta}$  and  $\tilde{\sigma}^2$ , the maximum likelihood estimates.

There are a number of practical issues in this process worthy of discussion. The first involves exploiting the Cholesky decomposition of  $\Omega(\rho, \gamma)$  denoted  $L(\rho, \gamma)$  such that

$$L(\rho, \gamma)L(\rho, \gamma)' = \Omega(\rho, \gamma)$$

where  $L(\rho, \gamma)$  is a lower triangular matrix. If we denote

$$y^*(\rho, \gamma) = L(\rho, \gamma)^{-1}y \quad (3.27)$$

$$X^*(\rho, \gamma) = L(\rho, \gamma)^{-1}X, \quad (3.28)$$

then (3.24) becomes

$$\tilde{\beta}_{\rho, \gamma} = (X^*(\rho, \gamma)'X^*(\rho, \gamma))^{-1}X^*(\rho, \gamma)'y^*(\rho, \gamma), \quad (3.29)$$

the ordinary least squares (OLS) estimator from the regression of  $y^*(\rho, \gamma)$  on  $X^*(\rho, \gamma)$ , and (3.25) becomes

$$\tilde{\sigma}_{\rho, \gamma}^2 = e^*(\rho, \gamma)'e^*(\rho, \gamma)/n, \quad (3.30)$$

the sum of squared residuals from this regression divided by  $n$ , where

$$e^*(\rho, \gamma) = y^*(\rho, \gamma) - X^*(\rho, \gamma)\tilde{\beta}_{\rho, \gamma}. \quad (3.31)$$

The problem of maximizing (3.26) with respect to  $\rho$  and  $\gamma$  now reduces to one of maximizing

$$-\frac{n}{2} \log (e^*(\rho, \gamma)'e^*(\rho, \gamma)) - \log |L(\rho, \gamma)| = -\frac{n}{2} \log (\tilde{e}^*(\rho, \gamma)' \tilde{e}^*(\rho, \gamma))$$

where

$$\tilde{e}^*(\rho, \gamma) = e^*(\rho, \gamma)|L(\rho, \gamma)|^{1/n}.$$

Thus the estimation problem reduces to minimizing the sum of squares

$$\tilde{s} = \sum_{i=1}^n \tilde{e}^*(\rho, \gamma)_i^2 \quad (3.32)$$

with respect to  $\rho$  and  $\gamma$ . This may be achieved by applying a standard nonlinear least squares algorithm to  $\tilde{s}$  and then using the resultant  $\tilde{\rho}$  and  $\tilde{\gamma}$  in (3.29) and (3.30) to obtain  $\tilde{\beta}$  and  $\tilde{\sigma}^2$ . Also observe that because  $L(\rho, \gamma)$  is a lower triangular matrix, its determinant is easily calculated as the product of the diagonal elements, i.e.

$$|L(\rho, \gamma)| = \prod_{i=1}^n L(\rho, \gamma)_{ii}. \quad (3.33)$$

The remaining issue is the construction of  $\Omega(\rho, \gamma)$  or more importantly  $L(\rho, \gamma)$  or  $L^{-1}(\rho, \gamma)$ . Note that through (3.27), (3.28), (3.29), (3.30), and (3.32), maximum likelihood estimates can be obtained without the need to calculate  $\Omega(\rho, \gamma)$  or  $\Omega^{-1}(\rho, \gamma)$ . In the case of AR(1) disturbances

$$L^{-1}(\rho, \lambda) = \begin{bmatrix} (1 - \rho^2)^{1/2} & 0 & \dots & \dots & \dots & 0 & 0 \\ -\rho & 1 & & & & & \vdots \\ \vdots & & \ddots & & & & \vdots \\ \vdots & & & \ddots & & & \vdots \\ \vdots & & & & \ddots & & \vdots \\ 0 & & & & & 1 & 0 \\ 0 & \dots & \dots & \dots & \dots & -\rho & 1 \end{bmatrix}$$

which was first derived by Prais and Winsten (1954) and is also known as the Prais–Winsten transformation.

For AR( $p$ ) disturbances, van der Leeuw (1994) has shown that

$$\Omega(\rho, \gamma) = [P'P - NN']^{-1}$$

where  $P$  is the  $n \times n$  triangular matrix

$$P = \begin{bmatrix} 1 & 0 & 0 & \dots & \dots & \dots & 0 & 0 & \dots & \dots & \dots & \dots & 0 & 0 \\ -\rho_1 & 1 & 0 & & & & \vdots & & & & & & & 0 \\ -\rho_2 & -\rho_1 & 1 & & & & \vdots & & & & & & & \vdots \\ & & \ddots & & & & \vdots & & & & & & & \vdots \\ & & & \ddots & & & \vdots & & & & & & & \vdots \\ & & & & \ddots & & \vdots & & & & & & & \vdots \\ -\rho_p & -\rho_{p-1} & \dots & \dots & \dots & \dots & 1 & 0 & \dots & \dots & \dots & \dots & \dots & 0 \\ 0 & -\rho_p & \dots & \dots & \dots & \dots & -\rho_1 & 1 & & & & & & 0 \\ \vdots & & & & & & & \ddots & & & & & & \vdots \\ \vdots & & & & & & \vdots & & \ddots & & & & & \vdots \\ \vdots & & & & & & \vdots & & & \ddots & & & & \vdots \\ \vdots & & & & & & \vdots & & & & 1 & 0 & & \\ 0 & 0 & \dots & \dots & \dots & \dots & 0 & 0 & \dots & -\rho_p & \dots & \dots & -\rho_1 & 1 \end{bmatrix}$$

and  $N$  is an  $n \times p$  matrix that has all elements zero except for its top  $p \times p$  block which has the triangular form

$$\begin{bmatrix} -\rho_p & -\rho_{p-1} & \cdots & \cdots & \cdots & \cdots & -\rho_1 \\ 0 & -\rho_p & & & & & -\rho_2 \\ 0 & 0 & & & & & \\ & & \ddots & & & & \\ & & & \ddots & & & \\ & & & & \ddots & & \\ 0 & 0 & & & 0 & -\rho_p \end{bmatrix}$$

As noted by Ara (1995), in this case  $L(\rho, \gamma)^{-1}$  has the same form as  $P$  but with the top left  $p \times p$  block replaced by the lower triangular matrix

$$\begin{bmatrix} a_{11} & 0 & \cdots & \cdots & \cdots & 0 \\ a_{21} & a_{22} & & & & 0 \\ \vdots & & \ddots & & & \vdots \\ \vdots & & & \ddots & & \vdots \\ \vdots & & & & \ddots & \vdots \\ a_{p1} & a_{p2} & \cdots & \cdots & \cdots & a_{pp} \end{bmatrix}.$$

The  $a_{ij}$  values can be calculated recursively in the following order

$$a_{pp} = (1 - \rho_p^2)^{1/2},$$

$$a_{p,p-i} = (-\rho_i - \rho_{p-i}\rho_p)/a_{pp}, \quad i = 1, \dots, p-1,$$

for  $j = p-1, p-2, \dots, 2$  with  $m = p-j$

$$a_{jj} = \left( 1 + \sum_{i=1}^{j-1} \rho_i^2 - \sum_{i=m+1}^p \rho_i^2 - \sum_{i=1}^m a_{j+i,j}^2 \right)^{1/2},$$

for  $\ell = 1, \dots, j-2$

$$a_{j,j-\ell} = \left( -\rho_\ell - \sum_{i=1}^{j-\ell-1} \rho_i \rho_{i+\ell} - \sum_{i=j+1}^p a_{ij} a_{i,j-\ell} \right) / a_{jj}$$

and

$$a_{j1} = \left( -\rho_{j-1} - \rho_{m+1} \rho_p - \sum_{i=j+1}^p a_{ij} a_{i1} \right) / a_{jj};$$

and

$$a_{11} = \left( 1 - \rho_p^2 - \sum_{i=1}^{p-1} a_{i+1,1}^2 \right)^{1/2}.$$

Note that in this case

$$|L(\rho, \gamma)|^{-1} = \prod_{i=1}^p a_{ii}.$$

For MA( $q$ ) disturbances, it is more convenient to construct  $L(\rho, \gamma)$  which is the band triangular matrix:

$$L(\rho, \gamma) = \begin{bmatrix} a_{11} & 0 & 0 & & \dots & \dots & \dots & \dots & 0 \\ a_{21} & a_{22} & 0 & & & & & & 0 \\ a_{31} & a_{32} & a_{33} & & & & & & 0 \\ \vdots & \vdots & \vdots & \ddots & & & & & \vdots \\ \vdots & \vdots & \vdots & & \ddots & & & & \vdots \\ \vdots & \vdots & \vdots & & & \ddots & & & \vdots \\ a_{q+1,1} & a_{q+1,2} & a_{q+1,3} & & & & a_{q+1,q+1} & & 0 \\ 0 & a_{q+2,2} & a_{q+2,3} & & & & & & 0 \\ \vdots & & & & & & \ddots & & \vdots \\ 0 & & & & & & & \ddots & 0 \\ 0 & 0 & 0 & \dots & \dots & \dots & a_{n,n-q} & \dots & a_{nn} \end{bmatrix}. \quad (3.34)$$

We solve for the nonzero elements of this matrix recursively via

$$w_0 = (1 + \gamma_1^2 + \dots + \gamma_q^2)$$

$$w_i = (\gamma_i + \gamma_1 \gamma_{i+1} + \dots + \gamma_{q-i} \gamma_q), \quad i = 1, \dots, q,$$

$$a_{11} = w_0^{1/2}$$

$$a_{i+1,1} = w_i / a_{11}, \quad i = 1, \dots, q,$$

$$a_{jj} = \left( w_0 - \sum_{i=1}^{j-1} a_{ji}^2 \right)^{1/2}$$

$$a_{i-j} = \left( w_{i-j} - \sum_{\ell=1}^{j-1} a_{j\ell} a_{i\ell} \right) / a_{jj}, \quad i = j+1, \dots, q+j,$$

for  $j = 2, \dots, q - 1$  (using  $a_{ij} = 0$  for all zero values of the elements of the above matrix in these formulae). For  $i = q, \dots, n$

$$a_{ii} = \left( w_0 - \sum_{j=1}^{q-1} a_{i,i-j}^2 \right)^{1/2}$$

and

$$a_{ji} = \left( w_{j-i} - \sum_{\ell=i+1}^{j-q+1} a_{i\ell} a_{j\ell} \right) / a_{ii}, \quad j = i + 1, \dots, i + q.$$

Observe that given the band triangular nature of (3.34), the transformation of  $\mathbf{z}$  to  $\mathbf{z}^* = L(\rho, \gamma)^{-1}\mathbf{z}$ , where  $\mathbf{z}$  denotes an  $n \times 1$  vector (either  $\mathbf{y}$  in order to get  $\mathbf{y}^*(\rho, \gamma)$  or a column of  $\mathbf{X}$  in order to compute  $\mathbf{X}^*(\rho, \gamma)$ ) can be performed recursively by

$$\begin{aligned} z_1^* &= z_1 / a_{11} \\ z_i^* &= \left( z_i - \sum_{j=1}^{i-1} a_{ij} z_j^* \right) / a_{ii}, \quad i = 2, \dots, q, \\ z_i^* &= \left( z_i - \sum_{j=i-q+1}^{i-1} a_{ij} z_j^* \right) / a_{ii}, \quad i = q + 1, \dots, n. \end{aligned}$$

The generalization of this approach to ARMA( $p, q$ ) disturbances is discussed by Ansley (1979) and in the special case of ARMA(1, 1) disturbances, is outlined by Rahman and King (1993). The above approach to maximum likelihood estimation is a general approach which requires numerical methods for optimization of the profile or concentrated loglikelihood. In the special cases of AR(1) disturbances and AR(2) disturbances, Beach and MacKinnon (1978a, 1978b) have derived an iterative method for solving the first-order conditions for the full maximum likelihood estimates. For further discussion of this and other methods, also see Davidson and MacKinnon (1993).

### 3.2 Maximum marginal likelihood estimation

There is a mounting literature that suggests that the method of maximum likelihood estimation as outlined in Section 3.1 can lead to biased estimates and inaccurate asymptotic test procedures based on these estimates. The problem is that, for the purpose of estimating  $\rho$  and  $\gamma$ ,  $\beta$  and  $\sigma^2$  are nuisance parameters. An early contribution on methods of overcoming the problems of nuisance parameters was made by Kalbfleisch and Sprott (1970) who proposed the use of marginal likelihood estimation. This approach does not work for all cases of nuisance

parameters but fortunately works very well for our problem of estimating  $\rho$  and  $\gamma$  in the presence of  $\beta$  and  $\sigma^2$ .

An important contribution to this literature was made by Tunnicliffe Wilson (1989) who showed that the marginal loglikelihood for  $\rho$  and  $\gamma$  in (3.22) is

$$f_m(\rho, \gamma) = -\frac{1}{2} \log |\Omega(\rho, \gamma)| - \frac{1}{2} \log |X^*(\rho, \gamma)' X^*(\rho, \gamma)| - \frac{n-k}{2} \log e^*(\rho, \gamma)' e^*(\rho, \gamma) \quad (3.35)$$

where  $X^*(\rho, \gamma)$  and  $e^*(\rho, \gamma)$  are given by (3.28) and (3.31), respectively. Maximum marginal likelihood (MML) estimates can be obtained from maximizing (3.35) with respect to  $\rho$  and  $\gamma$ . The various tricks outlined in Section 3.1 can be used to evaluate (3.35) for given  $\rho$  and  $\gamma$  in an efficient manner.

Levenbach (1972) considered MML estimation for the AR(1) model and Cooper and Thompson (1977) demonstrated its use reduces estimation bias for  $\gamma_1$  in the MA(1) model. Cordua (1986) demonstrated that MML estimation removes estimation bias in estimates of  $\rho_1$  when estimating regressions with trending regressors and an AR(1) error term. Tunnicliffe Wilson (1989) also presented evidence that the MML reduces estimation bias. In addition, see Rahman and King (1998) and Laskar and King (1998). The evidence is clear. In order to reduce estimation bias in estimates of  $\rho$  and  $\gamma$ , it is better to use the MML rather than the profile or concentrated likelihood.

## 4 HYPOTHESIS TESTING

Much has been written on the problem of testing for serial correlation, particularly in the disturbances of the linear regression model. Given the nonexperimental nature of almost all economic data and also the strong potential for disturbances to be autocorrelated in economic time series applications, it is extremely important to test for autocorrelation in this context.

### 4.1 The Durbin–Watson test

The von Neumann (1941, 1942) ratio is an important test of the independence of successive Gaussian time series observations, with unknown but constant mean. Its test statistic is of the form

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

where  $\bar{y}$  is the sample mean. Hart (1942a, 1942b) tabulated the distribution and critical values of  $\eta$  under the null hypothesis of independent Gaussian  $y_t$ s.

The most well known test for autocorrelation in regression disturbances is the Durbin–Watson (DW) test. Durbin and Watson (1950, 1951, 1971) considered the problem of testing  $H_0 : \rho = 0$  in the linear regression model (3.12) with stationary AR(1) disturbances (3.13) and Gaussian errors, i.e.  $\varepsilon_t \sim IN(0, \sigma^2)$ . The Durbin–Watson test statistic is of the form

$$d_1 = \sum_{t=2}^n (e_t - e_{t-1})^2 \Bigg/ \sum_{t=1}^n e_t^2$$

where  $e$  is the OLS residual vector from (3.12). Unfortunately the null distribution of  $d_1$  is a function of the  $X$  matrix through the projection matrix  $M = I_n - X(X'X)^{-1}X'$ . This meant that a single set of critical values could not be tabulated as was the case for the von Neumann ratio.

Durbin and Watson (1951) overcame this problem by tabulating bounds for the critical values. When the calculated value of the test statistic is between the two bounds, the test is inconclusive. A number of approximations to the DW critical values have been suggested, see King (1987). There are also extensive alternative tables of bounds, see Savin and White (1977), Farebrother (1980) and King (1981). These days computer methods such as simulation or Imhof's (1961) algorithm allow  $p$ -values for the test to be calculated as a matter of routine.

The Durbin–Watson test has a number of remarkable small sample power properties. The test is approximately uniformly most powerful invariant (UMPI) when the column space of  $X$  is spanned by  $k$  of the eigenvectors of the tridiagonal matrix,

$$A_1 = \begin{bmatrix} 1 & -1 & 0 & \cdots & \cdots & \cdots & 0 & 0 \\ -1 & 2 & -1 & & & & & 0 \\ 0 & -1 & 2 & & & & & 0 \\ \vdots & & & \ddots & & & & \vdots \\ \vdots & & & & \ddots & & & \vdots \\ \vdots & & & & & \ddots & & \vdots \\ \vdots & & & & & & 2 & -1 \\ 0 & 0 & 0 & & & & -1 & 1 \end{bmatrix}$$

and approximately locally best invariant (LBI) against AR(1) disturbances. The exact LBI test can be obtained by adding the first and last OLS residual squared to the numerator of the DW statistic. This is known as the modified DW test. The modified DW test is also LBI against MA(1) errors and is uniformly LBI against ARMA(1, 1) disturbances and sums of independent ARMA(1, 1) error components. Consequently, the DW test is approximately LBI or approximately uniformly LBI against these disturbance processes (see King and Evans, 1988). In summary, the literature suggests that the DW test generally has good power against any form of serial correlation provided there is a strong first-order component present.

## 4.2 Some other tests

The DW test has low power when testing against the simple AR(4) disturbance process given by (3.14) with Gaussian errors. Wallis (1972) and Vinod (1973) separately developed the fourth-order analogue to the DW test which has the test statistic

$$d_4 = \frac{\sum_{t=5}^n (e_t - e_{t-4})^2}{\sum_{t=1}^n e_t^2}.$$

This has been a popular test for use with quarterly time series data and, as Vinod (1973) demonstrated, can be easily generalized to test against any simple AR( $p$ ) disturbance process.

Another weakness of the DW test against AR(1) disturbances is the potential for its power to decline to zero as  $\rho_1$  tends to one. This was first noted by Tillman (1975) and explained convincingly by Bartels (1992). It means that the DW test can be at its weakest (in terms of power) just when it is needed most. The class of point optimal tests (King, 1985) provide a solution to this problem. They allow the researcher to fix a value of  $\rho_1$  at which power is optimized. For some  $X$  matrices (similar to those for the DW test), this test is approximately UMPI, so optimizing power at a particular  $\rho_1$  value does not mean the test has low power for other  $\rho_1$  values.

The Lagrange multiplier (LM) test is a popular test for AR( $p$ ), MA( $q$ ) or ARMA( $p, q$ ) regression disturbances. A strength of the DW test is that it is one-sided in the sense that it can be applied to test for either positive or negative autocorrelation by use of the appropriate tail of the distribution of  $d_1$  under  $H_0$ . The LM test is a general two-sided test and therefore cannot be expected to be as powerful against specific forms of autocorrelation. For further discussion, see Godfrey (1988) and the references therein.

Of course any of the classical tests, such as the likelihood ratio, Wald, and LM tests, can be applied to these testing problems. They can also be used to test one form of disturbance process against a more general form. Again there is mounting evidence to suggest that in terms of accuracy (more accurate sizes and better centered power curves), it is better to construct these tests using the marginal likelihood rather than the full likelihood function (see Rahman and King, 1998; Laskar and King, 1998).

With respect to non-nested testing of disturbance processes, there is a growing literature on testing MA(1) disturbances against AR(1) disturbances and vice versa. See for example King and McAleer (1987), Godfrey and Tremayne (1988) and Silvapulle and King (1991).

## 4.3 Testing disturbances in the dynamic linear regression model

Finally we turn our attention to the problem of testing for autocorrelation in the disturbances of the dynamic regression model (3.19). Whether the DW test can

be used in these circumstances has been an area of controversy in the literature. Durbin and Watson (1950) and others have warned against its use in the dynamic model and this is a theme taken up by many textbooks until recently. The problem has been one of finding appropriate critical values.

Based on the work of Inder (1985, 1986), King and Wu (1991) observed that the small disturbance distribution (the limiting distribution as  $\sigma^2 \rightarrow 0$ ) of the DW statistic is the exact distribution of  $d_1$  for the corresponding regression with the lagged dependent variables replaced by their expected values. This provides a justification for the use of the familiar tables of bounds when the DW test is applied to a dynamic regression model. It also highlights a further difficulty. Because  $E(y_{t-1})$  is a function of the regression parameters, it is clear the null distribution of the DW test also depends on these parameters, so the test can have different sizes for different parameter values under the null hypothesis. It appears this is a property shared by many alternative tests.

Durbin (1970) suggested his  $h$ -test and  $t$ -test as alternatives to the DW test. The  $h$ -test can suffer from problems in small samples caused by the need to take the square root of what can sometimes be a negative estimate of a variance. The  $t$ -test appears more reliable and can be conducted in a simple manner. Let  $e_1, \dots, e_n$  be the OLS residuals from (3.19).  $H_0 : \rho = 0$  can be tested using OLS regression to test the significance of the coefficient of  $e_{t-1}$  in the regression of  $e_t$  on  $e_{t-1}, y_{t-1}, y_{t-2}, \dots, x_t$ .

That this can be a difficult testing problem is best illustrated by considering (3.18) and (3.13). If  $x_t$  is lag invariant then if we switch  $\rho$  and  $\alpha_1$ , we will end up with the same value of the likelihood function which indicates a local identification problem. For near lag invariant  $x_t$  vectors, it is therefore difficult to distinguish between  $\rho$  and  $\alpha_1$ . This causes problems for the small-sample properties of asymptotic tests in this case and explains why it has been difficult to find a satisfactory test for this testing problem.

## 5 MODEL SELECTION

A difficult question when modeling economic behavior is to decide on what lags should be in the ARIMA model, the ARMA disturbance model, or the dynamic regression model. It is tempting to use hypothesis testing to help make model specification decisions based on the data, but as discussed by Granger, King, and White (1995), there are disadvantages in doing so. They and others recommend the use of a model selection procedure to make these decisions, the most common of which are information criteria (IC). For each of the models under consideration, one calculates the maximized loglikelihood and then penalizes this value to take account of the number of free parameters in the model which we will denote by  $j$ . Akaike's (1973) IC (AIC) uses  $j$  as the penalty whereas Schwarz's (1978) Bayesian IC (BIC) uses  $j\log(n)/2$ . There are a range of other IC procedures, but these two have become the most popular.

These days, BIC seems to be the favored procedure because it is consistent, which means that as the sample size goes to infinity, the probability that it will choose the correct model from a finite number of models goes to one. An unfortunate consequence of this property is that in small samples, BIC tends

to wrongly choose underfitting models and can have very low probabilities of correctly selecting a model which has a large number of free parameters. AIC seems more balanced in this regard in small samples, but can suffer from a tendency to overfit in larger samples.

## References

- Akaike, H. (1973). Information theory and an extension of the maximum likelihood principle. In B.N. Petro and F. Csaki (eds.) *2nd International Symposium on Information Theory*. Budapest: Akademiai Kiado, 267–81.
- Ansley, C.F. (1979). An algorithm for the exact likelihood of a mixed autoregressive – moving average process. *Biometrika* 66, 59–65.
- Ara, I. (1995). Marginal likelihood based tests of regression disturbances. Unpublished Ph.D. thesis, Monash University.
- Baillie, R.T. (1996). Long memory processes and fractional integration in econometrics. *Journal of Econometrics* 73, 5–59.
- Bartels, R. (1992). On the power function of the Durbin–Watson test. *Journal of Econometrics* 51, 101–12.
- Beach, C.M., and J. MacKinnon (1978a). Full maximum likelihood procedure for regression with autocorrelated errors. *Econometrica* 46, 51–8.
- Beach, C.M., and J. MacKinnon (1978b). Full maximum likelihood estimation of second-order autoregressive error models. *Journal of Econometrics* 7, 187–98.
- Box, G.E.P., and G.M. Jenkins (1970). *Time Series Analysis, Forecasting and Control*. San Francisco: Holden-Day.
- Box, G.E.P., and G.M. Jenkins (1976). *Time Series Analysis, Forecasting and Control*, 2nd edn. San Francisco: Holden-Day.
- Cochrane, D., and G. Orcutt (1949). Application of least squares regression to relationships containing autocorrelated error terms. *Journal of the American Statistical Association* 44, 32–61.
- Cooper, D.M., and R. Thompson (1977). A note on the estimation of parameters of the autoregressive-moving average process. *Biometrika* 64, 625–8.
- Corduas, M. (1986). The use of the marginal likelihood in testing for serial correlation in time series regression. Unpublished M.Phil. thesis, University of Lancaster.
- Davidson, R., and J.G. MacKinnon (1993). *Estimation and Inference in Econometrics*. Oxford: Oxford University Press.
- Durbin, J. (1970). Testing for serial correlation in least squares regression when some of the regressors are lagged dependent variables. *Econometrica* 38, 410–21.
- Durbin, J., and G.S. Watson (1950). Testing for serial correlation in least squares regression I. *Biometrika* 37, 409–28.
- Durbin, J., and G.S. Watson (1951). Testing for serial correlation in least squares regression II. *Biometrika* 38, 159–78.
- Durbin, J., and G.S. Watson (1971). Testing for serial correlation in least squares regression III. *Biometrika* 58, 1–19.
- Farebrother, R.W. (1980). The Durbin–Watson test for serial correlation when there is no intercept in the regression. *Econometrica* 48, 1553–63 and 49, 227.
- Godfrey, L.G. (1988). *Misspecification Tests in Econometrics: The Lagrange Multiplier Principle and Other Approaches*. Cambridge: Cambridge University Press.
- Godfrey, L.G., and A.R. Tremayne (1988). Checks of model adequacy for univariate time series models and their application to econometric relationships. *Econometric Reviews* 7, 1–42.

- Granger, C.W.J., M.L. King, and H. White (1995). Comments on testing economic theories and the use of model selection criteria. *Journal of Econometrics* 67, 173–87.
- Hart, B.I. (1942a). Tabulation of the probabilities for the ratio of the mean square successive difference to the variance. *Annals of Mathematical Statistics* 13, 207–14.
- Hart, B.I. (1942b). Significance levels for the ratio of the mean square successive difference to the variance. *Annals of Mathematical Statistics* 13, 445–7.
- Imhof, P.J. (1961). Computing the distribution of quadratic forms in normal variables. *Biometrika* 48, 419–26.
- Inder, B.A. (1985). Testing for first-order autoregressive disturbances in the dynamic linear regression model. Unpublished Ph.D. thesis, Monash University.
- Inder, B.A. (1986). An approximation to the null distribution of the Durbin–Watson statistic in models containing lagged dependent variables. *Econometric Theory* 2, 413–28.
- Johnston, J. (1972). *Econometric Methods*, 2nd edn. New York: McGraw-Hill.
- Kalbfleisch, J.D., and D.A. Sprott (1970). Application of likelihood methods to models involving large numbers of parameters. *Journal of the Royal Statistical Society B* 32, 175–94.
- King, M.L. (1981). The Durbin–Watson test for serial correlation: Bounds for regressions with trend and/or seasonal dummy variables. *Econometrica* 49, 1571–81.
- King, M.L. (1985). A point optimal test for autoregressive disturbances. *Journal of Econometrics* 27, 21–37.
- King, M.L. (1987). Testing for autocorrelation in linear regression models: A survey. In M.L. King and D.E.A. Giles (eds.) *Specification Analysis in the Linear Model*, London: Routledge and Kegan Paul, 19–73.
- King, M.L., and M.A. Evans (1988). Locally optimal properties of the Durbin–Watson test. *Econometric Theory* 4, 509–16.
- King, M.L., and M. McAleer (1987). Further results on testing AR(1) against MA(1) disturbances in the linear regression model. *Review of Economic Studies* 54, 649–63.
- King, M.L., and P.X. Wu (1991). Small-disturbance asymptotics and the Durbin–Watson and related tests in the dynamic regression model. *Journal of Econometrics* 47, 145–52.
- Laskar, M.R., and M.L. King (1998). Estimation and testing of regression disturbances based on modified likelihood functions. *Journal of Statistical Planning and Inference* 71, 75–92.
- Levenbach, H. (1972). Estimation of autoregressive parameters from a marginal likelihood function. *Biometrika* 59, 61–71.
- Nichols, D.F., A.R. Pagan, and R.D. Terrell (1975). The estimation and use of models with moving average disturbance terms: A survey. *International Economic Review* 16, 113–34.
- Prais, S.J., and C.B. Winsten (1954). Trend estimators and serial correlation. Unpublished Cowles Commission Discussion Paper, University of Chicago.
- Rahman, S., and M.L. King (1993). Testing for ARMA(1, 1) disturbances in the linear regression model. *Australian Economic Papers* 32, 284–98.
- Rahman, S., and M.L. King (1998). Marginal-likelihood score-based tests of regression disturbances in the presence of nuisance parameters. *Journal of Econometrics* 82, 81–106.
- Savin, N.E., and K.J. White (1977). The Durbin–Watson test for serial correlation with extreme sample sizes or many regressors. *Econometrica* 45, 1989–96.
- Schwarz, G.W. (1978). Estimating the dimension of a model. *Annals of Statistics* 6, 461–4.
- Silvapulle, P., and M.L. King (1991). Testing moving average against autoregressive disturbances in the linear regression model. *Journal of Business and Economic Statistics* 9, 329–35.
- Thomas, J.J., and K.F. Wallis (1971). Seasonal variation in regression analysis. *Journal of the Royal Statistical Society A* 134, 57–72.

- Tillman, J.A. (1975). The power of the Durbin–Watson test. *Econometrica* 43, 959–74.
- Tse, Y.K. (1982). Edgeworth approximations in first-order stochastic difference equations with exogenous variables. *Journal of Econometrics* 20, 175–95.
- Tunnicliffe Wilson, G. (1989). On the use of marginal likelihood in time series model estimation. *Journal of the Royal Statistical Society B* 51, 15–27.
- Van der Leeuw, J. (1994). The covariance matrix of ARMA errors in closed form. *Journal of Econometrics* 63, 397–405.
- Vinod, H.D. (1973). Generalization of the Durbin–Watson statistic for higher order autoregressive processes. *Communications in Statistics* 2, 115–44.
- Von Neumann, J. (1941). Distribution of the ratio of the mean square successive difference to the variance. *Annals of Mathematical Statistics* 12, 367–95.
- Von Neumann, J. (1942). A further remark concerning the distribution of the ratio of the mean square successive difference to the variance. *Annals of Mathematical Statistics* 13, 86–8.
- Wallis, K.F. (1972). Testing for fourth order autocorrelation in quarterly regression equations. *Econometrica* 40, 617–36.

CHAPTER FOUR

# Heteroskedasticity

*William E. Griffiths\**

## 1 INTRODUCTION

A random variable  $y$  is said to be heteroskedastic if its variance can be different for different observations. Conversely, it is said to be homoskedastic if its variance is constant for all observations. The most common framework in which heteroskedasticity is studied in econometrics is in the context of the general linear model

$$y_i = x'_i \beta + e_i, \quad (4.1)$$

where  $x_i$  is a  $K$ -dimensional vector of observations on a set of explanatory variables,  $\beta$  is a  $K$ -dimensional vector of coefficients which we wish to estimate and  $y_i$  denotes the  $i$ th observation ( $i = 1, 2, \dots, N$ ) on a dependent variable. In a heteroskedastic model the error term  $e_i$  is assumed to have zero-mean and variance  $\sigma_i^2$ , the  $i$  subscript on  $\sigma_i^2$  reflecting possibly different variances for each observation. Conditional on  $x'_i \beta$ , the dependent variable  $y_i$  has mean  $x'_i \beta$  and variance  $\sigma_i^2$ . Thus, in the heteroskedastic general linear model, the mean and variance of the random variable  $y$  can both change over observations.

Heteroskedasticity can arise empirically, through theoretical considerations and from model misspecification. Empirically, heteroskedasticity is often encountered when using cross-sectional data on a number of microeconomic units such as firms or households. A common example is the estimation of household expenditure functions. Expenditure on a commodity is more easily explained by conventional variables for households with low incomes than it is for households with high incomes. The lower predictive ability of the model for high incomes can be captured by specifying a variance  $\sigma_i^2$  which is larger when income is larger. Data on firms invariably involve observations on economic units of varying sizes. Larger firms are likely to be more diverse and flexible with respect to the way in which values for  $y_i$  are determined. This additional diversity is captured through an

error term with a larger variance. Theoretical considerations, such as randomness in behavior, can also lead to heteroskedasticity. Brown and Walker (1989, 1995) give examples of how it arises naturally in demand and production models. Moreover, as chapter 19 in this volume (by Swamy and Tavlas) illustrates, heteroskedasticity exists in all models with random coefficients. Misspecifications such as incorrect functional form, omitted variables and structural change are other reasons that a model may exhibit heteroskedasticity.

In this chapter we give the fundamentals of sampling theory and Bayesian estimation, and sampling theory hypothesis testing, for a linear model with heteroskedasticity. For sampling theory estimation it is convenient to first describe estimation for a known error covariance matrix and to then extend it for an unknown error covariance matrix. No attempt is made to give specific details of developments beyond what we consider to be the fundamentals. However, references to such developments and how they build on the fundamentals are provided. Autoregressive conditional heteroskedasticity (ARCH) which is popular for modeling volatility in time series is considered elsewhere in this volume and not discussed in this chapter.

## 2 SAMPLING THEORY INFERENCE WITH KNOWN COVARIANCE MATRIX

Writing  $y_i = x'_i\beta + e_i$  so that all  $N$  observations are included yields the familiar matrix expression

$$y = X\beta + e, \quad (4.2)$$

where  $y$  and  $e$  are of dimension  $(N \times 1)$  and  $X$  is of dimension  $(N \times K)$ , and rank  $K$ . The assumption of heteroskedastic  $y$  can be written as

$$E[(y - X\beta)(y - X\beta)'] = E[ee'] = V = \sigma^2\Lambda, \quad (4.3)$$

where

$$\begin{aligned} V &= \text{diagonal}(\sigma_1^2, \sigma_2^2, \dots, \sigma_N^2) \\ &= \sigma^2 \text{diagonal}(\lambda_1, \lambda_2, \dots, \lambda_N) \\ &= \sigma^2\Lambda. \end{aligned} \quad (4.4)$$

In equation (4.4) a constant  $\sigma^2$  has been factored out of  $V$  yielding a matrix  $\Lambda$  of ratios  $\lambda_i = \sigma_i^2/\sigma^2$ . This factoring device is useful (i) when  $\Lambda$  is known, but  $\sigma^2$  is not, and (ii) if a heteroskedastic specification has a constant component ( $\sigma^2$  in this case) and a component that varies over observations. The constant that is factored out is arbitrary, and, in practice, is chosen for convenience.

The generalized least squares estimator for  $\beta$  which, from the Gauss–Markov Theorem is known to be the best linear unbiased estimator, is given by

$$\begin{aligned}\hat{\beta} &= (\mathbf{X}' \mathbf{V}^{-1} \mathbf{X})^{-1} \mathbf{X}' \mathbf{V}^{-1} \mathbf{y} = \left( \sum_{i=1}^N \frac{x_i x'_i}{\sigma_i^2} \right)^{-1} \sum_{i=1}^N \frac{x_i y_i}{\sigma_i^2} \\ &= (\mathbf{X}' \Lambda^{-1} \mathbf{X})^{-1} \mathbf{X}' \Lambda^{-1} \mathbf{y} = \left( \sum_{i=1}^N \frac{x_i x'_i}{\lambda_i} \right)^{-1} \sum_{i=1}^N \frac{x_i y_i}{\lambda_i}.\end{aligned}\quad (4.5)$$

The right-hand expressions in equations (4.4) emphasize the *weighted* nature of the generalized least squares estimator. Each observation ( $x_i$  and  $y_i$ ) is weighted by the inverse standard deviation  $\sigma_i^{-1}$ , or a quantity proportional to it,  $\lambda_i^{-1/2}$ . Observations that are less reliable because they come from a distribution with a large variance are weighted less than more reliable observations where  $\sigma_i^2$  is smaller. The mean and covariance matrix of the generalized least squares estimator are given by  $E[\hat{\beta}] = \beta$  and  $V_{\hat{\beta}}$ , respectively, where

$$V_{\hat{\beta}} = (\mathbf{X}' \mathbf{V}^{-1} \mathbf{X})^{-1} = \left( \sum_{i=1}^N \frac{x_i x'_i}{\sigma_i^2} \right)^{-1} = \sigma^2 (\mathbf{X}' \Lambda^{-1} \mathbf{X})^{-1} = \sigma^2 \left( \sum_{i=1}^N \frac{x_i x'_i}{\lambda_i} \right)^{-1}.\quad (4.6)$$

Practical application of (4.5) and (4.6) requires knowledge of at least  $\Lambda$ . For inference purposes, an unbiased estimator for  $\sigma^2$  can be found from

$$\hat{\sigma}^2 = \frac{(y - \mathbf{X}\hat{\beta})' \Lambda^{-1} (y - \mathbf{X}\hat{\beta})}{N - K}.\quad (4.7)$$

Although most applications proceed by refining the specification of  $\Lambda$  into one that contains a reduced number of parameters that is constant for changing sample size, there are some scenarios where knowledge of  $\Lambda$  is a reasonable assumption. To illustrate one such example, suppose that we are interested in an industry cost function that can be written as

$$y_{ij} = x'_{ij} \beta + e_{ij},\quad (4.8)$$

where the double subscript  $(i, j)$  refers to the  $j$ th firm in the  $i$ th industry. Suppose also that the  $e_{ij}$  are independent with  $\text{var}(e_{ij}) = \sigma^2$  (a constant) and that there are  $n_i$  firms in the  $i$ th industry. A model for data obtained by averaging over all firms in each industry is given by

$$\frac{1}{n_i} \sum_{j=1}^{n_i} y_{ij} = \frac{1}{n_i} \sum_{j=1}^{n_i} x'_{ij} \beta + \frac{1}{n_i} \sum_{j=1}^{n_i} e_{ij}$$

or

$$\bar{y}_i = \bar{x}'_i \beta + \bar{e}_i.\quad (4.9)$$

The variance of the error term is

$$\text{var}(\bar{e}_i) = \frac{1}{n_i^2} \sum_{j=1}^{n_i} \text{var}(e_{ij}) = \frac{1}{n_i^2} n_i \sigma^2 = \frac{\sigma^2}{n_i}.$$

That is,  $\bar{e}_i$  is heteroskedastic with its variance depending on the number of firms used to compute the average industry data. Providing this number is available, the matrix  $\Lambda$  is known with its inverse given by

$$\Lambda^{-1} = \text{diagonal}(n_1, n_2, \dots, n_N).$$

The generalized least squares procedure can be applied. It recognizes that industry observations obtained by averaging a large number of firms are more reliable than those obtained by averaging a small number of firms.

To construct confidence intervals for the elements in  $\beta$  or to test hypotheses about the elements in  $\beta$ , one can assume the error vector  $e$  is normally distributed and proceed with finite sample inference procedures, or one can use large sample approximate inference procedures without the assumption of normally distributed errors. When the errors are normally distributed the following results hold:

$$\frac{(N - K)\hat{\sigma}^2}{\sigma^2} \sim \chi^2_{(N-K)} \quad (4.10)$$

$$R\hat{\beta} \sim N[R\beta, \sigma^2 R(X'\Lambda^{-1}X)^{-1}R'] \quad (4.11)$$

$$\frac{(R\hat{\beta} - R\beta)'[R(X'\Lambda^{-1}X)^{-1}R']^{-1}(R\hat{\beta} - R\beta)}{\sigma^2} \sim \chi^2_{(J)} \quad (4.12)$$

$$F = \frac{(R\hat{\beta} - R\beta)'[R(X'\Lambda^{-1}X)^{-1}R']^{-1}(R\hat{\beta} - R\beta)/J}{\hat{\sigma}^2} \sim F_{[J, (N-K)]}. \quad (4.13)$$

In the above expressions  $R$  is a  $(J \times K)$  matrix of rank  $J$  whose elements define the quantity for which inference is sought. These results parallel those for the general linear model with independent, identically distributed error terms, and can be derived from them by using a straightforward transformation. For details of the transformation, and more details of how equations (4.10)–(4.13) are used for hypothesis testing and interval estimation, see, for example, Judge *et al.* (1988, ch. 8). When the errors are not assumed to be normally distributed, approximate large sample inference is based on equation (4.12) with  $\sigma^2$  replaced by  $\hat{\sigma}^2$ .

For inferences about nonlinear functions of the elements of  $\beta$  that cannot be written as  $R\beta$ , we consider functions of the form  $g(\beta) = 0$  where  $g(\cdot)$  is a  $J$ -dimensional vector function. Inference can be based on the approximate result

$$\frac{g(\hat{\beta})'[G(X'\Lambda^{-1}X)^{-1}G']^{-1}g(\hat{\beta})}{\hat{\sigma}^2} \sim \chi^2_{(J)}, \quad (4.14)$$

where  $G$  is the  $(J \times K)$  matrix of partial derivatives, with rank  $J$ ,

$$G = \left. \frac{\partial g(\beta)}{\partial \beta'} \right|_{\hat{\beta}}. \quad (4.15)$$

Three categories of tests frequently used in econometrics are the Wald, Lagrange multiplier, and likelihood ratio tests. In the context of the scenarios discussed so far (hypothesis tests about  $\beta$  in a model with covariance matrix  $\sigma^2 \Lambda$ , with  $\Lambda$  known), all three testing principles lead to the results given above. The only difference is that, in a Lagrange multiplier test, the estimate for  $\sigma^2$  is based on the restricted rather than unrestricted generalized least squares residuals.

Further details on estimation and hypothesis testing for the case of a known error covariance matrix can be found in standard textbooks such as Judge *et al.* (1988, chs. 8, 9), Greene (1997, ch. 12) and Baltagi (1998, chs. 5, 9). Of particular interest might be the consequences of using the ordinary least squares (OLS) estimator  $b = (X'X)^{-1}X'y$  in the presence of heteroskedastic errors. It is well known that, under these circumstances, the OLS estimator is inefficient and that the estimated covariance matrix  $\hat{\sigma}^2(X'X)^{-1}$  is a biased estimate of the true covariance matrix  $\sigma^2(X'X)^{-1}X'\Lambda X(X'X)^{-1}$ . Examples of inefficiencies and bias are given in most textbook treatments.

### 3 SAMPLING THEORY ESTIMATION AND INFERENCE WITH UNKNOWN COVARIANCE MATRIX

Consider again the linear model  $y = X\beta + e$  with error-covariance matrix  $V = \sigma^2 \Lambda$ . In this section we relax the assumption that  $\Lambda$  is known. As we saw in the previous section, there are some circumstances where such an assumption is reasonable. However, there are also many where it is not. For example, in a household expenditure function, we may be willing to assume the variance of expenditure depends on total expenditure and the demographic composition of the household, but not willing to specify the values of parameters that describe the dependence. Thus, we could write, for example,

$$\sigma_i^2 = \theta_0 + \theta_1 z_{1i} + \theta_2 z_{2i}, \quad (4.16)$$

where  $z_{1i}$  and  $z_{2i}$  are total expenditure and demographic composition, respectively, and  $(\theta_0, \theta_1, \theta_2)$  are unknown parameters. If  $(\sigma_1^2, \sigma_2^2, \dots, \sigma_N^2)$  are not known, then some kind of reparameterization such as that in (4.16) is necessary to reduce the number of parameters to a manageable number that does not increase with sample size. We will work in terms of the general notation

$$\sigma_i^2 = \sigma^2 h_i(\alpha) = \sigma^2 h(z'_i \alpha), \quad (4.17)$$

where  $\alpha$  is an  $(S \times 1)$  vector of unknown parameters, and  $h_i(\cdot)$  is a differentiable function of those parameters and an  $(S \times 1)$  vector  $z_i$  which could be identical to or different from  $x_i$ . To write (4.16) in terms of the general notation in (4.17), we re-express it as

$$\sigma_i^2 = \theta_0 \left( 1 + \frac{\theta_1}{\theta_0} z_{1i} + \frac{\theta_2}{\theta_0} z_{2i} \right) = \sigma^2 (1 + \alpha_1 z_{1i} + \alpha_2 z_{2i}) = \sigma^2 h_i(\alpha). \quad (4.18)$$

In this example, and others which we consider,  $h_i(0) = 1$ , implying that  $\alpha = 0$  describes a model with homoskedastic errors.

Several alternative specifications of  $h_i(\alpha)$  have been suggested in the literature. See Judge *et al.* (1985, p. 422) for a review. One of these is that given in (4.18), namely

$$h_i(\alpha) = 1 + \alpha_1 z_{1i} + \dots + \alpha_S z_{Si}. \quad (4.19)$$

This model has been considered by, among others, Goldfeld and Quandt (1972) and Amemiya (1977), and, in the context of a random coefficient model, by Hildreth and Houck (1968) and Griffiths (1972). Note that, if  $(z_{1i}, \dots, z_{Si})$  are non-overlapping dummy variables, then the specification in (4.19) describes a partition of the sample into  $(S+1)$  subsamples, each one with a different error variance. Such a model could be relevant if parts of the sample came from different geographical regions or there exists some other way of naturally creating sample separations. Examples where estimation within this framework has been considered are Griffiths and Judge (1992) and Hooper (1993).

One potential difficulty with the specification in (4.19) is that the requirement  $h_i(\alpha) > 0$  can mean that restrictions must be placed on  $\alpha$  to ensure that negative variances are not possible. Two possible specifications which avoid this problem are

$$h_i(\alpha) = (1 + \alpha_1 z_{1i} + \dots + \alpha_S z_{Si})^2 \quad (4.20)$$

and

$$h_i(\alpha) = \exp(\alpha_1 z_{1i} + \dots + \alpha_S z_{Si}). \quad (4.21)$$

The specification in (4.20) has received attention from Rutenmiller and Bowers (1968) and Jobson and Fuller (1980). The specification in (4.21) was introduced by Harvey (1976) under the heading “multiplicative heteroskedasticity.” For applications and extensions, see Griffiths and Anderson (1982), and Hill *et al.* (1997).

A class of models which has been popular, but which does not fit within the framework of equation (4.17), is that where the location parameter vector  $\beta$  also appears within the variance function. Authors who have considered this class of models under varying degrees of generality include Amemiya (1973), Jobson and Fuller (1980), and Welsh *et al.* (1994).

### 3.1 Maximum likelihood estimation

Two-step estimation of heteroskedastic error models was popular prior to the development of modern software. These techniques use the residuals from least

squares estimation to estimate  $\alpha$ , and then use the estimate of  $\alpha$  in a generalized least squares estimator. See Judge *et al.* (1985, pp. 431–41) for details. However, it is now more common to assume normally distributed errors and proceed with maximum likelihood estimation. Working in this direction, the loglikelihood function can be written as

$$\begin{aligned} L &= -\frac{N}{2} \ln(2\pi) - \frac{N}{2} \ln(\sigma^2) - \frac{1}{2} \sum_{i=1}^N \ln(h_i(\alpha)) - \frac{1}{2\sigma^2} \sum_{i=1}^N \frac{(y_i - x'_i \beta)^2}{h_i(\alpha)} \\ &= -\frac{N}{2} \ln(2\pi) - \frac{N}{2} \ln(\sigma^2) - \frac{1}{2} \ln|\Lambda| - \frac{1}{2\sigma^2} (y - X\beta)' \Lambda^{-1} (y - X\beta). \end{aligned} \quad (4.22)$$

Differentiating this function with respect to  $\sigma^2$  and  $\beta$  and setting these derivatives equal to zero gives the results

$$\hat{\beta}(\alpha) = (X'\Lambda^{-1}X)^{-1}X'\Lambda^{-1}y \quad (4.23)$$

$$\hat{\sigma}^2(\alpha) = \frac{1}{N} (y - X\hat{\beta}(\alpha))' \Lambda^{-1} (y - X\hat{\beta}(\alpha)). \quad (4.24)$$

Since these estimators are conditional on knowing  $\alpha$ , they resemble those provided in the previous section, the only difference being the divisor  $N$ , instead of  $(N - K)$ , in equation (4.24).

Differentiating  $L$  with respect to  $\alpha$  yields

$$\frac{\partial L}{\partial \alpha} = -\frac{1}{2} \sum_{i=1}^N \frac{1}{h_i(\alpha)} \frac{\partial h_i}{\partial \alpha} + \frac{1}{2\sigma^2} \sum_{i=1}^N \frac{(y_i - x'_i \beta)^2}{[h_i(\alpha)]^2} \frac{\partial h_i}{\partial \alpha}. \quad (4.25)$$

Setting this derivative equal to zero does not yield a convenient solution for  $\alpha$ , although it does simplify for specific definitions of  $h_i(\alpha)$ . For example, for the specification in equation (4.21), written in matrix algebra notation as  $h_i(\alpha) = \exp\{z'_i \alpha\}$ , equating (4.25) to zero yields

$$\sum_{i=1}^N \frac{(y_i - x'_i \beta)^2}{\exp(z'_i \alpha)} z_i = \sigma^2 \sum_{i=1}^N z_i. \quad (4.26)$$

Substituting (4.23) and (4.24) into (4.22) yields the concentrated loglikelihood function

$$L^*(\alpha) = \text{constant} - \frac{N}{2} \ln [(y - X\hat{\beta}(\alpha))' \Lambda^{-1} (y - X\hat{\beta}(\alpha))] - \frac{1}{2} \ln |\Lambda|. \quad (4.27)$$

Thus, maximum likelihood estimation can proceed by numerically finding the value of  $\hat{\alpha}$  that maximizes  $L^*$ , and then substituting that value into equations (4.23) and (4.24).

The information matrix is given by

$$\begin{aligned}
 I(\beta, \alpha, \sigma^2) &= -E \left[ \begin{array}{ccc} \frac{\partial^2 L}{\partial \beta \partial \beta'} & \frac{\partial^2 L}{\partial \beta \partial \alpha'} & \frac{\partial^2 L}{\partial \beta \partial \sigma^2} \\ \frac{\partial^2 L}{\partial \alpha \partial \beta'} & \frac{\partial^2 L}{\partial \alpha \partial \alpha'} & \frac{\partial^2 L}{\partial \alpha \partial \sigma^2} \\ \frac{\partial^2 L}{\partial \sigma^2 \partial \beta'} & \frac{\partial^2 L}{\partial \sigma^2 \partial \alpha'} & \frac{\partial^2 L}{\partial (\sigma^2)^2} \end{array} \right] \\
 &= \left[ \begin{array}{ccc} \frac{X' \Lambda^{-1} X}{\sigma^2} & 0 & 0 \\ 0 & \frac{1}{2} \sum_{i=1}^N \frac{1}{[h_i(\alpha)]^2} \frac{\partial h_i}{\partial \alpha} \frac{\partial h_i}{\partial \alpha'} & \frac{1}{\partial \sigma^2} \sum_{i=1}^N \frac{1}{h_i(\alpha)} \frac{\partial h_i}{\partial \alpha} \\ 0 & \frac{1}{2\sigma^2} \sum_{i=1}^N \frac{1}{h_i(\alpha)} \frac{\partial h_i}{\partial \alpha'} & \frac{N}{2\sigma^4} \end{array} \right] \quad (4.28)
 \end{aligned}$$

The inverse of this matrix is the asymptotic covariance matrix for the maximum likelihood estimators of  $\beta$ ,  $\alpha$  and  $\sigma^2$ . Its block-diagonal nature means the asymptotic covariance matrix for the maximum likelihood estimator  $\hat{\beta}$  is given by the familiar expression

$$V_{\hat{\beta}} = \sigma^2 (X' \Lambda^{-1} X)^{-1}.$$

Equation (4.28) can be simplified considerably once  $h_i(\alpha)$  is specified explicitly. For example, for the case where  $h_i(\alpha) = \exp\{z'_i \alpha\}$ ,

$$I(\beta, \alpha, \sigma^2) = \left[ \begin{array}{ccc} \frac{X' \Lambda^{-1} X}{\sigma^2} & 0 & 0 \\ 0 & \frac{1}{2} Z' Z & \frac{1}{\partial \sigma^2} \sum z_i \\ 0 & \frac{1}{2\sigma^2} \sum z'_i & \frac{N}{2\sigma^4} \end{array} \right] \quad (4.29)$$

where  $Z' = (z_1, z_2, \dots, z_N)$ .

### 3.2 Testing hypotheses about $\beta$

To test hypotheses about  $\beta$ , the large sample result given in equation (4.12) can be used. The only differences are that  $\hat{\beta}$  becomes the maximum likelihood estimator, and  $\sigma^2$  and  $\Lambda$  are replaced by their maximum likelihood estimators.

Using the maximum likelihood estimator for  $\beta$  and its corresponding covariance matrix to test hypotheses about  $\beta$  requires knowledge of the function  $h_i(\alpha)$ . That is, the form of the heteroskedasticity is required. As an alternative, tests can be based on the ordinary least squares estimator for  $\beta$  and an estimate of its covariance matrix. Specifically, the least squares estimator  $b = (X'X)^{-1}X'y$  has covariance matrix

$$V_b = (X'X)^{-1} X' V X (X'X)^{-1},$$

which White (1980) has shown can be consistently estimated by

$$\hat{V}_b = (X'X)^{-1} X' \hat{V} X (X'X)^{-1},$$

where  $\hat{V}$  is a diagonal matrix containing the squares of the ordinary least squares residuals. Thus, the result

$$(Rb - R\beta)'(R\hat{V}_b R')^{-1}(Rb - R\beta) \stackrel{d}{\sim} \chi_{(J)}^2$$

can be used to make approximate inferences about  $\beta$ . The finite sample properties of such inferences have been questioned however and ways for improving the test have been investigated. For access to the literature on this issue see Keener *et al.* (1991) and Davidson and MacKinnon (1993).

### 3.3 Testing for heteroskedasticity

Assuming that  $h_i(\alpha)$  is such that  $h_i(0) = 1$ , tests for heteroskedasticity can be formulated in terms of the hypotheses

$$H_0 : \alpha = 0 \quad H_1 : \alpha \neq 0.$$

We will describe the likelihood ratio, Wald, and Lagrange multiplier test statistics for these hypotheses, and then refer to other tests and evaluations that have appeared in the literature.

Using equation (4.27), the *likelihood ratio (LR) test* statistic is given by

$$\begin{aligned} \gamma_{LR} &= 2[L(\hat{\alpha}) - L(0)] \\ &= N \ln \left( \frac{\hat{e}'_0 \hat{e}_0}{\hat{e}' \hat{\Lambda}^{-1} \hat{e}} \right) - \sum_{i=1}^N \ln [h_i(\hat{\alpha})], \end{aligned} \tag{4.30}$$

where  $\hat{e}_0 = y - Xb$  are the OLS residuals and  $\hat{e} = y - X\hat{\beta}(\hat{\alpha})$  are the maximum likelihood residuals. When the null hypothesis of homoskedasticity holds,  $\gamma_{LR}$  has an approximate  $\chi_{(s)}^2$  distribution.

The *Wald (W) test* statistic is given by

$$\gamma_W = \hat{\alpha}' \hat{V}_{\hat{\alpha}}^{-1} \hat{\alpha}, \tag{4.31}$$

where, applying partitioned-inverse results to equation (4.28), it can be shown that

$$V_{\hat{\alpha}}^{-1} = \frac{1}{2} \sum_{i=1}^N \frac{1}{[h_i(\alpha)]^2} \frac{\partial h_i}{\partial \alpha} \frac{\partial h_i}{\partial \alpha'} - \frac{1}{2N} \left( \sum_{i=1}^N \frac{1}{h_i(\alpha)} \frac{\partial h_i}{\partial \alpha} \right) \left( \sum_{i=1}^N \frac{1}{h_i(\alpha)} \frac{\partial h_i}{\partial \alpha'} \right). \quad (4.32)$$

The statistic  $\gamma_W$  has an approximate  $\chi_{(S)}^2$  distribution when  $\alpha = 0$ .

The *Lagrange multiplier (LM) test* statistic is given by

$$\gamma_{LM} = s_0' I_0^{-1}(\alpha, \sigma^2) s_0, \quad (4.33)$$

where

$$s_0 = \begin{pmatrix} \partial L / \partial \alpha \\ \partial L / \partial \sigma^2 \end{pmatrix} \text{ is evaluated at } \alpha = 0, \sigma^2 = \hat{\sigma}^2(0) \text{ and } \beta = b.$$

$I_0^{-1}(\alpha, \sigma^2)$  is the inverse of the bottom-right block in equation (4.28), evaluated at

$$\alpha = 0 \text{ and } \sigma^2 = \hat{\sigma}^2(0).$$

Recognizing that  $\partial h_i / \partial \alpha$  evaluated at  $\alpha = 0$  is equal to  $z_i$ , equation (4.25) can be used to yield

$$s_0 = \begin{pmatrix} \frac{1}{2} \sum_{i=1}^N z_i \left( \frac{\hat{e}_{0i}^2}{\hat{\sigma}_0^2} - 1 \right) \\ 0 \end{pmatrix} \quad (4.34)$$

where  $\hat{\sigma}_0^2 = \sum_{i=1}^N \hat{e}_{0i}^2 / N = \hat{\sigma}^2(0)$ .

Then, utilizing (4.32) evaluated at  $\alpha = 0$ , the Lagrange multiplier statistic becomes

$$\gamma_{LM} = \frac{\sum_{i=1}^N z_i' (\hat{e}_{0i}^2 - \hat{\sigma}_0^2) \left( \sum_{i=1}^N (z_i - \bar{z})(z_i - \bar{z})' \right)^{-1} \sum_{i=1}^N z_i (\hat{e}_{0i}^2 - \hat{\sigma}_0^2)}{2\hat{\sigma}_0^4}.$$

This statistic is conveniently calculated as one-half of the regression sum-of-squares of  $\hat{e}_{0i}^2 / \hat{\sigma}_0^2$  on  $z_i$  and a constant term. It has an approximate  $\chi_{(S)}^2$  distribution under  $H_0 : \alpha = 0$ . The Lagrange multiplier test statistic was derived by Breusch and Pagan (1979) and Godfrey (1978). To make it more robust to departures from normality, replacement of the denominator  $2\hat{\sigma}_0^4$  by  $N^{-1} \sum_{i=1}^N (\hat{e}_{0i}^2 - \hat{\sigma}_0^2)^2$  has been suggested (Koenker and Bassett, 1982).

Many more tests for heteroskedasticity have been suggested in the literature. See Pagan and Pak (1993) for a review and for details on how the various tests

can be classified as conditional moment tests. One popular test that we have not yet mentioned is the Goldfeld–Quandt (1965) test which uses the error variances from two separate least squares regressions to construct a finite sample  $F$ -statistic. Other classes of tests have been described by Szroeter (1978) and Farebrother (1987). Lee (1992) suggests a test where the mean function is estimated nonparametrically and hence does not have to be precisely specified. Orme (1992) describes tests in the context of censored and truncated regression models. Also, tests for heteroskedasticity in these and other nonlinear models, such as discrete choice models and count data models, are reviewed by Pagan and Pak (1993). Numerous Monte Carlo studies have compared the finite sample size and power of existing and new test statistics. Typically, authors uncover problems with existing test statistics such as poor finite sample size or power, or lack of robustness to misspecification and nonnormality, and suggest alternatives to correct for such problems. A study by Godfrey and Orme (1999) suggests that bootstrapping leads to favorable outcomes. Other examples of Monte Carlo studies that have appeared are Evans and King (1988), Griffiths and Surekha (1986), Griffiths and Judge (1992) and Godfrey (1996). See Farebrother (1987) for some insightful comments on the results of Griffiths and Surekha (1986).

### 3.4 Estimation with unknown form of heteroskedasticity

The work of White (1980) on testing for heteroskedasticity and testing hypotheses about  $\beta$  without specifying the precise form of the heteroskedasticity motivated others to seek *estimators* for  $\beta$  that did not require specification of the form of heteroskedasticity. Attempts have been made to specify estimators which are more efficient than OLS, while at the same time recognizing that the efficiency of GLS may not be achievable (Cragg, 1992; Amemiya, 1983). Carroll (1982) and Robinson (1987) develop adaptive estimators that assume no particular form of heteroskedasticity but nevertheless have the same asymptotic distribution as the generalized least squares estimator that uses a correct parametric specification. These adaptive estimators have been evaluated in terms of a second-order approximation by Linton (1996) and extended to time series models by Hidalgo (1992), to nonlinear multivariate models by Delgado (1992), and to panel data by Li and Stengos (1994). Szroeter (1994) suggests weighted least squares estimators that have better finite sample efficiency than OLS when the observations can be ordered according to increasing variances but no other information is available.

### 3.5 Other extensions

Rilestone (1991) has compared the relative efficiency of semiparametric and parametric estimators of  $\beta$  under different types of heteroskedasticity, whereas Surekha and Griffiths (1984) compare the relative efficiency of some Bayesian and sampling theory estimators using a similar Monte Carlo setup. Donald (1995) examines heteroskedasticity in sample selection models, and provides access to

that literature. In truncated and censored models heteroskedasticity impacts on the consistency of estimators, not just their efficiency. Heteroskedasticity in the context of seemingly unrelated regressions has been studied by Mandy and Martins-Filho (1993). Further details appear in chapter 5 by Fiebig. A useful reference that brings together much of the statistical literature on heteroskedasticity is Carroll and Ruppert (1988).

## 4 BAYESIAN INFERENCE

With Bayesian inference post-sample information about unknown parameters is summarized via posterior probability density functions (pdfs) on the parameters of interest. Representing parameter uncertainty in this way is (arguably) more natural than the point estimates and standard errors produced by sampling theory, and provides a flexible way of including additional prior information. In the heteroskedastic model that we have been discussing, namely,  $y_i = x'_i\beta + e_i$ , with  $\text{var}(e_i) = \sigma^2 h_i(\alpha)$ , the parameters of interest are  $\beta$ ,  $\alpha$ , and  $\sigma$ , with particular interest usually centering on  $\beta$ . The starting point for Bayesian inference is the specification of prior pdfs for  $\beta$ ,  $\sigma$ , and  $\alpha$ . Since noninformative prior pdfs carry with them the advantage of objective reporting of results, we adopt the conventional ones for  $\beta$  and  $\sigma$  (see Zellner, 1971)

$$f(\beta, \sigma) = f(\beta)f(\sigma) \propto \text{constant} \frac{1}{\sigma}. \quad (4.35)$$

The choice of prior for  $\alpha$  is likely to depend on the function  $h(\cdot)$ . Possible choices are a uniform prior or a prior based on the information matrix. See Zellner (1971, p. 47) for details on the latter. Leaving the precise nature of the prior for  $\alpha$  unspecified, the joint prior pdf for all unknown parameters can be written as

$$f(\beta, \sigma, \alpha) = f(\beta, \sigma)f(\alpha) \propto \frac{f(\alpha)}{\sigma}. \quad (4.36)$$

Assuming normally distributed observations, the likelihood function can be written as

$$f(y|\beta, \sigma, \alpha) \propto \frac{1}{\sigma^N} |\Lambda|^{-1/2} \exp \left\{ -\frac{1}{2\sigma^2} (y - X\beta)' \Lambda^{-1} (y - X\beta) \right\}. \quad (4.37)$$

The joint posterior pdf for  $(\beta, \sigma, \alpha)$  is

$$\begin{aligned} f(\beta, \sigma, \alpha | y) &\propto f(y|\beta, \sigma, \alpha)f(\beta, \sigma, \alpha) \\ &\propto \frac{f(\alpha)}{\sigma^{N+1}} |\Lambda|^{-1/2} \exp \left\{ -\frac{1}{2\sigma^2} (y - X\beta)' \Lambda^{-1} (y - X\beta) \right\}. \end{aligned} \quad (4.38)$$

Once this joint posterior pdf for all parameters has been obtained, the major task is to derive marginal posterior pdfs for each single parameter. The information in a marginal posterior pdf can then be represented in a diagram or summarized via the moments of the pdf. Where possible, marginal posterior pdfs are obtained by integrating out the remaining parameters. Where analytical integration is not possible, numerical methods are used to estimate the marginal posterior pdfs. There are a variety of ways in which one could proceed with respect to equation (4.38). The steps for one way that is likely to work well are:

1. Integrate  $\sigma$  out to obtain the joint posterior pdf  $f(\beta, \alpha | y)$ .
2. Integrate  $\beta$  out of the result in step 1 to obtain the posterior pdf  $f(\alpha | y)$ .
3. Use a Metropolis algorithm to draw observations from the density  $f(\alpha | y)$ .
4. Construct the conditional posterior pdf  $f(\beta | \alpha, y)$  from the joint posterior pdf that was obtained in step 1; note the conditional mean  $E[\beta | \alpha, y]$  and conditional variance  $\text{var}[\beta | \alpha, y]$ .
5. From step 4, note the conditional posterior pdf and corresponding moments for each element, say  $\beta_k$ , in the vector  $\beta$ .
6. Find estimates of the marginal posterior pdf for  $\beta_k$ , ( $k = 1, 2, \dots, K$ ), and its moments, by averaging the conditional quantities given in step 5, over the conditioning values of  $\alpha$  drawn in step 3.

We will consider each of these steps in turn.

## STEP 1

The joint posterior pdf for  $(\beta, \alpha)$  is given by

$$\begin{aligned} f(\beta, \alpha | y) &= \int f(\beta, \alpha, \sigma | y) d\sigma \\ &\propto f(\alpha) |\Lambda|^{-1/2} [N\hat{\sigma}^2(\alpha) + (\beta - \hat{\beta}(\alpha))' X' \Lambda^{-1} X (\beta - \hat{\beta}(\alpha))]^{-N/2} \end{aligned} \quad (4.39)$$

where  $\hat{\sigma}^2(\alpha)$  and  $\hat{\beta}(\alpha)$  are defined in equations (4.23) and (4.24). The pdf in (4.39) is not utilized directly; it provides an intermediate step for obtaining  $f(\alpha | y)$  and  $f(\beta | \alpha, y)$ .

## STEP 2

The marginal posterior pdf for the parameters in the variance function is given by

$$\begin{aligned} f(\alpha | y) &= \int f(\beta, \alpha | y) d\beta \\ &\propto f(\alpha) |\Lambda|^{-1/2} [\hat{\sigma}(\alpha)]^{-(N-K)} |X' \Lambda^{-1} X|^{-1/2} \end{aligned} \quad (4.40)$$

### STEP 3

The pdf in equation (4.40) is not of a recognizable form, even when imaginative choices for the prior  $f(\alpha)$  are made. Thus, it is not possible to perform further analytical integration to isolate marginal posterior pdfs for single elements such as  $\alpha_s$ . Instead, a numerical procedure, the Metropolis algorithm, can be used to indirectly draw observations from the pdf  $f(\alpha | y)$ . Once such draws are obtained, they can be used to form histograms as estimates of the posterior pdfs for single elements in  $\alpha$ . As we shall see, the draws are also useful for obtaining the posterior pdfs for the  $\beta_k$ .

The random walk Metropolis algorithm which we describe below in the context of the heteroskedastic model is one of many algorithms which come under the general heading of Markov Chain Monte Carlo (MCMC). A recent explosion of research in MCMC has made Bayesian inference more practical for models that were previously plagued by intractable integrals. For access to this literature, see Geweke (1999).

The first step towards using a convenient random walk Metropolis algorithm is to define a suitable "candidate generating function." Assuming that the prior  $f(\alpha)$  is relatively noninformative, and not in conflict with the sample information, the maximum likelihood estimate  $\hat{\alpha}$  provides a suitable starting value  $\alpha_{(0)}$  for the algorithm; and the maximum likelihood covariance matrix  $V_{\hat{\alpha}}$  provides the basis for a suitable covariance matrix for the random walk generator function. The steps for drawing the  $(m + 1)$ th observation  $\alpha_{(m+1)}$  are as follows:

1. Draw  $\alpha^* = \alpha_{(m)} + \varepsilon$  where  $\varepsilon \sim N(0, c V_{\hat{\alpha}})$  and  $c$  is scalar set so that  $\alpha^*$  is accepted approximately 50 percent of the time.
2. Compute

$$r = \frac{f(\alpha^* | y)}{f(\alpha_{(m)} | y)}$$

Note that this ratio can be computed without knowledge of the normalizing constant for  $f(\alpha | y)$ .

3. Draw a value  $u$  for a uniform random variable on the interval  $(0, 1)$ .
4. If  $u \leq r$ , set  $\alpha_{(m+1)} = \alpha^*$ .  
If  $u > r$ , set  $\alpha_{(m+1)} = \alpha_{(m)}$ .
5. Return to step 1, with  $m$  set to  $m + 1$ .

By following these steps, one explores the posterior pdf for  $\alpha$ , generating larger numbers of observations in regions of high posterior probability and smaller numbers of observations in regions of low posterior probability. Markov Chain Monte Carlo theory suggests that, after sufficient observations have been drawn, the remaining observations are drawn from the pdf  $f(\alpha | y)$ . Thus, by drawing a large number of values, and discarding early ones, we obtain draws from the required pdf.

**STEP 4**

The conditional posterior pdf  $f(\beta | \alpha, y)$  is obtained from the joint pdf  $f(\beta, \alpha | y)$  by simply treating  $\alpha$  as a constant in equation (4.39). However, for later use we also need to include any part of the normalizing constant that depends on  $\alpha$ . Recognizing that, when viewed only as a function of  $\beta$ , equation (4.39) is in the form of a multivariate student- $t$  pdf (Judge *et al.*, 1988, p. 312), we have

$$f(\beta | \alpha, y) \propto |X' \Lambda^{-1} X|^{1/2} [\hat{\sigma}(\alpha)]^{N-K} [N\hat{\sigma}^2(\alpha) + (\beta - \hat{\beta}(\alpha))' X' \Lambda^{-1} X (\beta - \hat{\beta}(\alpha))]^{-N/2} \quad (4.41)$$

This pdf has

$$\text{mean} = E(\beta | \alpha, y) = \hat{\beta}(\alpha) = (X' \Lambda^{-1} X)^{-1} X' \Lambda^{-1} y \quad (4.42)$$

$$\text{covariance matrix} = \left( \frac{N}{N - K - 2} \right) \hat{\sigma}^2(\alpha) (X' \Lambda^{-1} X)^{-1} \quad (4.43)$$

$$\text{degrees of freedom} = N - K.$$

**STEP 5**

Let  $a^{kk}(\alpha)$  be the  $k$ th diagonal element of  $(X' \Lambda^{-1} X)^{-1}$ , and  $\hat{\beta}_k(\alpha)$  be the  $k$ th element of  $\hat{\beta}(\alpha)$ . The conditional marginal posterior pdf for  $\beta_k$  given  $\alpha$  is the univariate- $t$  pdf

$$f(\beta_k | \alpha, y) = k^* [\hat{\sigma}(\alpha)]^{N-K} [a^{kk}(\alpha)]^{(N-K)/2} [N\hat{\sigma}^2(\alpha)a^{kk}(\alpha) + (\beta_k - \hat{\beta}_k(\alpha))^2]^{-(N-K+1)/2} \quad (4.44)$$

where  $k^*$  is a normalizing constant independent of  $\alpha$ . This pdf has

$$\text{mean} = E(\beta_k | \alpha, y) = \hat{\beta}_k(\alpha) \quad (4.45)$$

$$\text{variance} = \left( \frac{N}{N - K + 2} \right) \hat{\sigma}^2(\alpha) a^{kk}(\alpha) \quad (4.46)$$

$$\text{degrees of freedom} = N - K.$$

Equations (4.42) and (4.45) provide Bayesian quadratic-loss point estimates for  $\beta$  given  $\alpha$ . Note that they are identical to the generalized least squares estimator for known  $\alpha$ . It is the unknown  $\alpha$  case where sampling theory and Bayesian inference results for point estimation of  $\beta$  diverge. The sampling theory point estimate in this case is  $\hat{\beta}(\hat{\alpha})$ . The Bayesian point estimate is the mean of the marginal posterior pdf  $f(\beta | y)$ . It can be viewed as a weighted average of the  $\hat{\beta}(\alpha)$  over all  $\alpha$  with  $f(\alpha | y)$  used as the weighting pdf. The mechanics of this procedure are described in the next step.

## STEP 6

An estimate of the marginal posterior pdf  $f(\beta_k | y)$  is given by

$$\begin{aligned}\hat{f}(\beta_k | y) &= \frac{1}{M} \sum_{m=1}^M f(\beta_k | \alpha_{(m)}, y) \\ &= \frac{k^*}{M} \sum_{m=1}^M ([\hat{\sigma}(\alpha_{(m)})]^{N-K} [a^{kk}(\alpha_{(m)})]^{(N-K)/2} \\ &\quad \times [N\hat{\sigma}^2(\alpha_{(m)})a^{kk}(\alpha_{(m)}) + (\beta_k - \hat{\beta}_k(\alpha_{(m)}))^2]^{-(N-K+1)/2})\end{aligned}\quad (4.47)$$

where  $\alpha_{(1)}, \alpha_{(2)}, \dots, \alpha_{(M)}$  are the draws from  $f(\alpha | y)$  that were obtained in step 3. To graph  $\hat{f}(\beta_k | y)$  a grid of values of  $\beta_k$  is chosen and the average in equation (4.47) is calculated for each value of  $\beta_k$  in the grid. The mean and variance of the marginal posterior pdf  $f(\beta_k | y)$  can be estimated in a similar way. The mean is given by the average of the conditional means

$$\bar{\beta} = \hat{E}(\beta | y) = \frac{1}{M} \sum_{m=1}^M \hat{\beta}_k(\alpha_{(m)}). \quad (4.48)$$

The variance is given by the average of the conditional variances plus the variance of the conditional means. That is,

$$\text{var}(\beta | y) = \left( \frac{N}{N - K + 2} \right) \frac{1}{M} \sum_{m=1}^M \hat{\sigma}^2(\alpha_{(m)})a^{kk}(\alpha_{(m)}) + \frac{1}{M-1} \sum_{m=1}^M (\beta_k(\alpha_{(m)}) - \bar{\beta})^2 \quad (4.49)$$

Presenting information about parameters in terms of posterior pdfs rather than point estimates provides a natural way of representing uncertainty. In the process just described, the marginal posterior pdfs also provide a proper reflection of finite sample uncertainty. Maximum likelihood estimates (or posterior pdfs conditional on  $\hat{\alpha}$ ) ignore the additional uncertainty created by not knowing  $\alpha$ .

There are, of course, other ways of approaching Bayesian inference in heteroskedastic models. The approach will depend on specification of the model and prior pdf, and on the solution to the problem of intractable integrals. Gibbs sampling is another MCMC technique that is often useful; and importance sampling could be used to obtain draws from  $f(\alpha | y)$ . However, the approach we have described is useful for a wide range of problems, with specific cases defined by specification of  $h_i(\alpha)$  and  $f(\alpha)$ . Other studies which utilize Bayesian inference in heteroskedastic error models include Griffiths, Drynan, and Prakash (1979) and Boscardin and Gelman (1996).

## 5 CONCLUDING REMARKS

Recent sampling theory research on heteroskedastic models seems to be concentrated on methods for estimation and hypothesis testing that do not require

specification of a particular parametric form of heteroskedasticity. They are motivated by our inability to be certain about the most appropriate variance specification. However, methodology suggested along these lines is generally asymptotic and may not perform well in finite samples. What is likely to be important, and what seems to have been neglected, is whether the types of inferences we make in practice are very sensitive to the assumed form of heteroskedasticity. If they are not, then efforts to develop alternative methods, that do not require an explicit variance function, may be misplaced.

Bayesian estimation has several advantages. Results are presented in terms of intuitively meaningful posterior pdfs. Marginal posterior pdfs reflect all the parameter uncertainty in a model and do not condition on point estimates of nuisance parameters. Predictive pdfs for future values can also be constructed without conditioning on point estimates (Boscardin and Gelman, 1996). The advent of MCMC techniques means that many more practical applications of Bayesian inference to heteroskedastic models are now possible.

### Note

- \* The author acknowledges valuable comments on an earlier version by three anonymous reviewers.

### References

- Amemiya, T. (1973). Regression analysis when the variance of the dependent variable is proportional to the square of its expectation. *Journal of the American Statistical Association* 68, 928–34.
- Amemiya, T. (1977). A note on a heteroscedastic model. *Journal of Econometrics* 6, 365–70; and Corrigenda. *Journal of Econometrics* 8, 275.
- Amemiya, T. (1983). Partially generalized least squares and two-stage least squares estimators. *Journal of Econometrics* 23, 275–83.
- Baltagi, B.H. (1998). *Econometrics*. New York: Springer-Verlag.
- Boscardin, W.J., and A. Gelman (1996). Bayesian computation for parametric models of heteroscedasticity in the linear model. In R.C. Hill (ed.) *Advances in Econometrics Volume 11A: Bayesian Computational Methods and Applications*. Greenwich: JAI Press.
- Breusch, T.S., and A.R. Pagan (1979). A simple test for heteroscedasticity and random coefficient variation. *Econometrica* 47, 1287–94.
- Brown, B.W., and M.B. Walker (1989). The random utility hypothesis and inference in demand systems. *Econometrica* 57, 815–29.
- Brown, B.W., and M.B. Walker (1995). Stochastic specification in random production models of cost minimizing firms. *Journal of Econometrics* 66, 175–205.
- Carroll, R.J. (1982). Adapting for heteroscedasticity in linear models. *Annals of Statistics* 10, 1224–33.
- Carroll, R.J., and D. Ruppert (1988). *Transformation and Weighting in Regression*. New York: Chapman and Hall.
- Cragg, J.G. (1992). Quasi-Aitken estimation for heteroskedasticity of unknown form. *Journal of Econometrics* 54, 179–202.
- Davidson, R., and J.G. MacKinnon (1993). *Estimation and Inference in Econometrics*. New York: Oxford University Press.

- Delgado, M.A. (1992). Semiparametric generalized least squares in the multivariate non-linear regression model. *Econometric Theory* 8, 203–22.
- Donald, S.G. (1995). Two-step estimation of heteroskedastic sample selection models. *Journal of Econometrics* 65, 347–80.
- Evans, M.A., and M.L. King (1988). A further class of tests for heteroscedasticity. *Journal of Econometrics* 37, 265–76.
- Farebrother, R.W. (1987). The statistical foundations of a class of parametric tests for heteroskedasticity. *Journal of Econometrics* 36, 359–68.
- Geweke, J. (1999). Using simulation methods for Bayesian econometric models: inference, development and communication. *Econometric Reviews* 18, 1–74.
- Godfrey, L.G. (1978). Testing for multiplicative heteroskedasticity. *Journal of Econometrics* 8, 227–36.
- Godfrey, L.G. (1996). Some results on the Glejser and Koenker tests for heteroskedasticity. *Journal of Econometrics* 72, 275–99.
- Godfrey, L.G., and C.D. Orme (1999). The robustness, reliability and power of heteroskedasticity tests. *Econometric Reviews* 18, 169–94.
- Goldfeld, S.M., and R.E. Quandt (1965). Some tests for homoscedasticity. *Journal of the American Statistical Association* 60, 539–47.
- Goldfeld, S.M., and R.E. Quandt (1972). *Nonlinear Methods in Econometrics*. Amsterdam: North-Holland.
- Greene, W. (1997). *Econometric Analysis*, 3rd edn. Upper Saddle River: Prentice Hall.
- Griffiths, W.E. (1972). Estimation of actual response coefficients in the Hildreth-Houck random coefficient model. *Journal of the American Statistical Association* 67, 633–5.
- Griffiths, W.E., and J.R. Anderson (1982). Using time-series and cross-section data to estimate a production function with positive and negative marginal risks. *Journal of the American Statistical Association* 77, 529–36.
- Griffiths, W.E., and G.G. Judge (1992). Testing and estimating location vectors when the error covariance matrix is unknown. *Journal of Econometrics* 54, 121–38.
- Griffiths, W.E., and K. Surekha (1986). A Monte Carlo evaluation of the power of some tests for heteroscedasticity. *Journal of Econometrics* 31, 219–31.
- Griffiths, W.E., R.G. Drynan, and S. Prakash (1979). Bayesian estimation of a random coefficient model. *Journal of Econometrics* 10, 201–20.
- Harvey, A.C. (1976). Estimating regression models with multiplicative heteroscedasticity. *Econometrica* 44, 461–5.
- Hidalgo, J. (1992). Adaptive estimation in time series regression models with heteroskedasticity of unknown form. *Econometric Theory* 8, 161–87.
- Hildreth, C., and J.P. Houck (1968). Some estimators for a linear model with random coefficients. *Journal of the American Statistical Association* 63, 584–95.
- Hill, R.C., J.R. Knight, and C.F. Sirmans (1997). Estimating capital asset price indexes. *Review of Economics and Statistics* 80, 226–33.
- Hooper, P.M. (1993). Iterative weighted least squares estimations in heteroscedastic linear models. *Journal of the American Statistical Association* 88, 179–84.
- Jobson, J.D., and W.A. Fuller (1980). Least squares estimation when the covariance matrix and parameter vector are functionally related. *Journal of the American Statistical Association* 75, 176–81.
- Judge, G.G., W.E. Griffiths, R.C. Hill, and T.-C. Lee (1985). *The Theory and Practice of Econometrics*. New York: John Wiley and Sons.
- Judge, G.G., R.C. Hill, W.E. Griffiths, H. Lütkepohl, and T.-C. Lee (1988). *An Introduction to the Theory and Practice of Econometrics*. New York: John Wiley and Sons.

- Keener, R.W., J. Kmenta, and N.C. Weber (1991). Estimation of the covariance matrix of the least-squares regression coefficients when the disturbance covariance matrix is of unknown form. *Econometric Theory* 7, 22–43.
- Koenker, R., and G. Bassett, Jr. (1982). Robust tests for heteroscedasticity based on regression quantiles. *Econometrica* 50, 43–61.
- Lee, B.-J. (1992). A heteroskedasticity test robust to conditional mean specification. *Econometrica* 60, 159–72.
- Li, Q., and T. Stengos (1994). Adaptive estimation in the panel data error model with heteroskedasticity of unknown form. *International Economic Review* 35, 981–1000.
- Linton, O.B. (1996). Second order approximation in a linear regression with the heteroskedasticity of unknown form. *Econometric Reviews* 15, 1–32.
- Mandy, D.M., and C. Martins-Filho (1993). Seemingly unrelated regressions under additive heteroscedasticity: theory and share equation applications. *Journal of Econometrics* 58, 315–46.
- Orme, C. (1992). Efficient score tests for heteroskedasticity in microeconomics. *Econometric Reviews* 11, 235–52.
- Pagan, A., and Y. Pak (1993). Testing for heteroskedasticity. In G.S. Maddala, C.R. Rao, and H.D. Vinod (eds.) *Handbook of Statistics II: Econometrics*. Amsterdam: North-Holland, 489–518.
- Rilestone, P. (1991). Some Monte Carlo evidence on the relative efficiency of parametric and semiparametric EGLS estimators. *Journal of Business and Economic Statistics* 9, 179–87.
- Robinson, P.M. (1987). Asymptotically efficient estimation in the presence of heteroscedasticity of unknown form. *Econometrica* 55, 875–91.
- Rutemiller, H.C., and D.A. Bowers (1968). Estimation in a heteroscedastic regression model. *Journal of the American Statistical Association* 63, 552–7.
- Surekha, K., and W.E. Griffiths (1984). A Monte Carlo comparison of some Bayesian and sampling theory estimators in two heteroscedastic error models. *Communications in Statistics B* 13, 85–105.
- Szroeter, J. (1978). A class of parametric tests for heteroskedasticity in linear econometric models. *Econometrica* 46, 1311–28.
- Szroeter, J. (1994). Exact finite-sample relative efficiency of sub-optimality weighted least squares estimators in models with ordered heteroscedasticity. *Journal of Econometrics* 64, 29–44.
- Welsh, A.H., R.J. Carroll, and D. Ruppert (1994). Fitting heteroscedastic regression models. *Journal of the American Statistical Association* 89, 100–16.
- White, H. (1980). A heteroscedasticity-consistent covariance matrix estimators and a direct test for heteroscedasticity. *Econometrica* 48, 817–38.
- Zellner, A. (1971). *An Introduction to Bayesian Inference in Econometrics*. New York: John Wiley and Sons.

CHAPTER FIVE

# Seemingly Unrelated Regression

*Denzil G. Fiebig\**

## 1 INTRODUCTION

Seemingly unrelated regression (SUR) is one of the econometric developments that has found considerable use in applied work. Not all theoretical developments find their way immediately into the toolkit of the practitioner, but SUR is one of those which has. The popularity of the model is related to its applicability to a large class of modeling and testing problems and also the relative ease of estimation. Empirical studies that utilize SUR have been supported by a considerable amount of theoretical development; consequently much is known about the basic model and the extensions needed to accommodate more complex problems.

As noted by Griliches and Intriligator (1983, p. xiii) "The historical evolution of econometrics was driven both by the increased availability of data and by the desire of generations of scientists to analyze such data in a rigorous and coherent fashion." SUR falls neatly into this characterization. The increased availability of data representing a sample of cross-sectional units observed over several time periods provides researchers with a potentially rich source of information. At the same time, the nature of the data necessitates careful consideration of how regression parameters vary (if at all) over the cross-sectional and time series dimensions and the appropriate specification of the disturbance covariance matrix. SUR is able to provide estimates of how relationships can potentially vary over the data dimensions as well as providing a convenient vehicle for testing hypotheses about these relationships. The specification of the basic SUR model as first introduced by Zellner (1962) remains an important option in any modeling exercise using pooled data.

Given the voluminous growth in the SUR literature, this chapter is necessarily selective. Material found in the survey paper of Srivastava and Dwivedi (1979)

and the book by Srivastava and Giles (1987) serve to provide a good coverage of the literature until the mid to late 1980s. We concentrate on the more recent developments. While the SUR model has had a significant impact outside economics and business, these studies have largely been ignored to provide further focus for this current chapter.

## 2 BASIC MODEL

Suppose we have a set of  $N$  cross-sections with  $T$  time series observations on each. The classic data introduced in Zellner's (1962) initial work comprised firm-level investment data collected annually for 20 years. More recently Batchelor and Gulley (1995) analysed the determinants of jewelry demand for a sample of six countries over 16 years. In both cases, the disturbances from different regression equations, at a given point in time, were correlated because of common unobservable factors.

It is convenient to continue with the cross-section, time series characterization of the data, but clearly what is distinctive is that the data have two dimensions. In general we are dealing with data fields. Often in demand studies a system of demand equations is specified to explain household level consumption of several commodities. Here the disturbance covariances arise because of potential correlations between household specific unobservables associated with each household's commodity demand. In the area of energy demand Bartels, Fiebig, and Plumb (1996) consider household expenditures on gas, and two types of electricity, while Fiebig, Bartels, and Aigner (1991) and Henley and Peirson (1994) consider electricity demands at different times of the day.

The structure of the multi-dimensional data focuses attention on two important specification issues: (i) what should be the appropriate parameter variation across the two dimensions, and (ii) what should be the appropriate stochastic specification. In the case of the basic SUR model these specification issues are resolved by forming a system of  $N$  equations each containing  $T$  observations:

$$y_i = X_i \beta_i + u_i \quad i = 1, \dots, N \quad (5.1)$$

where  $y_i$  and  $u_i$  are  $T$ -dimensional vectors,  $X_i$  is  $T \times K_i$  and  $\beta_i$  is a  $K_i$ -dimensional vector. Stacking all  $N$  equations yields:

$$\begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_N \end{bmatrix} = \begin{bmatrix} X_1 & 0 & \cdots & 0 \\ 0 & X_2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & X_N \end{bmatrix} \begin{bmatrix} \beta_1 \\ \beta_2 \\ \vdots \\ \beta_N \end{bmatrix} + \begin{bmatrix} u_1 \\ u_2 \\ \vdots \\ u_N \end{bmatrix}$$

which can be written compactly as:

$$y = X\beta + u \quad (5.2)$$

where  $\beta$  is a  $K$ -dimensional vector of unknown parameters that needs to be estimated and  $K = \sum_{i=1}^N K_i$ . For the  $NT \times 1$  vector of stacked disturbances the assumptions are (i)  $E(u) = 0$ , and (ii) the  $NT \times NT$  covariance matrix is comprised of  $N^2$  blocks of the form  $E(u_i u_j') = \sigma_{ij} I_T$  where  $I_T$  is a  $T \times T$  identity matrix. These assumptions mean that the  $T$  disturbances in each of the  $N$  equations have zero mean, equal variance, and are uncorrelated and that covariances between contemporaneous disturbances for a pair of equations are potentially nonzero but equal, while non-contemporaneous covariances are all zero. Thus the full covariance matrix of  $u$  is given by  $\Omega = \Sigma \otimes I_T$  where  $\Sigma = [\sigma_{ij}]$  is the  $N \times N$  contemporaneous covariance matrix and  $\otimes$  denotes the Kronecker product.

Each of the  $N$  equations is individually assumed to satisfy the classical assumptions associated with the linear regression model and can be estimated separately. Of course this ignores the correlation between the disturbances of different equations, which can be exploited by joint estimation. The individual equations are related, even though superficially they may not seem to be; they are only seemingly unrelated. The GLS (generalized least squares) estimator is readily defined as

$$\hat{\beta}(\Sigma) = [X'(\Sigma^{-1} \otimes I_T)X]^{-1} X'(\Sigma^{-1} \otimes I_T)y \quad (5.3)$$

with a covariance matrix given by

$$\text{var}[\hat{\beta}(\Sigma)] = [X'(\Sigma^{-1} \otimes I_T)X]^{-1}. \quad (5.4)$$

It is well known that the GLS estimator reduces to OLS (ordinary least squares) when: (i) there is an absence of contemporaneous correlations ( $\sigma_{ij} = 0$ ,  $i \neq j$ ); or (ii) the same set of explanatory variables are included in each equation ( $X_1 = X_2 = \dots = X_N$ ). A more complete characterization of when OLS is equivalent to GLS is given in Baltagi (1989) and Bartels and Fiebig (1991).

In his original article, Zellner (1962) recognized that the efficiency gains resulting from joint estimation tended to be larger when the explanatory variables in different equations were not highly correlated but the disturbances from these equations were highly correlated. Work by Binkley (1982) and Binkley and Nelson (1988) has led to an important qualification to this conventional wisdom. They show that even when correlation among variables across equations is present, efficiency gains from joint estimation can be considerable when there is multicollinearity within an equation.

Consider the class of feasible GLS (FGLS) estimators that differ only in the choice of the estimator used for the contemporaneous covariance matrix, say  $\hat{\beta}(\hat{\Sigma})$ . The estimator is given by:

$$\hat{\beta}(\hat{\Sigma}) = [X'(\hat{\Sigma}^{-1} \otimes I_T)X]^{-1} X'(\hat{\Sigma}^{-1} \otimes I_T)y, \quad (5.5)$$

and inferences are based on the estimator of the asymptotic covariance matrix of  $\hat{\beta}(\hat{\Sigma})$  given by:

$$\text{a var}[\hat{\beta}(\hat{\Sigma})] = [X'(\hat{\Sigma}^{-1} \otimes I_T)X]^{-1}. \quad (5.6)$$

There are many variants of this particular FGLS estimator. Obviously, OLS belongs to the class with  $\hat{\Sigma} = I_N$ , but Zellner (1962) proposed the first operational estimator that explicitly utilized the SUR structure. He suggested an estimated covariance matrix calculated from OLS residuals obtained from (5.1); namely  $S = (s_{ij})$  where  $s_{ij} = (y_i - X_i b_i)'(y_j - X_j b_j)/\tau$  and  $b_i$  is the OLS estimator of  $\beta_i$ . For consistent estimation division by  $\tau = T$  suffices but other suggestions have also been made; see for example Srivastava and Giles (1987).

$S$  has been referred to as the restricted estimator of  $\Sigma$ , but estimation can also be based on the unrestricted residuals derived from OLS regressions which include all explanatory variables from the SUR system. Considerable theoretical work has been devoted to the comparison of respective finite sample properties of the restricted and unrestricted SUR estimators associated with the different estimators of  $\Sigma$ . All of the results discussed in Srivastava and Giles (1987) were based on the assumption of normally distributed disturbances. More recently, Hasegawa (1995) and Srivastava and Maekawa (1995) have presented comparisons between the restricted and unrestricted estimators allowing for nonnormal errors. None of this work produces a conclusive choice between the two alternative estimators.

While theoretical work continues on both restricted and unrestricted estimators, software designers typically make the choice for practitioners. SAS, SHAZAM, TSP, and LIMDEP all use restricted residuals in the estimation of  $\Sigma$ . Moreover, there is limited scope to opt for alternatives with only LIMDEP and TSP allowing one to input their own choice of estimator for  $\Sigma$ . Where the software packages do vary is in the default choice of  $\tau$  and whether to iterate or not. See Silk (1996) for further discussion of software comparisons between SAS, SHAZAM, and TSP in terms of systems estimation.

### 3 STOCHASTIC SPECIFICATION

Many of the recent developments in estimation of the parameters of the SUR model have been motivated by the need to allow for more general stochastic specifications. These developments are driven in part by the usual diagnostic procedures of practitioners, but also by the strong theoretical arguments that have been presented for behavioral models of consumers and producers. Chavas and Segerson (1987) and Brown and Walker (1989, 1995) argue that behavioral models should include a stochastic component as an integral part of the model. When this is done, however, each equation in the resultant input share system or system of demand equations will typically exhibit heteroskedasticity. The basic SUR specification allows for heteroskedasticity across but not within equations and thus will be inappropriate for these systems.

Examples of where heteroskedastic SUR specifications have been discussed include the share equations system of Chavas and Segerson (1987); the SUR random coefficients model of Fiebig *et al.* (1991); and the groupwise heteroskedasticity model described in Bartels *et al.* (1996). Each of these examples are members of what Bartels and Fiebig (1992) called generalized SUR (GSUR).

Recall that the covariance matrix of the basic SUR model is  $\Omega = \Sigma \otimes I_T$ . GSUR allows for a covariance matrix with the following structure:

$$\Omega \equiv \Omega(\theta, \Sigma) = R(\Sigma \otimes I_T)R', \quad (5.7)$$

where  $R = R(\theta)$  is block-diagonal, non-singular, and the parameters  $\theta$  and  $\Sigma$  are assumed to be separable. With this specification, GLS estimation can be viewed as proceeding in two stages: the first stage involves transforming the GSUR model on an equation-by-equation basis so that the classical SUR structure appears as the transformed model; and in the second stage the usual SUR estimator is applied to the transformed model.

The distinguishing feature of the GSUR class of models is the convenience of estimating  $\theta$  and  $\Sigma$  in separate stages with both stages involving familiar estimation techniques. Bollerslev (1990) also notes the computational convenience of this type of simplification for an SUR model that allows for time varying conditional variances and covariances. In this case estimation by maximum likelihood is proposed.

While computational convenience is important, this attractive feature of GSUR must be weighed against the potentially restrictive covariance structure shown in (5.7). While any symmetric, non-singular matrix  $\Omega$  can be written as  $R(\Sigma \otimes I_T)R'$ , the matrix  $R$  will in general depend on  $\Sigma$  as well as  $\theta$  and/or  $R$  may not be block diagonal. Ultimately, the validity of the assumption regarding the structure of the covariance matrix remains an empirical matter.

Mandy and Martins-Filho (1993) extended the basic SUR specification by assuming a contemporaneous covariance matrix that varies across observations within an equation. If we collect the  $N$  disturbances associated with the  $t$ th time period into the vector  $u_{(t)} = (u_{1t}, \dots, u_{Nt})'$ , they assume

$$E(u_{(t)}u'_{(t)}) = \Omega_t = \begin{bmatrix} \sigma_{11}^t & \sigma_{12}^t & \cdots & \sigma_{1N}^t \\ \sigma_{21}^t & \sigma_{22}^t & \cdots & \sigma_{2N}^t \\ \vdots & \vdots & \ddots & \vdots \\ \sigma_{N1}^t & \sigma_{N2}^t & \cdots & \sigma_{NN}^t \end{bmatrix}. \quad (5.8)$$

Specifically, they consider an SUR model subject to additive heteroskedasticity of the form

$$\sigma_{ij}^t = \alpha_{ij}' z_{ij}^t \quad (5.9)$$

where  $\alpha_{ij}$  is a vector of unknown parameters and  $z_{ij}^t$  a conformable vector of explanatory variables.

The framework of Mandy and Martins-Filho (1993) is less restrictive than GSUR but in general does not have the simplified estimation solutions of GSUR. Instead they develop an FGLS procedure that represents an SUR generalization of Amemiya's (1977) efficient estimator for the parameters of the covariance matrix in a single-equation additive heteroskedastic model. A practical problem with

this class of estimators is the appearance of estimated covariance matrices which are not positive definite. Such problems should disappear with large enough samples but for any particular data set there are no guarantees. With some specifications, such as SUR with groupwise heteroskedasticity, the structure ensures that nonnegative variance estimates are obtained. Alternatively, for their SUR random coefficients model, Fiebig *et al.* (1991) employ an estimation procedure that automatically ensures the same result.

When the observations within an equation have a grouped structure it is reasonable to consider a random effects specification for each equation of the SUR system. Avery (1977) and Baltagi (1980) undertook the initial work on SUR models with error component disturbances. An extension allowing for the error components to be heteroskedastic has recently been proposed by Wan, Griffiths, and Anderson (1992) in their study of rice, maize, and wheat production in 28 regions of China over the period 1980–83. In another application, Kumbhakar and Heshmati (1996) consider a system comprising a cost equation and several cost share equations estimated using 26 annual observations for several Swedish manufacturing industries. For the cost equation, but not the share equations, they also specify disturbances comprising heteroskedastic error components. Estimation proceeds along the same lines as discussed in the context of GSUR. The cost equation is first transformed, and the usual SUR estimation can proceed for the share equations and the transformed cost equation.

As has been apparent from the discussion, a major use of SUR is in the estimation of systems of consumer or factor demands. Often these are specified in share form which brings with it special problems. In particular the shares should be restricted to lie between zero and one and should sum to unity. By far the most common approach to the specification of the stochastic component in such models is to append an additive error term to the deterministic component of the model that is obtained from economic theory. Typically the error term is assumed to be multivariate normal. Even if the deterministic component respects the constraint of lying between zero and one, it is clear that assuming normality for the stochastic component means that the modeled share can potentially violate the constraint. One approach is to choose a more appropriate distribution. Woodland (1979), who chose the Dirichlet distribution, seems to be the only example of this approach. An alternative approach advocated by Fry, Fry, and McLaren (1996) is to append a multivariate normal error after performing a logratio transformation of the observed shares.

A second property of share equations, that of adding up, implies that additive errors sum identically to zero across equations. The induced singularity of the covariance matrix is typically accommodated in estimation by deleting one of the equations. Conditions under which the resultant estimators are invariant to the equation that is dropped have long been known; see Bewley (1986) for a summary of these early contributions. More recently, McLaren (1990) and Dhrymes (1994) have provided alternative treatments, which they argue, provide a more transparent demonstration of the conditions for invariance. Dhrymes (1994) is the more general discussion as it allows for the added complication of auto-correlated errors.

Singularity of the disturbance covariance matrix places strong restrictions on any autocorrelation structure that is specified. If we consider the  $N$ -vector of disturbances associated with the  $t$ th time period then adding-up implies  $\mathbf{1}'\mathbf{u}_{(t)} = 0$  where  $\mathbf{1}$  is a vector of ones. If a first order autoregressive process is assumed, i.e.

$$\mathbf{u}_{(t)} = \mathbf{A}\mathbf{u}_{(t-1)} + \boldsymbol{\varepsilon}_{(t)} \quad (5.10)$$

then adding up requires that  $\mathbf{1}'\boldsymbol{\varepsilon}_{(t)} = 0$  and  $\mathbf{1}'\mathbf{A} = k\mathbf{1}'$  where  $k$  is a scalar constant, implying that the columns of  $\mathbf{A}$  sum to the same constant. If  $\mathbf{A}$  is specified to be diagonal, all equations will need to have the same autocorrelation parameter. At the other extreme,  $\mathbf{A}$  is a full matrix with  $(N - 1)^2$  identifiable parameters. A series of contributions by Moschini and Moro (1994), McLaren (1996) and Holt (1998) have suggested alternative specifications involving  $(N - 1)$  parameters, and hence represent a compromise between the very restrictive one-parameter specification and the computationally demanding full specification.

It is not unusual to observe empirical results where systems of static equations have been estimated to yield results that exhibit serially correlated residuals. Common advice that is often followed is to re-estimate assuming an autoregressive error structure. For example, Judge *et al.* (1985, p. 497) conclude that: "as a general recommendation for a researcher estimating a set of equations, we suggest that possible contemporaneous correlation should always be allowed for and, if the number of observations is sufficient, some kind of autocorrelation process could also be assumed."

In the single equation context there has been a movement away from the simple to specific modeling approach, involving correcting for autocorrelation in the errors of static regression models. For example, Anderson and Blundell (1982, p. 1560) note that: "in the context of a single equation model, it has been argued that an autoregressive error specification whilst being a convenient simplification, when appropriate, may be merely accommodating a dynamic structure in the model which could be better represented by a general unrestricted dynamic formulation." Mizon (1995) is even more forceful as evidenced by his paper's title: "A simple message for autocorrelation correctors: Don't." Such advice has largely gone unheeded in the SUR literature where there has been little work on dynamic SUR models. Support for this contention is provided by Moschini and Moro (1994) who report that they found 24 papers published over the period 1988–92 that estimated singular systems with autocorrelated disturbances and that out of these only three did so as special cases of the general dynamic model of Anderson and Blundell (1982).

Deschamps (1998) is one exception where a general dynamic model has been used. He proposes and illustrates a methodology for estimating long-run demand relationships by maximum likelihood. Unlike previous work such as Anderson and Blundell (1982), Deschamps formulates the full likelihood function that is not conditional on the first observations of the dependent variables.

One other contribution that recognizes the need for more work in the area of dynamic systems is that of Kiviet, Phillips, and Schipp (1995). They employ

asymptotic expansions to compare the biases of alternative estimators of an SUR model comprised of dynamic regression equations.

As we have previously noted, the GLS estimator of a basic SUR model reduces to OLS when the design matrices are the same in each equation. Baltagi (1980), Bartels and Fiebig (1992) and Mandy and Martins-Filho (1993) all mention that this well known result needs modification when dealing with more general stochastic structures. The fact that each equation contains an identical set of explanatory variables is not a sufficient condition for joint GLS to collapse to OLS performed on each equation separately. The two-stage estimation process of GSUR highlights the intuition. The first stage involves transforming the GSUR model on an equation-by-equation basis so that the classical SUR structure appears as the transformed model. Even if the original explanatory variables were the same in each equation, the explanatory variables to be used in the second stage are generally not identical after being transformed. A related question is under what conditions does joint GLS collapse to GLS performed on each equation separately. Bartels and Fiebig (1992) provide necessary and sufficient conditions for the first-stage GLS estimates to be fully efficient. Lee (1995) provides some specific examples where this occurs in the estimation of singular systems with autoregressive errors.

## 4 TESTING

### 4.1 Testing linear restrictions

Under the standard assumptions of the basic SUR model, Atkinson and Wilson (1992) prove that:

$$\text{var}[\hat{\beta}(\tilde{\Sigma})] \geq \text{var}[\hat{\beta}(\Sigma)] \geq E[X'(\tilde{\Sigma}^{-1} \otimes I_T)X]^{-1} \quad (5.11)$$

where  $\tilde{\Sigma}$  is any unbiased estimator of  $\Sigma$  and the inequalities refer to matrix differences. The first inequality indicates that FGLS is inefficient relative to GLS. The second inequality provides an indication of the bias in the conventional estimator of the asymptotic covariance matrix of FGLS. While the result requires unbiased estimation of  $\Sigma$ , which does not hold for most conventional estimators of SUR models, it conveniently highlights an important testing problem in SUR models. Asymptotic Wald (W), Lagrange multiplier (LM), and likelihood ratio (LR) tests are prone to be biased towards overrejection.

Fiebig and Theil (1983) and Freedman and Peters (1984) reported Monte Carlo work and empirical examples where the asymptotic standard errors tended to underestimate the true variability of FGLS. Rosalsky, Finke, and Theil (1984), and Jensen (1995) report similar understatement of asymptotic standard errors for maximum likelihood estimation of nonlinear systems. Work by Laitinen (1978), Meisner (1979), and Bewley (1983) alerted applied researchers to the serious size problems of testing cross-equation restrictions in linear demand systems especially when the number of equations specified was large relative to the available number of observations. Similar problems have been observed in finance in

relation to testing restrictions associated with the CAPM (capital asset pricing model); see for example MacKinlay (1987).

In some special cases exact tests have been provided. For example, Laitinen (1978) derived an exact test based on Hotelling's  $T^2$  statistic for testing homogeneity in demand systems. Bewley (1983), de Jong and Thompson (1990), and Stewart (1997) discuss a somewhat wider class of problems where similar results are available but these are very special cases and the search for solutions that involve test statistics with tractable distributions remains an open area. One contribution to this end is provided by Hashimoto and Ohtani (1990) who derive an exact test for linear restrictions on the regression coefficients in an SUR model. Apart from being confined to an SUR system with the same regressors appearing in each equation, the practical drawbacks of this test are that it is computationally complicated, has low power, and to be feasible requires a large number of observations. This last problem is especially troublesome, as this is exactly the situation where the conventional asymptotic tests are the most unreliable.

One approach designed to produce tests with improved sampling properties is to use bootstrap methods. Following the advice of Freedman and Peters (1984), the studies by Williams (1986) and Eakin, McMillen, and Buono (1990) both used bootstrap methods for their estimation of standard errors. Unconditional acceptance of this advice was questioned by Atkinson and Wilson (1992) who compared the bias in the conventional and bootstrap estimators of coefficient standard errors in SUR models. While their theoretical results were inconclusive, their simulation results cautioned that neither of the estimators uniformly dominated and hence bootstrapping provides little improvement in the estimation of standard errors for the regression coefficients in an SUR model.

Rilstone and Veall (1996) argue that an important qualification needs to be made to this somewhat negative conclusion of Atkinson and Wilson (1992). They demonstrated that bootstrapping could result in an improvement in inferences. Rather than using the bootstrap to provide an estimate of standard errors, Rilstone and Veall (1996) recommend bootstrapping the  $t$ -ratios. The appropriate percentiles of the bootstrap distribution of standardized FGLS estimators are then used to construct the bootstrap percentile- $t$  interval. This is an example of what are referred to as pivotal methods. A statistic is (asymptotically) pivotal if its "limiting" distribution does not depend on unknown quantities. Theoretical work indicates that pivotal bootstrap methods provide a higher order of accuracy compared to the basic non-pivotal methods and, in particular, provide confidence intervals with superior coverage properties; see, for example, work cited in Jeong and Maddala (1993). This is precisely what Rilstone and Veall (1996) found.

Another potential approach to providing better inferences, involves the use of improved estimators for the disturbance covariance matrix. Fiebig and Theil (1983) and Ullah and Racine (1992) use nonparametric density estimation as an alternative source of moment estimators. In the tradition of the method of moments, the covariance matrix of a nonparametric density estimator is advocated as an estimator of the population covariance matrix. The proposed SUR estimators have the same structure as the FGLS estimator of equation (5.5) differing only in the choice of estimator for  $\Sigma$ . Ullah and Racine (1992) prove that their

nonparametric density estimator of  $\Sigma$  can be expressed as the usual estimator  $S$  plus a *positive definite* matrix that depends on the smoothing parameter chosen for the density estimation. It is this structure of the estimator that suggests the approach is potentially useful in large equation systems.

Fiebig and Kim (2000) investigate the combination of both approaches, bootstrapping and the improved estimation of the covariance matrix especially in the context of large systems. They conclude that using the percentile- $t$  method of bootstrapping in conjunction with the kernel-based estimator introduced by Ullah and Racine (1992) provides a very attractive estimator for large SUR models.

These studies that evaluate the effectiveness of the bootstrap typically rely on Monte Carlo simulations to validate the procedure and hence require large amounts of computations. Hill, Cartwright, and Arbaugh (1997) investigate the possibility of using Efron's (1992) jackknife-after-bootstrap as an alternative approach. Unfortunately their results indicate that the jackknife-after-bootstrap substantially overestimates the standard deviation of the bootstrap standard errors.

Yet another approach to the testing problem in SUR models is the use of Bartlett-type corrections. For the basic SUR model, Attfield (1998) draws on his earlier work in Attfield (1995) to derive a Bartlett adjustment to the likelihood ratio test statistic for testing linear restrictions. Since the derivations require the assumption of normality, and the absence of lagged dependent variables, the approach of Rocke (1989) may be a useful alternative. He suggests a computational rather than analytical approach for calculating the Bartlett adjustment using the bootstrap. In conclusion Rocke (1989) indicates that while his approach "achieves some improvement, this behavior is still not satisfactory when the sample size is small relative to the number of equations". But as we have observed, this is exactly the situation where the adjustment is most needed.

Silver and Ali (1989) also consider corrections that are derived computationally for the specific problem of testing Slutsky symmetry in a system of demand equations. On the basis of extensive Monte Carlo simulations they conclude that the covariance structure and the form of the regressors are relatively unimportant in describing the exact distribution of the  $F$ -statistic. Given this they use their simulation results to suggest "average" corrections that are only based on the sample size,  $T$ , and the number of equations,  $N$ . A second approach they consider is approximating the exact distribution of the  $F$ -statistic for symmetry using the Pearson class of distributions.

In many modeling situations it is difficult to confidently specify the form of the error covariance matrix. In the single equation context it has become very popular amongst practitioners to construct tests based on the OLS parameter estimators combined with a heteroskedastic and autocorrelation consistent (HAC) or "robust" covariance matrix estimator. Creel and Farell (1996) have considered extensions of this approach to the SUR model. They argue that in many applications the basic SUR stochastic specification will provide a reasonable approximation to a more general error structure. Proceeding in this manner, the usual estimator will be quasi-FGLS and will require a HAC covariance matrix to deal with the remaining heteroskedasticity and serial correlation.

Dufour and Torres (1998) are also concerned with obtaining robust inferences. They show how to use union-intersection techniques to combine tests or confidence intervals, which are based on different subsamples. Amongst their illustrations is an example showing how to test the null hypothesis that coefficients from different equations are equal without requiring any assumptions on how the equations are related as would be needed in the basic SUR specification.

## 4.2 Diagnostic testing

A cornerstone of the SUR model is the presence of contemporaneous correlation. When  $\Sigma$  is diagonal, joint estimation is not required, which simplifies computations. Shiba and Tsurumi (1988) provide a complete set of LM, W, and LR tests of the null hypothesis that  $\Sigma$  is block diagonal. When there are only two equations their LM test reduces to the popular test proposed by Breusch and Pagan (1980). They also derive a Bayesian test.

Classical tests such as the Breusch and Pagan (1980) LM test rely on large- $T$  asymptotics. Frees (1995) recognizes that such tests may be inappropriate with other data configurations and explores alternative tests concentrating on issues that arise when  $N$  and possibly  $T$  are large.

Work in urban and regional economics is distinguished by consideration of spatial aspects. In SUR models where the observations within equations refer to regions, careful consideration needs to be given to the potential for spatial autocorrelation. Anselin (1990) stresses this point in the context of tests for regional heterogeneity. Neglecting the presence of spatial autocorrelation is shown to distort the size and power of conventional Chow-type tests of parameter constancy across equations. On the basis of his Monte Carlo simulations Anselin (1990) advocates a pre-test approach where the first step involves testing for spatial autocorrelation using the LM test proposed in Anselin (1988).

## 5 OTHER DEVELOPMENTS

### 5.1 Unequal observations and missing data

Extending the standard SUR model to allow for an unequal number of observations in different equations causes some problems for estimation of the disturbance covariance matrix. (Problems associated with sample selection bias are avoided by assuming that data are missing at random.) If there are at least  $T_0 < T$  observations available for all equations then the key issue is how to utilize the "extra" observations in the estimation of the disturbance covariance matrix. Monte Carlo comparisons between alternative estimators led Schmidt (1977) to conclude that estimators utilizing less information did not necessarily perform poorly relative to those using more of the sample information. Baltagi, Garvin, and Kerman (1989) provide an extensive Monte Carlo evaluation of several alternative covariance matrix estimators and the associated FGLS estimators of  $\beta$ , attempting to shed some further light on conclusions made by Schmidt (1977). They conclude that while the use of extra observations may lead to better estimates of  $\Sigma$

and  $\Sigma^{-1}$ , this does not necessarily translate into better estimates of  $\beta$ . Baltagi *et al.* (1989) considered both  $\Sigma$  and  $\Sigma^{-1}$  because of Hwang (1990) who noted that the mathematical form of the alternative estimators of  $\Sigma$  gave a misleading impression of their respective information content. This was clarified by considering the associated estimators of  $\Sigma^{-1}$ . Hwang (1990) also proposes a modification of the Telser (1964) estimator, which performs well when the contemporaneous correlation is high.

When there are unequal observations in equations, concern with the use of the "extra" observations arises because of two types of restrictions: (i) imposing equality of variances across groups defined by complete and incomplete observations; and (ii) the need to maintain a positive definite estimate of the covariance matrix. In the groupwise heteroskedasticity model employed by Bartels *et al.* (1996) the groupings corresponded to the divisions between complete and incomplete observations, and, provided that there are no across group restrictions on the parameters of the variance–covariance matrices, positive definite covariance matrix estimates can be readily obtained by applying a standard SUR estimation to each group of data separately.

Consider a two-equation SUR system of the form

$$y_i = X_i\beta_i + u_i \quad i = 1, 2 \quad (5.12)$$

where  $y_1$  and  $u_1$  are  $T$ -dimensional vectors,  $X_1$  is  $T \times k$ ,  $\beta_i$  are  $k$ -dimensional vectors and

$$y_2 = \begin{bmatrix} y_{21} \\ y_{2e} \end{bmatrix}, X_2 = \begin{bmatrix} X_1 \\ X_e \end{bmatrix}, u_2 = \begin{bmatrix} u_{21} \\ u_{2e} \end{bmatrix},$$

where the  $e$  subscript denotes  $m$  extra observations that are available for the second equation. If  $m = 0$  there will be no gain from joint estimation because the system reduces to a basic SUR model with each equation containing an equal number of observations and common regressors. Conniffe (1985) and Im (1994) demonstrate that this conclusion no longer holds when there are an unequal number of observations, because  $y_{2e}$  and  $X_e$  are available. OLS for the second equation is the best linear unbiased estimator (BLUE) as one would expect but joint estimation delivers a more efficient estimator of  $\beta_1$  than OLS applied to the first equation.

Suppose that  $y_2$  and  $X$  are fully observed but  $y_1$  is not. Instead, realizations of a dummy variable  $D$  are available where  $D = 1$  if  $y_1 > 0$  and otherwise  $D = 0$ . Under an assumption of bivariate normality, a natural approach is to estimate the first equation by probit and the second by OLS. Chesher (1984) showed that joint estimation can deliver more efficient estimates than the "single-equation" probit, but, again as you would expect, there is no gain for the other equation. What if these two situations are combined? Conniffe (1997) examines this case, where the system comprises of a probit and a regression equation, but where there are more observations available for the probit equation. In this case estimates for both equations can be improved upon by joint estimation.

Meng and Rubin (1996) were also concerned with SUR models containing latent variables. They use an extension of the expectation maximization (EM) algorithm called the expectation conditional maximization (ECM) algorithm to discuss estimation and inference in SUR models when latent variables are present or when there are observations missing because of nonresponse. One application of this work is to seemingly unrelated tobit models; see Hwang, Sloan, and Adamache (1987) for an example.

## 5.2 Computational matters

With the continuing increase in computer power it may appear strange to be concerned with computational matters. However, the need to use computationally intensive methods, such as the bootstrap, in conjunction with regular estimation procedures, provides an incentive to look for computational efficiencies of the type discussed by Hirschberg (1992) and Kontoghiorghe and Clarke (1995). Hirschberg (1992) provides a simplified solution to the Liapunov matrix equation proposed by Byron (1982) to estimate a class of SUR models that are often encountered in demand modeling. Kontoghiorghe and Clarke (1995) propose an alternative numerical procedure for generating SUR estimators that avoids directly computing the inverse of  $\Sigma$ .

When some structure is assumed for the disturbance covariance matrix, or when there are particular patterns in the regressors, general results may be simplified to yield computational gains. For example, Kontoghiorghe and Clarke (1995) develop their approach for the case where the regressors in each equation contain the regressors from the previous equations as a proper subset. Also Seaks (1990) reminds practitioners of the computational simplifications available when cross-equation restrictions need to be tested in SUR models which contain the same set of regressors in each equation.

## 5.3 Bayesian methods

While increased computational power has had important implications for many areas of econometrics, the impact has probably been most dramatic in the area of Bayesian econometrics. It has long been accepted that the implementation of Bayesian methods by practitioners has been hindered by the unavailability of flexible prior densities that admit analytical treatment of exact posterior and predictive densities. For the SUR model, the problem is that the joint posterior distribution,  $f(\beta, \Sigma^{-1} | y, X)$ , has complicated marginal posteriors. Approximate inferences can be based on a conditional posterior,  $f(\beta | \Sigma^{-1} = \hat{\Sigma}^{-1}, y, X)$ , but exact inferences using the marginal posterior distribution,  $f(\beta | y, X)$ , are problematic.

Richard and Steel (1988) and Steel (1992) have been somewhat successful in extending the exact analytical results that are available for SUR models. Steel (1992) admits, these extensions fall short of providing models that would be of interest to practitioners, but suggests ways in which their analytical results may be effectively used in conjunction with numerical methods.

A better understanding and availability of computational approaches has meant that there are fewer impediments to the routine use of Bayesian methods amongst practitioners. Percy (1992) demonstrated how Gibbs sampling could be used to approximate the predictive density for a basic SUR model. This and other work on Bayesian approaches to SUR estimation is briefly reviewed in Percy (1996).

More recently, Markov chain Monte Carlo methods have enabled Bayesian analyses of even more complex SUR models. Chib and Greenberg (1995) consider a Bayesian hierarchical SUR model and allow the errors to follow a vector autoregressive or vector moving average process. Their contribution aptly illustrates the power of Markov chain Monte Carlo methods in evaluating marginal posterior distributions, which previously have been intractable.

Joint estimation of a SUR model is typically motivated by the presence of disturbance covariation. Blattberg and George (1991) suggest that joint estimation may also be justified in the absence of such dependence if one feels that there are similarities between the regression parameters. When individual estimation leads to nonsensical parameter estimates, they suggest the use of a Bayesian hierarchical model to shrink estimates across equations toward each other thereby producing less estimator variation. They refer to seemingly unrelated equations, SUE rather than SUR.

With these kinds of developments it is not surprising to see more Bayesian applications of SUR models. Examples include Bauwens, Fiebig, and Steel (1994), Griffiths and Chotikapanich (1997) and Griffiths and Valenzuela (1998).

## 5.4 Improved estimation

In a series of papers Hill, Cartwright, and Arbaugh (1990, 1991, 1992) consider the performance of conventional FGLS compared to various improved estimators when applied to a basic SUR model. The improved estimators include several variants of the Stein-rule family and the hierarchical Bayes estimator of Blattberg and George (1991). The primary example is the estimation of a price-promotion model, which captures the impact of price reductions and promotional activities on sales.

In Hill, Cartwright, and Arbaugh (1996) they extend their previous work on estimator performance by investigating the possibility of estimating the finite sample variability of these alternative estimators using bootstrap standard errors. Conclusions based on their Monte Carlo results obtained for conventional FGLS are found to be somewhat contrary to those of Atkinson and Wilson (1992) that we discussed previously. Hill *et al.* (1996, p. 195) conclude, "the bootstrap standard errors are generally less downward biased than the nominal standard errors" concluding that the former are more reliable than the latter. For the Stein-rule estimators they find that the bootstrap may either overestimate or underestimate the estimator variability depending on whether the specification errors in the restrictions are small or not.

An SUR pre-test estimator can be readily defined based on an initial test of the null hypothesis that  $\Sigma$  is diagonal. Ozcam, Judge, Bera, and Yancey (1993) define such an estimator for a two-equation system using the Lagrange multiplier test

of Breusch and Pagan (1980) and Shiba and Tsurumi (1988) and evaluate its risk properties under squared error loss.

## 5.5 Misspecification

Green, Hassan, and Johnson (1992) and Buse (1994) investigate the impact of model misspecification on SUR estimation. In the case of Green *et al.* (1992) it is the omission of income when estimating demand functions, the motivating example being the use of scanner data where demographic information such as income is not typically available. They conclude that anomalous estimates of own-price elasticities are likely due to this misspecification. Buse (1994) notes that the popular linearization of the almost ideal demand system introduces an errors-in-variables problem, which renders the usual SUR estimator inconsistent.

## 5.6 Robust estimation

In the context of single-equation modeling there has been considerable work devoted to the provision of estimators that are robust to small perturbations in the data. Koenker and Portnoy (1990) and Peracchi (1991) extend this work by proposing robust alternatives to standard FGLS estimators of the basic SUR model. Both papers illustrate how their estimators can guard against the potential sensitivity due to data contamination. Neither paper addresses the equally important issue of drawing inferences from their robust estimators.

## 5.7 Model extensions

A natural extension to the basic SUR model is to consider systems that involve equations which are not standard regression models. In the context of time series modeling Fernandez and Harvey (1990) consider a multivariate structural time series model comprised of unobserved components that are allowed to be contemporaneously correlated. King (1989) and Ozuna and Gomez (1994) develop a seemingly unrelated Poisson regression model, which somehow is given the acronym SUPREME. Both applications were to two-equation systems, with extensions to larger models not developed. King (1989) analyses the number of presidential vetoes per year for the period 1946–84 allowing for different explanations to be relevant for social welfare and defense policy vetoes. Ozuna and Gomez (1994) apply the approach to estimate the parameters of a two-equation system of recreation demand functions representing the number of visits to one of two sites.

## 6 CONCLUSION

The SUR model has been the source of much interest from a theoretical standpoint and has been an extremely useful part of the toolkit of applied econometricians and applied statisticians in general. According to Goldberger (1991, p. 323), the SUR model “plays a central role in contemporary econometrics.” This

is evidenced in our chapter by the breadth of the theoretical and applied work that has appeared since the major surveys of Srivastava and Dwivedi (1979) and Srivastava and Giles (1987). Hopefully this new summary of recent research will provide a useful resource for further developments in the area.

### Note

- \* I gratefully acknowledge the excellent research assistance of Hong Li and Kerri Hoffman. Badi Baltagi, Bob Bartels, Mike Smith, and three anonymous referees also provided helpful comments.

### References

- Amemiya, T. (1977). A note on a heteroscedastic model. *Journal of Econometrics* 6, 365–70.
- Anderson, G.J., and R.W. Blundell (1982). Estimation and hypothesis testing in dynamic singular equation systems. *Econometrica* 50, 1559–72.
- Anselin, L. (1988). A test for spatial autocorrelation in seemingly unrelated regressions. *Economics Letters* 28, 335–41.
- Anselin, L. (1990). Spatial dependence and spatial structural instability in applied regression analysis. *Journal of Regional Science* 30, 185–207.
- Atkinson, S.E., and P.W. Wilson (1992). The bias of bootstrapped versus conventional standard errors in the general linear and SUR models. *Econometric Theory* 8, 258–75.
- Attfield, C.L.F. (1995). Bartlett adjustment to the likelihood ratio test for a system of equations. *Journal of Econometrics* 66, 207–24.
- Attfield, C.L.F. (1998). Bartlett adjustments for systems of linear equations with linear restrictions. *Economics Letters* 60, 277–83.
- Avery, R. (1977). Error components and seemingly unrelated regressions. *Econometrica* 45, 199–209.
- Baltagi, B.H. (1980). On seemingly unrelated regressions with error components. *Econometrica* 48, 1547–51.
- Baltagi, B.H. (1989). Applications of a necessary and sufficient condition for OLS to be BLUE. *Statistics and Probability Letters* 8, 457–61.
- Baltagi, B.H., S. Garvin, and S. Kerman (1989). Further evidence on seemingly unrelated regressions with unequal number of observations. *Annales D'Economie et de Statistique* 14, 103–15.
- Bartels, R., and D.G. Fiebig (1991). A simple characterization of seemingly unrelated regressions models in which OLS is BLUE. *American Statistician* 45, 137–40.
- Bartels, R., and D.G. Fiebig (1992). Efficiency of alternative estimators in generalized seemingly unrelated regression models. In R. Bewley, and T.V. Hao (eds.) *Contributions to Consumer Demand and Econometrics: Essays in Honour of Henri Theil*. London: Macmillan Publishing Company, 125–39.
- Bartels, R., D.G. Fiebig, and M. Plumb (1996). Gas or electricity, which is cheaper? An econometric approach with application to Australian expenditure data. *Energy Journal* 17, 33–58.
- Batchelor, R., and D. Gulley (1995). Jewellery demand and the price of gold. *Resources Policy* 21, 37–42.
- Bauwens, L., D.G. Fiebig, and M.F.J. Steel (1994). Estimating end-use demand: A Bayesian approach. *Journal of Business and Economic Statistics* 12, 221–31.

- Bewley, R.A. (1983). Tests of restrictions in large demand systems. *European Economic Review* 20, 257–69.
- Bewley, R.A. (1986). *Allocation Models: Specification, Estimation and Applications*. Cambridge, MA: Ballinger Publishing Company.
- Binkley, J.K. (1982). The effect of variable correlation on the efficiency of seemingly unrelated regression in a two equation model. *Journal of the American Statistical Association* 77, 890–5.
- Binkley, J.K., and C.H. Nelson (1988). A note on the efficiency of seemingly unrelated regression. *American Statistician* 42, 137–9.
- Blattberg, R.C., and E.I. George (1991). Shrinkage estimation of price and promotional elasticities: Seemingly unrelated equations. *Journal of the American Statistical Association* 86, 304–15.
- Bollerslev, T. (1990). Modeling the coherence in short-run nominal exchange rates: A multivariate generalized ARCH model. *Review of Economics and Statistics* 72, 498–505.
- Breusch, T.S., and A.R. Pagan (1980). The Lagrange multiplier test and its applications to model specification in econometrics. *Review of Economic Studies* 47, 239–53.
- Brown, B.W., and M.B. Walker (1989). The random utility hypothesis and inference in demand systems. *Econometrica* 57, 815–29.
- Brown, B.W., and M.B. Walker (1995). Stochastic specification in random production models of cost minimizing firms. *Journal of Econometrics* 66, 175–205.
- Buse, A. (1994). Evaluating the linearized almost ideal demand system. *American Journal of Agricultural Economics* 76, 781–93.
- Byron, R.P. (1982). A note on the estimation of symmetric systems. *Econometrica* 50, 1573–5.
- Chavas, J.-P., and K. Segerson (1987). Stochastic specification and estimation of share equation systems. *Journal of Econometrics* 35, 337–58.
- Chesher, A. (1984). Improving the efficiency of probit estimators. *Review of Economics and Statistics* 66, 523–7.
- Chib, S., and E. Greenberg (1995). Hierarchical analysis of SUR models with extensions to correlated serial errors and time-varying parameter models. *Journal of Econometrics* 68, 339–60.
- Conniffe, D. (1985). Estimating regression equations with common explanatory variables but unequal numbers of observations. *Journal of Econometrics* 27, 179–96.
- Conniffe, D. (1997). Improving a linear regression through joint estimation with a probit model. *The Statistician* 46, 487–93.
- Creel, M., and M. Farell (1996). SUR estimation of multiple time-series models with heteroscedasticity and serial correlation of unknown form. *Economic Letters* 53, 239–45.
- de Jong, P., and R. Thompson (1990). Testing linear hypothesis in the SUR framework with identical explanatory variables. *Research in Finance* 8, 59–76.
- Deschamps, P.J. (1998). Full maximum likelihood estimation of dynamic demand models. *Journal of Econometrics* 82, 335–59.
- Dhrymes, P.J. (1994). Autoregressive errors in singular systems of equations. *Econometric Theory* 10, 254–85.
- Dufour, J.M., and O. Torres (1998). Union-intersection and sample-split methods in econometrics with applications to MA and SURE models. In A. Ullah, and D.E.A. Giles (eds.) *Handbook of Applied Economic Statistics*. New York: Marcel Dekker, 465–505.
- Eakin, B.K., D.P. McMillen, and M.J. Buono (1990). Constructing confidence intervals using the bootstrap: An application to a multi-product cost function. *Review of Economics and Statistics* 72, 339–44.

- Efron, B. (1992). Jackknife-after-bootstrap standard errors and influence functions. *Journal of the Royal Statistical Society, B* 54, 83–127.
- Fernandez, F.J., and A.C. Harvey (1990). Seemingly unrelated time series equations and a test for homogeneity. *Journal of Business and Economic Statistics* 8, 71–82.
- Fiebig, D.G., and J.H. Kim (2000). Estimation and inference in SUR models when the number of equations is large. *Econometric Reviews* 19, 105–130.
- Fiebig, D.G., and Theil, H. (1983). The two perils of symmetry-constrained estimation of demand systems. *Economics Letters* 13, 105–11.
- Fiebig, D.G., R. Bartels, and D.J. Aigner (1991). A random coefficient approach to the estimation of residential end-use load profiles. *Journal of Econometrics* 50, 297–327.
- Freedman, D.A., and S.C. Peters (1984). Bootstrapping a regression equation: Some empirical results. *Journal of the American Statistical Association* 79, 97–106.
- Frees, E.W. (1995). Assessing cross-sectional correlation in panel data. *Journal of Econometrics* 69, 393–414.
- Fry, J.M., T.R.L. Fry, and K.R. McLaren (1996). The stochastic specification of demand share equations: Restricting budget shares to the unit simplex. *Journal of Econometrics* 73, 377–86.
- Goldberger, A.S. (1991). *A Course in Econometrics*. Cambridge, MA: Harvard University Press.
- Green, R., Z.A. Hassan, and S.R. Johnson (1992). The bias due to omitting income when estimating demand functions. *Canadian Journal of Agricultural Economics* 40, 475–84.
- Griffiths, W.E., and D. Chotikapanich (1997). Bayesian methodology for imposing inequality constraints on a linear expenditure system with demographic factors. *Australian Economic Papers* 36, 321–41.
- Griffiths, W.E., and R. Valenzuela (1998). Missing data from infrequency of purchase: Bayesian estimation of a linear expenditure system. In T.B. Fomby, and R.C. Hill (eds.) *Advances in Econometrics, 13: Messy Data – Missing Observations, Outliers and Mixed-Frequency Data*. Greenwich CT: JAI Press, 47–74.
- Griliches, Z., and M.D. Intriligator (1983). Preface. In Z. Griliches, and M.D. Intriligator (eds.) *Handbook of Econometrics*. Amsterdam: Elsevier Science Publishers B.V., xi–xvii.
- Hasegawa, H. (1995). On small sample properties of Zellner's estimator for the case of two SUR equations with compound normal disturbances. *Communications in Statistics, Simulation and Computation* 24, 45–59.
- Hashimoto, N., and K. Ohtani (1990). An exact test for linear restrictions in seemingly unrelated regressions with the same regressors. *Economics Letters* 32, 243–6.
- Henley, A., and J. Peirson (1994). Time-of-use electricity pricing: Evidence from a British experiment. *Economics Letters* 45, 421–6.
- Hill, R.C., P.A. Cartwright, and J.F. Arbaugh (1990). Using aggregate data to estimate micro-level parameters with shrinkage rules. *American Statistical Association: Proceedings of the Business and Economic Statistics Section* 339–44.
- Hill, R.C., P.A. Cartwright, and J.F. Arbaugh (1991). Using aggregate data to estimate micro-level parameters with shrinkage rules: More results. *American Statistical Association: Proceedings of the Business and Economic Statistics Section* 155–60.
- Hill, R.C., P.A. Cartwright, and J.F. Arbaugh (1992). The finite sample properties of shrinkage estimators applied to seemingly unrelated regressions. *American Statistical Association: Proceedings of the Business and Economic Statistics Section* 17–21.
- Hill, R.C., P.A. Cartwright, and J.F. Arbaugh (1996). Bootstrapping estimators for the seemingly unrelated regressions model. *Journal of Statistical Computation and Simulation* 54, 177–96.
- Hill, R.C., P.A. Cartwright, and J.F. Arbaugh (1997). Jackknifing the bootstrap: Some Monte Carlo evidence. *Communications in Statistics, Simulation and Computation* 26, 125–239.

- Hirschberg, J.G. (1992). A computationally efficient method for bootstrapping systems of demand equations: A comparison to traditional techniques. *Statistics and Computing* 2, 19–24.
- Holt, M.T. (1998). Autocorrelation specification in singular equation systems: A further look. *Economics Letters* 58, 135–41.
- Hwang, H.S. (1990). Estimation of a linear SUR model with unequal numbers of observations. *Review of Economics and Statistics* 72, 510–15.
- Hwang, C.J., F.A. Sloan, and K.W. Adamache (1987). Estimation of seemingly unrelated Tobit regressions via the EM algorithm. *Journal of Business and Economic Statistics* 5, 425–30.
- Im, E.I. (1994). Unequal numbers of observations and partial efficiency gain. *Economics Letters* 46, 291–4.
- Jensen, M.J. (1995). A Monte Carlo study on two methods of calculating the MLE's covariance matrix in a seemingly unrelated nonlinear regression. *Econometric Reviews* 14, 315–30.
- Jeong, J., and G.S. Maddala (1993). A perspective on application of bootstrap methods in econometrics. In G.S. Maddala, C.R. Rao, and H.D. Vinod (eds.) *Handbook of Statistics, Volume 11*. Amsterdam: Elsevier Science Publishers B.V., 573–610.
- Judge G.G., W.E. Griffiths, R.C. Hill, H. Lütkepohl, and T.-C. Lee (1985). *The Theory and Practice of Econometrics*, 2nd edn. New York: John Wiley and Sons.
- King, G. (1989). A seemingly unrelated Poisson regression model. *Sociological Methods and Research* 17, 235–55.
- Kiviet, J.F., G.D.A. Phillips, and B. Schipp (1995). The bias of OLS, GLS and ZEF estimators in dynamic SUR models. *Journal of Econometrics* 69, 241–66.
- Koenker, R., and S. Portnoy (1990). M estimation of multivariate regressions. *Journal of the American Statistical Association* 85, 1060–8.
- Kontoghiorghes, E.J. and M.R.B. Clarke (1995). An alternative approach to the numerical solution of seemingly unrelated regression equation models. *Computational Statistics and Data Analysis* 19, 369–77.
- Kumbhakar, S.C., and A. Heshmati (1996). Technical change and total factor productivity growth in Swedish manufacturing industries. *Econometric Reviews* 15, 275–98.
- Laitinen, K. (1978). Why is demand homogeneity so often rejected? *Economics Letters* 1, 187–91.
- Lee, B.-J. (1995). Seemingly unrelated regression on the autoregressive (AR(p)) singular equation system. *Econometric Reviews* 14, 65–74.
- MacKinley, A.C. (1987). On multivariate tests of the CAPM. *Journal of Financial Economics* 18, 341–71.
- Mandy, D.M., and C. Martins-Filho (1993). Seemingly unrelated regressions under additive heteroscedasticity. *Journal of Econometrics* 58, 315–46.
- McLaren, K.R. (1990). A variant on the arguments for the invariance of estimators in a singular system of equations. *Econometric Reviews* 9, 91–102.
- McLaren, K.R. (1996). Parsimonious autocorrelation corrections for singular demand systems. *Economics Letters* 53, 115–21.
- Meisner, J.F. (1979). The sad fate of the asymptotic Slutsky symmetry test for large systems. *Economics Letters* 2, 231–3.
- Meng, X.L., and D.B. Rubin (1996). Efficient methods for estimation and testing with seemingly unrelated regressions in the presence of latent variables and missing observations. In D.A. Berry, K.M. Chaloner, and J.K. Geweke (eds.) *Bayesian Analysis in Statistics and Econometrics: Essays in Honor of Arnold Zellner*. New York: John Wiley and Sons, 215–27.

- Mizon, G.E. (1995). A simple message for autocorrelation correctors: Don't. *Journal of Econometrics* 69, 267–88.
- Moschini, G., and D. Moro (1994). Autocorrelation specification in singular equation systems. *Economics Letters* 46, 303–9.
- Ozcam, A., G. Judge, A. Bera, and T. Yancey (1993). The risk properties of a pre-test estimator for Zellner's seemingly unrelated regression model. *Journal of Quantitative Economics* 9, 41–52.
- Ozuna, T., and I.A. Gomez (1994). Estimating a system of recreation demand functions using a seemingly unrelated Poisson regression approach. *Review of Economics and Statistics* 76, 356–60.
- Peracchi, F. (1991). Bounded-influence estimators for the SURE model. *Journal of Econometrics* 48, 119–34.
- Percy, D.F. (1992). Prediction for seemingly unrelated regressions. *Journal of the Royal Statistical Society, B* 54, 243–52.
- Percy, D.F. (1996). Zellner's influence on multivariate linear models. In D.A. Berry, K.M. Chaloner, and J.K. Geweke (eds.) *Bayesian Analysis in Statistics and Econometrics: Essays in Honor of Arnold Zellner*. New York: John Wiley and Sons, 203–13.
- Richard, J.F., and M.F.J. Steel (1988). Bayesian analysis of systems of seemingly unrelated regression equations under a recursive extended natural conjugate prior density. *Journal of Econometrics* 38, 7–37.
- Rilstone, P., and M. Veall (1996). Using bootstrapped confidence intervals for improved inferences with seemingly unrelated regression equations. *Econometric Theory* 12, 569–80.
- Rocke, D.M. (1989). Bootstrap Bartlett adjustment in seemingly unrelated regression. *Journal of the American Statistical Association* 84, 598–601.
- Rosalsky, M.C., R. Finke, and H. Theil (1984). The downward bias of asymptotic standard errors of maximum likelihood estimates of non-linear systems. *Economics Letters* 14, 207–11.
- Schmidt, P. (1977). Estimation of seemingly unrelated regressions with unequal numbers of observations. *Journal of Econometrics* 5, 365–77.
- Seaks, T.G. (1990). The computation of test statistics for multivariate regression models in event studies. *Economics Letters* 33, 141–5.
- Shiba, T., and H. Tsurumi (1988). Bayesian and non-Bayesian tests of independence in seemingly unrelated regressions. *International Economic Review* 29, 377–95.
- Silk, J. (1996). Systems estimation: A comparison of SAS, SHAZAM and TSP. *Journal of Applied Econometrics* 11, 437–50.
- Silver, J.L., and M.M. Ali (1989). Testing Slutsky symmetry in systems of linear demand equations. *Journal of Econometrics* 41, 251–66.
- Srivastava, V.K., and T.D. Dwivedi (1979). Estimation of seemingly unrelated regression equations: a brief survey. *Journal of Econometrics* 10, 15–32.
- Srivastava, V.K., and D.E.A. Giles (1987). *Seemingly Unrelated Regression Models: Estimation and Inference*. New York: Marcel Dekker.
- Srivastava, V.K., and K. Maekawa (1995). Efficiency properties of feasible generalized least squares estimators in SURE models under non-normal disturbances. *Journal of Econometrics* 66, 99–121.
- Steel, M.F. (1992). Posterior analysis of restricted seemingly unrelated regression equation models: A recursive analytical approach. *Econometric Reviews* 11, 129–42.
- Stewart, K.G. (1997). Exact testing in multivariate regression. *Econometric Reviews* 16, 321–52.
- Telser, L.G. (1964). Iterative estimation of a set of linear regression equations. *Journal of the American Statistical Association* 59, 845–62.

- Ullah, A., and J. Racine (1992). Smooth improved estimators of econometric parameters. In W.E. Griffiths, H. Lütkepohl, and M.E. Bock (eds.) *Readings in Econometric Theory and Practice*. Amsterdam: Elsevier Science Publishers B.V., 198–213.
- Wan, G.H., W.E. Griffiths, and J.R. Anderson (1992). Using panel data to estimate risk effects in seemingly unrelated production functions. *Empirical Economics* 17, 35–49.
- Williams, M.A. (1986). An economic application of bootstrap statistical methods: Addyston Pipe revisited. *American Economist* 30, 52–8.
- Woodland, A.D. (1979). Stochastic specification and the estimation of share equations. *Journal of Econometrics* 10, 361–83.
- Zellner, A. (1962). An efficient method of estimating seemingly unrelated regressions and tests of aggregation bias. *Journal of the American Statistical Association* 57, 348–68.

CHAPTER SIX

# Simultaneous Equation Model Estimators: Statistical Properties and Practical Implications

*Roberto S. Mariano*

## 1 THE LINEAR SIMULTANEOUS EQUATIONS MODEL

This chapter deals with the statistical properties of estimators in simultaneous equation models. The discussion covers material that extends from the standard large sample asymptotics and the early works on finite-sample analysis to recent work on exact small-sample properties of IV (instrumental variable) estimators and the behavior of IV estimators when instruments are weak. This section introduces the linear simultaneous equations model and the notation for the rest of the chapter. Sections 2 and 3 then cover limited information and full information estimators of structural parameters. Section 4 moves on to large sample properties of the estimators while Sections 5 and 6 summarize results obtained in finite sample analysis. Section 7 ends the chapter with a summary of practical implications of the finite sample results and alternative asymptotics that involve increasing the number of instruments or reducing the correlation between instruments and endogenous regressors in instrumental variable estimation.

Consider the classical linear simultaneous equations model (SEM) of the form:

$$By_t + \Gamma x_t = u_t; \quad t = 1, 2, \dots, T \quad \text{or} \quad YB' + X\Gamma' = U, \quad (6.1)$$

where

- $B = G \times G$  matrix of fixed parameters (some of which are unknown),
- $\Gamma = G \times K$  matrix of fixed parameters (some unknown),
- $y_t = G \times 1$  vector of observations on endogenous variables at "time"  $t$ ,
- $x_t = K \times 1$  vector of observations on exogenous variables at "time"  $t$ ,
- $u_t = G \times 1$  vector of structural disturbances in "time"  $t$ ,
- $Y = T \times G$  matrix whose  $t$ th row is  $y'_t$ ,
- $X = T \times K$  matrix whose  $t$ th row is  $x'_t$ ,
- $U = T \times G$  matrix whose  $t$ th row is  $u'_t$ ,
- $T$  = sample size.

The system described in (6.1) consists of  $G$  linear equations. Each equation is linear in the components of  $y_t$ , with a particular row of  $B$  and  $\Gamma$  containing the coefficients of an equation in the system. Each equation may be stochastic or nonstochastic depending on whether or not the corresponding component of the disturbance vector  $u_t$  has a nondegenerate probability distribution.

The following assumptions together with (6.1) comprise the *classical linear simultaneous equations model*:

- A1.  $B$  is nonsingular. Thus, the model is complete in the sense that we can solve for  $y_t$  in terms of  $x_t$  and  $u_t$ .
- A2.  $X$  is exogenous. That is,  $X$  and  $U$  are independently distributed of each other.
- A3.  $X$  is of full column rank with probability 1.
- A4. The disturbances  $u_t$ , for  $t = 1, 2, \dots, T$ , are uncorrelated and identically distributed with mean zero and positive definite covariance matrix  $\Sigma$ .

At certain times, we will replace A4 with the stronger assumption,

- A4'. The disturbances  $u_t$  are independent and identically distributed as multivariate normal with mean zero and covariance matrix  $\Sigma$ .

We refer to the  $G$  equations in (6.1) as being structural in that they comprise a simultaneous system which explains the mutual interdependence among  $G$  endogenous variables and their relationship to  $K$  exogenous variables whose behavior, by assumption A2, is in turn explained by factors outside the system.

This model differs from the standard multivariate linear regression model in the statistics literature to the extent that  $B$  is generally not diagonal and  $Y$  and  $U$  are correlated. Under these conditions, the structural equations are such that, after normalization, some explanatory or right-hand side variables are correlated with the disturbance terms. (One can thus claim that the canonical equivalence is between this structural system and a multivariate linear regression model with measurement errors.)

The structural system in (6.1) also leads to a multivariate linear regression system with coefficient restrictions. Premultiplying (6.1) by  $B^{-1}$ , we get the so-called reduced form equations of the model:

$$y_t = -B^{-1}\Gamma x_t + B^{-1}u_t = \Pi x_t + v_t \quad \text{or} \quad Y = -X\Gamma' B'^{-1} + UB'^{-1} = X\Pi' + V \quad (6.2)$$

where  $\Pi = -B^{-1}\Gamma$ ,  $v_t = B^{-1}u_t$ , and  $V = UB'^{-1}$ .

It follows from A2 and (6.2) that  $X$  and  $V$  are independently distributed of each other and that, for  $\Omega = B^{-1}\Sigma B'^{-1}$ ,  $v_t \sim \text{uncorrelated}(0, \Omega)$ , if A4 holds; and  $v_t \sim \text{iid } N(0, \Omega)$  if A4' holds. Thus,  $y_t | X \sim \text{uncorrelated}(\Pi x_t, \Omega)$  under A4 and  $y_t | X \sim \text{independent } N(\Pi x_t, \Omega)$  under A4'.

The standard literature on identification of simultaneous equations models shows that identification of (6.1) through prior restrictions on the structural parameters ( $B$ ,  $\Gamma$ ,  $\Sigma$ ) generally will imply restrictions on the matrix  $\Pi$  of reduced form coefficients. For more details on identification of simultaneous equations models, see Hsiao (1983), Bekker and Dijkstra (1990), and Bekker and Wansbeek (chapter 7) in this volume.

## 2 LIMITED- INFORMATION ESTIMATORS OF STRUCTURAL PARAMETERS

We now consider the estimation of an equation in the linear simultaneous system. Note that the  $i$ th rows of  $B$  and  $\Gamma$  contain the coefficients in the  $i$ th structural equation of the system. For the moment, let us consider the first equation of the system and, after imposing zero restrictions and a normalization rule, write it as

$$y_{t1} = -\sum_{i=2}^{G_1} \beta_{1i} y_{ti} - \sum_{j=1}^{K_1} \gamma_{1j} x_{tj} + u_{t1} \quad \text{or} \quad y_1 = Y_1\beta + X_1\gamma + u_1 = Z\delta + u_1, \quad (6.3)$$

where  $y_{ti} = (t, i)$  element of the observation matrix  $Y$ ,  $x_{tj} = (t, j)$  element of  $X$ ,  $u_{ti} = (t, i)$  element of  $U$ ,  $\beta_{ij} = (i, j)$  element of  $B$ ,  $\gamma_{ij} = (i, j)$  element of  $\Gamma$ ,  $u_1 = 1$ st column of  $U$ ,  $X = (X_1, X_2)$ ,  $Y = (y_1, Y_1, Y_2)$ ,  $Z = (Y_1, X_1)$ ,  $\delta' = (\beta', \gamma')$ ,  $-\beta' = (\beta_{12}, \beta_{13}, \dots, \beta_{1G_1})$ , and  $-\gamma' = (\gamma_{11}, \gamma_{12}, \dots, \gamma_{1K_1})$ .

Thus, the normalization rule is  $\beta_{11} = 1$  and the prior restrictions impose the exclusion of the last  $G - G_1$  endogenous variables and the last  $K - K_1$  exogenous variables from the first equation.

From assumption A4 it follows that  $u_1 \sim (0, \sigma_{11}I)$  where  $\sigma_{11}$  is the  $(1, 1)$  element of  $\Sigma$ . By assumption A2,  $X_1$  and  $u_1$  are independent of each other. However, in general  $Y_1$  and  $u_1$  will be correlated.

We shall consider two general categories of estimators of (6.3):

1. limited information or single equation methods, and
2. full information or system methods.

Together with the specification (6.3) for the first equation, all that is needed for the limited information estimators are the reduced form equations for the "included endogenous variables," namely;  $(y_1, Y_1)$ . Under the classical assumptions

we have listed for the linear SEM (including A4'), this requirement reduces to a specification of the exogenous variables in the system since reduced form equations are linear in the exogenous variables with additive disturbances which are identically distributed as mutually independent multivariate Gaussian across sample observations. Full information methods, on the other hand, require the exact specification of the other structural equations composing the system.

In (6.3), the *ordinary least squares* (OLS) estimator of  $\delta$  is obtained directly from the linear regression of  $y_1$  on  $Z$ :

$$\hat{\delta}_{\text{OLS}} = (Z'Z)^{-1}Z'y_1 = \delta + (Z'Z)^{-1}Z'u_1. \quad (6.4)$$

Because of the nonzero covariance between  $Y_1$  and  $u_1$ , this estimator is inconsistent, in general:

$$\begin{aligned} \text{plim } \hat{\delta}_{\text{OLS}} - \delta &= \text{plim } [(Z'Z/T)^{-1}(Z'u_1/T)] \\ &= [\text{plim } (Z'Z/T)^{-1}][\text{plim } (Z'u_1/T)] \\ &= [\text{plim } (Z'Z/T)^{-1}] \begin{pmatrix} \Omega_{21} - \Omega_{22}\beta \\ 0 \end{pmatrix} \end{aligned}$$

where  $\Omega_{22}$  is the covariance matrix of  $t$ th row of  $Y_1$  and  $\Omega_{21}$  is the covariance vector between the  $t$ th rows of  $Y_1$  and  $y_1$ .

A generic class of estimators of  $\delta$  in (6.4) can be obtained by using an instrument matrix, say  $W$ , of the same size as  $Z$  to form the so-called *limited information instrumental variable* (IV) estimator satisfying the modified normal equations

$$W'y_1 = W'Z\hat{\delta}_{\text{IV}} \quad \text{or} \quad \hat{\delta}_{\text{IV}} = (W'Z)^{-1}W'y_1. \quad (6.5)$$

The minimal requirements on the instrument matrix  $W$ , which can be either stochastic or not, are

$$|W'Z| \neq 0 \text{ for any } T \text{ and } \text{plim } (W'Z/T) \text{ is nonsingular.} \quad (6.6)$$

The ordinary least squares estimator belongs to this class – with the instrument matrix, say  $W_{\text{OLS}}$ , equal to  $Z$  itself.

With (6.6) satisfied, a necessary and sufficient condition for  $\hat{\delta}_{\text{IV}}$  to be consistent is

$$\text{plim } (W'u_1/T) = 0. \quad (6.7)$$

Condition (6.7) is not satisfied in the case of OLS and hence OLS would be inconsistent in general. Note that a nonstochastic matrix  $W$  satisfying (6.6) will satisfy (6.7) as well.

The limited information estimators which we discuss presently and which have been developed as alternatives to OLS, can be interpreted as members of this class of IV estimators corresponding to stochastic matrices. First of all, since

$X_1$  is exogenous, we have  $\text{plim } X'_1 u_1 / T = 0$  and so, for consistency, we can take an instrument matrix of the form

$$W = (W_y, X_1), \quad (6.8)$$

where  $W_y$  is the instrument matrix for  $Y_1$  chosen so that

$$\text{plim } W'_y u_1 / T = 0. \quad (6.9)$$

One possibility is to use  $E(Y_1)$  for  $W_y$ , since  $E(Y_1) = X\Pi'_1$  and hence  $\text{plim } (\Pi_1 X' u_1 / T) = \Pi_1 \lim (X' u_1 / T) = 0$ . However, there is still one problem:  $\Pi_1$  is not known and  $E(Y_1)$  is not observable. To remedy this, we can use an estimate of  $\Pi_1$ , say  $\hat{\Pi}_1$  and as long as  $\text{plim } \hat{\Pi}_1$  is finite, we get  $\text{plim } (\hat{\Pi}_1 X' u_1 / T) = (\text{plim } \hat{\Pi}_1) [ \text{plim } (X' u_1 / T) ] = 0$ , and consequently, the instrument matrix  $(\hat{X}\hat{\Pi}'_1, X_1)$  provides a subclass of consistent instrumental variable estimators of  $\delta$ . Note in this discussion that  $\hat{\Pi}_1$  need not be consistent for  $\Pi_1$ ; all that is needed is that  $\hat{\Pi}_1$  have a finite probability limit and produces instruments that satisfy (6.6).

One member of this subclass is the *two-stage least squares* (2SLS) estimator where  $\Pi_1$  is estimated from the unrestricted least squares regression of  $Y_1$  on  $X$ ; that is,  $\hat{\Pi}'_1 = (X'X)^{-1}X'Y_1$  and the instrument matrix for  $Z$  is the regression of  $Z$  on  $X$ :

$$W_{\text{2SLS}} = (X(X'X)^{-1}X'Y_1, X_1) = P_X Z. \quad (6.10)$$

The terminology for this estimator derives from the fact that it can be interpreted as least squares applied twice. The procedure, which first regresses  $Y_1$  on  $X$  to get  $P_x Y_1$  and then regresses  $y_1$  on  $P_x Y_1$  and  $X_1$ , produces in the second-step regression exactly the 2SLS estimator which we have defined above.

The 2SLS estimator also can be interpreted as a *generalized least squares* (GLS) estimator in the linear model obtained by premultiplying (6.3) by  $X'$ .

Other alternative preliminary estimates of  $\Pi_1$  also have been proposed in the literature. One alternative is to estimate each structural equation by ordinary least squares, thus obtaining  $\hat{B}_{\text{OLS}}$  and  $\hat{\Gamma}_{\text{OLS}}$  and then using the appropriate submatrix of the derived reduced form coefficient matrix  $\hat{\Pi} = -(\hat{B}_{\text{OLS}})^{-1}\hat{\Gamma}_{\text{OLS}}$ , say,  $\hat{\Pi}^{(0)}$ , to construct the instrument matrix

$$(X\hat{\Pi}_1^{(0)\prime}, X_1). \quad (6.11)$$

Although  $\hat{\Pi}_1^{(0)}$  itself is inconsistent, the IV estimate of  $\delta$  based on (6.11) as instrument for  $Z$  will be consistent since (6.9) is satisfied.

The *limited information instrumental variable efficient* (LIVE) estimator is based on the instrument matrix

$$W_{\text{LIVE}} = (X\hat{\Pi}_1^{(1)\prime}, X_1),$$

where  $\hat{\Pi}_1^{(1)}$  is a consistent estimator of  $\Pi_1$  derived from some initial consistent estimates of  $B$  and  $\Gamma$ . This would be a two-step procedure if the initial consistent

estimates of  $B$  and  $\Gamma$  are themselves obtained by an instrumental variable procedure. In fact, one way of getting these initial consistent estimates of  $B$  and  $\Gamma$  is by using  $\hat{\Pi}^{(0)}$ , as in (6.11), to generate instruments for the first-step consistent estimation of  $B$  and  $\Gamma$ . Further iteration of this sequential procedure leads to the so-called *iterated instrumental variable* (IIV) procedure.

Yet another alternative to 2SLS is the so-called modified two-stage least squares (M2SLS). Like 2SLS, this is a two-step regression procedure; but here, in the first stage, we would regress  $Y_1$  on  $H$  instead of  $X$ , where  $H$  is a  $T \times h$  data matrix of full column rank and rank  $(\bar{P}_{X_1}H) \geq G - 1$  for  $\bar{P}_{X_1} = I - P_{X_1}$ . We can further show that this estimator will be exactly equivalent to the instrumental variable method with  $(P_H Y_1, X_1)$  as the instrument matrix if the column space of  $H$  contains the column space of  $X_1$  (see Mariano, 1977). Because of this, a suggested manner of constructing  $H$  is to start with  $X_1$  and then add at least  $(G_1 - 1)$  more of the remaining  $K_2$  exogenous variables or, alternatively, the first  $G_1 - 1$  principal components of  $\bar{P}_{X_1}X_2$ .

Another instrumental variable estimator is Theil's  $k$ -class estimator. In this estimator, the instrument matrix is a linear combination of the instrument matrices for OLS and 2SLS. Thus, for  $W_{(k)} = kW_{2SLS} + (1 - k)W_{OLS} = kP_xZ + (1 - k)Z$ , the  $k$ -class estimator of  $\delta$  is

$$\hat{\delta}_{(k)} = (W'_{(k)}Z)^{-1}W'_{(k)}y_1 = \delta + (W'_{(k)}Z)^{-1}W'_{(k)}u_1. \quad (6.12)$$

For consistency, we see from (6.12) that

$$\text{plim } W'_{(k)}u_1/T = [\text{plim } (1 - k)] [\text{plim } Z'u_1/T], \quad (6.13)$$

assuming that  $\text{plim } k$  is finite. Thus, for (6.13) to be equal to zero (and consequently, the consistency of  $k$ -class), a necessary and sufficient condition is  $\text{plim } (1 - k) = 0$ .

The *limited information maximum likelihood* (LIML) estimator, though based, as the term connotes, on the principle of maximizing a certain likelihood function, can also be given an instrumental variable interpretation. In fact as we shall show presently, it is a member of the  $k$ -class of estimators in (6.12). Essentially, the LIML estimator of  $\beta$  and  $\gamma$  maximizes the likelihood of the included endogenous variables subject to the identifiability restrictions imposed on the equation being estimated. This is limited information (rather than full information) maximum likelihood in the sense that the likelihood function considered pertains only to those endogenous variables appearing in the estimated equation; endogenous variables excluded from this equation are thus disregarded. Also, identifiability restrictions on other equations in the system are not taken into account in the constrained maximization of the appropriate likelihood function.

For an explicit development of the LIML estimator, we start with the non-normalized version of the equation to be estimated;  $Y_1^*\beta^* = (y_1, Y_1)\beta^* = X_1\gamma + u_1$ . Thus, for the moment, we take the first row of  $B$  as  $(\beta^{*'}, 0')$ . The reduced form equations for  $Y_1^*$  are  $Y_1^* = X\Pi_1^{*'} + V_1^* = X_1\Pi_{11}^* + X_2\Pi_{12}^{*'} + V_1^*$ . From the relationship

$B\Pi = -\Gamma$ , we get  $\Pi_{11}^{*\prime}\beta^* = \gamma$  and  $\Pi_{12}^{*\prime}\beta^* = 0$ . For identifiability of the first equation, a necessary and sufficient condition is rank  $\Pi_{12}^* = G_1 - 1$ .

The LIML estimator thus maximizes the likelihood function for  $Y_1^*$  subject to the restriction that  $\Pi_{12}^{*\prime}\beta^* = 0$ . This constrained maximization process reduces to the minimization of

$$v = (\beta^{*\prime} A \beta^*) / (\beta^{*\prime} S \beta^*) = 1 + (\beta^{*\prime} W \beta^*) / (\beta^{*\prime} S \beta^*) \quad (6.14)$$

with respect to  $\beta^*$ , for

$$S = Y_1^{*\prime} \bar{P}_X Y_1^*, \quad W = Y_1^{*\prime} (P_X - P_{X_1}) Y_1^*, \quad A = S + W = Y_1^{*\prime} \bar{P}_{X_1} Y_1^*. \quad (6.15)$$

Solving this minimization problem we get  $\hat{\beta}_{\text{LIML}}^*$  = a characteristic vector of A (with respect to S) corresponding to h, where h is the smallest root of  $|A - vS| = 0$  and is equal to  $(\hat{\beta}_{\text{LIML}}^{*\prime} A \hat{\beta}_{\text{LIML}}^*) / (\hat{\beta}_{\text{LIML}}^{*\prime} S \hat{\beta}_{\text{LIML}}^*)$ .

The above derivation of LIML also provides a least variance ratio interpretation for it. In (6.14),  $\beta^{*\prime} W \beta^*$  is the marginal contribution (regression sum of squares) of  $X_2$ , given  $X_1$ , in the regression of  $Y_1^* \beta$  on  $X$ , while  $E\{\beta^{*\prime} S \beta^* / (T - K)\} = \sigma_{11}$ .

Thus, LIML minimizes the explained sum of squares of  $Y_1^* \beta^*$  due to  $X_2$  given  $X_1$ , relative to a stochastic proxy for  $\sigma_{11}$ . On the other hand,  $\hat{\beta}_{\text{2SLS}}^*$  simply minimizes  $\beta^{*\prime} W \beta^*$  in absolute terms.

For the estimator  $\hat{\beta}_{\text{LIML}}^*$  as described above to be uniquely determined, a normalization rule needs to be imposed. We shall use the normalization that the first element of  $\hat{\beta}_{\text{LIML}}^*$  is equal to unity as in the case of all the limited information estimators which we have discussed so far. In this case, it can be easily shown that the LIML estimator of  $\beta$  and  $\gamma$  in the normalized equation (6.3) is a  $k$ -class estimator. The value of  $k$  which gives the LIML estimator is h, where h is the smallest characteristic root of A with respect to S. Note that because  $A = S + W$ , we also have  $h = 1 + \ell$  where  $\ell$  is the smallest characteristic root of W (wrt S). Thus we can interpret LIML as a linear combination of OLS and 2SLS, with  $k > 1$ . Indeed, it can be shown formally that the 2SLS estimate of  $\beta$  lies between the OLS and LIML estimates.

Also note that  $\hat{\beta}_{\text{LIML}}^*$  and  $\hat{\gamma}_{\text{LIML}}^*$  can be characterized equivalently as the maximum likelihood estimates of  $\beta$  and  $\gamma$  based on the "limited information" model

$$y_1 = Y_1 \beta + X_1 \gamma + u_1$$

$$Y_1 = X \Pi_1' + V_1.$$

If the equation being estimated is exactly identified by zero restrictions, then the *indirect least squares* (ILS) estimator of  $\beta$  and  $\gamma$  is well-defined. This is obtained directly from the unrestricted least squares estimate  $\tilde{\Pi}_1^*$  and is the solution to the system of equations taken from  $\tilde{\Pi}_{12}^{*\prime} \beta^* = 0$  and  $\tilde{\Pi}_{11}^{*\prime} \beta^* = \gamma$  after setting  $\beta^{*\prime} = (1, \beta')$ . We can further verify that if the equation being estimated is exactly identified, then the following estimators are exactly equivalent: 2SLS, LIML, ILS, and IV using  $(X_1, X_2)$  as the instrument matrix for the regressors  $(X_1, Y_1)$ .

### 3 FULL INFORMATION METHODS

In this section, we change the notation somewhat and write the  $j$ th equation as

$$y_j = Y_j \beta_j + X_j \gamma_j + u_j = Z_j \delta_j + u_j. \quad (6.16)$$

Here,  $y_j$  is  $T \times 1$ ,  $Y_j$  is  $T \times (G_j - 1)$ ,  $X_j$  is  $T \times K_j$ , and  $Z_j$  is  $T \times (G_j + K_j - 1)$ .

We can further write the whole linear simultaneous system in "stacked" form as

$$y = Z\delta + u, \quad (6.17)$$

where, now,  $y$ ,  $Z$ ,  $\delta$ , and  $u$  are defined differently from previous sections. If we start with  $Y$  and  $U$  as defined in Section 1, then  $y' = (\text{vec } Y)' = (y'_1, y'_2, \dots, y'_G)$ ,  $y_j = j$ th column of  $Y$ ,  $u' = (\text{vec } U)' = (u'_1, u'_2, \dots, u'_G)$ ,  $u_j = j$ th column of  $U$ ,  $\delta' = (\delta'_1, \delta'_2, \dots, \delta'_G)$ ,  $\delta'_j = (\beta'_j, \gamma'_j)$ ,  $Z_j = (Y_j, X_j)$ , and  $Z = \text{diagonal}(Z_j)$ .

Premultiplying both sides of (6.16) by  $X'$ , we get

$$X'y_j = X'Z_j \delta_j + X'u_j; \quad j = 1, 2, \dots, G \quad (6.18)$$

or, for the whole system,

$$(I \otimes X')y = (I \otimes X')Z\delta + (I \otimes X')u. \quad (6.19)$$

Recall that the application of generalized least squares to equation (6.18) separately for each  $j$ , with  $X'u_j \sim (0, \sigma_{jj}^2 X'X)$  produces the 2SLS estimator of  $\delta_j$ . On the other hand, feasible generalized least squares applied to the whole system in (6.19), with  $(I \otimes X')u \sim (0, \Sigma \otimes X'X)$ , leads to the *three-stage least squares* (3SLS) estimator of  $\delta$ :

$$\hat{\delta}_{3SLS} = (Z'(\hat{\Sigma}^{-1} \otimes P_X)Z)^{-1}Z'(\hat{\Sigma}^{-1} \otimes P_X)y, \quad (6.20)$$

where  $\hat{\Sigma}$  is estimated from calculated 2SLS structural residuals.

In contrast, note that the 2SLS estimator of  $\delta$ , obtained by applying 2SLS separately to each equation, is  $\hat{\delta}_{2SLS} = (Z'(I \otimes P_X)Z)^{-1}Z'(I \otimes P_X)y$ . Thus 2SLS is a special case of 3SLS where  $I$  is taken to be the estimate of  $\Sigma$ . Also, 2SLS and 3SLS would be exactly equivalent if  $\Sigma$  is diagonal.

Being a generalized least squares estimator, 3SLS also can be interpreted as an IV estimator of  $\delta$  in (6.17), where the instrument matrix for  $Z$  is

$$W_{3SLS} = (\hat{\Sigma}^{-1} \otimes P_X)Z. \quad (6.21)$$

The  $(i, j)$ th block of  $W_{3SLS}$  can be written further as

$$\hat{\sigma}^{ij}(P_X Y_j, X_j) = \hat{\sigma}^{ij}(X \tilde{\Pi}'_j, X_j), \quad (6.22)$$

where  $\tilde{\Pi}_j$  is the unrestricted least squares estimate of  $\Pi_j$ .

As in the limited information case, instrumental variable procedures have been developed as alternatives to 3SLS. One method is the *full information instrumental*

*variables efficient* estimator (FIVE). In this procedure, instead of (6.21), the instrument matrix for  $Z$  would be  $s^{ij}(X\hat{\Pi}'_j, X_j)$ . As in the case of LIVE,  $\hat{\Pi}_j$  is the appropriate sub-matrix of the restricted reduced form estimate  $\hat{\Pi} = -\hat{B}^{-1}\hat{\Gamma}$  based on some preliminary consistent estimates  $\hat{B}$  and  $\hat{\Gamma}$ , and  $s^{ij}$  is the  $(i, j)$ th element of the inverse of  $\hat{S} = \hat{U}'\hat{U}/T = (Y\hat{B}' + \hat{X}\Gamma')'(Y\hat{B}' + \hat{X}\Gamma')/T$ .

Note that FIVE utilizes all the restrictions in the system to construct the instrument matrix. 3SLS, on the other hand, does not, in the sense that no restrictions on  $\Pi$  are imposed in the calculation of  $X\tilde{\Pi}'_j$  in (6.22).

Another system method is the full information maximum likelihood (FIML) estimator. This is obtained by maximizing the likelihood of  $y$  defined in (6.17) subject to all prior restrictions in the model.

Solving  $\partial \log L / \partial \Sigma^{-1} = 0$  yields the FIML estimate of  $\Sigma : \hat{\Sigma} = (Y\hat{B}' + \hat{X}\Gamma')'(Y\hat{B}' + \hat{X}\Gamma')/T$ . Thus, the concentrated loglikelihood with respect to  $\Sigma$  is  $\log L_C = C + T \log \|B\| - T/2 \log |(YB' + X\Gamma')'(YB' + X\Gamma')|$ , where  $C = -GT/2(1 + \log 2\pi)$ . The FIML estimate of  $B$  and  $\Gamma$  are then solutions to  $0 = \partial \log L_C / \partial \delta = f(\hat{\delta}_{\text{FIML}})$ , say.

A Taylor series expansion of  $f(\hat{\delta}_{\text{FIML}})$  around some consistent estimator  $\tilde{\delta}$  to get

$$0 = f(\hat{\delta}_{\text{FIML}}) \approx f(\tilde{\delta}) + (H(\tilde{\delta}))(\hat{\delta} - \tilde{\delta}) \quad (6.23)$$

gives the *linearized FIML* estimator:

$$\hat{\delta}_{\text{LFIML}} = \tilde{\delta} - (H(\tilde{\delta}))^{-1}f(\tilde{\delta}), \quad (6.24)$$

where  $H$  is the Hessian matrix (of second order partial derivatives) of  $\log L_C$  with respect to  $\delta$ .

This linearized FIML estimator provides one way of numerically approximating  $\hat{\delta}_{\text{FIML}}$ . Such a procedure, as we shall see in the next section, will be asymptotically equivalent to FIML under certain regularity conditions. If carried through to convergence, linearized FIML, if it converges, will coincide exactly with FIML. Furthermore, at convergence,  $(H(\tilde{\delta}))^{-1}$  provides an estimate of the asymptotic covariance matrix of  $\hat{\delta}_{\text{FIML}}$ .

Other iterative procedures also have been devised for the numerical calculation of the FIML estimator. One such algorithm proceeds from a given estimate of  $\delta$ , say  $\hat{\delta}_{(p)}$  at the  $p$ th iteration:

$$\hat{\sigma}_{ij(p)} = (y_i - Z_i \hat{\delta}_{i(p)})'(y_j - Z_j \hat{\delta}_{j(p)})/T \quad \text{and} \quad \hat{\Sigma}_{(p)} = (\hat{\sigma}_{ij(p)}),$$

where  $\hat{Z}_{(p)} = \text{diag}(\hat{Z}_1, \dots, \hat{Z}_G)$ ,  $\hat{Z}_j = (\hat{Y}_j, X_j)$  and  $\hat{Y}_j$  comes from the solution values of the estimated system based on the estimate  $\hat{\delta}_{(p)}$ .

The formula for  $\hat{\delta}_{p+1}$  follows the 3SLS formula in (6.20) with the following major differences:

1.  $\hat{\Sigma}$  is updated through the iteration rounds.
2. The endogenous components  $\hat{Y}_j$  in  $\hat{Z}_{(p)}$  are constructed as *solutions* to the structural system (as in FIVE) and not as in 3SLS.

We end this section with some additional algebraic relationships among the estimators.

1. 3SLS reduces to 2SLS when  $\Sigma$  is diagonal or when all equations are just identified.
2. If some equations are overidentified and some are just identified, then
  - (a) the 3SLS estimates of the overidentified equations can be obtained by the application of 3SLS to the set of over identified equations, ignoring all just identified equations.
  - (b) the 3SLS estimate of each just identified equation differs from the 2SLS estimate by a vector which is a linear function of the 3SLS residuals of the overidentified equations. In particular, if we are dealing with a system consisting of one just identified equation and one overidentified, then 2SLS and 3SLS will be exactly equivalent for the overidentified but not for the just identified equation.

## 4 LARGE SAMPLE PROPERTIES OF ESTIMATORS

The statistical behavior of these estimators has been analyzed in finite samples and in the following alternative parameter sequences:

1. Sample size  $T \rightarrow \infty$  – the standard asymptotic analysis.
2. Concentration parameter  $\rightarrow \infty$  with  $T$  fixed. For example, see Basmann (1961), Mariano (1975), Anderson (1977) and Staiger and Stock (1997). The concentration parameter is defined in Section 6.2 of this chapter.
3. Structural error variances go to zero – the so-called small- $\sigma$  asymptotics discussed in Kadane (1971), and Morimune (1978).
4. Number of instruments,  $L$ , goes to infinity with sample size such that  $L/T \rightarrow \alpha$  ( $0 < \alpha < \infty$ ). See Anderson (1977), Kunitomo (1980), Morimune and Kunitomo (1980), Morimune (1983), and Bekker (1994).
5. Weak instrument asymptotics. Here  $L$  is fixed and the coefficients of instruments in the first stage regression in IV or modified 2SLS estimators are assumed to be  $O(T^{-1/2})$  as a way of representing weak correlation between the instruments and the endogenous explanatory variables – see Staiger and Stock (1997), and Bound, Jager and Baker (1995). Related discussion of structural testing, model diagnostics and recent applications is in Wang and Zivot (1998), Angrist (1998), Angrist and Krueger (1992), Donald and Newey (1999), and Hahn and Hausman (1999).

In this section, we summarize results under large sample asymptotics, then consider finite sample properties and end with a discussion of the practical implications of the analysis based on these alternative approaches.

### 4.1 Large sample properties of limited information estimators

We now consider the large sample asymptotic behavior of the limited information estimators described in Section 2. Because of space limitation, theorems are stated without proof. We start with a theorem for limited information instrumental variable estimators in general.

Thus, first consider the IV estimator defined in (6.5):  $\hat{\delta}_{IV} = (W'Z)^{-1}W'y_1$ , where the instrument matrix  $W$  is of the same dimension as  $Z \equiv (Y_1, X_1)$  such that

$$(1) \quad \text{plim } W'u_1/T = 0 \quad (6.25)$$

$$(2) \quad \text{plim } W'W/T = Q_W, \text{ positive definite and finite} \quad (6.26)$$

$$(3) \quad \text{plim } W'Z/T = M', \text{ nonsingular} \quad (6.27)$$

$$(4) \quad W'u_1/\sqrt{T} \text{ converges in distribution to } N(0, \sigma^2 Q_w). \quad (6.28)$$

Note that (6.25) together with (6.27) implies that  $\hat{\delta}_{IV}$  is consistent. Also, (6.25) is a consequence of (6.28) and hence (6.25) would be unnecessary if (6.28) is assumed. (6.26) is not necessarily satisfied by the original model; for example, when  $X$  contains a trending variable. For cases like these, the theorems we discuss presently will require further modifications. Both (6.26) and (6.27) will need minimally the assumption that  $(X'X/T)$  has a finite positive definite limit as  $T$  approaches infinity. (6.28) depends on a multivariate version (e.g. Lindeberg–Feller) of the central limit theorem for independent but nonidentically distributed random vectors.

Assuming that the instrument matrix  $W$  satisfies the three properties (6.26)–(6.28), we have

### Theorem 1.

1.  $\sqrt{T}(\hat{\delta}_{IV} - \delta) \xrightarrow{d} N(0, \sigma^2(MQ_W^{-1}M')^{-1}) \equiv N[0, \sigma^2 \text{plim}(Z'P_W Z/T)^{-1}]$
2. A consistent estimator of  $\sigma^2$  is  $(y_1 - Z\hat{\delta}_{IV})'(y_1 - Z\hat{\delta}_{IV})/T$ .

Applying the above result to two-stage least squares, we get

**Theorem 2.**  $\sqrt{T}(\hat{\delta}_{2SLS} - \delta) \xrightarrow{d} N(0, \sigma^2(RQ^{-1}R')^{-1}) \equiv N[0, \sigma^2 \text{plim}(Z'P_X Z/T)^{-1}]$ , where  $R = \text{plim}(Z'X/T)$  and  $Q = \text{plim}(X'X/T)$ , under the assumptions that

1.  $R$  exists, is finite, and has full row rank,
2.  $Q$  exists and is finite and positive definite,
3.  $X'u_1/(T^{1/2}) \rightarrow N(0, \sigma^2 Q)$ .

**Corollary 1.** A consistent estimator of  $\sigma^2$  is  $(y_1 - Z\hat{\delta}_{2SLS})'(y_1 - Z\hat{\delta}_{2SLS})/T$ .

For the  $k$ -class estimator, we have already indicated that  $\text{plim } k = 1$  is a sufficient condition for consistency. A sharper result derives from the two previous theorems.

**Theorem 3.** If  $k$  is such that  $\text{plim } T^{1/2}(k - 1) = 0$ , then  $T^{1/2}(\hat{\delta}_{2SLS} - \delta)$  and  $T^{1/2}(\hat{\delta}_{(k)} - \delta)$  are asymptotically equivalent.

For the LIML estimator, we have the following result as a consequence of the preceding theorem.

**Theorem 4.** If the data matrix  $X$  is such that  $(X'X/T)$  has a finite positive definite limit, then, under assumptions A1, A2, and A4' in Section 1, LIML is consistent and asymptotically equivalent to 2SLS; that is  $\sqrt{T}(\hat{\delta}_{\text{LIML}} - \delta) \xrightarrow{d} N(0, \sigma^2(RQ^{-1}R')^{-1})$  where  $R$  and  $Q$  are as defined in Theorem 2.

Under the conditions of the theorem, it can be shown that  $\text{plim } \sqrt{T}\ell = 0$  where  $\ell$  is the smallest characteristic root of  $|W - \lambda S| = 0$ , where  $W$  and  $S$  are the second moment residual matrices defined in (6.15). Thus, Theorem 4 now follows directly from Theorem 3 and the fact that LIML is equivalent to  $k$ -class with  $k$  equal to  $1 + \ell$ .

Note that  $\sqrt{T}\ell \xrightarrow{p} 0$  is a consequence of the stronger result that  $T\ell$  converges in distribution to a central chi-squared distribution with degrees of freedom equal to the number of overidentifying restrictions.

If the structural errors  $u_t$  are normally distributed, then LIML is asymptotically efficient among all consistent and uniformly asymptotic normal (CUAN) estimators of  $\beta$  and  $\gamma$  based on the “limited information” model discussed in the preceding section. By Theorem 4, under normality assumptions, 2SLS shares with LIML this property of being asymptotically efficient within this class of CUAN estimators of  $\beta$  and  $\gamma$ .

**Theorem 5.** Let  $\hat{\delta}_{\text{LIVE}}$  be the IV estimator of  $\delta$  based on the instrument matrix  $W_L = (X\hat{\Pi}_1', X_1)$ , where  $\hat{\Pi}_1$  is any consistent estimator of  $\Pi_1$ . Then  $\hat{\delta}_{\text{LIVE}}$  is asymptotically equivalent to  $\hat{\delta}_{\text{2SLS}}$ .

**Theorem 6.** Consider the class of IV estimators of the first equation where instruments are of the form  $W = XF$ , where  $F$  is either stochastic or nonstochastic and of size  $K \times (K_1 + G_1 - 1)$  and of full rank. Within this class, two-stage least squares is asymptotically efficient.

**Theorem 7.** For the modified 2SLS estimator  $\hat{\delta}_{\text{M2SLS}}$ , where the first stage regressor matrix is  $H$  instead of  $X$ ,

1.  $\hat{\delta}_{\text{M2SLS}}$  is exactly equivalent to the IV estimator of  $\delta$  using the instrument matrix  $(P_H Y_1, X_1)$  if  $X_1$  is contained in the column space of  $H$ .
2.  $\hat{\delta}_{\text{M2SLS}}$  is consistent if and only if the column space of  $H$  contains the column space of  $X_1$ .

## 4.2 Large sample properties of full information estimators

**Theorem 8.** For a linear simultaneous equations model satisfying the assumptions in Theorem 2 and with a nonsingular error covariance matrix, the following asymptotic properties of the 3SLS estimator hold:

$$(a) \quad \sqrt{T}(\hat{\delta}_{\text{3SLS}} - \delta) \xrightarrow{d} N(0, V_{\text{3SLS}})$$

$$(b) \quad V_{\text{3SLS}} \leq V_{\text{2SLS}}$$

where

$$V_{3SLS} = \text{plim } (Z'(\Sigma^{-1} \otimes P_X)Z/T)^{-1}$$

$$V_{2SLS} = \text{plim } \{1/T(Z'(I \otimes P_X)Z)^{-1}(Z'(\Sigma^{-1} \otimes P_X)Z)(Z'(I \otimes P_X)Z)^{-1}\}.$$

**Theorem 9.** Under regularity conditions which are satisfied by the classical simultaneous equations model

$$\sqrt{T}(\hat{\delta}_{FIML} - \delta) \xrightarrow{d} N\{0, \text{plim } [(1/T)(\partial^2 \log L / \partial \delta \partial \delta')]^{-1}\}.$$

Furthermore, if there are no restrictions on  $\Sigma$ , then the asymptotic covariances of the 3SLS and FIML estimators coincide and thus, 3SLS and FIML are asymptotically equivalent. Finally, FIML and 3SLS are asymptotically efficient within the class of all consistent, uniformly asymptotically Gaussian estimators of  $\delta$ .

**Theorem 10.** FIML and linearized FIML are asymptotically equivalent.

For the instrumental variable method, suppose the instrument matrix for  $Z$  in  $y = Z\delta + u$  is  $\hat{Z}$ , of the form

$$\hat{Z} = \begin{pmatrix} \hat{Z}_{11} & \hat{Z}_{12} & \dots & \hat{Z}_{1G} \\ \hat{Z}_{21} & \hat{Z}_{22} & \dots & \hat{Z}_{2G} \\ \vdots & \vdots & \ddots & \vdots \\ \hat{Z}_{G1} & \hat{Z}_{G2} & \dots & \hat{Z}_{GG} \end{pmatrix}.$$

Recall that the data matrix  $Z$  is diagonal ( $Z_1, Z_2, \dots, Z_G$ ).

Following the partitioning of  $Z_j$  into  $(Y_j, X_j)$ , we can further decompose each submatrix

$$\hat{Z} \text{ as } \hat{Z}_{ij} = (\hat{Z}_{ij1}, \hat{Z}_{ij2})$$

where  $\hat{Z}_{ij1}$  and  $\hat{Z}_{ij2}$  are parts of the instrument matrices for the right-hand side endogenous and exogenous variables, respectively, appearing in the  $j$ th equation with respective dimensions  $T \times (G_j - 1)$  and  $T \times K_j$ .

**Theorem 11.** For the simultaneous system  $y = Z\delta + u$ , let  $\hat{\delta}_{IV} = (\hat{Z}'Z)^{-1}\hat{Z}'y$ , where  $\hat{Z}$  is defined in the preceding paragraph. Suppose

1.  $\Pr(|\hat{Z}'Z| \neq 0) = 1$
2.  $\text{plim } (\hat{Z}'u/T) = 0$
3.  $\text{plim } (\hat{Z}'Z/T)$  is nonsingular and finite
4.  $\text{plim } (\hat{Z}'Z/T) = \text{plim } (\hat{Z}'(\Sigma \otimes I)\hat{Z}/T)$ .

Then  $\hat{\delta}_{IV}$  and  $\hat{\delta}_{3SLS}$  are asymptotically equivalent if and only if the following two conditions hold:

1.  $\text{plim } (\hat{Z}'_{ij1}X/T) = \sigma^{ij}\Pi_j \text{plim } (X'X/T)$
2.  $\text{plim } (\hat{Z}'_{ij2}X/T) = \sigma^{ij} \text{plim } (X'_jX/T)$ .

## 5 STRUCTURE OF LIMITED INFORMATION ESTIMATORS AS REGRESSION FUNCTIONS

The structure of the SEM estimators is discussed at length in Hendry (1976) and Hausman (1983), and Phillips (1983). In this section, we develop a perspective on the structure of limited information estimators which is particularly helpful in the finite sample analysis of these procedures.

Most of the limited information estimators we have discussed so far can be related to the regression moment matrices  $S$ ,  $W$ , and  $A$  defined in (6.15):

$$\begin{aligned}\hat{\beta}_{2SLS} &= \arg \min \beta_*' W \beta_* \\ \hat{\beta}_{OLS} &= \arg \min \beta_*' A \beta_* \\ \hat{\beta}_{LIML} &= \arg \min (\beta_*' W \beta_*/\beta_*' S \beta_*).\end{aligned}$$

If we partition  $S$  (and  $A$  and  $W$  similarly) as

$$S = \begin{pmatrix} S_{11} & S_{12} \\ S_{21} & S_{22} \end{pmatrix},$$

we then have

$$\begin{aligned}\hat{\beta}_{OLS} &= A_{22}^{-1} A_{21} \\ \hat{\beta}_{2SLS} &= W_{22}^{-1} W_{21}\end{aligned}\tag{6.29}$$

and for the  $k$ -class estimator, with  $\bar{k} = 1 - k$ ,

$$\hat{\beta}_{(k)} = (W_{22} + \bar{k}S_{22})^{-1}(W_{21} + \bar{k}S_{21}).\tag{6.30}$$

Thus, we also have

$$\hat{\beta}_{LIML} = (W_{22} - \ell S_{22})^{-1}(W_{21} - \ell S_{21}),\tag{6.31}$$

where  $\ell = \text{smallest eigenvalue of } S^{-1}W$ .

We can also give a similar characterization to the modified 2SLS estimator where the data matrix for the first stage regressor is  $H$  (constrained to contain  $X_1$ ) instead of  $X$ . Equivalently, this is the IV estimator using  $P_H Y_1$  as the instrument matrix for  $Y_1$  and its characterization is  $\hat{\beta}_{M2SLS} = F_{22}^{-1} F_{21}$ , where  $F = Y_1'(P_H - P_{X_1})Y_1$ . Thus  $\hat{\beta}_{OLS}$ ,  $\hat{\beta}_{2SLS}$  and  $\hat{\beta}_{M2SLS}$  are regression functions of moment matrices in  $Y_1$  – namely,  $A$ ,  $W$ , and  $F$ , which differ among themselves only in their associated projection matrices.

In the scalar Gaussian case,  $A$ ,  $W$ , and  $F$  are all proportional to noncentral Chi-squared variates with different degrees of freedom. In higher dimensional cases, they would have a so-called noncentral Wishart distribution which is indexed by the following parameters:

1. the order or size of the matrix (in our case  $G_1 + 1$ ),
2. the degrees of freedom or rank of the associated matrix  $Q$  (say  $q$ ),
3. the common covariance matrix of the rows of  $Y$  (say  $\Omega$ ),
4. the so-called means–sigma matrix, which is the generalization of the non-centrality parameter and is equal to  $(E(Y))'Q(E(Y)) = M$ , and
5. the rank of the means–sigma matrix, say  $m$ .

The Wishart distributions of  $A$ ,  $W$ , and  $F$  differ in their degrees of freedom. The means–sigma matrices for  $A$  and  $W$  are identical; that for  $F$  takes a different expression but it has the same rank as the first – one less than the order of the matrices (see Mariano, 1977). Consequently, the probability density functions for  $A$ ,  $W$ , and  $F$  are all of the same form. Because of this, we can say that OLS, 2SLS, and M2SLS are “distributionally equivalent” in the sense that the problem of deriving analytical results for one is of the same mathematical form as for the other two.

Distributional equivalence in the same vein exists between two-stage least squares in the just identified case and the instrumental variable estimator based on nonstochastic instruments for  $Y_1$  (see Mariano, 1977, 1982). The argument turns on the fact that 2SLS applied to a just identified equation can be interpreted as an IV estimator that uses the excluded exogenous variables as instruments for  $Y_1$ .

Expressions (6.29), (6.30), and (6.31) of the  $k$ -class estimators in terms of the matrices  $S$ ,  $A$ , and  $W$  also lead to a generalized method-of-moments (GMM) interpretation for these estimators. The moment conditions come from the asymptotic orthogonality of the instruments relative to the structural disturbances. Under standard large-sample asymptotics, these moment conditions are satisfied by 2SLS, LIML and the  $k$ -class with  $k$  converging to 1 in probability.

Moment conditions can also be derived by expressing  $\beta$  in terms of expectations of the matrices  $A$ ,  $W$ , and  $S$ ; see Bekker (1994). Under Bekker’s (1994) alternative asymptotics, where the number of instruments increases with sample size, LIML satisfies these moment conditions; but 2SLS and OLS do not. Consequently, under this alternative asymptotics LIML remains consistent while 2SLS and OLS are both inconsistent. This result provides a partial intuitive explanation for better LIML finite sample properties than 2SLS under conditions given in Section 7 of this chapter.

## 6 FINITE SAMPLE PROPERTIES OF ESTIMATORS

### 6.1 Arbitrary number of included endogenous variables

Let us first consider the available results dealing with the general case where there are no further restrictions on the linear system nor on the equation being estimated apart from the condition of identifiability and the classical Gaussian assumptions. The characterization in (6.29)–(6.31) is the starting point of the analysis.

The first group of results here deals with the existence and nonexistence of moments of various estimators and is best summarized in the following theorem.

**Theorem 12.** In an identified equation in a linear simultaneous system satisfying assumptions A1–A4', absolute moments of positive order for indicated estimators of coefficients in this structural equation are finite up to (and including) order

1.  $K_2 - G_1$  for 2SLS and 3SLS;
2.  $h - K_1 - G_1$ , for modified 2SLS (the class of stochastic IV estimators) where  $h$  is the number of linearly independent first-stage regressors;
3.  $T - K_1 - G_1$ , for the  $k$ -class estimators with  $k$  nonstochastic and  $0 \leq k < 1$ ;
4. 0, for the  $k$ -class with nonstochastic  $k$  exceeding 1;
5. 0, for instrumental variable estimators with nonstochastic instruments; and
6. 0, for LIML and FIML.

Because of the distributional equivalence results discussed in the preceding section, (2) follows from (1). As a corollary to (1), 2SLS in the just identified case will have no finite absolute moments of positive order; as a consequence (5) also follows from (1).

References to specific authors who derived these results are given in Mariano (1982) and Phillips (1983).

From Theorem 12, we see that the LIML and FIML estimators (as well as 2SLS if the degree of over-identification is less than two) are inadmissible under a strictly quadratic loss function. Of course, admissibility comparisons may change under alternative loss functions with finite risks for the estimators. Such alternatives could be based on probability concentration around the true parameter value or other loss functions that increase at a slower rate than quadratic.

Beyond these moment existence results, various authors have derived closed-form expressions for the exact moments and probability density functions of these estimators – see Phillips (1983) for specific references. The expressions are rather complicated, in terms of zonal polynomials.

Another class of results in the general case deals with asymptotic expansions for the estimators themselves – namely, expressions of the type

$$\hat{\alpha} = \sum_{j=0}^r a_j + O_p(N^{-(r+1)/2}),$$

where  $a_j = O_p(N^{-j/2})$  and  $a_0$  is the probability limit of  $\hat{\alpha}$  as  $N \rightarrow \infty$ . See Rothenberg (1984) for an extensive review. For the  $k$ -class estimator, such an expansion can be obtained by noting that we can write  $\hat{\beta}_{(k)} - \beta = H(I + \Delta)^{-1}\eta$ , where  $H$  is nonstochastic,  $\Delta = O_p(N^{-1/2})$  and  $\eta = O_p(1)$  and then applying an infinite series expansion for the matrix inverse. One of the earliest papers on this in the econometric literature is Nagar (1959), where expansions of this type are formally obtained for the  $k$ -class estimator when  $k$  is of the form  $1 + c/T$  where  $c$  is nonstochastic.

Given an asymptotic expansion like

$$\hat{\beta}_{(k)} - \beta = F_0 + F_{-1/2} + F_{-1} + O_p(T^{-3/2}) = F + O_p(T^{-3/2}),$$

we can then proceed to define the “asymptotic moments” of  $(\hat{\beta}_{(k)} - \beta)$  to be the moments of the stochastic approximation up to a certain order; for example, the moments of  $F$ . Thus, asymptotic moments determined this way relate directly to the moments of an approximation to the estimator itself and will not necessarily be approximations to the exact moments of  $\hat{\beta}_{(k)}$ . For example, this will not hold for LIML and other estimators for which moments of low positive order do not exist. In cases where moments do exist, Sargan (1974, 1976) gives conditions under which these asymptotic moments are valid approximations to the moments of the estimator.

## 6.2 The case of two included endogenous variables

For a more concrete discussion, let us assume now that the equation being estimated contains only two endogenous variables. The early work on the analytical study of finite sample properties of SEM estimators dealt with this case. The equation now takes the following form:

$$y_1 = y_2\beta + X_1\gamma + u_1, \quad (6.32)$$

where  $\beta$  is scalar and  $y_2$  is  $T \times 1$ . The characterization of the limited information estimators in (6.29)–(6.31) can be reduced further to canonical form for a better understanding of how parameter configurations affect the statistical behavior of the estimators (for example, see Mariano, 1977 and 1982). One reduction to canonical form simplifies in the case of two included endogenous variables to

$$\hat{\beta}_{(k)} - \beta = (\sigma/\omega)\sqrt{1 - \rho^2} (\hat{\beta}_{(k)}^* - \lambda), \quad (6.33)$$

where  $\sigma^2$  = variance of the structural disturbance  $u_{t1}$ ,  $\omega^2$  = variance of  $y_{t2}$ ,  $\rho$  = correlation coefficient between  $y_{t2}$  and  $u_{t1}$ ,  $\lambda = \rho/\sqrt{1 - \rho^2}$ , and

$$\hat{\beta}_{(k)}^* = (x'y + \bar{k}u'v)/(y'y + \bar{k}v'v). \quad (6.34)$$

The vectors  $x$ ,  $y$ ,  $u$ , and  $v$  are mutually independent multivariate normal (the first two are  $K_2 \times 1$  and the last two are  $(T - K) \times 1$ ) with unit variances, mean vector equal to zero for  $u$  and  $v$  and equal to  $\alpha_x = (0, \dots, 0, \mu\lambda)'$  and  $\alpha_y = (0, \dots, 0, \mu)'$  for  $x$  and  $y$ , with  $\mu^2 = [(E(y_2))'(P_x - P_{X_1})(E(y_2))]/\omega^2$ .

The last quantity,  $\mu^2$ , has been called the “concentration parameter” in the literature. This derives from the fact that as this parameter increases indefinitely, with sample size staying fixed, for  $k$  nonstochastic as well as for LIML,  $\hat{\beta}_{(k)}$  converges in probability to the true parameter value  $\beta$ ; Basmann (1961) and Mariano (1975). Also, in large sample asymptotics with the usual assumption that  $X'X/T$  tends to a finite, positive definite limit, the variance in the limiting normal distribution of  $\sqrt{T}(\hat{\beta}_{2SLS} - \beta)$  is inversely proportional to the limit of  $\mu^2/T$ . (The asymptotic variance is  $(\sigma^2/\omega^2)(\lim(\mu^2/T))^{-1}$ .)

One can then use (6.34) to derive exact expressions for moments and probability distributions of the  $k$ -class estimator – e.g. see Mariano (1982) and Phillips

(1983) for a more detailed survey. The critical parameters are the correlation between error and endogenous regressor, the concentration parameter, the degree of over-identification, the difference between sample size and number of exogenous variables, and the size of structural error variance relative to the variance of the endogenous regressor.

## 7 PRACTICAL IMPLICATIONS

Instead of going into more complicated formulas for pdfs or cdfs or moments of estimators, we devote this section to a discussion of the practical implications of the finite sample results and alternative asymptotics – covering material that extends from the early works in the 1960s to the recent results on exact small sample properties of IV estimators and the behavior of IV estimators when instruments are weak. For finite sample results, most of the conclusions come from the study of the case of two included endogenous variables.

1. For the  $k$ -class estimator (nonstochastic  $k \in [0, 1]$ ) bias is zero if and only if  $\rho$ , the correlation between the structural error and the endogenous regressor, is equal to zero. The direction of bias is the same for all  $k$  and it follows the direction of  $\rho$ . Negative correlation implies a downward bias; positive correlation implies an upward bias.

2. Absolute bias is an increasing function of the absolute value of  $\rho$ , a decreasing function of the concentration parameter  $\mu^2$  and a decreasing concave function of  $k$ . Thus, whenever both exist, OLS bias is always greater in absolute value than 2SLS bias. Of course, if the equation is just identified, then 2SLS bias does not exist while OLS bias is finite.

3. For two-stage least squares, absolute bias is an increasing function of the degree of over-identification. Since the 2SLS probability distribution depends on sample size only through the concentration parameter and since the value of the concentration parameter increases with additional observations, then 2SLS bias in absolute value decreases upon inclusion of more observations in the sample. The total effect of additional observations on OLS bias, on the other hand, is indeterminate. An increase in sample size produces a positive effect on absolute OLS bias (because of the increase in degrees of freedom) and a negative indirect effect through the increase in the concentration parameter.

4. The size of the OLS bias relative to 2SLS gets larger with higher  $\mu^2$ , lower degree of overidentification, bigger sample size, and higher absolute  $\rho$ .

5. For the  $k$ -class, the optimal value of nonstochastic  $k$  over  $[0, 1]$  for minimizing mean squared error varies over a wide range according to slight changes in parameter values and the sample size.

6. For the whole  $k$ -class,  $k$  nonstochastic and in  $[0, 1]$ , exact mean squared error is a decreasing function of the concentration parameter, an increasing function of the absolute value of  $\rho$  and an indefinite function of the degrees of freedom parameter ( $K_2 - 1$  or the degree of overidentification for 2SLS,  $T - K_1$  for OLS,  $h - K_1 - 1$  for M2SLS). Interpreting M2SLS as an IV method, keep in mind that  $h$  represents the number of instruments. Note also the *ceteris paribus* conditions here: all other parameters are kept fixed as a specific one, say the concentration

parameter, changes. Furthermore, because the 2SLS distribution depends on sample size only through the concentration parameter (see item 3), it follows that the exact mean squared error for 2SLS is a nonincreasing function of sample size. This does not apply, however, to the other  $k$ -class estimators; for them the net effect of increasing sample size is indefinite.

7. In terms of relative magnitudes of MSE, large values of  $\mu^2$  and large  $T$  favor 2SLS over OLS. One would expect this since the usual large-sample asymptotics would be taking effect and the dominant term would be the inconsistency in OLS. However, there are cases (small values of  $\rho$  and  $T$ ) where OLS would dominate 2SLS even for large values of  $\mu^2$ .

8. When the degree of overidentification gets large, the 2SLS and OLS distributions tend to be similar. This follows from the fact that the only difference in the distributions of 2SLS and OLS lies in the degrees of freedom parameter. These are the degree of overidentification for 2SLS and sample size less ( $K_1 + 1$ ) for OLS so that the smaller ( $T - K$ ) is, the more similar the 2SLS and OLS distributions will be. Of course, there will be no perfect coincidence since sample size is strictly greater than  $K$ .

9. The OLS and 2SLS distributions are highly sensitive to  $\rho$  and the 2SLS distribution is considerably asymmetric while the OLS distribution is almost symmetric.

10. The extensive tabulations of the 2SLS distribution function in Anderson and Sawa (1979) provide considerable insight into the degree of asymmetry and skewness in the 2SLS distribution. Bias in the direction of  $\rho$  is quite pronounced. For some combinations of parameter values (such as  $K_2 \geq 20$ , low concentration parameter and high numerical value of  $\rho$ ), the probability is close to 1 that the 2SLS estimator will be on one side of the true value: e.g. less than the true value if  $\rho$  is negative. With regard to convergence to normality, when either  $\rho$  or  $K_2$  or both are large, the 2SLS distribution tends to normality quite slowly. In comparison with 2SLS, the LIML distribution is far more symmetric though more spread out and it approaches normality faster.

11. Up to terms of order  $T^{-1}$ , the approximate LIML distribution (obtained from large sample asymptotic expansions) is median unbiased. For 2SLS, the median is  $\beta$  only if the equation is just identified or if  $\rho = 0$ . Up to order  $T^{-1/2}$ , the approximate distribution functions for both 2SLS and LIML assign the same probability as the normal to an interval which is symmetric about  $\beta$ . Also, the asymptotic mean squared errors, up to  $T^{-1/2}$ , reproduce that implied by the limiting normal distribution.

12. Anderson (1974) compares asymptotic mean squared errors of 2SLS and LIML up to order  $T^{-1}$  and finds that for a degree of overidentification ( $v$ ) strictly less than 7, 2SLS would have a smaller asymptotic mean squared error than LIML. For  $v$  greater than or equal to 7 and for  $\alpha^2 = \rho^2 \omega^2 \sigma^2 / (\omega_{11} \omega^2 - \omega_{12})$  not too small, LIML will have the smaller AMSE. Calculation of probabilities of absolute deviations around  $\beta$  leads to the same conclusion: small  $\rho^2$  or little simultaneity favors 2SLS while a high degree of over-identification favors LIML. The condition for 2SLS to have the advantage over LIML in this case is

$$\alpha^2 \leq 2 / (K_2 - 1) = 2/v.$$

13. In dealing with the case of one explanatory endogenous variable and one instrument (the just-identified case), Nelson and Startz (1990) find that

1. the probability distribution of the IV estimator can be bimodal. Maddala and Jeong (1992) show that this is a consequence of near singularity of reduced form error covariance matrices but not necessarily of poor instruments.
2. The asymptotic distribution of the IV estimator is a poor approximation to the exact distribution when the instrument has low correlation with the regressor and when the number of observations is small.

14. As a further amplification of item 2, in dealing with the bias of instrumental variable estimators, Buse (1992) derives conditions under which Phillips' (1980, 1983) observation would hold – that  $\hat{\beta}_{IV}$  would display more bias as the number of instruments increases. Buse shows that IV bias would increase or decrease with increased number of IV instruments depending on an inequality based on quadratic forms of incremental regression moment matrices of the right-hand side endogenous variables. In the case where there is only one right-hand side endogenous variable, the result simplifies to the conclusion that the estimated IV bias will increase with the number of excess instrumental variables “only if the proportional increase in the instruments is faster than the rate of increase in  $R^2$  measured relative to the fit of  $Y_1$  on  $X_1$ .” Thus, adding less important instrumental variables later will add little to  $R^2$  and increase IV bias. On the other hand, one could start with weak instruments and find that  $R^2$  rises dramatically (with a decline in IV bias) as important instruments are added. Consequently, whether or not there is an improvement in efficiency tradeoff between bias and variability in IV as more instruments are added depends on the IV selection sequence.

15. In Bekker's (1994) asymptotic analysis where the number of instruments increases at the same rate as sample size, there is numerical evidence that approximations to distributions of IV estimators under this parameter sequence are more accurate than large sample approximations, even if the number of instruments is small. Confidence regions based on this alternative asymptotic analysis also produce more accurate coverage rates when compared to standard IV confidence regions. Under this alternative asymptotics, 2SLS becomes inconsistent while LIML remains consistent. The asymptotic Gaussian distribution of LIML depends on  $\alpha$ , the limit of  $L/T$ , but the LIML asymptotic covariance matrix can be estimated by a strictly positive definite matrix without estimating  $\alpha$  or specifying  $L$  – see Bekker (1994). Bekker's numerical analysis shows that inference based on this estimated limiting distribution is more accurate than that based on large sample asymptotics (where  $\alpha = 0$ ).

16. From their weak-instrument asymptotics (number of instruments is fixed, coefficients of instruments in the first stage regression go to zero at the rate of  $T^{-1/2}$ ), Staiger and Stock (1997) conclude that

1. Conventional asymptotic results are invalid, even when sample size is large. The  $k$ -class estimator is not consistent and has a nonstandard asymptotic

distribution. Similarly, Bound *et al.* (1995) find large inconsistencies in IV estimates when instruments are weak.

2. 2SLS and LIML are not asymptotically equivalent. 2SLS can be badly biased and can produce confidence intervals with severely distorted coverage rates. In light of this, nonstandard methods for interval estimation should be considered.
3. Estimator bias is less of a problem for LIML than 2SLS (when there are two included endogenous variables).
4. When doing IV estimation, the  $R^2$  or F-statistic in the first stage regression should be reported. Bound *et al.* (1995) also recommend this as a useful indicator of the quality of IV estimates.

## References

- Anderson, T.W. (1974). An asymptotic expansion of the distribution of the limited information maximum likelihood estimate of a coefficient in a simultaneous equation system. *Journal of the American Statistical Association* 69, 565–73.
- Anderson, T.W. (1977). Asymptotic expansions of the distributions of estimates in simultaneous equations for alternative parameter sequences. *Econometrica* 45, 509–18.
- Anderson, T.W., and T. Sawa (1979). Evaluation of the distribution function of the two-stage least squares estimate. *Econometrica* 47, 163–82.
- Angrist, J. (1998). Estimating the labor market impact of voluntary military service using social security data on military applicants. *Econometrica* 66, 249–88.
- Angrist, J., and A. Krueger (1992). The effect of age of school entry on educational attainment: An application of instrumental variables with moments from two samples. *Journal of the American Statistical Association* 87, 328–36.
- Basman, R.L. (1961). A note on the exact finite sample frequency functions of GCL estimators in two leading over-identified cases. *Journal of the American Statistical Association* 56, 619–36.
- Bekker, P.A. (1994). Alternative approximations to the distribution of instrumental variable estimators. *Econometrica* 62, 657–81.
- Bekker, P.A., and T.K. Dijkstra (1990). On the nature and number of constraints on the reduced form as implied by the structural form. *Econometrica* 58, 507–14.
- Bound, J., D.A. Jaeger, and R.M. Baker (1995). Problems with instrumental variables estimation when the correlation between the instruments and the endogenous explanatory variable is weak. *Journal of the American Statistical Association* 90, 443–50.
- Buse, A. (1992). The bias of instrumental variable estimators. *Econometrica* 60, 173–80.
- Donald, S., and W. Newey (1999). Choosing the number of instruments. M.I.T. Working Paper, Department of Economics.
- Hahn, J., and J. Hausman (1999). A new specification test for the validity of instrumental variables. M.I.T. Working Paper, Department of Economics.
- Hausman, J. (1983). Specification and estimation of simultaneous equation models. *The Handbook of Econometrics, Volume I* pp. 393–448. North-Holland Publishing Company.
- Hendry, D.F. (1976). The structure of simultaneous equations estimators. *Journal of Econometrics* 4, pp. 51–88.
- Hsiao, C. (1983). Identification. *The Handbook of Econometrics, Volume I* pp. 223–83. North-Holland Publishing Company.
- Kadane, J. (1971). Comparison of  $k$ -class estimators when the disturbances are small. *Econometrica* 39, 723–37.

- Kunitomo, N. (1980). Asymptotic expansions of the distributions of estimators in a linear functional relationship and simultaneous equations. *Journal of the American Statistical Association* 75, 693–700.
- Maddala, G.S., and J. Jeong (1992). On the exact small sample distribution of the instrumental variable estimator. *Econometrica* 60, 181–3.
- Mariano, R.S. (1982). Analytical small-sample distribution theory in econometrics: The simultaneous-equations case. *International Economic Review* 23, 503–34.
- Mariano, R.S. (1975). Some large-concentration-parameter asymptotics for the  $k$ -class estimators. *Journal of Econometrics* 3, 171–7.
- Mariano, R.S. (1977). Finite-sample properties of instrumental variable estimators of structural coefficients. *Econometrica* 45, 487–96.
- Morimune, K. (1978). Improving the limited information maximum likelihood estimator when the disturbances are small. *Journal of the American Statistical Association* 73, 867–71.
- Morimune, K. (1983). Approximate distribution of  $k$ -class estimators when the degree of overidentifiability is large compared with the sample size. *Econometrica* 51, 821–41.
- Morimune, K., and N. Kunitomo (1980). Improving the maximum likelihood estimate in linear functional relationships for alternative parameter sequences. *Journal of the American Statistical Association* 75, 230–7.
- Nagar, A.L. (1959). The bias and moment matrix of the general  $k$ -class estimators of the parameters in structural equations. *Econometrica* 27, 575–95.
- Nelson, C.R., and R. Startz (1990). Some further results on the exact small sample properties of the instrumental variable estimator. *Econometrica* 58, 967–76.
- Phillips, P.C.B. (1980). The exact finite-sample density of instrumental variable estimators in an equation with  $n + 1$  endogenous variables. *Econometrica* 48, 861–78.
- Phillips, P.C.B. (1983). Exact small sample theory in the simultaneous equations model. *The Handbook of Econometrics, Volume I* pp. 449–516. North-Holland Publishing Company.
- Rothenberg, T. (1984). Approximating the distributions of econometric estimators and test statistics. *Handbook of Econometrics, Volume II* pp. 881–935. Elsevier Science Publishers.
- Sargan, J.D. (1974). The validity of Nagar's expansion for the moments of econometric estimators. *Econometrica* 42, 169–76.
- Sargan, J.D. (1976). Econometric estimators and the Edgeworth approximation. *Econometrica* 44, 421–48.
- Staiger, D., and J. Stock (1997). Instrumental variables regression with weak instruments. *Econometrica* 65, 557–86.
- Wang, J., and E. Zivot (1998). Inference on structural parameters in instrumental variables regression with weak instruments. *Econometrica* 66, 1389–404.

CHAPTER SEVEN

# Identification in Parametric Models

*Paul Bekker and Tom Wansbeek\**

## 1 INTRODUCTION

Identification is a notion of essential importance in quantitative empirical branches of science like economics and the social sciences. To the extent that statistical inference in such branches of science extends beyond a mere exploratory analysis, the generic approach is to use the subject matter theory to construct a stochastic model where the parameters in the distributions of the various random variables have to be estimated from the available evidence. Roughly stated, a model is then called identified when meaningful estimates for these parameters can be obtained. If that is not the case, the model is called underidentified. In an underidentified model different sets of parameter values agree equally well with the statistical evidence. Hence, preference of one set of parameter values over other ones is arbitrary. Scientific conclusions drawn on the basis of such arbitrariness are in the best case void and in the worst case dangerous.

So assessing the state of identification of a model is crucial. In this chapter we present a self-contained treatment of identification in parametric models. Some of the results can also be found in e.g. Fisher (1966), Rothenberg (1971), Bowden (1973), Richmond (1974), and Hsiao (1983, 1987). The pioneering work in the field is due to Haavelmo (1943), which contained the first identification theory for stochastic models to be developed in econometrics; see Aldrich (1994) for an extensive discussion.

The set-up of the chapter is as follows. In Section 2 we introduce the basic concepts of observational equivalence of two parameter points, leading to the definitions of local and global identification. The motivating connection between the notions of identification on the one hand and the existence of a consistent estimator on the other hand is discussed. In Section 3 an important theorem is presented that can be employed to assess the identification of a particular model.

It provides the link between identification and the rank of the information matrix. A further step towards practical usefulness is taken in Section 4, where the information matrix criterion is elaborated and an identification criterion is presented in terms of the rank of a Jacobian matrix. In Section 5 the role played by additional restrictions is considered.

All criteria presented have the practical drawback that they involve the rank evaluation of a matrix whose elements are functions of the parameters. The relevant rank is the rank for the true values of the parameters. These, however, are obviously unknown. Section 6 shows that this is fortunately not a matter of great concern due to considerations of rank constancy.

Up till then, the discussion involved the identification of the whole parameter vector. Now it may happen that the latter is not identified but some individual elements are. How to recognize such a situation is investigated in Section 7. The classical econometric context in which the identification issue figures predominantly is the simultaneous equations model. This issue has become a standard feature of almost every econometric textbook. See, e.g., Pesaran (1987) for a brief overview. In Section 8 we give the relevant theory for the classical simultaneous equations model. Section 9 concludes.

As the title shows, the chapter is restricted to identification in parametric models.<sup>1</sup> It is moreover limited in a number of other respects. It essentially deals with the identification of “traditional” models, i.e. models for observations that are independently identically distributed (iid). Hence, dynamic models, with their different and often quite more complicated identification properties, are not discussed. See, e.g., Deistler and Seifert (1978), Hsiao (1983, 1997), Hannan and Deistler (1988), and Johansen (1995). Also, the models to be considered here are linear in the variables. For a discussion of nonlinear models, which in general have a more favorable identification status, see, e.g., McManus (1992).

We consider identification based on sample information and on exact restrictions on the parameters that may be assumed to hold. We do not pay attention to a Bayesian approach where non-exact restrictions on the parameters in the form of prior distributions are considered. For this approach, see, e.g., Zellner (1971), Drèze (1975), Kadane (1975), Leamer (1978), and Poirier (1998).

As to notation, we employ the following conventions. A superscript 0, as in  $\beta^0$ , indicates the “true” value of a parameter, i.e. its value in the data generating process. When no confusion is possible, however, we may omit this superscript. We use the semicolon in stacking subvectors or submatrices, as the horizontal delimiter of subvectors and submatrices:

$$(A_1; A_2) \equiv (A'_1, A'_2)' = \begin{bmatrix} A_1 \\ A_2 \end{bmatrix}.$$

We will use this notation in particular for  $a_1$  and  $a_2$  being vectors. Covariance matrices are indicated by  $\Sigma$ . When it has a single index,  $\Sigma$  is the variance–covariance matrix of the vector in the subscript. When it has a double subscript,  $\Sigma$  is the matrix of covariances between two random vectors, as indicated by the subscripts.

## 2 BASIC CONCEPTS

Consider a structure  $s$  that describes the probability distribution function  $P_s(y)$  of a random vector  $Y$ . The set of all a priori possible structures is called a model. We assume that  $Y$  is generated by a known parametric probability function  $P(\cdot)$  conditional on a parameter vector  $\theta \in S$ , where  $S$  is an open subset of  $\mathbf{R}^l$ . So a structure is described by a parameter point  $\theta$  and a model is defined by a set  $\{P(y, \theta) | \theta \in S\}$ . Submodels  $\{P(y, \theta) | \theta \in H\}$  are defined by sets of structures  $H$  that are subsets of  $S$ :  $H \subset S$ . Hence, a structure is described by a parametric point  $\theta$ , and a model is a set of points  $H \subset \mathbf{R}^l$ . So the problem of distinguishing between structures is reduced to the problem of distinguishing between parameter points.

**Definition 1.** The sets of structures  $S_1$  and  $S_2$  are *observationally equivalent* if  $\{P(y, \theta) | \theta \in S_1\} = \{P(y, \theta) | \theta \in S_2\}$ . In particular, two parameter points  $\theta_1$  and  $\theta_2$  are observationally equivalent if  $P(y, \theta_1) = P(y, \theta_2)$  for all  $y$ .

**Definition 2.** The element  $\theta_k^0$  of the parameter vector  $\theta^0 \in S$  is said to be *locally identified* in  $S$  if there exists an open neighborhood of  $\theta^0$  containing no point  $\theta \in S$ , with  $\theta_k \neq \theta_k^0$ , that is observationally equivalent to  $\theta^0$ .

The notion of identification is related to the existence of an unbiased or consistent estimator. That is, if  $\theta_k^0$  is locally not identified, there exist points  $\theta$  arbitrarily close to  $\theta^0$  with  $\theta_k^0 \neq \theta_k$  and  $P(y, \theta) = P(y, \theta^0)$ . Hence exact knowledge of  $P(y, \theta^0)$  is not sufficient to distinguish between  $\theta_k^0$  and  $\theta_k$ .<sup>2</sup>

Now consider an estimator  $\hat{\theta}_k$  of  $\theta_k$ . Its distribution function is a function of  $P(y, \theta^0)$  so that, again, exact knowledge of this distribution function is not sufficient to distinguish between  $\theta_k^0$  and  $\theta_k$ . Asymptotically the same holds with respect to the limit distribution of  $\hat{\theta}_k$ . In that case  $\theta_k^0$  cannot be expressed as a function of the small- or large-sample distribution of  $\hat{\theta}_k$ . In particular,  $\theta_k^0$  cannot be expressed as the expectation or probability limit of  $\hat{\theta}_k$ .

On the other hand, if  $\theta_k^0$  is locally identified and if we restrict the parameter space to a sufficiently small open neighborhood of  $\theta^0$ , we find that  $P(y, \theta^0)$  corresponds uniquely to a single value  $\theta_k = \theta_k^0$ . In fact we have the following theorem.

**Theorem 1.** Let  $P(y, \theta)$  be a continuous function of  $\theta \in S$  for all  $y$ , then  $\theta_k^0$  is locally identified (in  $S$ ) if and only if there exists an open neighborhood  $O_{\theta^0}$  of  $\theta^0$  such that any sequence  $\theta^i$ ,  $i = 1, 2, \dots$  in  $S \cap O_{\theta^0}$  for which  $P(y, \theta^i) \rightarrow P(y, \theta^0)$ , for all  $y$ , also satisfies  $\theta_k^i \rightarrow \theta_k^0$ .

**Proof.** The proof is in two parts.

*Necessity.* If  $\theta_k^0$  is locally not identified, then for any open neighborhood  $O_{\theta^0}$  there exists a point  $\theta \in S \cap O_{\theta^0}$  with  $P(y, \theta) = P(y, \theta^0)$  and  $\theta_k \neq \theta_k^0$ . Thus if we take  $\theta^i = \theta$ ,  $i = 1, 2, \dots$  we find  $\theta_k^i = \theta_k \neq \theta_k^0$ .

*Sufficiency.* If for any open neighborhood  $O_{\theta^0}$  there exists a sequence  $\theta^i$ ,  $i = 1, 2, \dots$  in  $S \cap O_{\theta^0}$  for which  $P(y, \theta^i) \rightarrow P(y, \theta^0)$  and  $\theta_k^i$  does not converge to  $\theta_k^0$ , we

may consider converging subsequences in compact neighborhoods with  $\theta_k^i \rightarrow \theta_k^*$   $\neq \theta_k^0$ . Due to the continuity we find that  $P(y, \theta^*) = P(y, \theta^0)$  so that  $\theta_k^0$  is locally not identified. ■

Hence, if  $P(y, \theta^0)$  can be consistently estimated, for example in the case of iid observations, then  $\theta_k^0$ , the  $k$ th element of  $\theta^0 \in \mathcal{O}_{\theta^0}$ , can be consistently estimated if and only if it is identified. Thus, if one considers a sample as a single observation on a random vector with probability distribution  $P(y, \theta^0)$  and uses an asymptotic parameter sequence consisting of repeated samples, i.e. iid observations on this random vector, then  $\theta_k^0$  can be consistently estimated if and only if it is identified.<sup>3</sup>

So far for the identification of a single parameter. Definition 2 can be extended to the definition of the whole parameter vector straightforwardly.

**Definition 3.** If all elements of  $\theta^0$  are locally identified then  $\theta^0$  is said to be locally identified.

Although the notion of *local* identification plays the predominant role, we will occasionally refer to *global* identification.

**Definition 4.** If the open neighborhood referred to in definition 2 is equal to  $\mathcal{S}$ , then the identification is said to be *global* in  $\mathcal{S}$ .

Definitions 2 and 3 are obviously difficult to apply in practice. In the following section we present a much more manageable tool for the characterization of local identification.

### 3 IDENTIFICATION AND THE RANK OF THE INFORMATION MATRIX

When analyzing local identification of a model, the information matrix can be used conveniently. The following theorem, due to Rothenberg (1971) but with a slightly adapted proof, contains the essential result.

**Definition 5.** Let  $M(\theta)$  be a continuous matrix function of  $\theta \in \mathbf{R}^l$  and let  $\theta^0 \in \mathbf{R}^l$ . Then  $\theta^0$  is a *regular point* of  $M(\theta)$  if the rank of  $M(\theta)$  is constant for points in  $\mathbf{R}^l$  in an open neighborhood of  $\theta^0$ .

**Theorem 2.** Let  $\theta^0$  be a regular point of the information matrix  $\Psi(\theta)$ . Assume that the distribution of  $y$  has a density function  $f(y, \theta)$ , and assume that  $f(y, \theta)$  and  $\log f(y, \theta)$  are continuously differentiable in  $\theta$  for all  $\theta \in \mathcal{S}$  and for all  $y$ . Then  $\theta^0$  is locally identified if and only if  $\Psi(\theta^0)$  is nonsingular.

**Proof.** Let

$$g(y, \theta) \equiv \log f(y, \theta)$$

$$h(y, \theta) \equiv \partial \log f(y, \theta) / \partial \theta.$$

Then the mean value theorem implies

$$g(y, \theta) - g(y, \theta^0) = h(y, \theta^*)'(\theta - \theta^0), \quad (7.1)$$

for all  $\theta$  in a neighborhood of  $\theta^0$ , for all  $y$ , and with  $\theta^*$  between  $\theta$  and  $\theta^0$  (although  $\theta^*$  may depend on  $y$ ). Now suppose that  $\theta^0$  is *not* locally identified. Then any open neighborhood of  $\theta^0$  will contain parameter points that are observationally equivalent to  $\theta^0$ . Hence we can construct an infinite sequence  $\theta^1, \theta^2, \dots, \theta^k, \dots$ , such that  $\lim_{k \rightarrow \infty} \theta^k = \theta^0$ , with the property that  $g(y, \theta^k) = g(y, \theta^0)$ , for all  $k$  and all  $y$ . It then follows from (7.1) that for all  $k$  and all  $y$  there exist points  $\theta^{*k}$  (which again may depend on  $y$ ), such that

$$h(y, \theta^{*k})'\delta^k \equiv h(y, \theta^{*k})'(\theta^k - \theta^0)/\|\theta^k - \theta^0\| = 0, \quad (7.2)$$

with  $\theta^{*k}$  between  $\theta^k$  and  $\theta^0$ .

Since  $\theta^k \rightarrow \theta^0$ , there holds  $\theta^{*k} \rightarrow \theta^0$  for all  $y$ . Furthermore, the sequence  $\delta^1, \delta^2, \dots, \delta^k, \dots$  is an infinite sequence on the unit sphere, so there must be at least one limit point. Let  $\delta^0$  be such a limit point. Then (7.2) implies  $h(y, \theta^0)'\delta^0 = 0$  for all  $y$ . This gives for the information matrix

$$E\{(h(y, \theta^0)'\delta^0)^2\} = \delta^0'E\{h(y, \theta^0)h(y, \theta^0)'\}\delta^0 = \delta^0'\Psi(\theta^0)\delta^0 = 0,$$

so that indeed nonidentification of  $\theta^0$  implies singularity of the information matrix.

Conversely, if  $\theta^0$  is a regular point but  $\Psi(\theta^0)$  is singular, then there exists a vector  $c(\theta)$  such that in an open neighborhood of  $\theta^0$

$$c(\theta)'\Psi(\theta)c(\theta) = E\{(h(y, \theta)'c(\theta))^2\} = 0.$$

This implies, for all  $\theta$  in this neighborhood, that  $h(y, \theta)'c(\theta) = 0$  for all  $y$ . Since  $\Psi(\theta)$  is continuous and of constant rank,  $c(\theta)$  can be chosen to be continuous in a neighborhood of  $\theta^0$ . We use this property to define a curve  $\theta(t)$  which solves for  $0 \leq t \leq t_*$  the differential equation  $d\theta(t)/dt = c(\theta)$ ,  $\theta(0) = \theta^0$ . This gives

$$\frac{dg(y, \theta)}{dt} = h(y, \theta)'\frac{d\theta}{dt} = h(y, \theta)'c(\theta) = 0$$

for all  $y$ . So  $g(y, \theta)$  is constant along the curve for  $0 \leq t \leq t_*$ , hence  $\theta^0$  is not identified. ■

#### 4 THE JACOBIAN MATRIX CRITERION

The advantage of Theorem 2 is that we do not have to operate on the joint probability distribution of the underlying random variables directly when analyzing the identification of a particular model. It suffices to consider the information matrix. There is a further simplification possible when the underlying

distribution admits a sufficient statistic (of a size that does not depend on the sample size). Under suitable regularity conditions, which are essentially those for the Cramér–Rao theorem, such a sufficient statistic exists if and only if the distribution belongs to the exponential family. Then, its density function can be written as

$$f(y, \theta) = a(y)e^{b(y)\tau(\theta)+c(\theta)},$$

for a suitable choice of functions  $a(\cdot)$ ,  $b(\cdot)$ ,  $c(\cdot)$  and  $\tau(\cdot)$ , where  $\tau(\cdot)$  and  $b(\cdot)$  are vector functions. Without loss of generality we assume that the covariance matrix of  $b(y)$  is nonsingular.

When  $y_1, \dots, y_n$  denote the vectors of observations, a sufficient statistic is given by  $s(y) \equiv \sum_{i=1}^m b(y_i)$ , as follows from the factorization theorem for jointly sufficient statistics. The first-order derivative of the loglikelihood is

$$\frac{\partial \log l(\theta)}{\partial \theta} = Q(\theta)s(y) + n \frac{\partial c(\theta)}{\partial \theta},$$

where

$$Q(\theta) \equiv \frac{\partial \tau(\theta)'}{\partial \theta}. \quad (7.3)$$

Since  $E\{\partial \log l(\theta)/\partial \theta\} = 0$ , the information matrix is given by the variance of the derivative of the loglikelihood:

$$\Psi(\theta) = Q(\theta)\Sigma_{s(y)}Q(\theta)'.$$

Since  $\Sigma_{s(y)}$  is of full rank, the information matrix is of full rank if and only if  $Q(\theta)$  is of full row rank. So we have established the following result.

**Theorem 3.** Let  $f(y, \theta)$  belong to the exponential family. Let  $\theta^0$  be a regular point of the information matrix  $\Psi(\theta)$ . Let  $Q(\cdot)$  be as defined in (7.3). Then  $\theta^0$  is locally identified if and only if  $Q(\theta^0)$  has full row rank.

As a byproduct of this theorem, we note that  $\tau(\theta)$  (sometimes called the canonical or natural parameter) is identified since the corresponding information matrix is simply the covariance matrix of  $s(y)$ , assumed to be of full rank.

A major application of Theorem 3 concerns the  $k$ -dimensional normal distribution with parameters  $\mu$  and  $\Sigma$  whose elements are functions of a parameter vector  $\theta$ . For the normal we can write

$$\begin{aligned} f(y, \theta) &= (2\pi)^{-k/2} |\Sigma|^{-\frac{1}{2}} e^{-\frac{1}{2}(y-\mu)'\Sigma^{-1}(y-\mu)} \\ &= (2\pi)^{-k/2} e^{(y', -\frac{1}{2}y' \otimes y')(\Sigma^{-1}\mu, \text{vec}\Sigma^{-1}) + \log|\Sigma|^{-1/2} - \frac{1}{2}\mu'\Sigma^{-1}\mu}. \end{aligned}$$

This gives the normal distribution in the form of the exponential family but the term  $y \otimes y$  contains redundant elements and hence its covariance matrix is singular. To eliminate this singularity, let  $N_k$  (of order  $k^2 \times k^2$ ),  $D_k$  (of order  $k^2 \times \frac{1}{2}k(k+1)$ ), and  $L_k$  (of order  $\frac{1}{2}k(k+1) \times k^2$ ) be matrices with properties

$$N_k \text{vec} A = \text{vec} \frac{1}{2}(A + A'), \quad D_k v(A) = \text{vec} A, \quad L'_k v(B) = \text{vec} B$$

for every  $k \times k$ -matrix  $A$  and for every lower triangular  $k \times k$ -matrix  $B$ , and the  $\frac{1}{2}k(k+1)$ -vector  $v(A)$  is the vector obtained from  $\text{vec} A$  by eliminating all supradiagonal elements of  $A$ . Then  $D_k L_k N_k = N_k$  (Magnus, 1988, theorem 5.5(ii)), and

$$\begin{aligned} (y \otimes y)' \text{vec} \Sigma^{-1} &= (y \otimes y)' N_k \text{vec} \Sigma^{-1} \\ &= (y \otimes y)' N_k L'_k D'_k \text{vec} \Sigma^{-1} \\ &= (y \otimes y)' L'_k D'_k \text{vec} \Sigma^{-1} \\ &= (L_k(y \otimes y))' D'_k \text{vec} \Sigma^{-1}. \end{aligned}$$

So the normal density fits in the  $k$ -dimensional exponential family with

$$b(y) = (y; -\frac{1}{2}L_k(y \otimes y)), \quad \tau(\theta) = (\Sigma^{-1}\mu; D'_k \text{vec} \Sigma^{-1}).$$

The identification of a normality-based model with parameterized mean and variance hence depends on the column rank of the matrix of derivatives of  $\tau(\theta)$  with respect to  $\theta'$ , or equivalently on the column rank of the matrix of derivatives of

$$\sigma(\theta) \equiv (\mu; v(\Sigma)). \quad (7.4)$$

The equivalence is due to the fact that the Jacobian matrix of the transformation from  $\tau(\theta)$  to  $\sigma(\theta)$  is equal to

$$\begin{bmatrix} \Sigma^{-1} & -(\mu' \Sigma^{-1} \otimes \Sigma^{-1}) D_k \\ 0 & -D'_k (\Sigma^{-1} \otimes \Sigma^{-1}) D_k \end{bmatrix}$$

and is hence nonsingular. If, as is often the case in practice, the mean of the distribution is zero, (7.4) reduces to  $\sigma(\theta) = v(\Sigma)$ . So, the identification of a model when the underlying distribution is multivariate normal with zero means depends on the structure of the covariance matrix only.

While frequently normality is an assumption of convenience, not justified by theory or data, it should be stressed that, when dealing with problems of identification, it is also a conservative assumption (cf. Aigner *et al.*, 1984; Bekker, 1986). When the underlying distribution is nonnormal, models that are not identified under normality may as yet be identified since higher order moments can then be added to  $\sigma(\theta)$ , which either leaves the rank of the Jacobian matrix unaffected or increases it.

For the remainder of this chapter we do not make a specific assumption as to the form of the distribution. We merely assume in generality that there exists a vector function  $\sigma$  of order  $n$ ,  $\sigma(\theta) : \mathbf{R}^l \rightarrow \mathbf{R}^n$ , such that there exists a one-to-one relation between the elements of  $\{P(y, \theta) \mid \theta \in \mathbf{R}^l\}$  and  $\{\sigma(\theta) \mid \theta \in \mathbf{R}^l\}$ ; the identification of the parameters from the underlying distribution is “transmitted” through  $\sigma(\theta)$ .

## 5 PRIOR INFORMATION

The information on the parameters can come from two sources. One is from the observations, which are informative about  $\sigma(\theta)$ . But we may also have a priori information. Let this be of the following form:  $\theta^0$  satisfies  $p(\theta^0) = 0$ , where  $p(\cdot)$  is an  $r$ -dimensional vector function. Therefore we study the submodel  $\mathcal{H} \subset S$ , where

$$\mathcal{H} = \{\theta \mid \theta \in S \subset \mathbf{R}^l, p(\theta) = 0\}. \quad (7.5)$$

Now the problem of local identification of the parameter vector  $\theta^0$  reduces to the problem of verifying whether the equations system

$$f(\theta) = \begin{bmatrix} \sigma(\theta) \\ p(\theta) \end{bmatrix} - \begin{bmatrix} \sigma(\theta^0) \\ p(\theta^0) \end{bmatrix} = 0 \quad (7.6)$$

has a locally unique solution: if we can find parameter points  $\theta$  arbitrarily close to  $\theta^0$  for which  $f(\theta) = 0$ , then  $\theta^0$  is locally not identified. In order to find out whether  $\theta^0$  is locally identified we will apply the implicit function theorem. We will now first discuss this theorem.

Let  $C'$  be the class of functions that are  $r$  times continuously differentiable in  $\theta^0$ . Furthermore a  $C^\infty$  function is analytic if the Taylor series expansion of each component converges to that function. In practice most functions are analytic. They share the important quality of being either equal to zero identically or equal to zero only on a set of Lebesgue measure zero. We can now formulate the implicit function theorem.

**Theorem 4.** *Implicit function theorem.* Let  $f(\theta) = f(\theta_1; \theta_{II})$  be a  $C^p$  function ( $p \geq 1$ ),  $f : \mathbf{R}^{n+m} \rightarrow \mathbf{R}^n$ ;  $\theta_1 \in \mathbf{R}^n$ ,  $\theta_{II} \in \mathbf{R}^m$  such that  $f(\theta_1^0; \theta_{II}^0) = 0$  and the  $n \times n$ -matrix  $\partial f(\theta)/\partial \theta'_1|_{\theta^0}$  is nonsingular, then there exists an open neighborhood  $U \subset \mathbf{R}^n$  of  $\theta_1^0$  and an open neighborhood  $V \subset \mathbf{R}^m$  of  $\theta_{II}^0$  such that there exists a unique  $C^p$  function  $g : V \rightarrow U$  with  $g(\theta_{II}^0) = \theta_1^0$  and  $f(g(\theta_{II}); \theta_{II}) = 0$  for all  $\theta_{II} \in V$ .

If indeed  $\partial f(\theta)/\partial \theta'_1|_{\theta^0}$  is nonsingular, so that  $f(g(\theta_{II}); \theta_{II}) = 0$  for all  $\theta_{II} \in V$ , then we also have for points  $(\theta_1; \theta_{II}) = (g(\theta_{II}); \theta_{II})$  with  $\theta_{II} \in V$ :

$$\frac{\partial f}{\partial \theta'_{II}} + \frac{\partial f}{\partial g'} \frac{\partial g}{\partial \theta'_{II}} = 0.$$

Returning to the question of identification of  $\theta^0$  a first important result is easily derived if we assume that  $f(\theta)$ , as defined in (7.6), is continuously differentiable in  $\theta^0$ . Let the Jacobian matrix  $J(\theta)$  be

$$J(\theta) = \frac{\partial f(\theta)}{\partial \theta'}.$$

Now assume that  $J(\theta^0)$  has full column rank. Then, possibly after a rearrangement of elements of  $f(\theta) = (f_1(\theta); f_2(\theta))$ , we have

$$J(\theta) = \begin{bmatrix} J_1(\theta) \\ J_2(\theta) \end{bmatrix} = \begin{bmatrix} \partial f_1(\theta)/\partial \theta' \\ \partial f_2(\theta)/\partial \theta' \end{bmatrix}, \quad (7.7)$$

where  $J_2(\theta^0)$  is nonsingular. If we consider the function  $h(\theta; v) \equiv f_2(\theta) + 0 \cdot v$ , we find, by using the implicit function theorem, that the function  $g(v) = \theta^0$ , for which  $h(g(v); v) = 0$ , is unique. Consequently, there is an open neighborhood of  $\theta^0$  where  $\theta = \theta^0$  is the only solution of  $f_2(\theta) = 0$ , or  $f(\theta) = 0$ . Thus, if  $J(\theta^0)$  has full column rank, then  $\theta^0$  is locally identified.

Now assume that  $J(\theta^0)$  is not of full column rank. Does this imply that  $\theta^0$  is not locally identified? The answer is negative. For example, consider the special case where  $f(\theta) = \theta_1^2 + \theta_2^2$ , and  $\theta^0 = (0, 0)$ . Here the Jacobian matrix  $J(\theta)$  is given by  $(2\theta_1, 2\theta_2)$  and  $J(\theta^0) = (0, 0)$ , which is clearly not of full column rank. However,  $\theta^0$  is the only point in  $\mathbf{R}^2$  for which  $f(\theta) = 0$ , so that  $\theta^0$  is, in fact, globally identified.

This may seem a pathological case since the (row) rank of the Jacobian matrix simply was 0. What can we say if  $J(\theta^0)$  is not of full column rank while it is of full row rank? In that case  $\theta^0$  will not be identified as a direct consequence of the implicit function theorem. And so we are left with the general case where  $J(\theta^0)$  has both a deficient column rank and a deficient row rank. Without loss of generality we can rearrange the rows of  $J(\theta)$  as in (7.7) where  $\text{rank}\{J(\theta^0)\} = \text{rank}\{J_2(\theta^0)\}$  while now  $J_2(\theta^0)$  has full row rank. According to the implicit function theorem we can rearrange the elements of  $\theta = (\theta_I; \theta_{II})$  so that locally there exists a unique function  $g(\theta_{II})$  so that  $f_2(g(\theta_{II}); \theta_{II}) = 0$ . However,  $f_1(g(\theta_{II}); \theta_{II}) = 0$  does not necessarily hold, and we have not yet established a lack of identification.

Let  $h(\theta_{II}) = f(g(\theta_{II}); \theta_{II})$ . Then we have locally

$$\begin{aligned} \frac{\partial h(\theta_{II})}{\partial \theta'_{II}} &= \frac{\partial f}{\partial \theta'_{II}} + \frac{\partial f}{\partial g} \frac{\partial g}{\partial \theta'_{II}} \\ &= \begin{bmatrix} J_1(g(\theta_{II}); \theta_{II}) \\ J_2(g(\theta_{II}); \theta_{II}) \end{bmatrix} \begin{bmatrix} \partial g / \partial \theta'_{II} \\ I_k \end{bmatrix} \\ &= \begin{bmatrix} J_1(g(\theta_{II}); \theta_{II}) \begin{bmatrix} \partial g / \partial \theta'_{II} \\ I_k \end{bmatrix} \\ 0 \end{bmatrix}, \end{aligned}$$

where  $k \equiv l - \text{rank}\{J(\theta^0)\}$ . If it can be established that the rows of  $J_1(g(\theta_{II}); \theta_{II})$  are linearly dependent on the rows of  $J_2(g(\theta_{II}); \theta_{II})$  for all  $\theta_{II}$  in an open neighborhood of  $\theta_{II}^0 \in \mathbf{R}^k$ , then  $\partial h(\theta_{II})/\partial\theta_{II}' = 0$  in that neighborhood, which can be taken to be convex, so that  $f(g(\theta_{II}); \theta_{II}) = h(\theta_{II}) = h(\theta_{II}^0) = 0$  in an open neighborhood of  $\theta_{II}^0$ . As a result  $\theta^0$  will not be locally identified.

Notice that  $\theta^0$  is a regular point of  $J(\theta)$  if  $J(\theta^0)$  has full row (or column) rank since in that case  $|J(\theta^0)J(\theta^0)'| \neq 0$  (or  $|J(\theta^0)'J(\theta^0)| \neq 0$ ) so that  $J(\theta)$  will have full row (or column) rank for points close enough to  $\theta^0$ . If  $\theta^0$  is a regular point of  $J(\theta)$  and  $\text{rank}\{J(\theta^0)\} = \text{rank}\{J_2(\theta^0)\}$ , where  $J(\theta)$  has been partitioned as in (7.7), and  $J_2(\theta^0)$  is of full row rank, then  $\theta^0$  is a regular point of both  $J(\theta)$  and  $J_2(\theta)$ . So  $\text{rank}\{J(\theta)\} = \text{rank}\{J_2(\theta)\}$  in an open neighborhood of  $\theta^0$ . In other words, for all points in an open neighborhood of  $\theta^0$  the rows of  $J_1(\theta)$  will depend linearly on the rows of  $J_2(\theta)$ .

Summarizing these results we may give the following theorem, which is also given by Fisher (1966, theorem 5.9.2).

**Theorem 5.** Let  $J(\theta)$  be the Jacobian matrix of order  $(n + r) \times l$  formed by taking partial derivatives of  $(\sigma(\theta); \rho(\theta))$  with respect to  $\theta$ ,

$$J(\theta) \equiv \begin{bmatrix} \partial\sigma(\theta)/\partial\theta' \\ \partial\rho(\theta)/\partial\theta' \end{bmatrix}.$$

If  $\theta^0$  is a regular point of  $J(\theta)$ , then a necessary and sufficient condition for  $\theta^0$  to be locally identified is that  $J(\theta)$  has rank  $l$  at  $\theta^0$ .

## 6 HANDLING THE RANK IN PRACTICE

The question is how one should apply this theorem in practice. Since  $\theta^0$  is an unknown parameter vector it is not yet clear how one should compute the rank of  $J(\theta^0)$ . Furthermore, a question can be raised as to the restrictiveness of the assumption of regularity of  $\theta^0$ . It appears that these problems are related. That is, if  $\theta^0$  is a regular point of  $J(\theta)$ , then the rank of  $J(\theta^0)$  can be computed even though  $\theta^0$  itself is unknown. On the other hand, the assumption of regularity of  $\theta^0$ , as it is stated in Theorem 5, is unnecessarily restrictive.

The assumption of regularity of  $\theta^0$  makes it possible to compute the rank of  $J(\theta^0)$ . In the sequel we assume that  $f(\theta)$  is an analytic function, in which case almost all points  $\theta \in \mathcal{S}$  are regular points: the irregular points constitute a set in  $\mathbf{R}^l$  of Lebesgue measure zero. The proof of this statement is as follows. Let

$$\tilde{\theta} \equiv \underset{\theta \in \mathbf{R}^l}{\arg\max} \{\text{rank}\{J(\theta)\}\},$$

and let

$$J(\tilde{\theta}) \equiv \begin{bmatrix} J_1(\tilde{\theta}) \\ J_2(\tilde{\theta}) \end{bmatrix}'$$

so that  $\text{rank}\{J(\tilde{\theta})\} = \text{rank}\{J_2(\tilde{\theta})\}$  and  $J_2(\tilde{\theta})$  is of full row rank. Then, due to continuity,  $|J_2(\theta)J_2(\theta)'| > 0$  in an open neighborhood of  $\tilde{\theta}$ . That is, the analytic function  $|J_2(\theta)J_2(\theta)'|$  is not equal to zero on a set of positive Lebesgue measure. Hence  $|J_2(\theta)J_2(\theta)'| > 0$  almost everywhere, since an analytic function is either identical to zero or equal to zero on a set of Lebesgue measure zero only. Thus  $J_2(\theta)$  will have full row rank almost everywhere in  $S$  so that  $\text{rank}\{J(\theta)\} = \max_{\theta \in R^l} \{\text{rank}\{J(\theta)\}\}$  almost everywhere. Consequently, if  $\theta^0$  is a regular point of  $J(\theta)$ , then  $\text{rank}\{J(\theta)\} = \text{rank}\{J(\theta^0)\}$  for points  $\theta$  in a neighborhood with positive Lebesgue measure. Hence,  $\text{rank}\{J(\theta^0)\} = \max_{\theta \in R^l} \{\text{rank}\{J(\theta)\}\}$ , which can be computed without knowing  $\theta^0$ . We thus have shown the following result, see also Johansen (1995, theorem 2).

**Theorem 6.** Let  $\sigma(\cdot)$  and  $\rho(\cdot)$  be analytic functions. Then  $\theta^0$  is a regular point of  $J(\theta)$  if and only if

$$\text{rank}\{J(\theta^0)\} = \max_{\theta \in R^l} \{\text{rank}\{J(\theta)\}\},$$

which holds for almost all  $\theta^0 \in R^l$ .

Thus if we know nothing about  $\theta^0$  except that  $\theta^0 \in S \subset R^l$ , where  $S$  is some open set in  $R^l$ , then it would make sense to assume that  $\theta^0$  is a regular point, since almost all points of  $S$  are regular points. However, when doing so we would ignore the prior information contained in  $\rho(\theta^0) = 0$ . The set  $\mathcal{H}$  (cf. (7.5)) constitutes a manifold of dimension  $l - r$  in  $R^l$ , where we assume that  $\partial\rho(\theta)/\partial\theta'$  has full row rank in  $\theta^0$ . Let

$$R(\theta) \equiv \frac{\partial\rho(\theta)}{\partial\theta'},$$

then  $\text{rank}\{R(\theta^0)\} = r$ . The set  $\mathcal{H}$  is locally homeomorphic to an open set in  $R^{l-r} : r$  elements of  $\theta$ , collected in the vector  $\theta_I$ , say, are locally unique functions of the remaining elements collected in  $\theta_{II}$ :  $\theta_I = g(\theta_{II})$ , say. Without loss of generality we assume that  $\theta_I$  constitutes the first  $r$  elements of  $\theta$ . Now it does make sense to assume that  $\theta_{II}^0$  is a regular point of  $J(g(\theta_{II}))$  since almost all points in  $R^{l-r}$  are regular points and we do not have any further relevant prior information with respect to  $\theta_{II}^0$ . This assumption is less restrictive than the assumption in Theorem 5 since the constancy of  $\text{rank}\{J(\theta)\}$  in an open neighborhood of  $\theta^0$  implies the constancy of this rank for points close to  $\theta^0$  that satisfy  $\theta_I = g(\theta_{II})$ . Moreover, the rank of  $J(\theta^0)$  should be computed by  $\text{rank}\{J(\theta^0)\} = \max_{\theta \in \mathcal{H}} \{\text{rank}\{J(\theta)\}\}$ , which may be less than the maximum over  $\theta \in R^l$ . To formalize this, we generalize Definition 5 and generalize Theorem 5.

**Definition 6.** Let  $M(\theta)$  be a continuous matrix function and let  $\theta^0 \in \mathcal{H} \subset R^l$ . Then  $\theta^0$  is a *regular point* of  $M(\theta) | \mathcal{H}$  if the rank of  $M(\theta)$  is constant for points in  $\mathcal{H}$  in an open neighborhood of  $\theta^0$ .

**Theorem 7.** Let  $J(\theta)$  be the Jacobian matrix of order  $(n + r) \times l$  formed by taking the partial derivatives of  $(\sigma(\theta); \rho(\theta))$  with respect to  $\theta \in S \subset \mathbf{R}^l$ :

$$J(\theta) = \begin{bmatrix} \partial\sigma(\theta)/\partial\theta' \\ R(\theta) \end{bmatrix}. \quad (7.8)$$

Let  $\text{rank}\{R(\theta^0)\} = r$ . If  $\theta^0$  is a regular point of  $J(\theta) | \mathcal{H}$ , then  $\theta^0$  is locally identified in  $\mathcal{H}$  if and only if  $\max_{\theta \in \mathcal{H}}\{\text{rank}\{J(\theta)\}\} = l$ .

**Proof.** Since  $\theta^0$  is a regular point of  $J(\theta) | \mathcal{H}$ , the rank of  $J(\theta^0)$  is equal to  $\max_{\theta \in \mathcal{H}}\{\text{rank}\{J(\theta)\}\}$ . So if the latter rank is equal to  $l$ , then  $J(\theta^0)$  has full column rank and  $\theta^0$  is identified. If  $\text{rank}\{J(\theta^0)\} < l$ , then there is a partitioning as in (7.7), where the elements of  $\rho(\theta)$  are elements of  $f_2(\theta)$  and  $J_2(\theta^0)$  has full row rank and a deficient column rank. Hence we may apply the implicit function theorem to show that there exist points  $\theta = (g(\theta_{II}); \theta_{II})$  arbitrarily close to  $\theta^0$  so that  $f_2(\theta) = f_2(\theta^0)$ . Since the elements of  $\rho(\theta)$  are also elements of  $f_2(\theta)$ , these points  $\theta$  satisfy  $\rho(\theta) = \rho(\theta^0)$ , so that they are elements of  $\mathcal{H}$ .

Since  $\theta^0$  is a regular point of  $J(\theta) | \mathcal{H}$  and the points  $\theta = (g(\theta_{II}); \theta_{II})$  are located in  $\mathcal{H}$ ,  $\theta_{II}^0$  is a regular point of  $J(g(\theta_{II}); \theta_{II})$ , so that close to  $\theta_{II}^0$  the rows of  $J_1(g(\theta_{II}); \theta_{II})$  are linearly dependent on the rows of  $J_2(g(\theta_{II}); \theta_{II})$  and so, by the same argument as in the proof of Theorem 5,  $f_1(g(\theta_{II}); \theta_{II}) = f_1(\theta^0)$ . As a result,  $\theta^0$  will not be locally identified in  $\mathcal{H}$ . ■

It should be noted that, if  $\sigma(\theta)$  and  $\rho(\theta)$  are linear functions, so that  $J(\theta) = J$  does not depend on  $\theta$  and  $f(\theta) - f(\theta^0) = J \cdot (\theta - \theta^0)$ , then  $\theta^0$  is globally identified if and only if  $J(\theta^0) = J$  is of full column rank.

## 7 PARTIAL IDENTIFICATION

Up till now we have been concerned with identification of  $\theta^0$  as a whole. However, if  $\theta^0$  is locally not identified, it may still be the case that some separate elements of  $\theta^0$  are identified. If, for example, the  $i$ th element of  $\theta^0$  is identified, this means that for a point  $\theta$  observationally equivalent to  $\theta^0$  there holds  $\theta_i = \theta_i^0$  if  $\theta$  is close enough to  $\theta^0$ .

Insight into such partial identification could be relevant for estimation purposes, if one is interested in estimating the single parameter  $\theta_i^0$ . It could also be relevant for the purpose of model specification in the sense that it may suggest a further restriction of the parameter space such that  $\theta^0$  is identified as yet.

In order to describe conditions for the local identification of single parameters we denote by

$$\mathcal{A}_{(i)} \equiv \{\theta \mid \theta \in \mathcal{H}, \theta_i = \theta_i^0\}$$

the restricted parameter space whose elements satisfy the a priori restrictions  $\rho(\theta) = 0$  and  $\theta_i = \theta_i^0$ , cf. (7.5). Furthermore, let  $J_{(i)}(\theta)$  consist of the columns of  $J(\theta)$

except the  $i$ th one, so that derivatives have been taken with respect to all elements of  $\theta$  apart from the  $i$ th one. The main results on the identification of single parameters are contained in the following two theorems.

**Theorem 8.** Let  $\mathcal{A}_{(i)} \equiv \{\theta \mid \theta \in \mathcal{H}, \theta_i = \theta_i^0\}$ . If  $\theta^0$  is a regular point of  $J(\theta) \mid \mathcal{H}$  and  $\text{rank}\{J_{(i)}(\theta^0)\} = \text{rank}\{J(\theta^0)\}$ , then  $\theta_i^0$  is locally *not* identified in  $\mathcal{H}$ .

**Proof.** Let the rows of  $J(\theta)$  be rearranged and partitioned as in (7.7) so that  $J_2(\theta^0)$  is of full row rank equal to the rank of  $J(\theta^0)$ . Furthermore let the columns be rearranged so that

$$J_2(\theta) = (J_{21}(\theta), J_{22}(\theta)) = (\partial f_2 / \partial \theta'_I, \partial f_2 / \partial \theta'_{II}),$$

where  $J_{21}(\theta^0)$  is nonsingular. Now, if  $\text{rank}\{J(\theta^0)\} = \text{rank}\{J_{(i)}(\theta^0)\}$ , then  $\theta_{II}$  can be taken such that  $\theta_i$  is an element of  $\theta_{II}$ . Application of the implicit function theorem shows that there is a unique function  $\theta_I = g(\theta_{II})$  so that  $f_2(g(\theta_{II}); \theta_{II}) = 0$  for all  $\theta_{II}$  in an open neighborhood of  $\theta_{II}^0$ . Since the elements of  $\rho(\theta)$  are elements of  $f_2(\theta)$ , the points  $(g(\theta_{II}); \theta_{II})$  are located in  $\mathcal{H}$ . Furthermore, since  $\theta^0$  is a regular point of  $J(\theta) \mid \mathcal{H}$  it follows, by the same argument as in the proof of Theorem 5, that also  $f_1(g(\theta_{II}), \theta_{II}) = 0$ . Hence all elements of  $\theta_{II}$ , including  $\theta_i$ , are locally not identified in  $\mathcal{H}$ . ■

We assume that  $\text{rank}\{\partial \rho(\theta) / \partial \theta' \mid \theta^0\} = \text{rank}\{R(\theta^0)\} = \text{rank}\{R_{(i)}(\theta^0)\} = r$  so that  $\theta_i^0$  is locally not identified by the a priori restrictions  $\rho(\theta^0) = 0$  alone. In other words,  $\mathcal{H} = \{\theta \mid \theta \in S, \rho(\theta) = 0\}$  constitutes a manifold of dimension  $l - r$  in  $\mathbb{R}^l$  and  $\mathcal{A}_{(i)} = \{\theta \in \mathcal{H}, \theta_i = \theta_i^0\}$  constitutes a manifold of dimension  $l - r - 1$  in  $\mathbb{R}^l$ .

**Theorem 9.** Let  $\mathcal{A}_{(i)} \equiv \{\theta \mid \theta \in \mathcal{H}, \theta_i = \theta_i^0\}$ . If  $\theta^0$  is a regular point of  $J_{(i)}(\theta) \mid \mathcal{A}_{(i)}$  and  $\text{rank}\{J_{(i)}(\theta^0)\} < \text{rank}\{J(\theta^0)\}$ , then  $\theta_i^0$  is locally identified in  $\mathcal{H}$ .

**Proof.** Applying the same rearrangement and partitioning of  $J(\theta)$  as in the proof of Theorem 8, we find that  $\theta_i$  is an element of  $\theta_I$ . Again there is a unique function so that  $f_2(g(\theta_{II}), \theta_{II}) = 0$  for  $\theta_{II}$  close to  $\theta_{II}^0$ . Now, if we let  $\theta_i = \theta_i^0$  and differentiate  $f_2$  with respect to the remaining elements of  $\theta$ , we get the Jacobian matrix  $J_{2(i)}(\theta)$ . Since  $\text{rank}\{J_{(i)}(\theta^0)\} < \text{rank}\{J(\theta^0)\}$ , the matrix  $J_{2(i)}(\theta^0)$  is not of full row rank and the rank of  $\{J_{2(i)}(\theta^0)\}$  equals the rank of  $\{J_{2(i)}(\theta^0)\}$ . So if  $\theta^0$  is a regular point of  $J_{(i)}(\theta) \mid \mathcal{A}_{(i)}$  it is also a regular point of  $J_{2(i)}(\theta) \mid \mathcal{A}_{(i)}$ . So we may use the same argument as in the proof of Theorem 8, where we assume that  $\text{rank}\{R_{(i)}(\theta^0)\} = r$ , to verify that if  $\theta_i = \theta_i^0$ , then there exist unique functions  $\theta_j = h_j(\theta_{II})$ ,  $j \neq i$ , so that, if we write  $h_i(\theta_{II}) \equiv \theta_i^0$ ,  $f_2(h(\theta_{II}), \theta_{II}) = 0$  for  $\theta_{II}$  close to  $\theta_{II}^0$ . However, we already verified that  $g(\theta_{II})$  is a unique function, so  $g_i(\theta_{II}) = \theta_i^0$ . Consequently  $\theta_i^0$  is locally identified. ■

Again the ranks that occur in these theorems can be evaluated by computing the maximum rank over the relevant parameter space (or manifold), which is  $\mathcal{H}$  in Theorem 8 and  $\mathcal{A}_{(i)}$  in Theorem 9.

The importance of these two theorems in practice is best brought out by the following reformulation.

**Corollary 1.** Let the regularity assumptions of Theorems 8 and 9 be satisfied (which holds for almost all  $\theta^0 \in \mathcal{H}$ ). Let  $N(\theta^0)$  be a basis for the null-space of  $J(\theta^0)$ , i.e.  $J(\theta^0)N(\theta^0) = 0$  and let  $e_i$  be the  $i$ th unit vector. Then  $\theta_i^0$  is locally identified if and only if  $e_i'N(\theta^0) = 0$ .

So a zero-row in a null-space indicates an identified parameter.

## 8 THE CLASSICAL SIMULTANEOUS EQUATIONS MODEL

We now turn to identification in the classical simultaneous equations model. Estimation in this model is comprehensively reviewed in chapter 6 by Mariano. The model is

$$By + \Gamma x = \zeta, \quad (7.9)$$

where  $y$  is an  $m$ -vector of endogenous variables,  $x$  is a stochastic  $k$ -vector of exogenous variables, and  $\zeta$  is an  $m$ -vector of disturbances. It is assumed that  $E(\zeta) = 0$  and  $E(x\zeta') = 0$ , so that  $x$  and  $\zeta$  are uncorrelated, and  $\Sigma_x \equiv E(xx')$  is nonsingular. The coefficient matrices  $B$  and  $\Gamma$  are of order  $m \times m$  and  $m \times k$ , respectively, and  $B$  is nonsingular.

Under normality, the distribution of the observations  $(y; x)$  is uniquely determined by the first two moments. Since model (7.9) can be rewritten as

$$(y', x') = (\zeta', x') \begin{bmatrix} B' & 0 \\ \Gamma' & I_k \end{bmatrix}^{-1},$$

the moment equations are given by

$$(\mu_y', \mu_x') \equiv E(y', x') = (0', \mu_x') \begin{bmatrix} B' & 0 \\ \Gamma' & I_k \end{bmatrix}^{-1} \quad (7.10)$$

$$\begin{aligned} \begin{bmatrix} \Sigma_y & \Sigma_{yx} \\ \Sigma_{xy} & \Sigma_x \end{bmatrix} &\equiv E \left\{ \begin{bmatrix} y \\ x \end{bmatrix} (y', x') \right\} \\ &= \begin{bmatrix} B & \Gamma \\ 0 & I_k \end{bmatrix}^{-1} \begin{bmatrix} \Sigma_\zeta & 0 \\ 0 & \Sigma_x \end{bmatrix} \begin{bmatrix} B' & 0 \\ \Gamma' & I_k \end{bmatrix}^{-1}. \end{aligned} \quad (7.11)$$

Thus, equations (7.10) and (7.11) contain all observational information with regard to the structural parameter matrices  $B$ ,  $\Gamma$  and  $\Sigma_\zeta \equiv E(\zeta\zeta')$ .

However, for the identification of the structural parameter matrices we need only consider a subset of the equations in (7.10) and (7.11). The equations in (7.11) can be separated into

$$(B, \Gamma) \begin{bmatrix} \Sigma_y & \Sigma_{yx} \\ \Sigma_{xy} & \Sigma_x \end{bmatrix} \begin{bmatrix} B' \\ \Gamma' \end{bmatrix} - \Sigma_\zeta = 0 \quad (7.12)$$

$$(\Sigma_{xy}, \Sigma_x) \begin{bmatrix} B' \\ \Gamma' \end{bmatrix} = 0, \quad (7.13)$$

and, in fact, these equations contain all observational information relevant for the identification of  $B$ ,  $\Gamma$  and  $\Sigma_\zeta$ . That is to say, since the equations (7.10) and (7.11) are satisfied by the true parameter point, we find that  $(\mu'_y, \mu'_x) = \mu'_x \Sigma_x^{-1} (\Sigma_{xy}, \Sigma_x)$  so that, if any matrices  $B$  and  $\Gamma$  satisfy (7.13), they will also satisfy (7.10). The first moment equations do not provide additional information.

Now let prior information be given by a set of restrictions on the coefficient matrices  $B$  and  $\Gamma$ :

$$\rho(B, \Gamma) = 0, \quad (7.14)$$

where  $\rho(B, \Gamma)$  is a vector function of the matrices  $B$  and  $\Gamma$ . The parameter matrix  $\Sigma_\zeta$  is assumed to be unrestricted. In that case all information relevant for the identification of  $B$  and  $\Gamma$  is given by equation (7.13), which after stacking in a vector can be written as

$$\sigma(B, \Gamma) \equiv (I_m \otimes \Sigma_{xy}) \text{vec}(B') + (I_m \otimes \Sigma_x) \text{vec}(\Gamma') = 0,$$

and by equation (7.14). The Jacobian matrix is

$$J(B, \Gamma) = \frac{\partial(\sigma(B, \Gamma); \rho(B, \Gamma))}{\partial(\text{vec}'(B'), \text{vec}'(\Gamma'))} = \begin{bmatrix} (I_m \otimes \Sigma_{xy}, I_m \otimes \Sigma_x) \\ R_{B\Gamma} \end{bmatrix},$$

where  $R_{B\Gamma}$  is defined implicitly and is assumed to be of full row rank. Post-multiplication of the Jacobian by a conveniently chosen nonsingular matrix:

$$J(B, \Gamma) \begin{bmatrix} I_m \otimes B' & 0 \\ I_m \otimes \Gamma' & I_{mk} \end{bmatrix} = \begin{bmatrix} 0 & I_m \otimes \Sigma_x \\ R_{B\Gamma} \begin{bmatrix} I_m \otimes B' \\ I_m \otimes \Gamma' \end{bmatrix} & R_{B\Gamma} \begin{bmatrix} 0 \\ I_{mk} \end{bmatrix} \end{bmatrix}$$

shows that  $J(B, \Gamma)$  has full column rank if and only if

$$\tilde{J}(B, \Gamma) \equiv R_{B\Gamma} \begin{bmatrix} I_m \otimes B' \\ I_m \otimes \Gamma' \end{bmatrix} \quad (7.15)$$

has full column rank. Furthermore  $J(B, \Gamma)$  and  $\tilde{J}(B, \Gamma)$  share regular points. Thus, by Theorem 7 we have the following result.

**Theorem 10.** Let  $\mathcal{H} \equiv \{(B, \Gamma) \mid (B, \Gamma) \in \mathbf{R}^{m^2+mk}, \rho(B, \Gamma) = 0\}$  and let  $(B, \Gamma)^0$  be a regular point of  $\tilde{J}(B, \Gamma) \mid \mathcal{H}$ . Then  $(B, \Gamma)^0$  is locally identified if and only if  $\tilde{J}(B^0, \Gamma^0)$  has full row rank  $m^2$ .

This result corresponds to the classical condition for identification in a simultaneous equation model. If the restrictions on  $(B, \Gamma)$  are linear, i.e. when  $\rho(B, \Gamma)$  is a linear function, for example when separate elements of  $(B, \Gamma)$  are restricted to be fixed, then both sets of identifying equations (7.13) and (7.14) are linear so that  $J(B, \Gamma)$  is constant over the parameter space. In that case local identification implies global identification and we have the following corollary.

**Corollary 2.** Let  $\rho(B, \Gamma)$  be linear, then  $(B, \Gamma)^0$  is globally identified if and only if  $\text{rank}\{\tilde{J}(B, \Gamma)\} = m^2$ .

This constitutes the well-known rank condition for identification in simultaneous equations models, as developed in the early work of the Cowles Commission, e.g. Koopmans, Rubin, and Leipnik (1950), Koopmans (1953), and Koopmans and Hood (1953). Johansen (1995, theorem 3) gives an elegant formulation of the identification of the coefficients of a single equation.

## 9 CONCLUDING REMARKS

In this chapter we have presented a rigorous, self-contained treatment of identification in parametric models with iid observations. The material is essentially from the book by Bekker, Merckens, and Wansbeek (1994); see Rigdon (1997) for an extended review. The reader is referred to this book for a number of further topics. For example, it discusses identification of two extensions of the classical simultaneous equations model in two directions, viz. restrictions on the covariance matrix of the disturbances, and the measurement error in the regressors. It also discusses local identification of the equally classical factor analysis model.<sup>4</sup> These two models have been integrated in the literature through the hugely popular Lisrel model, which however often confronts researchers with identification problems which are hard to tackle analytically since, in the rank condition for identification, inverse matrices cannot be eliminated as in the classical simultaneous equations model. The book tackles this issue by parameterizing the restrictions on the reduced form induced by the restrictions on the structural form.

A distinctive feature of the book is its use of symbolic manipulation of algebraic structures by the computer. Essentially, all identification and equivalence results are couched in terms of ranks of structured matrices containing unknown parameters. To assess such ranks has become practically feasible by using computer algebra. The book contains a diskette with a set of computer algebra programs that can be used for rank evaluation of parameterized matrices for the models discussed.

## Notes

- \* This chapter is largely based on material adapted from P.A. Bekker, A. Merckens, and T.J. Wansbeek, *Identification, Equivalent Models and Computer Algebra* (Orlando: Academic Press, 1994). Reproduced by kind permission of the publisher. We are grateful to Badi Baltagi, Bart Boon, and an anonymous referee for their useful comments.
- 1 Identification in nonparametric models is a much different field, see, e.g., Prakasa Rao (1992).
- 2 Of course, it may be the case that exact knowledge of  $P(y, \theta^0)$  is sufficient to derive bounds on  $\theta_k^0$  (see, e.g., Bekker *et al.*, 1987; Manski, 1989; Manski, 1995). In such a case the sample information can be used to increase knowledge about  $\theta_k^0$  even though this parameter is not locally identified.
- 3 However, if one uses a “natural” parameter sequence, it may happen that  $\theta_k^0$  is identified, whereas no estimator converges in probability to  $\theta_k^0$ . For example, Gabrielsen (1978) discussed the model  $y_i = \beta r^i + u_i$ ,  $i = 1, \dots, n$ , where the  $u_i$  are iid  $N(0, 1)$ ,  $r$  is known and  $|r| < 1$ , and  $\beta$  is an unknown parameter. Here the OLS estimator  $\hat{\beta} \sim N(\beta, (1 - r^2)/(r^2(1 - r^{2n})))$  is unbiased, so clearly  $\beta$  is identified, but it is not consistent in the natural sequence defined by the model where  $n \rightarrow \infty$ . Since  $\hat{\beta}$  is efficient, there does not exist a consistent estimator.
- 4 For a discussion of global identification in factor analysis see Bekker and ten Berge (1997).

## References

- Aigner D.J., C. Hsiao, A. Kapteyn, and T.J. Wansbeek (1984). Latent variable models in econometrics. In Z. Griliches and M.D. Intriligator (eds.) *Handbook of Econometrics volume 2*. Amsterdam: North-Holland.
- Aldrich, J. (1994). Haavelmo’s identification theory. *Econometric Theory* 10, 198–219.
- Bekker, P.A. (1986). Comment on identification in the linear errors in variables model. *Econometrica* 54, 215–17.
- Bekker, P.A., A. Kapteyn, and T.J. Wansbeek (1987). Consistent sets of estimates for regressions with correlated or uncorrelated measurement errors in arbitrary subsets of all variables. *Econometrica* 55, 1223–30.
- Bekker, P.A., A. Merckens, and T.J. Wansbeek (1994). *Identification, Equivalent Models and Computer Algebra*. Orlando: Academic Press.
- Bekker, P.A., and J.M.F. ten Berge (1997). Generic global identification in factor analysis. *Linear Algebra and its Applications* 264, 255–63.
- Bowden, R. (1973). The theory of parametric identification. *Econometrica* 41, 1069–74.
- Deistler, M., and H.-G. Seifert (1978). Identifiability and consistent estimability in econometric models. *Econometrica* 46, 969–80.
- Drèze, J. (1975). Bayesian theory of identification in simultaneous equations models. In S.E. Fienberg and A. Zellner (eds.) *Studies in Bayesian Econometrics and Statistics*. Amsterdam: North-Holland.
- Fisher, F.M. (1966). *The Identification Problem in Econometrics*. New York: McGraw-Hill.
- Gabrielsen, A. (1978). Consistency and identifiability. *Journal of Econometrics* 8, 261–3.
- Haavelmo, T. (1943). The statistical implications of a system of simultaneous equations. *Econometrica* 11, 1–12.
- Hannan, E.J., and M. Deistler (1988). *The Statistical Theory of Linear Systems*. New York: Wiley.
- Hsiao, C. (1983). Identification. In Z. Griliches and M.D. Intriligator (eds.) *Handbook of Econometrics Volume 1*. Amsterdam: North-Holland.

- Hsiao, C. (1987). Identification. In J. Eatwell, M. Millgate, and P. Newman (eds.) *The New Palgrave: A Dictionary of Economics*. London: Macmillan.
- Hsiao, C. (1997). Cointegration and dynamic simultaneous equations model. *Econometrica* 65, 647–70.
- Johansen, S. (1995). Identifying restrictions of linear equations with applications to simultaneous equations and cointegration. *Journal of Econometrics* 69, 111–32.
- Kadane, J.B. (1975). The role of identification in Bayesian theory. In S.E. Fienberg and A. Zellner (eds.) *Studies in Bayesian Econometrics and Statistics*. Amsterdam: North-Holland.
- Koopmans, T.C., H. Rubin, and R.B. Leipnik (1950). Measuring the equation systems of dynamic economics. In T.C. Koopmans (ed.) *Statistical Inference in Dynamic Economic Models*. New York: Wiley.
- Koopmans, T.C. (1953). Identification problems in economic model construction. In W.C. Hood and T.C. Koopmans (eds.) *Studies in Econometric Methods*. New York: Wiley.
- Koopmans, T.C., and W.C. Hood (1953). The estimation of simultaneous linear economic relationships. In W.C. Hood and T.C. Koopmans (eds.) *Studies in Econometric Methods*. New York: Wiley.
- Leamer, E.E. (1978). *Specification Searches, ad hoc Inference with Nonexperimental Data*. New York: Wiley.
- Magnus, J.R. (1988). *Linear Structures*. London: Griffin.
- Manski, C.F. (1989). Anatomy of the selection problem. *The Journal of Human Resources* 24, 343–60.
- Manski, C.F. (1995). *Identification Problems in the Social Sciences*. Cambridge, MA: Harvard University Press.
- McManus, D.A. (1992). How common is identification in parametric models? *Journal of Econometrics* 53, 5–23.
- Pesaran, M.H. (1987). Econometrics. In J. Eatwell, M. Millgate, and P. Newman (eds.) *The New Palgrave: A Dictionary of Economics*. London: Macmillan.
- Poirier, D.J. (1998). Revising beliefs in nonidentified models. *Econometric Theory* 14, 483–509.
- Prakasa Rao, B.L.S. (1992). *Identifiability in Stochastic Models: Characterization of Probability Distributions*. Boston: Academic Press.
- Richmond, J. (1974). Identifiability in linear models. *Econometrica* 42, 731–6.
- Rigdon, E.E. (1997). Identification of structural equation models with latent variables: a review of contributions by Bekker, Merckens, and Wansbeek. *Structural Equation Modeling* 4, 80–5.
- Rothenberg, T.J. (1971). Identification in parametric models. *Econometrica* 39, 577–92.
- Zellner, A. (1971). *An Introduction to Bayesian Inference in Econometrics*. New York: Wiley.

CHAPTER EIGHT

# Measurement Error and Latent Variables

*Tom Wansbeek and Erik Meijer\**

## 1 INTRODUCTION

Traditionally, an assumption underlying econometric models is that the regressors are observed without measurement error. In practice, however, economic observations, micro and macro, are often imprecise (Griliches, 1986). This may be due to clearly identifiable factors. If these are known, we may apply a better measurement procedure on a later occasion. However, it may also be the case that no better procedure is possible, not even in a perfect world. The variable concerned may be a purely mental construct that does not correspond to a variable that can, at least in principle, be observed in practice. In fact, quite often economic theorizing involves such *latent* variables.

Typical examples of latent variables appearing in economic models are utility, the productivity of a worker, permanent income, consumer satisfaction, financial health of a firm, the weather condition in a season, socioeconomic status, or the state of the business cycle. Although we use the epithet "latent" for these variables, we can, for each of these examples, think of related *observable* variables, so some kind of indirect measurement is possible. In this sense the latency of variables is a generalization of measurement error, where the relation between the observed variable and its true or latent counterpart is just of the simple kind: observed = true + measurement error.

Clearly, many variables economists work with are latent, due to measurement error or intrinsically so. In this chapter, we will discuss the problems that are invoked by the presence of measurement error and latent variables in econometric models, possible solutions to these problems, and the opportunities offered by latent variable models. Related references are Aigner, Hsiao, Kapteyn, and Wansbeek (1984), who give an extensive overview of latent variable models, and Fuller (1987) and Cheng and Van Ness (1999), which are book-length treatments of measurement error models.

## 2 THE LINEAR REGRESSION MODEL WITH MEASUREMENT ERROR

The standard linear multiple regression model can be written as

$$y = \Xi\beta + \varepsilon, \quad (8.1)$$

where  $y$  is an observable  $N$ -vector,  $\varepsilon$  an unobservable  $N$ -vector of random variables, the elements of which are independently identically distributed (iid) with zero expectation and variance  $\sigma_\varepsilon^2$ , and  $N$  is the sample size. The  $g$ -vector  $\beta$  is fixed but unknown. The  $N \times g$ -matrix  $\Xi$  contains the regressors, which are assumed to be independent of  $\varepsilon$ . For simplicity, variables are assumed to be measured in deviations from their means.<sup>1</sup>

If there are measurement errors in the explanatory variables,  $\Xi$  is not observed. Instead, we observe the matrix  $X$ :

$$X = \Xi + V, \quad (8.2)$$

where  $V (N \times g)$  is a matrix of measurement errors. Its rows are assumed to be iid with zero expectation and covariance matrix  $\Omega (g \times g)$  and independent of  $\Xi$  and  $\varepsilon$ . Columns of  $V$  (and corresponding rows and columns of  $\Omega$ ) are zero when the corresponding regressors are measured without error.

We consider the consequences of neglecting the measurement errors. Let

$$\begin{aligned} b &\equiv (X'X)^{-1}X'y \\ s_e^2 &\equiv \frac{1}{N-g} (y - Xb)'(y - Xb) = \frac{1}{N-g} y'(I_N - X(X'X)^{-1}X')y \end{aligned}$$

be the ordinary least squares (OLS) estimators of  $\beta$  and  $\sigma_e^2$ . Substitution of (8.2) into (8.1) yields

$$y = (X - V)\beta + \varepsilon = X\beta + u, \quad (8.3)$$

with  $u \equiv \varepsilon - V\beta$ . This means that (8.3) has a disturbance term which shares a stochastic term ( $V$ ) with the regressor matrix. Thus,  $u$  is correlated with  $X$  and  $E(u | X) \neq 0$ . This lack of orthogonality means that a crucial assumption underlying the use of ordinary least squares regression is violated. As we shall see below, the main consequence is that  $b$  and  $s_e^2$  are no longer consistent estimators of  $\beta$  and  $\sigma_e^2$ . In order to analyze the inconsistency, let  $S_\Xi \equiv \Xi'\Xi/N$  and  $S_X \equiv X'X/N$ . Note that  $S_X$  is observable but  $S_\Xi$  is not.

We can interpret (8.1) in two ways. It is either a *functional* or a *structural* model. Under the former interpretation, we do not make explicit assumptions regarding the distribution of  $\Xi$ , but consider its elements as unknown fixed parameters. Under the latter interpretation, the elements of  $\Xi$  are supposed to be random variables. For both cases, we assume  $\text{plim}_{N \rightarrow \infty} S_\Xi = \Sigma_\Xi$  with  $\Sigma_\Xi$  a positive definite  $g \times g$ -matrix. Hence,

$$\Sigma_X \equiv \text{plim}_{N \rightarrow \infty} \frac{1}{N} X'X = \text{plim}_{N \rightarrow \infty} S_X = \Sigma_\Xi + \Omega.$$

Note that since  $\Sigma_{\varepsilon}$  is positive definite and  $\Omega$  is positive semidefinite,  $\Sigma_X$  is also positive definite.

## 2.1 Inconsistency and bias of the OLS estimators

Given the setup, the probability limits of  $b$  and  $s_e^2$ , for both the structural and the functional model, are

$$\kappa \equiv \text{plim}_{N \rightarrow \infty} b = \beta - \Sigma_X^{-1}\Omega\beta = \Sigma_X^{-1}\Sigma_{\varepsilon}\beta. \quad (8.4)$$

$$\gamma \equiv \text{plim}_{N \rightarrow \infty} s_e^2 = \sigma_e^2 + \beta'(\Sigma_{\varepsilon} - \Sigma_{\varepsilon}\Sigma_X^{-1}\Sigma_{\varepsilon})\beta \geq \sigma_e^2. \quad (8.5)$$

Hence  $b$  is inconsistent, with inconsistency equal to  $\kappa - \beta = -\Sigma_X^{-1}\Omega\beta$ , and  $s_e^2$  is also inconsistent, the inconsistency being nonnegative since  $\Sigma_{\varepsilon} - \Sigma_{\varepsilon}\Sigma_X^{-1}\Sigma_{\varepsilon} = \Omega - \Omega\Sigma_X^{-1}\Omega \geq 0$ .

Consider the case that there is only one regressor ( $g = 1$ ), which is measured with error. Then  $\Sigma_X > \Sigma_{\varepsilon} > 0$  are scalars and  $\kappa/\beta = \Sigma_X^{-1}\Sigma_{\varepsilon}\beta/\beta$  is a number between 0 and 1. So asymptotically the regression coefficient estimate is biased towards zero. This phenomenon is called *attenuation*. The size of the effect of the regressor on the dependent variable is underestimated.

In the multiple regression case the characterization of the attenuation is slightly more complicated. The inequality  $\Sigma_{\varepsilon} - \Sigma_{\varepsilon}\Sigma_X^{-1}\Sigma_{\varepsilon} \geq 0$  and (8.4) together imply

$$\beta'\Sigma_{\varepsilon}\beta \geq \kappa'\Sigma_X\kappa \quad (8.6)$$

or, using  $\Sigma_X\kappa = \Sigma_{\varepsilon}\beta$ ,  $(\beta - \kappa)'\Sigma_X\kappa \geq 0$ . This generalizes  $\beta - \kappa \geq 0$  for the case  $g = 1$  (assuming  $\kappa > 0$ ). So, given  $\beta$ ,  $\kappa$  is located in the half space bounded by the hyperplane  $\beta'c = \kappa'c$  that includes the origin, where  $c \equiv \Sigma_{\varepsilon}\beta = \Sigma_X\kappa$ , which is a hyperplane through  $\beta$  perpendicular to  $c$ . It is, however, possible that  $\kappa$  is farther from the origin than  $\beta$ .

The term  $\beta'\Sigma_{\varepsilon}\beta$  in (8.6) is the variance of the systematic part of the regression. The term  $\kappa'\Sigma_X\kappa$  is its estimate if measurement error is neglected. Thus the variance of the systematic part is underestimated. This also has a direct bearing on the properties of  $R^2 \equiv (b'S_Xb)/(\frac{1}{N}y'y)$ . This statistic converges to  $\rho^2$ , where

$$\rho^2 \equiv \frac{\kappa'\Sigma_X\kappa}{\sigma_e^2 + \beta'\Sigma_{\varepsilon}\beta} \leq \frac{\beta'\Sigma_{\varepsilon}\beta}{\sigma_e^2 + \beta'\Sigma_{\varepsilon}\beta},$$

and the right-hand side is the “true”  $R^2$ . So the explanatory power of the model is underestimated.

When there is more than one regressor, but only one is measured with error, generally the estimators of all regression coefficients are biased. The coefficient of the error-ridden regressor (the first one, say) is biased towards zero, and the sign of the biases of the other parameters can be estimated consistently. Let  $e_i$  denote the  $i$ th unit vector,  $\beta_1 = e_1'\beta$  (assumed positive), and  $\psi \equiv e_1'\Omega e_1 > 0$ , then the bias of the  $i$ th element of the estimator of  $\beta$  is  $e_i'(\kappa - \beta) = -e_i'\Sigma_X^{-1}\Omega\beta = -\psi\beta_1 \cdot e_i'\Sigma_X^{-1}e_1$ . Thus, the first regression coefficient is underestimated whereas the signs of the

biases in the other coefficients depend on the sign of the elements of the first column of  $\Sigma_x^{-1}$ . Even when  $\Omega$  is unknown, these signs can be consistently estimated from the signs of the corresponding elements of  $S_x^{-1}$ .

## 2.2 Bounds on the parameters

Let us return to the bivariate regression model (no intercept, all variables having mean zero), written in scalar notation:

$$y_n = \beta \xi_n + \varepsilon_n \quad (8.7a)$$

$$x_n = \xi_n + v_n, \quad (8.7b)$$

where  $n$  denotes a typical element and the other notation is obvious. Assume for simplicity that  $\beta > 0$  (the case with  $\beta < 0$  is similar). OLS yields

$$\text{plim}_{N \rightarrow \infty} (x'x)^{-1}x'y = \beta(1 - \sigma_v^2/\sigma_x^2) = \kappa < \beta, \quad (8.8)$$

the well known bias towards zero. But it also holds that

$$\text{plim}_{N \rightarrow \infty} (x'y)^{-1}y'y = \beta + \frac{\sigma_\varepsilon^2}{\beta \sigma_\xi^2} > \beta. \quad (8.9)$$

The left-hand side of (8.9) is the probability limit of the inverse of the coefficient of the regression of  $x$  on  $y$  (the "reverse regression"). This regression also gives an inconsistent estimator of  $\beta$ , but with a bias away from zero. Thus, (8.8) and (8.9) bound the true  $\beta$  from below and above, respectively. Since these bounds can be estimated consistently, by the regression and the reverse regression, we can take  $(x'x)^{-1}x'y$  and  $(x'y)^{-1}y'y$  as bounds between which  $\beta$  should lie in the limit. The bounds are obtained without making assumptions about the size of the measurement error.

These results on bounds without additional information carry over to the multiple regression case to a certain limited extent only:  $\beta$  lies anywhere in the convex hull of the elementary regression vectors if these are all positive, where the  $g + 1$  *elementary* regression vectors are defined as the regression vectors of each of the  $g + 1$  variables on the  $g$  other variables (scaled properly). This condition can be formulated slightly more generally by saying that it suffices that all regression vectors are in the same orthant since by changing signs of variables this can simply be translated into the previous condition (see Bekker, Wansbeek, and Kapteyn, 1985, for a discussion).

Reverse regression has drawn much attention in the context of the analysis of discrimination; see, e.g., Goldberger (1984a, 1984b). In its simplest form, the model is an extension of (8.7):

$$y_n = \beta \xi_n + \alpha d_n + \varepsilon_n$$

$$x_n = \xi_n + v_n,$$

$$\xi_n = \mu d_n + \zeta_n,$$

where  $y_n$  is wage,  $d_n$  is a dummy indicating race or gender, and  $\xi_n$  is productivity. This variable can only be measured imperfectly through the indicator  $x_n$ . The last equation in this model reflects the different average level of productivity between race or gender groups. The crucial parameter is  $\alpha$ , since a nonzero value (i.e. a wage differential even after controlling for productivity) may be interpreted as a sign of discrimination. Regressing  $y$  on  $x$  and  $d$  can be shown to give an overestimate of  $\alpha$ . Reverse regression, i.e. regressing  $x$  on  $y$  and  $d$ , is a useful technique here, since it can be shown to give an underestimate of  $\alpha$ . The primary research question then is whether the two estimates have the same sign.

### 3 SOLUTIONS TO THE MEASUREMENT ERROR PROBLEM

In Section 2, it was shown that in the linear regression model with measurement errors, the OLS estimators are biased and inconsistent. There is no “quick fix” since the inconsistency is due to an identification problem. Essentially, identification is equivalent to the existence of a consistent estimator; see Bekker, Merckens, and Wansbeek (1994, p. 18), and in a generic case the measurement error model given by (8.1) and (8.2) is not identified if all random variables are jointly normally distributed or attention is confined to the first and second order moments of the observed variables only.

Let the elements of  $\epsilon$  and the rows of  $V$  be iid normal. We consider first the case of a structural model. Then, according to Bekker (1986),  $\beta$  is identified if and only if there does not exist a nonsingular matrix  $A = (A_1, A_2)$  such that  $\xi' A_1$  is distributed normally and independently of  $\xi' A_2$ , where  $\xi'$  is a typical row of  $\Xi$ . In particular, this implies that if  $\xi$  is normally distributed,  $\beta$  is not identified. Due to a result by Wald (1940) this result applies in the functional case as well (cf. Aigner *et al.*, 1984, p. 1335). This means that, when searching for consistent estimators, additional information, instruments, nonnormality, or additional structure (typically through panel data) are desirable. In this section we consider these cases in turn. In a Bayesian context, incidentally, the outlook is different and inference on  $\beta$  is possible without identifying information, see, e.g., Poirier (1998).

#### 3.1 Restrictions on the parameters

Equations (8.4) and (8.5) show that the inconsistency of  $b$  and  $s_\epsilon^2$  could be removed if  $\Omega$  were known. For example, rather than  $b$  we would take  $(I_g - S_X^{-1}\Omega)^{-1}b$  as an estimator of  $\beta$ , and from (8.4) it is clear this estimator is consistent. In general,  $\Omega$  is unknown. If, however, we have a consistent estimator of  $\Omega$ , we can replace  $\Omega$  by its consistent estimate and obtain an estimator of  $\beta$  that by virtue of Slutsky’s theorem is consistent. The resulting statistic is then a least squares estimator that is adjusted to attain consistency.

As a generalization, assume that a system of just identifying restrictions on the unknown parameters  $\beta$ ,  $\sigma_\epsilon^2$ , and  $\Omega$  is available:

$$F(\beta, \sigma_\epsilon^2, \Omega) = 0, \quad (8.10)$$

with  $F$  a totally differentiable vector-function of order  $g^2$ . In view of the symmetry of  $\Omega$ ,  $\frac{1}{2}g(g-1)$  of the restrictions in (8.10) are of the form  $\Omega_{ij} - \Omega_{ji} = 0$ .

If we combine the sample information with the prior information and add hats to indicate estimators we obtain the following system of equations:

$$(I_g - S_X^{-1}\hat{\Omega})\hat{\beta} - b = 0, \quad (8.11a)$$

$$\hat{\sigma}_\epsilon^2 + \hat{\beta}'\hat{\Omega}b - s_\epsilon^2 = 0, \quad (8.11b)$$

$$F(\hat{\beta}, \hat{\sigma}_\epsilon^2, \hat{\Omega}) = 0. \quad (8.11c)$$

When  $F$  is such that this system admits a unique solution for  $\hat{\beta}$ ,  $\hat{\sigma}_\epsilon^2$ , and  $\hat{\Omega}$ , this solution will be a consistent estimator of  $\beta$ ,  $\sigma_\epsilon^2$ , and  $\Omega$  since, asymptotically,  $S_X$  tends to  $\Sigma_X$  and the system then represents the relationship between the true parameters on the one hand and  $\text{plim } b$  and  $\text{plim } s_\epsilon^2$  on the other. This solution is called the *consistent adjusted least squares* (CALS) estimator (Kapteyn and Wansbeek, 1984).

The CALS estimator is easy to implement. One can use a standard regression program to obtain  $b$  and  $s_\epsilon^2$  and then employ a computer program for the solution of a system of nonlinear equations. In many cases it will be possible to find an explicit solution for (8.11), which then obviates the necessity of using a computer program for the solution of nonlinear equations.

It can be shown that, in both the structural model and the functional model, the CALS estimator has asymptotic distribution

$$\sqrt{N}(\hat{\theta}_{\text{CALS}} - \theta) \xrightarrow{D} \mathcal{N}(0, H_\theta^{-1}H_\delta\Delta H'_\delta(H_\theta^{-1})'),$$

where  $\theta = (\beta', \sigma_\epsilon^2, (\text{vec } \Omega)')'$  and  $\hat{\theta}_{\text{CALS}}$  is the CALS estimator of  $\theta$ ,

$$H_\theta \equiv \begin{bmatrix} \Sigma_X^{-1}\Sigma_\varepsilon & 0 & -\beta' \otimes \Sigma_X^{-1} \\ \kappa'\Omega & 1 & \kappa' \otimes \beta' \\ \frac{\partial F}{\partial \beta'} & \frac{\partial F}{\partial \sigma_\epsilon^2} & \frac{\partial F}{\partial (\text{vec } \Omega)'} \end{bmatrix}$$

$$H_\delta \equiv \begin{bmatrix} -I_g & 0 & (\beta'\Omega \otimes I_g)(\Sigma_X^{-1} \otimes \Sigma_X^{-1})Q_g \\ \beta'\Omega & -1 & 0 \\ 0 & 0 & 0 \end{bmatrix}$$

$$\Delta \equiv \begin{bmatrix} \gamma\Sigma_X^{-1} & 0 & 0 \\ 0 & 2\gamma^2 & 0 \\ 0 & 0 & 2Q_g(\Sigma_X \otimes \Sigma_X) \end{bmatrix},$$

and  $Q_g \equiv \frac{1}{2}(I_{g^2} + P_g)$ , where  $P_g$  is the *commutation matrix* (Wansbeek, 1989).

Let us now consider the case in which inexact information on the measurement error variances is available of the following form:

$$0 \leq \Omega \leq \Omega^* \leq \Sigma_X,$$

with  $\Omega^*$  given. The motivation behind such a bound is that researchers who have reason to suppose that measurement error is present may not know the actual size of its variance but may have an idea of an upper bound to that variance. This makes it possible to derive a bound on  $\beta$  that can be consistently estimated. If  $\Omega > 0$ , i.e. there is measurement error in all variables,  $\beta$  should satisfy the inequality

$$(\beta - \frac{1}{2}(\kappa + \kappa^*))' \Sigma_X (\Omega^{*-1} - \Sigma_X^{-1}) \Sigma_X (\beta - \frac{1}{2}(\kappa + \kappa^*)) \leq \frac{1}{4}(\kappa^* - \kappa)' \Sigma_X \kappa,$$

where  $\kappa^* \equiv (\Sigma_X - \Omega^*)^{-1} \Sigma_X \kappa$  is the probability limit of the estimator under the assumption that  $\Omega = \Omega^*$ . This is an ellipsoid with midpoint  $\frac{1}{2}(\kappa + \kappa^*)$ , passing through  $\kappa$  and  $\kappa^*$  and tangent to the hyperplane  $\kappa' \Sigma_X (\beta - \kappa) = 0$ . If  $\Omega$  is singular (some variables are measured without error), a similar but more complicated inequality can be derived, see Bekker, Kapteyn, and Wansbeek (1984, 1987). In practical cases such ellipsoid bounds will be hard to use and bounds on the elements of  $\beta$  separately will be of more interest. Let  $a$  be an arbitrary  $g$ -vector. The ellipsoid bound implies for the linear combination  $a'\beta$  that

$$\frac{1}{2}a'(\kappa + \kappa^*) - \frac{1}{2}\sqrt{c} \leq a'\beta \leq \frac{1}{2}a'(\kappa + \kappa^*) + \frac{1}{2}\sqrt{c},$$

with  $c \equiv (\kappa^* - \kappa)' \Sigma_X \kappa \cdot a' F^* a$  and  $F^* \equiv (\Sigma_X - \Omega^*)^{-1} - \Sigma_X^{-1}$ . Bounds on separate elements of  $\beta$  are obtained when  $a$  is set equal to any of the  $g$  unit vectors. Erickson (1993) derived similar results by bounding the error *correlation* matrix, rather than the error covariance matrix. Bounds on quantities that are not identified are not often considered in econometrics and run against conventional wisdom. Yet, their usefulness has been recently advocated in the monograph by Manski (1995).

### 3.2 Instrumental variables

To introduce the notion of instrumental variables, we start from (8.3):  $y = X\beta + u$ , where consistent estimation was hampered by the correlation between  $X$  and  $u$ . If there are observations on variables, collected in  $Z$ , say, that correlate with the variables measured in  $X$  but do not correlate with  $u$  we have that the last term in  $Z'y/N = (Z'X/N)\beta + Z'u/N$  vanishes asymptotically and hence may lead to consistent estimation of  $\beta$ . This is the idea behind *instrumental variables* (IV) estimation,  $Z$  being the matrix of instrumental variables. Of all the methods used to obtain consistent estimators in models with measurement error or with endogenous regressors in general, this is undisputedly the most popular one.

If  $Z$  is of the same order as  $X$  and  $Z'X/N$  converges to a finite, nonsingular matrix, the IV estimator  $b_{IV}$  of  $\beta$ , defined as

$$b_{IV} \equiv (Z'X)^{-1}Z'y, \quad (8.12)$$

is consistent. When  $Z$  has  $h > g$  columns, the IV estimator is

$$b_{IV} = (X'P_Z X)^{-1}X'P_Z y,$$

where  $P_Z \equiv Z(Z'Z)^{-1}Z'$ . For  $h = g$  this reduces to (8.12). Letting  $\hat{X} \equiv P_Z X$ , we have alternatively  $b_{IV} = (\hat{X}'\hat{X})^{-1}\hat{X}'y$ , so it can be computed by OLS after transforming  $X$ , which comes down to computing the predicted value of  $X$  after regressing each of its columns on  $Z$ . Therefore,  $b_{IV}$  is also called the two-stage least squares (2SLS) estimator.

Under some standard regularity conditions,  $b_{IV}$  is asymptotically normally distributed (Bowden and Turkington, 1984, p. 26):

$$\sqrt{N}(b_{IV} - \beta) \xrightarrow{\text{d}} \mathcal{N}(0, \sigma_u^2(\Sigma'_{ZX}\Sigma_{ZZ}^{-1}\Sigma_{ZX})^{-1}),$$

where  $\sigma_u^2 \equiv \sigma_\epsilon^2 + \beta'\Omega\beta$ ,  $\Sigma_{ZX} \equiv \text{plim}_{N \rightarrow \infty} Z'X/N$ , and  $\Sigma_{ZZ} \equiv \text{plim}_{N \rightarrow \infty} Z'Z/N$ . The asymptotic covariance matrix can be consistently estimated by inserting  $Z'X/N$  for  $\Sigma_{ZX}$ ,  $Z'Z/N$  for  $\Sigma_{ZZ}$ , and the consistent estimator  $\hat{\sigma}_u^2 = (y - Xb_{IV})'(y - Xb_{IV})/N$  for  $\sigma_u^2$ . The residual variance  $\sigma_\epsilon^2$  can be consistently estimated by  $\hat{\sigma}_\epsilon^2 = y'(y - Xb_{IV})/N$ , which differs from  $\hat{\sigma}_u^2$  because, in contrast to the OLS case, the residuals  $y - Xb_{IV}$  are not perpendicular to  $X$ . The matrix  $\Omega$  cannot be estimated consistently unless additional assumptions are made. For example, if it is assumed that  $\Omega$  is diagonal, a consistent estimator of  $\Omega$  can be obtained.

The availability of instrumental variables is not only useful for consistent estimation in the presence of measurement error, it can also offer the scope for testing whether measurement error is present. An obvious testing strategy is to compare  $b$  from OLS with  $b_{IV}$  and to see whether they differ significantly. Under the null hypothesis of no measurement error, the difference between the two will be purely random, but since they have different probability limits under the alternative hypothesis, a significant difference might be indicative of measurement error.

If normality of the disturbance term  $u$  is assumed, a test statistic for the null hypothesis of no measurement error is (Bowden and Turkington, 1984, p. 51)

$$\frac{q/g}{(q^* - q)/(N - 2g)},$$

assuming  $X$  and  $Z$  do not share columns, where

$$\begin{aligned} q &= (b - b_{IV})'((X'P_Z X)^{-1} - (X'X)^{-1})(b - b_{IV}) \\ q^* &= (y - Xb)'(y - Xb) \end{aligned}$$

and  $b$  is the OLS estimator  $(X'X)^{-1}X'y$ . Under the null hypothesis, this test statistic follows an  $F$ -distribution with  $g$  and  $N - 2g$  degrees of freedom. If the

disturbances are not assumed to be normally distributed, a test statistic for no measurement error is  $q/\hat{\sigma}_u^2$ , which under the null hypothesis converges in distribution to a chi-square variate with  $g$  degrees of freedom (Bowden and Turkington, 1984, p. 51).

Most of the above discussion has been asymptotic. It appears that in finite samples, instrumental variables estimators do not possess the desirable properties they have asymptotically, especially when the instruments correlate only weakly with the regressors (e.g. Nelson and Startz, 1990a, 1990b; Bound, Jaeger, and Baker, 1995; Staiger and Stock, 1997). This happens often in practice, because good instruments are frequently hard to find. Consider, for example, household income. If that is measured with error, typical instruments may be years of schooling or age of the head of the household. While these are obviously correlated with income, the relation will generally be relatively weak. Bekker (1994) proposed an alternative estimator that has better small sample properties than the standard IV estimator. His method of moments (MM) estimator is a generalization of the well known LIML estimator for simultaneous equations models. Its formula is given by

$$b_{\text{MM}} \equiv (X'P_Z^*X)^{-1}X'P_Z^*y, \quad (8.13)$$

where  $P_Z^* \equiv P_Z + \lambda_{\text{MM}}(I_N - P_Z)$  and  $\lambda_{\text{MM}}$  is the smallest solution  $\lambda$  of the generalized eigenvalue equation

$$(\bar{S} - \lambda S^\perp) \begin{bmatrix} 1 \\ -\beta \end{bmatrix} = 0, \quad (8.14)$$

where  $\bar{S} \equiv (y, X)'P_Z(y, X)$  and  $S^\perp \equiv (y, X)'(I_N - P_Z)(y, X)$ . Equivalently,  $\lambda_{\text{MM}}$  is the minimum of

$$\lambda \equiv \frac{(y - X\beta)'P_Z(y - X\beta)}{(y - X\beta)'(I_N - P_Z)(y - X\beta)}. \quad (8.15)$$

The solution vector  $\beta$  in (8.14) or (8.15) is equivalent to the estimator  $b_{\text{MM}}$  in (8.13). Under the usual assumptions of  $N \rightarrow \infty$  and  $g$  constant,  $\lambda_{\text{MM}} \rightarrow 0$  and, consequently,  $P_Z^* \rightarrow P_Z$  and the IV estimator and MM estimator are asymptotically equivalent. In finite samples, however, MM performs better than IV.

Other proposals for alternative estimators that should have better small sample properties than standard IV estimators can, for example, be found in Alonso-Borrego and Arellano (1999) and Angrist, Imbens, and Krueger (1999).

### 3.3 Nonnormality

Consider the bivariate regression model with measurement errors (8.7). Further, assume that  $\phi_3 \equiv E(\xi_n^3) \neq 0$  and that  $\xi_n$ ,  $\varepsilon_n$ , and  $v_n$  are independently distributed (similar assumptions can be formulated for the functional model). Now,

$$\operatorname{plim}_{N \rightarrow \infty} \left( \frac{1}{N} \sum_{n=1}^N y_n^2 x_n \middle/ \frac{1}{N} \sum_{n=1}^N y_n x_n^2 \right) = \frac{\beta^2 \phi_3}{\beta \phi_3} = \beta. \quad (8.16)$$

This illustrates that nonnormality may be exploited to obtain consistent estimators. It was first shown by Reiersøl (1950) that the model is identified under nonnormality. The precise condition, as stated at the beginning of Section 3, was derived by Bekker (1986).

There are many ways in which nonnormality can be used to obtain a consistent estimator of  $\beta$ . The estimator (8.16) will generally not be efficient and its small sample properties are usually not very good. Asymptotically more efficient estimators may be obtained by combining the equations for the covariances and several higher order moments and then using a nonlinear GLS procedure (see Section 4). If  $E(\xi_n^3) = 0$ , the third order moments do not give information about  $\beta$  and fourth or higher order moments may be used. An extensive discussion of the ways in which higher order moments can be used to obtain consistent estimators of  $\beta$  is given by Van Montfort, Mooijaart, and De Leeuw (1987).

### 3.4 Panel data

Panel data are repeated measurements over time for a set of cross-sectional units (e.g. households, firms, regions). The additional time dimension allows for consistent estimation when there is measurement error. As a start, consider the simple case with a single regressor and consider the impact of measurement error. For a typical observation indexed by  $n$ ,  $n = 1, \dots, N$ , the model is

$$\begin{aligned} y_n &= \xi_n \beta + \iota_T \cdot \alpha_n + \varepsilon_n \\ x_n &= \xi_n + v_n, \end{aligned}$$

where  $y_n$ ,  $\xi_n$ ,  $\varepsilon_n$ , and  $v_n$  are vectors of length  $T$  and  $T$  is the number of observed time points;  $\xi_n$ ,  $\varepsilon_n$ , and  $v_n$  are mutually independent, and their distributions are independent of  $\alpha_n$ , which is the so-called individual effect, a time-constant latent characteristic (fixed or random) of the cross-sectional units commonly included in a panel data model; see, e.g., Baltagi (1995) for an overview. The vector  $\iota_T$  is a  $T$ -vector of ones. The random vectors  $\xi_n$ ,  $\varepsilon_n$  and  $v_n$  are iid with zero expectation and variances  $E(\xi_n \xi_n') = \Sigma_\xi$ ,  $E(\varepsilon_n \varepsilon_n') = \sigma_\varepsilon^2 I_T$ , and  $E(v_n v_n') = \Sigma_v$ , respectively. Consequently,  $E(x_n x_n') \equiv \Sigma_x = \Sigma_\xi + \Sigma_v$ . Eliminating  $\xi_n$  gives  $y_n = x_n \beta + \iota_T \cdot \alpha_n + u_n$ , where  $u_n \equiv \varepsilon_n - v_n \beta$ .

Let  $Q$  be a symmetric  $T \times T$ -matrix which is as yet unspecified apart from the property  $Q\iota_T = 0$ , so that  $Qy_n$  does not contain the individual effect  $\alpha_n$  anymore. We consider estimators of  $\beta$  of the form

$$\hat{\beta} = \frac{\Sigma_n x_n' Q y_n}{\Sigma_n x_n' Q x_n}. \quad (8.17)$$

This general formulation includes the so-called within-estimator when  $Q = I_T - \mathbf{1}_T \mathbf{1}_T' / T$  is the centering operator and the first-difference estimator when  $Q = RR'$ , where  $R'$  is the matrix taking first differences. Now,

$$\text{plim}_{N \rightarrow \infty} \hat{\beta} = \beta - \beta \frac{\text{tr } Q \Sigma_v}{\text{tr } Q \Sigma_x}.$$

Again we have an estimator that is asymptotically biased towards zero. Whatever the precise structure of  $\Sigma_v$  and  $\Sigma_x$ , it seems reasonable to assume that the true regressor values  $\xi$  will be much stronger correlated over time than the measurement errors  $v$ . Therefore,  $x$  will also have a stronger correlation over time. Hence the variance matrix  $\Sigma_x$  will be more reduced than  $\Sigma_v$  by eliminating the means over time by the  $Q$  matrix and the bias of the estimator (8.17) will be worse than the bias of the OLS estimator.

The main virtue of the panel data structure is, however, that in some cases, several of these estimators (with different  $Q$  matrices) can be combined into a consistent estimator. The basic results on measurement error in panel data are due to Griliches and Hausman (1986); see also Wansbeek and Koning (1991). Further elaboration for a variety of cases is given by Biørn (1992a, 1992b).

## 4 LATENT VARIABLE MODELS

Consider again the bivariate measurement error model (8.7), where the unobservable random variables  $\xi_n$ ,  $\varepsilon_n$ , and  $v_n$  are assumed to be mutually independent with expectation zero. The variables  $y_n$  and  $x_n$  are observable. Now let us assume that there is a third observable variable,  $z_n$ , say, that is linearly related to  $\xi_n$  in the same way as  $y_n$  is:

$$z_n = \gamma \xi_n + u_n, \quad (8.18)$$

where  $u_n$  is independent of  $\xi_n$ ,  $\varepsilon_n$ , and  $v_n$ , and has also mean zero. From this extended model, we obtain the following equations for the variances and covariances of the observable variables (the "covariance" equations):

$$\begin{aligned} \sigma_y^2 &= \sigma_\xi^2 \beta^2 + \sigma_\varepsilon^2 & \sigma_{yx} &= \sigma_\xi^2 \beta & \sigma_{yz} &= \sigma_\xi^2 \beta \gamma \\ \sigma_x^2 &= \sigma_\xi^2 + \sigma_v^2 & \sigma_{xz} &= \sigma_\xi^2 \gamma & \\ \sigma_z^2 &= \sigma_\xi^2 \gamma^2 + \sigma_u^2. \end{aligned}$$

This system of six equations in six unknown parameters can be solved uniquely for the unknown parameters. Since the left-hand variables are consistently estimated by their sample counterparts consistent estimators for the parameters follow immediately. For example, the estimator of  $\beta$  is  $\hat{\beta} = \hat{\sigma}_{yz} / \hat{\sigma}_{xz} = y'z / x'z$ , which is equivalent to the IV estimator with  $z$  as instrumental variable.

## 4.1 Factor analysis

A generalization of the model discussed above is the *factor analysis* model. It is written, in a somewhat different notation, as

$$y_n = \Lambda \xi_n + \varepsilon_n,$$

where  $\xi_n$  is a vector of (common) *factors*,  $y_n$  is a vector of  $M$  *indicators* of these factors,  $\Lambda$  is an  $M \times k$  matrix of *factor loadings*, and  $\varepsilon_n$  is a vector of  $M$  errors. It is assumed that  $E(\xi_n) = 0$ ,  $E(\xi_n \xi_n') \equiv \Phi$ ,  $E(\varepsilon_n) = 0$ ,  $\Psi \equiv E(\varepsilon_n \varepsilon_n')$  is diagonal, and  $E(\xi_n \varepsilon_n') = 0$ . Under this model, the covariance matrix of the observations is

$$\Sigma \equiv E(y_n y_n') = \Lambda \Phi \Lambda' + \Psi.$$

The diagonality of  $\Psi$  implies that any correlation that may exist between different elements of  $y$  is solely due to the common factors  $\xi$ .

The unrestricted model is not identified, but frequently, identifying restrictions on the parameters are available from substantive theory. This is the case when an economic theory forms the base of the model and for every concept in that theory (e.g. productivity of a worker) several well-chosen variables are used that should reflect the concept in question as well as possible. In that case, the loadings of the indicators with respect to the factors (concepts) that they are supposed to reflect are free parameters, whereas the other loadings are fixed to zero. For a given set of restrictions (i.e. a given model), the program IDFAC of Bekker *et al.* (1994) can be used to check whether the model is identified.

## 4.2 MIMIC and reduced rank regression

Above, we considered elaborations of equation (8.18), which offered additional information about the otherwise unidentified parameters in the form of an additional indicator. The latent variable appeared once more as the exogenous variable in yet another regression equation. Another way in which additional information may be available is in the form of a regression equation with the latent variable as the endogenous variable:

$$\xi_n = w_n' \alpha + u_n, \quad (8.19)$$

where  $w_n$  is an  $l$ -vector of observable variables that "cause"  $\xi$ ,  $\alpha$  is an  $l$ -vector of regression coefficients and  $u_n$  is an iid disturbance term with mean zero and variance  $\sigma_u^2$ . Note that in this case,  $w_n$  can also be used as a vector of instrumental variables.

We now show that an additional relation of the form (8.19) can help identification. To that end, it is convenient to write the model in matrix format. For (8.19) this gives  $\xi = W\alpha + u$  and for the factor analysis structure  $Y = \xi\lambda' + E$  in self-evident notation. The model consisting of these two elements is known as the *multiple indicators-multiple causes* (MIMIC) model (Jöreskog and Goldberger, 1975)

and relates a number of exogenous variables (causes) to a number of endogenous variables (indicators) through a single latent variable. In reduced form, it is

$$Y = W\alpha\lambda' + (E + u\lambda').$$

This multivariate regression system has two kinds of restriction on its parameters. First, the matrix of regression coefficients  $\alpha\lambda'$ , which is of order  $l \times M$ , has rank one. Second, the disturbance vector has a covariance matrix of the form  $\Sigma \equiv \Psi + \sigma_u^2\lambda\lambda'$ , with  $\Psi$  diagonal, which is a one-factor FA structure. One normalization on the parameters  $\alpha$ ,  $\lambda$ , and  $\sigma_u^2$  is required since, for any  $c > 0$ , multiplying  $\alpha$  by  $c$ , dividing  $\lambda$  by  $c$  and multiplying  $\sigma_u^2$  by  $c^2$  has no observable implications. Under this normalization, the model is identified.

A frequently used generalization of the MIMIC model is the *reduced rank regression* (RRR) model. It extends MIMIC in two ways. First, the rank of the coefficient matrix can be larger than one, and the error covariance matrix need not be structured. Then

$$Y = WA\Lambda' + F,$$

where  $A$  and  $\Lambda$  are  $l \times r$  and  $M \times r$  matrices, respectively, both of full column rank  $r < \min(M, l)$ , and  $F$  has iid rows with expectation zero and unrestricted covariance matrix  $\Psi$ . See, e.g., Cragg and Donald (1997), for tests of the rank  $r$  of the matrix of regression coefficients. Bekker, Dobbelsstein, and Wansbeek (1996) showed that the *arbitrage pricing theory* model can be written as an RRR model with rank one. In its general form,  $A$  and  $\Lambda$  are not identified, because  $A^*\Lambda^{*'} \equiv (AT)(T^{-1}\Lambda') = A\Lambda'$  for every nonsingular  $r \times r$  matrix  $T$ . In some cases, the identification problem may be resolved by using restrictions derived from substantive theory, whereas in others an arbitrary normalization can be used. Ten Berge (1993, section 4.6) and Reinsel and Velu (1998) give an extensive discussion of the reduced rank regression model and its relations to other multivariate statistical methods.

### 4.3 General linear structural equation models

Up till now, we have discussed several models that specify linear relations among observed and/or latent variables. Such models are called (linear) *structural equation models*. A general formulation of structural equation models can be given by the following equations.

$$x_n = \Lambda_x \xi_n + \delta_n \quad (8.20a)$$

$$y_n = \Lambda_y \eta_n + \varepsilon_n, \quad (8.20b)$$

$$\eta_n = B\eta_n + \Gamma\xi_n + \zeta_n, \quad (8.20c)$$

where  $\eta_n$  is a vector of latent endogenous variables for subject  $n$ ,  $\xi_n$  is a vector of latent exogenous variables for subject  $n$ ,  $\zeta_n$  is a vector of random residuals,  $B$  and  $\Gamma$  are matrices of regression coefficients,  $\Lambda_x$  and  $\Lambda_y$  are matrices of factor loadings,

and  $\delta_n$  and  $\varepsilon_n$  are vectors of errors. The random vectors  $\delta_n$ ,  $\varepsilon_n$ ,  $\zeta_n$ , and  $\xi_n$  are assumed mutually independent. The formulation (8.20) is known as the LISREL model, named after the widely used LISREL program in which it was implemented (Jöreskog and Sörbom, 1996) and consists of a simultaneous equations system in latent endogenous and exogenous variables (8.20c), where (8.20a) and (8.20b) relate the latent variables to observable variables through an FA structure. The theory of structural equation modeling is discussed by Bollen (1989) and Hoyle (1995), which also contains some applications and practicalities. An overview with more recent topics is given by Bentler and Dudgeon (1996).

It turns out that it is possible to write a large number of models as submodels of this model. Examples of submodels are standard linear regression models, simultaneous equations linear regression models, linear regression models with measurement errors, MANOVA, factor analysis, MIMIC. The general model is, of course, highly underidentified. In practice, many restrictions are imposed on the parameters, for example many loadings and regression coefficients are fixed to zero, the scales of the latent variables are fixed by setting a factor loading or a variance parameter to one. The advantage of the general formulation is that all restricted models can be easily estimated by the same computer program and that theoretical properties of estimators can be derived for a large class of models at the same time. For a given set of restrictions (i.e. a given model), the identification of the model can be checked by the program IDLIS (Bekker *et al.*, 1994). For the important special case of simultaneous equations with measurement error (i.e.  $x_n$  and  $\xi_n$  of the same order,  $\Lambda_x = I$ , and analogously for  $y_n$ ,  $\eta_n$ , and  $\Lambda_y$ ), identification conditions are given by Merckens and Bekker (1993) and estimation is discussed by Wooldridge (1996).

The most well known software packages for structural equation modeling are LISREL, including the preprocessor PRELIS (Jöreskog and Sörbom, 1996), EQS (Bentler, 1995), AMOS (Arbuckle, 1997), and SAS/CALIS. A new software package, which can also estimate latent class models, is *Mplus* (Muthén and Muthén, 1998).

## ESTIMATION

If a specific distribution of the observed variables is assumed, typically the normal distribution, the model can be estimated with maximum likelihood (ML). An alternative estimation method is (nonlinear) generalized least squares (GLS). Assume that we have a vector of sample statistics  $s_N$ , which usually consists of the diagonal and subdiagonal elements of the sample covariance matrix  $S_N$  of  $z_n \equiv (x'_n, y'_n)'$ . Further, assume that

$$\sqrt{N}(s_N - \sigma(\theta)) \xrightarrow{\mathcal{L}} \mathcal{N}(0, Y),$$

where the vector  $\sigma(\theta) = \text{plim } s_N$  and  $\theta$  is the vector of free parameters. This assumption is usually satisfied under very mild regularity conditions. The estimator is obtained by minimizing the function

$$F(\theta) = (s_N - \sigma(\theta))' W (s_N - \sigma(\theta)), \quad (8.21)$$

where  $W$  is a symmetric positive definite matrix. If  $\text{plim } W^{-1} = Y$ ,  $W$  is optimal in the sense that the estimator has the smallest asymptotic covariance matrix in the Löwner sense.

If  $s_N$  consists of the nonduplicated elements of the sample covariance matrix, the elements of the matrix  $Y$  are given by the formula  $Y_{ij,kl} = \sigma_{ijkl} - \sigma_{ij}\sigma_{kl}$ , where  $\sqrt{N}S_N$ ,  $\sigma_{ijkl} \equiv E(z_{ni}z_{nj}z_{nk}z_{nl})$  and  $\sigma_{ij} \equiv E(z_{ni}z_{nj})$ . An asymptotically optimal  $W$  is given by letting  $W^{-1}$  have elements  $s_{ijkl} - s_{ij}s_{kl}$ . This estimator is called the *asymptotically distribution free* (ADF) estimator (Browne, 1984), denoted by  $\hat{\theta}_{\text{ADF}}$  (although in EQS this is called AGLS and in LISREL it is called WLS). The asymptotic distribution of the ADF estimator is given by

$$\sqrt{N}(\hat{\theta} - \theta) \xrightarrow{d} \mathcal{N}(0, (\Delta'Y^{-1}\Delta)^{-1}),$$

where  $\Delta \equiv \partial\sigma/\partial\theta'$ , evaluated in the true value of  $\theta$ . The asymptotic covariance matrix can be consistently estimated by evaluating  $\Delta$  in  $\hat{\theta}$  and inserting  $W$  for  $Y^{-1}$ . Note that ADF, as well as all estimators discussed before, are all special cases of *generalized method of moments* (GMM) estimators, see Hall (2001).

## MODEL FIT

If a structural equation model has been estimated, it is important to assess the *fit* of the model, i.e. whether the model and the data agree. Many statistics have been proposed for assessing model fit. Most of these are functions of  $\hat{F} = F(\hat{\theta})$ , where  $F$  denotes the function (8.21) that is minimized. In this section, it is assumed that  $\text{plim } W^{-1} = Y$ . The statistic most frequently used is the *chi-square statistic*  $\chi^2 \equiv N\hat{F}$ , which is a formal test statistic for the null hypothesis that the model is correct in the population against the alternative hypothesis that the model is not correct in the population. Under the null hypothesis, this test statistic converges to a chi-square variate with  $\text{df} \equiv p^* - q$  degrees of freedom, where  $p^*$  is the number of elements of  $\sigma(\theta)$  and  $q$  is the number of elements of  $\theta$ .

In practice, however, models are obviously rarely entirely correct in the population. For the GLS estimators,  $\hat{F}$  converges to  $F_+ = (\sigma_+ - \sigma(\theta_+))'Y^{-1}(\sigma_+ - \sigma(\theta_+))$  for some  $\theta_+$ , where  $\sigma_+ = \text{plim } s_N$ . If the model is correct in the population,  $\sigma_+ = \sigma(\theta_+)$  and  $F_+ = 0$ . If the model is not entirely correct in the population,  $\sigma_+ \neq \sigma(\theta_+)$  and  $F_+ > 0$ . Hence,  $\chi^2 \rightarrow NF_+ \rightarrow +\infty$ . This illustrates the empirical finding that for large sample sizes, nonsaturated models (i.e. models with  $\text{df} > 0$ ) tend to be rejected, although they may describe the data very well. Therefore, alternative measures of fit have been developed. The quality of the model may be defined by the quantity

$$\frac{F_0 - F_1}{F_0}, \quad (8.22)$$

where  $F_0$  is defined similar to  $F_+$ , but for a highly restrictive baseline model or *null model* and  $F_1$  is  $F_+$  for the *target model*. It is customary to use the independence model, in which all variables are assumed to be independently distributed, as the

null model. Clearly, (8.22) is very similar to  $R^2$ . It is always between zero and one, higher values indicating better fit. It may be estimated by the (Bentler-Bonett) *normed fit index*  $NFI = (\hat{F}_0 - \hat{F}_1)/\hat{F}_0$ . The NFI has been widely used since its introduction by Bentler and Bonett (1980).

However, simulation studies and theoretical derivations have shown that NFI is biased in finite samples and that its mean is generally an increasing function of  $N$ . By approximating the distribution of  $N\hat{F}$  by a noncentral chi-square distribution, a better estimator of (8.22) has been derived. This is the *relative noncentrality index* (RNI)

$$RNI \equiv \frac{\hat{\delta}_0 - \hat{\delta}_1}{\hat{\delta}_0},$$

where  $\hat{\delta}_i \equiv \hat{F}_i - df_i/N$  (McDonald and Marsh, 1990). A disadvantage of RNI is that it is not necessarily between zero and one, although usually it is. This disadvantage is overcome by the *comparative fit index* (CFI; Bentler, 1990), which is generally equal to the RNI, but if  $RNI > 1$ ,  $CFI = 1$ , and if  $RNI < 0$ ,  $CFI = 0$ , provided  $\hat{\delta}_0 > 0$ , which is usually the case.

## Notes

- \* The authors would like to thank Anne Boomsma, Bart Boon, Jos ten Berge, Michel Wedel, and an anonymous referee for their helpful comments on an earlier version of this paper.
- 1 Strictly speaking, this violates the iid assumptions used in this chapter. It would be theoretically better to specify the model with nonzero means and intercepts. The practical consequences of this violation are, however, negligible, whereas the formulas are considerably less complicated. Therefore, in this chapter we ignore the resulting theoretical subtleties.

## References

- Aigner, D.J., C. Hsiao, A. Kapteyn, and T.J. Wansbeek (1984). Latent variable models in econometrics. In Z. Griliches and M.D. Intriligator (eds.) *Handbook of Econometrics, Volume 2*, pp. 1321–93. Amsterdam: North-Holland.
- Alonso-Borrego, C., and M. Arellano (1999). Symmetrically normalized instrumental-variable estimation using panel data. *Journal of Business & Economic Statistics* 17, 36–49.
- Angrist, J.D., G.W. Imbens, and A.B. Krueger (1999). Jackknife instrumental variables estimation. *Journal of Applied Econometrics* 14, 57–67.
- Arbuckle, J.L. (1997). *Amos User's Guide*. Version 3.6. Chicago: Smallwaters.
- Baltagi, B.H. (1995). *Econometric Analysis of Panel Data*. Chichester: Wiley.
- Bekker, P.A. (1986). Comment on identification in the linear errors in variables model. *Econometrica* 54, 215–17.
- Bekker, P.A. (1994). Alternative approximations to the distributions of instrumental variable estimators. *Econometrica* 62, 657–81.
- Bekker, P.A., P. Doppelstein, and T.J. Wansbeek (1996). The APT model as reduced rank regression. *Journal of Business & Economic Statistics* 14, 199–202.

- Bekker, P.A., A. Kapteyn, and T.J. Wansbeek (1984). Measurement error and endogeneity in regression: bounds for ML and 2SLS estimates. In T.K. Dijkstra (ed.) *Misspecification Analysis*. pp. 85–103. Berlin: Springer.
- Bekker, P.A., A. Kapteyn, and T.J. Wansbeek (1987). Consistent sets of estimates for regressions with correlated or uncorrelated measurement errors in arbitrary subsets of all variables. *Econometrica* 55, 1223–30.
- Bekker, P.A., A. Merckens, and T.J. Wansbeek (1994). *Identification, Equivalent Models, and Computer Algebra*. Boston: Academic Press.
- Bekker, P.A., T.J. Wansbeek, and A. Kapteyn (1985). Errors in variables in econometrics: New developments and recurrent themes. *Statistica Neerlandica* 39, 129–41.
- Bentler, P.M. (1990). Comparative fit indexes in structural models. *Psychological Bulletin* 107, 238–46.
- Bentler, P.M. (1995). *EQS Structural Equations Program Manual*. Encino, CA: Multivariate Software.
- Bentler, P.M., and D.G. Bonett (1980). Significance tests and goodness of fit in the analysis of covariance structures. *Psychological Bulletin* 88, 588–606.
- Bentler, P.M., and P. Dudgeon (1996). Covariance structure analysis: Statistical practice, theory, and directions. *Annual Review of Psychology* 47, 563–92.
- Biørn, E. (1992a). The bias of some estimators for panel data models with measurement errors. *Empirical Economics* 17, 51–66.
- Biørn, E. (1992b). Panel data with measurement errors. In L. Mátyás and P. Sevestre (eds.) *The Econometrics of Panel Data*. Dordrecht: Kluwer.
- Bollen, K.A. (1989). *Structural Equations with Latent Variables*. New York: Wiley.
- Bound, J., D.A. Jaeger, and R.M. Baker (1995). Problems with instrumental variables estimation when the correlation between the instruments and the endogenous explanatory variable is weak. *Journal of the American Statistical Association* 90, 443–50.
- Bowden, R.J., and D.A. Turkington (1984). *Instrumental Variables*. Cambridge, UK: Cambridge University Press.
- Browne, M.W. (1984). Asymptotically distribution-free methods for the analysis of covariance structures. *British Journal of Mathematical and Statistical Psychology* 37, 62–83.
- Cheng, C.-L., and J.W. Van Ness (1999). *Statistical Regression with Measurement Error*. London: Arnold.
- Cragg, J.G., and S.G. Donald (1997). Inferring the rank of a matrix. *Journal of Econometrics* 76, 223–50.
- Erickson, T. (1993). Restricting regression slopes in the errors-in-variables model by bounding the error correlation. *Econometrica* 61, 959–69.
- Fuller, W.A. (1987). *Measurement Error Models*. New York: Wiley.
- Goldberger, A.S. (1984a). Redirecting reverse regression. *Journal of Business & Economic Statistics* 2, 114–16.
- Goldberger, A.S. (1984b). Reverse regression and salary discrimination. *The Journal of Human Resources* 19, 293–319.
- Griliches, Z. (1986). Economic data issues. In Z. Griliches and M.D. Intriligator (eds.) *Handbook of Econometrics, Volume 3*. Amsterdam: North-Holland.
- Griliches, Z., and J.A. Hausman (1986). Errors in variables in panel data. *Journal of Econometrics* 32, 93–118.
- Hall, A.R. (2001). Generalized method of moments. In B.H. Baltagi (ed.) *A Companion to Theoretical Econometrics*. Oxford: Blackwell Publishing. (this volume)
- Hoyle, R. (ed.). (1995). *Structural Equation Modeling: Concepts, Issues, and Applications*. Thousand Oaks, CA: Sage.

- Jöreskog, K.G., and A.S. Goldberger (1975). Estimation of a model with multiple indicators and multiple causes of a single latent variable. *Journal of the American Statistical Association* 70, 631–9.
- Jöreskog, K.G., and D. Sörbom (1996). *LISREL 8 User's Reference Guide*. Chicago: Scientific Software International.
- Kapteyn, A., and T.J. Wansbeek (1984). Errors in variables: Consistent Adjusted Least Squares (CALIS) estimation. *Communications in Statistics – Theory and Methods* 13, 1811–37.
- Manski, C.F. (1995). *Identification Problems in the Social Sciences*. Cambridge, MA: Harvard University Press.
- McDonald, R.P., and H.W. Marsh (1990). Choosing a multivariate model: Noncentrality and goodness of fit. *Psychological Bulletin* 107, 247–55.
- Merckens, A., and P.A. Bekker (1993). Identification of simultaneous equation models with measurement error: A computerized evaluation. *Statistica Neerlandica* 47, 233–44.
- Muthén, B.O., and L.K. Muthén (1998). *Mplus User's Guide*. Los Angeles: Muthén & Muthén.
- Nelson, C.R., and R. Startz (1990a). Some further results on the exact small sample properties of the instrumental variables estimator. *Econometrica* 58, 967–76.
- Nelson, C.R., and R. Startz (1990b). The distribution of the instrumental variables estimator and its *t*-ratio when the instrument is a poor one. *Journal of Business* 63, 125–40.
- Poirier, D.J. (1998). Revising beliefs in nonidentified models. *Econometric Theory* 14, 483–509.
- Reiersøl, O. (1950). Identifiability of a linear relation between variables which are subject to error. *Econometrica* 18, 375–89.
- Reinsel, G.C., and R.P. Velu (1998). *Multivariate Reduced Rank Regression: Theory and Applications*. New York: Springer.
- Staiger, D., and J.H. Stock (1997). Instrumental variables regression with weak instruments. *Econometrica* 65, 557–86.
- Ten Berge, J.M.F. (1993). *Least Squares Optimization in Multivariate Analysis*. Leiden, The Netherlands: DSWO Press.
- Van Montfort, K., A. Mooijaart, and J. De Leeuw (1987). Regression with errors in variables: Estimators based on third order moments. *Statistica Neerlandica* 41, 223–39.
- Wald, A. (1940). The fitting of straight lines if both variables are subject to error. *Annals of Mathematical Statistics* 11, 284–300.
- Wansbeek, T.J. (1989). Permutation matrix – II. In S. Kotz and N.L. Johnson (eds.) *Encyclopedia of Statistical Sciences, Supplement Volume*. pp. 121–2. New York: Wiley.
- Wansbeek, T.J., and R.H. Koning (1991). Measurement error and panel data. *Statistica Neerlandica* 45, 85–92.
- Wooldridge, J.M. (1996). Estimating systems of equations with different instruments for different equations. *Journal of Econometrics* 74, 387–405.

CHAPTER NINE

# Diagnostic Testing

*Jeffrey M. Wooldridge\**

## 1 INTRODUCTION

Diagnostic testing has become an integral part of model specification in econometrics, with several important advances over the past 20 years. Some of these advances involve new insights into older diagnostics, such as the Durbin–Watson (1950) statistic, Ramsey's (1969) regression error specification test (RESET), general Lagrange multiplier (LM) or score statistics, and White's (1982, 1994) information matrix (IM) test. Other advances have focused on the implicit alternatives of a diagnostic test (Davidson and MacKinnon, 1987) and the related topic of the robustness of diagnostic tests to auxiliary assumptions – that is, assumptions that are imposed under the null but which are not being tested (Wooldridge, 1990a, 1991a, 1991b).

How does diagnostic testing, which is also called *specification testing* (and sometimes *misspecification testing*), differ from classical testing? The difference is arguably a matter of perspective. In classical hypothesis testing we typically assume that we have, in an appropriate sense, a correctly specified parametric model. This could either be a model for a conditional mean, a conditional median, or a fully specified conditional distribution. We then use standard statistics, such as Wald, Lagrange multiplier, and likelihood (or quasi-likelihood) ratio statistics to test restrictions on the parameters.

With diagnostic testing our concern is with testing our maintained model for various misspecifications. We can do this by nesting the model within a more general model, testing it against a nonnested alternative, or using an omnibus test intended to detect a variety of misspecifications.

The purpose of this chapter is to develop diagnostic testing from a modern perspective, omitting some of the intricacies of large sample theory. Rather than considering a general, abstract setting and a unified approach to diagnostic testing – such as in Newey (1985), Tauchen (1985), Pagan and Vella (1989), Wooldridge (1990a), and White (1994) – I instead consider diagnostic testing in relatively simple settings that nevertheless arise often in applied work.

It is useful to organize our thinking on diagnostic testing before jumping into any analysis. The following questions arise, and, ideally, are answered, when devising and analyzing a diagnostic test:

1. What null hypothesis is the statistic intended to test?
2. What auxiliary assumptions are maintained under  $H_0$ ? If an auxiliary assumption fails, does it cause the test to simply have the wrong null limiting distribution, or does the statistic reject with probability approaching one as the sample size grows?
3. Against which alternatives is the test consistent? (That is, against which alternatives does the test have unit asymptotic power?)
4. Against which local alternatives does the test have asymptotic power greater than asymptotic size?
5. Are unusual regularity conditions needed for the test to have a standard (i.e. normal or chi-square) limiting distribution under  $H_0$ ?
6. In problems with a time dimension, what do we need to assume about stability and weak dependence of the time series processes? For example, when are unit roots allowed? How does the presence or absence of unit roots and cointegrating relationships affect the limiting distributions?

This chapter is broken down by the two most popular kinds of data sets used in econometrics. Section 2 covers cross section applications and Section 3 treats time series. Section 4 contains some concluding remarks, including a brief summary of some topics we did not discuss, some of which are fruitful areas for further research.

## 2 DIAGNOSTIC TESTING IN CROSS SECTION CONTEXTS

To obtain a unified view of diagnostic testing, it is important to use a modern perspective. This requires facility with concepts from basic probability, such as conditional expectations and conditional variances, and application of tools such as the law of iterated expectations. While sloppiness in stating assumptions is often innocuous, in some cases it is not. In the following sections we show how to properly state the null hypothesis for diagnostic testing in cross section applications.

### 2.1 Diagnostic tests for the conditional mean in the linear regression model

We begin with the standard linear regression model because it is still the work-horse in empirical economics, and because it provides the simplest setting for a modern approach to diagnostic testing. Even in this basic setup, we must be careful in stating assumptions. The following statement, or a slight variant on it, appears in numerous textbooks and research papers:

Consider the model

$$y_i = \beta_0 + x_i\beta + u_i, \quad i = 1, 2, \dots, N, \quad (9.1)$$

where  $x_i$  is a  $1 \times k$  vector of explanatory variables and the  $u_i$  are iid zero-mean errors.

This formulation may be so familiar that we do not even think to question it. Unfortunately, for econometric analysis, the statement of the model and assumptions is almost useless, as it omits the most important consideration: What is the relationship between the error,  $u_i$ , and the explanatory variables,  $x_i$ ? If we assume random sampling, the errors *must* be independent and identically distributed because  $u_i$  is a function of  $y_i$  and  $x_i$ . But this tells us nothing of value for estimating  $\beta$ .

Often, the explanatory variables in (9.1) are assumed to be fixed or nonrandom. Then we can obtain an unbiased estimator of  $\beta$  because the model satisfies the Gauss–Markov assumptions. But assuming fixed regressors assumes away the interesting problems that arise with analyzing nonexperimental data.

What is a better model formulation? For most cross section applications it is best to start with a *population model*, which in the linear regression case is written as

$$y = \beta_0 + x\beta + u, \quad (9.2)$$

where  $x$  is a  $1 \times k$  vector of explanatory variables and  $u$  is the error. If we supplement this model with random sampling – which, for econometric applications with cross section data, is more realistic than the fixed regressor assumption – we can forget about the realizations of the random variables in (9.2) and focus entirely on the assumptions in the population.

In order to consistently estimate  $\beta$  by OLS (ordinary least squares), we need to make assumptions about the relationship between  $u$  and  $x$ . There are several possibilities. The weakest useful assumptions are

$$E(u) = 0 \quad (9.3)$$

$$E(x'u) = 0, \quad (9.4)$$

where we assume throughout that all expectations are well-defined.

Assumption (9.3) is for free because we have included an intercept in the model. Assumption (9.4) is equivalent to assuming  $u$  is uncorrelated with each  $x_j$ . Under (9.3), (9.4), random sampling, and the assumption that  $\text{var}(x)$  has full rank – i.e. there is no perfect collinearity in the population – the OLS estimator is consistent and  $\sqrt{N}$ -asymptotically normal for  $\beta_0$  and  $\beta$ .

As a minimal set of assumptions, (9.4) is fine. But it is not generally enough to interpret the  $\beta_j$  as the partial effect of  $x_j$  on the expected value of  $y$ . A stronger assumption is that  $u$  has a zero conditional mean:

$$E(u | x) = 0, \quad (9.5)$$

which implies that the population regression function,  $E(y | x)$ , is linear:

$$E(y | x) = \beta_0 + x\beta. \quad (9.6)$$

Under (9.6), no additional functions of  $x$  appear in a linear regression model. More formally, if for any  $1 \times h$  function  $g(x)$  we write

$$y = \beta_0 + x\beta + g(x)\gamma + u,$$

and (9.5) holds, then  $\gamma = 0$ . Tests for functional form always maintain (9.6) as the null hypothesis. If we assume only (9.3) and (9.4), there is nothing to test: by definition,  $\beta_0$  and  $\beta$  appear in the linear projection of  $y$  on  $x$ .

In testing for omitted variables,  $z$ , that are not exact functions of  $x$ , the null hypothesis in (9.2) is

$$E(u | x, z) = 0, \quad (9.7)$$

which is equivalent to

$$E(y | x, z) = E(y | x) = \beta_0 + x\beta. \quad (9.8)$$

The first equality in equation (9.8) has a very natural interpretation: once  $x$  has been controlled for,  $z$  has no effect on the mean value of  $y$ . The most common way to test (9.8) is to specify the extended model

$$y = \beta_0 + x\beta + z\gamma + u \quad (9.9)$$

and to test  $H_0 : \gamma = 0$ .

Similar considerations arise when we test a maintained model against a nonnested alternative. One use of nonnested tests is to detect misspecified functional form. A traditional way of specifying competing models that are linear in parameters is

$$y = \beta_0 + a(x)\beta + u \quad (9.10)$$

and

$$y = \gamma_0 + h(x)\gamma + v, \quad (9.11)$$

where  $a(x)$  and  $h(x)$  are row-vectors of functions of  $x$  that may or may not contain elements in common and can be of different dimensions. For example, in (9.10), all explanatory variables may be in level or quadratic form and in (9.11) some or all may be in logarithmic form. When we take the null to be (9.10), assumption (9.5) is the weakest assumption that makes sense: we are testing  $H_0 : E(y | x) = \beta_0 + a(x)\beta$  against the alternative  $H_1 : E(y | x) = \gamma_0 + h(x)\gamma$ . Of course, there are different methods of testing  $H_0$  against  $H_1$ , but before we choose a test we must agree on the proper statement of the null.

Stating the proper null requires even more care when using nonnested tests to choose between models with explanatory variables that are not functionally related to one another. If we agree to treat explanatory variables as random

and focus on conditional expectations, no confusion can arise. The traditional approach to specifying the competing models is

$$y = \beta_0 + x\beta + u \quad (9.12)$$

and

$$y = \gamma_0 + z\gamma + v. \quad (9.13)$$

We are thinking of cases where not all elements of  $z$  are functionally related to  $x$ , and vice versa. For example, in an equation to explain student performance,  $x$  contains one set of school and family characteristics and  $z$  contains another set (where there might be some overlap). Specifying (9.12) as the null model gives us nothing to test, as we have assumed nothing about how  $u$  relates to  $x$  and  $z$ . Instead, the null hypothesis is exactly as in equation (9.8): once the elements of  $x$  have been controlled for,  $z$  has no effect on  $y$ . This is the sense in which (9.12) is the correct model, and (9.13) is not. Similarly, if (9.13) is the null model, we are really saying  $E(y|x, z) = E(y|z) = \gamma_0 + z\gamma$ .

It may be tempting to specify the null model as  $E(y|x) = \beta_0 + x\beta$  and the competing model as  $E(y|z) = \gamma_0 + z\gamma$ , but then we have nothing to test. Both of these hypotheses can be true, in which case it makes no sense to test one against the other.

So far, we have said nothing about actually computing conditional mean diagnostics. A common method is based on *variable addition statistics* or *artificial regressions*. (See Davidson and MacKinnon, Chapter 1 this volume, for a survey of artificial regressions.) A general variable addition statistic is obtained by regressing the OLS residuals obtained under the null on the  $x_i$  and some additional test variables. In particular, the statistic is  $N \cdot R_u^2$  from the regression

$$\hat{u}_i \text{ on } 1, x_i, \hat{g}_i, i = 1, 2, \dots, N, \quad (9.14)$$

where  $\hat{u}_i \equiv y_i - \hat{\beta}_0 - x_i\hat{\beta}$  are the OLS residuals and  $\hat{g}_i \equiv g(x_i, z_i, \hat{\delta})$  is a  $1 \times q$  vector of *misspecification indicators*. Notice that  $\hat{g}_i$  is allowed to depend on some estimated *nuisance parameters*,  $\hat{\delta}$ , an  $r \times 1$  vector.

When testing for misspecified functional form in (9.6),  $\hat{g}_i$  depends only on  $x_i$  (and possible nuisance parameter estimates). For example, we can take  $\hat{g}_i = g_i = g(x_i)$ , where  $g(x)$  is a row vector of nonlinear functions of  $x$ . (Squares and cross products are common, but we can use more exotic functions, such as  $g(x) = \exp(x)/ (1 + \exp(x))$ .) For RESET,  $\hat{g}_i$  consists of powers of the OLS fitted values,  $\hat{y}_i = \hat{\beta}_0 + x_i\hat{\beta}$ , usually  $\hat{y}_i^2, \hat{y}_i^3$ , and possibly  $\hat{y}_i^4$ . The Davidson–MacKinnon (1981) test of (9.6) against the nonnested alternative (9.11) takes  $\hat{g}_i$  to be the scalar  $\hat{\gamma}_0 + h(x_i)\hat{y}_i$ , the fitted values from an OLS regression of  $y_i$  on 1,  $h(x_i)$ ,  $i = 1, 2, \dots, N$ . Wooldridge (1992a), obtains the LM test of  $\lambda = 1$  in the more general model  $E(y|x) = (1 + \lambda(\beta_0 + x\beta))^{1/\lambda}$  when  $y \geq 0$ . Then,  $\hat{g}_i = \hat{y}_i \log(\hat{y}_i)$ , assuming that  $\hat{y}_i > 0$  for all  $i$ .

As discussed above, the null hypothesis we are testing is given by (9.6). What auxiliary assumptions are needed to obtain a usable limiting distribution for

$LM \equiv N \cdot R_u^2$ ? (We denote this statistic by  $LM$  because, when we are testing the null model against a more general alternative, it is the most popular form of the LM statistic.) By auxiliary assumptions we do not mean regularity conditions. In this setup, regularity conditions consist of assuming enough finite moments on the elements of  $x_i$ ,  $g_i = g(x_i, \delta)$ , and  $u_i$ , sufficient differentiability of  $g(x, \cdot)$  over the interior of the nuisance parameter space, and  $\sqrt{N}$ -consistency of  $\hat{\delta}$  for  $\delta$  in the interior of the nuisance parameter space. We will say no more about these kinds of assumptions because they are rarely checked.

The key auxiliary assumption when testing for functional form, using the standard  $N\text{-}R^2$  statistic, is homoskedasticity (which, again, is stated in terms of the population):

$$\text{var}(y | x) = \text{var}(u | x) = \sigma^2. \quad (9.15)$$

(Notice how we do not restrict the variance of  $u$  given explanatory variables that do not appear in  $x$ , as no such restrictions are needed.) It follows from Wooldridge (1991a) that, under (9.6) and (9.15),

$$N \cdot R_u^2 \xrightarrow{d} \chi_q^2, \quad (9.16)$$

where we assume that there are no redundancies in  $g(x_i, \delta)$ . (Technically, the population residuals from the population regression of  $g(x, \delta)$  on  $1, x$  must have variance matrix with full rank.)

When we test for omitted variables in equation (9.9),  $\hat{g}_i = z_i$ . If we test (9.8) against the nonnested alternative  $E(y | x, z) = \gamma_0 + z\gamma$ , there are two popular tests. Let  $w$  be the elements of  $z$  not in  $x$ . The  $N\text{-}R^2$  version of the F-statistic proposed by Mizon and Richard (1986) simply takes  $\hat{g}_i = w_i$ . In other words, we consider a composite model that contains all explanatory variables, and then test for joint significance of those variables not in  $x$ . The Davidson–MacKinnon test again takes  $\hat{g}_i$  to be the fitted values from the alternative model. In these cases, the auxiliary assumption under  $H_0$  is

$$\text{var}(y | x, z) = \text{var}(u | x, z) = \sigma^2. \quad (9.17)$$

Note that (9.15) is no longer sufficient.

Rather than use an  $N\text{-}R^2$  statistic, an F-test is also asymptotically valid. We simply obtain the F-statistic for significance of  $\hat{g}_i$  in the artificial model

$$y_i = \beta_0 + x_i\beta + \hat{g}_i\gamma + \text{error}_i. \quad (9.18)$$

The F-statistic is asymptotically valid in the sense that  $q \cdot F \xrightarrow{d} \chi_q^2$  under  $H_0$  and the appropriate homoskedasticity assumption ((9.15) or (9.17)). From now on we focus on the LM form of the statistic.

Interestingly, when we test for omitted variables, the asymptotic result in (9.16) does not generally hold under (9.17) if we only assume  $u$  and  $z$  (and  $u$  and  $x$ ) are uncorrelated under  $H_0$ . To see this, as well as gain other insights into the

asymptotic behavior of the LM statistic, it is useful to sketch the derivation of (9.16). First, straightforward algebra shows that  $LM$  can be written as

$$LM = \left( N^{-1/2} \sum_{i=1}^N \hat{g}'_i \hat{u}_i \right)' \left[ \hat{\sigma}^2 \left( N^{-1} \sum_{i=1}^N \hat{r}'_i \hat{r}_i \right) \right]^{-1} \left( N^{-1/2} \sum_{i=1}^N \hat{g}'_i \hat{u}_i \right), \quad (9.19)$$

where  $\hat{\sigma}^2 = N^{-1} \sum_{i=1}^N \hat{u}_i^2$  and  $\hat{r}_i \equiv \hat{g}_i - \hat{\pi}_0 - x_i \hat{\Pi}$  are the OLS residuals from a multivariate regression of  $\hat{g}_i$  on 1,  $x_i$ ,  $i = 1, 2, \dots, N$ . Equation (9.19) makes it clear that the statistic is based on the sample covariance between  $\hat{g}_i$  and  $\hat{u}_i$ . From (9.19), we see that the asymptotic distribution depends on the asymptotic distribution of

$$N^{-1/2} \sum_{i=1}^N \hat{g}'_i \hat{u}_i.$$

As shown in Wooldridge (1990a, 1991a), under  $H_0$  (either (9.6) or (9.8)),

$$N^{-1/2} \sum_{i=1}^N \hat{g}'_i \hat{u}_i = N^{-1/2} \sum_{i=1}^N r'_i u_i + o_p(1) \quad (9.20)$$

where  $r_i \equiv g_i - \pi_0 - x_i \Pi$  are the population residuals from the population regression of  $g_i$  on 1,  $x_i$ . Under  $H_0$ ,  $E(u_i | r_i) = 0$  (since  $r_i$  is either a function of  $x_i$  or of  $(x_i, z_i)$ ), and so  $E(r'_i u_i) = 0$ . It follows that the second term in (9.20) has a limiting  $q$ -variate normal distribution. Therefore, whether  $LM$  has a limiting chi-square distribution under  $H_0$  hinges on whether  $\hat{\sigma}^2(N^{-1} \sum_{i=1}^N \hat{r}'_i \hat{r}_i)$  is a consistent estimator of  $\text{var}(r'_i u_i) = E(u_i^2 r'_i r_i)$ . By the law of iterated expectations,

$$E(u_i^2 r'_i r_i) = E[E(u_i^2 | r_i) | r'_i r_i] = \sigma^2 E(r'_i r_i), \quad (9.21)$$

where the second equality holds provided  $E(u_i^2 | r_i) = E(u_i^2) = \sigma^2$ . For testing (9.8) under (9.17),  $E(u_i^2 | x_i, z_i) = \text{var}(u_i | x_i, z_i) = \sigma^2$ , and  $r_i$  is a function of  $(x_i, z_i)$ , so  $E(u_i^2 | r_i) = \sigma^2$ .

If we only assume  $E(r'_i u_i) = 0$  – for example,  $E(u) = 0$ ,  $E(x'u) = 0$ , and  $E(z'u) = 0$  in (9.9) – then  $E(u^2 | r)$  and  $\text{var}(u | r)$  are not necessarily the same, and (9.17) is no longer enough to ensure that  $LM$  has an asymptotic chi-square distribution.

We can also use (9.19) and (9.20) to see how  $LM$  behaves when the conditional mean null hypothesis holds but the auxiliary homoskedasticity assumption does not. An important point is that the representation in (9.20) holds with or without homoskedasticity, which implies that  $LM$  has a well-defined limiting distribution even if the conditional variance is not constant. Therefore, the rejection frequency tends to some number strictly less than one (typically, substantially below one), which means that a diagnostic test for the conditional mean has no systematic power to detect heteroskedasticity. Intuitively, it makes sense that a conditional mean test makes a poor test for heteroskedasticity. However, some authors have

claimed that conditional mean diagnostics, such as RESET, have the ability to detect heteroskedasticity; the previous argument shows this simply is not true.

Without (9.21), the matrix in the quadratic form is not a consistent estimator of  $\text{var}(\mathbf{r}'_i \hat{u}_i)$ , and so the limiting distribution of  $LM$  is not chi-square. The resulting test based on chi-square critical values may be asymptotically undersized or oversized, and it is difficult to know which is the case.

It is now fairly well known how to adjust the usual LM statistic to allow for heteroskedasticity of unknown form under  $H_0$ . A computationally simple regression-based test has its roots in the Messer and White (1984) method for obtaining the heteroskedasticity-robust variance matrix of the OLS estimator, and was first proposed by Davidson and MacKinnon (1985). Subsequently, it was shown to be valid quite generally by Wooldridge (1990a). The simplest way to derive a heteroskedasticity-robust test is to note that a consistent estimator of  $\text{var}(\mathbf{r}'_i \hat{u}_i)$ , with or without homoskedasticity, is  $N^{-1} \sum_{i=1}^N \hat{u}_i^2 \hat{\mathbf{r}}'_i \hat{\mathbf{r}}_i$ . A useful algebraic fact is

$$\sum_{i=1}^N \hat{\mathbf{g}}'_i \hat{u}_i = \sum_{i=1}^N \hat{\mathbf{r}}'_i \hat{u}_i,$$

because the  $\hat{\mathbf{r}}_i$  are residuals from an OLS regression of  $\hat{\mathbf{g}}_i$  on 1,  $\mathbf{x}_i$  and  $\mathbf{x}_i$  is orthogonal to  $\hat{u}_i$  in sample. Therefore, the robust LM statistic is

$$\left( N^{-1/2} \sum_{i=1}^N \hat{\mathbf{r}}'_i \hat{u}_i \right)' \left( N^{-1} \sum_{i=1}^N \hat{u}_i^2 \hat{\mathbf{r}}'_i \hat{\mathbf{r}}_i \right)^{-1} \left( N^{-1/2} \sum_{i=1}^N \hat{\mathbf{r}}'_i \hat{u}_i \right). \quad (9.22)$$

As a computational device, this can be obtained as  $N \cdot R_0^2 = N - \text{SSR}_0$  from the regression

$$1 \text{ on } \hat{u}_i \hat{\mathbf{r}}_i, \quad i = 1, 2, \dots, N, \quad (9.23)$$

where  $R_0^2$  is now the uncentered  $R^2$  and  $\text{SSR}_0$  is the usual sum of squared residuals. Under (9.6) or (9.8), the heteroskedasticity-robust LM statistic has an asymptotic  $\chi_q^2$  distribution.

As we mentioned in the introduction, for any diagnostic test it is important to know the alternatives against which it is consistent. Before we leave this subsection, we provide an example of how the previous tools can be used to shed light on conflicting claims about specification tests in the literature. Ramsey's RESET has often been touted as a general diagnostic that can detect, in addition to functional form problems, omitted variables. (See, e.g., Thursby (1979, 1989) and Godfrey (1988, section 4.2.2).) In fact, RESET, or any other test where the misspecification indicators are functions of  $\mathbf{x}_i$  (and possibly nuisance parameters) make poor tests for omitted variables. To see why, suppose that  $E(y | \mathbf{x}, q) = \beta_0 + \mathbf{x}\beta + \gamma q$ , where  $\gamma \neq 0$ . We start with this model to emphasize that we are interested in the partial effect of each  $x_j$ , holding  $q$ , and the other elements of  $\mathbf{x}$ , fixed. Now, suppose that  $q$  is not observed. If  $q$  is correlated with one or more elements of  $\mathbf{x}$ , the OLS regression  $y$  on 1,  $\mathbf{x}$ , using a random sample of data, is biased and

inconsistent for the  $\beta_j$ . What happens if we apply RESET or some other functional form test? Suppose that  $q$  has a linear conditional expectation given  $x$ :  $E(q|x) = \pi_0 + x\pi$ . This is certainly the leading case; after all, we started with a linear model for  $E(y|x, q)$ . Then, by iterated expectations,

$$\begin{aligned} E(y|x) &= E[E(y|x, q)|x] = \beta_0 + x\beta + \gamma E(q|x) \\ &= (\beta_0 + \gamma\pi_0) + x(\beta + \gamma\pi) \equiv \theta_0 + x\theta. \end{aligned}$$

In other words, regardless of the size of  $\gamma$  or the amount of correlation between  $q$  and  $x$ ,  $E(y|x)$  is linear in  $x$ , and so RESET generally converges in distribution to a quadratic form in a normal random vector. This means it is inconsistent against the omitted variable alternative. If  $\text{var}(y|x)$  is constant, as is the case if  $\text{var}(y|x, q)$  is constant and  $\text{var}(q|x)$  is constant, then RESET has a limiting chi-square distribution: its asymptotic power is equal to its asymptotic size. RESET would only detect the omission of  $q$  if  $E(q|x)$  is nonlinear. But then we could never distinguish between  $\gamma \neq 0$  and  $E(y|x)$  being nonlinear in  $x$ .

Finally, we can compare (9.19) and (9.22) to see that, when the homoskedasticity assumption holds, the statistics have the same limiting distribution under local alternatives. (See Davidson and MacKinnon (1993, section 12.2) for a discussion of local alternatives.) The statistics only differ in the  $q \times q$  matrix in the quadratic form. Under homoskedasticity, these both converge in probability to  $\sigma^2 E(r'r)$  under local alternatives. See Wooldridge (1990a) for a more general discussion.

## 2.2 Testing for heteroskedasticity

If we only care about the conditional mean, it makes sense to make conditional mean diagnostics robust to arbitrary forms of heteroskedasticity. (Just as it is very common now in microeconometric studies to report heteroskedasticity-robust standard errors and  $t$ -statistics.) Some researchers prefer to compute nonrobust conditional mean diagnostics and then later test the homoskedasticity assumption. There are some drawbacks to such a strategy. If there is heteroskedasticity, then the conditional mean tests were carried out at the incorrect size, and so the overall size of the testing procedure is difficult to determine. Plus, the standard regression-based tests for heteroskedasticity impose their own auxiliary assumptions.

Wooldridge (1991a) suggests first testing the conditional mean, making any conditional mean diagnostics robust to arbitrary heteroskedasticity. Then, if we are still interested in knowing whether homoskedasticity is violated, we can test that subsequently.

An alternative to sequentially testing the mean and variance is to test them jointly. Devising such a test is a good illustration of the role auxiliary assumptions play in obtaining a valid test statistic. For concreteness, consider a joint test of (9.6) and (9.15). (If we were testing for additional variables in the mean or variance function, the analysis would be essentially the same.) A general class of diagnostics is based on sample averages of the form  $N^{-1}\sum_{i=1}^N \hat{g}_i'\hat{u}_i$  and  $N^{-1}\sum_{i=1}^N \hat{h}_i'(\hat{u}_i^2 - \hat{\sigma}^2)$ ,

where the first of these is familiar from the conditional mean tests in Section 2.1, and the second is intended to test (9.15). Like  $\hat{g}_i$ ,  $\hat{h}_i$  is a function of  $x_i$  and possibly some nuisance parameter estimates. For the Breusch–Pagan (1979) test for heteroskedasticity,  $\hat{h}_i = x_i$ . For the White (1980) test,  $\hat{h}_i$  consists of  $x_i$  and all nonredundant squares and cross products of elements of  $x_i$ . A useful special case of the White test is when  $\hat{h}_i = (\hat{y}_i, \hat{y}_i^2)$ , where the  $\hat{y}_i$  are the OLS fitted values. For concreteness, let  $\hat{g}_i$  be  $1 \times q$  and let  $\hat{h}_i$  be  $1 \times p$ . Following Wooldridge (1991a), we can show that, under (9.6) and (9.15), equation (9.20) holds and

$$N^{-1/2} \sum_{i=1}^N \hat{h}'_i (\hat{u}_i^2 - \hat{\sigma}^2) = N^{-1/2} \sum_{i=1}^N (h_i - \mu_h)' (u_i^2 - \sigma^2) + o_p(1), \quad (9.24)$$

where  $\mu_h = E(h_i)$ . Therefore, we need to obtain the limiting distribution of

$$N^{-1/2} \begin{pmatrix} \sum_{i=1}^N r'_i u_i \\ \sum_{i=1}^N c'_i v_i \end{pmatrix}, \quad (9.25)$$

where  $c_i \equiv (h_i - \mu_h)$  and  $v_i \equiv (u_i^2 - \sigma^2)$ . The limiting distribution is, of course, multivariate normal; what we need is to obtain the  $(q+p) \times (q+p)$  variance–covariance matrix. Under (9.6) and (9.15), this is easily done. We already obtained the upper  $q \times q$  block:  $\sigma^2 E(r'_i r_i)$  (because we are maintaining homoskedasticity under  $H_0$ ). The lower  $p \times p$  block is simply

$$E(v_i^2 c'_i c_i) = E((u_i^2 - \sigma^2)^2 c'_i c_i). \quad (9.26)$$

The  $q \times p$  upper right block is

$$E(u_i v_i r'_i c_i) = E(u_i^3 r'_i c_i), \quad (9.27)$$

where we use the fact that  $E(u_i | x_i) = 0$  under (9.6) and  $r_i$  and  $c_i$  are both functions of  $x_i$ .

From (9.26) and (9.27), we see immediately that there are some convenient simplifications if we make standard auxiliary assumptions. First, if we assume *conditional symmetry*, then  $E(u_i^3 | x_i) = 0$ , and so  $E(u_i^3 r'_i c_i) = 0$  by the usual iterated expectations argument. This means that the variance–covariance matrix of (9.25) is block diagonal, which has the important implication that the conditional mean and conditional variance tests are asymptotically independent. If our estimate of the asymptotic variance imposes block diagonality, the joint test statistic is simply the sum of the conditional mean and conditional variance statistics.

If we impose the *homokurtosis* assumption, that is,  $E(u^4 | x) = \tau^2$ , then  $E((u^2 - \sigma^2)^2 | x)$  is constant, say  $\kappa^2$ , and the diagnostic for heteroskedasticity simplifies. A consistent estimator of the lower  $p \times p$  block of the asymptotic variance

matrix is  $\hat{\kappa}^2(N^{-1}\sum_{i=1}^N(\hat{h}_i - \bar{h})'(\hat{h}_i - \bar{h}))$ , where  $\bar{h}$  is the sample average of  $\hat{h}_i$  and  $\hat{\kappa}^2 = N^{-1}\sum_{i=1}^N(\hat{u}_i^2 - \hat{\sigma}^2)^2$ . The test for heteroskedasticity can be computed as  $N \cdot R_v^2$  from the regression

$$\hat{u}_i^2 \text{ on } 1, \hat{h}_i, \quad i = 1, 2, \dots, N, \quad (9.28)$$

where  $R_v^2$  is the usual  $R^2$ . Under (9.6), (9.15), and the auxiliary homokurtosis assumption,  $N \cdot R_v^2 \stackrel{a}{\sim} \chi_p^2$ . (This is true whether or not the symmetry condition holds.)

We can easily relax the homokurtosis assumption in testing for heteroskedasticity. As shown in Wooldridge (1991a), or as can be derived from the above representations, a robust test for heteroskedasticity is obtained as  $N \cdot R_1^2 = N - \text{SSR}_1$  from the regression

$$1 \text{ on } (\hat{u}_i^2 - \hat{\sigma}^2)(\hat{h}_i - \bar{h}), \quad i = 1, 2, \dots, N, \quad (9.29)$$

where  $R_1^2$  is the uncentered  $R^2$  and  $\text{SSR}_1$  is the sum of squared residuals. Under (9.6) and (9.15),  $N \cdot R_1^2 \stackrel{a}{\sim} \chi_p^2$ .

In some cases, we may want to test  $\text{var}(y | x, z) = \sigma^2$ , where  $z$  is an additional set of variables that does not show up in the mean equation. However, the test for heteroskedasticity only makes sense if  $E(y | x, z) = E(y | x)$ . The misspecification indicator  $\hat{h}_i$  would depend on  $z_i$  as well as on  $x_i$ . When  $E(y | x, z) = E(y | x)$ , heteroskedasticity that depends only on  $z$  does not invalidate the usual OLS inference procedures.

Now we can combine various sets of auxiliary assumptions to find when different statistics for testing the joint null are valid. If (9.6), (9.15), symmetry, and homokurtosis all hold, then

$$N \cdot R_u^2 + N \cdot R_v^2 \sim \chi_{q+p}^2. \quad (9.30)$$

Notice that assuming  $u$  is independent of  $x$  and normally distributed is sufficient, but not necessary, for the auxiliary assumptions. (See Bera and Jarque (1982) for a similar test under normality.) If we maintain symmetry but relax homokurtosis,  $N \cdot R_u^2 + N \cdot R_1^2 \stackrel{a}{\sim} \chi_{q+p}^2$ ; this appears to be a new result.

If we relax symmetry, there are no simple versions of the joint test statistic because the mean and variance tests are asymptotically correlated. To obtain a fully robust test – that is, a test that maintains only (9.6) and (9.15) under  $H_0$  – we need to obtain the quadratic form with a general asymptotic variance estimator. For example, the upper  $q \times p$  off-diagonal block can be estimated as  $N^{-1}\sum_{i=1}^N \hat{u}_i^2 \hat{r}'_i(\hat{h}_i - \bar{h})$ .

A joint test of the conditional mean and conditional variance is an example of an *omnibus test*. Such tests must be used with caution because it is difficult to know where to look if a statistic rejects. More importantly, it gives relatively unimportant forms of misspecification, such as the existence of heteroskedasticity, parity with important forms of misspecification, such as misspecification of the conditional mean.

## 2.3 Diagnostic testing in nonlinear models of conditional means

Much of what we discussed in Section 2.1 carries over to nonlinear models. With a nonlinear conditional mean function, it becomes even more important to state hypotheses in terms of conditional expectations; otherwise the null model has no interesting interpretation. For example, if  $y$  is a nonnegative response, a convenient regression function is exponential:

$$E(y|x) = \exp(x\beta), \quad (9.31)$$

where, for simplicity, unity is included in the  $1 \times k$  vector  $x$ . Importantly, (9.31) puts no restrictions on the nature of  $y$  other than that it is nonnegative. In particular,  $y$  could be a count variable, a continuous variable, or a variable with discrete and continuous characteristics. We can construct a variety of alternatives to (9.31). For example, in the more general model

$$E(y|x) = \exp(x\beta + \delta_1(x\beta)^2 + \delta_2(x\beta)^3), \quad (9.32)$$

we can test  $H_0 : \delta_1 = 0, \delta_2 = 0$ .

Generally, if we nest  $E(y|x) = m(x, \beta)$  in the model  $\mu(x, \beta, \delta)$ , where  $m(x, \beta) = \mu(x, \beta, \delta_0)$ , for a known value  $\delta_0$ , then the LM statistic is easy to compute:  $LM = N \cdot R_u^2$  from the regression

$$\hat{u}_i \text{ on } \nabla_\beta \hat{m}_i, \nabla_\delta \hat{\mu}_i, \quad i = 1, 2, \dots, N, \quad (9.33)$$

where  $\hat{u}_i = y_i - m(x_i, \hat{\beta})$ ,  $\nabla_\beta \hat{m}_i = \nabla_\beta m(x_i, \hat{\beta})$ , and  $\nabla_\delta \hat{\mu}_i = \nabla_\delta \mu(x_i, \hat{\beta}, \delta_0)$ ;  $\hat{\beta}$  is the nonlinear least squares (NLS) estimator obtained under  $H_0$ . (The  $R^2$  is generally the uncentered  $R^2$ .) Under  $H_0 : E(y|x) = m(x, \beta)$  and the homoskedasticity assumption (9.15),  $LM \stackrel{a}{\sim} \chi_q^2$ , where  $q$  is the dimension of  $\delta$ . For testing (9.31) against (9.32),  $\nabla_\beta \hat{m}_i = x_i \exp(x_i \hat{\beta})$  and  $\nabla_\delta \hat{\mu}_i = ((x_i \hat{\beta})^2 \exp(x_i \hat{\beta}), (x_i \hat{\beta})^3 \exp(x_i \hat{\beta}))$ ; the latter is a  $1 \times 2$  vector.

Davidson and MacKinnon (1985) and Wooldridge (1991a) show how to obtain a heteroskedasticity-robust form of the LM statistic by first regressing  $\nabla_\delta \hat{\mu}_i$  on  $\nabla_\beta \hat{m}_i$  and obtaining the  $1 \times q$  residuals,  $\hat{r}_i$ , and then computing the statistic exactly as in (9.23).

For more general tests,  $\nabla_\delta \hat{\mu}_i$  is replaced by a set of misspecification indicators, say  $\hat{g}_i$ . For example, if we are testing  $H_0 : E(y|x) = m(x, \beta)$  against  $H_1 : E(y|x) = \mu(x, \gamma)$ , the Davidson–MacKinnon (1981) test takes  $\hat{g}_i = \hat{\mu}_i - \hat{m}_i$ , the difference in the fitted values from the two models. Wooldridge's (1990b) conditional mean encompassing (CME) test takes  $\hat{g}_i = \nabla_\gamma \hat{\mu}_i = \nabla_\gamma \mu(x_i, \hat{\gamma})$ , the  $1 \times q$  estimated gradient from the alternative mean function.

It is straightforward to obtain conditional mean tests in the context of maximum likelihood or quasi-maximum likelihood estimation for densities in the linear exponential family (LEF). As shown by Gouriéroux, Monfort, and Trognon

(1984), the quasi-MLE is consistent and asymptotically normal provided only that the conditional mean is correctly specified. The tests have the same form as those for nonlinear regression, except that all quantities are weighted by the inverse of the estimated conditional standard deviation – just as in weighted nonlinear least squares. So, in (9.33), we would divide each quantity by  $\hat{v}_i^{1/2}$ , where  $\hat{v}_i \equiv v(x_i, \beta)$  is the estimated conditional variance function from the LEF density, under  $H_0$ . For example, in the case of the binary response density,  $\hat{v}_i = \hat{m}_i(1 - \hat{m}_i)$ , where  $\hat{m}_i$  would typically be the logit or probit function. (See Moon (1988) for some examples of alternative conditional mean functions for the logit model.) For Poisson regression,  $\hat{v}_i = \hat{m}_i$ , where  $\hat{m}_i = \exp(x_i \hat{\beta})$  is the usual conditional mean function under  $H_0$  (see Wooldridge (1997) for further discussion). The statistic from (9.33), after quantities have been appropriately weighted, is valid when  $\text{var}(y_i | x_i)$  is proportional to the variance implied by the LEF density. The statistic obtained from (9.23) is robust to arbitrary variance misspecification, provided  $\hat{u}_i$  is replaced with  $\hat{u}_i/\hat{v}_i^{1/2}$  and  $\hat{r}_i$  is replaced with the residuals from the multivariate regression  $\nabla_{\delta}\hat{\mu}_i/\hat{v}_i^{1/2}$  on  $\nabla_{\beta}\hat{m}_i/\hat{v}_i^{1/2}$ ; see Wooldridge (1991b, 1997) for details.

### 3 DIAGNOSTIC TESTING IN TIME SERIES CONTEXTS

All of the tests we covered in Section 2 can be applied in time series contexts. However, because we can no longer assume that the observations are independent of one another, the discussion of auxiliary assumptions under  $H_0$  is more complicated. We assume in this section that the weak law of large numbers and central limit theorems can be applied, so that standard inference procedures are available. This rules out processes with unit roots, or fractionally integrated processes. (See Wooldridge (1994) or Pötscher and Prucha (chapter 10 in this volume) for a discussion of the kinds of dependence allowed.) For notational simplicity, we assume that the process is strictly stationary so that moment matrices do not depend on  $t$ .

#### 3.1 Conditional mean diagnostics

To illustrate the issues that arise in obtaining conditional mean diagnostics for time series models, let  $x_t$  be a vector of conditioning variables, which can contain contemporaneous variables,  $z_t$ , as well as lagged values of  $y_t$  and  $z_t$ . Given  $x_t$ , we may be interested in testing linearity of  $E(y_t | x_t)$ , which is stated as

$$E(y_t | x_t) = \beta_0 + x_t \beta \quad (9.34)$$

for some  $\beta_0 \in \mathbb{R}$  and  $\beta \in \mathbb{R}^k$ . For example,  $y_t$  might be the return on an asset, and  $x_t$  might contain lags of  $y_t$  and lagged economic variables. The same kinds of tests we discussed in Section 2.1 can be applied here, including RESET, the Davidson–MacKinnon test, and LM tests against a variety of nonlinear alternatives. Let  $\hat{g}_t = g(x_t, \hat{\delta})$  be the  $1 \times q$  vector of misspecification indicators. The LM statistic is obtained exactly as in (9.14), with standard notational changes (the  $t$  subscript

replaces  $i$  and the sample size is denoted  $T$  rather than  $N$ ). Just as in the cross section case, we need to assume homoskedasticity conditional on  $x_t$ :

$$\text{var}(y_t | x_t) = \sigma^2. \quad (9.35)$$

If  $x_t$  contains lagged  $y_t$ , this rules out dynamic forms of heteroskedasticity, such as ARCH (Engle, 1982) and GARCH (Bollerslev, 1986), as well as static forms of heteroskedasticity if  $x_t$  contains  $z_t$ .

Because of the serial dependence in time series data, we must add another auxiliary assumption in order for the usual LM statistic to have an asymptotic  $\chi_q^2$  distribution. If we write the model in error form as

$$y_t = \beta_0 + x_t \beta + u_t, \quad (9.36)$$

then a useful auxiliary assumption is

$$E(u_t | x_t, u_{t-1}, x_{t-1}, \dots) = 0. \quad (9.37)$$

Assumption (9.37) implies that  $\{u_t\}$  is serially uncorrelated, but it implies much more. For example,  $u_t$  and  $u_s$  are uncorrelated conditional on  $(x_t, x_s)$ , for  $t \neq s$ . Also,  $u_t$  is uncorrelated with any function of  $(x_t, u_{t-1}, x_{t-1}, \dots)$ .

We can easily see why (9.37) is sufficient, along with (9.35), to apply the usual LM test. As in the cross section case, we can show under (9.34) that (9.20) holds with the obvious changes in notation. Now, for (9.19) to have an asymptotic chi-square distribution, we need  $\hat{\sigma}^2(T^{-1} \sum_{t=1}^T r_t' \hat{r}_t)$  to consistently estimate

$$\lim_{T \rightarrow \infty} \text{var}\left(T^{-1/2} \sum_{t=1}^T r_t' u_t\right). \quad (9.38)$$

Assumption (9.37) ensures that all of the covariance terms in this asymptotic variance are zero. For  $s < t$ ,  $E(u_t u_s r_t' r_s) = E[E(u_t | r_t, u_s, r_s) u_s r_t' r_s] = 0$  because  $E(u_t | r_t, u_s, r_s) = 0$  under (9.37). The last statement follows because  $(r_t, u_s, r_s)$  is a function of  $(x_t, u_{t-1}, x_{t-1}, \dots)$ . When we add the homoskedasticity assumption, we see that the usual  $T$ -R-squared statistic is asymptotically valid.

It is easily seen that (9.37) is equivalent to

$$E(y_t | x_t, y_{t-1}, x_{t-1}, y_{t-2}, \dots) = E(y_t | x_t) = \beta_0 + x_t \beta, \quad (9.39)$$

which we call *dynamic completeness* of the conditional mean. Under (9.39), all of the dynamics are captured by what we have put in  $x_t$ . For example, if  $x_t = (y_{t-1}, z_{t-1})$ , then (9.39) becomes

$$E(y_t | y_{t-1}, z_{t-1}, y_{t-2}, z_{t-2}, \dots) = E(y_t | y_{t-1}, z_{t-1}), \quad (9.40)$$

which means at most one lag each of  $y_t$  and  $z_t$  are needed to fully capture the dynamics. (Because  $z_t$  is not in  $x_t$  in this example, (9.40) places no restrictions on

any contemporaneous relationship between  $y_t$  and  $z_t$ .) Generally, if  $x_t$  contains lags of  $y_t$  and possibly lags of other variables, we are often willing to assume dynamic completeness of the conditional mean when testing for nonlinearities. In any case, we should know that the usual kind of test essentially requires this assumption.

If  $x_t = z_t$  for a vector of contemporaneous variables, (9.39) is very strong:

$$E(y_t | z_t, y_{t-1}, z_{t-1}, y_{t-2}, \dots) = E(y_t | z_t), \quad (9.41)$$

which means that once contemporaneous  $z_t$  has been controlled for, lags of  $y_t$  and  $z_t$  are irrelevant. If we are just interested in testing for nonlinearities in a static linear model for  $E(y_t | z_t)$ , we might not want to impose dynamic completeness.

Relaxing the homoskedasticity assumption is easy: the same heteroskedasticity-robust statistic from regression (9.23) is valid, provided (9.39) holds. This statistic is not generally valid in the presence of serial correlation.

Wooldridge (1991a) discusses different ways to make conditional mean diagnostics robust to serial correlation (as well as to heteroskedasticity). One approach is to obtain an estimator of (9.38) that allows general serial correlation; see, for example, Newey and West (1987) and Andrews (1991). Perhaps the simplest approach is to prewhiten  $\hat{k}_t \equiv \hat{u}_t \hat{r}_t$ , where the  $\hat{u}_t$  are the OLS residuals from estimating the null model and the  $\hat{r}_t$  are the  $1 \times q$  residuals from the multivariate regression of  $\hat{g}_t$  on  $1, x_t, t = 1, 2, \dots, T$ . If  $\hat{e}_t, t = (p+1), \dots, T$  are the  $1 \times q$  residuals from a vector autoregression (VAR) of  $\hat{k}_t$  on  $1, \hat{k}_t, \dots, \hat{k}_{t-p}$ , then the test statistic is

$$\left( \sum_{t=p+1}^T \hat{e}_t \right) \left( \sum_{t=p+1}^T \hat{e}_t' \hat{e}_t \right)^{-1} \left( \sum_{t=p+1}^T \hat{e}_t' \right);$$

under (9.34), the statistic has an asymptotic  $\chi_q^2$  distribution, provided the VAR adequately captures the serial correlation in  $\{\hat{k}_t \equiv u_t r_t\}$ .

We can gain useful insights by studying the appropriate asymptotic representation of the LM statistic. Under (9.34), regularity conditions, and strict stationarity and weak dependence, we can write the  $T-R^2$  LM statistic as

$$LM = \left( T^{-1/2} \sum_{t=1}^T r_t' u_t \right)' \left( \sigma^2 E(r_t' r_t) \right)^{-1} \left( T^{-1/2} \sum_{t=1}^T r_t' u_t \right) + o_p(1). \quad (9.42)$$

This representation does not assume either (9.35) or (9.37), but if either fails then  $\sigma^2 E(r_t' r_t)$  does not generally equal (9.38). This is why, without (9.35) and (9.37), the usual LM statistic does not have a limiting chi-square distribution.

We can use (9.42) to help resolve outstanding debates in the literature. For example, there has long been a debate about whether RESET in a model with strictly exogenous explanatory variables is robust to serial correlation (with homoskedasticity maintained). The evidence is based on simulation studies. Thursby (1979) claims that RESET is robust to serial correlation; Porter and Kashyap (1984)

find that it is not. We can help reconcile the disagreement by studying (9.42). With strictly exogenous regressors,  $\{u_t\}$  is independent of  $\{x_t\}$ , and the  $\{u_t\}$  are always assumed to have a constant variance (typically,  $\{u_t\}$  follows a stable AR(1) model). Combined, these assumptions imply (9.35). Therefore, RESET will have a limiting chi-square distribution when the covariance terms in (9.38) are all zero, that is,

$$E(u_t u_s \mathbf{r}'_t \mathbf{r}_s) = 0, \quad t \neq s. \quad (9.43)$$

When  $\{x_t\}$  is independent of  $\{u_t\}$ ,

$$E(u_t u_s \mathbf{r}'_t \mathbf{r}_s) = E(u_t u_s) E(\mathbf{r}'_t \mathbf{r}_s),$$

because  $\mathbf{r}_t$  is a function of  $x_t$ . Recall that  $\mathbf{r}_t$  is a population residual from a regression that includes an intercept, and so it has zero mean. Here is the key: if  $\{x_t\}$  is an independent sequence, as is often the case in simulation studies, then  $E(\mathbf{r}'_t \mathbf{r}_s) = 0, t \neq s$ . But then (9.43) holds, regardless of the degree of serial correlation in  $\{u_t\}$ . Therefore, if  $\{x_t\}$  is generated to be strictly exogenous and serially independent, RESET is asymptotically robust to arbitrary serial correlation in the errors. (We have also shown that (9.37) is not necessary for the  $T-R^2$  LM statistic to have a limiting chi-square distribution, as (9.37) is clearly false when  $\{u_t\}$  is serially correlated. Instead, strict exogeneity and serial independence of  $\{x_t\}$  are sufficient.)

If  $\{x_t\}$  is serially correlated, the usual RESET statistic is not robust. However, what matters is serial correlation in  $\mathbf{r}_t$ , and this might be small even with substantial serial correlation in  $\{x_t\}$ . For example,  $x_t^2$  net of its linear projection on to  $(1, x_t)$  might not have much serial correlation, even if  $\{x_t\}$  does.

Earlier we emphasized that, in general, the usual LM statistic, or its heteroskedasticity-robust version, maintain dynamic completeness under  $H_0$ . Because dynamic completeness implies that the errors are not serially correlated, serially correlated errors provide evidence against (9.39). Therefore, testing for serial correlation is a common specification test.

The most common form of the LM statistic for  $AR(p)$  serial correlation – see, e.g., Breusch (1978), Godfrey (1978), or Engle (1984) – is  $LM = (T - p)R_u^2$ , where  $R_u^2$  is the usual  $R^2$  from the regression

$$\hat{u}_t \text{ on } 1, x_t, \hat{u}_{t-1}, \hat{u}_{t-2}, \dots, \hat{u}_{t-p}, \quad t = (p+1), \dots, T. \quad (9.44)$$

Under what assumptions is  $LM$  asymptotically  $\chi_q^2$ ? In addition to (9.39) (equivalently, (9.37)), sufficient is the homoskedasticity assumption

$$\text{var}(u_t | x_t, u_{t-1}, \dots, u_{t-p}) = \sigma^2. \quad (9.45)$$

Notice that (9.35) is no longer sufficient; we must rule out heteroskedasticity conditional on lagged  $u_t$  as well.

More generally, we can test for misspecified dynamics, misspecified functional form, or both, by using specification indicators  $g(w_t, \hat{\delta})$ , where  $w_t$  is a subset of

$(x_t, y_{t-1}, x_{t-1}, \dots)$ . If we want to ensure the appropriate no serial correlation assumption holds, we take the null to be (9.39), which implies that  $E(y_t | x_t, w_t) = E(y_t | x_t) = \beta_0 + x_t\beta$ . The homoskedasticity assumption is  $\text{var}(y_t | x_t, w_t) = \sigma^2$ . The adjustment for heteroskedasticity is the same as described for pure functional form tests (see equation (9.23)).

In this section we have focused on a linear null model. Once we specify hypotheses in terms of conditional means, there are no special considerations for nonlinear regression functions with weakly dependent data. All of the tests we discussed for cross section applications can be applied. The statement of homoskedasticity is the same in the linear case, and the dynamic completeness assumption is stated as in (9.39) but with the linear regression function replaced by  $m(x_t, \beta)$ . The standard LM test discussed in Section 2.3 is valid, and both heteroskedasticity and serial correlation robust forms are easily computed (see Wooldridge (1991a) for details).

### 3.2 Testing for heteroskedasticity

As the discussion of testing for serial correlation in Section 3.1 suggests, different kinds of heteroskedasticity have different implications in time series models. Suppose we start with the linear model written in error form as in (9.36). The weakest useful homoskedasticity assumption is

$$H_0 : \text{var}(u_t | x_t) = \sigma^2, \quad (9.46)$$

which only makes sense once we have maintained  $E(u_t | x_t) = 0$ . Then, the conditional mean is correctly specified as  $E(y_t | x_t) = \beta_0 + x_t\beta$ . A popular class of test statistics for heteroskedasticity is based  $T \cdot R_v^2$ , where  $R_v^2$  is the usual  $R^2$  from the regression

$$\hat{u}_t^2 \text{ on } 1, \hat{h}(x_t), \quad t = 1, 2, \dots, T, \quad (9.47)$$

which is exactly the kind of statistic we covered for testing the homoskedasticity assumption in a cross section application. For the  $T \cdot R_v^2$  statistic to be valid, we need the homokurtosis assumption  $E(u_t^4 | x_t) = \kappa^2$ , just as in Section 2.2. However, we need more. While the statistic from (9.47) clearly only has power against violations of (9.46), we need to rule out serial correlation in  $\{u_t^2\}$ . In particular, it suffices that  $E((u_t^2 - \sigma^2)(u_s^2 - \sigma^2) | x_t, x_s) = 0$ ,  $t \neq s$ . Sufficient for this is dynamic completeness of the conditional mean, (9.39), along with the dynamic homoskedasticity assumption

$$\text{var}(y_t | x_t, y_{t-1}, x_{t-1}, \dots) = \sigma^2. \quad (9.48)$$

While (9.48) is not technically necessary, it would be weird to have (9.48) fail but have  $T \cdot R_v^2 \stackrel{\text{a}}{\sim} \chi_q^2$ . Thus, in time series contexts, the usual Breusch-Pagan and White-type tests for heteroskedasticity essentially maintain dynamic completeness of the variance under the null in order to obtain a limiting chi-square statistic.

Sometimes we want to test for dynamic heteroskedasticity even if the conditional mean is static. Consider, for example, (9.41), where the conditional mean model is linear. Then (9.47) becomes  $\text{var}(u_t | z_t) = \sigma^2$ , which does not restrict the variance conditional on past values. It could be that (9.46) holds, which, along with (9.8), means that the usual inference procedures are asymptotically valid. However, we might want to know if  $\text{var}(u_t | z_t, u_{t-1}, z_{t-1}, \dots)$  depends on, say,  $u_{t-1}$ .

With general  $x_t$  under (9.39), Engle's (1982) ARCH model of order one implies

$$\text{var}(u_t | x_t, u_{t-1}, x_{t-1}, \dots) = \text{var}(u_t | u_{t-1}) = \delta_0 + \delta_1 u_{t-1}^2 \quad (9.49)$$

The ARCH( $p$ ) model has  $p$  lags of  $u_t^2$ , and the LM test for ARCH is obtained as  $(T - p)R_v^2$  from the regression

$$\hat{u}_t^2 \text{ on } 1, \hat{u}_{t-1}^2, \dots, \hat{u}_{t-p}^2, \quad t = p + 1, \dots, T. \quad (9.50)$$

As with other tests for heteroskedasticity, this statistic uses an auxiliary homokurtosis assumption, in this case,  $E(u_t^4 | x_t, u_{t-1}, x_{t-1}, \dots) = \kappa^2$ . The regression in equation (9.29), which makes the test robust to heterokurtosis, is valid here as well.

### 3.3 Omnibus tests on the errors in time series regression models

Omnibus tests on the errors in time series regression models have recently become popular. A good example is the so-called BDS test (Brock, Dechert, LeBaron, and Scheinkman, 1996), which has been viewed as a general test for "nonlinearity." The null hypothesis for the BDS test is that the errors are independent and identically distributed, and the test has power against a variety of departures from the iid assumption, including neglected nonlinearities in the conditional mean, dynamic forms of heteroskedasticity, and even dynamics in higher order moments. Unfortunately, no economic theory implies that errors are iid. In many applications, especially in finance, it is often easy to reject the iid assumption using a simple test for dynamic heteroskedasticity, such as ARCH.

As with other omnibus tests, BDS gives equal weight to hypotheses that have very different practical importance. Finding that the errors in, say, an asset pricing equation are serially correlated – which usually means a violation of the efficient markets hypothesis – is more important than finding dynamic heteroskedasticity, which in turn is more important than finding, say, a nonconstant conditional fourth moment.

## 4 FINAL COMMENTS

Our focus in this chapter has been on the most common setting for diagnostic tests, namely, in univariate parametric models of conditional means and conditional variances. Recently, attention has turned to testing when some aspect of the

estimation problem is nonparametric. For example, we might wish to construct a test of a parametric model that has unit asymptotic power against all alternatives that satisfy fairly weak regularity conditions. Bierens (1990), Wooldridge (1992b), Hong and White (1995), de Jong (1996), Fan and Li (1996), and Zheng (1996) are some examples. Or, the estimated model may be semiparametric in nature, depending on an infinite dimensional parameter in addition to a finite dimensional parameter (Stoker, 1992; Fan and Li, 1996). The alternative is an infinite dimensional parameter space. In some cases the null model may be fully nonparametric, in which case the alternative is also nonparametric (e.g. Lewbel, 1995; and Fan and Li, 1996).

For diagnostic testing in time series contexts, we assumed that the underlying stochastic processes were weakly dependent. Currently, there is no general theory of diagnostic testing when the processes are not weakly dependent. Wooldridge (1999) considers a particular class of diagnostic tests in linear models with integrated processes and shows that, when the misspecification indicator is cointegrated, in a generalized sense, with the included explanatory variables, LM-type statistics have asymptotic chi-square distributions.

Another important topic we have omitted is diagnostic testing for panel data models. Panel data raises some additional important considerations, most of which revolve around our ability to control, to some extent, for time-constant heterogeneity. Strict exogeneity assumptions on the regressors, especially conditional on the unobserved effect, are important. Dynamic models with unobserved effects raise even more issues for estimation and diagnostic testing. (See Hsiao (1986) and Baltagi (1995) for discussions of these issues.)

### Note

- \* Two anonymous referees and Badi Baltagi provided helpful, timely comments on the first draft.

### References

- Andrews, D.W.K. (1991). Heteroskedasticity and autocorrelation consistent covariance matrix estimation. *Econometrica* 59, 817–58.
- Baltagi, B.H. (1995). *Econometric Analysis of Panel Data*. New York: Wiley.
- Bera, A.K., and C.M. Jarque (1982). Model specification tests: A simultaneous approach. *Journal of Econometrics* 20, 59–82.
- Bierens, H.J. (1990). A conditional moment test of functional form. *Econometrica* 58, 1443–58.
- Bollerslev, T. (1986). Generalized autoregressive conditional heteroskedasticity. *Journal of Econometrics* 31, 307–27.
- Breusch, T.S. (1978). Testing for autocorrelation in dynamic linear models. *Australian Economic Papers* 17, 334–55.
- Breusch, T.S., and A.R. Pagan (1979). A simple test for heteroskedasticity and random coefficient variation. *Econometrica* 50, 987–1007.
- Brock, W., W.D. Dechert, B. LeBaron, and J. Scheinkman (1996). A test for independence based on the correlation dimension. *Econometric Reviews* 15, 197–235.

- Davidson, R., and J.G. MacKinnon (1981). Several tests of model specification in the presence of alternative hypotheses. *Econometrica* 49, 781–93.
- Davidson, R., and J.G. MacKinnon (1985). Heteroskedasticity-robust tests in regression directions. *Annales de l'INSEÉ* 59/60, 183–218.
- Davidson, R., and J.G. MacKinnon (1987). Implicit alternatives and the local power of test statistics. *Econometrica* 55, 1305–29.
- Davidson, R., and J.G. MacKinnon (1993). *Estimation and Inference in Econometrics*. New York: Oxford University Press.
- de Jong, R.M. (1996). The Bierens test under data dependence. *Journal of Econometrics* 72, 1–32.
- Durbin, J., and G.S. Watson (1950). Testing for serial correlation in least squares regressions I. *Biometrika* 37, 409–28.
- Engle, R.F. (1982). Autoregressive conditional heteroskedasticity with estimates of the variance of United Kingdom inflation. *Econometrica* 50, 987–1008.
- Engle, R.F. (1984). Wald, likelihood, ratio, and Lagrange multiplier tests in econometrics. In Z. Griliches and M. Intriligator, eds., *Handbook of Econometrics*, vol. 2, Amsterdam: North-Holland.
- Fan, Y., and Q. Li (1996). Consistent model specification tests: Omitted variables and semiparametric functional forms. *Econometrica* 64, 865–90.
- Godfrey, L.G. (1978). Testing for higher order serial correlation in regression equations when the regressors include lagged dependent variables. *Econometrica* 46, 1303–10.
- Godfrey, L.G. (1988). *Misspecification Tests in Econometrics*. Cambridge: Cambridge University Press.
- Gouriéroux, C., A. Monfort, and C. Trognon (1984). Pseudo-maximum likelihood methods: Theory. *Econometrica* 52, 681–700.
- Hong, Y., and H. White (1995). Consistent specification testing via nonparametric series regression. *Econometrica* 63, 1133–59.
- Hsiao, C. (1986). *Analysis of Panel Data*. Cambridge: Cambridge University Press.
- Lewbel, A. (1995). Consistent nonparametric hypothesis tests with an application to Slutsky symmetry. *Journal of Econometrics* 67, 379–401.
- Messer, K., and H. White (1984). A note on computing the heteroskedasticity consistent covariance matrix using instrumental variable techniques. *Oxford Bulletin of Economics and Statistics* 46, 181–4.
- Mizon, G.E., and J.-F. Richard (1986). The encompassing principle and its application to testing non-nested hypotheses. *Econometrica* 54, 657–78.
- Moon, C.-G. (1988). Simultaneous specification test in a binary logit model: Skewness and heteroskedasticity. *Communications in Statistics* 17, 3361–87.
- Newey, W.K. (1985). Maximum likelihood specification testing and conditional moment tests. *Econometrica* 53, 1047–70.
- Newey, W.K., and K.D. West (1987). A simple positive semi-definite heteroskedasticity and autocorrelation consistent covariance matrix. *Econometrica* 55, 703–8.
- Pagan, A.R., and F. Vella (1989). Diagnostic tests for models based on individual data: A survey. *Journal of Applied Econometrics* 4, S29–59.
- Porter, R.D., and A.K. Kashyap (1984). Autocorrelation and the sensitivity of RESET. *Economics Letters* 14, 229–33.
- Ramsey, J.B. (1969). Tests for specification errors in the classical linear least squares regression analysis. *Journal of the Royal Statistical Society Series B* 31, 350–71.
- Stoker, T.M. (1992). *Lectures on Semiparametric Econometrics*. Louvain-la-Neuve, Belgium: CORE Lecture Series.

- Tauchen, G. (1985). Diagnostic testing and evaluation of maximum likelihood models. *Journal of Econometrics* 30, 415–43.
- Thursby, J.G. (1979). Alternative specification error tests: A comparative study. *Journal of the American Statistical Association* 74, 222–5.
- Thursby, J.G. (1989). A comparison of several specification error tests for a general alternative. *International Economic Review* 30, 217–30.
- White, H. (1980). A heteroskedasticity-consistent covariance matrix estimator and a direct test for heteroskedasticity. *Econometrica* 48, 817–38.
- White, H. (1982). Maximum likelihood estimation of misspecified models. *Econometrica* 50, 1–26.
- White, H. (1994). *Estimation, Inference and Specification Analysis*. Cambridge: Cambridge University Press.
- Wooldridge, J.M. (1990a). A unified approach to robust, regression-based specification tests. *Econometric Theory* 6, 17–43.
- Wooldridge, J.M. (1990b). An encompassing approach to conditional mean tests with applications to testing nonnested hypotheses. *Journal of Econometrics* 45, 331–50.
- Wooldridge, J.M. (1991a). On the application of robust, regression-based diagnostics to models of conditional means and conditional variances. *Journal of Econometrics* 47, 5–46.
- Wooldridge, J.M. (1991b). Specification testing and quasi-maximum likelihood estimation. *Journal of Econometrics* 48, 29–55.
- Wooldridge, J.M. (1992a). Some alternatives to the Box–Cox regression model. *International Economic Review* 33, 935–55.
- Wooldridge, J.M. (1992b). A test for functional form against nonparametric alternatives. *Econometric Theory* 8, 452–75.
- Wooldridge, J.M. (1994). Estimation and inference for dependent processes. In R.F. Engle and D.L. McFadden (eds.) *Handbook of Econometrics*, Volume 4. pp. 2639–2738. Amsterdam: North-Holland.
- Wooldridge, J.M. (1997). Quasi-likelihood methods for count data. In M.H. Pesaran and P. Schmidt (eds.) *Handbook of Applied Econometrics*, Volume 2. pp. 352–406. Oxford: Blackwell.
- Wooldridge, J.M. (1999). Asymptotic properties of some specification tests in linear models with integrated processes. In R.F. Engle and H. White (eds.) *Cointegration, Causality, and Forecasting*. Oxford: Oxford University Press.
- Zheng, J.X. (1996). A consistent test of functional form via nonparametric estimation techniques. *Journal of Econometrics* 75, 263–89.

CHAPTER TEN

# Basic Elements of Asymptotic Theory

Benedikt M. Pötscher and Ingmar R. Prucha

## 1 INTRODUCTION

Consider the estimation problem where we would like to estimate a parameter vector  $\theta$  from a sample  $Y_1, \dots, Y_n$ . Let  $\hat{\theta}_n$  be an estimator for  $\theta$ , i.e. let  $\hat{\theta}_n = h(Y_1, \dots, Y_n)$  be a function of the sample.<sup>1</sup> In the important special case where  $\hat{\theta}_n$  is a *linear* function of  $Y_1, \dots, Y_n$ , i.e.  $\hat{\theta}_n = Ay$ , where  $A$  is a nonrandom matrix and  $y = (Y_1, \dots, Y_n)'$ , we can easily express the expected value and the variance-covariance matrix of  $\hat{\theta}_n$  in terms of the first and second moments of  $y$  (provided those moments exist). Also, if the sample is normally distributed, so is  $\hat{\theta}_n$ . Well known examples of linear estimators are the OLS- and the GLS-estimator of the linear regression model. Frequently, however, the estimator of interest will be a *nonlinear* function of the sample. In principle, the distribution of  $\hat{\theta}_n$  can then be found from the distribution of the sample, if the model relating the parameter  $\theta$  to the observables  $Y_1, \dots, Y_n$  fully specifies the distribution of the sample. For example in a linear regression model with independently and identically distributed errors this would require assuming a specific distribution for the errors. However, even if the researcher is willing to make such a specific assumption, it will then still often be impossible – for all practical purposes – to obtain an exact expression for the distribution of  $\hat{\theta}_n$  because of the complexity of the necessary calculations. (Even if  $\hat{\theta}_n$  is linear, but the distribution of  $y$  is nonnormal, it will typically be difficult to obtain the exact distribution of  $\hat{\theta}_n$ .) Similarly, obtaining expressions for, say, the first and second moments of  $\hat{\theta}_n$  will, for practical purposes, typically be unfeasible for nonlinear estimators; and even if it is feasible, the resulting expressions will usually depend on the entire distribution of the sample, and not only on the first and second moments as in the case of a linear estimator. A further complication arises in case the model relating  $\theta$  to the observables  $Y_1, \dots, Y_n$  does not fully specify the distribution of  $Y_1, \dots, Y_n$ . For

example in a linear regression model the errors may only be assumed to be identically and independently distributed with zero mean and finite variance, without putting any further restrictions on the distribution function of the disturbances. In this case we obviously cannot get a handle on the distribution of  $\hat{\theta}_n$  (even if  $\hat{\theta}_n$  is linear), in the sense that this distribution will depend on the unknown distribution of the errors.

In view of the above discussed difficulties in obtaining *exact* expressions for characteristics of estimators like their moments or distribution functions we will often have to resort to *approximations* for these exact expressions. Ideally, these approximations should be easier to obtain than the exact expressions and they should be of a simpler form. Asymptotic theory is one way of obtaining such approximations by essentially asking what happens to the exact expressions as the sample size tends to infinity. For example, if we are interested in the expected value of  $\hat{\theta}_n$  and an exact expression for it is unavailable or unwieldy, we could ask if the expected value of  $\hat{\theta}_n$  converges to  $\theta$  as the sample size increases (i.e. if  $\hat{\theta}_n$  is “asymptotically unbiased”). One could try to verify this by first showing that the estimator  $\hat{\theta}_n$  itself “converges” to  $\theta$  in an appropriate sense, and then by attempting to obtain the convergence of the expected value of  $\hat{\theta}_n$  to  $\theta$  from the “convergence” of the estimator. In order to properly pose and answer such questions we need to study various notions of convergence of random vectors.

The article is organized as follows: in Section 2 we define various modes of convergence of random vectors, and discuss the properties of and the relationships between these modes of convergence. Sections 3 and 4 provide results that allow us to deduce the convergence of certain important classes of random vectors from basic assumptions. In particular, in Section 3 we discuss laws of large numbers, including uniform laws of large numbers. A discussion of central limit theorems is given in Section 4. In Section 5 we suggest additional literature for further reading.

We emphasize that the article only covers material that lays the foundation for asymptotic theory. It does not provide results on the asymptotic properties of estimators for particular models; for references see Section 5. All of the material presented here is essentially textbook material. We provide proofs for some selected results for the purpose of practice and since some of the proofs provide interesting insights. For most results given without a proof we provide references to widely available textbooks. Proofs for some of the central limit theorems presented in Section 4 are given in a longer mimeographed version of this article, which is available from the authors upon request.

We adopt the following notation and conventions: throughout this chapter  $Z_1, Z_2, \dots$ , and  $Z$  denote random vectors that take their values in a Euclidean space  $\mathbb{R}^k$ ,  $k \geq 1$ . Furthermore, all random vectors involved in a particular statement are assumed to be defined on a common probability space  $(\Omega, \mathcal{F}, P)$ , except when noted otherwise. With  $|.|$  we denote the absolute value and with  $\|.\|$  the Euclidean norm. All matrices considered are real matrices. If  $A$  is a matrix, then  $A'$  denotes its transpose; if  $A$  is a square matrix, then  $A^{-1}$  denotes the inverse of  $A$ . The norm of a matrix  $A$  is denoted by  $\|A\|$  and is taken to be  $\|\text{vec}(A)\|$ , where  $\text{vec}(A)$  stands for the columnwise vectorization of  $A$ . If  $C_n$  is a sequence of sets,

then  $C_n \uparrow C$  stands for  $C_n \subseteq C_{n+1}$  for all  $n \in \mathbb{N}$  and  $C = \bigcup_{n=1}^{\infty} C_n$ . Similarly,  $C_n \downarrow C$  stands for  $C_n \supseteq C_{n+1}$  for all  $n \in \mathbb{N}$  and  $C = \bigcap_{n=1}^{\infty} C_n$ . Furthermore, if  $B$  is a set, then  $\mathbf{1}(B)$  denotes the indicator function of  $B$ .

## 2 MODES OF CONVERGENCE FOR SEQUENCES OF RANDOM VECTORS

In this section we define and discuss various modes of convergence for sequences of random vectors taking their values in  $\mathbb{R}^k$ .

### 2.1 Convergence in probability, almost surely, and in $r$ th mean

We first consider the case where  $k = 1$ , i.e. the case of real valued random variables. Extension to the vector case are discussed later. We start by defining convergence in probability.

**Definition 1.** (Convergence in probability) The sequence of random variables  $Z_n$  converges in probability (or stochastically) to the random variable  $Z$  if for every  $\varepsilon > 0$

$$\lim_{n \rightarrow \infty} P(|Z_n - Z| \leq \varepsilon) = 1. \quad (10.1)$$

We then write  $\text{plim}_{n \rightarrow \infty} Z_n = Z$ , or  $Z_n \xrightarrow{P} Z$ , or  $Z_n \rightarrow Z$  i.p. as  $n \rightarrow \infty$ .

We next define almost sure convergence.

**Definition 2.** (Almost sure convergence) The sequence of random variables  $Z_n$  converges almost surely (or strongly or with probability one) to the random variable  $Z$  if there exists a set  $N \in \mathcal{F}$  with  $P(N) = 0$  such that  $\lim_{n \rightarrow \infty} Z_n(\omega) = Z(\omega)$  for every  $\omega \in \Omega - N$ , or equivalently

$$P(\{\omega \in \Omega : \lim_{n \rightarrow \infty} Z_n(\omega) = Z(\omega)\}) = 1. \quad (10.2)$$

We then write  $Z_n \xrightarrow{a.s.} Z$ , or  $Z_n \rightarrow Z$  a.s., or  $Z_n \rightarrow Z$  w.p.1 as  $n \rightarrow \infty$ .

The following theorem provides an alternative characterization of almost sure convergence.

**Theorem 1.** The sequence of random variables  $Z_n$  converges almost surely to the random variable  $Z$  if and only if

$$\lim_{n \rightarrow \infty} P(\{|Z_i - Z| \leq \varepsilon \text{ for all } i \geq n\}) = 1 \quad (10.3)$$

for every  $\varepsilon > 0$ .

**Proof.** Let

$$A = \{\omega \in \Omega : \lim_{n \rightarrow \infty} Z_n(\omega) = Z(\omega)\}$$

and

$$A_n^\varepsilon = \{\omega \in \Omega : |Z_i(\omega) - Z(\omega)| \leq \varepsilon \text{ for all } i \geq n\},$$

then (10.2) and (10.3) can be written equivalently as  $P(A) = 1$  and  $\lim_{n \rightarrow \infty} P(A_n^\varepsilon) = 1$ . Next define  $A^\varepsilon = \bigcup_{n=1}^{\infty} A_n^\varepsilon$  and observe that  $A_n^\varepsilon \uparrow A^\varepsilon$ . By construction  $A^\varepsilon$  is the set of all  $\omega \in \Omega$  for which there exists some finite index  $n_\varepsilon(\omega)$  such that  $|Z_i(\omega) - Z(\omega)| \leq \varepsilon$  for all  $i \geq n_\varepsilon(\omega)$ . Consequently  $A \subseteq A^\varepsilon$ ; in fact  $A = \bigcap_{\varepsilon > 0} A^\varepsilon$ . Now suppose (10.2) holds, i.e.  $P(A) = 1$ . Then, using the continuity theorem for probability measures given, e.g., in Billingsley (1979, p. 21), we have  $P(A^\varepsilon) = \lim_{n \rightarrow \infty} P(A_n^\varepsilon) \geq P(A) = 1$ , i.e. (10.3) holds. Conversely, suppose (10.3) holds, then  $P(A^\varepsilon) = 1$ . Observe that  $A^\varepsilon \downarrow A$  as  $\varepsilon \downarrow 0$ . Choosing  $\varepsilon = 1/k$  we have  $A = \bigcap_{k=1}^{\infty} A^{1/k}$  and, using again the continuity theorem for probability measures,  $P(A) = \lim_{k \rightarrow \infty} P(A^{1/k}) = 1$ . ■

The above theorem makes it evident that almost sure convergence implies convergence in probability.

**Theorem 2.** If  $Z_n \xrightarrow{a.s.} Z$ , then  $Z_n \xrightarrow{p} Z$ .

**Proof.** Obviously the event  $B_n^\varepsilon = \{\omega \in \Omega : |Z_n(\omega) - Z(\omega)| \leq \varepsilon\}$  contains the event  $A_n^\varepsilon = \{\omega \in \Omega : |Z_i(\omega) - Z(\omega)| \leq \varepsilon \text{ for all } i \geq n\}$ . Hence Theorem 1 implies that  $\lim_{n \rightarrow \infty} P(B_n^\varepsilon) = 1$ , i.e. that (10.1) holds. ■

The converse of the above theorem does not hold. That is, in general, convergence in probability does not imply almost sure convergence as is demonstrated by the following well known example.

**Example 1.<sup>2</sup>** Let  $\Omega = [0, 1]$ , let  $\mathcal{F}$  be the corresponding Borel  $\sigma$ -field, and let  $P(\cdot)$  be the uniform distribution on  $\Omega$ , i.e.  $P([a, b]) = b - a$ . Define

$$Z_n(\omega) = \begin{cases} 1 & \text{if } \omega \in [m_n 2^{-k_n}, (m_n + 1) 2^{-k_n}) \\ 0 & \text{otherwise} \end{cases}$$

where the integers  $m_n$  and  $k_n$  satisfy  $n = m_n + 2^{k_n}$  and  $0 \leq m_n < 2^{k_n}$ . That is,  $k_n$  is the largest integer satisfying  $2^{k_n} \leq n$ . Let  $Z = 0$  and let  $A_n^\varepsilon$  and  $B_n^\varepsilon$  be defined as above. Then for  $\varepsilon < 1$  we have  $B_n^\varepsilon = \Omega - [m_n 2^{-k_n}, (m_n + 1) 2^{-k_n})$  and hence  $P(B_n^\varepsilon) = 1 - 2^{-k_n} \rightarrow 1$  as  $n \rightarrow \infty$ . This establishes that  $Z_n$  converges to zero in probability. Observe further that  $A_n^\varepsilon = \bigcap_{i=n}^{\infty} B_i^\varepsilon = \emptyset$ . Consequently  $Z_n$  does not converge to zero almost surely. In fact, in this example  $Z_n(\omega)$  does not converge to 0 for any  $\omega \in \Omega$ , although  $Z_n \xrightarrow{p} 0$ .

We next define convergence in  $r$ th mean.

**Definition 3.** (Convergence in  $r$ th mean) The sequence of random variables  $Z_n$  converges in  $r$ th mean to the random variable  $Z$ ,  $0 < r < \infty$ , if

$$\lim_{n \rightarrow \infty} E|Z_n - Z|^r = 0.$$

We then write  $Z_n \xrightarrow{r\text{th}} Z$ . For  $r = 2$  we say the sequence converges in quadratic mean or mean square.<sup>3</sup>

**Remark 1.** For all three modes of convergence introduced above one can show that the limiting random variable  $Z$  is unique up to null sets. That is, suppose  $Z$  and  $Z^*$  are both limits of the sequence  $Z_n$ , then  $P(Z = Z^*) = 1$ .

Lyapounov's inequality implies that  $E|Z_n - Z|^s \leq \{E|Z_n - Z|^r\}^{s/r}$  for  $0 < s \leq r$ . As a consequence we have the following theorem, which tells us that the higher the value of  $r$ , the more stringent the condition for convergence in  $r$ th mean.

**Theorem 3.**  $Z_n \xrightarrow{r\text{th}} Z$  implies  $Z_n \xrightarrow{s\text{th}} Z$  for  $0 < s \leq r$ .

The following theorem gives conditions under which convergence in  $r$ th mean implies convergence of the  $r$ th moments.

**Theorem 4.**<sup>4</sup> Suppose  $Z_n \xrightarrow{r\text{th}} Z$  and  $E|Z|^r < \infty$ . Then  $E|Z_n|^r \rightarrow E|Z|^r$ . If, furthermore,  $Z_n^r$  and  $Z^r$  are well-defined for all  $n$  (e.g. if  $Z_n \geq 0$  and  $Z \geq 0$ , or if  $r$  is a natural number), then also  $EZ_n^r \rightarrow EZ^r$ .

By Chebyshev's inequality we have  $P\{|Z_n - Z| \geq \varepsilon\} \leq E|Z_n - Z|^r / \varepsilon^r$  for  $r > 0$ . As a consequence, convergence in  $r$ th mean implies convergence in probability, as stated in the following theorem.

**Theorem 5.** If  $Z_n \xrightarrow{r\text{th}} Z$  for some  $r > 0$ , then  $Z_n \xrightarrow{p} Z$ .

The corollary below follows immediately from Theorem 5 with  $r = 2$  by utilizing the decomposition  $E|Z_n - c|^2 = \text{var}(Z_n) + (EZ_n - c)^2$ .

**Corollary 1.** Suppose  $EZ_n \rightarrow c$  and  $\text{var}(Z_n) \rightarrow 0$ , then  $Z_n \xrightarrow{p} c$ .

The corollary is frequently used to show that for an estimator  $\hat{\theta}_n$  with  $E\hat{\theta}_n \rightarrow \theta$  (i.e. an asymptotically unbiased estimator) and with  $\text{var}(\hat{\theta}_n) \rightarrow 0$  we have  $\hat{\theta}_n \xrightarrow{p} \theta$ .

**Example 2.** Let  $y_t$  be a sequence of iid distributed random variables with  $Ey_t = \theta$  and  $\text{var}(y_t) = \sigma^2 < \infty$ . Let  $\hat{\theta}_n = n^{-1} \sum_{t=1}^n y_t$  denote the sample mean. Then  $E\hat{\theta}_n = \theta$  and  $\text{var}(\hat{\theta}_n) = \sigma^2/n \rightarrow 0$ , and hence  $\hat{\theta}_n \xrightarrow{p} \theta$ .

Theorem 5 and Corollary 1 show how convergence in probability can be implied from the convergence of appropriate moments. The converse is not true in general, and in particular  $Z_n \xrightarrow{p} Z$  does not imply  $Z_n \xrightarrow{r\text{th}} Z$ . In fact even  $Z_n \xrightarrow{as} Z$  does not imply  $Z_n \xrightarrow{r\text{th}} Z$ . These claims are illustrated by the following example.

**Example 3.** Let  $\Omega$ ,  $\mathcal{F}$ , and  $P$  be as in Example 1 and define

$$Z_n(\omega) = \begin{cases} 0 & \text{for } \omega \in [0, 1 - 1/n), \\ n & \text{for } \omega \in [1 - 1/n, 1]. \end{cases}$$

Then  $Z_n(\omega) \rightarrow 0$  for all  $\omega \in \Omega$  and hence  $Z_n \xrightarrow{a.s.} 0$ . However,  $E|Z_n| = 1$  for all  $n$  and hence  $Z_n$  does not converge to 0 in  $r$ th mean with  $r = 1$ .

The above example shows in particular that an estimator that satisfies  $\hat{\theta}_n \xrightarrow{p} \theta$  (or  $\hat{\theta}_n \xrightarrow{a.s.} \theta$ ) need not satisfy  $E\hat{\theta}_n \rightarrow \theta$ , i.e. need not be asymptotically unbiased. Additional conditions are needed for such a conclusion to hold. Such conditions are given in the following theorem. The theorem states that convergence in probability implies convergence in  $r$ th mean, given that the convergence is dominated.

**Theorem 6.<sup>5</sup>** (Dominated convergence theorem) Suppose  $Z_n \xrightarrow{p} Z$ , and there exists a random variable  $Y$  satisfying  $|Z_n| \leq Y$  a.s. for all  $n$  and  $EY^r < \infty$ . Then  $Z_n \xrightarrow{r\text{th}} Z$  and  $E|Z|^r < \infty$ . (Of course, the theorem also holds if  $Z_n \xrightarrow{p} Z$  is replaced by  $Z_n \xrightarrow{a.s.} Z$ , since the latter implies the former.)

Under the assumptions of the above theorem also convergence of the  $r$ th moments follows in view of Theorem 4. We also note that the existence of a random variable  $Y$  satisfying the requirements in Theorem 6 is certainly guaranteed if there exists a real number  $M$  such that  $|Z_n| \leq M$  a.s. for all  $n$  (choose  $Y = M$ ).

Now let  $Z_n$  be a sequence of random vectors taking their values in  $\mathbb{R}^k$ . Convergence in probability, almost surely, and in the  $r$ th mean are then defined exactly as in the case  $k = 1$  with the only difference that the absolute value  $|.|$  has to be replaced by  $\|.\|$ , the Euclidean norm on  $\mathbb{R}^k$ . Upon making this replacement all of the results presented in this subsection generalize to the vector case with two obvious exceptions: first, in Corollary 1 the condition  $\text{var}(Z_n) \rightarrow 0$  has to be replaced by the conditions that the variances of the components of  $Z_n$  converge to zero, or equivalently, that the variance–covariance matrix of  $Z_n$  converges to zero. Second, the last claim in Theorem 4 continues to hold if the symbol  $Z_n^r$  is interpreted so as to represent the vector of the  $r$ th power of the components of  $Z_n$ . Instead of extending the convergence notions to the vector case by replacing the absolute value  $|.|$  by the norm  $\|.\|$ , we could have defined convergence in probability, almost surely and in  $r$ th mean for sequences of random vectors by requiring that each component of  $Z_n$  satisfies Definition 1, 2, or 3, respectively. That this leads to an equivalent definition is shown in the following theorem.

**Theorem 7.** Let  $Z_n$  and  $Z$  be random vectors taking their values in  $\mathbb{R}^k$ , and let  $Z_n^{(i)}$  and  $Z^{(i)}$  denote their  $i$ th component, respectively. Then  $Z_n \xrightarrow{p} Z$  if and only if  $Z_n^{(i)} \xrightarrow{p} Z^{(i)}$  for  $i = 1, \dots, k$ . An analogous statement holds for almost sure convergence and for convergence in  $r$ th mean.

The theorem follows immediately from the following simple inequality:

$$|Z_n^{(i)} - Z^{(i)}| \leq \|Z_n - Z\| \leq \sqrt{k} \max_{i=1,\dots,k} (|Z_n^{(i)} - Z^{(i)}|).$$

For sequences of random  $k \times l$ -matrices  $W_n$  convergence in probability, almost surely, and in  $r$ th mean is defined as the corresponding convergence of the sequence  $\text{vec}(W_n)$ .

We finally note the following simple fact: Suppose  $Z_1, Z_2, \dots$ , and  $Z$  are nonrandom vectors, then  $Z_n \xrightarrow{p} Z$ ,  $Z_n \xrightarrow{a.s.} Z$ , and  $Z_n \xrightarrow{r\text{th}} Z$  each hold if and only if  $Z_n \rightarrow Z$  as  $n \rightarrow \infty$ . That is, in this case all of the concepts of convergence of random vectors introduced above coincide with the usual convergence concept for sequences of vectors in  $\mathbb{R}^k$ .

## 2.2 Convergence in distribution

Let  $\hat{\theta}_n$  be an estimator for a real-valued parameter  $\theta$  and assume  $\hat{\theta}_n \xrightarrow{p} \theta$ . If  $G_n$  denotes the cumulative distribution function (CDF) of  $\hat{\theta}_n$ , i.e.,  $G_n(z) = P(\hat{\theta}_n \leq z)$ , then as  $n \rightarrow \infty$

$$G_n(z) \rightarrow \begin{cases} 0 & \text{for } z < \theta \\ 1 & \text{for } z > \theta. \end{cases} \quad (10.4)$$

To see this observe that  $P(\hat{\theta}_n \leq z) = P(\hat{\theta}_n - \theta \leq z - \theta) \leq P(|\hat{\theta}_n - \theta| \geq \theta - z)$  for  $z < \theta$ , and  $P(\hat{\theta}_n \leq z) = 1 - P(\hat{\theta}_n > z) = 1 - P(\hat{\theta}_n - \theta > z - \theta) \geq 1 - P(|\hat{\theta}_n - \theta| > z - \theta)$  for  $z > \theta$ . The result in (10.4) shows that the distribution of  $\hat{\theta}_n$  “collapses” into the degenerate distribution at  $\theta$ , i.e., into

$$G(z) = \begin{cases} 0 & \text{for } z < \theta \\ 1 & \text{for } z \geq \theta. \end{cases} \quad (10.5)$$

Consequently, knowing that  $\hat{\theta}_n \xrightarrow{p} \theta$  does not provide information about the shape of  $G_n$ . As a point of observation note that  $G_n(z) \rightarrow G(z)$  for  $z \neq \theta$ , but  $G_n(z)$  may not converge to  $G(z) = 1$  for  $z = \theta$ . For example, if  $\hat{\theta}_n$  is distributed symmetrically around  $\theta$ , then  $G_n(\theta) = 1/2$  and hence does not converge to  $G(\theta) = 1$ .

This raises the question of how we can obtain information about  $G_n$  based on some limiting process. Consider, for example, the case where  $\hat{\theta}_n$  is the sample mean of iid random variables with mean  $\theta$  and variance  $\sigma^2 > 0$ . Then  $\hat{\theta}_n \xrightarrow{p} \theta$  in light of Corollary 1, since  $E\hat{\theta}_n = \theta$  and  $\text{var}(\hat{\theta}_n) = \sigma^2/n \rightarrow 0$ . Consequently, as discussed above, the distribution of  $\hat{\theta}_n$  “collapses” into the degenerate distribution at  $\theta$ . Observe, however, that the rescaled variable  $\sqrt{n}(\hat{\theta}_n - \theta)$  has mean zero and variance  $\sigma^2$ . This indicates that the distribution of  $\sqrt{n}(\hat{\theta}_n - \theta)$  will not collapse to a degenerate distribution. Hence, if  $\sqrt{n}(\hat{\theta}_n - \theta)$  “converges,” the limiting CDF can be expected to be non-degenerate. To formalize these ideas we need to define an appropriate notion of convergence of CDFs.<sup>6</sup>

**Definition 4.** (Convergence in distribution) Let  $F_1, F_2, \dots$ , and  $F$  denote CDFs on  $\mathbb{R}$ . Then  $F_n$  converges weakly to  $F$  if

$$\lim_{n \rightarrow \infty} F_n(z) = F(z)$$

for all  $z \in \mathbb{R}$  that are continuity points of  $F$ .

Let  $Z_1, Z_2, \dots$ , and  $Z$  denote random variables with corresponding CDFs  $F_1, F_2, \dots$ , and  $F$ , respectively. We then say that  $Z_n$  converges in distribution (or in law) to  $Z$ , if  $F_n$  converges weakly to  $F$ . We write  $Z_n \xrightarrow{d} Z$  or  $Z_n \xrightarrow{L} Z$ .

Consider again the sample mean  $\hat{\theta}_n$  of iid random variables with mean  $\theta$  and variance  $\sigma^2 > 0$ . As demonstrated above,  $\hat{\theta}_n \xrightarrow{P} \theta$  only implies weak convergence of the CDF of  $\hat{\theta}_n$  to a degenerate distribution, which is not informative about the shape of the distribution function of  $\hat{\theta}_n$ . In contrast the limiting distribution of  $\sqrt{n}(\hat{\theta}_n - \theta)$  is found to be non-degenerate. In fact, using Theorem 24 below, it can be shown that  $\sqrt{n}(\hat{\theta}_n - \theta)$  converges in distribution to a  $N(0, \sigma^2)$  distributed random variable. As a result we can take  $N(0, \sigma^2)$  as an approximation for the finite sample distribution of  $\sqrt{n}(\hat{\theta}_n - \theta)$ , and consequently take  $N(\theta, \sigma^2/n)$  as an approximation for the finite sample distribution of  $\hat{\theta}_n$ .

## Remark 2

- (a) The reason for requiring in the above definition that  $F_n(z) \rightarrow F(z)$  converges only at the continuity points of  $F$  is to accommodate situations as, e.g., in (10.4). Of course, if  $F$  is continuous, then  $F_n$  converges weakly to  $F$  if and only if  $F_n(z) \rightarrow F(z)$  for all  $z \in \mathbb{R}$ .
- (b) As is evident from the definition, the concept of convergence in distribution is defined completely in terms of the convergence of distribution functions. In fact, the concept of convergence in distribution remains well defined even for sequences of random variables that are not defined on a common probability space.
- (c) To further illustrate what convergence in distribution does not mean consider the following example: Let  $Y$  be a random variable that takes the values  $+1$  and  $-1$  with probability  $1/2$ . Define  $Z_n = Y$  for  $n \geq 1$  and  $Z = -Y$ . Then clearly  $Z_n \xrightarrow{d} Z$  since  $Z_n$  and  $Z$  have the same distribution, but  $|Z_n - Z| = 2$  for all  $n \geq 1$ . That is, convergence in distribution does not necessarily mean that the difference between random variables vanishes in the limit. More generally, if  $Z_n \xrightarrow{d} Z$  and one replaces the sequence  $Z_n$  by a sequence  $Z_n^*$  that has the same marginal distributions, then also  $Z_n^* \xrightarrow{d} Z$ .

The next theorem provides several equivalent characterizations of weak convergence.

**Theorem 8.<sup>7</sup>** Consider the cumulative distribution functions  $F, F_1, F_2, \dots$ . Let  $Q, Q_1, Q_2, \dots$  denote the corresponding probability measures on  $\mathbb{R}$ , and let  $\phi, \phi_1, \phi_2, \dots$  denote the corresponding characteristic functions. Then the following statements are equivalent:

- (a)  $F_n$  converges weakly to  $F$ .
- (b)  $\lim_{n \rightarrow \infty} Q_n(A) = Q(A)$  for all Borel sets  $A \subseteq \mathbb{R}$  that are  $Q$ -continuous, i.e. for all Borel sets  $A$  whose boundary  $\partial A$  satisfies  $Q(\partial A) = 0$ .
- (c)  $\lim_{n \rightarrow \infty} \int f dF_n = \int f dF$  for all bounded and continuous real valued functions  $f$  on  $\mathbb{R}$ .
- (d)  $\lim_{n \rightarrow \infty} \phi_n(t) = \phi(t)$  for all  $t \in \mathbb{R}$ .

If, furthermore, the cumulative distribution functions  $F, F_1, F_2, \dots$  have moment generating functions  $M, M_1, M_2, \dots$  in some common interval  $[-t_*, t_*]$ ,  $t_* > 0$ , then (a), (b), (c) or (d) are, respectively, equivalent to

$$(e) \quad \lim_{n \rightarrow \infty} M_n(t) = M(t) \text{ for all } t \in [-t_*, t_*].$$

**Remark 3.** The equivalence of (a) and (b) of Theorem 8 can be reformulated as  $Z_n \xrightarrow{d} Z \Leftrightarrow P(Z_n \in A) \rightarrow P(Z \in A)$  for all Borel sets  $A$  with  $P(Z \in \partial A) = 0$ . The equivalence of (a) and (c) can be expressed equivalently as  $Z_n \xrightarrow{d} Z \Leftrightarrow Ef(Z_n) \rightarrow Ef(Z)$  for all bounded and continuous real valued functions  $f$  on  $\mathbb{R}$ .

The following theorem relates convergence in probability to convergence in distribution.

**Theorem 9.**  $Z_n \xrightarrow{p} Z$  implies  $Z_n \xrightarrow{d} Z$ . (Of course, the theorem also holds if  $Z_n \xrightarrow{p} Z$  is replaced by  $Z_n \xrightarrow{a.s.} Z$  or  $Z_n \xrightarrow{rth} Z$ , since the latter imply the former.)

**Proof.** Let  $f(z)$  be any bounded and continuous real valued function, and let  $C$  denote the bound. Then  $Z_n \xrightarrow{p} Z$  implies  $f(Z_n) \xrightarrow{p} f(Z)$  by the results on convergence in probability of transformed sequences given in Theorem 14 in Section 2.3. Since  $|f(Z_n(\omega))| \leq C$  for all  $n$  and  $\omega \in \Omega$  it then follows from Theorems 6 and 4 that  $Ef(Z_n) \rightarrow Ef(Z)$ , and hence  $Z_n \xrightarrow{d} Z$  by Theorem 8. ■

The converse of the above theorem does not hold in general, i.e.  $Z_n \xrightarrow{d} Z$  does not imply  $Z_n \xrightarrow{p} Z$ . To see this consider the following example: let  $Z \sim N(0, 1)$  and put  $Z_n = (-1)^n Z$ . Then  $Z_n$  does not converge almost surely or in probability. But since each  $Z_n \sim N(0, 1)$ , evidently  $Z_n \xrightarrow{d} Z$ .

Convergence in distribution to a constant is, however, equivalent to convergence in probability to that constant.

**Theorem 10.** Let  $c \in \mathbb{R}$ , then  $Z_n \xrightarrow{d} c$  is equivalent to  $Z_n \xrightarrow{p} c$ .

**Proof.** Because of Theorem 9 we only have to show that  $Z_n \xrightarrow{d} c$  implies  $Z_n \xrightarrow{p} c$ . Observe that for any  $\varepsilon > 0$

$$\begin{aligned} P(|Z_n - c| > \varepsilon) &= P(Z_n - c < -\varepsilon) + P(Z_n - c > \varepsilon) \\ &\leq P(Z_n \leq c - \varepsilon) - P(Z_n \leq c + \varepsilon) + 1 \\ &= F_n(c - \varepsilon) - F_n(c + \varepsilon) + 1 \end{aligned}$$

where  $F_n$  is the CDF of  $Z_n$ . The CDF of  $Z = c$  is

$$F(z) = \begin{cases} 0 & z < c \\ 1 & z \geq c \end{cases}.$$

Hence,  $c - \varepsilon$  and  $c + \varepsilon$  are continuity points of  $F$ . Since  $Z_n \xrightarrow{d} Z$  it follows that  $F_n(c - \varepsilon) \rightarrow F(c - \varepsilon) = 0$  and  $F_n(c + \varepsilon) \rightarrow F(c + \varepsilon) = 1$ . Consequently,

$$0 \leq P(|Z_n - c| > \varepsilon) \leq F_n(c - \varepsilon) + 1 - F_n(c + \varepsilon) \rightarrow 0 + 1 - 1 = 0.$$

This shows  $Z_n \xrightarrow{p} Z = c$ . ■

In general convergence in distribution does not imply convergence of moments; in fact the moments may not even exist. However, we have the following result.

**Theorem 11.**<sup>8</sup> Suppose  $Z_n \xrightarrow{d} Z$  and suppose that  $\sup_n E|Z_n|^r < \infty$  for some  $0 < r < \infty$ . Then for all  $0 < s < r$  we have  $E|Z|^s < \infty$  and  $\lim_{n \rightarrow \infty} E|Z_n|^s = E|Z|^s$ . If, furthermore,  $Z^s$  and  $Z_n^s$  are well-defined for all  $n$ , then also  $\lim_{n \rightarrow \infty} EZ_n^s = EZ^s$ .

**Remark 4.** Since  $Z_n \xrightarrow{p} Z$  and  $Z_n \xrightarrow{a.s.} Z$  imply  $Z_n \xrightarrow{d} Z$ , Theorem 11 provides sufficient conditions under which  $Z_n \xrightarrow{p} Z$  and  $Z_n \xrightarrow{a.s.} Z$  imply convergence of moments. These conditions are an alternative to those of Theorems 6 and 4.

The concept of convergence in distribution can be generalized to sequences of random vectors  $Z_n$  taking their values in  $\mathbb{R}^k$ . Contrary to the approach taken in generalizing the notions of convergence in probability, almost surely, and in  $r$ th mean to the vector case, the appropriate generalization is here *not* obtained by simply requiring that the component sequences  $Z_n^{(i)}$  converge in distribution for  $i = 1, \dots, k$ . Such an attempt at generalizing the notion of convergence in distribution would yield a nonsensical convergence concept as is illustrated by Example 4 below. The proper generalization is given in the following definition.

**Definition 5.** Let  $F_1, F_2, \dots$ , and  $F$  denote CDFs on  $\mathbb{R}^k$ . Then  $F_n$  converges weakly to  $F$  if

$$\lim_{n \rightarrow \infty} F_n(z) = F(z)$$

for all  $z \in \mathbb{R}^k$  that are continuity points of  $F$ .

Let  $Z_1, Z_2, \dots$ , and  $Z$  denote random vectors taking their values in  $\mathbb{R}^k$  with corresponding CDFs  $F_1, F_2, \dots$ , and  $F$ , respectively. We then say that  $Z_n$  converges in distribution (or in law) to  $Z$ , if  $F_n$  converges weakly to  $F$ . We write  $Z_n \xrightarrow{d} Z$  or  $Z_n \xrightarrow{L} Z$ .

All the results presented in this subsection so far also hold for the multivariate case (with  $\mathbb{R}^k$  replacing  $\mathbb{R}$ ). Convergence in distribution of a sequence of random matrices  $W_n$  is defined as convergence in distribution of  $\text{vec}(W_n)$ .

The next theorem states that weak convergence of the joint distributions implies weak convergence of the marginal distributions.

**Theorem 12.** Weak convergence of  $F_n$  to  $F$  implies weak convergence of  $F_n^{(i)}$  to  $F^{(i)}$  and  $Z_n \xrightarrow{d} Z$  implies  $Z_n^{(i)} \xrightarrow{d} Z^{(i)}$ , where  $F_n^{(i)}$  and  $F^{(i)}$  denote the  $i$ th marginal

distribution of  $F_n$  and  $F$ , and  $Z_n^{(i)}$  and  $Z^{(i)}$  denote the  $i$ th component of  $Z_n$  and  $Z$ , respectively.

**Proof.** The result follows from Theorem 14 below, since projections are continuous. ■

However, as alluded to in the above discussion, the converse of Theorem 12 is not true. That is, weak convergence of the marginal distributions is not equivalent to weak convergence of the joint distribution, as is illustrated by the following counter example.

**Example 4.** Let  $Z \sim N(0, 1)$  and let

$$Z_n = \begin{pmatrix} Z \\ (-1)^n Z \end{pmatrix}.$$

Clearly, the marginal distributions of each component of  $Z_n$  converge weakly to  $N(0, 1)$ . However, for  $n$  even the distribution of  $Z_n$  is concentrated on the line  $\{(z, z) : z \in \mathbb{R}\}$ , whereas for  $n$  odd the distribution of  $Z_n$  is concentrated on the line  $\{(z, -z) : z \in \mathbb{R}\}$ . Consequently, the random vectors  $Z_n$  do not converge in distribution, i.e. the distributions of  $Z_n$  do not converge weakly.

The following result is frequently useful in reducing questions about convergence in distribution of random vectors to corresponding questions about convergence in distribution of random variables.

**Theorem 13.** (Cramér–Wold device) Let  $Z_1, Z_2, \dots$ , and  $Z$  denote random vectors taking their values in  $\mathbb{R}^k$ . Then the following statements are equivalent:

- (a)  $Z_n \xrightarrow{d} Z$
- (b)  $\alpha' Z_n \xrightarrow{d} \alpha' Z$  for all  $\alpha \in \mathbb{R}^k$ .
- (c)  $\alpha' Z_n \xrightarrow{d} \alpha' Z$  for all  $\alpha \in \mathbb{R}^k$  with  $\|\alpha\| = 1$ .

**Proof.** The equivalence of (b) and (c) is obvious. We now prove the equivalence of (a) with (c). Let  $\phi_n(t)$  and  $\phi(t)$  denote, respectively, the characteristic functions of  $Z_n$  and  $Z$ . According to the multivariate version of Theorem 8 we have  $Z_n \xrightarrow{d} Z$  if and only if  $\phi_n(t) \rightarrow \phi(t)$  for all  $t = (t_1, \dots, t_k)' \in \mathbb{R}^k$ . Let  $\phi_n^\alpha(s)$  and  $\phi^\alpha(s)$  denote the characteristic functions of  $\alpha' Z_n$  and  $\alpha' Z$ , respectively. Again,  $\alpha' Z_n \xrightarrow{d} \alpha' Z$  if and only if  $\phi_n^\alpha(s) \rightarrow \phi^\alpha(s)$  for all  $s \in \mathbb{R}$ . Observe that for  $t \neq 0$  we have

$$\phi_n(t) = E(\exp(it' Z_n)) = E(\exp(is\alpha' Z_n)) = \phi_n^\alpha(s)$$

with  $\alpha = t/\|t\|$  and  $s = \|t\|$ . Note that  $\|\alpha\| = 1$ . Similarly,  $\phi(t) = \phi^\alpha(s)$ . Consequently,  $\phi_n(t) \rightarrow \phi(t)$  for all  $t \neq 0$  if and only if  $\phi_n^\alpha(s) \rightarrow \phi^\alpha(s)$  for all  $s \neq 0$  and all  $\alpha$  with  $\|\alpha\| = 1$ . Since  $\phi_n(0) = \phi(0) = 1$  and  $\phi_n^\alpha(0) = \phi^\alpha(0) = 1$ , the proof is complete observing that  $t = 0$  if and only if  $s = 0$ . ■

## 2.3 Convergence properties and transformations

We are often interested in the convergence properties of transformed random vectors or variables. In particular, suppose  $Z_n$  converges to  $Z$  in a certain mode, then given a function  $g$  we may ask the question whether or not  $g(Z_n)$  converges to  $g(Z)$  in the same mode. The following theorem answers the question in the affirmative, provided  $g$  is continuous (in the sense specified below). Part (a) of the theorem is commonly referred to as Slutsky's theorem.

**Theorem 14.<sup>9</sup>** Let  $Z_1, Z_2, \dots$ , and  $Z$  be random vectors in  $\mathbb{R}^k$ . Furthermore, let  $g : \mathbb{R}^k \rightarrow \mathbb{R}^s$  be a Borel-measurable function and assume that  $g$  is continuous with  $P_Z$ -probability one (where  $P_Z$  denotes the probability measure induced by  $Z$  on  $\mathbb{R}^k$ ).<sup>10</sup> Then

- (a)  $Z_n \xrightarrow{p} Z$  implies  $g(Z_n) \xrightarrow{p} g(Z)$ ,
- (b)  $Z_n \xrightarrow{\text{a.s.}} Z$  implies  $g(Z_n) \xrightarrow{\text{a.s.}} g(Z)$ ,
- (c)  $Z_n \xrightarrow{d} Z$  implies  $g(Z_n) \xrightarrow{d} g(Z)$ .

In the special case where  $Z = c$  is a constant or a vector of constants, the continuity condition on  $g$  in the above theorem only requires that the function  $g$  is continuous at  $c$ .

As special cases of Theorem 14 we have, for example, the following corollaries.

**Corollary 2.** Let  $W_n$  and  $V_n$  be sequences of  $k$ -dimensional random vectors. Suppose  $W_n \rightarrow W$  and  $V_n \rightarrow V$  i.p. [a.s.], then

$$W_n \pm V_n \rightarrow W \pm V \quad \text{i.p. [a.s.],}$$

$$W'_n V_n \rightarrow W' V \quad \text{i.p. [a.s.].}$$

In case  $k = 1$ ,

$$W_n/V_n \rightarrow W/V \quad \text{i.p. [a.s.]}$$

if  $V \neq 0$  with probability one, and where  $W_n/V_n$  is set to an arbitrary value on the event  $\{V_n = 0\}$ .<sup>11</sup>

**Proof.** The assumed convergence of  $W_n$  and  $V_n$  implies that  $Z_n = (W'_n, V'_n)'$  converges to  $Z = (W', V')'$  i.p. [a.s.] in view of Theorem 7. The corollary then follows from Theorem 14(a), (b) since the maps  $g_1(w, v) = w + v$ ,  $g_2(w, v) = w - v$ ,  $g_3(w, v) = w'v$  are continuous on all of  $\mathbb{R}^{2k}$ , and since the map  $g_4(w, v) = w/v$  if  $v \neq 0$  and  $g_4(w, v) = c$  for  $v = 0$  (with  $c$  arbitrary) is continuous on  $A = \mathbb{R} \times (\mathbb{R} - \{0\})$ , observing furthermore that  $P_Z(A) = 1$  provided  $V \neq 0$  with probability 1. ■

The proof of the following corollary is completely analogous.

**Corollary 3.** Let  $W_n$  and  $V_n$  be sequences of random matrices of fixed dimension. Suppose  $W_n \rightarrow W$  and  $V_n \rightarrow V$  i.p. [a.s.], then

$$W_n \pm V_n \rightarrow W \pm V \quad \text{i.p. [a.s.],}$$

$$W_n V_n \rightarrow WV \quad \text{i.p. [a.s.].}$$

Furthermore

$$W_n V_n^{-1} \rightarrow WV^{-1} \quad \text{and} \quad V_n^{-1} W_n \rightarrow V^{-1} W \quad \text{i.p. [a.s.]}$$

if  $V$  is nonsingular with probability one, and where  $W_n V_n^{-1}$  and  $V_n^{-1} W_n$  are set to an arbitrary matrix of appropriate dimension on the event  $\{V_n \text{ singular}\}$ . (The matrices are assumed to be of conformable dimensions.)

The following example shows that convergence in probability or almost surely in Corollaries 2 and 3 *cannot* be replaced by convergence in distribution.

**Example 5.** Let  $U \sim N(0, 1)$  and define  $W_n = U$  and  $V_n = (-1)^n U$ . Then

$$W_n + V_n = \begin{cases} 2U \sim N(0, 4) & \text{if } n \text{ is even} \\ 0 & \text{if } n \text{ is odd.} \end{cases}$$

Clearly,  $W_n + V_n$  does not converge in distribution, although  $W_n \xrightarrow{d} U$  and  $V_n \xrightarrow{d} U$ .

The reason behind this negative result is again the fact that convergence in distribution of the components of a random vector does in general not imply convergence in distribution of the entire random vector. Of course, if the entire random vector  $Z_n = (W'_n, V'_n)'$  converges in distribution to  $Z = (W', V)'$  then  $W_n \pm V_n \xrightarrow{d} W \pm V$ ,  $W'_n V'_n \xrightarrow{d} W' V$  as a consequence of Theorem 14; also, if  $k = 1$  and  $V \neq 0$  with probability 1, then  $W_n / V_n \xrightarrow{d} W / V$ .

However, there is an important special case in which we can conclude that  $Z_n = (W'_n, V'_n)' \xrightarrow{d} Z = (W', V)'$  from knowing that  $W_n \xrightarrow{d} W$  and  $V_n \xrightarrow{d} V$ : this is the case where  $V = c$  and  $c$  is a constant vector.

**Theorem 15.** Let  $W_n$  and  $V_n$  be sequences of  $k \times 1$  and  $l \times 1$  random vectors, respectively. Let  $W$  be a  $k \times 1$  random vector and let  $V = c$  be a constant vector in  $\mathbb{R}^l$ . Suppose  $W_n \xrightarrow{d} W$  and  $V_n \xrightarrow{d} c$  (or equivalently  $V_n \xrightarrow{p} c$  in light of Theorem 10). Then  $Z_n = (W'_n, V'_n)' \xrightarrow{d} Z = (W', V)' = (W', c)'$ .

**Proof.** Let  $\phi_n(t)$  and  $\phi(t)$  denote, respectively, the characteristic function of  $Z_n$  and  $Z$ . To show that  $Z_n \xrightarrow{d} Z$  it suffices to show that  $\phi_n(t) \rightarrow \phi(t)$  for all  $t \in \mathbb{R}^{k+l}$  in light of the multivariate version of Theorem 8. Let  $t = (s', u')'$  with  $s \in \mathbb{R}^k$  and  $u \in \mathbb{R}^l$  arbitrary. Observing that  $|\exp(is' W_n)| = 1 = |\exp(iu' c)|$ , we have

$$\begin{aligned}
|\phi_n(t) - \phi(t)| &= |E(e^{is'W_n}e^{iu'V_n} - e^{is'W}e^{iu'c})| \\
&\leq E[|e^{is'W_n}| |e^{iu'V_n} - e^{iu'c}|] + |e^{iu'c}| |E(e^{is'W_n} - e^{is'W})| \\
&\leq E|e^{iu'V_n} - e^{iu'c}| + |E(e^{is'W_n} - e^{is'W})| \\
&= E|e^{iu'V_n} - e^{iu'c}| + |\phi_n^W(s) - \phi^W(s)|,
\end{aligned} \tag{10.6}$$

where  $\phi_n^W(s)$  and  $\phi^W(s)$  denote, respectively, the characteristic function of  $W_n$  and  $W$ . Since  $V_n \xrightarrow{p} c$  it follows from Theorem 14 that  $\exp(iu'V_n) - \exp(iu'c) \xrightarrow{p} 0$ . Observing that  $|\exp(iu'V_n) - \exp(iu'c)| \leq 2$  it follows furthermore from Theorem 6 that  $E|\exp(iu'V_n) - \exp(iu'c)| \rightarrow 0$ . By assumption  $W_n \xrightarrow{d} W$ . It then follows again from the multivariate version of Theorem 8 that  $\phi_n^W(s) \rightarrow \phi^W(s)$ . Thus both terms in the last line of (10.6) converge to zero, and hence  $\phi_n(t) \rightarrow \phi(t)$ . ■

Given Theorem 15 the following result follows immediately from Theorem 14.

**Corollary 4.** Let  $W_n$  and  $V_n$  be sequences of  $k \times 1$  and  $l \times 1$  random vectors, respectively. Let  $W$  be a  $k \times 1$  random vector and  $c$  a constant vector in  $\mathbb{R}^l$ . Suppose  $W_n \xrightarrow{d} W$  and  $V_n \xrightarrow{d} c$  (or equivalently  $V_n \xrightarrow{p} c$ ). Let  $g : \mathbb{R}^k \times \mathbb{R}^l \rightarrow \mathbb{R}^s$  be a Borel measurable function and assume that  $g$  is continuous in every point of  $A \times \{c\}$  where  $A \subseteq \mathbb{R}^k$  satisfies  $P(W \in A) = 1$ . Then  $g(W_n, V_n) \xrightarrow{d} g(W, c)$ .

As a further corollary we have the following useful results.

**Corollary 5.** Let  $W_n$  and  $V_n$  be sequences of  $k \times 1$  and  $l \times 1$  random vectors, let  $A_n$  and  $B_n$  be sequences of  $l \times k$  and  $k \times k$  random matrices, respectively. Furthermore, let  $W$  be a  $k \times 1$  random vector, let  $c$  be a  $l \times 1$  nonstochastic vector, and let  $A$  and  $B$  be some nonstochastic  $l \times k$  and  $k \times k$  matrices.

(a) For  $k = l$

$$\begin{aligned}
W_n \xrightarrow{d} W, V_n \xrightarrow{p} c \quad \text{implies} \quad W_n \pm V_n \xrightarrow{d} W \pm c \\
W'_n V_n \xrightarrow{d} W'c.
\end{aligned}$$

(If  $c = 0$ , then  $W'_n V_n \xrightarrow{d} 0$  and hence also  $W'_n V_n \xrightarrow{p} 0$ ).

(b) For  $k = l = 1$

$$\begin{aligned}
W_n \xrightarrow{d} W, V_n \xrightarrow{p} c \quad \text{implies} \quad W_n/V_n \xrightarrow{d} W/c \quad \text{if } c \neq 0, \\
V_n/W_n \xrightarrow{d} c/W \quad \text{if } P(W = 0) = 0.
\end{aligned}$$

(c)  $W_n \xrightarrow{d} W, V_n \xrightarrow{p} c, A_n \xrightarrow{p} A \quad \text{implies} \quad A_n W_n + V_n \xrightarrow{d} AW + c,$

(d)  $W_n \xrightarrow{d} W, B_n \xrightarrow{p} B \quad \text{implies} \quad W'_n B_n W_n \xrightarrow{d} W'BW.$

Of course, if in the above corollary  $W \sim N(\mu, \Sigma)$ , then  $AW + c \sim N(A\mu + c, A\Sigma A')$ . If  $W \sim N(0, I_k)$  and  $B$  is idempotent of rank  $p$ , then  $W'BW \sim \chi^2(p)$ .

## 2.4 Orders of magnitude

In determining the limiting behavior of sequences of random variables it is often helpful to employ notions of orders of relative magnitudes. We start with a review of the concepts of order of magnitudes for sequences of real numbers.

**Definition 6.** (Order of magnitude of a sequence of real numbers) Let  $a_n$  be a sequence of real numbers and let  $c_n$  be a sequence of positive real numbers. We then say  $a_n$  is at most of order  $c_n$ , and write  $a_n = O(c_n)$ , if there exists a constant  $M < \infty$  such that  $c_n^{-1} |a_n| \leq M$  for all  $n \in \mathbb{N}$ . We say  $a_n$  is of smaller order than  $c_n$ , and write  $a_n = o(c_n)$ , if  $c_n^{-1} |a_n| \rightarrow 0$  as  $n \rightarrow \infty$ . (The definition extends to vectors and matrices by applying the definition to their norm.)

The following results concerning the algebra of order in magnitude operations are often useful.

**Theorem 16.** Let  $a_n$  and  $b_n$  be sequences of real numbers, and let  $c_n$  and  $d_n$  be sequences of positive real numbers.

- (a) If  $a_n = o(c_n)$  and  $b_n = o(d_n)$ , then  $a_n b_n = o(c_n d_n)$ ,  $|a_n|^s = o(c_n^s)$  for  $s > 0$ ,  
 $a_n + b_n = o(\max\{c_n, d_n\}) = o(c_n + d_n)$ .
- (b) If  $a_n = O(c_n)$  and  $b_n = O(d_n)$ , then  $a_n b_n = O(c_n d_n)$ ,  $|a_n|^s = O(c_n^s)$  for  $s > 0$ ,  
 $a_n + b_n = O(\max\{c_n, d_n\}) = O(c_n + d_n)$ .
- (c) If  $a_n = o(c_n)$  and  $b_n = O(d_n)$ , then  $a_n b_n = o(c_n d_n)$ .

We now generalize the concept of order of magnitude from sequences of real numbers to sequences of random variables.

**Definition 7.** (Order in probability of a sequence of random variables) Let  $Z_n$  be a sequence of random variables, and let  $c_n$  be a sequence of positive real numbers. We then say  $Z_n$  is at most of order  $c_n$  in probability, and write  $Z_n = O_p(c_n)$ , if for every  $\varepsilon > 0$  there exists a constant  $M_\varepsilon < \infty$  such that  $P(c_n^{-1} |Z_n| \geq M_\varepsilon) \leq \varepsilon$ . We say  $Z_n$  is of smaller order in probability than  $c_n$ , and write  $Z_n = o_p(c_n)$ , if  $c_n^{-1} |Z_n| \xrightarrow{P} 0$  as  $n \rightarrow \infty$ . (The definition extends to vectors and matrices by applying the definition to their norm.)

The algebra of order in probability operations  $O_p$  and  $o_p$  is identical to that of order in magnitude operations  $O$  and  $o$  presented in the theorem above; see, e.g., Fuller (1976, p. 184).

A sequence of random variables  $Z_n$  that is  $O_p(1)$  is also said to be “stochastically bounded” or “bounded in probability”. The next theorem gives sufficient conditions for a sequence to be stochastically bounded.

### Theorem 17.

- (a) Suppose  $E |Z_n|^r = O(1)$  for some  $r > 0$ , then  $Z_n = O_p(1)$ .
- (b) Suppose  $Z_n \xrightarrow{d} Z$ , then  $Z_n = O_p(1)$ .

**Proof.** Part (a) follows readily from Markov's inequality. To prove part (b) fix  $\varepsilon > 0$ . Now choose  $M_\varepsilon^*$  such that  $F$  is continuous at  $-M_\varepsilon^*$  and  $M_\varepsilon^*$ , and  $F(-M_\varepsilon^*) \leq \varepsilon/4$  and  $F(M_\varepsilon^*) \geq 1 - \varepsilon/4$ . Since every CDF has at most a countable number of discontinuity points, such a choice is possible. By assumption  $F_n(z) \rightarrow F(z)$  for all continuity points of  $F$ . Let  $n_\varepsilon$  be such that for all  $n \geq n_\varepsilon$

$$|F_n(-M_\varepsilon^*) - F(-M_\varepsilon^*)| \leq \varepsilon/4$$

and

$$|F_n(M_\varepsilon^*) - F(M_\varepsilon^*)| \leq \varepsilon/4.$$

Then for  $n \geq n_\varepsilon$

$$\begin{aligned} P(|Z_n| \geq M_\varepsilon^*) &\leq F_n(-M_\varepsilon^*) - F_n(M_\varepsilon^*) + 1 \\ &\leq F(-M_\varepsilon^*) - F(M_\varepsilon^*) + 1 + \varepsilon/2 \leq \varepsilon. \end{aligned}$$

Since  $\lim_{M \rightarrow \infty} P(|Z_i| \geq M) = 0$  for each  $i \in \mathbb{N}$  we can find an  $M_\varepsilon^{**}$  such that  $P(|Z_i| \geq M_\varepsilon^{**}) \leq \varepsilon$  for  $i = 1, \dots, n_\varepsilon - 1$ . Now let  $M_\varepsilon = \max\{M_\varepsilon^*, M_\varepsilon^{**}\}$ . Then  $P(|Z_n| \geq M_\varepsilon) \leq \varepsilon$  for all  $n \in \mathbb{N}$ . ■

### 3 LAWS OF LARGE NUMBERS

Let  $Z_t$ ,  $t \in \mathbb{N}$ , be a sequence of random variables with  $EZ_t = \mu_t$ . Furthermore let  $\bar{Z}_n = n^{-1} \sum_{t=1}^n Z_t$  denote the sample mean, and let  $\bar{\mu}_n = E\bar{Z}_n = n^{-1} \sum_{t=1}^n \mu_t$ . A law of large numbers (LLN) then specifies conditions under which

$$\bar{Z}_n - E\bar{Z}_n = n^{-1} \sum_{t=1}^n (Z_t - \mu_t)$$

converges to zero either in probability or almost surely. If the convergence is in probability we speak of a weak LLN, if the convergence is almost surely we speak of a strong LLN. We note that in applications the random variables  $Z_t$  may themselves be functions of other random variables.

The usefulness of LLNs stems from the fact that many estimators can be expressed as (continuous) functions of sample averages of random variables, or differ from such a function only by a term that can be shown to converge to zero i.p. or a.s. Thus to establish the probability or almost sure limit of such an estimator we may try to establish in a first step the limits for the respective averages by means of LLNs. In a second step we may then use Theorem 14 to derive the actual limit for the estimator.

**Example 6.** As an illustration consider the linear regression model  $y_t = x_t \theta + \varepsilon_t$ ,  $t = 1, \dots, n$ , where  $y_t$ ,  $x_t$  and  $\varepsilon_t$  are all scalar and denote the dependent variable, the independent variable, and the disturbance term in period  $t$ . The ordinary least squares estimator for the parameter  $\theta$  is then given by

$$\hat{\theta}_n = \frac{\sum_{t=1}^n x_t y_t}{\sum_{t=1}^n x_t^2} = \theta + \frac{n^{-1} \sum_{t=1}^n x_t \varepsilon_t}{n^{-1} \sum_{t=1}^n x_t^2}$$

and thus  $\hat{\theta}_n$  is seen to be a function of the sample averages of  $x_t \varepsilon_t$  and  $x_t^2$ .

### 3.1 Independent processes

In this subsection we discuss LLNs for independent processes.

**Theorem 18.**<sup>12</sup> (Kolmogorov's strong LLN for iid random variables) Let  $Z_t$  be a sequence of identically and independently distributed (iid) random variables with  $E|Z_1| < \infty$  and  $EZ_1 = \mu$ . Then  $\bar{Z}_n \xrightarrow{a.s.} \mu$  (and hence  $\bar{Z}_n \xrightarrow{i.p.} \mu$ ) as  $n \rightarrow \infty$ .

We have the following trivial but useful corollary.

**Corollary 6.** Let  $Z_t$  be a sequence of iid random variables, and let  $f$  be a Borel-measurable real function satisfying  $E|f(Z_1)| < \infty$ , then  $n^{-1} \sum_{t=1}^n f(Z_t) \xrightarrow{a.s.} Ef(Z_1)$  as  $n \rightarrow \infty$ .

The corollary can, in particular, be used to establish convergence of sample moments of, say, order  $p$  to the corresponding population moment by choosing  $f(Z_t) = Z_t^p$ .

We now derive the probability limit of the ordinary least squares estimator considered in Example 6 as an illustration.

**Example 7.** Assume the setup of Example 6. Assume furthermore that the processes  $x_t$  and  $\varepsilon_t$  are iid with  $Ex_t^2 = Q_x$ ,  $0 < Q_x < \infty$ ,  $E|\varepsilon_t| < \infty$ , and  $E\varepsilon_t = 0$ , and that the two processes are independent of each other. Then  $x_t \varepsilon_t$  is iid, has finite expectation and satisfies  $Ex_t \varepsilon_t = Ex_t E\varepsilon_t = 0$ . Hence it follows from Theorem 18 that  $n^{-1} \sum_{t=1}^n x_t \varepsilon_t \xrightarrow{a.s.} 0$ . Corollary 6 implies  $n^{-1} \sum_{t=1}^n x_t^2 \xrightarrow{a.s.} Q_x$ . Applying Theorem 14 then yields  $\hat{\theta}_n \xrightarrow{a.s.} \theta + 0/Q_x = \theta$ .

The assumption in Theorem 18 that the random variables are identically distributed can be relaxed at the expense of maintaining additional assumptions on the second moments.

**Theorem 19.**<sup>13</sup> (Kolmogorov's strong LLN for ID random variables) Let  $Z_t$  be a sequence of independently distributed (ID) random variables with  $EZ_t = \mu_t$  and  $\text{var}(Z_t) = \sigma_t^2 < \infty$ . Suppose  $\sum_{t=1}^{\infty} \sigma_t^2/t^2 < \infty$ . Then  $\bar{Z}_n - \bar{\mu}_n \xrightarrow{a.s.} 0$  as  $n \rightarrow \infty$ .

The condition  $\sum_{t=1}^n \sigma_t^2/t^2 < \infty$  puts a restriction on the permissible variation in the  $\sigma_t^2$ . For example, it is satisfied if the sequence  $\sigma_t^2$  is bounded.

### 3.2 Dependent processes

The following weak LLN follows immediately from Corollary 1. In contrast to the above LLNs this theorem does not require the variables to be independently distributed, but only requires uncorrelatedness.

**Theorem 20.** (Chebychev's weak LLN for uncorrelated random variables) Let  $Z_t$  be a sequence of uncorrelated random variables with  $EZ_t = \mu_t$  and  $\text{var}(Z_t) = \sigma_t^2 < \infty$ . Suppose  $\text{var}(\bar{Z}_n) = n^{-2} \sum_{t=1}^n \sigma_t^2 \rightarrow 0$  as  $n \rightarrow \infty$ , then  $\bar{Z}_n - \bar{\mu}_n \xrightarrow{P} 0$ .

The condition on the variance in Theorem 20 is weaker than the corresponding condition in Theorem 19 in view of Kronecker's lemma; see, e.g., Shirayev (1984, p. 365). The condition is clearly satisfied if the sequence  $\sigma_t^2$  is bounded.

A class of dependent processes that is important in econometrics and statistics is the class of martingale difference sequences. For example, the score of the maximum likelihood estimator evaluated at the true parameter value represents (under mild regularity conditions) a martingale difference sequence.

**Definition 8.** (Martingale difference sequence) Let  $\mathcal{F}_t$ ,  $t \geq 0$ , be a sequence of  $\sigma$ -fields such that  $\mathcal{F}_0 \subseteq \mathcal{F}_1 \subseteq \dots \subseteq \mathcal{F}$ . Let  $Z_t$ ,  $t \geq 1$ , be a sequence of random variables, then  $Z_t$  is said to be a martingale difference sequence (wrt the sequence  $\mathcal{F}_t$ ), if  $Z_t$  is  $\mathcal{F}_t$ -measurable,  $E|Z_t| < \infty$  and

$$E(Z_t | \mathcal{F}_{t-1}) = 0$$

for all  $t \geq 1$ .

We note that if  $Z_t$  is a martingale difference sequence then  $E(Z_t) = E(E(Z_t | \mathcal{F}_{t-1})) = 0$  by the law of iterated expectations. Furthermore, since  $E(Z_t Z_{t+k}) = E(Z_t E(Z_{t+k} | \mathcal{F}_{t+k-1})) = 0$  for  $k \geq 1$ , we see that every martingale difference sequence is uncorrelated, provided the second moments are finite. We also note, if  $Z_t$  is a martingale difference sequence wrt the  $\sigma$ -fields  $\mathcal{F}_t$ , then it is also a martingale difference sequence wrt the  $\sigma$ -fields  $\mathcal{R}_t$  where  $\mathcal{R}_t$  is generated by  $\{Z_t, Z_{t-1}, \dots, Z_1\}$  and  $\mathcal{R}_0 = \{\emptyset, \Omega\}$ .

We now present a strong LLN for martingale difference sequences.

**Theorem 21.<sup>14</sup>** Let  $Z_t$  be a martingale difference sequence with  $\text{var}(Z_t) = \sigma_t^2 < \infty$ . Suppose  $\sum_{t=1}^n \sigma_t^2 / t^2 < \infty$ . Then  $\bar{Z}_n \xrightarrow{a.s.} 0$  as  $n \rightarrow \infty$ .

The above LLN contains Kolmogorov's strong LLN for independent random variables as a special case (with  $Z_t$  replaced by  $Z_t - \mu_t$ ).

Many processes of interest in econometrics and statistics are correlated, and hence are not covered by the above LLNs. In the following we present a strong LLN for strictly stationary processes, which allows for a wide range of correlation structures.

**Definition 9.** (Strict stationarity) The sequence of random variables  $Z_t$ ,  $t \geq 1$ , is said to be strictly stationary if  $(Z_1, Z_2, \dots, Z_n)$  has the same distribution as  $(Z_{1+k}, Z_{2+k}, \dots, Z_{n+k})$  for all  $k \geq 1$  and  $n \geq 1$ .

**Definition 10.<sup>15</sup>** (Invariance and ergodicity of strictly stationary sequences) Let  $Z_t$ ,  $t \geq 1$ , be a strictly stationary sequence.

(a) Consider the event

$$A = \{\omega \in \Omega : (Z_1(\omega), Z_2(\omega), \dots) \in B\}$$

with  $B \in \mathcal{B}^\infty$ , where  $\mathcal{B}^\infty$  are the Borel sets of  $\mathbb{R}^\infty$ . Then  $A$  is said to be invariant if

$$A = \{\omega \in \Omega : (Z_{1+k}(\omega), Z_{2+k}(\omega), \dots) \in B\}$$

for all  $k \geq 1$ .

(b) The sequence  $Z_t$  is ergodic if every invariant event has probability one or zero.

We note that every iid sequence of random variables is strictly stationary and ergodic. Furthermore, if  $Z_t$  is strictly stationary and ergodic, and  $g : \mathbb{R}^\infty \rightarrow \mathbb{R}$  is measurable, then the sequence  $Y_t$  with  $Y_t = g(Z_t, Z_{t+1}, \dots)$  is again strictly stationary and ergodic.

We can now give the following strong LLN, which is often referred to as the Ergodic Theorem. This theorem contains Kolmogorov's strong LLN for iid random variables as a special case.

**Theorem 22.<sup>16</sup>** Let  $Z_t$  be a strictly stationary and ergodic sequence with  $E|Z_1| < \infty$  and  $EZ_1 = \mu$ . Then  $\bar{Z}_n \xrightarrow{a.s.} \mu$  as  $n \rightarrow \infty$ .

There is a large literature on LLNs for dependent processes beside the LLNs presented above. LLNs for weakly stationary processes, including linear processes and ARMA (autoregressive moving average) processes, can be found in Hannan (1970, ch. IV.3); see also Phillips and Solo (1992). Important classes of dependent processes considered in econometrics and statistics are  $\alpha$ -mixing,  $\phi$ -mixing, near epoch dependent and  $L_p$ -approximable processes. LLNs for such processes are discussed in some detail in, e.g., Davidson (1994, Part IV) and Pötscher and Prucha (1997, ch. 6), and in the references given therein; see also Davidson and de Jong (1997) for recent extensions.

### 3.3 Uniform laws of large numbers

In the previous sections we have been concerned with various notions of convergence for sequences of random variables and random vectors. Sometimes one is confronted with sequences of random functions, say  $Q_n(\theta)$ , that depend on a parameter vector  $\theta$  contained in some parameter space  $\Theta$ . That is,  $Q_n(\theta)$  is a random variable for each fixed  $\theta \in \Theta$ .<sup>17</sup> As an example of a random function

consider, for example, the loglikelihood function of iid random variables  $Z_1, \dots, Z_n$  with a density that depends on a parameter vector  $\theta$ :

$$Q_n(\theta) = \frac{1}{n} \sum_{t=1}^n q(Z_t, \theta) \quad (10.7)$$

where  $q$  is the logarithm of the density of  $Z_t$ . Clearly, for every fixed  $\theta \in \Theta$  we can apply the notions of convergence for random variables discussed above to  $Q_n(\theta)$ . However, for many purposes these notions of “pointwise” convergence are not sufficient and stronger notions of convergence are needed. For example, those stronger notions of convergence are often useful in proving consistency of maximum likelihood estimators: in many cases an explicit expression for the maximum likelihood estimator will not be available. In those cases one may try to deduce the convergence behavior of the estimator from the convergence behavior of the loglikelihood objective function  $Q_n(\theta)$ . By Kolmogorov’s LLN we have

$$Q_n(\theta) \xrightarrow{a.s.} Q(\theta) \text{ for all } \theta \in \Theta, \quad (10.8)$$

where  $Q(\theta) = E q(Z_t, \theta)$ , provided  $E |q(Z_t, \theta)| < \infty$ . A well-established result from the theory of maximum likelihood estimation tells us furthermore that the limiting objective function  $Q(\theta)$  is uniquely maximized at the true parameter value, say  $\theta_0$ , provided  $\theta_0$  is identified. It is tempting to conclude that these two facts imply a.s. convergence of the maximum likelihood estimators, i.e. of the maximizers of the objective functions  $Q_n(\theta)$ , to  $\theta_0$ . Unfortunately, this line of reasoning is not conclusive in general, as can be seen from counter examples; see, e.g., Amemiya (1985, p. 109). However, this line of reasoning can be salvaged if we can establish not only “pointwise” convergence a.s., i.e. (10.8), but even uniform convergence a.s., i.e.,

$$\sup_{\theta \in \Theta} |Q_n(\theta) - Q(\theta)| \xrightarrow{a.s.} 0, \quad (10.9)$$

and if, for example,  $\Theta$  is compact and  $Q$  is continuous.<sup>18</sup>

The above discussion motivates interest in results that establish uniform convergence of random functions  $Q_n(\theta)$ . In the important special case where  $Q_n(\theta) = n^{-1} \sum_{t=1}^n q(Z_t, \theta)$  and  $Q(\theta) = EQ(\theta)$  such results are called uniform laws of large numbers (ULLNs). We next present an ULLN for functions of iid random variables.

**Theorem 23.**<sup>19</sup> Let  $Z_t$  be a sequence of identically and independently distributed  $k \times 1$  random vectors, let  $\Theta$  be a compact subset of  $\mathbb{R}^p$ , and let  $q$  be a real valued function on  $\mathbb{R}^k \times \Theta$ . Furthermore, let  $q(., \theta)$  be Borel-measurable for each  $\theta \in \Theta$ , and let  $q(z, .)$  be continuous for each  $z \in \mathbb{R}^k$ . If  $E \sup_{\theta \in \Theta} |q(Z_t, \theta)| < \infty$ , then

$$\sup_{\theta \in \Theta} \left| \frac{1}{n} \sum_{t=1}^n [q(Z_t, \theta) - Eq(Z_t, \theta)] \right| \xrightarrow{a.s.} 0 \text{ as } n \rightarrow \infty,$$

i.e.  $Q_n(\theta) = n^{-1} \sum_{t=1}^n q(Z_t, \theta)$  satisfies a ULLN.

The theorem also holds if the assumption that  $Z_t$  is iid is replaced by the assumption that  $Z_t$  is stationary and ergodic; see, e.g., Pötscher and Prucha (1986, Lemma A.2). Uniform laws of large numbers that also cover functions of dependent and heterogeneous random vectors are, for example, given in Andrews (1987) and Pötscher and Prucha (1989); for additional references see Pötscher and Prucha (1997, ch. 5).

## 4 CENTRAL LIMIT THEOREMS

Let  $Z_t$ ,  $t \in \mathbb{N}$ , be a sequence of iid random variables with  $EZ_t = \mu$  and  $\text{var}(Z_t) = \sigma^2$ ,  $0 < \sigma^2 < \infty$ . Let  $\bar{Z}_n = n^{-1} \sum_{t=1}^n Z_t$  denote the sample mean. By Kolmogorov's strong LLN for iid random variables (Theorem 18) it then follows that  $\bar{Z}_n - E\bar{Z}_n$  converges to zero a.s. and hence i.p. This implies that the limiting distribution of  $\bar{Z}_n - E\bar{Z}_n$  is degenerate at zero, and thus no insight is gained from this limiting distribution regarding the shape of the distribution of the sample mean for finite  $n$ ; compare the discussion at the beginning of Section 2.2. Suppose we consider the rescaled quantity

$$\sqrt{n}(\bar{Z}_n - E\bar{Z}_n) = n^{-1/2} \sum_{t=1}^n (Z_t - \mu). \quad (10.10)$$

Then the variance of the rescaled expression is  $\sigma^2 > 0$  for all  $n$ , indicating that its limiting distribution will not be degenerate. Theorems that provide results concerning the limiting distribution of expressions like (10.10) are called central limit theorems (CLTs). Rather than to center the respective random variables, as is done in (10.10), we assume in the following, without loss of generality, that the respective random variables have mean zero.

### 4.1 Independent processes

#### SOME CLASSICAL CLTs

In this subsection we will present several classical CLTs, starting with the Lindeberg–Lévy CLT.

**Theorem 24.**<sup>20</sup> (Lindeberg–Lévy CLT) Let  $Z_t$  be a sequence of iid random variables with  $EZ_t = 0$  and  $\text{var}(Z_t) = \sigma^2 < \infty$ . Then  $n^{-1/2} \sum_{t=1}^n Z_t \xrightarrow{d} N(0, \sigma^2)$ . (In case  $\sigma^2 = 0$  the limit  $N(0, 0)$  should be interpreted as the degenerate distribution having all its probability mass concentrated at zero.<sup>21</sup>)

Of course, if  $\sigma^2 > 0$  the conclusion of the theorem can be written equivalently as  $n^{-1/2} \sum_{t=1}^n Z_t / \sigma \xrightarrow{d} N(0, 1)$ . Extensions of Theorem 24 and of any of the following central limit theorems to the vector case are readily obtained using the Cramér–Wold device (Theorem 13). To illustrate this we exemplarily extend Theorem 24 to the vector case.

**Example 8.** Let  $Z_t$  be a sequence of iid  $k$ -dimensional random vectors with zero mean and finite variance covariance matrix  $\Sigma$ . Let  $\xi_n = n^{-1/2} \sum_{t=1}^n Z_t$ , let  $\xi \sim N(0, \Sigma)$  (where  $N(0, \Sigma)$  denotes a singular normal distribution if  $\Sigma$  is singular), and let  $\alpha$  be some element of  $\mathbb{R}^k$ . Now consider the scalar random variables  $\alpha' \xi_n = n^{-1/2} \sum_{t=1}^n \alpha' Z_t$ . Clearly the summands  $\alpha' Z_t$  are iid with mean zero and variance  $\alpha' \Sigma \alpha$ . It hence follows from Theorem 24 that  $\alpha' \xi_n$  converges in distribution to  $N(0, \alpha' \Sigma \alpha)$ . Of course  $\alpha' \xi \sim N(0, \alpha' \Sigma \alpha)$ , and hence  $\alpha' \xi_n \xrightarrow{d} \alpha' \xi$ . Since  $\alpha$  was arbitrary it follows from Theorem 13 that  $\xi_n \xrightarrow{d} \xi$ , which shows that the random vector  $n^{-1/2} \sum_{t=1}^n Z_t$  converges in distribution to  $N(0, \Sigma)$ .

Theorem 24 postulates that the random variables  $Z_t$  are iid. The following theorems relax this assumption to independence. It proves helpful to define

$$\sigma_{(n)}^2 = \sum_{t=1}^n \sigma_t^2 \quad (10.11)$$

where  $\sigma_t^2 = \text{var}(Z_t)$ . For independent  $Z_t$ s clearly  $\sigma_{(n)}^2 = n^2 \text{var}(\bar{Z}_n)$ , and in case the  $Z_t$ s are iid with variance  $\sigma^2$  we have  $\sigma_{(n)}^2 = n\sigma^2$ . To connect Theorem 24 with the subsequent CLTs observe that within the context of Theorem 24 we have  $n^{-1/2} \sum_{t=1}^n Z_t / \sigma = \sum_{t=1}^n Z_t / \sigma_{(n)}$  (given  $\sigma^2 > 0$ ).

**Theorem 25.<sup>22</sup>** (Lindeberg–Feller CLT) Let  $Z_t$  be a sequence of independent random variables with  $EZ_t = 0$  and  $\text{var}(Z_t) = \sigma_t^2 < \infty$ . Suppose that  $\sigma_{(n)}^2 > 0$ , except for finitely many  $n$ . If for every  $\varepsilon > 0$

$$\lim_{n \rightarrow \infty} \frac{1}{\sigma_{(n)}^2} \sum_{t=1}^n E[|Z_t|^2 \mathbf{1}(|Z_t| \geq \varepsilon \sigma_{(n)})] = 0, \quad (\text{L})$$

then  $\sum_{t=1}^n Z_t / \sigma_{(n)} \xrightarrow{d} N(0, 1)$ .

Condition (L) is called the Lindeberg condition. The next theorem employs in place of the Lindeberg condition a condition that is stronger but easier to verify.

**Theorem 26.<sup>23</sup>** (Lyapounov CLT) Let  $Z_t$  be a sequence of independent random variables with  $EZ_t = 0$  and  $\text{var}(Z_t) = \sigma_t^2 < \infty$ . Suppose that  $\sigma_{(n)}^2 > 0$ , except for finitely many  $n$ . If for some  $\delta > 0$

$$\lim_{n \rightarrow \infty} \sum_{t=1}^n E|Z_t/\sigma_{(n)}|^{2+\delta} = 0, \quad (\text{P})$$

then  $\sum_{t=1}^n Z_t / \sigma_{(n)} \xrightarrow{d} N(0, 1)$ .

Condition (P) is called the Lyapounov condition. Condition (P) implies condition (L). It is readily seen that a sufficient condition for (P) is that

$$n^{-1}\sigma_{(n)}^2 = n^{-1} \sum_{t=1}^n \sigma_t^2 \geq \text{const} > 0$$

for  $n$  sufficiently large and that

$$\lim_{n \rightarrow \infty} \sum_{t=1}^n E|Z_t/\sqrt{n}|^{2+\delta} = 0.$$

In turn, sufficient conditions for those two conditions are, respectively,

$$\lim_{n \rightarrow \infty} n^{-1}\sigma_{(n)}^2 = \psi, \quad 0 < \psi < \infty, \quad (10.12)$$

and

$$\sup_n n^{-1} \sum_{t=1}^n E|Z_t|^{2+\delta} < \infty. \quad (10.13)$$

We note that the conclusions of Theorems 25 and 26 can be stated equivalently as  $n^{-1/2} \sum_{t=1}^n Z_t \xrightarrow{d} N(0, \psi)$ , whenever (10.12) holds. In this context we also make the trivial observation, that for a sequence of independent random variables  $Z_t$  with zero mean and finite variances  $\sigma_t^2 \geq 0$  the condition  $n^{-1}\sigma_{(n)}^2 \rightarrow \psi = 0$  implies  $n^{-1/2} \sum_{t=1}^n Z_t \xrightarrow{P} 0$  (Corollary 1), which can also be rewritten as  $n^{-1/2} \sum_{t=1}^n Z_t \xrightarrow{d} N(0, \psi), \psi = 0$ .

The above CLTs were given for sequences of random variables  $(Z_t, t \geq 1)$ . They can be readily generalized to cover triangular arrays of random variables  $(Z_{tn}, 1 \leq t \leq n, n \geq 1)$ . In fact Theorems 25 and 26 hold with  $Z_t$  replaced by  $Z_{tn}$  and  $\sigma_t^2$  replaced by  $\sigma_{tn}^2$ ; see, e.g., Billingsley (1979, pp. 310–12).

The need for CLTs for triangular arrays arises frequently in econometrics. One example is the derivation of the limiting distribution of the least squares estimator when different regressors grow at different rates. In this case one can still obtain a limiting normal distribution for the least squares estimator if the usual  $\sqrt{n}$ -norming is replaced with a normalization by an appropriate diagonal matrix. In essence, this entails renormalizing the  $i$ th regressor by the square root of  $\sum_{t=1}^n x_{ti}^2$ , whose obvious dependence on  $n$  leads to the consideration of a CLT for quantities of the form  $\sum_{t=1}^n c_{tn} u_t$  with  $u_t$  iid; see Theorem 28 below.

### CLTs FOR REGRESSION ANALYSIS

In this subsection we present some CLTs that are geared towards regression analysis. As discussed above, within this context we will often need CLTs for a sequence of iid random variables multiplied by some time-varying scale factors, that may also depend on the sample size. We first give a general CLT that covers such situations as a corollary to the Lindeberg–Feller CLT.

**Theorem 27.<sup>24</sup>** Let  $\underline{Z}_t$  be a sequence of iid random variables with  $E\underline{Z}_t = 0$  and  $\text{var}(\underline{Z}_t) = 1$ . Furthermore, let  $(\sigma_{tn}, 1 \leq t \leq n, n \geq 1)$  be a triangular array of real

numbers, and define the triangular array  $Z_{tn}$  by  $Z_{tn} = \sigma_{tn} Z_t$ . Suppose that  $\sigma_{(n)}^2 = \sum_{t=1}^n \sigma_{tn}^2 > 0$ , except for finitely many  $n$ . If

$$\lim_{n \rightarrow \infty} \frac{\max_{1 \leq t \leq n} \sigma_{tn}^2}{\sum_{t=1}^n \sigma_{tn}^2} = 0, \quad (M)$$

then  $\sum_{t=1}^n Z_{tn} / \sigma_{(n)} \xrightarrow{d} N(0, 1)$ .

All of the subsequent CLTs in this section are based on Theorem 27. Explicit proofs are given in a longer mimeographed version of this article, which is available from the authors upon request.

**Theorem 28.** Let  $u_t$ ,  $t \geq 1$ , be a sequence of iid random variables with  $E u_t = 0$  and  $E u_t^2 = \sigma^2 < \infty$ . Let  $X_n$ ,  $n \geq 1$ , with  $X_n = (x_{ti})$  be a sequence of real non-stochastic  $n \times k$  matrices with

$$\lim_{n \rightarrow \infty} \frac{\max_{1 \leq t \leq n} x_{ti}^2}{\sum_{t=1}^n x_{ti}^2} = 0 \quad \text{for } i = 1, \dots, k, \quad (10.14)$$

where it is assumed that  $\sum_{t=1}^n x_{ti}^2 > 0$  for all but finitely many  $n$ . Define  $W_n = X_n S_n^{-1}$  where  $S_n$  is a  $k \times k$  diagonal matrix with the  $i$ th diagonal element equal to  $[\sum_{t=1}^n x_{ti}^2]^{1/2}$ , and assume that  $\lim_{n \rightarrow \infty} W_n' W_n = \Phi$  is finite. Let  $u_n = [u_1, \dots, u_n]'$ , then  $W_n' u_n \xrightarrow{d} N(0, \sigma^2 \Phi)$ .

The above theorem is given in Amemiya (1985, p. 97), for the case of nonsingular  $\sigma^2 \Phi$ .<sup>25</sup> The theorem allows for trending (nonstochastic) regressors. For example, (10.14) holds for  $x_{ti} = t^p$ ,  $p > 0$ . We note that in case of a single regressor  $W_n' W_n = \Phi = 1$ .

**Theorem 29.** Let  $u_t$ ,  $t \geq 1$ , be a sequence of iid random variables with  $E u_t = 0$  and  $E u_t^2 = \sigma^2 < \infty$ . Let  $X_n$ ,  $n \geq 1$ , with  $X_n = (x_{ti})$  be a sequence of real nonstochastic  $n \times k$  matrices with  $\lim_{n \rightarrow \infty} n^{-1} X_n' X_n = Q$  finite. Let  $u_n = [u_1, \dots, u_n]'$ , then  $n^{-1/2} X_n' u_n \xrightarrow{d} N(0, \sigma^2 Q)$ .

The theorem is, for example, given in Theil (1971, p. 380), for the case of non-singular  $\sigma^2 Q$ . The theorem does not require that the elements of  $X_n$  are bounded in absolute value, as is often assumed in the literature.

We now use Theorems 28 and 29 to exemplarily give two asymptotic normality results for the least squares estimator.

**Example 9.** (Asymptotic normality of the least squares estimator) Consider the linear regression model

$$y_t = \sum_{i=1}^k x_{ti} \beta_i + u_t, \quad t \geq 1.$$

Suppose  $u_t$  and  $X_n = (x_{ti})$  satisfy the assumption of Theorem 28. Assume furthermore that the matrix  $\Phi$  in Theorem 28 is nonsingular. Then  $\text{rank}(X_n) = k$  for large  $n$  and the least squares estimator for  $\beta = (\beta_1, \dots, \beta_k)'$  is then given by  $\hat{\beta}_n = (X'_n X_n)^{-1} X'_n y_n$  with  $y_n = (y_1, \dots, y_n)'$ . Since  $\hat{\beta}_n - \beta = (X'_n X_n)^{-1} X'_n u_n$  we have

$$S_n(\hat{\beta}_n - \beta) = S_n(X'_n X_n)^{-1} S'_n S_n^{-1} X'_n u_n = (W'_n W_n)^{-1} W'_n u_n,$$

where  $S_n$  is defined in Theorem 28. Since  $\lim_{n \rightarrow \infty} W'_n W_n = \Phi$  and  $\Phi$  is assumed to be nonsingular, we obtain

$$S_n(\hat{\beta}_n - \beta) \xrightarrow{d} N(0, \sigma^2 \Phi^{-1})$$

as a consequence of Theorem 28. Note that this asymptotic normality result allows for trending regressors.

Now suppose that  $u_t$  and  $X_n = (x_{ti})$  satisfy the assumptions of Theorem 29 and that furthermore  $Q$  is nonsingular. Then we obtain by similar argumentation

$$\sqrt{n}(\hat{\beta}_n - \beta) = (n^{-1} X'_n X_n)^{-1} (n^{-\frac{1}{2}} X'_n u_n) \xrightarrow{d} N(0, \sigma^2 Q^{-1}).$$

We note that Theorem 29 does not hold in general if the regressors are allowed to be triangular arrays, i.e. the elements are allowed to depend on  $n$ . For example, suppose  $k = 1$  and  $X_n = [x_{11,n}, \dots, x_{n1,n}]'$  where

$$x_{t1,n} = \begin{cases} 0 & t < n \\ \sqrt{n} & t = n, \end{cases}$$

then  $n^{-1} X'_n X_n = 1$  and  $n^{-1/2} X'_n u_n = u_n$ . The limiting distribution of this expression is just the distribution of the  $u_i$ s, and hence not necessarily normal, violating the conclusion of Theorem 29.

We now give a CLT where the elements of  $X_n$  are allowed to be triangular arrays, but where we assume additionally that the elements of the  $X_n$  matrices are bounded in absolute value.

**Theorem 30.** Let  $u_t$ ,  $t \geq 1$ , be a sequence of iid random variables with  $E u_t = 0$  and  $E u_t^2 = \sigma^2 < \infty$ . Let  $(x_{ti,n}, 1 \leq t \leq n, n \geq 1)$ ,  $i = 1, \dots, k$ , be triangular arrays of real numbers that are bounded in absolute value, i.e.  $\sup_n \sup_{1 \leq t \leq n, 1 \leq i \leq k} |x_{ti,n}| < \infty$ . Let  $X_n = (x_{ti,n})$  denote corresponding sequences of  $n \times k$  real matrices and let  $\lim_{n \rightarrow \infty} n^{-1} X'_n X_n = Q$  be finite. Furthermore, let  $u_n = [u_1, \dots, u_n]'$ , then  $n^{-1/2} X'_n u_n \xrightarrow{d} N(0, \sigma^2 Q)$ .

Inspection of the proof of Theorem 30 shows that the uniform boundedness condition is stronger than is necessary and that it can be replaced by the condition  $\max_{1 \leq i \leq n} |x_{ti,n}| = o(n^{1/2})$  for  $i = 1, \dots, k$ .

## 4.2 Dependent processes

There is a large literature on CLTs for dependent processes. Due to space limitation we will only present here – analogously as in our discussion of LLNs – two CLTs for dependent processes. Both CLTs are given for martingale difference sequences. As discussed, martingale difference sequences represent an important class of stochastic processes in statistics. The first of the subsequent two theorems assumes that the process is strictly stationary.

**Theorem 31.**<sup>26</sup> Let  $Z_t$  be a strictly stationary and ergodic martingale difference sequence with  $\text{var}(Z_t) = \sigma^2 < \infty$ . Then  $n^{-1/2} \sum_{t=1}^n Z_t \xrightarrow{d} N(0, \sigma^2)$ . (In case  $\sigma^2 = 0$ , the limit  $N(0, 0)$  should be interpreted as the degenerate distribution having all its probability mass concentrated at zero.<sup>27</sup>)

The above theorem contains the Lindeberg–Lévy CLT for iid random variables as a special case. The usefulness of Theorem 31 is illustrated by the following example.

**Example 10.** Suppose  $y_t$  is a stationary autoregressive process of order one satisfying

$$y_t = ay_{t-1} + \varepsilon_t,$$

where  $|a| < 1$  and the  $\varepsilon_t$ s are iid with mean zero and variance  $\sigma^2$ ,  $0 < \sigma^2 < \infty$ . Then  $y_t = \sum_{j=0}^{\infty} a^j \varepsilon_{t-j}$  is strictly stationary and ergodic. The least squares estimator calculated from a sample  $y_0, y_1, \dots, y_n$  is given by  $\hat{a}_n = \sum_{t=1}^n y_t y_{t-1} / \sum_{t=1}^n y_{t-1}^2$  (with the convention that we set  $\hat{a} = 0$  on the event  $\{\sum_{t=0}^n y_{t-1}^2 = 0\}$ ). Thus

$$n^{1/2}(\hat{a}_n - a) = \left( n^{-1/2} \sum_{t=1}^n \varepsilon_t y_{t-1} \right) / \left( n^{-1} \sum_{t=1}^n y_{t-1}^2 \right).$$

The denominator converges a.s. to  $E(y_{t-1}^2) = \sigma^2 / (1 - a^2) > 0$  by the Ergodic Theorem (Theorem 22). Observe that  $Z_t = \varepsilon_t y_{t-1}$  satisfies  $E(Z_t | \varepsilon_{t-1}, \varepsilon_{t-2}, \dots) = y_{t-1} E(\varepsilon_t | \varepsilon_{t-1}, \varepsilon_{t-2}, \dots) = y_{t-1} E(\varepsilon_t) = 0$  since  $y_{t-1}$  is a (linear) function of  $\varepsilon_{t-1}, \varepsilon_{t-2}, \dots$  and since  $\varepsilon_t$  is independent of  $\{\varepsilon_{t-1}, \varepsilon_{t-2}, \dots\}$ . Hence,  $Z_t$  is a martingale difference sequence wrt  $\mathcal{F}_t$ , where  $\mathcal{F}_t$  denotes the  $\sigma$ -field generated by  $\varepsilon_t, \varepsilon_{t-1}, \dots$ . As a function of  $\varepsilon_t$  and  $y_{t-1}$  the sequence  $Z_t$  is clearly strictly stationary and ergodic. Furthermore,  $\text{var}(Z_t) = E(\varepsilon_t^2 y_{t-1}^2) = E(\varepsilon_t^2) E(y_{t-1}^2) = \sigma^4 / (1 - a^2) < \infty$ . Theorem 31 then implies

$$n^{-1/2} \sum_{t=1}^n \varepsilon_t y_{t-1} \xrightarrow{d} N(0, \sigma^4 / (1 - a^2)).$$

Combining this with the already established convergence of the denominator implies the asymptotic normality result

$$n^{1/2}(\hat{a}_n - a) \xrightarrow{d} N(0, 1 - a^2).$$

**Theorem 32.**<sup>28</sup> Let  $Z_t$  be a martingale difference sequence (wrt  $\mathcal{F}_t$ ) with conditional variances  $E(Z_t^2 | \mathcal{F}_{t-1}) = \sigma_t^2$ . Let  $\sigma_{(n)}^2 = \sum_{t=1}^n \sigma_t^2$ . Suppose

$$n^{-1}\sigma_{(n)}^2 \xrightarrow{P} \psi, \quad 0 < \psi < \infty, \quad (10.15)$$

and

$$\sum_{t=1}^n E\left(\left|Z_t/\sqrt{n}\right|^{2+\delta} | \mathcal{F}_{t-1}\right) \xrightarrow{P} 0 \quad (10.16)$$

as  $n \rightarrow \infty$  for some  $\delta > 0$ , then  $n^{-1/2} \sum_{t=1}^n Z_t \xrightarrow{d} N(0, \psi)$ .

Condition (10.16) is a conditional Lyapounov condition. A sufficient condition for (10.16) is

$$\sup_n n^{-1} \sum_{t=1}^n E(|Z_t|^{2+\delta} | \mathcal{F}_{t-1}) = O_p(1). \quad (10.17)$$

As mentioned above, there is an enormous body of literature on CLTs for dependent processes. For further CLTs for martingale difference sequences and related results see Hall and Heyde (1980). Central limit theorems for  $m$ -dependent and linear processes can, for example, be found in Hannan (1970, ch. IV.4), Anderson (1971, ch. 7.7), or Phillips and Solo (1992). Classical references to central limit theorems for mixingales (including  $\alpha$ -mixing,  $\phi$ -mixing, and near epoch dependent processes) are McLeish (1974, 1975). For additional discussions of CLTs see, e.g., Davidson (1994, Part V) and Pötscher and Prucha (1997, ch. 10) and the references given therein.

## 5 FURTHER READINGS

There is a large number of books available that provide further in-depth discussions of the material (or parts of the material) presented in this article. The list of such books includes texts by Billingsley (1968, 1979), Davidson (1994), Serfling (1980) and Shirayev (1984), to mention a few. Hall and Heyde (1980) give a thorough discussion of martingale limit theory.

Recent books on asymptotic theory for least mean distance estimators (including maximum likelihood estimators) and generalized method of moments estimators for general classes of nonlinear models include texts by Bierens (1994), Gallant (1987), Gallant and White (1988), Pötscher and Prucha (1997), and White (1994). For recent survey articles see, e.g., Newey and McFadden (1994), and Wooldridge (1994).

## Notes

1 Strictly speaking,  $h(\cdot)$  has to be Borel measurable; for a definition of Borel measurability see Billingsley (1979), section 13.

- 2 For an understanding of the example it proves helpful to plot  $Z_1(\omega)$ ,  $Z_2(\omega), \dots$  for  $0 \leq \omega < 1$ .
- 3 The space of all random variables with finite  $r$ th moment is often referred to as the space  $L^r$ . On this space convergence in  $r$ th mean is often referred to as  $L^r$ -convergence.
- 4 For a proof see, e.g., Serfling (1980, p. 15).
- 5 See Billingsley (1979, p. 180 and Example 21.21).
- 6 We note that the rescaled quantities like  $\sqrt{n}(\hat{\theta}_n - \theta)$  typically do not converge a.s. or i.p., and thus a new notion of convergence is needed for these quantities.
- 7 See, e.g., Billingsley (1968, p. 12), Billingsley (1979, p. 345), and Serfling (1980, p. 16).
- 8 See, e.g., Serfling (1980, pp. 13–14).
- 9 See, e.g., Serfling (1980, p. 24).
- 10 That is, let  $A \subseteq \mathbb{R}^k$  denote the set of continuity points of  $g$ , then  $P_Z(A) = P(Z \in A) = 1$ . Of course, if  $g$  is continuous on  $\mathbb{R}^k$ , then  $A = \mathbb{R}^k$  and the condition  $P_Z(A) = 1$  is trivially satisfied.
- 11 The event  $\{V_n = 0\}$  has probability approaching zero, and hence it is irrelevant which value is assigned to  $W_n/V_n$  on this event.
- 12 See, e.g., Shirayev (1984, p. 366).
- 13 See, e.g., Shirayev (1984, p. 364).
- 14 See, e.g., Shirayev (1984, p. 487), or Davidson (1994, p. 314).
- 15 See, e.g., Stout (1974, p. 180).
- 16 See, e.g., Stout (1974, p. 181).
- 17 More precisely,  $Q_n$  is a function from  $\Omega \times \Theta$  into  $\mathbb{R}$  such that  $Q(\cdot, \theta)$  is  $\mathcal{F}$ -measurable for each  $\theta \in \Theta$ .
- 18 The same line of argument holds if (uniform) convergence a.s. is replaced by (uniform) convergence i.p.
- 19 For a proof see Jennrich (1969, Theorem 2).
- 20 See, e.g., Billingsley (1979, p. 308).
- 21 Of course, the case  $\sigma^2 = 0$  is trivial since in this case  $Z_t = 0$  a.s.
- 22 See, e.g., Billingsley (1979, p. 310).
- 23 See, e.g., Billingsley (1979, p. 312).
- 24 The theorem is given as Problem 27.6 in Billingsley (1979, p. 319).
- 25 The proof given in Amemiya seems not to be entirely rigorous in that it does not take into account that the elements of  $S_n$  and hence those of  $W_n$  depend on the sample size  $n$ .
- 26 See, e.g., Gänssler and Stute (1977, p. 372).
- 27 See footnote 21.
- 28 See, e.g., Gänssler and Stute (1977, p. 365 and 370).

## References

- Amemiya, T. (1985). *Advanced Econometrics*. Cambridge: Harvard University Press.
- Anderson, T.W. (1971). *The Statistical Analysis of Time Series*. New York: Wiley.
- Andrews, D.W.K. (1987). Consistency in nonlinear econometric models: A generic uniform law of large numbers. *Econometrica* 55, 1465–71.
- Bierens, H.J. (1994). *Topics in Advanced Econometrics*. Cambridge: Cambridge University Press.
- Billingsley, P. (1968). *Convergence of Probability Measures*. New York: Wiley.
- Billingsley, P. (1979). *Probability and Measure*. New York: Wiley.
- Davidson, J. (1994). *Stochastic Limit Theory*. Oxford: Oxford University Press.

- Davidson, J., and R. de Jong (1997). Strong laws of large numbers for dependent heterogeneous processes: A synthesis of recent new results. *Econometric Reviews* 16, 251–79.
- Fuller, W.A. (1976). *Introduction to Statistical Time Series*. New York: Wiley.
- Gallant, A.R. (1987). *Nonlinear Statistical Models*. New York: Wiley.
- Gallant, A.R., and H. White (1988). *A Unified Theory of Estimation and Inference in Nonlinear Dynamic Models*. New York: Basil Blackwell.
- Gänssler, P., and W. Stute (1977). *Wahrscheinlichkeitstheorie*. New York: Springer Verlag.
- Hall, P., and C.C. Heyde (1980). *Martingale Limit Theory and Its Application*. New York: Academic Press.
- Hannan, E.J. (1970). *Multiple Time Series*. New York: Wiley.
- Jennrich, R.I. (1969). Asymptotic properties of nonlinear least squares estimators. *Annals of Mathematical Statistics* 40, 633–43.
- McLeish, D.L. (1974). Dependent central limit theorems and invariance principles. *Annals of Probability* 2, 620–8.
- McLeish, D.L. (1975). Invariance principles for dependent variables. *Zeitschrift für Wahrscheinlichkeitstheorie und Verwandte Gebiete* 32, 165–78.
- Newey, W.K., and D.L. McFadden (1994). Large sample estimation and hypothesis testing. In R.F. Engle, and D.L. McFadden (eds.) *Handbook of Econometrics, Volume 4*. pp. 2113–245, New York: Elsevier Science B.V.
- Phillips, P.C.B., and V. Solo (1992). Asymptotics for linear processes. *Annals of Statistics* 20, 971–1001.
- Pötscher, B.M., and I.R. Prucha (1986). A class of partially adaptive one-step M-estimators for the non-linear regression model with dependent observations. *Journal of Econometrics* 32, 219–51.
- Pötscher, B.M., and I.R. Prucha (1989). A uniform law of large numbers for dependent and heterogeneous data processes. *Econometrica* 57, 675–83.
- Pötscher, B.M., and I.R. Prucha (1997). *Dynamic Nonlinear Econometric Models, Asymptotic Theory*. New York: Springer Verlag.
- Serfling, R.J. (1980). *Approximation Theorems of Mathematical Statistics*. New York: Wiley.
- Shiryayev, A.N. (1984). *Probability*. New York: Springer Verlag.
- Stout, W.F. (1974). *Almost Sure Convergence*. New York: Academic Press.
- Theil, H. (1971). *Principles of Econometrics*. New York: Wiley.
- White, H. (1994). *Estimation, Inference and Specification Analysis*. Cambridge: Cambridge University Press.
- Wooldridge, J.M. (1994). Estimation and inference for dependent processes. In R.F. Engle and D.L. McFadden (eds.) *Handbook of Econometrics, Volume 4*. pp. 2641–738, New York: Elsevier Science B.V.

CHAPTER ELEVEN

# Generalized Method of Moments

*Alastair R. Hall\**

## 1 INTRODUCTION

Generalized method of moments (GMM) was first introduced into the econometrics literature by Lars Hansen in 1982. Since then, GMM has had considerable impact on the theory and practice of econometrics. For theoreticians, the main advantage is that GMM provides a very general framework for considering issues of statistical inference because it encompasses many estimators of interest in econometrics. For applied researchers, it provides a computationally convenient method of estimating nonlinear dynamic models without complete knowledge of the probability distribution of the data. These applications have been in very diverse areas spanning macroeconomics, finance, agricultural economics, environmental economics, and labour economics. Depending on the context, GMM has been applied to time series, cross-sectional, and panel data. In this chapter we provide a survey of the GMM estimation framework and its properties in correctly specified models.<sup>1</sup> Inevitably, GMM builds from earlier work, and its most obvious statistical antecedents are method of moments (Pearson, 1893, 1894, 1895) and instrumental variables estimation (Wright, 1925; Reiersol, 1941; Geary, 1942; Sargan, 1958).

To introduce the basic idea behind the GMM framework, it is useful to consider briefly the structure of method of moments (MM) estimation. Suppose that an economic and/or statistical model implies a vector of observed variables,  $v_t$ , and a  $p \times 1$  vector of unknown parameters,  $\theta_0$ , satisfy a  $p \times 1$  vector of population moment conditions,

$$E[f(v_t, \theta_0)] = 0. \quad (11.1)$$

The MM estimator of  $\theta_0$ , is found by solving the analogous sample moment condition. So if the MM estimator is denoted by  $\hat{\theta}_T$  then it is defined by

$$g_T(\hat{\theta}_T) = T^{-1} \sum_{t=1}^T f(v_t, \hat{\theta}_T) = 0, \quad (11.2)$$

where  $T$  is the sample size. Notice that (11.2) represents a set of  $p$  equations in  $p$  unknowns and so has a unique solution under certain conditions. This approach has a natural appeal, and intuition suggests – correctly – that the solution to (11.2),  $\hat{\theta}_T$ , converges in probability to the solution to (11.1),  $\theta_0$ , subject to appropriate regularity conditions. Now suppose that  $f(\cdot)$  is a  $q \times 1$  vector and that  $q > p$ . In this case (11.2) represents a set of  $q$  equations in  $p < q$  unknowns. Such a system typically does not possess a solution and so MM estimation is rendered infeasible. *Generalized* method of moments circumvents this problem by choosing the value of  $\theta$  which is closest to satisfying (11.2) as the estimator for  $\theta_0$ . To make the approach operational, it is necessary to define a measure of how far  $g_T(\theta)$  is from zero. In GMM, the measure of distance is

$$Q_T(\theta) = g_T(\theta)' W_T g_T(\theta), \quad (11.3)$$

where  $W_T$  is a  $q \times q$  weighting matrix which must satisfy certain conditions that need not concern us for the moment. So the GMM estimator is defined to be

$$\hat{\theta}_T = \operatorname{argmin}_{\theta \in \Theta} Q_T(\theta), \quad (11.4)$$

where  $\Theta$  denotes the parameter space.

If it were always the case that  $q = p$  in econometric applications then there would be no need for a separate GMM theory because GMM would reduce to MM. However,  $q$  is greater than  $p$  in many situations of interest, and it is this possibility which leads to the unique features of the GMM framework. In this chapter we concentrate on issues pertaining to estimation. This means we will largely ignore the considerable literature on hypothesis testing based on GMM estimators. However, the interested reader can find discussion of various aspects of hypothesis testing elsewhere in this volume.<sup>2</sup>

Throughout this chapter, the analysis abstracts to the general form of population moment condition given in (11.1). However, before we begin, it is useful to present two examples which help to illustrate both how population moment conditions arise and also the forms they can take. The first example is taken from a study in the education literature based on cross-sectional data. The second example is a study from the empirical finance and macroeconomic literatures based on time series data.

1. *Education example:* Angrist and Krueger (1992) investigate the impact of age at school entry on educational attainment using the model,

$$y_{i,n} = \alpha + \beta a_{i,n} + \varepsilon_{i,n}$$

where  $y_{i,n}$  is the average number of years of education completed by students born in quarter  $i$  of year  $n$ , and  $a_{i,n}$  is the average age of school entry for

members of that cohort. Within this model, the marginal response of attainment to age of entry is captured by  $\beta$ , and so this represents the parameter of interest. Estimation of this parameter is complicated by a correlation between the explanatory variable and the error which arises because many children who start school at a younger age do so because they show above average learning potential. This correlation means the ordinary least squares estimator is inconsistent. However, the error is anticipated to be uncorrelated with the quarter of birth. This logic leads to the population moment condition  $E[z_{i,n}(y_{i,n} - \alpha - \beta a_{i,n})] = 0$  where  $z'_{i,n} = [Q_{1,i}, Q_{2,i}, Q_{3,i}, Q_{4,i}]$  and  $Q_{j,i} = 1$  if  $i = j$  and 0 otherwise.

2. *Empirical finance example:* Hansen and Singleton (1982) estimate a model which seeks to explain the relationship between asset prices and their returns via the decisions of a representative consumer.<sup>3</sup> Within this framework, a representative consumer makes consumption and investment decisions to maximize his or her expected discounted lifetime utility. If it is assumed that the agent possesses a constant relative risk aversion utility function and invests at time  $t$  in an asset which matures at time  $t + 1$  then the asset return satisfies the equation

$$E[\delta(r_{t+1}/p_t)(c_{t+1}/c_t)^{\gamma-1} - 1 | \Omega_t] = 0 \quad (11.5)$$

where  $r_{t+1}$  is the return on the asset in period  $t + 1$ ,  $p_t$  is the price of the asset in period  $t$ ,  $c_t$  is consumption in period  $t$ ,  $\Omega_t$  is the information set available to the agent in period  $t$ ,  $\gamma$  is the agent's coefficient of relative risk aversion and  $\delta$  is his or her discount factor. To use this model for asset pricing, it is necessary to estimate  $\gamma$  and  $\delta$ . Unfortunately, the joint distribution of consumption growth and asset returns is unknown, and this makes maximum likelihood infeasible. However, (11.5) and an iterated expectations argument imply the population moment condition,

$$E[z_t(\delta(r_{t+1}/p_t)(c_{t+1}/c_t)^{\gamma-1} - 1)] = 0$$

where  $z_t$  is a vector of variables contained in  $\Omega_t$ .

An overview of the chapter is as follows. To begin, we return to the basic definition of the GMM estimation principle, and consider formally certain issues which were swept aside in the heuristic discussion above. Most notably, the discussion was predicated on the assumption that the population moment condition provides sufficient information to uniquely determine  $\theta_0$ . This need not be the case, and Section 2 introduces the concepts of global and local identification of the parameter vector. One important ramification of  $q > p$  is that the estimation effects a decomposition on the population moment condition into so-called identifying and overidentifying restrictions. Section 3 describes this decomposition and shows how these components are linked to the parameter estimator and estimated sample moment,  $g_T(\hat{\theta}_T)$ . Section 4 considers the asymptotic properties of the estimator and the estimated sample moment. For the estimator, this discussion

focuses on the consistency and asymptotic distribution of the estimator. The latter can be used to construct large sample confidence intervals for the elements of  $\theta_0$ . In practice, these intervals depend on the long-run variance of the sample moment, and so we briefly consider how this variance can be estimated. For the estimated sample moment, the discussion concentrates on its asymptotic distribution. Up to this point the analysis only restricts the weighting matrix to be a member of a certain class. However, it will emerge that the choice of  $W_T$  impacts on the estimator via its asymptotic variance. Section 5 characterizes the optimal choice of  $W_T$  and discusses certain issues involved in the calculation of the associated “optimal” GMM estimator. Although we restrict attention to correctly specified models, a researcher can never be sure in practice that this is the case. Section 6 describes how the estimated sample moment can be used to construct the “overidentifying restrictions test” for the adequacy of the model specification. Throughout the first six sections, the population moment condition is taken as given. The next two sections explore issues related to the choice of  $f(\cdot)$ . Section 7 shows how various other econometric estimators can be considered special cases of GMM. Section 8 considers two extremes: the optimal choice of  $f(\cdot)$  and what happens if the population moment condition provides no – or virtually no – information about  $\theta_0$ . Finally Section 9 provides a brief review of the available evidence on the finite sample behavior of GMM.

Due to space constraints, we present only heuristic arguments for the main results and provide references to appropriate sources for more formal analyses. A rigorous treatment of the material in the chapter – and many other aspects of the GMM framework – can also be found in Hall (2000b).

## 2 THE POPULATION MOMENT CONDITION AND IDENTIFICATION

In this section, we consider the conditions under which the population moment provides sufficient information to determine uniquely  $\theta_0$  from all other elements in the parameter space  $\Theta \subseteq \Re^p$ . If this is the case then  $\theta_0$  is said to be *identified*. To begin, it is necessary to define formally the information contained in the population moment condition. From the heuristic discussion above, it is clear that the population moment condition is a statement about a  $q \times 1$  vector of functions,  $f(\cdot, \cdot)$ , of the observable vector of random variables  $v_t$  and the unknown ( $p \times 1$ ) parameter vector,  $\theta_0$ . Certain restrictions must be placed on these constituents as we proceed with the analysis, and we shall impose the most important as they become necessary. However, space constraints forbid a complete accounting of all the required conditions; these can be found in Hansen (1982) or Wooldridge (1994). Throughout the chapter we follow Hansen’s (1982) original framework and consider the following case.

**Assumption 1. Strict Stationarity.** The  $(r \times 1)$  random vectors  $\{v_t; -\infty < t < \infty\}$  form a strictly stationary process with sample space  $V \subseteq \Re^r$ .

Recall that this assumption implies all expectations of functions of  $v_t$  are independent of time.<sup>4</sup>

**Assumption 2. Regularity Conditions for  $f(\cdot, \cdot)$ .** The function  $f: V \times \Theta \rightarrow \mathbb{R}^q$  satisfies: (i) it is continuous on  $\Theta$  for each  $v \in V$ ; (ii)  $E[f(v_t, \theta)]$  exists and is finite for every  $\theta \in \Theta$ ; (iii)  $E[f(v_t, \theta)]$  is continuous on  $\Theta$ .

With these two assumptions in place, we can now formally restate the population moment condition.

**Assumption 3. Population Moment Condition.** There exists  $\theta_0 \in \Theta$  such that the following ( $q \times 1$ ) population moment condition holds:  $E[f(v_t, \theta_0)] = 0$ .

Throughout this chapter, we focus on the properties of the GMM estimator of  $\theta_0$  based on this population moment condition. However, by itself, Assumption 3 does not provide sufficient information to identify  $\theta_0$ . This is only the case if there are no other values of  $\theta$  at which the population moment condition is satisfied. This condition for parameter identification can be stated as follows.

**Assumption 4. Global Identification.** The parameter vector  $\theta_0$  is globally identified by the population moment condition in Assumption 3 if and only if  $E[f(v_t, \bar{\theta})] \neq 0$  for all  $\bar{\theta} \in \Theta$  such that  $\bar{\theta} \neq \theta_0$ .

The adjective “global” emphasizes that the population moment condition only holds at one value in the *entire* parameter space. While this condition is easily stated, it is often hard to verify a priori in nonlinear models. Identification can fail due to the properties of the data,  $v_t$ , or due to the properties of  $f(\cdot)$  as a function of  $\theta$  or due to an interaction of the two. Fortunately, a more useful condition can be found if attention is limited to some suitably defined neighborhood of  $\theta_0$ . The price of this approach is that we are now deriving conditions for identification only within this neighborhood and so these are referred to as conditions for *local* identification. As the names suggest, local identification is necessary but not sufficient for global identification. Therefore, a more transparent condition for local identification is useful because it provides insights into the circumstances in which identification can fail.

The condition for local identification is based on a first order Taylor series expansion, and so it is necessary to introduce the following definition and assumption. An  $\varepsilon$ -neighborhood of  $\theta_0$  is defined to be the set  $N_\varepsilon$  which satisfies  $N_\varepsilon = \{\theta; \|\theta - \theta_0\| < \varepsilon\}$  where  $\|a\| = (a'a)^{1/2}$ .

**Assumption 5. Regularity Conditions on  $\partial f(v_t, \theta)/\partial\theta'$ .** (i) The derivative matrix  $\partial f(v, \theta)/\partial\theta'$  exists and is continuous on  $\Theta$  for each  $v \in V$ ; (ii)  $\theta_0$  is an interior point of  $\Theta$ ; (iii)  $E[\partial f(v_t, \theta_0)/\partial\theta']$  exists and is finite.

To derive the condition for local identification, we restrict attention to a sufficiently small  $\varepsilon$  such that  $f(\cdot)$  can be approximated by the following first order Taylor series expansion<sup>5</sup> in  $N_\varepsilon$

$$f(v_t, \theta) \approx f(v_t, \theta_0) + \{\partial f(v_t, \theta_0)/\partial\theta'\} (\theta - \theta_0), \quad (11.6)$$

where  $\partial f(v_t, \theta_0)/\partial\theta'$  is the  $q \times p$  matrix with  $i - j$ th element  $\partial f_i(v_t, \theta_0)/\partial\theta'_j$ . Taking expectations on both sides of (11.6) and using Assumptions 3 and 5 yields

$$E[f(v_t, \theta)] \approx \{E[\partial f(v_t, \theta_0)/\partial\theta']\} (\theta - \theta_0). \quad (11.7)$$

Equation (11.7) states that in the neighborhood of  $\theta_0$ ,  $E[f(v_t, \theta)]$  is essentially a linear function of  $\theta - \theta_0$ . This leads to the following condition for *local* identification.

**Lemma 1. Local Identification.** The parameter vector  $\theta_0$  is locally identified by the population moment condition in Assumption 3 if and only if  $E[\partial f(v_t, \theta_0)/\partial\theta']$  is of rank  $p$ .

While this condition needs to be verified on a case by case basis, it does provide some general insights into identification in nonlinear models. First, the rank condition immediately implies identification fails if there are fewer moment conditions than parameters, i.e.  $q < p$ . Second, notice that the dependence of the partial derivative matrix on  $\theta$  implies the population moment condition may provide enough information to identify the parameters at some values of  $\theta_0$  but not at others.

From this discussion, it is clear that the relationship between  $q$  and  $p$  is important. This has led to the introduction of the following terminology. If  $p > q$  and hence the local identification condition is not satisfied then  $\theta_0$  is said to be *un*-identified. If  $p = q$  and Assumption 4 is satisfied then  $\theta_0$  is said to be *just*-identified. Finally, if  $q > p$  and Assumption 4 is satisfied then  $\theta_0$  is said to be *over*-identified.

### 3 THE ESTIMATOR AND A FUNDAMENTAL DECOMPOSITION

The introduction provides the essence of GMM. In this section, we discuss the form of the estimator in more detail, and also describe how estimation effects a fundamental decomposition of the population moment condition.

Recall that the GMM estimator of  $\theta_0$  based on the population moment condition in Assumption 3 is:

$$\hat{\theta}_T = \operatorname{argmin}_{\theta \in \Theta} Q_T(\theta), \quad (11.8)$$

where  $Q_T(\theta) = g_T(\theta)'W_Tg_T(\theta)$ . For this approach to make sense,  $Q_T(\theta)$  must be a meaningful measure of distance and hence the weighting matrix must possess certain properties. Specifically,  $W_T$  must be positive semi-definite for finite  $T$  so that  $Q_T(\theta)$  is both nonnegative and equals zero if  $g_T(\theta) = 0$ . However, *semi*-definiteness leaves open the possibility that  $Q_T(\theta) = 0$  without  $g_T(\theta) = 0$ . Since all our statistical analysis is based on asymptotic theory, it is only necessary to rule out this possibility in the limit. Accordingly,  $W_T$  is assumed to satisfy:

**Assumption 6. Properties of the Weighting Matrix.**  $W_T$  is a positive semi-definite matrix which converges in probability to the positive definite matrix of constants  $W$ .

If Assumption 5 holds, and in most cases of interest it will, then the first order conditions for this minimization imply  $\partial Q_T(\hat{\theta}_T)/\partial \theta = 0$ . This condition yields<sup>6</sup>

$$\left\{ T^{-1} \sum_{t=1}^T \frac{\partial f(v_t, \hat{\theta}_T)}{\partial \theta'} \right\}' W_T \left\{ T^{-1} \sum_{t=1}^T f(v_t, \hat{\theta}_T) \right\} = 0 \quad (11.9)$$

However, there is typically no closed form solution for  $\hat{\theta}_T$  and so the estimator must be obtained via numerical optimization techniques.<sup>7</sup>

This characterization of the estimator yields an interesting interpretation of GMM. Inspection of (11.9) reveals that the GMM estimator based on  $E[f(v_t, \theta_0)] = 0$  can be interpreted as a method of moments estimator based on

$$\{E[\partial f(v_t, \theta_0)/\partial \theta']\}' W E[f(v_t, \theta_0)] = 0. \quad (11.10)$$

Therefore, GMM is an MM estimator based on the information that the  $p$  linear combinations of  $E[f(v_t, \theta_0)]$  in (11.10) are zero. Notice that if the assumptions behind Lemma 1 hold then these  $p$  linear combinations are linearly independent and so the MM interpretation emphasizes the fundamental connection between identification and estimation – that is, the  $p$  parameters are only *locally* identified if the estimation is based on  $p$  linearly independent equations. If  $p = q$  then (11.10) is equivalent to  $E[f(v_t, \theta_0)] = 0$ , and we note parenthetically that this means the weighting matrix plays no role in the analysis. However, if  $q > p$  then there is a difference between information used in estimation and the original population moment condition.

The advantage of this MM interpretation is that it makes explicit the information used in GMM information. However, it is not particularly amenable to the characterization of what information is left out because (11.10) is a system of  $p < q$  equations. Sowell (1996) shows that this problem can be circumvented if (11.10) is interpreted in terms of the transformed moment condition  $W^{1/2}E[f(v_t, \theta_0)]$ . Equation (11.10) can then be rewritten as

$$F(\theta_0)' W^{1/2} E[f(v_t, \theta_0)] = 0, \quad (11.11)$$

where  $F(\theta_0) = W^{1/2}E[\partial f(v_t, \theta_0)/\partial \theta']$ . Equation (11.11) states that  $W^{1/2}E[f(v_t, \theta_0)]$  lies in the null space of  $F(\theta_0)'$ . Sowell (1996) shows that the information about  $\theta_0$  in (11.10) is equivalent to the information in

$$F(\theta_0)[F(\theta_0)' F(\theta_0)]^{-1} F(\theta_0)' W^{1/2} E[f(v_t, \theta_0)] = 0. \quad (11.12)$$

Formally, (11.12) states that the least squares projection of  $W^{1/2}E[f(v_t, \theta_0)]$  onto the column space of  $F(\theta_0)$  is zero, and thereby places

$$\text{rank}\{F(\theta_0)[F(\theta_0)' F(\theta_0)]^{-1} F(\theta_0)'\} = p$$

restrictions on the transformed population moment condition. Sowell (1996) refers to (11.12) as the *identifying restrictions*.

The advantage of this alternative characterization is that (11.12) is a system of  $q$  equations and so it is now possible to characterize the part of  $E[f(v_t, \theta_0)] = 0$  which is ignored in estimation. By definition, this remainder is

$$\{I_q - F(\theta_0)[F(\theta_0)'F(\theta_0)]^{-1}F(\theta_0)'\}W^{1/2}E[f(v_t, \theta_0)] = 0, \quad (11.13)$$

which represents what Hansen referred to as the *overidentifying restrictions* in his original article. Equation (11.13) states that the projection of  $W^{1/2}E[f(v_t, \theta_0)]$  on to the orthogonal complement of  $F(\theta_0)$  is zero, and thereby places  $q - p$  restrictions on the transformed population moment condition.

Equations (11.12)–(11.13) indicate that GMM estimation effects a fundamental decomposition of the  $q \times 1$  population moment condition into  $p$  identifying restrictions upon which estimation is based, and  $q - p$  overidentifying restrictions which are ignored in estimation. Notice also that these two sets of restrictions are orthogonal due to the projection matrix structure.

The roles of the two sets of restrictions are reflected in their sample counterparts. Since the identifying restrictions represent the information used in estimation, their sample analogs are satisfied at  $\hat{\theta}_T$  by construction. In contrast, the overidentifying restrictions are ignored in estimation and so their sample analog is not satisfied. However, they can be used to give a useful interpretation to the GMM minimand. From (11.9), it follows that

$$\begin{aligned} W_T^{1/2}T^{-1}\sum_{t=1}^T f(v_t, \hat{\theta}_T) &= \{I_q - F_T(\hat{\theta}_T)[F_T(\hat{\theta}_T)'F_T(\hat{\theta}_T)]^{-1}F_T(\hat{\theta}_T)'\} \\ &\quad \times W_T^{1/2}T^{-1}\sum_{t=1}^T f(v_t, \hat{\theta}_T) \end{aligned} \quad (11.14)$$

where  $F_T(\theta) = T^{-1}\sum_{t=1}^T \partial f(v_t, \theta)/\partial\theta'$ . Therefore,  $Q_T(\hat{\theta}_T)$  can be interpreted as a measure of how far the sample is from satisfying the overidentifying restrictions.

#### 4 ASYMPTOTIC PROPERTIES

At the beginning, we presented an intuitive justification for the GMM estimation framework. In this section, we provide a more rigorous argument by establishing the consistency and asymptotic normality of the estimator. The latter facilitates the construction of large sample confidence intervals for  $\theta_0$ . These intervals depend on the long-run variance of the sample moment, and so we briefly discuss how this variance can be consistently estimated. We also derive the asymptotic distribution of the estimated sample moment. The latter analysis provides further evidence of the connection between  $g_T(\hat{\theta}_T)$  and the overidentifying restrictions, and plays an important role in the model specification test discussed in Section 6.

The analysis rests on applications of the laws of large numbers (LLN) and the central limit theorem (CLT) to functions of  $v_t$ . So far, we have only restricted  $v_t$  to be stationary, and this is insufficient by itself to guarantee these limit theorems. Accordingly, we impose the following condition.

**Assumption 7. Ergodicity.** The random process  $\{v_t; -\infty < t < \infty\}$  is ergodic.

A formal definition of ergodicity involves rather sophisticated mathematical ideas and is beyond the scope of this chapter. Instead we refer the interested reader to Davidson (1994, pp. 199–203). It is sufficient for ergodicity that the dependence between  $v_t$  and  $v_{t-m}$  decreases at a certain rate to zero as  $m \rightarrow \infty$ . If  $v_t$  satisfies this type of restriction then it is called a *mixing* process. Certain other regularity conditions must also be imposed but due to space constraints we shall not explore them here. Instead we refer the interested reader to Hansen's (1982) original article or the surveys by Newey and McFadden (1994) and Wooldridge (1994).

Recall that  $\hat{\theta}_T$  is consistent for  $\theta_0$  if  $\hat{\theta}_T \xrightarrow{P} \theta_0$ . If there is a closed form solution for  $\hat{\theta}_T$  then it is relatively straightforward to establish whether or not the estimator is consistent by examining the limiting behavior of appropriate functions of the data. Unfortunately, as remarked above, we do not have this luxury in most non-linear models. However, although there is no closed form,  $\hat{\theta}_T$  is clearly defined by (11.8). The key to a proof of consistency is the consideration of what happens if we perform a similar minimization on the population analog to  $Q_T(\theta)$ ,

$$Q_0(\theta) = \{E[f(v_t, \theta)]\}' W \{E[f(v_t, \theta)]\} \quad (11.15)$$

The answer follows directly from our earlier assumptions. The population moment condition implies  $Q_0(\theta_0) = 0$ . The global identification condition and the positive definiteness of  $W$ , imply  $Q_0(\theta) > 0$  for all  $\theta \neq \theta_0$ . Taken together these two properties imply  $Q_0(\theta)$  has a unique minimum at  $\theta = \theta_0$ . Intuition suggests that if  $\hat{\theta}_T$  minimizes  $Q_T(\theta)$  and  $Q_T(\theta)$  converges in probability to a function,  $Q_0(\theta)$ , with a unique minimum at  $\theta_0$ , then  $\hat{\theta}_T$  converges in probability to  $\theta_0$ . In essence this intuition is correct but there is one mathematical detail which needs to be taken into account. It is not necessarily the case that the minimum of a sequence of functions converges to the minimum of the limit of the sequence of functions. For this to be the case, it is sufficient that  $Q_T(\theta)$  converges *uniformly* to  $Q_0(\theta)$ .<sup>8</sup>

**Assumption 8. Uniform Convergence in Probability of  $Q_T(\theta)$ .**  
 $\sup_{\theta \in \Theta} |Q_T(\theta) - Q_0(\theta)| \xrightarrow{P} 0$ .

Once uniform convergence is imposed, then consistency can be established along the lines described above; *e.g.* see Hansen (1982) or Wooldridge (1994).

**Theorem 1. Consistency of the Parameter Estimator.** If Assumptions 1–4, 6–8 and certain other regularity conditions hold then  $\hat{\theta}_T \xrightarrow{P} \theta_0$ .

To develop the asymptotic distribution of the estimator, we require an asymptotically valid closed form representation for  $T^{1/2}(\hat{\theta}_T - \theta_0)$ . This representation comes from an application of the Mean Value Theorem<sup>9</sup> which relates  $f(\cdot)$  to its first derivatives  $\partial f(v_t, \theta)/\partial \theta'$ . So, to pursue this approach, it is necessary to impose Assumption 5. The Mean Value Theorem implies that

$$g_T(\hat{\theta}_T) = g_T(\theta_0) + G_T(\hat{\theta}_T, \theta_0, \lambda_T)(\hat{\theta}_T - \theta_0), \quad (11.16)$$

where  $G_T(\hat{\theta}_T, \theta_0, \lambda)$  is the  $(q \times p)$  matrix whose  $i$ th row is the corresponding row of  $G_T(\bar{\theta}_T^{(i)})$  where  $G_T(\theta) = T^{-1} \sum_{t=1}^T \partial f(v_t, \theta) / \partial \theta'$ ,  $\bar{\theta}_T^{(i)} = \lambda_{i,T} \theta_0 + (1 - \lambda_{i,T}) \hat{\theta}_T$  for some  $0 \leq \lambda_{i,T} \leq 1$ , and  $\lambda_T$  is the  $(q \times 1)$  vector with  $i$ th element  $\lambda_{i,T}$ . Premultiplication of (11.16) by  $G_T(\hat{\theta}_T)' W_T$  yields

$$G_T(\hat{\theta}_T)' W_T g_T(\hat{\theta}_T) = G_T(\hat{\theta}_T)' W_T g_T(\theta_0) + G_T(\hat{\theta}_T)' W_T G_T(\hat{\theta}_T, \theta_0, \lambda_T)(\hat{\theta}_T - \theta_0). \quad (11.17)$$

Now the first order conditions in (11.9) imply the left-hand side of (11.17) is zero and so with some rearrangement it follows from (11.17) that

$$\begin{aligned} T^{1/2}(\hat{\theta}_T - \theta_0) &= -[G_T(\hat{\theta}_T)' W_T G_T(\hat{\theta}_T, \theta_0, \lambda_T)]^{-1} G_T(\hat{\theta}_T)' W_T T^{1/2} g_T(\theta_0) \\ &= \tilde{M}_T T^{1/2} g_T(\theta_0) \text{ say.} \end{aligned} \quad (11.18)$$

Equation (11.18) implies that  $T^{1/2}(\hat{\theta}_T - \theta_0)$  behaves like the product of a random matrix,  $\tilde{M}_T$ , and a random vector,  $T^{1/2} g_T(\theta_0)$ . Therefore, we can derive the asymptotic distribution of the estimator from the limiting behavior of these two components. The asymptotic behavior of  $T^{1/2} g_T(\theta_0)$  is given by a version of the CLT.

**Assumption 9. Central Limit Theorem for  $T^{1/2} g_T(\theta_0)$ .**  $T^{1/2} g_T(\theta_0) \xrightarrow{d} N(0, S)$  where  $S$  is a positive definite matrix of constants.

Now consider  $\tilde{M}_T$ . Since  $\hat{\theta}_T \xrightarrow{p} \theta_0$  and  $\bar{\theta}_T^{(i)}$  lies on the line segment between  $\hat{\theta}_T$  and  $\theta_0$ , then it follows that  $\bar{\theta}_T^{(i)} \xrightarrow{p} \theta_0$  for  $i = 1, 2 \dots p$ . Intuition suggests that this should imply both  $G_T(\hat{\theta}_T)$  and  $G_T(\hat{\theta}_T, \theta_0, \lambda_T)$  converge in probability to  $G_0 = E[\partial f(v_t, \theta_0) / \partial \theta']$ . In essence this is correct, but the argument can only be formally justified if  $G_T(\theta)$  converges uniformly and certain other regularity conditions apply. For brevity, we adopt the high level assumption that the desired behavior occurs, and refer the interested reader to Newey and McFadden (1994) for the necessary underlying regularity conditions.

**Assumption 10. Convergence of  $G_T(\hat{\theta}_T)$  and  $G_T(\hat{\theta}_T, \theta_0, \lambda_T)$ .**  $G_T(\hat{\theta}_T) \xrightarrow{p} G_0$  and  $G_T(\hat{\theta}_T, \theta_0, \lambda_T) \xrightarrow{p} G_0$ .

Assumptions 6 and 10 can be combined with Slutsky's Theorem to deduce that  $\tilde{M}_T \xrightarrow{p} (G_0' W G_0)^{-1} G_0' W$ . Therefore,  $T^{1/2}(\hat{\theta}_T - \theta_0)$  is the product of a random matrix which converges in probability to a constant, and a random vector which converges to a normal distribution. This structure implies:<sup>10</sup>

**Theorem 2. Asymptotic distribution of the estimator.** If Assumptions 1–10 and certain other regularity conditions hold then:  $T^{1/2}(\hat{\theta}_T - \theta_0) \xrightarrow{d} N(0, MSM')$  where  $M = (G_0' W G_0)^{-1} G_0' W$ .

Theorem 2 implies that an approximate  $100(1 - \alpha)\%$  confidence interval for  $\theta_{0,i}$  in large samples is given by

$$\hat{\theta}_{T,i} \pm z_{\alpha/2} \sqrt{\hat{V}_{T,ii}/T}, \quad (11.19)$$

where  $\hat{V}_{T,ii}$  is the  $i - i$ th element of a consistent estimator of  $MSM'$ . In practice, the asymptotic variance can be consistently estimated by  $\hat{V}_T = \hat{M}_T \hat{S}_T \hat{M}'_T$  where  $\hat{M}_T = [G_T(\hat{\theta}_T)' W_T G_T(\hat{\theta}_T)]^{-1} G_T(\hat{\theta}_T)' W_T$ , and  $\hat{S}_T$  is a consistent estimator of  $S$ . The construction of  $\hat{S}_T$  depends on the time series properties of  $f(v_t, \theta_0)$ . With certain relatively mild additional conditions, it can be shown that<sup>11</sup>

$$S = \Gamma_0 + \sum_{i=1}^{\infty} (\Gamma_i + \Gamma'_i), \quad (11.20)$$

where  $\Gamma_j = E[(f_t - E[f_t])(f_{t-j} - E[f_{t-j}])']$  is known as the  $j$ th autocovariance matrix<sup>12</sup> of  $f_t = f(v_t, \theta_0)$ . For brevity, we distinguish only two cases of interest. First, if  $f_t$  is a martingale difference (MD) sequence and hence serially uncorrelated (that is,  $\Gamma_i = 0, i \neq 0$ ) then  $S$  can be consistently estimated by<sup>13</sup>

$$\hat{S}_{MD} = T^{-1} \sum_{t=1}^T \hat{f}_t \hat{f}'_t, \quad (11.21)$$

where  $\hat{f}_t = f(v_t, \hat{\theta}_T)$ . It can be shown that  $\hat{S}_{MD} \xrightarrow{P} S$  if the martingale difference assumption is valid; for example see White (1994, Theorem 8.27, p. 193). Second, and more generally,  $S$  can be estimated by a member of the class of *heteroskedasticity autocorrelation consistent covariance* (HACC) estimators,

$$\hat{S}_{HACC} = \hat{\Gamma}_0 + \sum_{i=1}^{b(T)} \omega_{iT} (\hat{\Gamma}_i + \hat{\Gamma}'_i), \quad (11.22)$$

where  $\hat{\Gamma}_i = T^{-1} \sum_{t=i+1}^T \hat{f}_t \hat{f}'_{t-i}$ ,  $\{\omega_{iT}\}$  are known as weights and  $b(T)$  is the bandwidth. The weights and bandwidth must satisfy certain conditions if  $\hat{S}_{HACC}$  is to be both positive semi-definite<sup>14</sup> and consistent. Various combinations have been proposed in the literature.<sup>15</sup> One example is the “Bartlett” kernel proposed in this context by Newey and West (1987) for which  $\omega_{iT} = 1 - i/[b(T) + 1]$ . Andrews (1991) shows that this choice yields a consistent estimator if  $b(T) \rightarrow \infty$  and  $b(T) = o(T^{1/2})$ . In practice, the researcher must choose both the bandwidth and the weights. While this choice can be guided by asymptotic theory, there is no consensus to date upon what choice is best in the sample sizes encountered in economics and finance.<sup>16</sup> It should be noted that the consistency of both  $\hat{S}_{MD}$  and  $\hat{S}_{HACC}$  is predicated on  $E[f(v_t, \theta_0)] = 0$ . If the model is misspecified, and hence Assumption 3 is violated, then neither estimator is consistent. This inconsistency can have important consequences for the model specification test described in Section 6, and we return to this issue there.

Finally, we consider the asymptotic distribution of the estimated sample moment. It is most convenient to work with the transformed moment,  $W_T^{1/2}T^{1/2}g_T(\hat{\theta}_T)$ . Equation (11.16) implies

$$W_T^{1/2}T^{1/2}g_T(\hat{\theta}_T) = W_T^{1/2}T^{1/2}g_T(\theta_0) + W_T^{1/2}G_T(\hat{\theta}_T, \theta_0, \lambda_T)T^{1/2}(\hat{\theta}_T - \theta_0). \quad (11.23)$$

If we substitute for  $T^{1/2}(\hat{\theta}_T - \theta_0)$  from (11.18) then (11.23) can be written as

$$W_T^{1/2}T^{1/2}g_T(\hat{\theta}_T) = N_T(\hat{\theta}_T)W_T^{1/2}T^{1/2}g_T(\theta_0), \quad (11.24)$$

where

$$N_T(\hat{\theta}_T) = I_q - W_T^{1/2}G_T(\hat{\theta}_T, \theta_0, \lambda_T)[G_T(\hat{\theta}_T)'W_TG_T(\hat{\theta}_T, \theta_0, \lambda_T)]^{-1}G_T(\hat{\theta}_T)'W_T^{1/2}'.$$

Equation (11.24) implies  $W_T^{1/2}T^{1/2}g_T(\hat{\theta}_T)$  has the same generic structure as the expression for  $T^{1/2}(\hat{\theta}_T - \theta_0)$  in (11.18) namely: a random matrix, which converges to a matrix of constants, times a random vector which converges to a normal distribution. Therefore, we can use the same logic as before to deduce the following result; see Hansen (1982).

**Theorem 3. Asymptotic distribution of the estimated sample moment.** If Assumptions 1–10 and certain other regularity conditions hold then:  $W_T^{1/2}T^{1/2}g_T(\hat{\theta}_T) \xrightarrow{d} N(0, NSN')$  where  $N = [I_q - P(\theta_0)]W^{1/2}$  and  $P(\theta_0) = F(\theta_0)'F(\theta_0)^{-1}F(\theta_0)'$ .

The connection between the estimated sample moment and the overidentifying restrictions manifests itself in the asymptotic distribution. Equation (11.24) implies that

$$W_T^{1/2}T^{1/2}g_T(\hat{\theta}_T) = [I_q - P(\theta_0)]W^{1/2}T^{1/2}g_T(\theta_0) + o_p(1). \quad (11.25)$$

Inspection of (11.25) reveals that the asymptotic behavior of the estimated sample moment is governed by the function of the data which appears in the overidentifying restrictions. Therefore, the mean of the asymptotic distribution in Theorem 3 is zero because the overidentifying restrictions are satisfied at  $\theta_0$ . This relationship also has an impact on the properties of the variance of the limiting distribution. Since  $W^{1/2}$  and  $S$  are nonsingular, it follows that<sup>17</sup>  $\text{rank}\{NSN'\} = \text{rank}\{I_q - P(\theta_0)\} = q - p$ , and so the covariance matrix is singular.<sup>18</sup> This rank is easily recognized to be the number of overidentifying restrictions.

## 5 THE OPTIMAL TWO-STEP OR ITERATED GMM ESTIMATOR

It is remarked in Section 3 that if  $q = p$  then GMM is equivalent to the MM estimator based on  $E[f(v_t, \theta_0)] = 0$  and so the estimator does not depend on the weighting matrix. However if  $q > p$  then it is clear from Theorem 2 that the asymptotic variance of  $\hat{\theta}_T$  depends on  $W_T$  via  $W$ .<sup>19</sup> This opens up the possibility

that inferences may be sensitive to  $W$ . It is desirable to base inference on the most precise estimator and so the optimal choice of  $W$  is the one which yields the minimum variance in a matrix sense. This choice is given in the following theorem which was first proved by Hansen (1982).

**Theorem 4. Optimal weighting matrix.** If Assumptions 1–10 and certain other regularity conditions hold then the minimum asymptotic variance of  $\hat{\theta}_T$  is  $(G_0' S^{-1} G_0)^{-1}$  and this can be obtained by setting  $W = S^{-1}$ .

Theorem 4 implies the optimal choice of  $W_T$  is  $\hat{S}_T^{-1}$  where  $\hat{S}_T$  is a consistent estimator of  $S$ . This appears to create a circularity because inspection of (11.21)–(11.22) reveals that  $\hat{S}_T$  depends on  $\hat{\theta}_T$  in general. However, this problem is easily resolved by using a two-step estimation. On the first step a sub-optimal choice of  $W_T$  is used to obtain a preliminary estimator,  $\hat{\theta}_T(1)$ . This estimator is used to obtain a consistent estimator of  $S$ , which is denoted  $\hat{S}_T(1)$ . On the second step  $\theta_0$  is re-estimated with  $W_T = \hat{S}_T(1)^{-1}$ . The resulting estimator,  $\hat{\theta}_T(2)$ , has the minimum asymptotic covariance matrix given in Theorem 4. However, this *two-step* estimator is based on a version of the optimal weighting matrix constructed using a sub-optimal estimator of  $\theta_0$ . This suggests there may be finite sample gains from using  $\hat{\theta}_T(2)$  to construct a new estimator of  $S$ ,  $\hat{S}_T(2)$  say, and then re-estimating  $\theta_0$  with  $W_T = \hat{S}_T(2)^{-1}$ . The resulting estimator,  $\hat{\theta}_T(3)$ , also has the same asymptotic distribution as  $\hat{\theta}_T(2)$  but it is anticipated to be more efficient in finite samples. This potential finite sample gain in efficiency provides a justification for updating the estimate of  $S$  again and re-estimating  $\theta_0$ . This process can be continued iteratively until the estimates converge; if this is done then it yields what has become known as the *iterated GMM estimator*.

The choice of  $W = S^{-1}$  has a second important implication for the asymptotic behavior of the estimator which is presented in the following theorem.<sup>20</sup>

**Theorem 5. Asymptotic independence of the estimator and estimated sample moment.** If (i) Assumptions 1–10 and certain other regularity conditions hold; (ii)  $W = S^{-1}$ ; then  $T^{1/2}(\hat{\theta}_T - \theta_0)$  and  $S^{-1/2}T^{1/2}g_T(\hat{\theta}_T)$  are asymptotically independent.

Since both  $T^{1/2}(\hat{\theta}_T - \theta_0)$  and  $S^{-1/2}T^{1/2}g_T(\hat{\theta}_T)$  are asymptotically normally distributed, Theorem 5 is established by showing that these two statistics are asymptotically uncorrelated. The latter can be deduced from (11.18) and (11.25). Using Assumption 10 and putting  $W = S^{-1}$ , it follows from (11.18) and (11.25) that

$$T^{1/2}(\hat{\theta}_T - \theta_0) = H_{1,T} + o_p(1), \quad (11.26)$$

$$W_T^{1/2}T^{1/2}g_T(\hat{\theta}_T) = H_{2,T} + o_p(1), \quad (11.27)$$

where  $H_{1,T} = -[F(\theta_0)'F(\theta_0)]^{-1}F(\theta_0)'S^{-1/2}T^{1/2}g_T(\theta_0)$  and  $H_{2,T} = [I_q - P(\theta_0)]S^{1/2}T^{1/2}g_T(\theta_0)$ . If we let  $C = \lim_{T \rightarrow \infty} \text{cov}[H_{1,T}, H_{2,T}]$  then it follows from Theorems 2 and 3 that

$$C = \lim_{T \rightarrow \infty} E[H_{1,T}H_{2,T}']. \quad (11.28)$$

Using (11.25) and (11.26) in (11.28), we obtain

$$\begin{aligned} C &= \lim_{T \rightarrow \infty} E[-[F(\theta_0)'F(\theta_0)]^{-1}F(\theta_0)'S^{-1/2}T^{1/2}g_T(\theta_0)T^{1/2}g_T(\theta_0)'S^{-1/2}[I_q - P(\theta_0)]] \\ &= -[F(\theta_0)'F(\theta_0)]^{-1}F(\theta_0)'S^{-1/2}\left\{\lim_{T \rightarrow \infty} \text{var}[T^{1/2}g_T(\theta_0)]\right\}S^{-1/2}[I_q - P(\theta_0)] \\ &= -[F(\theta_0)'F(\theta_0)]^{-1}F(\theta_0)'S^{-1/2}SS^{-1/2}[I_q - P(\theta_0)] \\ &= 0 \end{aligned}$$

because  $S = S^{1/2}S^{1/2}$  implies  $S^{-1/2}SS^{-1/2} = I_q$ , and  $F(\theta_0)'[I_q - P(\theta_0)] = 0$ .

In contrast, if  $W \neq S^{-1}$  then the same sequence of arguments yields the conclusion that  $C \neq 0$ . Therefore, Theorem 5 provides an interesting perspective on why this choice of  $W$  leads to an efficient estimator:  $W = S^{-1}$  is the only choice of weighting matrix for which the estimator is statistically independent of the part of the moment condition unused in estimation. In other words, by making this choice of  $W$ , we have extracted all possible information about the parameters contained in the sample moment.

The estimators described in this section are often described as “the optimal two-step GMM” or “optimal iterated GMM” estimator. It is important to realize that this optimality only refers to the choice of weighting matrix. These are the most precise GMM estimators which can be constructed from the given population moment condition  $E[f(v_t, \theta_0)] = 0$ . It does not imply that there is anything optimal about the population moment condition itself. The optimal choice of moment condition is discussed in Section 8.

## 6 THE OVERIDENTIFYING RESTRICTIONS TEST

The asymptotic theory so far has been predicated on the assumption that the model is correctly specified in the sense that  $E[f(v_t, \theta_0)] = 0$ . If this assumption is false then the argument behind Theorem 1 breaks down, and it is no longer possible to establish the consistency of the estimator. Since the validity of the population moment condition is central to GMM, it is desirable to develop methods for testing whether the data are consistent with this assumption. If  $p = q$  then it is not possible to examine this hypothesis directly because, as we have seen,  $g_T(\hat{\theta}_T) = 0$ . In this case, the validity of Assumption 3 can only be assessed indirectly using so called conditional moment tests; see Newey (1985) and Tauchen (1985). However, if  $q > p$  then the estimator does not set the sample moment to zero and so this leaves scope for testing  $E[f(v_t, \theta_0)] = 0$  directly. In Section 3, it is shown that GMM estimation effects a decomposition on the population moment condition into identifying and overidentifying restrictions. By definition, the estimator satisfies the sample analog to the identifying restrictions. However, the overidentifying restrictions are ignored in estimation and so are available as a basis for testing the validity of the population moment condition. This leads us to the overidentifying restrictions test, which is routinely reported in applications of GMM.

In practice, it is desirable to base inference on the two-step (or iterated) estimator because it yields the most efficient GMM estimator based on  $E[f(v_t, \theta_0)] = 0$ . Therefore, we restrict attention to this case and so substitute  $W = S^{-1}$ . In Section 3, it is shown that  $Q_T(\hat{\theta}_T)$  can be interpreted as a measure of how close the sample is to satisfying the overidentifying restrictions. This motivated Sargan (1958) to propose using the statistic

$$J_T = TQ_T(\hat{\theta}_T) \quad (11.29)$$

to test whether the overidentifying restrictions are satisfied. His analysis is restricted to linear models estimated by instrumental variables. However, Hansen (1982) extends this approach to nonlinear models. The distribution is given in the following theorem.

**Theorem 6. Asymptotic distribution of the overidentifying restrictions test.** If Assumptions 1–10 and certain other regularity conditions hold and  $W = S^{-1}$  then  $J_T \xrightarrow{d} \chi_{q-p}^2$ .

Notice that the degrees of freedom equal the number of overidentifying restrictions. The limiting distribution can be derived heuristically from our earlier analysis. From (11.25), we have that

$$\begin{aligned} TQ_T(\hat{\theta}_T) &= T^{1/2}g_T(\hat{\theta}_T)' \hat{S}_T^{-1}T^{1/2}g_T(\hat{\theta}_T) \\ &= T^{1/2}g_T(\theta_0)' S^{-1/2}[I_q - P(\theta_0)]S^{-1/2}T^{1/2}g_T(\theta_0) + o_p(1) \end{aligned} \quad (11.30)$$

where we have used the fact that  $I_q - P(\theta_0)$  is a projection matrix. Equation (11.30) implies  $J_T$  is asymptotically equivalent to a quadratic form in a random vector with an  $N(0, I_q)$  distribution and a projection matrix with rank  $q - p$ . The result then follows directly from Rao (1973, p. 186).

To consider the power properties of the test, it is necessary to briefly consider what it means for the model to be misspecified in our context. Taken together, Assumptions 3 and 4 imply the population moment condition is satisfied at some unique value in  $\Theta$ . Therefore, the model is misspecified if there is no value in  $\Theta$  at which the population moment condition holds. Such a situation is captured by the following assumption.

**Assumption 11. Misspecification.**  $E[f(v_t, \theta)] = \mu(\theta)$  where  $\mu : \Theta \rightarrow \mathcal{R}^q$  and  $\|\mu(\theta)\| > 0$  for all  $\theta \in \Theta$ .

One important consequence of Assumption 11 is that the two covariance matrix estimators in (11.21)–(11.22) are inconsistent estimators of the long-run variance because they are calculated under the assumption that  $E[f(v_t, \theta_0)] = 0$ .<sup>21</sup> This inconsistency is not by itself a cause for concern provided  $p \lim_{T \rightarrow \infty} \hat{S}_T^{-1}$  is a non-singular matrix. Such would be the case for the estimator  $\hat{S}_{MD}$ , and in consequence  $J_T$  is  $O_p(T)$  and a consistent test versus the misspecification characterized

in Assumption 11.<sup>22</sup> However, Hall (2000a) shows that  $p \lim_{T \rightarrow \infty} \hat{S}_{\text{HACC}}^{-1}$  is a singular matrix, and that this causes  $J_T$  to be only  $O_p(T/b(T))$  although still consistent. A more powerful test can be constructed by using a version of the HACC which is consistent under both null and alternative hypothesis. Hall (2000a) shows that such an estimator can be constructed by replacing  $\hat{\Gamma}_i$  in (11.22) with

$$\tilde{\Gamma}_i = T^{-1} \sum_{t=i+1}^T [f(v_t, \hat{\theta}_T(1)) - g_T(\hat{\theta}_T(1))] [f(v_{t-i}, \hat{\theta}_T(1)) - g_T(\hat{\theta}_T(1))]'$$

Once this change is made, the overidentifying restrictions test is consistent but now  $O_p(T)$ .

## 7 OTHER ESTIMATORS AS SPECIAL CASES OF GMM

It is remarked in the introduction that GMM estimation encompasses many estimators of interest in econometrics, and so provides a very convenient framework for the examination of various issues pertaining to inference. In this section, we justify this statement and illustrate it using maximum likelihood estimation.

Many econometric estimators are obtained by optimizing a scalar of the form

$$\sum_{t=1}^T N_t(\theta). \quad (11.31)$$

If  $N_t(\theta)$  is differentiable then the estimator,  $\tilde{\theta}$ , is the value which solves the associated first order conditions

$$\sum_{t=1}^T \partial N_t(\tilde{\theta}) / \partial \theta = 0. \quad (11.32)$$

Equation (11.32) implies that  $\tilde{\theta}$  is equivalent to the MM estimator based on the population moment condition

$$E[\partial N_t(\theta_0) / \partial \theta] = 0. \quad (11.33)$$

Since  $\partial N_t(\theta_0) / \partial \theta$  is a  $(p \times 1)$  vector it can be recalled from Section 3 that  $\tilde{\theta}$  is also the GMM estimator based on (11.33).

As an illustration, we now derive the population moment condition implicit in the GMM interpretation of maximum likelihood estimation. Suppose the conditional probability density function of the continuous stationary random vector  $v_t$  given  $\{v_{t-1}, v_{t-2}, \dots\}$  is  $p(v_t; \theta_0, V_{t-1})$  where  $V_{t-1} = (v'_{t-1}, v'_{t-2}, \dots, v'_{t-k})$ . The maximum likelihood estimator (MLE) of  $\theta_0$  based on the conditional log likelihood function is the value of  $\theta$  which maximizes,

$$L_T(\theta) = \sum_{t=1}^T \ln\{p(v_t; \theta, V_{t-1})\}. \quad (11.34)$$

This fits within our framework with  $N_t(\theta) = \ln\{p(v_t; \theta, V_{t-1})\}$  and so the MLE can be interpreted as a GMM estimator based on the population moment condition

$$E[\partial \ln\{p(v_t; \theta, V_{t-1})\} / \partial \theta] = 0. \quad (11.35)$$

Since MLE is derived from a perfectly valid estimation principle in its own right, it is reasonable to question whether there is any value to this GMM interpretation. The advantage of the GMM interpretation is that it focuses attention specifically on the information used in estimation, and thereby facilitates an analysis of the consequences of misspecification. For example, the implementation of MLE requires a specific assumption about the distribution of the data. In many cases economic theory does not provide such information and so it is natural to be concerned about the consequences of choosing the wrong distribution. The GMM interpretation reveals that the estimator is still consistent provided (11.35) holds when expectations are taken with respect to the true distribution. Furthermore, Theorem 2 can be used to deduce the asymptotic distribution of the MLE in misspecified models<sup>23</sup> for which (11.35) holds.

## 8 OPTIMAL MOMENTS AND NEARLY UNINFORMATIVE MOMENTS

Throughout the analysis of the GMM estimator, we have taken the population moment condition as given. However, in practice, a researcher is typically faced with a large set of alternatives from which  $q$  elements are chosen to make up the population moment. In this section we consider the two extreme scenarios in which the “best” choice is made and the “worst” choice is made. To understand what best and worst mean in this context, it is useful to consider two ways in which the choice of population moment condition impacts on the asymptotic analysis. Theorem 1 establishes that the consistency of GMM depends crucially on the identification condition in Assumption 4. Theorem 2 reveals that the asymptotic variance of the estimator depends directly on the choice of moment condition via both  $S$  and  $G_0$ . Therefore, the best choice is the population moment condition which leads to the estimator with the smallest asymptotic variance. Section 8.1 summarizes the main results in the literature on the best or *optimal* choice of population moment condition. The worst case scenario is when the population moment condition does not or nearly does not provide enough information to identify  $\theta_0$ . Section 8.2 describes both the consequences of (nearly) uninformative population moment conditions for the inference techniques discussed above and also how the problems can be circumvented.

### 8.1 The optimal choice

In its most general form, we have already answered this question in Section 7. It is shown there that MLE can be interpreted as a GMM estimator based on (11.35). Since the MLE is known to be asymptotically efficient in the class of consistent uniformly asymptotically normal estimators, the optimal choice of population moment condition is just the score function associated with the true probability

distribution of the data. Unfortunately, in many cases of interest in economics, the true probability distribution is unknown. One solution is to choose a distribution arbitrarily but this strategy can have undesirable consequences if the wrong choice is made. In this case, the estimator is no longer asymptotically efficient and may also be inconsistent in nonlinear models.<sup>24</sup> In many cases where MLE is infeasible, GMM is applied using a population moment condition which takes the form

$$E[z_t \otimes u_t(\theta_0)] = 0, \quad (11.36)$$

where  $z_t$  is a vector of observable instruments and  $u_t(\theta_0)$  is a vector of functions which depend on both the data and the unknown parameter vector. Hansen and Singleton (1982) refer to GMM estimation based on (11.36) as *generalized instrumental variables*. Notice that both our examples from the introduction fit into this class. Within this framework  $u_t(\theta_0)$  is usually determined by the model, and so the only difference between choices of moment condition arises from the choice of instrument vector. Therefore, the optimal moment condition is characterized by finding the optimal choice of instrument vector.

In the literature on optimal instruments, it is customary to work with a slightly modified version of the population moment condition.<sup>25</sup> Instead of (11.36), the population moment condition takes the form

$$E[f(v_t, \theta_0)] = E[Z(v_{2t})u_t(\theta_0)] = 0, \quad (11.37)$$

where  $u_t(\theta_0)$  is a  $(s \times 1)$  vector of functions which satisfies  $E[u_t(\theta_0) | \Omega_t] = 0$ ,  $\Omega_t$  represents the information set at time  $t$ ,  $Z(v_{2t})$  is a  $(q \times s)$  matrix whose elements are functions of  $v_{2t} \in \Omega_t$ , and we have partitioned  $v_t = (v_{1t}, v_{2t})'$ . The problem is then to find the optimal choice of  $Z(v_{2t})$ . This question is typically broken down into two parts: what is the optimal choice of  $Z(\cdot)$  for a given choice of  $v_{2t}$ ? and then what is the optimal choice of  $v_{2t}$ ? The answer to the second question is going to depend on the model in question and so we do not address that here. Instead we focus entirely on the first question.

It turns out that the optimal instrument is relatively easy to characterize in static models, but is much more difficult in time series models. We therefore introduce the following restriction.

**Assumption 12. Independence.**  $\{v_t; t = 1, 2 \dots T\}$  forms an independent sequence.

Notice that Assumptions 1 and 12 imply  $v_t$  forms an iid process.

If GMM estimation is based on the population moment condition in (11.37) with the optimal choice of weighting matrix, then from Theorems 2 and 4 the asymptotic covariance matrix of the  $\hat{\theta}_T$  is

$$V(Z) = \left\{ E \left[ \left( \frac{\partial u_t(\theta_0)}{\partial \theta'} \right)' Z_t' \right] S_Z^{-1} E \left[ Z_t \frac{\partial u_t(\theta_0)}{\partial \theta'} \right] \right\}^{-1}, \quad (11.38)$$

where for simplicity we have set  $Z_t = Z(v_{2t})$  and  $S_Z = E[Z_t u_t(\theta_0) u_t(\theta_0)' Z_t']$ . The optimal choice of  $Z(\cdot)$  given  $v_{2t}$  is the function which minimizes  $V(Z)$  in a matrix sense, and this is given by the next theorem.<sup>26</sup>

**Theorem 7. The optimal choice of instrument in static models.** If (i) Assumptions 1–10, 12 and certain other regularity conditions hold; (ii) the population moment condition is given by (11.37); then the optimal choice of  $Z(\cdot)$  given  $v_{2t}$  is

$$Z^0(v_{2t}) = E[\partial u_t(\theta_0)/\partial \theta' | v_{2t}]' \Sigma_{u|v2}^{-1}$$

where  $\Sigma_{u|v2} = E[u_t(\theta_0) u_t(\theta_0)' | v_{2t}]$ , and this choice leads to a GMM estimator with asymptotic covariance matrix

$$V(Z^0) = \{E[Z^0(v_{2t})] \Sigma_{u|v2} E[Z^0(v_{2t})']\}^{-1}.$$

An intuition for this result can be derived by relating the optimal estimator to the familiar case of two-stage least squares (2SLS) estimation in the linear model. For expositional simplicity, we consider the case in which  $s = 1$ , and so let  $\Sigma_{u|v2} = \sigma^2$ . To set up the analogy to 2SLS, it is necessary to return to the asymptotic behavior of  $T^{1/2}(\hat{\theta}_T - \theta_0)$ . Our previous analysis indicates that  $T^{1/2}(\hat{\theta}_T - \theta_0)$  is asymptotically equivalent to the function of the data in (11.26). Using (11.37), (11.26) becomes

$$\begin{aligned} T^{1/2}(\hat{\theta}_T - \theta_0) &= -\{[T^{-1}D_T(\theta_0)' \tilde{Z}_T] W_T [T^{-1}\tilde{Z}_T D_T(\theta_0)]\}^{-1} \\ &\quad \times [T^{-1}D_T(\theta_0)' \tilde{Z}_T] W_T T^{-1/2} \tilde{Z}_T' U_T(\theta_0) + o_p(1), \end{aligned} \quad (11.39)$$

where  $D_T(\theta_0)$  is the  $T \times p$  matrix with  $t$ th row  $\partial u_t(\theta_0)/\partial \theta'$ ,  $\tilde{Z}_T$  is the  $T \times q$  matrix with  $t$ th row  $Z_t$ , and  $U_T(\theta_0)$  is the  $T \times 1$  vector with  $t$ th element  $u_t(\theta_0)$ . Assumption 12 implies that the optimal choice of weighting matrix is  $S^{-1} = \sigma^{-2}\{E[Z_t Z_t']\}^{-1}$ . Since the scaling factor  $\sigma^{-2}$  cancels out in the formula for the estimator, the two-step GMM estimator can be obtained by setting  $W_T = (T^{-1}\tilde{Z}_T \tilde{Z}_T')^{-1}$ . Making this substitution in (11.39), we obtain

$$\begin{aligned} T^{1/2}(\hat{\theta}_T - \theta_0) &= -\{[T^{-1}D_T(\theta_0)' \tilde{Z}_T] [T^{-1}\tilde{Z}_T \tilde{Z}_T']^{-1} [T^{-1}\tilde{Z}_T D_T(\theta_0)]\}^{-1} \\ &\quad \times [T^{-1}D_T(\theta_0)' \tilde{Z}_T] [T^{-1}\tilde{Z}_T \tilde{Z}_T']^{-1} T^{-1/2} \tilde{Z}_T' U_T(\theta_0) + o_p(1). \end{aligned} \quad (11.40)$$

In the linear regression model,  $u_t(\theta_0) = y_t - x_t' \theta_0$  and so  $D_T(\theta_0) = -X$ , the matrix of observations on  $x_t$ . In this case, (11.40) reduces to the formula for the linear IV estimator and the optimal instrument is  $E[x_t | Z_t]$ . If  $x_t$  is assumed to be a linear function of  $Z_t$  then the feasible optimal IV estimator is just the two-stage least squares estimator.<sup>27</sup>

Now let us return to the original nonlinear setting. By analogy to the linear model, (11.40) implies that  $T^{1/2}(\hat{\theta}_T - \theta_0)$  behaves asymptotically like an IV estimator in a linear model with regressor vector,  $x_t = -\partial u_t(\theta_0)/\partial \theta'$  and error,  $u_t(\theta_0)$ . Now we have just argued that the optimal instrument in a linear model is given

by the conditional expectation of the regressor. Therefore, applying that logic here, the optimal instrument is given by  $-E[\partial u_t(\theta_0)/\partial\theta'|Z_t]$ , which is identical to the result in Theorem 7 except for the presence of the scaling factor,  $-\sigma^2$ . This difference is inconsequential because, as remarked above, the scaling factor cancels out and so does not effect the estimator. To construct a feasible optimal instrument, it is possible to follow a similar strategy to 2SLS and assume a model for  $\partial u_t(\theta_0)/\partial\theta'$ . However, this is likely to require an assumption about the distribution of  $v_{2t}$  in order to evaluate the expectation. This is undesirable here because it is the absence of this information which led us to generalized IV estimation in the first place. An alternative solution is to estimate  $Z^0(v_{2t})$  nonparametrically, and Newey (1993) provides a survey of various methods which have been proposed in this context.

The above discussion gives an intuition for the part of  $Z^0(v_{2t})$  involving the partial derivative, but does not explain the presence of  $\Sigma_{u|v}^{-1}$  in the formula because the  $\sigma^2$  factor canceled out. However, if  $s > 1$  and  $\Sigma_{u|v} \neq \sigma^2 I_s$  then it is necessary to employ a correction in the construction of the optimal instrument for either the unequal variances or any contemporaneous correlation (or both) of the elements of  $u_t(\theta_0)$ . It is for this reason that  $Z^0(v_{2t})$  is transformed by  $\Sigma_{u|v}^{-1}$ .<sup>28</sup>

The matrix  $V(Z^0)$  can be interpreted as a lower bound on the asymptotic covariance matrix for this class of estimators. It should be remembered that the optimal IV estimator is likely to be less efficient than maximum likelihood because (11.37) does not typically contain all the information in the true score function of the data. However, there is a sense in which  $V(Z^0)$  is the best we can do given the information available. Chamberlain (1987) shows that  $V(Z^0)$  is also the lower bound on the asymptotic covariance matrix of *any* consistent and asymptotically normal estimator of  $\theta_0$  in which the only substantive information used in estimation is the population moment condition in (11.39).<sup>29</sup>

It would be desirable to extend this theorem to time series, but so far there has only been limited success in this direction. Hansen and Singleton (1982), Hayashi and Sims (1983), Hansen (1985) and Hansen, Heaton, and Ogaki (1988) have all provided characterizations of a lower bound on the asymptotic variance under different assumptions about the functional form of  $u_t(\theta_0)$  and its dynamic structure. However, as yet, these results have not been translated into general algorithms for the calculation of a feasible optimal instrument in dynamic non-linear models.<sup>30</sup>

## 8.2 Nearly uninformative moment conditions

While it is desirable to base estimation on the optimal moment conditions, this is not necessary. Even if the population moment condition is sub-optimal, the GMM framework can be used to obtain consistent, asymptotically normal estimators *provided* that the parameter is identified. In recent years, there has been a growing awareness that this proviso may not be so trivial in situations which arise in practice. In a very influential paper, Nelson and Startz (1990) drew attention to this potential problem and provided the first evidence of the problems it causes for the inference framework we have described above. Their paper has prompted

considerable interest in the behavior of GMM in cases in which the population moment condition provided is nearly uninformative about  $\theta_0$ . In this section we concentrate on illustrating the nature of the problem, and then briefly consider a potential solution.

For expositional simplicity, we restrict attention to the simple linear regression model,

$$y_t = x_t \theta_0 + u_t, \quad (11.41)$$

in which  $u_t$  is an iid process with mean zero and variance  $\sigma^2$ . Suppose the scalar parameter  $\theta_0$  is estimated by instrumental variables which, as we have seen, is just GMM estimation based on the population moment condition

$$E[z_t u_t(\theta_0)] = 0, \quad (11.42)$$

where  $z_t$  is a  $q \times 1$  vector of instruments and  $u_t(\theta_0) = y_t - x_t \theta_0$ . From Lemma 1,  $\theta_0$  is identified<sup>31</sup> by (11.42) if  $\text{rank}\{E[z_t x_t]\} = 1$ . In this simple example,  $\theta_0$  is unidentified if  $E[z_t x_t]$  is the null vector, which would occur if  $z_t$  and  $x_t$  are uncorrelated and both possess zero means. In practice, it is unlikely that  $E[z_t x_t]$  is exactly zero. The contribution of Nelson and Startz's (1990) paper is to demonstrate that problems occur if  $E[z_t x_t]$  is nonzero but small.<sup>32</sup> It is this scenario which we refer to as "nearly uninformative moment conditions."

To proceed, it is necessary to develop a model which can capture the idea of nearly uninformative moment conditions. Following Staiger and Stock (1997), we solve this problem by assuming that

$$x_t = z_t' \gamma_T + \varepsilon_t, \quad (11.43)$$

where  $\gamma_T = T^{-1/2}c$ ,  $c$  is a nonzero  $q \times 1$  vector of constants, and  $\varepsilon_t$  is the unobserved error which has both a zero mean and is uncorrelated with  $z_t$ .<sup>33</sup> Notice that (11.43) implies that  $E_T[z_t x_t] = E[z_t z_t']T^{-1/2}c$  and so is nonzero for finite  $T$  but zero in the limit as  $T \rightarrow \infty$ .<sup>34</sup> So the concept of nearly uninformative moment conditions is captured by assuming that  $\{x_t; t = 1, 2, \dots, T\}$  is generated by a sequence of processes whose relationship to  $z_t$  disappears at rate  $T^{1/2}$ . This rate is chosen so that the effects of the nearly uninformative moment conditions manifest themselves in the limiting behavior of the estimator. Since  $p = 1$ , we have

$$\hat{\theta}_T - \theta_0 = \frac{x' Z(Z'Z)^{-1} Z'u}{x' Z(Z'Z)^{-1} Z'x}. \quad (11.44)$$

To analyze the limiting behavior of  $\hat{\theta}_T - \theta_0$  it is necessary to impose certain regularity conditions. We explicitly assume that  $z_t$  is independent of  $u_t$ , but leave the other necessary regularity conditions unstated for brevity. Using the weak law of large numbers and the central limit theorem respectively, it follows that: (i)  $T^{-1}Z'Z = M_{zz}$ , a positive definite matrix of constants; (ii)  $T^{-1/2}Z'u \xrightarrow{d} N(0, \sigma^2 M_{zz})$ . Notice that neither (i) nor (ii) involve the relationship between  $x_t$  and  $z_t$

and so would equally hold if  $\theta_0$  is properly identified. The key difference comes in the behavior of  $Z'x$ . From (11.43), it follows that

$$Z'x = T^{-1/2}Z'Zc + Z'\epsilon, \quad (11.45)$$

where  $\epsilon$  is the  $T \times 1$  vector with  $t$ th element  $\epsilon_t$ . Therefore,  $T^{-1}Z'x \xrightarrow{p} 0$  and  $T^{-1/2}Z'x \xrightarrow{d} N(M_{zz}c, \sigma_\epsilon^2 M_{zz})$ . The nature of this limiting behavior means that,

$$\begin{aligned} \hat{\theta}_T - \theta_0 &= \frac{T^{-1/2}x'Z(T^{-1}Z'Z)^{-1}T^{-1/2}Z'u}{T^{-1/2}x'Z(T^{-1}Z'Z)^{-1}T^{-1/2}Z'x} \\ &\xrightarrow{d} \frac{\Psi'_1 M_{zz}^{-1} \Psi_2}{\Psi'_1 M_{zz}^{-1} \Psi_1} \end{aligned} \quad (11.46)$$

where  $\Psi_1 \sim N(M_{zz}c, \sigma_\epsilon^2 M_{zz})$  and  $\Psi_2 \sim N(0, \sigma_\epsilon^2 M_{zz})$ . Therefore,  $\hat{\theta}_T$  converges to a random variable if the moment conditions are nearly uninformative in the sense of (11.43). This is in marked contrast to the case when  $\theta_0$  is identified in the sense of Assumption 4. In that case, Theorem 1 indicates  $\hat{\theta}_T$  converges in probability to  $\theta_0$ .

This analysis provides an indication that the asymptotic theory derived in Section 4 is inappropriate for the nearly uninformative moment condition case. It is unlikely to be known a priori if the population moment condition in question is informative – in the sense of Assumption 4 – or nearly uninformative. Therefore, it is useful to develop statistical tests to discriminate between the two cases. In our linear model example, a natural diagnostic is the F-statistic for the hypothesis  $x_t$  is linearly unrelated to  $z_t$ . If this hypothesis is not rejected then this can be interpreted as evidence that identification of  $\theta_0$  is suspect. Faced with an insignificant F-statistic, there are two possible responses. One strategy is to keep changing the instrument vector until the F-statistic is significant. However, Hall, Rudebusch and Wilcox (1996) report evidence that this approach does not solve the problem and in fact tends to make matters worse. A second, and more promising, strategy is to develop an inference theory which provides a better approximation in the nearly uninformative moment condition case. This line of research is still in its early stages but significant advances have been made by Staiger and Stock (1997), Stock and Wright (1997), and Wang and Zivot (1998).

## 9 FINITE SAMPLE BEHAVIOR

The foregoing discussion has rested upon asymptotic theory. In finite samples, such theory can only provide an approximation. It is therefore important to assess the quality of this approximation in the types of model and sample sizes that are encountered in economics and finance. Intuition suggests that the quality is going to vary from case to case depending on the form of the nonlinearity and the dynamic structure. A number of simulation studies have examined this question; see *inter alia* Tauchen (1986), Kocherlakota (1990) and the seven papers included in the July 1996 issue of *Journal of Business and Economics Statistics*. It is

beyond the scope of this article to provide a comprehensive review of these studies.<sup>35</sup> However, it should be noted that in certain circumstances of interest the quality of the approximation is poor. In view of this evidence, it is desirable to develop methods which improve the quality of finite sample inferences. One such method is the bootstrap, and this has been explored in the context of GMM by Hall and Horowitz (1996).

## Notes

- \* I am grateful to Atsushi Inoue, Fernanda Peixe, James Stock, and three anonymous reviewers for comments on an earlier draft of this paper.
- 1 Hall (1993) and Ogaki (1993) provide an overview of the areas in which GMM has been applied.
- 2 Also see Hall (1998) for a survey of hypothesis tests based specifically on GMM estimators.
- 3 This generic approach is known as the consumption based asset pricing model.
- 4 It is possible to generalize the arguments to allow for certain types of nonstationarity; see Gallant and White (1988), Pötscher and Prucha (1997).
- 5 E.g. see Apostol (1974, p. 361).
- 6 E.g. see Dhrymes (1984, Proposition 92, p. 111).
- 7 For example see Quandt (1983) or Gallant (1987, ch. 2).
- 8 This property is not guaranteed by pointwise convergence of  $Q_T(\theta)$ . See Apostol (1974, ch. 9) for a useful discussion of the difference between pointwise and uniform convergence.
- 9 E.g. see Apostol (1974, p. 355).
- 10 E.g. see Fuller (1976, p. 199). Hansen (1982) and Wooldridge (1994) provide formal proofs of the theorem.
- 11 For example see Hamilton (1994, pp. 279–80).
- 12 See Hamilton (1994, pp. 261–2) for a discussion of the properties of autocovariance matrices.
- 13 For example see White (1994, Theorem 8.27, p. 193).
- 14 The requirement that  $S_T$  be positive semi-definite (p.s.d.) is the matrix generalization of a nonnegative scalar variance. This property is not guaranteed for estimators of the generic form in (11.22). For example  $\omega_{it} = 1$  does not yield a p.s.d. matrix; see Newey and West (1987).
- 15 See *inter alia* Newey and West (1987), Gallant (1987), Andrews (1991), Andrews and Monahan (1992).
- 16 Andrews (1991) and Newey and West (1994) propose data-based methods for bandwidth selection.
- 17 See Dhrymes (1984, p. 17).
- 18 See Rao (1973, ch. 8) for a discussion of the singular normal distribution.
- 19 If  $p = q$  then the asymptotic variance of  $\hat{\theta}_T$  is  $MSM' = (G_0' S^{-1} G_0)^{-1}$ .
- 20 See Hall (2000b, ch. 3).
- 21 It is common to impose this assumption in both theoretical treatments and applications of these long-run variance estimators in the context of GMM.
- 22 See Hall (2000b, ch. 5).
- 23 White (1982) refers to such an estimator as *quasi*-maximum likelihood.
- 24 See Hansen and Singleton (1982).
- 25 This difference facilitates the analysis but makes no difference to the ultimate result.

- 26 See Newey (1993) for a formal proof.
- 27 See Hall (1993) or Theil (1971, pp. 451–3).
- 28 Notice that if  $\Sigma_{u|v_2} = \sigma^2 I_s$  then the  $\sigma^2$  factor cancels out as in our example.
- 29 Chamberlain's (1987) analysis is based on a form of semiparametric maximum likelihood subject to (11.37). Also see Newey (1993, pp. 423–4).
- 30 One exception is the case in which  $u_t(\theta_0)$  is a martingale difference case for which Hansen (1985) shows Theorem 7 extends directly with only a slight modification to make allowance for conditional heteroskedasticity.
- 31 In the linear model, global and local identification are equivalent because (11.6) is no longer an approximation but is an identity which holds over  $\Theta$ .
- 32 This terminology parallels the distinction between exact and near collinearity in the linear regression model.
- 33 Equation (11.43) implies the explanatory variable is a triangular array  $\{x_{t,T} : t = 1, 2 \dots T; T = 1, 2 \dots\}$  but we suppress the second subscript for notational brevity.
- 34 Notice that the data generation process for  $x_t$  changes with  $T$  and it is for this reason that the expectations operator is indexed by  $T$ .
- 35 The interested reader is referred to Hall (1999b, ch. 6).

## References

- Andrews, D.W.K. (1991). Heteroscedasticity and autocorrelation consistent covariance matrix estimation. *Econometrica* 59, 817–58.
- Andrews, D.W.K., and J.C. Monahan (1992). An improved heteroscedasticity and autocorrelation consistent covariance matrix. *Econometrica* 60, 953–66.
- Angrist, J.D., and A.B. Krueger (1992). The effect of age at school entry on educational attainment: an application of instrumental variables with moments from two samples. *Journal of the American Statistical Association* 87, 328–36.
- Apostol, T. (1974). *Mathematical Analysis*, 2nd edn. Reading, MA: Addison-Wesley.
- Chamberlain, G. (1987). Asymptotic efficiency in estimation with conditional moment restrictions. *Journal of Econometrics* 34, 305–34.
- Davidson, J. (1994). *Stochastic Limit Theory*. Oxford: Oxford University Press.
- Dhrymes, P.J. (1984). *Mathematics for Econometrics*, 2nd edn. New York: Springer-Verlag.
- Fuller, W.A. (1976). *Introduction to Statistical Time Series*. New York: Wiley.
- Gallant, A.R. (1987). *Nonlinear Statistical Models*. New York: Wiley.
- Gallant, A.R., and H. White (1988). *A Unified Theory of Estimation and Inference in Nonlinear Models*. Oxford: Basil Blackwell.
- Geary, R.C. (1942). Inherent relations between random variables. *Proceedings of the Royal Irish Academy, Section A* 47, 63–76.
- Hall, A.R. (1993). Some aspects of Generalized Method of Moments estimation. In G.S. Maddala, C.R. Rao, and H.D. Vinod (eds.) *Handbook of Statistics, Volume 11*. pp. 393–417. Amsterdam: Elsevier Science Publishers.
- Hall, A.R. (1998). Hypothesis testing in models estimated by Generalized Method of Moments. In L. Mátyás (ed.) *Generalized Method of Moments*. pp. 75–101. Cambridge: Cambridge University Press.
- Hall, A.R. (2000a). Covariance matrix estimation and the power of the overidentifying restrictions test. Discussion paper, Department of Economics, North Carolina State University, Raleigh NC.
- Hall, A.R. (2000b). *Generalized Method of Moments*. Manuscript in preparation, Oxford: Oxford University Press.

- Hall, A.R., G. Rudebusch, and D. Wilcox (1996). Judging instrument relevance in instrumental variables estimation. *International Economic Review* 37, 283–98.
- Hall, P., and J.L. Horowitz (1996). Bootstrap critical values for tests based on generalized Method of Moments. *Econometrica* 64, 891–917.
- Hamilton, J.D. (1994). *Time Series Analysis*. Princeton NJ: Princeton University Press.
- Hansen, L.P. (1982). Large sample properties of Generalized Method of Moments estimators. *Econometrica* 50, 1029–54.
- Hansen, L.P. (1985). A method of calculating bounds on the asymptotic covariance matrices of generalized method of moments estimators. *Journal of Econometrics* 30, 203–38.
- Hansen, L.P., J. Heaton, and M. Ogaki (1988). Efficiency bounds implied by multi-period conditional moment restrictions. *Journal of the American Statistical Association* 83, 863–71.
- Hansen, L.P., and K.S. Singleton (1982). Generalized instrumental variables estimation of nonlinear rational expectations models. *Econometrica* 50, 1269–86.
- Hayashi, F., and C. Sims (1983). Nearly efficient estimation of time series models with predetermined, but not exogenous instruments. *Econometrica* 51, 783–98.
- Kocherlakota, N.R. (1990). On tests of representative consumer asset pricing models. *Journal of Monetary Economics* 26, 285–304.
- Nelson, C.R., and R. Startz (1990). The distribution of the instrumental variables estimator and its *t* ratio when the instrument is a poor one. *Journal of Business* 63, S125–S140.
- Newey, W.K. (1985). Maximum likelihood specification testing and instrumented score tests. *Econometrica* 53, 1047–70.
- Newey, W.K. (1993). Efficient estimation of models with conditional moment restrictions. In G.S. Maddala, C.R. Rao, and H.D. Vinod (eds.) *Handbook of Statistics, Volume 2*. pp. 419–54. Amsterdam: Elsevier Science Publishers.
- Newey, W.K., and D.L. McFadden (1994). Large sample estimation and hypothesis testing. In R. Engle and D.L. McFadden (eds.) *Handbook of Econometrics, Volume 4*. pp. 2113–247. Amsterdam: Elsevier Science Publishers.
- Newey, W.K., and K.D. West (1987). A simple positive semi-definite heteroscedasticity and autocorrelation consistent covariance matrix. *Econometrica* 55, 703–8.
- Newey, W.K., and K.D. West (1994). Automatic lag selection in covariance matrix estimation. *Review of Economic Studies* 61, 631–53.
- Ogaki, M. (1993). Generalized Method of Moments: econometric applications. In G.S. Maddala, C.R. Rao, and H.D. Vinod (eds.) *Handbook of Statistics, Volume 11*. pp. 455–88. Amsterdam: Elsevier Science Publishers.
- Pearson, K.S. (1893). Asymmetrical frequency curves. *Nature* 48: 615–16.
- Pearson, K.S. (1894). Contributions to the mathematical theory of evolution. *Philosophical Transactions of the Royal Society of London (A)* 185, 71–110.
- Pearson, K.S. (1895). Contributions to the mathematical theory of evolution, II: skew variation. *Philosophical Transactions of the Royal Society of London (A)* 186, 343–414.
- Pötscher, B.M., and I.R. Prucha (1997). *Dynamic Nonlinear Econometric Models*. Berlin: Springer Verlag.
- Quandt, R.E. (1983). Computational problems and methods. In Z. Griliches and M.D. Intriligator (eds.) *Handbook of Econometrics, Volume 1*. pp. 699–764. Amsterdam: Elsevier Science Publishers.
- Rao, C.R. (1973). *Linear Statistical Inference and its Applications*, 2nd edn. New York: Wiley.
- Reiersol (1941). Confluence analysis by means of lag moments and other methods of confluence analysis. *Econometrica* 9, 1–24.
- Sargan, J.D. (1958). The estimation of economic relationships using instrumental variables. *Econometrica* 26, 393–415.

- Sowell, F. (1996). Optimal tests of parameter variation in the Generalized Method of Moments framework. *Econometrica* 64, 1085–108.
- Staiger, D., and J. Stock (1997). Instrumental variables regression with weak instruments. *Econometrica* 65, 557–86.
- Stock, J., and J. Wright (1997). GMM with weak identification. Discussion paper, Kennedy School of Government, Harvard University, Cambridge, MA.
- Tauchen, G. (1985). Diagnostic testing and evaluation of maximum likelihood models. *Journal of Econometrics* 30, 415–43.
- Tauchen, G. (1986). Statistical properties of Generalized Method of Moments estimators of structural parameters obtained from financial market data. *Journal of Business and Economic Statistics* 4, 397–416.
- Theil, H. (1971). *Principles of Econometrics*. New York: Wiley.
- Wang, J. and E. Zivot (1998). Inference on structural parameters in instrumental variables regression with weak instruments. *Econometrica* 66, 1389–404.
- White, H. (1982). Maximum likelihood in misspecified models. *Econometrica*. 50, 1–25.
- White, H. (1994). *Estimation, Inference and Specification Analysis*. New York: Cambridge University Press.
- Wooldridge, J.M. (1994). Estimation and inference for dependent processes. In R. Engle and D.L. McFadden (eds.) *Handbook of Econometrics, Volume 4*. pp. 2641–739. Amsterdam: Elsevier Science Publishers.
- Wright, S. (1925). Corn and hog correlations. Discussion paper, US Department of Agriculture Bulletin No. 1300, Washington, DC.

---

C H A P T E R   T W E L V E

# Collinearity

*R. Carter Hill and Lee C. Adkins\**

Multicollinearity is God's will, not a problem  
with OLS or statistical techniques in general.

*Blanchard (1987, p. 49)*

## 1 INTRODUCTION

Collinearity, a devilish problem to be sure, receives the blame for a substantial amount of inconclusive, weak, or unbelievable empirical work. Social scientists are, for the most part, nonexperimental scientists. We do not have the luxury of designing and carrying out the experiments that generate our data. Consequently our data are often weak and not up to the task of isolating the effect of changes in one economic variable upon another. In regression models the least squares estimates may have the wrong sign, be sensitive to slight changes in the data or the model specification, or may not yield statistically significant results for theoretically important explanatory variables. These symptoms may appear despite significant values for the overall F-test of model significance or high  $R^2$  values. These are commonly cited consequences of a "collinearity problem."

In the context of the linear regression model, collinearity takes three distinct forms. First, an explanatory variable may exhibit little variability. Intuitively, this is a problem for regression because we are trying to estimate the effect of changes in the explanatory variable upon the dependent variable. If an explanatory variable does not vary much in our sample, it will be difficult to estimate its effect. Second, two explanatory variables may exhibit a large correlation. In this case, the attempt to isolate the effect of one variable, all other things held constant, is made difficult by the fact that in the sample the variable exhibits little *independent* variation. The correlation between two explanatory variables implies that changes in one are linked to changes in the other, and thus separating out their individual effects may be difficult. Third, and generally, there may be one, or more, nearly exact linear relationship among the explanatory variables. As in the case when

two explanatory variables are correlated, such relationships obscure the effects of involved variables upon the dependent variable. These are the three faces of collinearity.

In this chapter we explore collinearity in linear and nonlinear models. In Section 2 we present the basics, examining the forms that collinearity may take and the damage it does to estimation. The variance decomposition of Belsley, Kuh, and Welsch (1980) (hereinafter BKW) is presented in Section 3, and other collinearity diagnostics and issues are considered in Section 4. Section 5 reviews suggested remedies for collinearity problems. In Section 6 we examine the problems of collinearity in nonlinear models. Summary remarks are contained in Section 7.

## 2 THE NATURE AND STATISTICAL CONSEQUENCES OF COLLINEARITY

Consider first a linear regression model with two explanatory variables,

$$y_t = \beta_1 + \beta_2 x_{t2} + \beta_3 x_{t3} + e_t. \quad (12.1)$$

Assume that the errors are uncorrelated, with mean zero and constant variance,  $\sigma^2$ , and that  $x_{t2}$  and  $x_{t3}$  are nonstochastic. Under these assumptions the least squares estimators are the best, linear, unbiased estimators of the regression parameters. The variance of the least squares estimator  $b_2$  of  $\beta_2$  is

$$\text{var}(b_2) = \frac{\sigma^2}{\sum_{t=1}^T (x_{t2} - \bar{x}_2)^2 (1 - r_{23}^2)}, \quad (12.2)$$

where  $\bar{x}_2$  is the sample mean of the  $T$  observations on  $x_{t2}$ , and  $r_{23}$  is the sample correlation between  $x_{t2}$  and  $x_{t3}$ . The formula for the variance of  $b_3$ , the least squares estimator of  $\beta_3$ , is analogous, but the variance of the intercept estimator is messier and we will not discuss it here. The covariance between  $b_2$  and  $b_3$  is

$$\text{cov}(b_2, b_3) = \frac{-r_{23}\sigma^2}{(1 - r_{23}^2)\sqrt{\sum_{t=1}^T (x_{t2} - \bar{x}_2)^2} \sqrt{\sum_{t=1}^T (x_{t3} - \bar{x}_3)^2}}. \quad (12.3)$$

The variance and covariance expressions reveal the consequences of two of the three forms of collinearity. First, suppose that  $x_{t2}$  exhibits little variation about its sample mean, so that  $\sum(x_{t2} - \bar{x}_2)^2$  is small. The less the variation in the explanatory variable  $x_{t2}$  about its mean, the larger will be the variance of  $b_2$ , and the larger will be the covariance, in absolute value, between  $b_2$  and  $b_3$ . Second, the larger the correlation between  $x_{t2}$  and  $x_{t3}$  the larger will be the variance of  $b_2$ , and the larger will be the covariance, in absolute value, between  $b_2$  and  $b_3$ . If the correlation is positive the covariance will be negative. This is the source of another conventional

observation about collinearity, namely that the coefficients of highly correlated variables tend to have opposite signs.

Exact, or perfect, collinearity occurs when the variation in an explanatory variable is zero,  $\sum(x_{i2} - \bar{x}_2)^2 = 0$ , or when the correlation between  $x_{i2}$  and  $x_{i3}$  is perfect, so that  $r_{23} = \pm 1$ . In these cases the least squares estimates are not unique, and, in absence of additional information, best linear unbiased estimators are not available for all the regression parameters. Fortunately, this extreme case rarely occurs in practice.

The commonly cited symptoms of collinearity, that least squares estimates have the wrong sign, are sensitive to slight changes in the data or the model specification, or are not statistically significant, follow from the large variances of the least squares estimators. The least squares estimators are unbiased under standard assumptions, so that  $E[b_k] = \beta_k$ , but how close an estimate might be to the true parameter value is determined by the estimator variance. Large variances for estimators imply that their sampling (probability) distributions are wide, meaning that in any particular sample the estimates we obtain may be far from the true parameter values.

## 2.1 Collinearity in the linear regression model

Denote the linear regression model as

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{e}, \quad (12.4)$$

where  $\mathbf{y}$  is a  $T \times 1$  vector of observations on the dependent variable,  $\mathbf{X}$  is a  $T \times K$  non-stochastic matrix of observations on  $K$  explanatory variables,  $\boldsymbol{\beta}$  is a  $K \times 1$  vector of unknown parameters, and  $\mathbf{e}$  is the  $T \times 1$  vector of uncorrelated random errors, with zero means and constant variances,  $\sigma^2$ .

In the general linear model exact, or perfect, collinearity exists when the columns of  $\mathbf{X}$ , denoted  $x_i$ ,  $i = 1, \dots, K$ , are linearly dependent. This occurs when there is at least one relation of the form  $a_1x_1 + a_2x_2 + \dots + a_Kx_K = 0$ , where the  $a_i$  are constants, not all equal to zero. In this case the column rank of  $\mathbf{X}$  is less than  $K$ , the normal equations  $\mathbf{X}'\mathbf{X}\boldsymbol{\beta} = \mathbf{X}'\mathbf{y}$  do not have a unique solution, and least squares estimation breaks down. Unique best linear unbiased estimators do not exist for all  $K$  parameters. However, even in this most severe of cases, all is not lost. Consider equation (12.1),  $y_t = \beta_1 + \beta_2x_{i2} + \beta_3x_{i3} + e_t$ . Suppose that  $a_2x_2 + a_3x_3 = 0$ , or more simply,  $x_2 = ax_3$ . Substituting this into (12.1) we obtain  $y_t = \beta_1 + \beta_2(ax_3) + \beta_3x_{i3} + e_t = \beta_1 + (a\beta_2 + \beta_3)x_{i3} + e_t = \beta_1 + \gamma x_{i3} + e_t$ . Thus we can obtain a best linear unbiased estimator of  $\gamma = a\beta_2 + \beta_3$ , a linear combination of the parameters. The classic paper by Silvey (1969) provides expressions for determining which linear combinations of parameters are estimable.

Exact collinearity is rare, and easily recognized. More frequently, one or more linear combinations of explanatory variables are *nearly* exact, so that  $a_1x_1 + a_2x_2 + \dots + a_Kx_K \approx 0$ . We now examine the consequences of such near exact linear dependencies.

## 2.2 Diagnosing collinearity using the singular value decomposition

The singular-value decomposition is a factorization of  $X$ . The matrix  $X$  may be decomposed as  $X = U\Lambda^{1/2}C'$ , where  $U'U = C'C = CC' = I_K$  and  $\Lambda^{1/2}$  is a diagonal matrix with nonnegative diagonal values  $\lambda_1^{1/2}, \lambda_2^{1/2}, \dots, \lambda_K^{1/2}$ , called the singular values of  $X$ . The relation to eigenanalysis is that the singular values are the positive square roots of the eigenvalues of  $X'X$ , and the  $K \times K$  matrix  $C$  is the matrix whose columns contain the eigenvectors of  $X'X$ . Thus

$$C'X'XC = \Lambda, \quad (12.5)$$

where  $\Lambda$  is a diagonal matrix with the real values  $\lambda_1, \lambda_2, \dots, \lambda_K$  on the diagonal. The matrix  $U$  is  $T \times K$ , and its properties are discussed in Belsley (1991, pp. 42–3). The columns of the matrix  $C$ , denoted  $c_i$ , are the eigenvectors (or characteristic vectors) of the matrix  $X'X$ , and the real values  $\lambda_i$  are the corresponding eigenvalues (or characteristic roots). It is customary to assume that the columns of  $C$  are arranged so that the eigenvalues are ordered by magnitude,  $\lambda_1 \geq \lambda_2 \geq \dots \geq \lambda_K$ .

If  $X$  is of full column rank  $K$ , so that there are no exact linear dependencies among the columns of  $X$ , then  $X'X$  is a positive definite and symmetric matrix, and all its eigenvalues are not only real but also positive. If we find a “small” eigenvalue,  $\lambda_i \approx 0$ , then  $c_i'X'Xc_i = (Xc_i)'(Xc_i) = \lambda_i \approx 0$  and therefore  $Xc_i \approx 0$ . Thus we have located a near exact linear dependency among the columns of  $X$ . If there is a single small eigenvalue, then the linear relation  $Xc_i \approx 0$  indicates the form of the linear dependency, and can be used to determine which of the explanatory variables are involved in the relationship.

## 2.3 Collinearity and the least squares estimator

Using equation (12.5), and the orthogonality of  $C$ ,  $C'C = CC' = I_K$ , we can write  $X'X = CAC'$ , and therefore

$$(X'X)^{-1} = C\Lambda^{-1}C' = \sum_{i=1}^K \lambda_i^{-1} c_i c_i'. \quad (12.6)$$

The covariance matrix of the least squares estimator  $b$  is  $\text{cov}(b) = \sigma^2(X'X)^{-1}$ , and using equation (12.6) the variance of  $b_j$  is

$$\text{var}(b_j) = \sigma^2 \left( \frac{c_{j1}^2}{\lambda_1} + \frac{c_{j2}^2}{\lambda_2} + \dots + \frac{c_{jK}^2}{\lambda_K} \right), \quad (12.7)$$

where  $c_{jk}$  is the element in the  $j$ th row and  $i$ th column of  $C$ . The orthogonality of  $C$  implies that  $\sum_{k=1}^K c_{jk}^2 = 1$ . Thus the variance of  $b_j$  depends upon three distinct factors. First, the magnitude of the error variance,  $\sigma^2$ ; second, the magnitudes

of the constants  $c_{jk}$ ; and third, the magnitude of the eigenvalues,  $\lambda_k$ . A small eigenvalue may cause a large variance for  $b_j$  if it is paired with a constant  $c_{jk}$  that is not close to zero. The constants  $c_{jk} = 0$  when  $x_j$  and  $x_k$ , the  $j$ th and  $k$ th columns of  $X$ , respectively, are orthogonal, so that  $x'_j x_k = 0$ . This fact is an important one for it will allow us to determine which variables are “not” involved in collinear relationships.

Suppose  $\beta_j$  is a critical parameter in the model, and there is one small eigenvalue,  $\lambda_K \approx 0$ . If  $x_j$  is not involved in the corresponding linear dependency  $Xc_K \approx 0$ , then  $c_{jk}$  will be small, and the fact that  $\lambda_K \approx 0$ , i.e. the  $K$ th eigenvalue is very small, will not adversely affect the precision of estimation of  $\beta_j$ . *The presence of collinearity in the data does not automatically mean that “all is lost.”* If  $X'X$  has one or more small eigenvalues, then you must think clearly about the objectives of your research, and determine if the collinearity reduces the precision of estimation of your key parameters by an unacceptable amount. We address the question What is a small eigenvalue? in Section 3. For more about the geometry of characteristic roots and vectors see Fomby, Hill, and Johnson (1984, pp. 288–93).

## 2.4 Collinearity and the least squares predictor

Another bit of conventional wisdom is that while collinearity may affect the precision of the least squares estimator, it need not affect the reliability of predictions based on it, if the collinearity in the sample extends to the forecast period. Suppose we wish to predict the value of  $y_0$ , given by  $y_0 = x'_0\beta + e_0$ , where  $x'_0$  is a  $1 \times K$  vector of regressor values, and  $e_0$  is a random disturbance with zero-mean, constant variance  $\sigma^2$ , and which is uncorrelated with the regression disturbances. Using equation (12.6), the best linear unbiased predictor of  $E(y_0)$ ,  $\hat{y}_0 = x'_0 b$ , has variance

$$\text{var}(\hat{y}_0) = \sigma^2 x'_0 (X'X)^{-1} x_0 = \sigma^2 x'_0 C \Lambda^{-1} C' x_0 = \sigma^2 \left( \frac{(x'_0 c_1)^2}{\lambda_1} + \frac{(x'_0 c_2)^2}{\lambda_2} + \cdots + \frac{(x'_0 c_K)^2}{\lambda_K} \right). \quad (12.8)$$

If, for example, there is a single linear dependence among the columns of  $X$ , then  $\lambda_K \approx 0$ , and  $Xc_K \approx 0$ . A small eigenvalue could make the last term of the sum in equation (12.8) large, producing a large prediction variance. However, if (and this is a big *if*) the new observation  $x'_0$  obeys the same collinearity pattern as the sample data, then it may also be true that  $x'_0 c_K \approx 0$ , effectively negating the small eigenvalue.

## 3 THE VARIANCE DECOMPOSITION OF BELSLEY, KUH, AND WELSCH (1980)

A property of eigenvalues is that  $\text{tr}(X'X) = \sum_{i=1}^K \lambda_i$ . This implies that the sizes of the eigenvalues are determined in part by the scaling of the data. Data matrices

**Table 12.1** Matrix of variance proportions

Condition index	Proportions of variances of least squares estimator			
	$var(b_1)$	$var(b_2)$	...	$var(b_K)$
$\eta_1$	$\pi_{11}$	$\pi_{12}$	...	$\pi_{1K}$
$\eta_2$	$\pi_{21}$	$\pi_{22}$	...	$\pi_{2K}$
.	.	.		.
.	.	.	...	.
.	.	.		.
$\eta_K$	$\pi_{K1}$	$\pi_{K2}$	...	$\pi_{KK}$

consisting of large numbers will have larger eigenvalues, in total, than data matrices with small numbers. To remove the effect of scaling BKW, whose collinearity diagnostic procedure we recommend, suggest scaling the columns of  $X$  to unit length. Define  $s_i = (\sum_{t=1}^T x_{it}^2)^{1/2}$ , and let  $S = \text{diag}(s_1, s_2, \dots, s_K)$ . Then the scaled  $X$  matrix is  $XS^{-1}$ . This scaling is only for the purpose of diagnosing collinearity, not for model estimation or interpretation.

To diagnose collinearity, examine the proportion of the variance of each least squares coefficient contributed by each individual eigenvalue. Define  $\phi_{jk} = \frac{c_{jk}^2}{\lambda_k}$ , and let  $\phi_j$  be the variance of  $b_j$ , apart from  $\sigma^2$ ,  $\phi_j = \left( \frac{c_{j1}^2}{\lambda_1} + \frac{c_{j2}^2}{\lambda_2} + \dots + \frac{c_{jk}^2}{\lambda_k} \right)$ . Then, the proportion of the variance of  $b_j$  associated with the  $k$ th eigenvalue  $\lambda_k$  is  $\pi_{kj} = \frac{\phi_{jk}}{\phi_j}$ , which appears in the  $k$ th row and the  $j$ th column of Table 12.1. The columns of the table correspond to the variances of individual least squares coefficients, and the sum of each column, because it contains the proportions  $\pi_{kj}$ , is one. The rows of this matrix correspond to the different eigenvalues, which have been scaled in a certain way. The “condition index” is the square root of the ratio of  $\lambda_1$ , the largest eigenvalue, to the  $k$ th largest,  $\lambda_k$ , that is,  $\eta_k = \left( \frac{\lambda_1}{\lambda_k} \right)^{1/2}$ . The condition indices are ordered in magnitude, with  $\eta_1 = 1$  and  $\eta_K$  being the largest, since its denominator is  $\lambda_K$ , the smallest eigenvalue. The largest condition index is often called “the condition number of  $X$ ” and denoted as  $\eta_K = \left( \frac{\lambda_1}{\lambda_K} \right)^{1/2} = \kappa$ .

Table 12.1 summarizes much of what we can learn about collinearity in data. BKW carried out extensive simulations to determine how large condition indexes affect the variances of the least squares estimators. Their diagnostic procedures, also summarized in Belsley (1991, ch. 5), are these:

**STEP 1**

Begin by identifying large condition indices. A small eigenvalue and a near exact linear dependency among the columns of  $X$  is associated with each large condition index. BKW's experiments lead them to the general guidelines that indices in the range 0–10 indicate weak near dependencies, 10–30 indicate moderately strong near dependencies, 30–100 a strong near dependency, and indices in excess of 100 are very strong. Thus when examining condition indexes values of 30 and higher should immediately attract attention.

**STEP 2 (IF THERE IS A SINGLE LARGE CONDITION INDEX)**

Examine the variance-decomposition proportions. If there is a single large condition index, indicating a single near dependency associated with one small eigenvalue, collinearity adversely affects estimation when *two or more* coefficients each have 50 percent or more of their variance associated with the large condition index, in the last row of Table 12.1. The variables involved in the near dependency have coefficients with large variance proportions.

**STEP 2 (IF THERE ARE TWO OR MORE LARGE CONDITION INDEXES OF RELATIVELY EQUAL MAGNITUDE)**

If there are  $J \geq 2$  large and roughly equal condition indexes, then  $X'X$  has  $J$  eigenvalues that are near zero and  $J$  near exact linear dependencies among the columns of  $X$  exist. Since the  $J$  corresponding eigenvectors span the space containing the coefficients of the true linear dependence, the "50 percent rule" for identifying the variables involved in the near dependencies must be modified.

If there are two (or more) small eigenvalues, then we have two (or more) near exact linear relations, such as  $Xc_i \approx 0$  and  $Xc_j \approx 0$ . These two relationships do not, necessarily, indicate the form of the linear dependencies, since  $X(a_1c_i + a_2c_j) \approx 0$  as well. In this case the two vectors of constants  $c_i$  and  $c_j$  define a two-dimensional vector space in which the two near exact linear dependencies exist. While we may not be able to identify the individual relationships among the explanatory variables that are causing the collinearity, we can identify the variables that appear in the two (or more) relations.

Thus variance proportions in a single row *do not* identify specific linear dependencies, as they did when there was but one large condition number. In this case, *sum* the variance proportions across the  $J$  large condition number rows in Table 12.1. The variables involved in the (set of) near linear dependencies are identified by summed coefficient variance proportions of greater than 50 percent.

**STEP 2 (IF THERE ARE  $J \geq 2$  LARGE CONDITION INDEXES, WITH ONE EXTREMELY LARGE)**

An extremely large condition index, arising from a very small eigenvalue, can "mask" the variables involved in other near exact linear dependencies. For example, if one condition index is 500 and another is 50, then there are two near exact linear dependencies among the columns of  $X$ . However, the variance decompositions associated with the condition index of 50 may not indicate that there are two or more variables involved in a relationship. Identify the variables

involved in the set of near linear dependencies by summing the coefficient variance proportions in the last  $J$  rows of Table 12.1, and locating the sums greater than 50 percent.

### STEP 3

Perhaps the most important step in the diagnostic process is determining which coefficients *are not* affected by collinearity. If there is a single large condition index, coefficients with variance proportions less than 50 percent in the last row of Table 12.1 are not adversely affected by the collinear relationship in the data. If there are  $J \geq 2$  large condition indexes, then sum the last  $J$  rows of variance proportions. Coefficients with summed variance proportions of less than 50 percent are not adversely affected by the collinear relationships. If the parameters of interest have coefficients unaffected by collinearity, then small eigenvalues and large condition numbers *are not a problem*.

### STEP 4

If key parameter estimates are adversely affected by collinearity, further diagnostic steps may be taken. If there is a single large condition index the variance proportions identify the variables involved in the near dependency. If there are multiple large condition indexes, auxiliary regressions may be used to further study the nature of the relationships between the columns of  $X$ . In these regressions one variable in a near dependency is regressed upon the other variables in the identified set. The usual  $t$ -statistics may be used as diagnostic tools to determine which variables are involved in specific linear dependencies. See Belsley (1991, p. 144) for suggestions. Unfortunately, these auxiliary regressions may also be confounded by collinearity, and thus they may not be informative.

## 4 OTHER DIAGNOSTIC ISSUES AND TOOLS

There are a number of issues related to the diagnosis of collinearity, and other diagnostic tools. In this section we summarize some of these.

### 4.1 The centering issue

Eigenvalue magnitudes are affected by the scale of the data. There is wide agreement that the  $X$  matrix should be scaled before analyzing collinearity, and scaling the columns of  $X$  to unit length is standard. A much more hotly debated issue, chronicled in Belsley (1984), is whether the data should be *centered*, and then scaled, prior to collinearity diagnosis. If the data are centered, by subtracting the mean, the origin is translated so that the regression, in terms of the centered data, has a  $y$ -intercept of zero. The least squares estimates of slopes are unaffected by centering. The least squares estimate of the intercept itself can be obtained after the slopes are estimated, as  $b_1 = \bar{y} - b_2\bar{x}_2 - \dots - b_K\bar{x}_K$ . So nothing is really gained, or lost, by centering. Let  $X_c$  be the  $X$  matrix after centering, scaling to unit length, and deleting the first column of zeros. Then  $X'_c X_c = R_c$  is the regressor correlation matrix.

The “pro-centering” point of view is summarized by Stewart (1987, p. 75), who suggests that the constant term is rarely of interest and its inclusion “masks” the real variables. “Centering simply shows the variable for what it is.” The “anti-centering” viewpoint is based on several points. First, as a practical matter, centering lowers the condition number of the data (Belsley, 1991, p. 189), usually by a large amount, and thus makes it an unreliable diagnostic. Second, and more importantly, centering the data makes it impossible to identify collinearities caused by linear combinations of explanatory variables which exhibit little variation. If a variable, or a linear combination of variables, exhibits little variation, then it will be “collinear” with the constant term, the column of 1s in the first column of  $X$ . That is, suppose  $a_2x_{t2} + a_3x_{t3} + \dots + a_Kx_{tK} \approx a$ , where  $a$  is a constant. If  $x_{t1} = 1$ , then  $a_2x_{t2} + a_3x_{t3} + \dots + a_Kx_{tK} - ax_{t1} \approx 0$ .

The pro-centering view is that the constant term is not interesting, and therefore such linear dependencies are not important. The anti-centering group notes that such a collinear relationship affects not only the intercept, but also affects the coefficients of the variables in the collinear relationship, whether the intercept is of theoretical importance or not.

We fall squarely into the anti-centering camp. The data should be scaled to unit length, but not centered, prior to examining collinearity diagnostics. The interested reader should see Belsley (1984), including comments, for the complete, lively, debate.

## 4.2 Other diagnostics

The expression for the variance of the least squares estimator in equation (12.2), for the regression model with two explanatory variables, can be extended to the multiple regression context, for all coefficients except the intercept, as

$$\text{var}(b_j) = \frac{\sigma^2}{\sum_{t=1}^T (x_{tj} - \bar{x}_j)^2} \cdot \frac{1}{1 - R_j^2}, \quad (12.9)$$

where  $R_j^2$  is the “ $R^2$ ” goodness-of-fit measure from the auxiliary regression of  $x_{tj}$  on all other regressors. The second factor in (12.9) is called the *variance-inflation factor* (VIF), as it shows the effect of linear associations among the explanatory variables upon the variances of the least squares estimators, as measured by  $R_j^2$ . Stewart (1987) proposes collinearity indexes that are the square roots of the VIFs. Fox and Monette (1992) generalize VIFs to measure variance-inflation in a subset of coefficients. Auxiliary regressions and VIFs have the same strengths and weaknesses as collinearity diagnostics. Their strength is their intuitive appeal. If  $R_j^2 \rightarrow 1$ , then  $x_{tj}$  is in some collinear relationship. The weaknesses are, apart from the pro-centering–anti-centering debate, that we have no measure of how close  $R_j^2$  must be to 1 to imply a collinearity problem, and these measures cannot determine the number of near linear dependencies in the data (Belsley, 1991, p. 30). Kennedy (1998, p. 90) suggests the rule of thumb that a  $\text{VIF} > 10$ , for scaled data,

indicates severe collinearity. The same critiques apply to the strategy of looking at the matrix  $R_c$  of correlations among the regressors as a diagnostic tool.

Another diagnostic that is often mentioned, though recognized as deficient, is the determinant of  $X'X$  (or  $X'_c X_c = R_c$ ). One or more near exact collinearities among the columns of  $X$  drive this determinant to zero. The problems with measuring collinearity this way include deciding how small the determinant must be before collinearity is judged a problem, the fact that using this measure we determine neither the number nor form of the collinearities, and the ever-present centering debate. Soofi (1990) offers an information theory-based approach for diagnosing collinearity in which the log determinant plays an important role. Unfortunately, his measures reduce the diagnosis of collinearity to the examination of a single index, which has the same flaws as the determinant.

### 4.3 Collinearity-influential observations

One or two observations can make a world of difference in a data set, substantially improving, or worsening, the collinearity in the data. Can we find these “collinearity-influential” observations? If we do, what, if anything, do we do with them? The answer to the former question is “Maybe.” The answer to the latter question is “It depends.”

Influential-data diagnostics are designed to find “unusual” observations in a data set and evaluate their impact upon regression analysis. Standard references include BKW, Cook and Weisberg (1982) and Chatterjee and Hadi (1988). Mason and Gunst (1985) illustrate the effect that individual observations can have on data collinearity. Belsley (1991, pp. 245–70) reviews and illustrates diagnostics that may be useful for detecting collinearity-*inducing* observations, whose inclusion worsens collinearity in the data, and collinearity-*breaking* observations, whose inclusion lessens collinearity in the data. If  $\kappa = (\lambda_1/\lambda_K)^{1/2}$  is the condition number of the  $X$  matrix, and if  $\kappa_{(i)}$  denotes the condition index of the matrix  $X$  with row  $i$  (or set of rows) deleted, then one measure of the effect of an observation upon collinearity is

$$\delta_{(i)} = \frac{\kappa_{(i)} - \kappa}{\kappa}. \quad (12.10)$$

A large negative value of  $\delta_{(i)}$  indicates a collinearity-inducing observation, while a positive value indicates a collinearity-breaking observation. Chatterjee and Hadi (1988), Hadi and Wells (1990) and Sengupta and Bhimasankaram (1997) study this measure and variations of it. See Belsley (1991, p. 251) for examples.

The question is what to do when collinearity-influential observations are found? As with all influential, or unusual, observations we must first determine if they are correct. If they are incorrect, then they should be corrected. If they are correct, then the observations deserve close examination, in an effort to determine why they are unusual, and exactly what effect their inclusion, or exclusion, has upon the signs, magnitudes, and significance of the coefficient estimates.

A second consideration concerns estimator choice. When collinearity is present, and deemed harmful to the least squares estimator, alternative estimators designed to improve the precision of estimation are sometimes suggested. We will review some of these estimators in Section 5. If collinearity is induced by a few influential observations, then a robust estimator may be an alternative to consider.

#### 4.4 Detecting harmful collinearity

We can determine the number of collinear relations, their severity, and the variables involved using the diagnostics in Section 3. This does not end the diagnostic process, because we must still determine if the collinearity present is actually harmful to our regression. Whether the collinearity matters depends on the magnitude of the regression parameters. The parameters matter in two regards. First, from equations (12.2) and (12.9), it is clear that a small value of the error variance,  $\sigma^2$ , can offset the effects of high correlation between the regressors or low regressor variability.

Second, the magnitudes of the  $\beta_k$  matter. If the variance of  $b_k$  is  $\sigma_{b_k}^2$ , then  $100(1 - \alpha)\%$  interval estimator for  $\beta_k$  is  $b_k \pm t_c \hat{\sigma}_{b_k}$ , where  $t_c$  is a critical value from the  $t$ -distribution. Suppose we diagnose severe collinearity affecting (and inflating) the variance of  $b_k$  and compute  $t_c \hat{\sigma}_{b_k} = 3$ . Is collinearity harmful when  $\beta_k = 1$ ? What if  $\beta_k = 1000$ ? If you answered "yes" to the first question, but "no" to the second, you are saying, and rightly so, that the magnitude of the parameter  $\beta_k$  also matters when determining if collinearity is harmful or not.

Belsley (1982) addresses these issues by developing tests for adequate "signal-to-noise," abbreviated  $s/n$ , in the regression model and data. For a single parameter Belsley defines an  $s/n$  parameter,

$$\tau = \frac{\beta_k}{\sigma_{b_k}}. \quad (12.11)$$

If  $\tau$  is small, then the error variance  $\sigma^2$  is not small enough, and/or  $\beta_k$  is not large enough, to offset the effects of collinearity and/or lack of regressor variability. Belsley proposes to test the hypothesis that  $|\tau| > \tau_*$ , where  $\tau_*$  is an adequate magnitude. For details of this, and a more general multiparameter test, see Belsley (1982).

In the end, Belsley (1982, p. 225) proposes that investigators (i) examine collinearity using the diagnostics described in Section 3, and (ii) carry out the test for adequate  $s/n$ . The conclusions one can draw are summarized in Table 12.2.

The four possible outcomes are these: (I) negligible collinearity and adequate  $s/n$ ; (II) collinearity present, but not harmful, since adequate  $s/n$  is present; (III) negligible collinearity, but inadequate  $s/n$  present, caused by lack of regressor variation; (IV) harmful collinearity, the joint occurrence of severe collinearity and inadequate  $s/n$ . In the next section we address what remedies are available in cases III and IV.

**Table 12.2** Harmful collinearity decision matrix

		Collinearity present?	
		no	yes
Inadequate	no	I	II
	yes	III	IV

## 5 WHAT TO Do?

In this section we address the question of what to do when harmful collinearity is found with respect to the important parameters in a regression model. This section is like a minefield. There is danger all around, and but two safe paths. We will identify the safe paths, though these are the roads less traveled, and we will mention some potentially dangerous and self-defeating methods for dealing with collinearity.

Since the collinearity problem is actually one of insufficient independent variation in the data, the first and most desirable solution is to obtain more and *better* data. If the new data possess the same dependencies found in the original sample, then they are unlikely to be of much help. On the other hand, if new data can be found in, as Belsley (1991, p. 297) calls it, “novel or underrepresented portions” of the sample space, then the new observations may mitigate the ill-effects of collinearity. Unfortunately, nonexperimental empirical researchers seldom have much if any control over the design of the data generation process, and hence this advice is, for the most part, empty of practical content. Should the occasion arise, however, Silvey (1969, p. 545) discusses the optimal choice, for the purpose of improving the estimation of a linear combination of the parameters  $c'\beta$ , of the values of the explanatory variables in a new observation. This classic treatment has been extended by Sengupta (1995).

Blanchard (1987, p. 449) says, “Only use of more economic theory in the form of additional restrictions may help alleviate the multicollinearity problem.” We agree that the only “cure” for collinearity, apart from additional data, is additional information about regression parameters. However, restrictions can come from economic theory or previous empirical research, which we collectively term *nonsample* sources. If harmful collinearity is present, we are admitting that the sample data are inadequate for the purpose of precisely estimating some or all of the regression parameters. Thus the second safe strategy for mitigating the effects of collinearity is to introduce *good* nonsample information about the parameters into the estimation process. When nonsample information is added during the estimation process, estimator variances are reduced, which is exactly what we want to happen in the presence of collinearity (and indeed, all the time.) The downside to using nonsample information is that estimator bias is introduced.

It is possible that small amounts of bias are acceptable in return for significant increases in precision. The most commonly used measure of the bias/precision tradeoff is mean-square-error (MSE),

$$\text{MSE}(\hat{\beta}) = E[(\hat{\beta} - \beta)'(\hat{\beta} - \beta)] = \sum_k \text{var}(\hat{\beta}_k) + \sum_k [E(\hat{\beta}_k) - \beta_k]^2, \quad (12.12)$$

which combines estimator variances with squared biases. This measure is also known as estimator “risk” in decision theory literature (Judge *et al.*, 1988, pp. 807–12).

Our general objective is to introduce nonsample information that improves upon the MSE of the OLS estimator. This is much easier said than done, and there is a huge literature devoted to methods for obtaining MSE improvement. See Judge and Bock (1978, 1983). Suffice it to say that MSE improvements occur only when the nonsample information we employ is good. How do we know if the information we introduce is good enough? We do *not* know, and can *never* know, if our nonsample information is good enough to ensure an MSE reduction, since that would require us to know the true parameter values. This is our conundrum. Below we briefly survey alternative methods for introducing non-sample information, all of which can be successful in reducing MSE, if the nonsample information is good enough.

## 5.1 Methods for introducing exact nonsample information

The most familiar method for introducing nonsample information into a regression model is to use restricted least squares (RLS). The restricted least squares estimator, which we denote as  $b^*$ , is obtained by minimizing the sum of squared errors subject to  $J$  exact linear parameter restrictions,  $R\beta = r$ . Examples of linear restrictions are  $\beta_2 + \beta_3 = 1$  and  $\beta_5 = \beta_6$ . The variances of the RLS estimator are smaller than those of the OLS estimator, but  $b^*$  is biased unless the parameter restrictions imposed are exactly true. As noted above, the restrictions do not have to be exactly true for RLS to be better than OLS under a criterion such as MSE, which trades off bias against variance reduction. A question that naturally arises is why such restrictions, if they exist, are not imposed at the outset. A classic example of RLS used to mitigate collinearity is the Almon (1965) polynomial distributed lag. To determine if the imposed restrictions improve the conditioning of  $X$ , substitute the restrictions into the model, via the method outlined in Fomby *et al.* (1984, p. 85), and apply the collinearity diagnostics.

Some familiar “tricks” employed in the presence of collinearity are, in fact, RLS estimators. The most common, and often ill-advised, strategy is to drop a variable if it is involved in a collinear relationship and its estimated coefficient is statistically insignificant. Dropping a variable,  $x_k$ , is achieved by imposing the linear constraint that  $\beta_k = 0$ . Unless  $\beta_k = 0$ , dropping  $x_k$  from the model generally biases *all* coefficient estimators. Similarly, two highly correlated variables are

often replaced by their sum, say  $z = x_k + x_m$ . How is this achieved? By imposing the restriction that  $\beta_k = \beta_m$ . Once again, if this constraint is not correct, reductions in variance are obtained at the expense of biasing estimation of all regression coefficients. Kennedy (1983) detects the failure of a similar collinearity trick used by Buck and Hakim (1981) in the context of estimating and testing differences in parameters between two groups of observations.

Economists recognize the bias/precision tradeoff and wish to impose constraints that are “good.” It is standard practice to check potential constraints against the data by testing them as if they were hypotheses. Should we drop  $x_k$ ? Test the hypothesis that  $\beta_k = 0$ . Should we sum  $x_k$  and  $x_m$ ? Test the hypothesis  $\beta_k = \beta_m$ . Belsley (1991, p. 212) suggests formally testing for MSE improvement. The MSE test amounts to comparing the usual F-statistic for a joint hypothesis to critical values tabled in Toro-Vizcarrondo and Wallace (1968). Following the tests a decision is made to abandon restrictions that are rejected, and use restrictions that are not rejected. Such a strategy prevents egregious errors, but actually defines a new, “pre-test” estimation rule. This rule, which chooses either the OLS or RLS estimator based on the outcome of a test, does not have desirable statistical properties, but it seems unlikely that this practice will be abandoned. See Judge *et al.* (1985, ch. 3).

Another alternative is the Stein-rule estimator, which is a “smart” weighted average of the OLS and RLS estimators, weighting the RLS estimator more when the restrictions are compatible with the data, and weighting the OLS estimator more when the restrictions are not compatible with the data. The Stein-rule usually provides an MSE gain over OLS, but it is not guaranteed to ameliorate the specific problems caused by collinearity. See Judge *et al.* (1985, ch. 22).

## 5.2 Methods for introducing inexact nonsample information

Economists usually bring general information about parameters to the estimation problem, but it is not like the exact restrictions discussed in the previous section. For example, we may know the signs of marginal effects, which translate into inequality restrictions on parameters. Or we may think that a parameter falls in the unit interval, and that there is a good chance it falls between 0.25 and 0.75. That is, we are able to suggest signs of parameters, and even perhaps ranges of reasonable values. While such information has long been available, it has been difficult to use in applications. Perhaps the biggest breakthrough in recent years has been the development of methods and the distribution of software that makes it feasible to estimate linear (and nonlinear) models subject to inequality restrictions, and to implement Bayesian statistical methods.

The theory of inequality restricted least squares was developed some time ago. See Judge and Yancey (1986). However, the numerical problems of minimizing the sum of squared regression errors or maximizing a likelihood function subject to general inequality restrictions are substantial. Recently major software packages (SAS, GAUSS, GAMS) have made algorithms for such constrained

optimization much more accessible. With inequality restrictions, such as  $\beta_k > 0$ , MSE gains require only that the direction of the inequality be correct.

The Bayesian paradigm is an alternative mode of thought. See Zellner (1971). In it we represent our uncertainty about parameter values using probability distributions. Inexact nonsample information is specified up front in the Bayesian world, by specifying a "prior" probability distribution for each parameter (in general a joint prior). The prior density can be centered over likely values. It can be a truncated distribution, putting zero prior probability on parameter values we rule out on theoretical grounds, and so on. When prior beliefs are combined with data a multivariate probability distribution of the parameters is generated, called the posterior distribution, which summarizes all available information about the parameters.

As noted in Judge *et al.* (1985, p. 908), Bayesians have no special problem dealing with the singularity or near-singularity of  $X'X$ . Their approach to the collinearity problem is to combine the prior densities on the parameters,  $\beta$  with the sample information contained in the data to form a posterior density (see Zellner, 1971, pp. 75–81). The problem for Bayesians, as noted by Leamer (1978), is that when data are collinear the posterior distribution becomes very sensitive to changes in the prior. Small changes in the prior density result in large changes in the posterior, which complicates the use and analysis of the results in much the same way that collinearity makes inference imprecise in the classical theory of inference.

Bayesian theory is elegant, and logically consistent, but it has been a nightmare in practice. Suppose  $g(\beta)$  is the multivariate posterior distribution for the vector of regression parameters  $\beta$ . The problem is how to extract the information about a single parameter of interest, say  $\beta_k$ . The brute force method is to obtain the posterior density for  $\beta_k$  by integrating all the other parameters out of  $g(\beta)$ . When the posterior distribution  $g(\beta)$  is complicated, as it usually is, this integration is a challenging problem.

The Bayesian miracle has been the development of computationally intensive, but logically simple, procedures for deriving the posterior densities for individual parameters. These procedures include the Gibbs sampler, the Metropolis and Metropolis–Hastings algorithms (Dorfman, 1997). These developments will soon make Bayesian analysis feasible in many economic applications.

In passing we note that non-Bayesians have tried to achieve the incorporation of similar information by making the exact restrictions in Section 5.1 inexact (Theil and Goldberger, 1961). This is achieved by adding a random disturbance  $v \sim (0, \Omega)$  to exact restrictions, to obtain  $r = R\beta + v$ . This additional information is combined with the linear model as

$$\begin{bmatrix} y \\ r \end{bmatrix} = \begin{bmatrix} X \\ R \end{bmatrix} \beta + \begin{bmatrix} e \\ v \end{bmatrix}. \quad (12.13)$$

The resulting model is estimated by generalized least squares, which is called "mixed estimation" in this context. The difficulty, of course, apart from specifying

the constraints, is the specification of the covariance matrix  $\Omega$ , reflecting parameter uncertainty.

Another estimation methodology has been introduced recently, based upon the maximum entropy principle (Golan, Judge, and Miller, 1996). This estimation method, instead of maximizing the likelihood function, or minimizing the sum of squared errors, maximizes the entropy function, subject to data and logical constraints. The method of maximum entropy is “nonparametric” in the sense that no specific probability distribution for the errors need be assumed. Like the Bayesian methodology, maximum entropy estimation requires the incorporation of prior information about the regression parameters at the outset. Golan, Judge and Miller find that the maximum entropy estimator, which like the Stein-rule is a shrinkage estimator, performs well in the presence of collinearity.

### 5.3 Estimation methods designed specifically for collinear data

A number of estimation methods have been developed to improve upon the least squares estimator when collinearity is present. We will briefly discuss two, ridge regression and principal components regression, if only to warn readers about their use.

The ridge family of estimators is

$$\hat{b}(k) = (X'X + kI)^{-1}X'y, \quad (12.14)$$

where  $k$  is a suitably chosen constant. When  $k = 0$  then the ridge estimator is just the OLS estimator of  $\beta$ . For nonstochastic values of  $k > 0$  the ridge estimator is biased, but has smaller variances than the least squares estimator. It achieves the variance reduction by “shrinking” the least squares estimates towards zero. That is, the (Euclidean) length of the ridge estimator is smaller than that of the least squares estimator. Choosing  $k$  is important since some values result in reductions of overall mean square error and others do not. Unfortunately, picking a value of  $k$  that assures reduction in overall MSE requires knowledge of  $\beta$  and  $\sigma^2$ , the original object of the regression analysis. Numerous methods for selecting  $k$  based on the data have been proposed, but choosing  $k$  using data makes  $k$  random, and completely alters the statistical properties of the resulting “adaptive” ridge estimator (Hoerl, Kennard, and Baldwin, 1975; Lawless and Wang, 1976). Finite sample inference using the ridge estimator is hindered by dependence of its sampling distribution on unknown parameters. There is a huge statistics literature on the ridge estimator, but the fundamental problems remain and we cannot recommend this estimator.

Principal components regression (Fomby *et al.*, 1984, pp. 298–300) is based upon eigenanalysis. Recall that the  $(K \times K)$  matrix  $C$ , whose columns are the eigenvectors of  $X'X$ , is an orthogonal matrix, such that  $C'C = CC' = I$ . The  $T \times K$  matrix  $Z = XC$  is called the matrix of principal components of  $X$ . The  $i$ th column of  $Z$ ,  $z_i = Xc_i$ , is called the  $i$ th principal component. From equation (12.5)  $z_i$  has the property that  $z_i'z_i = \lambda_i$ .

The “principal components” form of the linear regression model is

$$y = X\beta + e = XCC'\beta + e = Z\theta + e, \quad (12.15)$$

where  $Z = XC$  and  $\theta = C'\beta$ . The new set of explanatory variables  $Z$  are linear transformations of the original variables, and have the property that  $Z'Z = \Lambda = \text{diag}(\lambda_1, \lambda_2, \dots, \lambda_K)$ , where the  $\lambda_k$  are the ordered (in decreasing magnitude) eigenvalues of  $X'X$ . If we apply least squares to the transformed model we obtain  $\hat{\theta} = (Z'Z)^{-1}Z'y$ , which has covariance matrix  $\text{cov}(\hat{\theta}) = \sigma^2(Z'Z)^{-1} = \sigma^2\Lambda^{-1}$ , so that  $\text{var}(\hat{\theta}_k) = \sigma^2/\lambda_k$ . If the data are collinear then one or more of the eigenvalues will be near zero. If  $\lambda_K \approx 0$  then the eigenvector  $z_K \approx 0$ , and consequently it is difficult to estimate  $\theta_K$  precisely, which is reflected in the large variance of its estimator,  $\text{var}(\hat{\theta}_K) = \sigma^2/\lambda_K$ . Principal components regression deletes from equation (12.15) the  $z_k$  associated with small eigenvalues (usually based upon tests of significance, or some other model selection criterion, such as AIC or BIC). Partition the transformed model as  $y = Z\theta + e = Z_1\theta_1 + Z_2\theta_2 + e$ . Dropping  $Z_2$ , which contains the  $z_k$  to be deleted, and applying OLS yields  $\hat{\theta}_1 = (Z'_1Z_1)^{-1}Z'_1y$ . The principal components estimator of  $\beta$  is obtained by applying an inverse transformation

$$b_{pc} = C\theta = [C_1 \quad C_2] \begin{bmatrix} \hat{\theta}_1 \\ \theta_2 = 0 \end{bmatrix} = C_1\hat{\theta}_1. \quad (12.16)$$

The properties of this estimator follow directly from the observation that it is equivalent to the RLS estimator of  $\beta$  obtained by imposing the constraints  $C'_2\beta = 0$ . Thus the principal components estimator  $b_{pc}$  is biased, but has smaller variances than the OLS estimator. The data based constraints  $C'_2\beta = 0$  generally have no economic content, and are likely to induce substantial bias. One positive use of principal components regression is as a benchmark. The  $J$  constraints  $C'_2\beta = 0$  have the property that they provide the maximum variance reduction of any set of  $J$  linear constraints (Fomby, Hill, and Johnson, 1978). Thus researchers can measure the potential for variance reduction using linear constraints.

## 5.4 Artificial orthogonalization

Consider a regression model containing two explanatory variables,

$$y = \beta_1 + \beta_2x_2 + \beta_3x_3 + e, \quad (12.17)$$

where the regressors  $x_2$  and  $x_3$  are highly correlated. To “purge” the model of collinearity regress  $x_3$  on  $x_2$ , and compute the residuals  $x_3^* = x_3 - \hat{x}_3$ . It is argued that  $x_3^*$  contains the information in the variable  $x_3$  after the effects of collinearity are removed. Because least squares residuals are orthogonal to the regressors,  $x_3^*$  and  $x_2$  are uncorrelated, and thus collinearity is eliminated! Substituting  $x_3^*$  into the model we obtain,

$$y = \beta_1 + \beta_2x_2 + \beta_3^*x_3^* + e^*, \quad (12.18)$$

which is then estimated by OLS. Buse (1994) shows that the least squares estimates of  $\beta_1$  and  $\beta_3$  are unaffected by this substitution, as are their standard errors and  $t$ -statistics, and the residuals from (12.18) are identical to those from (12.17), hence statistics such as  $R^2$ , the Durbin–Watson  $d$ , and  $\hat{\sigma}^2$  are unaffected by the substitution. But what about the estimator of  $\beta_2$ ? Kennedy (1982) first points out the problems with this procedure, and Buse (1994) works out the details. Buse shows that the estimator of  $\beta_2$  from (12.18) is biased. Furthermore, instead of gaining a variance reduction in return for this bias, Buse shows that the variance of  $\hat{\beta}_2^*$  can be larger than the OLS variance of  $b_3$ , and he gives several examples. Thus artificial orthogonalization is not a cure for collinearity.

## 6 NONLINEAR MODELS

Assessing the severity and consequences of collinearity in nonlinear models is more complicated than in linear models. To illustrate, we first discuss its detection in nonlinear regression, and then in the context of maximum likelihood estimation.

### 6.1 The nonlinear regression model

Consider the nonlinear regression model

$$y = f(X, \beta) + e, \quad (12.19)$$

where  $e \sim (0, \sigma^2 I)$  and  $f(X, \beta)$  is some nonlinear function that relates the independent variables and parameters to form the systematic portion of the model. The nonlinear least squares estimator chooses  $\hat{\beta}$  to minimize

$$S(\beta) = [y - f(X, \beta)]' [y - f(X, \beta)].$$

The first order conditions yield the least squares solution,

$$Z(\beta)' [y - f(X, \beta)] = 0, \quad (12.20)$$

where the  $T \times K$  matrix  $Z(\beta) = \partial f(X, \beta) / \partial \beta'$ . Since equation (12.20) is nonlinear, the least squares estimates  $\hat{\beta}$  must be obtained using numerical methods.

A useful algorithm for finding the minimum of  $S(\beta)$  is the Gauss–Newton. The Gauss–Newton algorithm is based on a first order Taylor's series expansion of  $f(X, \beta)$  around a starting value  $\beta_1$ . From that we obtain the linearized model

$$\bar{y}(\beta_1) = Z(\beta_1)\beta + e, \quad (12.21)$$

where  $\bar{y}(\beta_1) = y - f(X, \beta_1) + Z(\beta_1)\beta_1$ . In (12.21) the dependent variable and the “regressors”  $Z(\beta_1)$  are completely determined given  $\beta_1$ . The next round estimate is obtained by applying least squares to (12.21), and in general the iterations are

$$\beta_{n+1} = [Z(\beta_n)' Z(\beta_n)]^{-1} Z(\beta_n)' \bar{y}(\beta_n). \quad (12.22)$$

The iterations continue until a convergence criterion is met, perhaps that  $\beta_n \approx \beta_{n+1} = \hat{\beta}$ , which defines the nonlinear least squares estimates of  $\beta$ . Given that  $f(X, \beta)$  is a nice function, then, asymptotically,

$$\hat{\beta} \sim N(\beta, \sigma^2[Z(\beta)'Z(\beta)]^{-1}) \quad (12.23)$$

and the asymptotic covariance matrix of  $\hat{\beta}$  is estimated as

$$\text{acov}(\hat{\beta}) = \hat{\sigma}^2[Z(\hat{\beta})'Z(\hat{\beta})]^{-1}, \quad (12.24)$$

where  $\hat{\sigma}^2 = S(\hat{\beta})/(T - K)$ . Equations (12.21)–(12.23) show that  $Z(\beta)$  in nonlinear regression plays the role of  $X$  in the linear regression model. Consequently, it is the columns of  $Z(\beta)$ , which we examine via the BKW diagnostics in Section 3, that we must consider when diagnosing collinearity in the nonlinear regression model.

## 6.2 Collinearity in nonlinear regression models

When examining  $Z(\beta)$  for collinearity a problem arises. That is,  $Z(\beta)$  depends not only on the data matrix  $X$  but also on the parameter values  $\beta$ . Thus collinearity changes from point to point in the parameter space, and the degree of collinearity among the columns of the data matrix  $X$  may or may not correspond to collinearity in  $Z(\beta)$ . This problem affects nonlinear regression in two ways. First, the Gauss–Newton algorithm itself may be affected by collinearity in  $Z(\beta)$ , because at each iteration the cross-product matrix  $Z(\beta_n)'Z(\beta_n)$  must be inverted. If the columns of  $Z(\beta_n)$  are highly collinear then the cross-product matrix may be difficult to invert, at which point the algorithm may fail. Second, the estimated asymptotic covariance matrix of the nonlinear least squares estimator, equation (12.24), contains the cross-product matrix  $Z(\hat{\beta})'Z(\hat{\beta})$ , and thus the estimated variances and covariances suffer from the usual consequences of collinearity, depending on the relationships between the columns of  $Z(\hat{\beta})$ . Computer software packages, such as SAS 6.12, compute and report the BKW diagnostics for the matrix  $Z(\beta_n)'Z(\beta_n)$  when the Gauss–Newton algorithm fails, so that the user may try to determine the source of the very nearly exact collinearity that leads to the failure, and it also computes the conditioning diagnostics for  $Z(\hat{\beta})'Z(\hat{\beta})$ , upon convergence of the algorithm. There remains, of course, the collinearity among the columns of  $Z(\beta)$ , which enters the true asymptotic covariance matrix of the nonlinear least squares estimator in equation (12.23), and which remains unknown.

What do we do if collinearity, or ill-conditioning, of  $Z(\beta_n)$  causes the Gauss–Newton algorithm to fail to converge? The conditioning of  $Z(\beta_n)$  can be affected by scaling the data. One common problem is that the columns of  $Z(\beta_n)$  have greatly different magnitudes. Recall that  $Z(\beta_n)$  contains the first derivatives of the function evaluated at  $\beta_n$ , so magnitudes in  $Z(\beta_n)$  are slopes of the functions  $f(X, \beta)$ . If these are greatly different then the function is steep in some directions and shallow in others. Such an irregular surface is difficult to work with. By rescaling the columns of  $X$ , it is sometimes possible to more nearly equalize the

columns of  $Z(\beta_n)$ , meaning that the function  $f(X, \beta)$  itself has been smoothed. This is usually advantageous.

When computing the BKW diagnostics the columns of  $Z(\beta_n)$  should be scaled to unit length. If, after the data are scaled, the condition number of  $Z(\beta_n)$  is still large, closer examination of the function, data, and parameter values are required. To illustrate, Greene (1997, p. 456) and Davidson and MacKinnon (1993, pp. 181–6) give the example of the nonlinear consumption function  $C = \alpha + \beta Y^\gamma + e$ , where  $C$  is consumption and  $Y$  is aggregate income. For this model the  $t$ th row of  $Z(\beta)$  is  $[1 \quad Y^\gamma \quad \beta Y^\gamma \ln Y]$ . What happens if during the Gauss–Newton iterations the value of  $\gamma$  approached zero? The second column approaches 1, and is collinear with the first column. What happens if  $\beta \rightarrow 0$ ? Then the third column approaches 0, making  $Z(\beta)$  ill-conditioned. In these cases collinearity is avoided by avoiding these parameter values, perhaps by selecting starting values wisely. For a numerical example see Greene (1997, pp. 456–8). There are alternative algorithms to use when convergence is a problem in nonlinear least squares regression. It is very useful to be aware of the alternatives offered by your software, as some may perform better than others in any given problem. See Greene (1997, ch. 5).

### 6.3 Collinearity in maximum likelihood estimation

Collinearity in the context of maximum likelihood estimation is similarly diagnosed. Instead of minimizing the sum of squared errors we maximize the loglikelihood function. Standard gradient methods for numerical maximization use first and/or second derivatives. As in the Gauss–Newton algorithm for nonlinear least squares, these methods involve an inversion: the Hessian for the Newton–Raphson, the Information matrix for the method of scoring, and the cross-product matrix of first derivatives for the method of Berndt, Hall, Hall, and Hausman. In these algorithms if the matrix to be inverted becomes singular, or nearly so, estimation fails. In each case we can apply the BKW diagnostics to the matrix we are inverting at each step of the nonlinear optimization, and to the estimate of the asymptotic covariance matrix. The same difficulties arise in diagnosing collinearity here as in nonlinear least squares, only it is worse, because while the condition numbers provide a measure of how ill-conditioned the matrix is, the rows of Table 12.1 no longer provide any information about which variables are involved in collinear relations. Similar remarks hold for collinearity diagnosis in generalized least squares and simultaneous equations models.

Some common maximum likelihood estimators, among others, probit, logit, tobit, Poisson regression, and multiplicative heteroskedasticity, have information matrices of a common form,

$$I(\beta) = X'WX, \tag{12.25}$$

where  $W$  is a  $T \times T$  diagonal weight matrix that often is a function of the unknown parameters,  $\beta$ , and the independent variables.

The class of generalized linear models (McCullagh and Nelder, 1989) contains many of these estimators as special cases, and have information matrices in the

form of equation (12.25), thus collinearity diagnostics for these models are relevant. Weissfeld and Sereika (1991) explore the detection of collinearity in the class of generalized linear models (GLM). Segerstedt and Nyquist (1992) observe that ill-conditioning in these models can be due to collinearity of the variables,  $X$ , the influence of the weights,  $W$ , or both. Weissfeld and Sereika suggest applying the BKW diagnostics to the scaled information matrix. Lee and Weissfeld (1996) do the same for the Cox regression model. Once again, while the variance decompositions can be computed in these instances, their interpretation is not straightforward, since collinearity can be due to the weights,  $W$ .

Lesaffre and Marx (1993) also investigate the problem of ill-conditioning in GLM and take a slightly different approach. Following Mackinnon and Puterman (1989), they suggest that only the columns of  $X$  be standardized to unit length, forming  $X_1$ . Then, conditioning diagnostics are computed on  $X_1' \hat{W} X_1$ , where  $\hat{W}$  is the estimated weight matrix based on the rescaled data. The square root of the ratio of the largest to smallest eigenvalue describes the worst relative precision with which linear combinations of the parameters can be estimated. Thus, this scaling gives a structural interpretation to the conditioning diagnostic. One problem with this scaling is that  $X_1' \hat{W} X_1$  could be ill-conditioned because of the effects of  $\hat{W}$ .

## 7 CLOSING REMARKS

We conclude by pointing out the main lessons of this essay. First, we have tools, the Belsley, Kuh, and Welsch (1980) collinearity diagnostics, which allow us to determine the form and severity of collinearity in the linear regression model. Most importantly, we know which variables are involved in collinear relationships, and which variables are *not* involved in collinear relationships. If the least squares estimator is severely affected by collinearity, but the model's variables of interest are not involved in the collinear relationships, then there is no call for remedial actions. Such a conclusion requires us to think clearly about our models, and to pinpoint key variables.

Since new and better data are rarely available, the only practical approach to mitigating harmful collinearity is the introduction of nonsample information about the parameters, based on prior empirical research or economic theory. However the information is introduced, whether it be via restricted least squares, the Bayesian approach, or maximum entropy estimation, we must endeavor to introduce "good" nonsample information. The difficulty with this statement is that we never *truly* know whether the information we introduce is good enough to reduce estimator mean square error, or not.

The analysis of collinearity in nonlinear models is difficult. Collinearity (ill-conditioning) in asymptotic covariance matrices may arise from collinearity in the matrix of explanatory variables  $X$ , and/or particular parameter values and function values. Identifying the cause of the ill-conditioning may, or may not, be possible, but again the use of good nonsample information would seem the only remedy. In nonlinear models the problem of collinearity spills over into the estimation process, because the iterative algorithms used for numerical optimization

may be sensitive to it. When this occurs, consider alternative algorithms, because how we find the maximum or minimum of our objective function is not important. Estimator properties only depend upon the successful location of the global maximum.

### Note

- \* The authors wish to thank three anonymous referees for their helpful comments. All remaining errors are the authors' own.

### References

- Almon, S. (1965). The distributed lag between capital appropriations and expenditures. *Econometrica* 33, 178–96.
- Belsley, D.A. (1982). Assessing the presence of harmful collinearity and other forms of weak data through a test for signal-to-noise. *Journal of Econometrics* 20, 211–53.
- Belsley, D.A. (1984). Demeaning conditioning diagnostics through centering. *American Statistician* 38, 73–93.
- Belsley, D.A. (1991). *Collinearity Diagnostics: Collinearity and Weak Data in Regression*. New York: Wiley.
- Belsley, D.A., E. Kuh, and R.E. Welsch (1980). *Regression Diagnostics: Identifying Influential Data and Sources of Collinearity*. New York: Wiley.
- Blanchard, O.J. (1987). Comment. *Journal of Business and Economic Statistics* 5, 449–51.
- Buck, A.J., and S. Hakim (1981). Appropriate roles for statistical decision theory and hypothesis testing in model selection: An exposition. *Regional Science and Urban Economics* 11, 135–47.
- Buse, A. (1994). Brickmaking and the collinear arts: a cautionary tale. *Canadian Journal of Economics* 27, 408–14.
- Chatterjee, S. and A.S. Hadi (1988). *Sensitivity Analysis in Linear Regression*. New York: Wiley.
- Cook, R.D. and S. Weisberg (1982). *Residuals and Influence in Regression*. London: Chapman & Hall.
- Davidson, R. and J.G. MacKinnon (1993). *Estimation and Inference in Econometrics*. New York: Oxford University Press.
- Dorfman, J.H. (1997). *Bayesian Economics through Numerical Methods: A Guide to Econometrics and Decision-Making with Prior Information*. New York: Springer.
- Fomby, T.B., R.C. Hill, and S.R. Johnson (1978). An optimality property of principal components regression. *Journal of the American Statistical Association* 73, 191–3.
- Fomby, T.B., R.C. Hill, and S.R. Johnson (1984). *Advanced Econometric Methods*. New York: Springer-Verlag.
- Fox, J. and G. Monette (1992). Generalized collinearity diagnostics. *Journal of the American Statistical Association* 87, 178–83.
- Golan, A., G.G. Judge, and D. Miller (1996). *Maximum Entropy Econometrics: Robust Estimation with Limited Data*. New York: John Wiley and Sons.
- Greene, W. (1997). *Econometric Analysis*, 3rd edn. Upper Saddle River, NJ: Prentice Hall.
- Hadi, A.S. and M.T. Wells (1990). Assessing the effects of multiple rows on the condition of a matrix. *Journal of the American Statistical Association* 85, 786–92.
- Hoerl, A.E., R.W. Kennard, and K.F. Baldwin (1975). Ridge regression: Some simulations. *Communications in Statistics, A*, 4 105–23.

- Judge, G.G. and M.E. Bock (1978). *Statistical Implications of Pretest and Stein-Rule Estimators in Econometrics*. Amsterdam: North-Holland.
- Judge, G.G. and M.E. Bock (1983). Biased Estimation. In Z. Griliches and M.D. Intrilligator (eds.), *Handbook of Econometrics, Volume 1*. Amsterdam: North-Holland.
- Judge, G.G., W.E. Griffiths, R.C. Hill, H. Lütkepohl, and T.C. Lee (1985). *The Theory and Practice of Econometrics*, 2nd edn. New York: John Wiley and Sons, Inc.
- Judge, G.G., R.C. Hill, W.E. Griffiths, H. Lütkepohl, and T.C. Lee (1988). *Introduction to the Theory and Practice of Econometrics*, 2nd edn. New York: John Wiley and Sons, Inc.
- Judge, G.G. and T.A. Yancy (1986). *Improved Methods of Inference in Econometrics*. Amsterdam: North-Holland.
- Kennedy, P. (1982). Eliminating problems caused by multicollinearity: A warning. *Journal of Economic Education* 13, 62–4.
- Kennedy, P. (1983). On an inappropriate means of reducing multicollinearity. *Regional Science and Urban Economics* 13, 579–81.
- Kennedy, P. (1998). *A Guide to Econometrics*, 4th edn. Cambridge: MIT Press.
- Lawless, J.F. and P. Wang (1976). A simulation study of ridge and other regression estimators. *Communications in Statistics A* 5, 307–23.
- Leamer, E.E. (1978). *Specification Searches: Ad Hoc Inference with Nonexperimental Data*. New York: Wiley.
- Lee, Kyung Yul and L.A. Weissfeld (1996). A multicollinearity diagnostic for the Cox model with time dependent covariates. *Communications in Statistics – Simulation* 25, 41–60.
- Lesaffre, E. and B.D. Marx (1993). Collinearity in generalized linear regression. *Communications in Statistics – Theory and Methods* 22, 1933–52.
- Mackinnon, M.J. and M.L. Puterman (1989). Collinearity in generalized linear models. *Communications in Statistics – Theory and Methods* 18, 3463–72.
- McCullagh, P. and J.A. Nelder (1989). *Generalized Linear Models*, 2nd edn. London: Chapman and Hall.
- Mason, R.L. and R.F. Gunst (1985). Outlier-induced collinearities. *Technometrics* 27, 401–7.
- Segerstedt, B. and H. Nyquist (1992). On the conditioning problem in generalized linear models. *Journal of Applied Statistics* 19, 513–22.
- Sengupta, D. (1995). Optimal choice of a new observation in a linear model. *Sankhya: The Indian Journal of Statistics Series A* 57, 137–53.
- Sengupta, D. and P. Bhimasankaram (1997). On the roles of observations in collinearity in the linear model. *Journal of the American Statistical Association* 92, 1024–32.
- Silvey, S. (1969). Multicollinearity and imprecise estimation. *Journal of the Royal Statistical Society B* 31, 539–52.
- Soofi, E.S. (1990). Effects of collinearity on information about regression coefficients. *Journal of Econometrics* 43, 255–74.
- Stewart, G.W. (1987). Collinearity and least squares regression. *Statistical Science* 1, 68–100.
- Theil, H. and A. Goldberger (1961). On pure and mixed statistical estimation in economics. *International Economic Review* 2, 65–78.
- Toro-Vizcarrondo, C. and T. Wallace (1968). A test of the mean square error criterion for restrictions in linear regression. *Journal of the American Statistical Association* 63, 558–76.
- Weissfeld, L.A. and S.M. Sereika (1991). A multicollinearity diagnostic for generalized linear models. *Communications in Statistics A* 20, 1183–98.
- Zellner, A. (1971). *An Introduction to Bayesian Inference in Econometrics*. New York: John Wiley and Sons.

CHAPTER THIRTEEN

# Nonnested Hypothesis Testing: An Overview

*M. Hashem Pesaran and Melvyn Weeks*

## 1 INTRODUCTION

This chapter focuses on the hypotheses testing problem when the hypotheses or models under consideration are “nonnested” or belong to “separate” families of distributions, in the sense that none of the individual models may be obtained from the remaining either by imposition of parameter restrictions or through a limiting process.<sup>1</sup> In econometric analysis nonnested models arise naturally when rival economic theories are used to explain the same phenomenon such as unemployment, inflation, or output growth. Typical examples from the economics literature are Keynesian and new classical explanations of unemployment, structural, and monetary theories of inflation, alternative theories of investment, and endogenous and exogenous theories of growth.<sup>2</sup> Nonnested models could also arise when alternative functional specifications are considered such as multinomial probit and logit distribution functions used in the qualitative choice literature, exponential, and power utility functions used in the asset pricing models, and a variety of nonnested specifications considered in the empirical analysis of income and wealth distributions. Finally, even starting from the same theoretical paradigm, it is possible for different investigators to arrive at different models if they adopt different conditioning or follow different paths to a more parsimonious model using the general-to-specific specification search methodology, advocated, for example by Hendry (1993).

The concept of an econometric model is discussed in Section 2, where a distinction is made between conditional and unconditional models. This is an important distinction since most applied work in econometrics takes place within a modeling framework where the behavior of one or more “endogenous” variables is often explained *conditional* on a set of “exogenous” variables. This discussion also highlights the importance of conditioning in the process of model evaluation.

Examples of nonnested models are given in Section 3. Section 4 discusses the differences that lie behind model selection and hypotheses testing. Although this chapter is primarily concerned with hypotheses testing involving nonnested models, a discussion of the differences and similarities of the two approaches to model evaluation can serve an important pedagogic purpose in clarifying the conditions under which one approach rather than the other could be appropriate.

The literature on nonnested hypothesis testing in statistics was pioneered by the seminal contributions of Cox (1961), Cox (1962), and Atkinson (1970), and was subsequently applied to econometric models by Pesaran (1974) and Pesaran and Deaton (1978). The analysis of nonnested regression models was further considered by Davidson and MacKinnon (1981), Fisher and McAleer (1981), Dastoor (1983), Deaton (1982), Sawyer (1983), Gouriéroux, Monfort, and Trognon (1983), and Godfrey and Pesaran (1983).<sup>3</sup> This literature is reviewed in Section 5 where we examine a number of alternative approaches to testing nonnested hypotheses, including the encompassing approach advanced by Mizon and Richard (1986), Gouriéroux and Monfort (1995), and Smith (1993).

Generally speaking, two models, say  $H_f$  and  $H_g$ , are said to be nonnested if it is not possible to derive  $H_f$  (or  $H_g$ ) from the other model either by means of an exact set of parametric restrictions or as a result of a limiting process. But for many purposes a more rigorous definition is needed. Section 6 examines this issue and focuses on the Kullback–Leibler divergence measure which has played a pivotal role in the development of a number of nonnested test statistics. The Vuong approach to model selection, viewed as a hypothesis testing problem is also discussed in this section (see Vuong, 1989). Section 7 deals with the practical problems involved in the implementation of the Cox procedure. Apart from a few exceptions, the centering of the loglikelihood ratio statistic required to construct the Cox statistic, will involve finding an estimate of the Kullback–Leibler measure of closeness of the alternative to the null hypothesis, which in most cases is not easy to compute using analytical techniques. Subsequently, we explore two methods which circumvent the problem. First, following work by Pesaran and Pesaran (1993), we examine the simulation approach which provides a consistent estimator of the KLIC measure. However, since this approach is predicated upon the adherence to a classical testing framework, we also examine the use of a parametric bootstrap approach. Whereas the use of simulation facilitates the construction of a pivotal test statistic with an asymptotically well-defined limiting distribution, the bootstrap procedure effectively replaces the theoretical distribution with the empirical distribution function. We also discuss the use of pivotal bootstrap statistics for testing nonnested models.

## 2 MODELS AND THEIR SPECIFICATION

Suppose the focus of the analysis is to consider the behavior of the  $n \times 1$  vector of random variables  $w_t = (w_{1t}, w_{2t}, \dots, w_{nt})'$  observed over the period  $t = 1, 2, \dots, T$ . A model of  $w_t$ , indexed by  $\mathcal{M}_i$ , is defined by the joint probability distribution function (pdf) of the observations

$$W = (w'_1, w'_2, \dots, w'_T)'$$

$$\mathcal{M}_i : f_i(w_1, w_2, \dots, w_T | w_0, \varphi_i) = f_i(W | w_0, \varphi_i), \quad i = 1, 2, \dots, m, \quad (13.1)$$

where  $f_i(\cdot)$  is the probability density function of the model (hypothesis)  $\mathcal{M}_i$ , and  $\varphi_i$  is a  $p_i \times 1$  vector of unknown parameters associated with model  $\mathcal{M}_i$ .<sup>4</sup>

The models characterized by  $f_i(W | w_0, \varphi_i)$  are *unconditional* in the sense that probability distribution of  $w_t$  is fully specified in terms of some initial values,  $w_0$ , and for a given value of  $\varphi_i$ . In econometrics the interest often centers on conditional models, where a vector of "endogenous" variables,  $y_t$ , is explained (or modeled) *conditional* on a set of "exogenous", variables,  $x_t$ . Such conditional models can be derived from (13.1) by noting that

$$\begin{aligned} f_i(w_1, w_2, \dots, w_T | w_0, \varphi_i) &= f_i(y_1, y_2, \dots, y_T | x_1, x_2, \dots, x_T, \psi(\varphi_i)) \\ &\quad \times f_i(x_1, x_2, \dots, x_T | w_0, \kappa(\varphi_i)), \end{aligned} \quad (13.2)$$

where  $w_t = (y'_t, x'_t)$ . The unconditional model  $\mathcal{M}_i$  is decomposed into a conditional model of  $y_t$  given  $x_t$  and a marginal model of  $x_t$ . Denoting the former by  $\mathcal{M}_{i,y|x}$  we have

$$\mathcal{M}_{i,y|x} : f_i(y_1, y_2, \dots, y_T | x_1, x_2, \dots, x_T, w_0, \psi(\varphi_i)) = f_i(Y | X, w_0, \psi(\varphi_i)), \quad (13.3)$$

where  $Y = (y'_1, y'_2, \dots, y'_T)'$  and  $X = (x'_1, x'_2, \dots, x'_T)'$ .

Confining attention to the analysis and comparison of conditional models is valid only if the variations in the parameters of the marginal model,  $\kappa(\varphi_i)$ , does not induce changes in the parameters of the conditional model,  $\psi(\varphi_i)$ . Namely  $\partial\psi(\varphi_i)/\partial'\kappa(\varphi_i) = 0$ . When this condition holds it is said that  $x_t$  is *weakly exogenous* for  $\psi_i$ . The parameters of the conditional model,  $\psi_i = \psi(\varphi_i)$ , are often referred to as the *parameters of interest*.<sup>5</sup>

The conditional models  $\mathcal{M}_{i,y|x}$   $i = 1, 2, \dots, m$  all are based on the same conditioning variables,  $x_t$ , and differ only in so far as they are based upon different pdfs. We may introduce an alternative set of models which share the same pdfs but differ with respect to the inclusion of exogenous variables. For any model,  $\mathcal{M}_i$  we may partition the set of exogenous variables  $x_t$  according to a simple included/excluded dichotomy. Therefore  $x_t = (x'_{it}, x'^*_{it})'$  writes the set of exogenous variables according to a subset  $x_{it}$  which are included in model  $\mathcal{M}_i$ , and a subset  $x^*_{it}$  which are excluded. We may then write

$$\begin{aligned} f_i(Y | x_1, x_2, \dots, x_T, w_0, \varphi_i) &= f_i(Y | x_{i1}, x_{i2}, \dots, x_{iT}, x^*_{i1}, x^*_{i2}, \dots, x^*_{iT}, w_0, \varphi_i) \\ &= f_i(Y | X_i, w_0, \psi_i(\varphi_i)) \times f_i(X_i^* | X_i, w_0, c_i(\varphi_i)), \end{aligned}$$

where  $X_i = (x'_{i1}, x'_{i2}, \dots, x'_{iT})'$  and  $X_i^* = (x'^*_{i1}, x'^*_{i2}, \dots, x'^*_{iT})'$ . As noted above in the case of models differentiated solely by different pdfs, a comparison of models based upon the partition of  $x_t$  into  $x_{it}$  and  $x^*_{it}$  should be preceded by determining whether  $\partial\psi_i(\varphi_i)/\partial c_i(\varphi_i) = 0$ .

The above setup allows consideration of rival models that could differ in the conditioning set of variables,  $\{x_{it}, i = 1, 2, \dots, m\}$  and/or the functional form of their underlying probability distribution functions,  $\{f_i(\cdot), i = 1, 2, \dots, m\}$ . In much of this chapter we will be concerned with two rival (conditional) models and for notational convenience we denote them by

$$H_f : \mathcal{F}_\theta = \{f(y_t | x_t, \Omega_{t-1}; \theta), \theta \in \Theta\}, \quad (13.4)$$

$$H_g : \mathcal{F}_\gamma = \{g(y_t | z_t, \Omega_{t-1}; \gamma), \gamma \in \Gamma\}, \quad (13.5)$$

where  $\Omega_{t-1}$  denotes the set of all past observations on  $y$ ,  $x$  and  $z$ ,  $\theta$  and  $\gamma$  are respectively  $k_f$  and  $k_g$  vectors of unknown parameters belonging to the non-empty compact sets  $\Theta$  and  $\Gamma$ , and where  $x$  and  $z$  represent the conditioning variables. For the sake of notational simplicity we shall also often use  $f_t(\theta)$  and  $g_t(\gamma)$  in place of  $f(y_t | x_t, \Omega_{t-1}; \theta)$  and  $g(y_t | z_t, \Omega_{t-1}; \gamma)$ , respectively.

Now given the observations  $(y_t, x_t, z_t, t = 1, 2, \dots, T)$  and conditional on the initial values  $w_0$ , the maximum likelihood (ML) estimators of  $\theta$  and  $\gamma$  are given by

$$\hat{\theta}_T = \arg \max_{\theta \in \Theta} L_f(\theta), \quad \hat{\gamma}_T = \arg \max_{\gamma \in \Gamma} L_g(\gamma), \quad (13.6)$$

where the respective loglikelihood functions are given by:

$$L_f(\theta) = \sum_{t=1}^T \ln f_t(\theta), \quad L_g(\gamma) = \sum_{t=1}^T \ln g_t(\gamma). \quad (13.7)$$

Throughout we shall assume that the conditional densities  $f_t(\theta)$  and  $g_t(\gamma)$  satisfy the usual regularity conditions as set out, for example, in White (1982) and Smith (1993), needed to ensure that  $\hat{\theta}_T$  and  $\hat{\gamma}_T$  have asymptotically normal limiting distributions under the data generating process (DGP). We allow the DGP to differ from  $H_f$  and  $H_g$ , and denote it by  $H_h$ ; thus admitting the possibility that both  $H_f$  and  $H_g$  could be misspecified and that both are likely to be rejected in practice. In this setting  $\hat{\theta}_T$  and  $\hat{\gamma}_T$  are referred to as quasi-ML estimators and their probability limits under  $H_h$ , which we denote by  $\theta_{h*}$  and  $\gamma_{h*}$  respectively, are known as (asymptotic) pseudo-true values. These pseudo-true values are defined by

$$\theta_{h*} = \arg \max_{\theta \in \Theta} E_h\{T^{-1}L_f(\theta)\}, \quad \gamma_{h*} = \arg \max_{\gamma \in \Gamma} E_h\{T^{-1}L_g(\gamma)\}, \quad (13.8)$$

where  $E_h(\cdot)$  denotes expectations are taken under  $H_h$ . In the case where  $w_t$  follows a strictly stationary process, (13.8) simplifies to

$$\theta_{h*} = \arg \max_{\theta \in \Theta} E_h\{\ln f_t(\theta)\}, \quad \gamma_{h*} = \arg \max_{\gamma \in \Gamma} E_h\{\ln g_t(\gamma)\}. \quad (13.9)$$

To ensure global identifiability of the pseudo-true values, it will be assumed that  $\theta_{f*}$  and  $\gamma_{f*}$  provide *unique* maxima of  $E_h\{T^{-1}L_f(\theta)\}$  and  $E_h\{T^{-1}L_g(\gamma)\}$ , respectively. Clearly, under  $H_f$ , namely assuming  $H_f$  is the DGP, we have  $\theta_{f*} = \theta_0$ , and  $\gamma_{f*} = \gamma_*(\theta_0)$  where  $\theta_0$  is the “true” value of  $\theta$  under  $H_f$ . Similarly, under  $H_g$  we have  $\gamma_{g*} = \gamma_0$ , and  $\theta_{g*} = \theta_*(\gamma_0)$  with  $\gamma_0$  denoting the “true” value of  $\gamma$  under  $H_g$ . The functions  $\gamma_*(\theta_0)$ , and  $\theta_*(\gamma_0)$  that relate the parameters of the two models under consideration are called the *binding* functions. These functions do not involve the true model,  $H_h$ , and only depend on the models  $H_f$  and  $H_g$  that are under consideration. As we shall see later a formal definition of encompassing is given in terms of the pseudo-true values,  $\theta_{h*}$  and  $\gamma_{h*}$ , and the binding functions  $\gamma_*(\theta_0)$ , and  $\theta_*(\gamma_0)$ .

Before proceeding further it would be instructive to consider some examples of nonnested models from the literature.

### 3 EXAMPLES OF NONNESTED MODELS

We start with examples of unconditional nonnested models. One such example, originally discussed by Cox (1961) is that of testing a lognormal versus an exponential distribution.

$$H_f : f(y_t | \theta) = f_t(\theta) = y_t^{-1}(2\pi\theta_2)^{-1/2} \exp\left\{-\frac{(\ln y_t - \theta_1)^2}{2\theta_2}\right\}, \quad \infty > \theta_2 > 0, \quad y_t > 0.$$

$$H_g : g(y_t | \gamma) = g_t(\gamma) = \gamma^{-1} \exp(-y_t/\gamma), \quad \gamma > 0, \quad y_t > 0.$$

These hypotheses (models) are globally nonnested, in the sense that neither can be obtained from the other either by means of suitable parametric restrictions or by a limiting process.<sup>6</sup> Under  $H_f$  the pseudo-true value of  $\gamma$ , denoted by  $\gamma_{f*}$  is obtained by solving the following maximization problem

$$\gamma_{f*} = \arg \max_{\gamma > 0} E_f\{\ln g_t(\gamma)\}.$$

But<sup>7</sup>

$$E_f\{\ln g_t(\gamma)\} = -\ln \gamma - E_f(y_t)/\gamma = -\ln \gamma - \exp(\theta_1 + 0.5\theta_2)/\gamma,$$

which yields

$$\gamma_{f*} = \gamma_*(\theta_0) = \exp(\theta_{10} + 0.5\theta_{20}).$$

Similarly, under  $H_g$  we have<sup>8</sup>

$$\theta_1^*(\lambda_0) = \ln \gamma_0 - 0.5772, \quad \theta_2^*(\gamma_0) = 1.6449.$$

Other examples of nonnested unconditional models include lognormal versus Weibull and Pereira (1984) and lognormal versus gamma distribution, Pesaran (1987).

The most prominent example of conditional nonnested models is linear normal regression models with "rival" sets of conditioning variables. As an example consider the following regression models:

$$H_f: y_t = \alpha' x_t + u_{tf}, \quad u_{tf} \sim N(0, \sigma^2), \quad \infty > \sigma^2 > 0, \quad (13.10)$$

$$H_g: y_t = \beta' z_t + u_{tg}, \quad u_{tg} \sim N(0, \omega^2), \quad \infty > \omega^2 > 0. \quad (13.11)$$

The conditional probability density associated with these regression models are given by

$$H_f: f(y_t | x_t; \theta) = (2\pi\sigma^2)^{-1/2} \exp \left\{ \frac{-1}{2\sigma^2} (y_t - \alpha' x_t)^2 \right\}, \quad (13.12)$$

$$H_g: g(y_t | z_t; \theta) = (2\pi\omega^2)^{-1/2} \exp \left\{ \frac{-1}{2\omega^2} (y_t - \beta' z_t)^2 \right\}, \quad (13.13)$$

where  $\theta = (\alpha', \sigma^2)',$  and  $\gamma = (\beta', \omega^2)'$ . These regression models are nonnested if it is not possible to write  $x_t$  as an exact linear function of  $z_t$  and vice versa, or more formally if  $x_t \not\equiv z_t$  and  $z_t \not\equiv x_t.$  Model  $H_f$  is said to be nested in  $H_g$  if  $x_t \subset z_t$  and  $z_t \not\equiv x_t.$  The two models are observationally equivalent if  $x_t \subset z_t$  and  $z_t \subset x_t.$  Suppose now that neither of these regression models is true and the DGP is given by

$$H_h: y_t = \delta' w_t + u_{th}, \quad u_{th} \sim N(0, v^2), \quad \infty > v^2 > 0. \quad (13.14)$$

It is then easily seen that conditional on  $\{x_t, z_t, w_t, t = 1, 2, \dots, T\}$

$$E_h\{T^{-1}L_f(\theta)\} = -\frac{1}{2} \ln(2\pi\sigma^2) - \frac{v^2}{2\sigma^2} - \frac{\delta' \hat{\Sigma}_{ww} \delta - 2\delta' \hat{\Sigma}_{wx} \alpha + \alpha' \hat{\Sigma}_{xx} \alpha}{2\sigma^2},$$

where

$$\hat{\Sigma}_{ww} = T^{-1} \sum_{t=1}^T w_t w_t', \quad \hat{\Sigma}_{xx} = T^{-1} \sum_{t=1}^T x_t x_t', \quad \hat{\Sigma}_{wx} = T^{-1} \sum_{t=1}^T w_t x_t'.$$

Maximizing  $E_h\{T^{-1}L_f(\theta)\}$  with respect to  $\theta$  now yields the conditional pseudo-true values:

$$\theta_{h*} = \begin{pmatrix} \alpha_{h*} \\ \sigma_{h*}^2 \end{pmatrix} = \begin{pmatrix} \hat{\Sigma}_{xx}^{-1} \hat{\Sigma}_{wx} \delta \\ v^2 + \delta' (\hat{\Sigma}_{ww} - \hat{\Sigma}_{wx} \hat{\Sigma}_{xx}^{-1} \hat{\Sigma}_{wx}) \delta \end{pmatrix}. \quad (13.15)$$

Similarly,

$$\gamma_{h*} = \begin{pmatrix} \beta_{h*} \\ w_{h*}^2 \end{pmatrix} = \begin{pmatrix} \hat{\Sigma}_{zz}^{-1} \hat{\Sigma}_{zw} \delta \\ v^2 + \delta' (\hat{\Sigma}_{ww} - \hat{\Sigma}_{wz} \hat{\Sigma}_{zz}^{-1} \hat{\Sigma}_{zw}) \delta \end{pmatrix}, \quad (13.16)$$

where

$$\hat{\Sigma}_{zz} = T^{-1} \sum_{t=1}^T z_t z'_t, \quad \hat{\Sigma}_{wz} = T^{-1} \sum_{t=1}^T w_t z'_t.$$

When the regressors are stationary, the unconditional counterparts of the above pseudo-true values can be obtained by replacing  $\hat{\Sigma}_{ww}$ ,  $\hat{\Sigma}_{xx}$ ,  $\hat{\Sigma}_{wx}$ , etc. by their population values, namely  $\Sigma_{ww} = E(w, w')$ ,  $\Sigma_{xx} = E(x, x')$ ,  $\Sigma_{wx} = E(w, x')$ , etc.

Other examples of nonnested regression models include models with endogenous regressors estimated by instrumental variables (see, for example, Ericsson, 1983; and Godfrey, 1983), nonnested nonlinear regression models and regression models where the left-hand side variables of the rival regressions are known transformations of a dependent variable of interest. One important instance of this last example is the problem of testing linear versus loglinear regression models and vice versa.<sup>9</sup> More generally we may have

$$H_f : f(y_t) = \alpha' x_t + u_{tf}, \quad u_{tf} \sim N(0, \sigma^2), \quad \sigma > 0,$$

$$H_g : g(y_t) = \beta' z_t + u_{tg}, \quad u_{tg} \sim N(0, \omega^2), \quad \omega > 0,$$

where  $f(y_t)$  and  $g(y_t)$  are known one-to-one functions of  $y_t$ . Within this more general regression framework testing a linear versus a loglinear model is characterized by  $f(y_t) = y_t$  and  $g(y_t) = \ln(y_t)$ ; a ratio model versus a loglinear model by  $f(y_t) = y_t/q_t$  and  $g(y_t) = \ln(y_t)$ , where  $q_t$  is an observed regressor, and a ratio versus a linear model by  $f(y_t) = y_t/q_t$  and  $g(y_t) = y_t$ . For example, in analysis of aggregate consumption a choice needs to be made between a linear and a loglinear specification of the aggregate consumption on the one hand, and between a loglinear and a saving rate formulation on the other hand. The testing problem is further complicated here due to the linear transformations of the dependent variable, and additional restrictions are required if the existence of pseudo-true values in the case of these models are to be ensured. For example, suitable truncation restrictions need to be imposed on the errors of the linear model when it is tested against a loglinear alternative.

Other examples where specification of an appropriate error structure plays an important role in empirical analysis include discrete choice and duration models used in microeconometric research. Although the analyst may utilize both prior knowledge and theory to select an appropriate set of regressors, there is generally little guidance in terms of the most appropriate probability distribution. Nonnested hypothesis testing is particularly relevant to microeconometric research where the same set of regressors are often used to explain individual decisions but based on different functional distributions, such as multinomial probit and logit specifications in the analysis of discrete choice, exponential and Weibull distributions in the analysis of duration data. In the simple case of a probit ( $H_f$ ) versus a logit model ( $H_g$ ) we have

$$H_f : \Pr(y_t = 1) = \Phi(\theta' x_t) = \int_{-\infty}^{\theta' x_t} \frac{1}{\sqrt{2\pi}} \exp\left\{-\frac{1}{2}v^2\right\} dv \quad (13.17)$$

$$H_g : \Pr(y_t = 1) = \Lambda(\gamma' z_t) = \frac{e^{\gamma' z_t}}{1 + e^{\gamma' z_t}} \quad (13.18)$$

where  $y_t$ ,  $t = 1, 2, \dots, T$  are independently distributed binary random variables taking the value of 1 or 0. In practice the two sets of regressors  $x_t$  used in the probit and logit specifications are likely to be identical, and it is only the form of the distribution functions that separate the two models. Other functional forms can also be entertained. Suppose, for example, that the true DGP for this simple discrete choice problem is given by the probability distribution function  $H(\delta' x_t)$ , then pseudo-true values for  $\theta$  and  $\gamma$  can be obtained as functions of  $\delta$ , but only in an implicit form. We first note that the loglikelihood function under  $H_f$ , for example, is given by

$$L_f(\theta) = \sum_{t=1}^T y_t \log[\Phi(\theta' x_t)] + \sum_{t=1}^T (1 - y_t) \log[1 - \Phi(\theta' x_t)],$$

and hence under the assumed DGP we have

$$\begin{aligned} E_h\{T^{-1}L_f(\theta)\} &= T^{-1} \sum_{t=1}^T H(\delta' x_t) \log[\Phi(\theta' x_t)] \\ &\quad + T^{-1} \sum_{t=1}^T [1 - H(\delta' x_t)] \log[1 - \Phi(\theta' x_t)]. \end{aligned}$$

Therefore, the pseudo-true value of  $\theta$ , namely  $\theta_*(\delta)$  or simply  $\theta_*$ , satisfies the following equation

$$T^{-1} \sum_{t=1}^T x_t \phi(\theta_*' x_t) \left\{ \frac{H(\delta' x_t)}{\Phi(\theta_*' x_t)} - \frac{1 - H(\delta' x_t)}{1 - \Phi(\theta_*' x_t)} \right\} = 0,$$

where  $\phi(\theta_*' x_t) = (2\pi)^{-1/2} \exp[-\frac{1}{2}(\theta_*' x_t)^2]$ . Using results in Amemiya (1985, pp. 271–2) it is easily established that the solution of  $\theta_*$  in terms of  $\delta$  is in fact unique, and  $\theta_* = \delta$  if and only if  $\Phi(\cdot) = H(\cdot)$ . Similar results also obtain for the logistic specification.

## 4 MODEL SELECTION VERSUS HYPOTHESIS TESTING

Hypothesis testing and model selection are different strands in the model evaluation literature. However, these strands differ in a number of important respects which are worth emphasizing here.<sup>10</sup> Model selection begins with a given set of models,  $\mathcal{M}$ , characterized by the (possibly) conditional pdfs

$$\mathcal{M} = \{f_i(Y | X_i, \psi_i), i = 1, 2, \dots, m\},$$

with the aim of *choosing* one of the models under consideration for a particular purpose with a specific loss (utility) function in mind. In essence model selection is a part of decision making and as argued in Granger and Pesaran (2000) ideally it should be fully integrated into the decision making process. However, most of the current literature on model selection builds on statistical measure of fit such as sums of squares of residuals or more generally maximized loglikelihood values, rather than economic value which one would expect to follow from a model choice.<sup>11</sup> As a result model selection seems much closer to hypothesis testing than it actually is in principle.

The model selection process treats all models under consideration symmetrically, while hypothesis testing attributes a different status to the null and to the alternative hypotheses and by design treats the models asymmetrically. Model selection always ends in a definite outcome, namely one of the models under consideration is selected for use in decision making. Hypothesis testing on the other hand asks whether there is any statistically significant evidence (in the Neyman–Pearson sense) of departure from the null hypothesis in the direction of one or more alternative hypotheses. Rejection of the null hypothesis does not necessarily imply acceptance of any one of the alternative hypotheses; it only warns the investigator of possible shortcomings of the null that is being advocated. Hypothesis testing does not seek a definite outcome and if carried out with due care need not lead to a favorite model. For example, in the case of nonnested hypothesis testing it is possible for all models under consideration to be rejected, or all models to be deemed as observationally equivalent.

Due to its asymmetric treatment of the available models, the choice of the null hypothesis plays a critical role in the hypothesis testing approach. When the models are nested the most parsimonious model can be used as the null hypothesis. But in the case of nonnested models (particularly when the models are globally nonnested) there is no natural null, and it is important that the null hypothesis is selected on a priori grounds.<sup>12</sup> Alternatively, the analysis could be carried out with different models in the set treated as the null. Therefore, the results of nonnested hypothesis testing is less clear cut as compared to the case where the models are nested.<sup>13</sup>

It is also important to emphasize the distinction between *paired* and joint nonnested hypothesis tests. Letting  $f_1$  denote the null model and  $f_i \in \mathcal{M}, i = 2, \dots, m$  index a set of  $m - 1$  alternative models, a paired test is a test of  $f_1$  against a *single* member of  $\mathcal{M}$ , whereas a joint test is a test of  $f_1$  against multiple alternatives in  $\mathcal{M}$ . McAleer (1995) is careful to highlight this distinction and in doing so points out a deficiency in many applied studies insofar as many authors have utilized a sequence of paired tests for problems characterized by multiple alternatives. Examples of studies which have applied nonnested tests to the choice between more than two models include Sawyer (1984), Smith and Maddala (1983) and Davidson and MacKinnon (1981). The paper by Sawyer is particularly relevant since he develops the multiple model equivalent of the Cox test.

The distinction between model selection and nonnested hypothesis tests can also be motivated from the perspective of Bayesian versus sampling-theory approaches to the problem of inference. For example, it is likely that with a large

amount of data the posterior probabilities associated with a particular hypothesis will be close to one. However, the distinction drawn by Zellner (1971) between "comparing" and "testing" hypothesis is relevant given that within a Bayesian perspective the progression from a set of prior to posterior probabilities on  $\mathcal{M}$ , mediated by the Bayes factor, does not necessarily involve a decision to accept or reject the hypothesis. If a decision is required it is generally based upon minimizing a particular expected loss function. Thus, model selection motivated by a decision problem is much more readily reconcilable with the Bayesian rather than the classical approach to model selection.

Finally, the choice between hypothesis testing and model selection clearly depends on the primary objective of the exercise. There are no definite rules. Model selection is more appropriate when the objective is decision making. Hypothesis testing is better suited to inferential problems where the empirical validity of a theoretical prediction is the primary objective. A model may be empirically adequate for a particular purpose but of little relevance for another use. Only in the unlikely event that the true model is known or knowable will the selected model be universally applicable. In the real world where the truth is elusive and unknowable both approaches to model evaluation are worth pursuing.

## 5 ALTERNATIVE APPROACHES TO TESTING NONNESTED HYPOTHESES WITH APPLICATION TO LINEAR REGRESSION MODELS

To provide an intuitive introduction to concepts which are integral to an understanding of nonnested hypothesis tests we consider testing of linear regression models as a convenient starting point. In the ensuing discussion we demonstrate that despite its special features nonnested hypothesis testing is firmly rooted within the Neyman–Pearson framework.

There are three general approaches to nonnested hypothesis testing all discussed in the pioneering contributions of Cox (1961) and Cox (1962). (i) The modified (centered) loglikelihood ratio procedure, also known as the Cox test. (ii) The comprehensive model approach, whereby the nonnested models are tested against an artificially constructed general model that includes the nonnested models as special cases. This approach was advocated by Atkinson (1970) and was later taken up under a different guise by Davidson and MacKinnon (1981) in developing their J-test and by Fisher and McAleer (1981) who proposed a related alternative procedure known as the JA-test. (iii) A third approach, originally considered by Deaton (1982) and Dastoor (1983) and further developed by Gouriéroux *et al.* (1983) and Mizon and Richard (1986) is the encompassing procedure where the ability of one model to explain particular features of an alternative model is tested directly. The Wald and score encompassing tests (usually denoted by WET and SET) are typically constructed under the assumption that one of the rival models is correct. Encompassing tests when the true model does not necessarily lie in the set of models (whether nested or nonnested) under consideration are proposed by Gouriéroux and Monfort (1995) and Smith (1993).

We shall now illustrate the main features of these three approaches in the context of the classical linear normal regression models (13.10) and (13.11) set out above. Rewriting these models in familiar matrix notations we have:

$$H_f : \mathbf{y} = \mathbf{X}\boldsymbol{\alpha} + \mathbf{u}_f, \quad \mathbf{u}_f \sim N(0, \sigma^2 \mathbf{I}_T), \quad (13.19)$$

$$H_g : \mathbf{y} = \mathbf{Z}\boldsymbol{\beta} + \mathbf{u}_g, \quad \mathbf{u}_g \sim N(0, \omega^2 \mathbf{I}_T), \quad (13.20)$$

where  $\mathbf{y}$  is the  $T \times 1$  vector of observations on the dependent variable,  $\mathbf{X}$  and  $\mathbf{Z}$  are  $T \times k_f$  and  $T \times k_g$  observation matrices for the regressors of models  $H_f$  and  $H_g$ ,  $\boldsymbol{\alpha}$  and  $\boldsymbol{\beta}$  are the  $k_f \times 1$  and  $k_g \times 1$  unknown regression coefficient vectors,  $\mathbf{u}_f$  and  $\mathbf{u}_g$  are the  $T \times 1$  disturbance vectors, and  $\mathbf{I}_T$  is an identity matrix of order  $T$ . In addition, throughout this section we assume that

$$T^{-1}\mathbf{X}'\mathbf{u}_f \xrightarrow{p} 0, \quad T^{-1}\mathbf{Z}'\mathbf{u}_g \xrightarrow{p} 0, \quad T^{-1/2}\mathbf{X}'\mathbf{u}_f \xrightarrow{d} N(0, \sigma^2 \Sigma_{xx}),$$

$$T^{-1}\mathbf{Z}'\mathbf{u}_g \xrightarrow{p} 0, \quad T^{-1}\mathbf{Z}'\mathbf{u}_f \xrightarrow{p} 0, \quad T^{-1/2}\mathbf{Z}'\mathbf{u}_g \xrightarrow{d} N(0, \omega^2 \Sigma_{zz}),$$

$$\hat{\Sigma}_{xx} = T^{-1}\mathbf{X}'\mathbf{X} \xrightarrow{p} \Sigma_{xx}, \quad \hat{\Sigma}_{zz} = T^{-1}\mathbf{Z}'\mathbf{Z} \xrightarrow{p} \Sigma_{zz}, \quad \hat{\Sigma}_{zx} = T^{-1}\mathbf{Z}'\mathbf{X} \xrightarrow{p} \Sigma_{zx},$$

where  $\xrightarrow{p}$  denotes convergence in probability, the matrices  $\hat{\Sigma}_{xx}$ ,  $\Sigma_{xx}$ ,  $\hat{\Sigma}_{zz}$ ,  $\Sigma_{zz}$  are non-singular,  $\Sigma_{zx} = \Sigma'_{xz} \neq 0$ , and set

$$\Sigma_f = \Sigma_{xx} - \Sigma_{xz}\Sigma_{zz}^{-1}\Sigma_{zx}, \quad \text{and} \quad \Sigma_g = \Sigma_{zz} - \Sigma_{zx}\Sigma_{xx}^{-1}\Sigma_{xz}.$$

## 5.1 Motivation for nonnested statistics

From a statistical view point the main difference between the nested and nonnested hypothesis testing lies in the fact that the usual loglikelihood ratio or Wald statistics used in the conventional hypothesis testing are automatically centered at zero under the null when the hypotheses under consideration are nested while this is not true in the case of nonnested hypotheses. However, once the conventional test statistics are appropriately centered (at least asymptotically) the same classical techniques can be applied to testing of nonnested hypotheses. Using the two nonnested linear regression models in (13.19) and (13.20) we first demonstrate the problems with standard test statistics by focusing on a simple comparison of sums of squared errors.

Consider the following test statistic:

$$\xi_T = \tilde{\sigma}_g^2 - \tilde{\sigma}_f^2, \quad (13.21)$$

where

$$\tilde{\sigma}_f^2 = \mathbf{e}'_f \mathbf{e}_f / (T - k_f)$$

$$\tilde{\sigma}_g^2 = \mathbf{e}'_g \mathbf{e}_g / (T - k_g),$$

and  $e_f$  is the OLS residual vector under  $H_f$  such that  $e_f = M_f y$ . Note that (13.21) represents a natural starting point being the difference between the mean sum of squared errors for the two models.

In general the exact distribution of  $\xi_T$  will depend on the unknown parameters. To see this, first note that under  $H_f$ ,  $e_f = M_f(u_f + X\alpha)$  therefore, (since  $M_f X = 0$ ), we have

$$(T - k_f)\tilde{\sigma}_f^2 = u_f' M_f u_f. \quad (13.22)$$

Now under  $H_g$ ,

$$e_g = M_g y = M_g(X\alpha + u_f),$$

or

$$e_g = M_g X\alpha + M_g u_f$$

and

$$\begin{aligned} (T - k_g)\tilde{\sigma}_g^2 &= e_g'e_g \\ &= (u_f' + \alpha'X')M_g(X\alpha + u_f) \\ &= u_f'M_g u_f + 2\alpha'X'M_g u_f + \alpha'X'M_g X\alpha. \end{aligned} \quad (13.23)$$

Using (13.22) and (13.23) in (13.21) and taking expectations (under  $H_f$ ) we have

$$E(\xi_T) = \frac{\alpha'X'M_g X\alpha}{T - k_g} \geq 0, \quad (13.24)$$

which we denote by  $\mu_T = (\alpha'X'M_g X\alpha)/(T - k_g)$ . Since  $\xi_T$  does not have mean zero under the null hypothesis  $H_f$ , then  $\xi_T$  cannot provide us with a suitable *test-statistic*. Notice, however that when  $H_f$  is nested within  $H_g$ , then  $M_g X = 0$  and  $\xi_T$  will have mean zero (exactly) under  $H_g$ . In this case, if we also assume that  $u_f$  is normally distributed, it can be easily shown that

$$\frac{(T - k_f)\xi_T}{r\tilde{\sigma}_g^2} = 1 - F_{r, T - k_g}$$

where  $F_{r, T - k_g}$  is distributed as a (central)  $F$  with  $r$  and  $T - k_g$  degrees of freedom;  $r$  here stands for the number of restrictions that we need to impose on  $H_g$  in order to obtain  $H_f$ .

A fundamental tenet of classical hypothesis testing is that the distribution of the test statistic is known under a well specified null hypothesis. Thus, in this context if  $H_f$  is nested within  $H_g$  then under the null of  $H_f$  the normalized difference between the sum of squared errors have a zero expectation. When  $H_f$  is not nested within  $H_g$  we may adopt a number of alternate approaches. First, a suitable test statistic that has zero mean asymptotically will be

$$z_T = \xi_T - \hat{\mu}_T$$

where  $\hat{\mu}_T$  is a consistent estimator of  $\mu_T$  under  $H_f$ . More specifically

$$z_T = \tilde{\sigma}_g^2 - \tilde{\sigma}_f^2 - \frac{\hat{\alpha}' X' M_g X \hat{\alpha}}{T - k_g}, \quad (13.25)$$

where  $\hat{\alpha} = (X'X)^{-1}X'y$ . Equation (13.25) represents an example of *centering* of a test statistic such that the distribution of  $z_T$  is known (asymptotically). Cox (1961, 1962) utilized this approach to center the loglikelihood ratio statistic for two nonnested models. When the models are nested the loglikelihood ratio statistic is properly centered (at least asymptotically). For example, if we let  $L_f(\theta)$  and  $L_g(\gamma)$  denote, respectively the loglikelihood functions for  $H_f$  and  $H_g$ , and we assume that  $H_f$  is nested within  $H_g$ , then under  $H_f$  the loglikelihood ratio statistic,  $2[L_g(\hat{\gamma}_T) - L_f(\hat{\theta}_T)]$ , does not require any centering and the test defined by the critical region

$$2[L_g(\hat{\gamma}_T) - L_f(\hat{\theta}_T)] \geq \chi^2_{(1-\alpha)}(r),$$

where  $r$  is the number of parameter restrictions required to obtain  $H_f$  from  $H_g$ , asymptotically has the size  $\alpha$  and is consistent. In the case of nonnested models the likelihood ratio statistic is not distributed as a chi-squared random variable. The reason for this is simple. The degrees of freedom of the chi-square statistic for the *LR* test is equal to the reduction in the size of the parameter space after imposing the necessary set of zero restrictions. Thus, if neither  $H_f$  nor  $H_g$  nests the other model, the attendant parameter spaces and hence the likelihoods are unrelated. In Section 5.2 we examine the application of centering (or mean adjustment) of the likelihood ratio statistic to obtain a test statistic that has a known asymptotic distribution. Given that in most instances the form of mean adjustment involves analytically intractable expectations in Section 7.1 we examine the use of simulation methods as a method of circumventing this problem.

Following seminal work by Efron (1979), an alternative approach conducts inference utilizing the empirical distribution function of the test statistic. In this instance there is, in general, no need to center  $\xi_T$  using  $\hat{\mu}_T$ . Instead we take  $\xi_T$  as the observed test statistic, and given a null hypothesis, we simulate a large number, say  $R$ , of the  $\tilde{\sigma}_g^2, \tilde{\sigma}_f^2$  pairs. The empirical distribution function for  $\xi_T$  is then constructed based on  $\hat{\sigma}_{gr}^2$  and  $\hat{\sigma}_{fr}^2, r = 1, 2, \dots, R$ . In Section 7.2 we examine the use of bootstrap procedures for conducting nonnested hypothesis tests. We also consider the case for combining the type of mean adjustment in (13.25) with bootstrap procedures.

## 5.2 The Cox procedure

This procedure focuses on the loglikelihood ratio statistic, and in the case of the above regression models is given by (using the notations of Section 2):

$$LR_{fg} = L_f(\hat{\theta}_T) - L_g(\hat{\gamma}_T) = \frac{T}{2} \ln \left( \frac{\hat{\sigma}_T^2}{\hat{\omega}_T^2} \right),$$

where

$$\begin{aligned}\hat{\sigma}_T^2 &= T^{-1} \mathbf{e}'_f \mathbf{e}_f, \quad \hat{\alpha}_T = (\mathbf{X}' \mathbf{X})^{-1} \mathbf{X}' \mathbf{y}, \\ \mathbf{e}_f &= \mathbf{y} - \mathbf{X} \hat{\alpha}_T = \mathbf{M}_x \mathbf{y}, \quad \mathbf{M}_x = \mathbf{I}_T - \mathbf{X} (\mathbf{X}' \mathbf{X})^{-1} \mathbf{X}',\end{aligned}\tag{13.26}$$

and

$$\begin{aligned}\hat{\omega}_T^2 &= T^{-1} \mathbf{e}'_g \mathbf{e}_g, \quad \hat{\beta}_T = (\mathbf{Z}' \mathbf{Z})^{-1} \mathbf{Z}' \mathbf{y}, \\ \mathbf{e}_g &= \mathbf{y} - \mathbf{Z} \hat{\beta}_T = \mathbf{M}_z \mathbf{y}, \quad \mathbf{M}_z = \mathbf{I}_T - \mathbf{Z} (\mathbf{Z}' \mathbf{Z})^{-1} \mathbf{Z}.\end{aligned}\tag{13.27}$$

In the general case where the regression models are nonnested the average loglikelihood ratio statistic,  $\frac{1}{2} \ln(\hat{\sigma}_T^2 / \hat{\omega}_T^2)$ , does not converge to zero even if  $T$  is sufficiently large. For example, under  $H_f$  we have

$$p\lim_{T \rightarrow \infty} (T^{-1} LR_{fg} | H_f) = \frac{1}{2} \ln \left( \frac{\sigma_0^2}{\omega_*^2(\theta_0)} \right) = \frac{1}{2} \ln \left( \frac{\sigma_0^2}{\sigma_0^2 + \alpha'_0 \Sigma_f \alpha_0} \right),$$

and under  $H_g$ :

$$p\lim_{T \rightarrow \infty} (T^{-1} LR_{fg} | H_g) = \frac{1}{2} \ln \left( \frac{\sigma_*^2(\gamma_0)}{\omega_0^2} \right) = \frac{1}{2} \ln \left( \frac{\omega_0^2 + \beta'_0 \Sigma_g \beta_0}{\omega_0^2} \right).$$

The LR statistic is naturally centered at zero if one or the other of the above probability limits is equal to zero; namely if either  $\Sigma_f = 0$  or  $\Sigma_g = 0$ .<sup>14</sup> When  $\Sigma_f = 0$  then  $X \subset Z$  and  $H_f$  is nested in  $H_g$ . Alternatively, if  $\Sigma_g = 0$ , then  $Z \subset X$  and  $H_g$  is nested in  $H_f$ . Finally, if both  $\Sigma_f = 0$  and  $\Sigma_g = 0$  then the two regression models are observationally equivalent. In the nonnested case where both  $\Sigma_f \neq 0$  or  $\Sigma_g \neq 0$ , the standard LR statistic will not be applicable and needs to be properly centered. Cox's contribution was to note that this problem can be overcome if a consistent estimate of  $\text{Plim}_{T \rightarrow \infty} (T^{-1} LR_{fg} | H_f)$ , which we denote by  $\hat{E}_f(T^{-1} LR_{fg})$ , is subtracted from  $T^{-1} LR_{fg}$ , which yields the new centered (modified) loglikelihood ratio statistic (also known as the Cox statistic) for testing  $H_f$  against  $H_g$ :

$$S_{fg} = T^{-1} LR_{fg} - \hat{E}_f(T^{-1} LR_{fg}) \tag{13.28}$$

$$\begin{aligned}&= \frac{1}{2} \ln \left( \frac{\hat{\sigma}_T^2}{\hat{\omega}_T^2} \right) - \frac{1}{2} \ln \left( \frac{\sigma_T^2}{\hat{\sigma}_T^2 + \hat{\alpha}'_T \hat{\Sigma}_f \hat{\alpha}_T} \right) \\ &= \frac{1}{2} \ln \left( \frac{\hat{\sigma}_T^2 + \hat{\alpha}'_T \hat{\Sigma}_f \hat{\alpha}_T}{\hat{\omega}_T^2} \right).\end{aligned}\tag{13.29}$$

It is now clear that by construction the Cox statistic,  $S_{fg}$ , has asymptotically mean zero under  $H_f$ . As was pointed out earlier, since there is no natural null hypothesis in this setup, one also needs to consider the modified loglikelihood ratio statistic for testing  $H_g$  against  $H_f$  which is given by

$$S_{gf} = \frac{1}{2} \ln \left( \frac{\hat{\omega}_T^2 + \hat{\beta}'_T \hat{\Sigma}_g \hat{\beta}_T}{\hat{\sigma}_T^2} \right).$$

Both of these test statistics (when appropriately normalized by  $\sqrt{T}$ ) are asymptotically normally distributed under their respective nulls with a zero mean and finite variances. For the test of  $H_f$  against  $H_g$  we have<sup>15</sup>

$$\widehat{\text{Asyvar}}(\sqrt{T} S_{fg}) = V_{fg} = \frac{\hat{\sigma}_T^2 (\hat{\alpha}'_T X' M_z M_x M_z X \hat{\alpha}_T)}{T (\hat{\sigma}_T^2 + \hat{\alpha}'_T \hat{\Sigma}_f \hat{\alpha}_T)^2}.$$

The associated standardized Cox statistic is given by

$$N_{fg} = \frac{\sqrt{T} S_{fg}}{\sqrt{V_{fg}}} \stackrel{d}{\sim} N(0, 1). \quad (13.30)$$

By reversing the role of the null and the alternative hypothesis a similar standardized Cox statistic can be computed for testing  $H_g$  against  $H_f$ , which we denote by  $N_{gf}$ . Denote the  $(1 - \alpha)$  percent critical value of the standard normal distribution by  $C_\alpha$ , then four outcomes are possible:

1. Reject  $H_g$  but not  $H_f$  if  $|N_{fg}| < C_\alpha$  and  $|N_{gf}| \geq C_\alpha$ ,
2. Reject  $H_f$  but not  $H_g$  if  $|N_{fg}| \geq C_\alpha$  and  $|N_{gf}| < C_\alpha$ ,
3. Reject both  $H_f$  and  $H_g$  if  $|N_{fg}| \geq C_\alpha$  and  $|N_{gf}| \geq C_\alpha$ ,
4. Reject neither  $H_f$  nor  $H_g$  if  $|N_{fg}| < C_\alpha$  and  $|N_{gf}| < C_\alpha$ .

These are to be contrasted to the outcomes of the nested hypothesis testing where the null is either rejected or not, which stem from the fact that when the hypotheses under consideration are nonnested there is no natural null (or maintained) hypothesis and one therefore needs to consider in turn each of the hypotheses as the null. So there are twice as many possibilities as there are when the hypotheses are nested. Note that if we utilize the information in the *direction of rejection*, that is instead of comparing the *absolute* value of  $N_{fg}$  with  $C_\alpha$  we determine whether rejection is in the direction of the null or the alternative, there are a total of eight possible test outcomes (see the discussion in Fisher and McAleer (1979) and Dastoor (1981)). This aspect of nonnested hypothesis testing has been criticized by some commentators, pointing out the test outcome can lead to ambiguities. (See, for example, Granger, King, and White, 1995.) However, this is a valid criticism only if the primary objective is to *select* a specific model for forecasting or decision making, but not if the aim is to learn about the comparative strengths and weaknesses of rival explanations. What is viewed as a weakness from the perspective of model selection now becomes a strength when placed in the

context of statistical inference and model building. For example, when both models are rejected the analysis points the investigator in the direction of developing a third model which incorporates the main desirable features of the original, as well as being theoretically meaningful. (See Pesaran and Deaton, 1978.)

### 5.3 The comprehensive approach

Another approach closely related to the Cox's procedure is the comprehensive approach advocated by Atkinson (1970) whereby tests of nonnested models are based upon a third comprehensive model, artificially constructed so that each of the nonnested models can be obtained from it as special cases. Clearly, there are a large number of ways that such a comprehensive model can be constructed. A prominent example is the exponential mixture,  $H_\lambda$ , which in the case of the nonnested models (13.4) and (13.5) is defined by

$$H_\lambda : c_\lambda(y_t | x_t, z_t, \Omega_{t-1}; \theta, \gamma) = \frac{f(y_t | x_t, \Omega_{t-1}; \theta)^{1-\lambda} g(y_t | z_t, \Omega_{t-1}; \gamma)^\lambda}{\int_{\mathcal{R}_y} f(y_t | x_t, \Omega_{t-1}; \theta)^{1-\lambda} g(y_t | z_t, \Omega_{t-1}; \gamma)^\lambda dy_t},$$

where  $\mathcal{R}_y$  represents the domain of variations of  $y_t$ , and the integral in the denominator ensures that the combined function,  $c_\lambda(y_t | x_t, z_t, \Omega_{t-1}; \theta, \gamma)$ , is in fact a proper density function integrating to unity over  $\mathcal{R}_y$ . The "mixing" parameter  $\lambda$  varies in the range [0, 1] and represents the weight attached to model  $H_f$ . A test of  $\lambda = 0$  ( $\lambda = 1$ ) against the alternative that  $\lambda \neq 0$  ( $\lambda \neq 1$ ) can now be carried out using standard techniques from the literature on nested hypothesis testing. (See Atkinson, 1970 and Pesaran, 1982a.) This approach is, however, subject to three important limitations. First, although the testing framework is nested, the test of  $\lambda = 0$  is still *nonstandard* due to the fact that under  $\lambda = 0$  the parameters of the alternative hypothesis,  $\gamma$ , disappear. This is known as the Davies problem. (See Davies, 1977.) The same also applies if the interest is in testing  $\lambda = 1$ . The second limitation is due to the fact that testing  $\lambda = 0$  against  $\lambda \neq 0$ , is not equivalent to testing  $H_f$  against  $H_g$ , which is the problem of primary interest. This implicit change of the alternative hypothesis can have unfavorable consequences for the power of nonnested tests. Finally, the particular functional form used to combine the two models is arbitrary and does not allow identification of the mixing parameter,  $\lambda$ , even if  $\theta$  and  $\gamma$  are separately identified under  $H_f$  and  $H_g$  respectively. (See Pesaran, 1981.)

The application of the comprehensive approach to the linear regression models (13.19) and (13.20) yields:

$$H_\lambda : y = \left\{ \frac{(1-\lambda)v^2}{\sigma^2} \right\} X\alpha + \left\{ \frac{\lambda v^2}{\omega^2} \right\} Z\beta + u, \quad u \sim N(0, v^2 I_T), \quad (13.31)$$

where  $v^{-2} = (1 - \lambda)\sigma^{-2} + \lambda\omega^{-2}$ . It is clear that the mixing parameter  $\lambda$  is not identified.<sup>16</sup> In fact setting  $\kappa = \lambda v^2 / \omega^2$  the above "combined" regression can also be written as

$$H_\kappa : y = (1 - \kappa)X\alpha + \kappa Z\beta + u, \quad (13.32)$$

and a test of  $\lambda = 0$  in (13.31) can be carried by testing  $\kappa = 0$  in (13.32). Since the error variances  $\sigma^2$  and  $\omega^2$  are strictly positive  $\lambda = 0$  will be equivalent to testing  $\kappa = 0$ . The Davies problem, of course, continues to apply and under  $H_f(\kappa = 0)$  the coefficients of the rival model,  $\beta$ , disappear from the combined model. To resolve this problem Davies (1977) proposes a two-stage procedure. First, for a given value of  $\beta$  a statistic for testing  $\kappa = 0$  is chosen. In the present application this is given by the  $t$ -ratio of  $\kappa$  in the regression of  $y$  on  $X$  and  $y_\beta = Z\beta$ , namely

$$t_\kappa(Z\beta) = \frac{\beta'Z'M_xy}{\hat{v}(\beta'Z'M_xZ\beta)^{1/2}},$$

$$\hat{v}^2 = \frac{1}{T - k_f - 1} \left\{ y'M_xy - \frac{(\beta'Z'M_xy)^2}{\beta'Z'M_xZ\beta} \right\},$$

and where  $M_x$  is already defined by (13.26). In the second stage a test is constructed based on the entire random function of  $t_\kappa(Z\beta)$  viewed as a function of  $\beta$ . One possibility would be to construct a test statistic based on

$$F_\kappa = \max_{\beta} \{t_\kappa(Z\beta)\}.$$

Alternatively, a test statistic could be based on the average value of  $t_\kappa(Z\beta)$  obtained using a suitable prior distribution for  $\beta$ . Following the former classical route it is then easily seen that  $F_\kappa$  becomes the standard  $F_{z^*}$  statistic for testing  $b_2 = 0$ , in the regression

$$y = Xb_1 + Z^*b_2 + v_f, \quad (13.33)$$

where  $Z^*$  is the set of regressors in  $Z$  but not in  $X$ , namely  $Z^* = Z - X \cap Z$ .<sup>17</sup> Similarly for testing  $H_g$  against  $H_f$  the comprehensive approach involves testing  $c_1 = 0$ , in the combined regression

$$y = X^*c_1 + Zc_2 + v_g, \quad (13.34)$$

where  $X^*$  is the set of variables in  $X$  but not in  $Z$ . Denoting the F-statistic for testing  $c_1 = 0$  in this regression by  $F_{x^*}$ , notice that there are still four possible outcomes to this procedure; in line with the ones detailed above for the Cox test. This is because we have two F-statistics,  $F_{x^*}$  and  $F_{z^*}$ , with the possibility of rejecting both hypotheses, rejecting neither, etc.

An altogether different approach to the resolution of the Davies problem would be to replace the regression coefficients,  $\beta$ , in (13.32) by an estimate, say  $\tilde{\beta}$ , and then proceed as if  $\tilde{y}_\beta = Z\tilde{\beta}$  is data. This is in effect what is proposed by Davidson and MacKinnon (1981) and Fisher and McAleer (1981). Davidson and MacKinnon suggest using the estimate of  $\beta$  under  $H_g$ , namely  $\hat{\beta}_T = (Z'Z)^{-1}Zy$ . This leads to the J-test which is the standard  $t$ -ratio of the estimate of  $\kappa$  in the artificial regression<sup>18</sup>

$$H_\kappa : y = X\alpha + \kappa Z\hat{\beta}_T + v_\kappa. \quad (13.35)$$

For testing  $H_g$  against  $H_f$ , the  $J$ -test will be based on the OLS regression of  $y$  on  $Z$  and  $X\hat{\alpha}_T$ , and the  $J$ -statistic is the  $t$ -ratio of the coefficient of  $X\hat{\alpha}_T$  (which is the vector of fitted values under  $H_f$ ) in this regression.

The test proposed by Fisher and McAleer (known as the  $JA$ -test) replaces  $\beta$  by the estimate of its pseudo-true value under  $H_f$ , given by  $\beta_*(\hat{\alpha}_T)$

$$\hat{\beta}_*(\hat{\alpha}_T) = (Z'Z)^{-1}Z'\hat{\alpha}_T.$$

In short the  $JA$ -test of  $H_f$  against  $H_g$  is the  $t$ -ratio of the coefficient of  $\hat{y}_{\beta\alpha} = Z(Z'Z)^{-1}Z'\hat{\alpha}_T$  in the OLS regression of  $y$  on  $X$  and  $\hat{y}_{\beta\alpha}$ . Similarly, a  $JA$ -test of  $H_g$  against  $H_f$  can be computed.

Both the  $J$ - and the  $JA$ -test statistics, as well as their various variations proposed in the literature can also be derived as linear approximations to the Cox test statistic. See (13.28).

Various extensions of nonnested hypothesis testing have also appeared in the literature. These include tests of nonnested linear regression models with serially correlated errors (McAleer *et al.*, 1990); models estimated by instrumental variables (Ericsson, 1983; Godfrey, 1983); models estimated by the generalized method of moments (Smith, 1992); nonnested Euler equations (Ghysels and Hall, 1990); autoregressive versus moving average models (Walker, 1967; King, 1983); the generalized autoregressive conditional heteroskedastic (GARCH) model against the exponential-GARCH model (McAleer and Ling, 1998); linear versus loglinear models (Aneuryn-Evans and Deaton, 1980; Davidson and MacKinnon, 1985; Pesaran and Pesaran, 1995); logit and probit models (Pesaran and Pesaran, 1993; Weeks, 1996; Duncan and Weeks, 1998); nonnested threshold autoregressive models (Altissimo and Violante, 1998; Pesaran and Potter, 1997; Kapetanios and Weeks, 1999).

## 5.4 The encompassing approach

This approach generalizes Cox's original idea and asks whether model  $H_f$  can explain one or more features of the rival model  $H_g$ . When *all* the features of model  $H_g$  can be explained by model  $H_f$  it is said that model  $H_f$  *encompasses* model  $H_g$ ; likewise model  $H_g$  is said to encompass model  $H_f$  if all the features of model  $H_f$  can be explained by model  $H_g$ . A formal definition of encompassing can be given in terms of the pseudo-true parameters and the binding functions defined in Section 2.

Model  $H_g$  is said to encompass model  $H_f$ , respectively defined by (13.5) and (13.4), if and only if

$$H_g \mathcal{E} H_f : \theta_{h*} = \theta_*(\gamma_{h*}). \quad (13.36)$$

Similarly,  $H_f$  is said to encompass  $H_g$  (or  $H_g$  is encompassed by  $H_f$ ) if and only if

$$H_f \mathcal{E} H_g : \gamma_{h*} = \gamma_*(\theta_{h*}).$$

Recall that  $\theta_{h*}$  and  $\gamma_{h*}$  are the pseudo-true values of  $\theta$ , and  $\gamma$  with respect to the true model  $H_h$ , and  $\theta_*(\cdot)$  are  $\gamma_*(\cdot)$  are the binding functions linking the parameters

of the models  $H_f$  and  $H_g$ . For example, in the case of the linear rival regression models (13.10) and (13.11), and assuming that the true model is given by (13.14) then it is easily seen that the functions that bind the parameters of model  $H_g$  to that of  $H_f$  are

$$\theta_*(\gamma_{h*}) = \begin{pmatrix} \hat{\Sigma}_{xx}^{-1} \hat{\Sigma}_{xz} \beta_{h*} \\ \omega_{h*}^2 + \beta'_{h*} (\hat{\Sigma}_{zz} - \hat{\Sigma}_{zx} \hat{\Sigma}_{xx}^{-1} \hat{\Sigma}_{xz}) \beta_{h*} \end{pmatrix}.$$

Using (13.16) to substitute for the pseudo-true values  $\beta_{h*}$  and  $\omega_{h*}^2$  we have

$$\theta_*(\gamma_{h*}) = \begin{pmatrix} \hat{\Sigma}_{xx}^{-1} \hat{\Sigma}_{xz} \hat{\Sigma}_{zz}^{-1} \hat{\Sigma}_{zw} \delta \\ v^2 + \delta' (\hat{\Sigma}_{ww} - \hat{\Sigma}_{wz} \hat{\Sigma}_{zz}^{-1} \hat{\Sigma}_{zw}) \delta + \delta' \hat{\Sigma}_{wz} \hat{\Sigma}_{zz}^{-1} (\hat{\Sigma}_{zz} - \hat{\Sigma}_{zx} \hat{\Sigma}_{xx}^{-1} \hat{\Sigma}_{xz}) \hat{\Sigma}_{zz}^{-1} \hat{\Sigma}_{zw} \delta \end{pmatrix}.$$

Therefore, conditional on the observation matrices  $X$ ,  $Z$ , and  $W$ , model  $H_f$  encompasses model  $H_g$  if and only if

$$\begin{aligned} & \begin{pmatrix} \hat{\Sigma}_{xx}^{-1} \hat{\Sigma}_{xz} \delta \\ v^2 + \delta' (\hat{\Sigma}_{ww} - \hat{\Sigma}_{wx} \hat{\Sigma}_{xx}^{-1} \hat{\Sigma}_{xw}) \delta \end{pmatrix} \\ &= \begin{pmatrix} \hat{\Sigma}_{xx}^{-1} \hat{\Sigma}_{xz} \hat{\Sigma}_{zz}^{-1} \hat{\Sigma}_{zw} \delta \\ v^2 + \delta' (\hat{\Sigma}_{ww} - \hat{\Sigma}_{wz} \hat{\Sigma}_{zz}^{-1} \hat{\Sigma}_{zw}) \delta + \delta' \hat{\Sigma}_{wz} \hat{\Sigma}_{zz}^{-1} (\hat{\Sigma}_{zz} - \hat{\Sigma}_{zx} \hat{\Sigma}_{xx}^{-1} \hat{\Sigma}_{xz}) \hat{\Sigma}_{zz}^{-1} \hat{\Sigma}_{zw} \delta \end{pmatrix}. \end{aligned}$$

These conditions are simplified to

$$\hat{\Sigma}_{xw} \delta = \hat{\Sigma}_{xz} \hat{\Sigma}_{zz}^{-1} \hat{\Sigma}_{zw} \delta, \quad (13.37)$$

and

$$\delta' \hat{\Sigma}_{wx} \hat{\Sigma}_{xx}^{-1} \hat{\Sigma}_{xw} \delta = \delta' \hat{\Sigma}_{wz} \hat{\Sigma}_{zz}^{-1} \hat{\Sigma}_{zx} \hat{\Sigma}_{xx}^{-1} \hat{\Sigma}_{xz} \hat{\Sigma}_{zz}^{-1} \hat{\Sigma}_{zw} \delta. \quad (13.38)$$

But it is easily verified that (13.37) implies (13.38), namely encompassing with respect to the regression coefficients imply encompassing with respect to the error variances. Therefore,  $H_f$  is encompassed by  $H_g$  if and only if  $(X'M_z W)\delta = 0$ . This condition is clearly satisfied if either  $H_f$  is nested within  $H_g$ , ( $X'M_z = 0$ ), or if  $H_g$  contains the true model, ( $M_z W = 0$ ). The remaining possibility, namely when  $(X'M_z W) = 0$ , but the true value of  $\delta$ , say  $\delta_0$ , is such that  $(X'M_z W)\delta_0 = 0$ , is a rather low probability event.

The encompassing hypothesis,  $H_g \mathcal{E} H_f$ , (or  $H_f \mathcal{E} H_g$ ) can now be tested using the encompassing statistics,  $\sqrt{T} [\hat{\theta}_T - \theta_*(\hat{\gamma}_T)]$ , (or  $\sqrt{T} [\hat{\gamma}_T - \gamma_*(\hat{\theta}_T)]$ ). Gouriéroux and Monfort (1995) show that under the encompassing hypothesis,  $\theta_{h*} = \theta_*(\gamma_{h*})$ , and assuming certain regularity conditions are met,  $\sqrt{T} [\hat{\theta}_T - \theta_*(\hat{\gamma}_T)]$  is asymptotically normally distributed with zero means and a variance covariance matrix that in general depends in a complicated way on the probability density functions of the rival models under consideration. Complications arise since  $H_g$  need not belong to  $H_h$ . Two testing procedures are proposed, the Wald encompassing test

(WET) and the score encompassing test (SET), both being difficult to implement. First, the binding functions  $\theta_*(\cdot)$  and  $\gamma_*(\cdot)$  are not always easy to derive. (But this problem also afflicts the implementation of the Cox procedure, see below.) Second, and more importantly, the variance–covariance matrices of  $\sqrt{T}[\hat{\theta}_T - \theta_*(\hat{\theta}_T)]$ , (or  $\sqrt{T}[\hat{\gamma}_T - \gamma_*(\hat{\theta}_T)]$ ), are, in general, non-invertible and the construction of WET and SET statistics involve the use of generalized inverse and this in turn requires estimation of the rank of these covariance matrices. Alternative ways of dealing with these difficulties are considered in Gouriéroux and Monfort (1995) and Smith (1993).

In the case of linear regression models full parameter encompassing (namely an encompassing exercise involving both regression coefficients and error variances) is unnecessary.<sup>19</sup> Focusing on regression coefficients the encompassing statistics for testing  $H_g \not\subseteq H_f$  are given by

$$\sqrt{T}[\hat{\alpha}_T - \alpha_*(\hat{\beta}_T)] = \sqrt{T}(X'X)^{-1}X'M_zy.$$

Under  $H_h$ , defined by (13.14),

$$\sqrt{T}[\hat{\alpha}_T - \alpha_*(\hat{\beta}_T)] = \sqrt{T}(X'X)^{-1}(X'M_zW)\delta + \sqrt{T}(X'X)^{-1}X'M_zu_h,$$

where  $u_h \sim N(0, v^2I_T)$ .<sup>20</sup> Hence, under the encompassing hypothesis,  $(X'M_zW)\delta = 0$ , the encompassing statistic  $\sqrt{T}[\hat{\alpha}_T - \alpha_*(\hat{\beta}_T)]$  is asymptotically normally distributed with mean zero and the covariance matrix  $v^2\Sigma_{xx}^{-1}(\Sigma_{xx} - \Sigma_{xz}\Sigma_{zz}^{-1}\Sigma_{zx})\Sigma_{xx}^{-1}$ . Therefore, the construction of a standardized encompassing test statistic requires a consistent estimate of  $v^2$ , the error variance of the true regression model, and this does not seem possible without further assumptions about the nature of the true model. In the literature it is often (implicitly) assumed that the true model is contained in the union intersection of the rival models under consideration (namely  $W \equiv X \cup Z$ ) and  $v^2$  is then consistently estimated from a regression of  $y$  on  $X \cup Z$ . Under this additional assumption, the WET statistic for testing  $H_g \not\subseteq H_f$ , is given by

$$\mathcal{E}_{gf} = \frac{y'M_zX(X'M_z\bar{X})X'M_zy}{\hat{v}^2},$$

where  $\hat{v}^2$  is the estimate of the error variance of the regression of  $y$  on  $X \cup Z$ , and  $(X'M_z\bar{X})$  is a generalized inverse of  $X'M_zX$ . This matrix is rank deficient whenever  $X$  and  $Z$  have variables in common, namely if  $X \cap Z = Q \neq 0$ . Let  $X = (X_1, Q)$  and  $Z = (Z_1, Q)$ , then

$$X'M_zX = \begin{pmatrix} X_1'M_zX_1 & 0 \\ 0 & 0 \end{pmatrix}.$$

But it is easily seen that  $\mathcal{E}_{gf}$  is invariant to the choice of the g-inverse used and is given by

$$\mathcal{E}_{gf} = \frac{y'M_zX_1(X_1'M_zX_1)^{-1}X_1'M_zy}{\hat{v}^2},$$

and is identical to the standard Wald statistic for testing the statistical significance of  $X_1$  in the OLS regression of  $y$  on  $Z$  and  $X_1$ . This is perhaps not surprising, considering the (implicit) assumption concerning the true model being a union intersection of the rival regression models  $H_f$  and  $H_g$ .

Other encompassing tests can also be developed depending on the parameters of interest or their functions. For example, a *variance* encompassing test of  $H_g \not\subseteq H_f$  compares a consistent estimate of  $\sigma^2$  with that of its pseudo-true value  $\sigma_{h*}^2$ , namely  $\hat{\sigma}_T^2 - \sigma_*^2(\hat{\gamma}_T) = \hat{\sigma}_T^2 - [\hat{\omega}_T^2 + T^{-1}\hat{\beta}_T'Z'M_xZ\hat{\beta}_T]$ .<sup>21</sup> Under the encompassing hypothesis this statistic tends to zero, but its asymptotic distribution in general depends on  $H_h$ . In the case where  $H_g$  contains the true model the variance encompassing test will become asymptotically equivalent to the Cox and the *J*-tests discussed above.

The encompassing approach can also be applied to the loglikelihood functions. For example, to test  $H_g \not\subseteq H_f$  one could use the encompassing loglikelihood ratio statistic  $T^{-1}\{L_f(\hat{\theta}_T) - L_f(\theta_*(\hat{\gamma}_T))\}$ . This test can also be motivated using Cox's idea of a centered loglikelihood ratio statistic, with the difference that the centering is now carried out under  $H_h$  rather than under  $H_g$  (or  $H_f$ ). See Gouriéroux and Monfort (1995) and Smith (1993) for details and difficulties involved in their implementation. Other relevant literature include Dastoor (1983), Gouriéroux *et al.* (1983) and Mizon and Richard (1986).

## 5.5 Power and finite sample properties

A number of studies have examined the small sample properties of nonnested tests. For a limited number of cases it is possible to determine the exact form of the test statistic and the sampling distribution. For example, Godfrey (1983) shows that under  $H_f$  if  $X$  and  $Z$  are non-stochastic with normal errors, then the *JA*-test has an exact  $t(T - k_f - 1)$  distribution.<sup>22</sup> In the majority of cases the finite sample properties have been examined using Monte Carlo studies. A recurrent finding is that many Cox-type tests for nonnested regression models have a finite sample size which is significantly greater than the nominal level. Modifications based upon mean and variance adjustments have been proposed in Godfrey and Pesaran (1983), and are shown to affect a substantial improvement in finite sample performance. The authors demonstrate that in experimental designs allowing for nonnested models with either nonnormal errors, different number of regressors, or a lagged dependent variable, the adjusted Cox-test performs favorably relative to the *J*-test or *F*-test.<sup>23</sup> In the case of nonnested linear regression models, Davidson and MacKinnon (1982) compared a number of variants of the Cox-test with *F*-, *JA*- and *J*-test.

An analysis of the power properties of non-tested tests has been undertaken using a number of approaches. In the case of nested models local alternatives are readily defined in terms of parameters that link the null to the alternative. Obviously in the case of models that are globally nonnested (i.e. the exponential and lognormal) this procedure is not possible. In the case of regression models Pesaran (1982a) is able to develop a asymptotic distribution of Cox-type tests under a sequence of local alternatives defined in terms of the degree of multicollinearity of the regressors from the two rival models. Under this sequence of local

alternatives he shows that the F-test based on the comprehensive model is less powerful than the Cox-type tests, unless the number of non-overlapping variables of the alternative over the null hypothesis is unity. An alternative approach to asymptotic power comparisons which does not require specification of local alternatives is advanced by Bahadur (1960) and Bahadur (1967) and holds the alternative hypothesis fixed but allows the size of the test to tend to zero as the sample size increases. Asymptotic power comparisons of nonnested tests by the Bahadur approach is considered in Gouriéroux (1982) and Pesaran (1984).

## 6 MEASURES OF CLOSENESS AND VUONG'S APPROACH

So far the concepts of nested and nonnested hypotheses have been loosely defined, but for a more integrated approach to nonnested hypothesis testing and model selection a more formal definition is required. This can be done by means of a variety of "closeness" criteria proposed in the literature for measuring the divergence of one distribution function with respect to another. A popular measure employed in Pesaran (1987) for this purpose is the Kullback–Leibler (Kullback, 1959) information criterion (KLIC). This criterion has been used extensively in the development of both nonnested hypotheses tests and model selection procedures. Given hypotheses  $H_f$  and  $H_g$ , defined by (13.4) and (13.5), the KLIC measure of  $H_g$  with respect to  $H_f$  is written as

$$\begin{aligned} I_{fg}(\theta, \gamma) &= E_f\{\ln f_t(\theta) - \ln g_t(\gamma)\} \\ &\quad \int_{R_f} \ln \left\{ \frac{f_t(\theta)}{g_t(\gamma)} \right\} f_t(\theta) dy. \end{aligned} \quad (13.39)$$

It is important to note that  $I_{fg}(\theta, \gamma)$  is not a distance measure. For example, in general  $I_{fg}(\theta, \gamma)$  is not the same as  $I_{gf}(\gamma, \theta)$ , and KLIC does not satisfy the triangular inequality, namely  $I_{fg} + I_{gh}$  need not exceed  $I_{fh}$  as required if KLIC were a distance measure. Nevertheless, KLIC has a number attractive properties:  $I_{fg}(\theta, \gamma) \geq 0$ , with the strict equality holding if and only if  $f(\cdot) = g(\cdot)$ . Assuming that observations on  $y_t$  are independently distributed then the KLIC measure is additive over sample observations.

To provide a formal definition of nonnested or nested hypothesis we define two "closeness" measures: one measuring the closeness of  $H_g$  to  $H_f$  (viewed from the perspective of  $H_f$ ), and another the closeness measure of  $H_f$  to  $H_g$ . These are respectively given by  $C_{fg}(\theta_0) = I_{fg}(\theta_0, \gamma_*(\theta_0))$ , and  $C_{gf}(\gamma_0) = I_{gf}(\gamma_0, \theta_*(\gamma_0))$ , where, as before,  $\gamma_*(\theta_0)$  is the pseudo-true value of  $\gamma$  under  $H_f$ , and  $\theta_*(\gamma_0)$  is pseudo-true value of  $\theta$  under  $H_g$ .

**Definition 1.**  $H_f$  is nested within  $H_g$  if and only if  $C_{fg}(\theta_0) = 0$ , for all values of  $\theta_0 \in \Theta$ , and  $C_{gf}(\gamma_0) \neq 0$  for some  $\gamma_0 \in \Gamma$ .

**Definition 2.**  $H_f$  and  $H_g$  are globally nonnested if and only if  $C_{fg}(\theta_0)$  and  $C_{gf}(\gamma_0)$  are both non-zero for all values of  $\theta_0 \in \Theta$  and  $\gamma_0 \in \Gamma$ .

**Definition 3.**  $H_f$  and  $H_g$  are partially nonnested if  $C_{fg}(\theta_0)$  and  $C_{gf}(\gamma_0)$  are both non-zero for some values of  $\theta_0 \in \Theta$  and  $\gamma_0 \in \Gamma$ .

**Definition 4.**  $H_f$  and  $H_g$  are observationally equivalent if and only if  $C_{fg}(\theta_0) = 0$  and  $C_{gf}(\gamma_0) = 0$  for all values of  $\theta_0 \in \Theta$  and  $\gamma_0 \in \Gamma$ .

Using the above definitions it is easily seen, for example, that linear or nonlinear rival regression models can at most be partially nonnested, but exponential and lognormal distributions discussed in Section 3 are globally nonnested. For further details see Pesaran (1987).

We may also define a closeness measure of  $H_g$  to  $H_f$  from the perspective of the true model  $H_h$  and in doing so are able to motivate Vuong's approach to hypothesis testing and model selection. (See Vuong, 1989.) The primary focus of Vuong's analysis is to test the hypothesis that the models under consideration are "equally" close to the true model. As Vuong (1989) notes "If the distance between a specified model and the true distribution is defined as the minimum of the KLIC over the distributions in the model, then it is natural to define the 'best' model among a collection of competing models to be the model that is closest to the true distribution". Thus, in contrast to the standard approach to model selection, a hypothesis testing framework is adopted and a *probabilistic* decision rule used to select a "best" model.

With our setup and notations the closeness of  $H_f$  to  $H_h$  viewed from the perspective of the true model,  $H_h$  is defined by

$$C_{hf}(\theta_{h*}) = E_h\{\ln h_t(\cdot) - \ln f_t(\theta_{h*})\}.$$

Similarly, the closeness of  $H_g$  to  $H_h$  is defined by

$$C_{hg}(\gamma_{h*}) = E_h\{\ln h_t(\cdot) - \ln g_t(\gamma_{h*})\}.$$

The null hypothesis underlying Vuong's approach is now given by

$$H_V : C_{hf}(\theta_{h*}) = C_{hg}(\gamma_{h*}),$$

which can also be written as

$$H_V : E_h\{\ln f_t(\theta_{h*})\} = E_h\{\ln g_t(\gamma_{h*})\}.$$

The quantity  $E_h\{\ln f_t(\theta_{h*}) - \ln g_t(\gamma_{h*})\}$  is unknown and depends on the unknown true distribution  $H_h$ , but can be consistently estimated by the average loglikelihood ratio statistic,  $T^{-1}\{L_f(\hat{\theta}_T) - L_g(\hat{\gamma}_T)\}$ . Vuong derives the asymptotic distribution of the average loglikelihood ratio under  $H_V$ , and shows that it crucially depends on whether  $f_t(\theta_{h*}) = g_t(\gamma_{h*})$ , namely whether the distributions in  $H_f$  and  $H_g$  that are closest to the true model are observationally equivalent or not. In view of this a sequential approach to hypothesis testing is proposed. See Vuong (1989) for further details.

## 7 PRACTICAL PROBLEMS

In Section 5 we noted that the motivation for the Cox test statistic was based upon the observation that unless two models, say  $f(\cdot)$  and  $g(\cdot)$  are nonnested then the expectation

$$T^{-1}E_f[L_f(\theta) - L_g(\gamma)], \quad (13.40)$$

does not evaluate to zero and as a result standard likelihood ratio statistics are not appropriate. Cox (1961, 1962) proposed a procedure such that a centered (modified) loglikelihood ratio has a well-defined limiting distribution. In Section 5.1 we demonstrated that in the case of the linear regression we may obtain a closed form consistent estimate of (13.40). However, this is the exception rather than the rule and the use of the Cox test has been restricted to a relatively small number of applications due to problems in constructing a consistent estimate of the expected loglikelihood ratio statistic. There are two principal problems. First, in order to estimate (13.40) we require a consistent estimate of the pseudo-true value,  $\gamma(\theta_0)$ . Second, in most cases even given such an estimate, the expectation (13.40) will still be intractable. An exception is the application of the Cox test to both binary and multinomial probit and logit models. Independent of the dimension of the choice set, the expected difference between the two loglikelihoods under the null has a relatively simple, closed form expression (see Pesaran and Pesaran, 1993).

Following the work of Pesaran and Pesaran (1993, 1995) and Weeks (1996), a simulation-based application of the modified likelihood principle has been used to affect adjustments to the test statistic in order to improve the finite sample size and power properties. A drawback of this approach is that it is still reliant upon a reference distribution which is valid asymptotically. In addition, Orme (1994) attests to the existence of a large number of asymptotically equivalent (AE) variants of the Cox test statistic which represents a formidable menu of choices for the applied econometrician. In the case of the numerator, various test statistics are based upon the use of alternative consistent estimators of the Kullback–Leibler measure of closeness. An additional set of variants of the Cox test statistic depends upon the existence of a number of AE ways of estimating the variance of the test statistic.

An alternative approach based upon the seminal work of Efron (1979), with contributions by Hall (1986), Beran (1988), Hinkely (1988), and Coulibaly and Brorsen (1998), applies bootstrap-based procedures to directly evaluate the empirical distribution function of the loglikelihood ratio statistic. In this context the focus is upon correcting the reference distribution rather than centering the loglikelihood ratio statistic and utilizing limiting distribution arguments. This type of adjustment may, in a number of cases, be theoretically justified through Edgeworth expansions and can under certain conditions result in improvements over classical asymptotic inference. The existence of a large menu of broadly equivalent test statistics is also relevant in the context of bootstrap-based inference. Recent surveys by Vinod (1993), Jeong and Maddala (1993), and Li and

Maddala (1996), review a large number of variants including the double, recursive, and weighted bootstrap. Similarly, Hall (1988) notes that in many applications the precise nature of the bootstrap design is not stated.

## 7.1 A simulation application of the modified likelihood principle

The essence of the Cox nonnested test is that the mean adjusted ratio of the maximized loglikelihoods of two nonnested models has a well-defined *limiting* distribution under the null hypothesis. Using the notation set out in Section 2 above we may write the numerator of the Cox test statistic as

$$S_{fg} = T^{-1}LR_{fg} - C_{fg}(\hat{\theta}_T, \tilde{\gamma}). \quad (13.41)$$

The last term on the right-hand side of (13.41),  $C_{fg}(\hat{\theta}_T, \tilde{\gamma})$ , represents a consistent estimator of  $C_{fg}(\theta_0, \gamma_*(\theta_0))$ , the KLIC measure of closeness of  $g(\cdot)$  to  $f(\cdot)$ . This may be written as  $C_{fg}(\hat{\theta}_T, \tilde{\gamma}) = \hat{E}_f[T^{-1}(L_f(\hat{\theta}_T) - L_g(\tilde{\gamma}))]$ , and is an estimator of the difference between the expected value of the two maximized loglikelihoods under the distribution given by  $f(\cdot)$ ;  $\tilde{\gamma}$  is any consistent estimator for  $\gamma_*(\theta_0)$ . Weeks (1996), in testing probit and logit models of discrete choice, distinguished between three variants,  $\tilde{\gamma} = \{\hat{\gamma}_T, \gamma_R(\hat{\theta}_T), \tilde{\gamma}_T\}$ .  $\hat{\gamma}_T$  is the MLE of  $\gamma$ ,  $\tilde{\gamma}$  is due to Kent (1986) and is an estimator derived from maximizing the fitted loglikelihood, and  $\gamma_{*R}(\hat{\theta}_T) = \frac{1}{R} \sum_{r=1}^R \gamma_r^*(\hat{\theta}_T)$  is a simulation-based estimator where  $\gamma_r^*(\hat{\theta}_T)$  is the solution to

$$\arg \max_{\gamma} \left\{ L_g^r(\gamma) = \sum_{t=1}^T \ln g(y_t^r(\hat{\theta}_T) | z_t, \Omega_{t-1}; \gamma) \right\}, \quad (13.42)$$

where  $y_t^r(\hat{\theta}_T)$  is the  $r$ th draw of  $y_t$  under  $H_f$  using  $\hat{\theta}_T$  and  $R$  is the number of simulations. Note that for both  $R \rightarrow \infty$  and  $T \rightarrow \infty$  then  $\gamma_{*R}(\hat{\theta}_T) \rightarrow \gamma_*(\theta_0)$ .

A simulation-based estimator of  $C_{fg}(\theta_0, \gamma_*(\theta_0))$  has been suggested by Pesaran and Pesaran (1993) and is given by

$$C_{fg,R}(\hat{\theta}_T, \gamma_{*R}(\hat{\theta}_T)) = \frac{1}{TR} \sum_{r=1}^R [L_f^r(\hat{\theta}_T) - L_g^r(\gamma_{*R}(\hat{\theta}_T))]. \quad (13.43)$$

However (13.43) represents one approach to centering the loglikelihood ratio statistic, whereby both  $\hat{\theta}_T$  and  $\gamma_{*R}(\hat{\theta}_T)$  are treated as *fixed* parameters. An alternative method of mean adjustment is given by the following estimator of KLIC

$$C_{fg,R}(\hat{\theta}_T^1, \dots, \hat{\theta}_T^R, \gamma_*^1(\hat{\theta}_T), \dots, \gamma_*^R(\hat{\theta}_T)) = \frac{1}{TR} \sum_{r=1}^R [L_f^r(\hat{\theta}_T^r) - L_g^r(\gamma_*^r(\hat{\theta}_T))], \quad (13.44)$$

where the parameter arguments to both  $L_f(\cdot)$  and  $L_g(\cdot)$  are allowed to *vary* across each  $r$ th replication. (See Coulibaly and Brorsen, 1998.)

## 7.2 Resampling the likelihood ratio statistic: bootstrap methods

The bootstrap is a data based simulation method for statistical inference. The bootstrap approach involves approximating the distribution of a function of the observed data by the bootstrap distribution of the quantity. This is done by substituting the empirical distribution function for the unknown distribution and repeating this process many times to obtain a simulated distribution. Its recent development follows from the requirement of a significant amount of computational power. Obviously there is no advantage to utilizing bootstrap procedures when the exact sampling distribution of the test statistic is known. However, it has been demonstrated that when the sampling distribution is not known, the substitution of computational intensive bootstrap resampling can offer an improvement over asymptotic theory. The use of *non-pivotal* bootstrap testing procedures does not require the mean adjustment facilitated by (13.43) and (13.44). However, pivotal (or bootstrap-t) procedures require both mean and variance adjustments in order to guarantee asymptotic pivotalness.

Utilizing a parametric bootstrap we present below a simple algorithm for resampling the likelihood ratio statistic which we then use to construct the empirical distribution function of the test statistic. For the purpose of exposition the algorithm is presented for the non-pivotal bootstrap.

1. Generate  $R$  samples of size  $T$  by sampling from the *fitted* null model  $f_t(\hat{\theta}_T)$ .
2. For each  $r$ th simulated sample, the pair  $(\hat{\theta}_T^r, \gamma_*^r(\hat{\theta}_T))$  represent the parameter estimates obtained by maximizing the loglikelihoods

$$L_f^r(\theta) = \sum_{t=1}^T \ln f_t(y_t^r(\hat{\theta}_T) | x_t, \Omega_{t-1}; \theta), \quad L_g^r(\gamma) = \sum_{t=1}^T \ln g_t(y_t^r(\hat{\theta}_T) | z_t, \Omega_{t-1}; \gamma), \quad (13.45)$$

where  $y_t^r(\hat{\theta}_T)$  denotes the  $r$ th bootstrap-sample conditional upon  $\theta = \hat{\theta}_T$ . We then compute the simulated loglikelihood ratio statistic

$$T_f^r = L_f(\hat{\theta}_T^r) - L_g(\gamma_*^r(\hat{\theta}_T)).$$

3. By constructing the empirical cdf of  $\{T_f^r : 1 \leq r \leq R\}$ , we can compare the *observed* test statistic,  $T_f = L_f(\hat{\theta}_T) - L_g(\gamma_*(\hat{\theta}_T))$ , with critical values obtained from the  $R$  independent (conditional) realizations of  $T_f^r$ . The  $p$ -value based upon the bootstrap procedure is given by<sup>24</sup>

$$P_R = \frac{\sum_{r=1}^R 1(T_f^r \geq T_f)}{R}, \quad (13.46)$$

where  $1(\cdot)$  is the indicator function.

The bootstrap procedure outlined above simply resamples the likelihood ratio statistic *without* pivoting. There are a number of alternative test statistics which by using pivotal methods are conjectured to represent an improvement over classical first order methods (see for example, Beran (1988) and Hall (1988)). An evaluation of both the size and power properties of a number of simulation and bootstrap-based tests applied to linear versus loglinear regression models and a number of variants of threshold autoregressive models is provided in Kapetanios and Weeks (1999).

## Notes

- 1 Therefore our focus is distinct from Chow (1981) who, in examining a similar problem, assumes that the set of models under consideration contains a general model from which all other competing models may be obtained by the imposition of suitable parameter restrictions.
- 2 See, for example, Friedman and Meiselman (1963) on alternative consumption models, Barro (1977), Pesaran (1982b) and McAleer, Pesaran, and Bera (1990) on alternative explanations of the unemployment rate; Jorgenson and Siebert (1968), Dixit and Pindyck (1994) and Bernanke, Bohn, and Reiss (1988) on alternative models of investment behavior; and McAleer, Fisher, and Volker (1982) and Smith and Smyth (1991) on nonnested money demand functions.
- 3 An excellent survey article on nonnested hypothesis testing can be found in Gouriéroux and Monfort (1994).
- 4 In cases where one or more elements of  $z_t$  are discrete, as in probit or tobit specifications cumulative probability distribution functions can be used instead of probability density functions.
- 5 See Engle, Hendry, and Richard (1983).
- 6 A formalization of the concept of globally nonnested models can be found in Pesaran (1987). Also see Section 6.
- 7 Note that under  $H_f$ ,  $E(y_i) = E\{\exp(\ln y_i)\} = \exp(\theta_1 + 0.5\theta_2)$ .
- 8 See, Pesaran (1984, pp. 249–50).
- 9 There is substantial literature on nonnested tests of linear versus loglinear regression models. Earlier studies include Aneuryn-Evans and Deaton (1980), Godfrey and Wickens (1981) and Davidson and MacKinnon (1985). In a more recent study Pesaran and Pesaran (1995) have examined the properties of a simulation-based variant of the Cox test.
- 10 A review of the model selection literature is beyond the scope of the present paper. See, for example, Leamer (1983) for an excellent review. A recent review focusing upon the selection of regressors problems is to be found in Lavergne (1998). Two excellent texts are Grasa (1989) and Linhart and Zucchini (1986). Maddala (1981) edited a special issue of the *Journal of Econometrics* which focuses on model selection.
- 11 For the case of the classical linear regression model examples of model selection criteria include Theil's  $\bar{R}^2$ , with more general loss functions based upon information criteria including Akaike's (1973) information criteria and Schwarz's (1978) Bayesian information criterion.
- 12 The concepts of globally and partially nonnested models are defined in Pesaran (1987).
- 13 See also Dastoor (1981) for further discussion.
- 14 The cases where  $\Sigma_f \neq 0$  (respectively  $\Sigma_g \neq 0$ ) but nevertheless  $\Sigma_f \alpha_0 = 0$  (respectively  $\Sigma_g \beta_0 = 0$ ) are discussed in Pesaran (1987, p. 74).

- 15 See Pesaran (1974) for details of the derivations.
- 16 For example, it is not possible to test whether  $\lambda = 1/2$ , which could have been of interest in assessing the relative weights attached to the two rival models.
- 17 For a proof see McAleer and Pesaran (1986).
- 18 Chao and Swanson (1997) provide some asymptotic results for the  $J$ -test in the case of nonnested models with  $I(1)$  regressors.
- 19 Recall that the encompassing condition (13.37) for the regression coefficients implies the condition (13.38) for error variance encompassing but not vice versa.
- 20 Notice that the normality assumption is not needed and can be relaxed.
- 21 Similarly, the variance encompassing statistic for testing  $H_f \not\equiv H_g$  is given by  $\hat{\omega}_T^2 - [\hat{\omega}_T^2 + T^{-1} \hat{\alpha}'_T X' M_z X \hat{\alpha}_T]$ .
- 22 See also McAleer (1983).
- 23 See McAleer and Pesaran (1986) for additional details.
- 24 If  $T$  is discrete then repeat values of  $T$  can occur requiring that we make an adjustment to (13.46).

## References

- Akaike, H. (1973). Information theory and an extension of the maximum likelihood principle. In N. Petrov and F. Csadki (eds.) *Proceedings of the 2nd International Symposium on Information Theory*. pp. 267–81. Budapest: Akademiai Kiado.
- Altissimo, F., and G.L. Violante (1998). The nonlinear dynamics of output and unemployment in the US. *Journal of Applied Econometrics* (forthcoming).
- Amemiya, T. (1985). *Advanced Econometrics*. Cambridge, MA: Harvard University Press.
- Aneuryn-Evans, G., and A.S. Deaton (1980). Testing linear versus logarithmic regression models. *Review of Economic Studies* 47, 275–91.
- Atkinson, A. (1970). A method for discriminating between models (with discussion). *Journal of the Royal Statistical Society, B* B32, 323–53.
- Bahadur, R.R. (1960). Stochastic comparison of tests. *Annals of Mathematics and Statistics* 31, 276–95.
- Bahadur, R.R. (1967). Rates of convergence of estimates and test statistics. *Annals of Mathematics and Statistics* 38, 303–24.
- Barro, R. (1977). Unanticipated money growth and unemployment in the United States. *American Economic Review* 67, 101–15.
- Beran, R. (1988). Prepivoting test statistics: A bootstrap view of asymptotic refinements. *Journal of the American Statistical Association* 83, 403.
- Bernanke, B., H. Boivin, and P. Reiss (1988). Alternative nonnested specification tests of time-series investment models. *Journal of Econometrics* 37, 293–326.
- Chao, J.C., and N.R. Swanson (1997). Tests of nonnested hypotheses in nonstationary regressions with an application to modeling industrial production. Working Paper, Department of Economics, Pennsylvania State University.
- Chow, G.C. (1981). A comparison of the information and posterior probability criteria for model selection. *Journal of Econometrics* 16, 21–33.
- Coulibaly, N., and B. Brorsen (1998). A Monte Carlo sampling approach to testing nonnested hypotheses: Monte Carlo results. *Econometric Reviews* 195–209.
- Cox, D. (1961). Tests of separate families of hypothesis. *Proceedings of the Fourth Berkeley Symposium on Mathematical Statistics and Probability*.
- Cox, D. (1962). Further results on tests of separate families of hypotheses. *Journal of Royal Statistical Society B* 24, 406–24.

- Dastoor, N. (1981). A note on the interpretation of the Cox procedure for nonnested hypotheses. *Economics Letters* 8, 113–19.
- Dastoor, N.K. (1983). Some aspects of testing nonnested hypotheses. *Journal of Econometrics* 21, 213–28.
- Davidson, R., and J. MacKinnon (1981). Several tests for model specification in the presence of alternative hypotheses. *Econometrica* 49, 781–93.
- Davidson, R., and J.G. MacKinnon (1982). Some nonnested hypothesis tests and the relations among them. *Review of Economic Studies* 49, 551–65.
- Davidson, R., and J.G. MacKinnon (1985). Testing linear and loglinear regressions against Box–Cox alternatives. *Canadian Journal of Economics* 18, 499–517.
- Davies, R.B. (1977). Hypothesis testing when a nuisance parameter is present only under the alternative. *Biometrika* 64(2), 247–54.
- Deaton, A.S. (1982). Model selection procedures, or, does the consumption function exist? In G.C. Show, and P. Corsi (eds.) *Evaluating the Reliability of Macroeconomic Models*, pp. 43–65. New York: Wiley.
- Dixit, A.V., and R.S. Pindyck (1994). *Investment Under Uncertainty*. Chichester, UK: Princeton University Press.
- Duncan, A., and M. Weeks (1998). Nonnested models of labour supply with discrete choices. Working Paper, Department of Economics, University of York.
- Efron, B. (1979). Bootstrap methods: Another look at the jackknife. *Annals of Statistics* 7, 1–26.
- Engle, R., D. Hendry, and J. Richard (1983). Exogeneity. *Econometrica* 51, 277–304.
- Ericsson, N. (1983). Asymptotic properties of instrumental variables statistics for testing nonnested hypotheses. *Review of Economic Studies* 50, 287–304.
- Fisher, G., and M. McAleer (1979). On the interpretation of the Cox test in econometrics. *Economic Letters* 4, 145–50.
- Fisher, G.R., and M. McAleer (1981). Alternative procedures and associated tests of significance for nonnested hypotheses. *Journal of Econometrics* 16, 103–19.
- Friedman, M., and D. Meiselman (1963). The relative stability of monetary velocity and the investment multiplier in the United States 1897–1958. In *Stabilization Policies*. Englewood Cliffs, NJ: Commission on Money and Credit Research Study.
- Ghysels, E., and A. Hall (1990). Testing nonnested Euler conditions with quadrature based method of approximation. *Journal of Econometrics* 46, 273–308.
- Godfrey, L.G. (1983). Testing nonnested models after estimation by instrumental variables or least squares. *Econometrica* 51(2), 355–65.
- Godfrey, L.G., and M.H. Pesaran (1983). Tests of nonnested regression models: Small sample adjustments and Monte Carlo evidence. *Journal of Econometrics* 21, 133–54.
- Godfrey, L.G., and M. Wickens (1981). Testing linear and loglinear regressions for functional form. *Review of Economic Studies* 48, 487–96.
- Gouriéroux, C. (1982). Asymptotic comparison of tests for nonnested hypotheses by Bahadur's A.R.E. Discussion Paper 8215, CEPREMAP, Paris.
- Gouriéroux, C., and A. Monfort (1994). Testing nonnested hypotheses. In R.F. Engle, and D.L. McFadden (eds.) *Handbook of Econometrics, Volume 4*. Oxford: Elsevier.
- Gouriéroux, C., and A. Monfort (1995). Testing, encompassing, and simulating dynamic econometric models. *Econometric Theory* 11, 195–228.
- Gouriéroux, C., A. Monfort, and A. Trognon (1983). Testing nested or nonnested hypotheses. *Journal of Econometrics* 21, 83–115.
- Granger, C.W.J., M.L. King, and H. White (1995). Comments on testing economic theories and the use of model selection criteria. *Journal of Econometrics* 67, 173–87.

- Granger, C.W.J., and M.H. Pesaran (2000). A decision theoretic approach to forecast evaluation. In W.S. Chan, W.K. Li, and H. Tong (eds.), *Statistics and Finance: An Interface*. London: Imperial College Press.
- Grasa, A.A. (1989). *Econometric Model Selection: A New Approach*. Spain: Kluwer Academic Publishers.
- Hall, P. (1986). On the number of bootstrap simulations required to construct a confidence interval. *The Annals of Statistics* 14(4).
- Hall, P. (1988). Theoretical comparison of bootstrap confidence intervals. *Annals of Statistics* 16, 927–53.
- Hendry, D.F. (1993). *Econometrics: Alchemy or Science?* Oxford: Blackwell Publishers.
- Hinkley, D. (1988). Bootstrap methods. *Journal of the Royal Statistical Society, Series B* 50, 321–37.
- Jeong, J., and G. Maddala (1993). A perspective on application of bootstrap methods in econometrics. *Handbook of Statistics* 11, 573–605.
- Jorgenson, D.W., and C.D. Siebert (1968). A comparison of alternative theories of corporate investment behavior. *American Economic Review* 58, 681–712.
- Kapetanios, G., and M. Weeks (1999). Nonnested models and the likelihood ratio statistic: A comparison of simulation and bootstrap-based tests. Department of Applied Economics Working Paper, University of Cambridge.
- Kent, J. (1986). The underlying structure of nonnested hypothesis tests. *Biometrika* 7, 333–43.
- King, M.L. (1983). Testing for autoregressive against moving average errors in the linear regression model. *Journal of Econometrics* 21, 35–51.
- Kullback, S. (1959). *Information Theory and Statistics*. New York: Wiley.
- Lavergne, P. (1998). Selection of regressors in econometrics: Parametric and nonparametric methods. *Econometric Reviews* 17(3), 227–73.
- Leamer, E.E. (1983). Model choice and specification analysis. In Z. Griliches, and M.D. Intriligator (eds.) *Handbook of Econometrics Volume 1*. University of California, LA: North-Holland Publishing Company.
- Li, H., and G. Maddala (1996). Bootstrapping time series models. *Econometric Reviews* 15(2), 115–58.
- Linhart, H., and W. Zucchini (1986). *Model Selection*. New York: Wiley and Sons.
- Maddala, G.S.E. (1981). Model selection. Special Issue of *Journal of Econometrics* 16(1).
- McAleer, M. (1983). Exact tests of a model against nonnested alternative. *Biometrika* 70, 285–88.
- McAleer, M. (1995). The significance of testing empirical nonnested models. *Journal of Econometrics* 67, 149–71.
- McAleer, M., and S. Ling (1998). A nonnested test for the GARCH and E-GARCH models. Working Paper, Department of Economics, University of Western Australia.
- McAleer, M., and M.H. Pesaran (1986). Statistical inference in nonnested econometric models. *Applied Mathematics and Computation* 20, 271–311.
- McAleer, M.G., G. Fisher, and P. Volker (1982). Separate misspecified regressions and U.S. long run demand for money function. *Review of Economics and Statistics* 64, 572–83.
- McAleer, M.J., M.H. Pesaran, and A.K. Bera (1990). Alternative approaches to testing nonnested models with autocorrelated disturbances: An application to models of U.S. unemployment. *Communications in Statistics series A*(19), 3619–44.
- Mizon, G.E., and J.F. Richard (1986). The encompassing principle and its application to nonnested hypotheses. *Econometrica* 54, 657–78.
- Orme, C. (1994). Nonnested tests for discrete choice models. Working Paper, Department of Economics, University of York.

- Pereira, B.D.B. (1984). On the choice of a Weibull model. *Journal of the Inter American Statistical Institute* 26, 157–63.
- Pesaran, M.H. (1974). On the general problem of model selection. *Review of Economic Studies* 41, 153–71.
- Pesaran, M.H. (1981). Pitfalls of testing nonnested hypotheses by the Lagrange multiplier method. *Journal of Econometrics* 17, 323–31.
- Pesaran, M.H. (1982a). Comparison of local power of alternative tests of nonnested regression models. *Econometrica*.
- Pesaran, M.H. (1982b). A critique of the proposed tests of the natural rate-rational expectations hypothesis. *The Economic Journal* 92, 529–54.
- Pesaran, M.H. (1984). Asymptotic power comparisons of tests of separate parametric families by Bahadur's approach. *Biometrika* 71(2), 245–52.
- Pesaran, M.H. (1987). Global and partial nonnested hypotheses and asymptotic local power. *Econometric Theory* 3, 69–97.
- Pesaran, M.H., and S. Deaton (1978). Testing nonnested nonlinear regression models. *Econometrica* 46, 677–94.
- Pesaran, M.H., and B. Pesaran (1993). A simulation approach to the problem of computing Cox's statistic for testing nonnested models. *Journal of Econometrics* 57, 377–92.
- Pesaran, M.H., and B. Pesaran (1995). A nonnested test of level differences versus log-differenced stationary models. *Econometric Reviews* 14(2), 213–27.
- Pesaran, M.H., and S. Potter (1997). A floor and ceiling model of US output. *Journal of Economic Dynamics and Control* 21(4–5), 661–96.
- Sawyer, K.R. (1983). Testing separate families of hypotheses: An information criterion. *Journal of the Royal Statistical Society B* 45, 89–99.
- Sawyer, K.R. (1984). Multiple hypothesis testing. *Royal Statistical Society B* 46(3), 419–24.
- Schwarz, G. (1978). Estimating the dimension of a model. *Annals of Statistics* 6, 461–4.
- Smith, M.A., and G.S. Maddala (1983). Multiple model testing for nonnested heteroskedastic censored regression models. *Journal of Econometrics* 21, 71–81.
- Smith, M.D., and D.J. Smyth (1991). Multiple and pairwise nonnested tests of the influence of taxes on money demand. *Journal of Applied Econometrics* 6, 17–30.
- Smith, R.J. (1992). Nonnested tests for competing models estimated by generalized method of moments. *Econometrica* 60, 973–80.
- Smith, R.J. (1993). Consistent tests for the encompassing hypothesis. Document de Travail No. 9403, INSEE, Paris.
- Vinod, H. (1993). Bootstrap methods: Applications in econometrics. *Handbook of Statistics* 11, 629–61.
- Vuong, Q.H. (1989). Likelihood ratio tests for model selection and nonnested hypothesis. *Econometrica* 57(2), 307–33.
- Walker, A.M. (1967). Some tests of separate families of hypotheses in time series analysis. *Biometrika* 54, 39–68.
- Weeks, M. (1996). Testing the binomial and multinomial choice models using Cox's nonnested test. *Journal of the American Statistical Association (Papers and Proceedings)*.
- White, H. (1982). Regularity conditions for Cox's test of nonnested hypothesis. *Journal of Econometrics* 19, 301–18.
- Zellner, A. (1971). *An Introduction to Bayesian Inference in Econometrics*. New York: John Wiley and Sons.

---

CHAPTER FOURTEEN

# Spatial Econometrics

*Luc Anselin\**

## 1 INTRODUCTION

Spatial econometrics is a subfield of econometrics that deals with spatial interaction (*spatial autocorrelation*) and spatial structure (*spatial heterogeneity*) in regression models for cross-sectional and panel data (Paelinck and Klaassen, 1979; Anselin, 1988a). Such a focus on location and spatial interaction has recently gained a more central place not only in applied but also in theoretical econometrics. In the past, models that explicitly incorporated “space” (or geography) were primarily found in specialized fields such as regional science, urban, and real estate economics and economic geography (e.g. recent reviews in Anselin, 1992a; Anselin and Florax, 1995a; Anselin and Rey, 1997; Pace *et al.*, 1998). However, more recently, spatial econometric methods have increasingly been applied in a wide range of empirical investigations in more traditional fields of economics as well, including, among others, studies in demand analysis, international economics, labor economics, public economics and local public finance, and agricultural and environmental economics.<sup>1</sup>

This new attention to specifying, estimating, and testing for the presence of spatial interaction in the mainstream of applied and theoretical econometrics can be attributed to two major factors. One is a growing interest within theoretical economics in *models* that move towards an explicit accounting for the interaction of an economic agent with other heterogeneous agents in the system. These new theoretical frameworks of “interacting agents” model strategic interaction, social norms, neighborhood effects, copy-catting, and other peer group effects, and raise interesting questions about how the individual interactions can lead to emergent collective behavior and aggregate patterns. Models used to estimate such phenomena require the specification of how the magnitude of a variable of interest (say crime) at a given location (say a census tract) is determined by the values of the same variable at other locations in the system (such as neighboring census tracts). If such a dependence exists, it is referred to as spatial autocorrelation. A second driver behind the increased interest in spatial econometric

techniques is the need to handle spatial *data*. This has been stimulated by the explosive diffusion of geographic information systems (GIS) and the associated availability of geocoded data (i.e. data sets that contain the location of the observational units). There is a growing recognition that standard econometric techniques often fail in the presence of spatial autocorrelation, which is commonplace in geographic (cross-sectional) data sets.<sup>2</sup>

Historically, spatial econometrics originated as an identifiable field in Europe in the early 1970s because of the need to deal with sub-country data in regional econometric models (e.g. Paelinck and Klaassen, 1979). In general terms, spatial econometrics can be characterized as the set of techniques to deal with methodological concerns that follow from the explicit consideration of *spatial effects*, specifically spatial autocorrelation and spatial heterogeneity. This yields four broad areas of interest: (i) the formal *specification* of spatial effects in econometric models; (ii) the *estimation* of models that incorporate spatial effects; (iii) specification *tests* and diagnostics for the presence of spatial effects; and (iv) spatial *prediction* (interpolation). In this brief review chapter, I will focus on the first three concerns, since they fall within the central preoccupation of econometric methodology.

The remainder of the chapter is organized as follows. In Section 2, I outline some foundations and definitions. In Section 3, the specification of spatial regression models is treated, including the incorporation of spatial dependence in panel data models and models with qualitative variables. Section 4 focuses on estimation and Section 5 on specification testing. In Section 6, some practical implementation and software issues are addressed. Concluding remarks are formulated in Section 7.

## 2 FOUNDATIONS

### 2.1 Spatial autocorrelation

In a regression context, spatial effects pertain to two categories of specifications. One deals with spatial dependence, or its weaker expression, *spatial autocorrelation*, and the other with *spatial heterogeneity*.<sup>3</sup> The latter is simply structural instability, either in the form of non-constant error variances in a regression model (heteroskedasticity) or in the form of variable regression coefficients. Most of the methodological issues related to spatial heterogeneity can be tackled by means of the standard econometric toolbox.<sup>4</sup> Therefore, given the space constraints for this chapter, the main focus of attention in the remainder will be on spatial dependence.

The formal framework used for the statistical analysis of spatial autocorrelation is a so-called spatial stochastic process (also often referred to as a spatial random field), or a collection of random variables  $y_i$ , indexed by location  $i$ ,

$$\{y_i, i \in D\}, \quad (14.1)$$

where the index set  $D$  is either a continuous surface or a finite set of discrete locations. (See Cressie (1993), for technical details.) Since each random variable is

"tagged" by a location, spatial autocorrelation can be formally expressed by the moment condition,

$$\text{cov}[y_i, y_j] = E[y_i y_j] - E[y_i] \cdot E[y_j] \neq 0, \quad \text{for } i \neq j \quad (14.2)$$

where  $i, j$  refer to individual observations (locations) and  $y_i$  ( $y_j$ ) is the value of a random variable of interest at that location. This covariance becomes meaningful from a spatial perspective when the particular configuration of nonzero  $i, j$  pairs has an interpretation in terms of spatial structure, spatial interaction or the spatial arrangement of the observations. For example, this would be the case when one is interested in modeling the extent to which technological innovations in a county spill over into neighboring counties.

The spatial covariance can be modeled in three basic ways. First, one can specify a particular functional form for a spatial stochastic process generating the random variable in (14.1), from which the covariance structure would follow. Second, one can model the covariance structure directly, typically as a function of a small number of parameters (with any given covariance structure corresponding to a class of spatial stochastic processes). Third, one can leave the covariance unspecified and estimate it nonparametrically.<sup>5</sup> I will review each of these approaches in turn.

### Spatial Stochastic Process Models

The most often used approach to formally express spatial autocorrelation is through the specification of a functional form for the spatial stochastic process (14.1) that relates the value of a random variable at a given location to its value at other locations. The covariance structure then follows from the nature of the process. In parallel to time series analysis, spatial stochastic processes are categorized as spatial autoregressive (SAR) and spatial moving average (SMA) processes, although there are several important differences between the cross-sectional and time series contexts.<sup>6</sup>

For example, for an  $N \times 1$  vector of random variables,  $y$ , observed across space, and an  $N \times 1$  vector of iid random errors  $\varepsilon$ , a simultaneous spatial autoregressive (SAR) process is defined as

$$(y - \mu i) = \rho W(y - \mu i) + \varepsilon, \quad \text{or} \quad (y - \mu i) = (I - \rho W)^{-1}\varepsilon, \quad (14.3)$$

where  $\mu$  is the (constant) mean of  $y_i$ ,  $i$  is an  $N \times 1$  vector of ones, and  $\rho$  is the spatial autoregressive parameter.

Before considering the structure of this process more closely, note the presence of the  $N \times N$  matrix  $W$ , which is referred to as a *spatial weights matrix*. For each location in the system, it specifies which of the other locations in the system affect the value at that location. This is necessary, since in contrast to the unambiguous notion of a "shift" along the time axis (such as  $y_{t-1}$  in an autoregressive model), there is no corresponding concept in the spatial domain, especially when observations are located irregularly in space.<sup>7</sup> Instead of the notion of shift, a *spatial lag operator* is used, which is a weighted average of random variables at "neighboring" locations.<sup>8</sup>

The spatial weights crucially depend on the definition of a neighborhood set for each observation. This is obtained by selecting for each location  $i$  (as the row) the neighbors as the columns corresponding to nonzero elements  $w_{ij}$  in a fixed (nonstochastic) and positive  $N \times N$  spatial weights matrix  $W$ .<sup>9</sup> A spatial lag for  $y$  at  $i$  then follows as

$$[Wy]_i = \sum_{j=1, \dots, N} w_{ij} \cdot y_j, \quad (14.4)$$

or, in matrix form, as

$$Wy. \quad (14.5)$$

Since for each  $i$  the matrix elements  $w_{ij}$  are only nonzero for those  $j \in S_i$  (where  $S_i$  is the neighborhood set), only the matching  $y_j$  are included in the lag. For ease of interpretation, the elements of the spatial weights matrix are typically row-standardized, such that for each  $i$ ,  $\sum_j w_{ij} = 1$ . Consequently, the spatial lag may be interpreted as a weighted average (with the  $w_{ij}$  being the weights) of the neighbors, or as a spatial smoother.

It is important to note that the elements of the weights matrix are nonstochastic and exogenous to the model. Typically, they are based on the geographic arrangement of the observations, or contiguity. Weights are nonzero when two locations share a common boundary, or are within a given distance of each other. However, this notion is perfectly general and alternative specifications of the spatial weights (such as economic distance) can be considered as well (Anselin, 1980, ch. 8; Case, Rosen, and Hines, 1993; Pinkse and Slade, 1998).

The constraints imposed by the weights structure (the zeros in each row), together with the specific form of the spatial process (autoregressive or moving average) determine the variance–covariance matrix for  $y$  as a function of two parameters, the variance  $\sigma^2$  and the spatial coefficient,  $\rho$ . For the SAR structure in (14.3), this yields (since  $E[y - \mu_i] = 0$ )

$$\text{cov}[(y - \mu_i), (y - \mu_i)] = E[(y - \mu_i)(y - \mu_i)'] = \sigma^2[(I - \rho W)'(I - \rho W)]^{-1}. \quad (14.6)$$

This is a full matrix, which implies that shocks at any location affect all other locations, through a so-called *spatial multiplier* effect (or, global interaction).<sup>10</sup>

A major distinction between processes in space compared to the time domain is that even with iid error terms  $\varepsilon_i$ , the diagonal elements in (14.6) are not constant.<sup>11</sup> Furthermore, the heteroskedasticity depends on the neighborhood structure embedded in the spatial weights matrix  $W$ . Consequently, the process in  $y$  is not covariance-stationary. Stationarity is only obtained in very rare cases, for example on regular lattice structures when each observation has an identical weights structure, but this is of limited practical use. This lack of stationarity has important implications for the types of central limit theorems (CLTs) and laws of large numbers (LLNs) that need to be invoked to obtain asymptotic properties for estimators and specification test, a point that has not always been recognized in the literature.

## DIRECT REPRESENTATION

A second commonly used approach to the formal specification of spatial autocorrelation is to express the elements of the variance–covariance matrix in a parsimonious fashion as a “direct” function of a small number of parameters and one or more exogenous variables. Typically, this involves an inverse function of some distance metric, for example,

$$\text{cov}[\varepsilon_i, \varepsilon_j] = \sigma^2 f(d_{ij}, \varphi), \quad (14.7)$$

where  $\varepsilon_i$  and  $\varepsilon_j$  are regression disturbance terms,  $\sigma^2$  is the error variance,  $d_{ij}$  is the distance separating observations (locations)  $i$  and  $j$ , and  $f$  is a distance decay function such that  $\frac{\partial f}{\partial d} < 0$  and  $|f(d_{ij}, \varphi)| \leq 1$ , with  $\varphi \in \Phi$  as a  $p \times 1$  vector of parameters on an open subset  $\Phi$  of  $R^p$ . This form is closely related to the variogram model used in geostatistics, although with stricter assumptions regarding stationarity and isotropy. Using (14.7) for individual elements, the full error covariance matrix follows as

$$E[\varepsilon\varepsilon'] = \sigma^2 \Omega(d_{ij}, \varphi), \quad (14.8)$$

where, because of the scaling factor  $\sigma^2$ , the matrix  $\Omega(d_{ij}, \varphi)$  must be a positive definite spatial correlation matrix, with  $\omega_{ii} = 1$  and  $|\omega_{ij}| \leq 1, \forall i, j$ .<sup>12</sup> Note that, in contrast to the variance for the spatial autoregressive model, the direct representation model does not induce heteroskedasticity.

In spatial econometrics, models of this type have been used primarily in the analysis of urban housing markets, e.g. in Dubin (1988, 1992), and Basu and Thibodeau (1998). While this specification has a certain intuition, in the sense that it incorporates an explicit notion of spatial clustering as a function of the distance separating two observations (i.e. positive spatial correlation), it is also fraught with a number of estimation and identification problems (Anselin, 2000a).

## NONPARAMETRIC APPROACHES

A nonparametric approach to estimating the spatial covariance matrix does not require an explicit spatial process or functional form for the distance decay. This is common in the case of panel data, when the time dimension is (considerably) greater than the cross-sectional dimension ( $T \gg N$ ) and the “spatial” covariance is estimated from the sample covariance for the residuals of each set of location pairs (e.g. in applications of Zellner’s SUR estimator; see Chapter 5 by Fiebig in this volume).

Applications of this principle to spatial autocorrelation are variants of the well known Newey-West (1987) heteroskedasticity and autocorrelation consistent covariance matrix and have been used in the context of generalized methods of moments (GMM) estimators of spatial regression models (see Section 4.3). Conley (1996) suggested a covariance estimator based on a sequence of weighted averages of sample autocovariances computed for subsets of observation pairs that fall within a given distance band (or spatial window). Although not presented as such, this has a striking similarity to the nonparametric estimation of a semi-variogram in geostatistics (see, e.g. Cressie, 1993, pp. 69–70), but the assumptions

of stationarity and isotropy required in the GMM approach are stricter than those needed in variogram estimation. In a panel data setting, Driscoll and Kraay (1998) use a similar idea, but avoid having to estimate the spatial covariances by distance bands. This is accomplished by using only the cross-sectional averages (for each time period) of the moment conditions, and by relying on asymptotics in the time dimension to yield an estimator for the spatial covariance structure.

## 2.2 Asymptotics in spatial stochastic processes

As in time series analysis, the properties of estimators and tests for spatial series are derived from the asymptotics for stochastic processes. However, these properties are not simply extensions to two dimensions of the time series results. A number of complicating factors are present and to date some formal results for the spatial dependence case are still lacking. While an extensive treatment of this topic is beyond the scope of the current chapter, three general comments are in order. First, the intuition behind the asymptotics is fairly straightforward in that regularity conditions are needed to limit the extent of spatial dependence (memory) and heterogeneity of the spatial series in order to obtain the proper (uniform) laws of large numbers and central limit theorems to establish consistency and asymptotic normality. In this context, it is important to keep in mind that both SAR and SMA processes yield heteroskedastic variances, so that the application of results for dependent *stationary* series are not applicable.<sup>13</sup> In addition to the usual moment conditions that are similar in spirit to those for heterogeneous dependent processes in time (e.g. Pötscher and Prucha, 1997), specific spatial conditions will translate into constraints on the spatial weights and on the parameter space for the spatial coefficients (for some specific examples, see, e.g. Anselin and Kelejian, 1997; Kelejian and Prucha, 1999b; Pinkse and Slade, 1998; Pinkse, 2000). In practice, these conditions are likely to be satisfied by most spatial weights that are based on simple contiguity, but this is not necessarily the case for general weights, such as those based on economic distance.

A second distinguishing characteristic of asymptotics in space is that the limit may be approached in two different ways, referred to as *increasing domain* asymptotics and *infill* asymptotics.<sup>14</sup> The former consists of a sampling structure where new “observations” are added at the edges (boundary points), similar to the underlying asymptotics in time series analysis. Infill asymptotics are appropriate when the spatial domain is bounded, and new observations are added in between existing ones, generating a increasingly denser surface. Many results for increasing domain asymptotics are not directly applicable to infill asymptotics (Lahiri, 1996). In most applications of spatial econometrics, the implied structure is that of an increasing domain.

Finally, for spatial processes that contain spatial weights, the asymptotics require the use of CLT and LLN for triangular arrays (Davidson, 1994, chs. 19, 24). This is caused by the fact that for the boundary elements the “sample” weights matrix changes as new data points are added (i.e. the new data points change the connectedness structure for existing data points).<sup>15</sup> Again, this is an additional degree of complexity, which is not found in time series models.

### 3 SPATIAL REGRESSION MODELS

#### 3.1 Spatial lag and spatial error models

In the standard linear regression model, spatial dependence can be incorporated in two distinct ways: as an additional regressor in the form of a spatially lagged dependent variable ( $Wy$ ), or in the error structure ( $E[\varepsilon_i \varepsilon_j] \neq 0$ ). The former is referred to as a *spatial lag* model and is appropriate when the focus of interest is the assessment of the existence and strength of spatial interaction. This is interpreted as substantive spatial dependence in the sense of being directly related to a *spatial model* (e.g. a model that incorporates spatial interaction, yardstick competition, etc.). Spatial dependence in the regression disturbance term, or a *spatial error* model is referred to as nuisance dependence. This is appropriate when the concern is with correcting for the potentially biasing influence of the spatial autocorrelation, due to the use of *spatial data* (irrespective of whether the model of interest is spatial or not).

Formally, a spatial lag model, or a mixed regressive, spatial autoregressive model is expressed as

$$y = \rho Wy + X\beta + \varepsilon, \quad (14.9)$$

where  $\rho$  is a spatial autoregressive coefficient,  $\varepsilon$  is a vector of error terms, and the other notation is as before.<sup>16</sup> Unlike what holds for the time series counterpart of this model, the spatial lag term  $Wy$  is correlated with the disturbances, even when the latter are iid. This can be seen from the reduced form of (14.9),

$$y = (I - \rho W)^{-1}X\beta + (I - \rho W)^{-1}\varepsilon, \quad (14.10)$$

in which each inverse can be expanded into an infinite series, including both the explanatory variables and the error terms at all locations (the spatial multiplier). Consequently, the spatial lag term must be treated as an endogenous variable and proper estimation methods must account for this endogeneity (OLS will be biased and inconsistent due to the simultaneity bias).

A spatial error model is a special case of a regression with a non-spherical error term, in which the off-diagonal elements of the covariance matrix express the structure of spatial dependence. Consequently, OLS remains unbiased, but it is no longer efficient and the classical estimators for standard errors will be biased. The spatial structure can be specified in a number of different ways, and (except for the non-parametric approaches) results in a error variance–covariance matrix of the form

$$E[\varepsilon \varepsilon'] = \Omega(\theta), \quad (14.11)$$

where  $\theta$  is a vector of parameters, such as the coefficients in an SAR error process.<sup>17</sup>

### 3.2 Spatial dependence in panel data models

When observations are available across space as well as over time, the additional dimension allows the estimation of the full covariance of one type of association, using the other dimension to provide the asymptotics (e.g. in SUR models with  $N \ll T$ ). However, as in the pure cross-sectional case, there is insufficient information in the  $NT$  observations to estimate the complete  $(NT)^2$  covariance matrix  $\text{cov}[y_{it}, y_{js}] \neq 0$ , (with  $i \neq j$  and  $t \neq s$ ) without imposing some structure. For small  $N$  and large  $T$ , the asymptotics in the time domain can be exploited to obtain a nonparametric estimate of cross-sectional dependence, while time dependence must be parameterized. Similarly, for large  $N$  and small  $T$ , the asymptotics in the spatial domain can be exploited to yield a nonparametric estimate of serial (time) dependence, while spatial dependence must be parameterized. As in the pure cross-sectional case, the latter requires the use of a spatial weights matrix. In each of these situations, asymptotics are only needed in one of the dimensions while the other can be treated as fixed.

When both spatial as well as serial dependence are parameterized, a range of specifications can be considered, allowing different combinations of the two. For ease of exposition, assume that the observations are stacked by time period, i.e. they can be considered as  $T$  time slices of  $N$  cross-sectional units. Restricting attention to “lag” dependence, and with  $f(z)$  as a generic designation for the regressors (which may be lagged in time and/or space), four types of models can be distinguished.

1. *pure space-recursive*, in which the dependence pertains to neighboring locations in a different period, or,

$$y_{it} = \gamma [W y_{t-1}]_i + f(z) + \varepsilon_{it}, \quad (14.12)$$

where, using the same notational convention as before,  $[W y_{t-1}]_i$  is the  $i$ th element of the spatial lag vector applied to the observations on the dependent variable in the previous time period (using an  $N \times N$  spatial weights matrix for the cross-sectional units).

2. *time-space recursive*, in which the dependence relates to the same location as well as the neighboring locations in another period, or,

$$y_{it} = \lambda y_{it-1} + \gamma [W y_{t-1}]_i + f(z) + \varepsilon_{it} \quad (14.13)$$

3. *time-space simultaneous*, with both a time-wise and a spatially lagged dependent variable, or,

$$y_{it} = \lambda y_{it-1} + \rho [W y_t]_i + f(z) + \varepsilon_{it} \quad (14.14)$$

where  $[W y_t]_i$  is the  $i$ th element of the spatial lag vector in the same time period.

4. *time-space dynamic*, with all forms of dependence, or,

$$y_{it} = \lambda y_{it-1} + \rho [W y_{it}]_i + \gamma [W y_{t-1}]_i + f(z) + \varepsilon_{it}. \quad (14.15)$$

In order to estimate the parameters of the time-space simultaneous model, asymptotics are needed in the cross-sectional dimension, while for the time-space dynamic model, asymptotics are needed in both dimensions. For the other models, the type of asymptotics required are determined by the dependence structure in the error terms. For example, the pure space-recursive model with iid errors satisfies the assumptions of the classical linear model and can be estimated by means of OLS.

Spatial lag and spatial error dependence can be introduced into the cross-sectional dimension of traditional panel data models in a straightforward way. For example, in a spatial SUR model, both autoregressive as well as regression parameters are allowed to vary by time period, in combination with a nonparametric serial covariance. The spatial lag formulation of such a model would be (in the same notation as before):

$$y_{it} = \rho_t [W y_t]_i + x'_{it} \beta_t + \varepsilon_{it} \quad (14.16)$$

with  $\text{var}[\varepsilon_{it}] = \sigma_t^2$  and  $E[\varepsilon_{it} \varepsilon_{is}] = \sigma_{is}$ .<sup>18</sup>

An important issue to consider when incorporating spatial dependence in panel data models is the extent to which fixed effects may be allowed. Since the estimation of the spatial process models requires asymptotics in the cross-sectional domain ( $N \rightarrow \infty$ ), fixed effects (i.e. a dummy variable for each location) would suffer from the incidental parameter problem and no consistent estimator exists. Hence, fixed cross-sectional effects are incompatible with spatial processes and instead a random effects specification must be considered.

### 3.3 Spatial dependence in models for qualitative data

Empirical analysis of interacting agents requires models that incorporate spatial dependence for discrete dependent variables, such as counts or binary outcomes (Brock and Durlauf, 1995). This turns out to be quite complex and continues to be an active area of research. While an extensive discussion of the technical aspects associated with spatial discrete choice models is beyond the scope of the current chapter, the salient issues may be illustrated with a spatial version of the probit model, which has recently received considerable attention.<sup>19</sup>

The point of departure is the familiar expression for a linear model in a latent (unobserved) dependent variable  $y_i^*$

$$y_i^* = x'_i \beta + \varepsilon_i, \quad (14.17)$$

where  $\varepsilon_i$  is a random variable for which a given distribution is assumed (e.g. the normal for the probit model). The realization of  $y_i^*$  is observed in the form of discrete events,  $y_i = 1$  for  $y_i^* \geq 0$ , and  $y_i = 0$  for  $y_i^* < 0$ . The discrete events are

related to the underlying probability model through the error term, for example,  $y_i^* \geq 0$  implies  $-x_i'\beta < \varepsilon_i$ , and, therefore,

$$E[y_i] = P[y_i = 1] = \Phi[x_i'\beta], \quad (14.18)$$

where  $\Phi$  is the cumulative distribution function for the standard normal.

Spatial autocorrelation can be introduced into this model in the form of a spatial autoregressive process for the error term  $\varepsilon_i$  in (14.17), or

$$\varepsilon_i = \lambda \sum_j w_{ij} \varepsilon_j + u_i, \quad (14.19)$$

where  $\lambda$  is an autoregressive parameter, the  $w_{ij}$  are the elements in the  $i$ th row of a spatial weights matrix, and  $u_i$  may be assumed to be iid standard normal. As a consequence of the spatial multiplier in the autoregressive specification, the random error at each location now becomes a function of the random errors at all other locations as well. Its distribution is multivariate normal with  $N \times N$  variance–covariance matrix

$$E[\varepsilon \varepsilon'] = [(I - \lambda W)'(I - \lambda W)]^{-1}. \quad (14.20)$$

As pointed out above, besides being non-diagonal, (14.20) is also heteroskedastic. Consequently, the usual inequality conditions that are at the basis of (14.18) no longer hold, since each location has a different variance. Moreover,  $P[-x_i'\beta < \varepsilon_i]$  can no longer be derived from the univariate standard normal distribution, but rather must be expressed explicitly as the marginal distribution of a  $N$ -dimensional multivariate normal vector, whose variance–covariance matrix contains off-diagonal elements that are a function of the autoregressive parameter  $\lambda$ . This is non-standard and typically not analytically tractable, which greatly complicates estimation and specification testing. Similar issues are faced in the spatial lag model for a latent variable.<sup>20</sup>

## 4 ESTIMATION

### 4.1 Maximum likelihood estimation

Maximum likelihood (ML) estimation of spatial lag and spatial error regression models was first outlined by Ord (1975).<sup>21</sup> The point of departure is an assumption of normality for the error terms. The joint likelihood then follows from the multivariate normal distribution for  $y$ . Unlike what holds for the classic regression model, the joint loglikelihood for a spatial regression does not equal the sum of the loglikelihoods associated with the individual observations. This is due to the two-directional nature of the spatial dependence, which results in a Jacobian term that is the determinant of a full  $N \times N$  matrix, e.g.  $|I - \rho W|$ .

For the SAR error model, the loglikelihood is based on the multivariate normal case, for example, as used in the general treatment of Magnus (1978). Since  $\varepsilon \sim MVN(0, \Sigma)$ , it follows that, with  $\varepsilon = y - X\beta$  and  $\Sigma = \sigma^2[(I - \lambda W)'(I - \lambda W)]^{-1}$ ,

$$\begin{aligned} \ln L = & -(N/2) \ln (2\pi) - (N/2) \ln \sigma^2 + \ln |I - \lambda W| \\ & -(1/2\sigma^2)(y - X\beta)'(I - \lambda W)'(I - \lambda W)(y - X\beta). \end{aligned} \quad (14.21)$$

Closer inspection of the last term in (14.21) reveals that, conditional upon  $\lambda$  (the spatial autoregressive parameter), a maximization of the loglikelihood is equivalent to the minimization of the sum of squared residuals in a regression of a spatially filtered dependent variable  $y^* = y - \lambda Wy$  on a set of spatially filtered explanatory variables  $X^* = X - \lambda WX$ . The first order conditions for  $\hat{\beta}_{ML}$  indeed yield the familiar generalized least squares estimator:

$$\hat{\beta}_{ML} = [(X - \lambda WX)'(X - \lambda WX)]^{-1}(X - \lambda WX)'(y - \lambda Wy) \quad (14.22)$$

and, similarly, the ML estimator for  $\sigma^2$  follows as:

$$\hat{\sigma}_{ML}^2 = (e - \lambda We)'(e - \lambda We)/N \quad (14.23)$$

with  $e = y - X\hat{\beta}_{ML}$ . However, unlike the time series case, a consistent estimator for  $\lambda$  cannot be obtained from the OLS residuals and therefore the standard two-step FGLS approach does not apply.<sup>22</sup> Instead, the estimator for  $\lambda$  must be obtained from an explicit maximization of a concentrated likelihood function (for details, see Anselin, 1988a, ch. 6, and Anselin and Bera, 1998).

The loglikelihood for the spatial lag model is obtained using the same general principles (see Anselin, 1988, ch. 6 for details) and takes the form

$$\begin{aligned} \ln L = & -(N/2) \ln (2\pi) - (N/2) \ln \sigma^2 + \ln |I - \rho W| \\ & -(1/2\sigma^2)(y - \rho Wy - X\beta)'(y - \rho Wy - X\beta). \end{aligned} \quad (14.24)$$

The minimization of the last term in (14.24) corresponds to OLS, but since this ignores the log Jacobian  $\ln |I - \rho W|$ , OLS is not a consistent estimator in this model. As in the spatial error model, there is no satisfactory two-step procedure and estimators for the parameters must be obtained from an explicit maximization of the likelihood. This is greatly simplified since both  $\hat{\beta}_{ML}$  and  $\hat{\sigma}_{ML}^2$  can be obtained conditional upon  $\rho$  from the first order conditions:

$$\hat{\beta}_{ML} = (X'X)^{-1}X'(y - \rho Wy), \quad (14.25)$$

or, with  $\hat{\beta}_0 = (X'X)^{-1}X'y$ ,  $e_0 = y - X\hat{\beta}_0$ ,  $\hat{\beta}_L = (X'X)^{-1}X'Wy$ ,  $e_L = y - X\hat{\beta}_L$ ,

$$\hat{\beta}_{ML} = \hat{\beta}_0 - \rho \hat{\beta}_L \quad (14.26)$$

and

$$\hat{\sigma}_{ML}^2 = (e_0 - \rho e_L)'(e_0 - \rho e_L)/N. \quad (14.27)$$

This yields a concentrated loglikelihood in a single parameter, which is straightforward to optimize by means of direct search techniques (see Anselin (1980, 1988a) for derivations and details).

Both spatial lag and spatial error models are special cases of a more general specification that may include forms of heteroskedasticity as well. This also provides the basis for ML estimation of spatial SUR models with spatial lag or spatial error terms (Anselin, 1980, ch. 10). Similarly, ML estimation of error components models with spatial lag or spatial error terms can be implemented as well. Spatial models with discrete dependent variables are typically not estimated by means of ML, given the prohibitive nature of evaluating multiple integrals to determine the relevant marginal distributions.<sup>23</sup>

Finally, it is important to note that models with spatial dependence do not fit the classical framework (e.g. as outlined in Rao, 1973) under which the optimal properties (consistency, asymptotic efficiency, asymptotic normality) of ML estimators are established. This implies that these properties do not necessarily hold and that careful consideration must be given to the explicit formulation of regularity conditions. In general terms, aside from the usual restrictions on the variance and higher moments of the model variables, these conditions boil down to constraints on the range of dependence embodied in the spatial weights matrix.<sup>24</sup> In addition, to avoid singularity or explosive processes, the parameter space for the coefficient in a spatial process model is restricted to an interval other than the familiar  $-1, +1$ . For example, for an SAR process, the parameter space is  $1/\omega_{\min} < \rho < 1/\omega_{\max}$ , where  $\omega_{\min}$  and  $\omega_{\max}$  are the smallest (on the real line) and largest eigenvalues of the spatial weights matrix  $W$ . For row-standardized weights,  $\omega_{\max} = 1$ , but  $\omega_{\min} > -1$ , such that the lower bound on the parameter space is less than  $-1$  (Anselin, 1980). This must be taken into account in practical implementations of estimation routines.

## 4.2 Spatial two-stage least squares

The endogeneity of the spatially lagged dependent variable can also be addressed by means of an instrumental variables or two-stage least squares (2SLS) approach (Anselin, 1980, 1988a, 1990; Kelejian and Robinson, 1993; Kelejian and Prucha, 1998). As demonstrated in Kelejian and Robinson (1993), the choice of an instrument for  $Wy$  follows from the conditional expectation in the reduced form (14.10),

$$E[y|X] = (I - \rho W)^{-1}X\beta = X\beta + \rho WX\beta + \rho^2 W^2X\beta + \dots \quad (14.28)$$

Apart from the exogenous variables  $X$  (which are always instruments), this includes their spatial lags as well, suggesting  $WX$  as a set of instruments.

Under a set of reasonable assumptions that are easily satisfied when the spatial weights are based on contiguity, the spatial two-stage least squares estimator achieves the consistency and asymptotic normality properties of the standard 2SLS (see, e.g. the theorems spelled out in Schmidt, 1976).<sup>25</sup> A straightforward extension is the application of 3SLS to the spatial SUR model with a spatial lag (Anselin, 1988a, ch. 10).

### 4.3 Method of moments estimators

Recently, a number of approaches have been outlined to estimate the coefficients in a spatial error model as an application of general principles underlying the method of moments. Kelejian and Prucha (1999a) develop a set of moment conditions that yield estimation equations for the parameter of an SAR error model. Specifically, assuming an iid error vector  $u$ , the following three conditions readily follow

$$\begin{aligned} E[u'u/N] &= \sigma^2 \\ E[u'W'Wu/N] &= \sigma^2(1/N)\text{tr}(W'W) \\ E[u'Wu/N] &= 0 \end{aligned} \tag{14.29}$$

where  $\text{tr}$  is the matrix trace operator. Replacing  $u$  by  $e - \lambda We$  (with  $e$  as the vector of OLS residuals) in (14.29) yields a system of three equations in the parameters  $\lambda$ ,  $\lambda^2$ , and  $\sigma^2$ . Kelejian and Prucha (1999a) suggest the use of nonlinear least squares to obtain a consistent generalized moment estimator for  $\lambda$  from this system, which can then be used to obtain consistent estimators for the  $\beta$  in an FGLS approach. Since the  $\lambda$  is considered as a nuisance parameter, its significance (as a test for spatial autocorrelation) cannot be assessed, but its role is to provide a consistent estimator for the regression coefficients.<sup>26</sup>

A different approach is taken in the application of Hansen's (1982) generalized method of moments estimator (GMM) to spatial error autocorrelation in Conley (1996). This estimator is the standard minimizer of a quadratic form in the sample moment conditions, where the covariance matrix is obtained in nonparametric form as an application of the ideas of Newey and West (1987). Specifically, the spatial covariances are estimated from weighted averages of sample covariances for pairs of observations that are within a given distance band from each other. Note that this approach requires covariance stationarity, which is only satisfied for a restricted set of spatial processes (e.g. it does not apply to SAR error models).

Pinkse and Slade (1998) use a set of moment conditions to estimate a probit model with SAR errors. However, they focus on the induced heteroskedasticity of the process and do not explicitly deal with the spatial covariance structure.<sup>27</sup>

The relative efficiency of the new methods of moments approaches relative to the more traditional maximum likelihood techniques remains an area of active investigation.

### 4.4 Other estimation methods

A number of other approaches have been suggested to deal with the estimation of spatial regression models. An early technique is the so-called coding method, originally examined in Besag and Moran (1975).<sup>28</sup> This approach consists of

selecting a subsample from the data such that the relevant neighbors are removed (a non-contiguous subsample). This in effect eliminates the simultaneity bias in the spatial lag model, but at the cost of converting the model to a conditional one and with a considerable reduction of the sample size (down to 20 percent of the original sample for irregular lattice data). The advantage of this approach is that standard methods may be applied (e.g. for discrete choice models). However, it is not an efficient procedure and considerable arbitrariness is involved in the selection of the coding scheme.

Another increasingly common approach consists of the application of computational estimators to spatial models. A recent example is the recursive importance sampling (RIS) estimator (Vijverberg, 1997) applied to the spatial probit model in Beron and Vijverberg (2000).

A considerable literature also exists on Bayesian estimation of spatial models, but a detailed treatment of this is beyond the current scope.

## 5 SPECIFICATION TESTS

### 5.1 Moran's I

The most commonly used specification test for spatial autocorrelation is derived from a statistic developed by Moran (1948) as the two-dimensional analog of a test for univariate time series correlation (see also Cliff and Ord, 1973). In matrix notation, Moran's  $I$  statistic is

$$I = [N/S_0](\mathbf{e}' \mathbf{W} \mathbf{e} / \mathbf{e}' \mathbf{e}), \quad (14.30)$$

with  $\mathbf{e}$  as a vector of OLS residuals and  $S_0 = \sum_i \sum_j w_{ij}$ , a standardization factor that corresponds to the sum of the weights for the nonzero cross-products. The statistic shows a striking similarity to the familiar Durbin–Watson test.<sup>29</sup>

Moran's  $I$  test has been shown to be locally best invariant (King, 1981) and consistently outperforms other tests in terms of power in simulation experiments (for a recent review, see Anselin and Florax, 1995b). Its application has been extended to residuals in 2SLS regression in Anselin and Kelejian (1997), and to generalized residuals in probit models in Pinkse (2000). General formal conditions and proofs for the asymptotic normality of Moran's  $I$  in a wide range of regression models are given in Pinkse (1998) and Kelejian and Prucha (1999b). The consideration of Moran's  $I$  in conjunction with spatial heteroskedasticity is covered in Kelejian and Robinson (1998, 2000).

### 5.2 ML based tests

When spatial regression models are estimated by maximum likelihood, inference on the spatial autoregressive coefficients may be based on a Wald or asymptotic  $t$ -test (from the asymptotic variance matrix) or on a likelihood ratio test (see Anselin, 1988a, ch. 6; Anselin and Bera, 1998). Both approaches require that the

alternative model (i.e. the spatial model) be estimated. In contrast, a series of test statistics based on the Lagrange Multiplier (LM) or Rao Score (RS) principle only require estimation of the model under the null. The LM/RS tests also allow for the distinction between a spatial error and a spatial lag alternative.<sup>30</sup>

An LM/RS test against a spatial error alternative was originally suggested by Burridge (1980) and takes the form

$$\text{LM}_{\text{err}} = [\mathbf{e}' \mathbf{W} \mathbf{e} / (\mathbf{e}' \mathbf{e} / N)]^2 / [\text{tr}(\mathbf{W}^2 + \mathbf{W}' \mathbf{W})]. \quad (14.31)$$

This statistic has an asymptotic  $\chi^2(1)$  distribution and, apart from a scaling factor, corresponds to the square of Moran's  $I$ .<sup>31</sup> From several simulation experiments (Anselin and Rey, 1991; Anselin and Florax, 1995b) it follows that Moran's  $I$  has slightly better power than the  $\text{LM}_{\text{err}}$  test in small samples, but the performance of both tests becomes indistinguishable in medium and large size samples. The LM/RS test against a spatial lag alternative was outlined in Anselin (1988c) and takes the form

$$\text{LM}_{\text{lag}} = [\mathbf{e}' \mathbf{W} \mathbf{y} / (\mathbf{e}' \mathbf{e} / N)]^2 / D, \quad (14.32)$$

where  $D = [(W\mathbf{X}\beta)'(I - X(X'X)^{-1}X')(W\mathbf{X}\beta)/\sigma^2] + \text{tr}(\mathbf{W}^2 + \mathbf{W}' \mathbf{W})$ . This statistic also has an asymptotic  $\chi^2(1)$  distribution.

Since both tests have power against the other alternative, it is important to take account of possible lag dependence when testing for error dependence and vice versa. This can be implemented by means of a joint test (Anselin, 1988c) or by constructing tests that are robust to the presence of local misspecification of the other form (Anselin *et al.*, 1996).

The LM/RS principle can also be extended to more complex spatial alternatives, such as higher order processes, spatial error components and direct representation models (Anselin, 2000), to panel data settings (Anselin, 1988b), and to probit models (Pinkse, 1998, 2000; Pinkse and Slade, 1998). A common characteristic of the LM/RS tests against spatial alternatives is that they do not lend themselves readily to a formulation as an  $NR^2$  expression based on an auxiliary regression. However, as recently shown in Baltagi and Li (2000a), it is possible to obtain tests for spatial lag and spatial error dependence in a linear regression model by means of Davidson and MacKinnon's (1988) double length artificial regression approach.

## 6 IMPLEMENTATION ISSUES

To date, spatial econometric methods are not found in the main commercial econometric and statistical software packages, although macro and scripting facilities may be used to implement some estimators (Anselin and Hudak, 1992). The only comprehensive software to handle both estimation and specification testing of spatial regression models is the special-purpose SpaceStat package (Anselin, 1992b, 1998). A narrower set of techniques, such as maximum likelihood

estimation of spatial models is included in the Matlab routines of Pace and Barry (1998), and estimation of spatial error models is part of the S+Spatialstats add-on to S-Plus (MathSoft, 1996).<sup>32</sup>

In contrast to maximum likelihood estimation, method of moments and 2SLS can easily be implemented with standard software, provided that spatial lags can be computed. This requires the construction of a spatial weights matrix, which must often be derived from information in a geographic information system. Similarly, once a spatial lag can be computed, the LM/RS statistics are straightforward to implement.

The main practical problem is encountered in maximum likelihood estimation where the Jacobian determinant must be evaluated for every iteration in a nonlinear optimization procedure. The original solution to this problem was suggested by Ord (1975), who showed how the log Jacobian can be decomposed in terms that contain the eigenvalues of the weights matrix  $\omega_i$ ,

$$\ln |I - \rho W| = \sum_{i=1}^n \ln(1 - \rho \omega_i). \quad (14.33)$$

This is easy to implement in a standard optimization routine by treating the individual elements in the sum as observations on an auxiliary term in the log-likelihood (see Anselin and Hudak, 1992). However, the computation of the eigenvalues quickly becomes numerically unstable for matrices of more than 1,000 observations. In addition, for large data sets this approach is inefficient in that it does not exploit the high degree of sparsity of the spatial weights matrix. Recently suggested solutions to this problem fall into two categories. Approximate solutions avoid the computation of the Jacobian determinant, but instead approximate it by a polynomial function or by means of simulation methods (e.g. Barry and Pace, 1999). Exact solutions are based on Cholesky or LU decomposition methods that exploit the sparsity of the weights (Pace and Barry, 1997a, 1997b), or use a characteristic polynomial approach (Smirnov and Anselin, 2000). While much progress has been made, considerable work remains to be done to develop efficient algorithms and data structures to allow for the analysis of very large spatial data sets.

## 7 CONCLUDING REMARKS

This review chapter has been an attempt to present the salient issues pertaining to the methodology of spatial econometrics. It is by no means complete, but it is hoped that sufficient guidance is provided to pursue interesting research directions. Many challenging problems remain, both methodological in nature as well as in terms of applying the new techniques to meaningful empirical problems. Particularly in dealing with spatial effects in models other than the standard linear regression, much needs to be done to complete the spatial econometric toolbox. It is hoped that the review presented here will stimulate statisticians and econometricians to tackle these interesting and challenging problems.

## Notes

- \* This paper benefited greatly from comments by Wim Vijverberg and two anonymous referees. A more comprehensive version of this paper is available as Anselin (1999).
- 1 A more extensive review is given in Anselin and Bera (1998) and Anselin (1999).
- 2 An extensive collection of recent applications of spatial econometric methods in economics can be found in Anselin and Florax (2000).
- 3 In this chapter, I will use the terms spatial dependence and spatial autocorrelation interchangeably. Obviously, the two are not identical, but typically, the weaker form is used, in the sense of a moment of a joint distribution. Only seldom is the focus on the complete joint density (a recent exception can be found in Brett and Pinkse (1997)).
- 4 See Anselin (1988a), for a more extensive discussion.
- 5 One would still need to establish the class of spatial stochastic processes that would allow for the consistent estimation of the covariance; see Frees (1995) for a discussion of the general principles.
- 6 See Anselin and Bera (1998) for an extensive and technical discussion.
- 7 On a square grid, one could envisage using North, South, East and West as spatial shifts, but in general, for irregular spatial units such as counties, this is impractical, since the number of neighbors for each county is not constant.
- 8 In Anselin (1988a), the term spatial lag is introduced to refer to this new variable, to emphasize the similarity to a distributed lag term rather than a spatial shift.
- 9 By convention,  $w_{ii} = 0$ , i.e. a location is never a neighbor of itself. This is arbitrary, but can be assumed without loss of generality. For a more extensive discussion of spatial weights, see Anselin (1988a, ch. 3), Cliff and Ord (1981), Upton and Fingleton (1985).
- 10 See Anselin and Bera (1998) for further details.
- 11 See McMillen (1992) for an illustration.
- 12 The specification of spatial covariance functions is not arbitrary, and a number of conditions must be satisfied in order for the model to be "valid" (details are given in Cressie (1993, pp. 61–3, 67–8 and 84–6)).
- 13 Specifically, this may limit the applicability of GMM estimators that are based on a central limit theorem for stationary mixing random fields such as the one by Bolthausen (1982), used by Conley (1996).
- 14 Cressie (1993, pp. 100–1).
- 15 See Kelejian and Prucha (1999a, 1999b).
- 16 For ease of exposition, the error term is assumed to be iid, although various forms of heteroskedasticity can be incorporated in a straightforward way (Anselin, 1988a, ch. 6).
- 17 Details and a review of alternative specifications are given in Anselin and Bera (1998).
- 18 For further details, see Anselin (1988a, 1988b). A recent application is Baltagi and Li (2000b).
- 19 Methodological issues associated with spatial probit models are considered in Case (1992), McMillen (1992), Pinkse and Slade (1998) and Beron and Vijverberg (2000).
- 20 For an extensive discussion, see Beron and Vijverberg (2000).
- 21 Other classic treatments of ML estimation in spatial models can be found in Whittle (1954), Besag (1974), and Mardia and Marshall (1984).
- 22 For a formal demonstration, see Anselin (1988a) and Kelejian and Prucha (1997).
- 23 For details, see, e.g. McMillen (1992), Pinkse and Slade (1998), Beron and Vijverberg (2000), and also, for general principles, Poirier and Ruud (1988).
- 24 For a careful consideration of these issues, see Kelejian and Prucha (1999a).
- 25 For technical details, see, e.g. Kelejian and Robinson (1993), Kelejian and Prucha (1998).

- 26 A recent application of this method is given in Bell and Bockstael (2000). An extension of this idea to the residuals of a spatial 2SLS estimation is provided in Kelejian and Prucha (1998).
- 27 See also Case (1992) and McMillen (1992) for a similar focus on heteroskedasticity in the spatial probit model.
- 28 See also the discussion in Haining (1990, pp. 131–3).
- 29 For example, for row-standardized weights,  $S_0 = N$ , and  $I = e'We/e'e$ . See Anselin and Bera (1998) for an extensive discussion.
- 30 Moran's  $I$  is not based on an explicit alternative and has power against both (see Anselin and Rey, 1991).
- 31 As shown in Anselin and Kelejian (1997) these tests are asymptotically equivalent.
- 32 Neither of these toolboxes include specification tests. Furthermore, S+Spatialstats has no routines to handle the spatial lag model.

## References

- Anselin, L. (1980). *Estimation Methods for Spatial Autoregressive Structures*. Regional Science Dissertation and Monograph Series 8. Field of Regional Science, Cornell University, Ithaca, N.Y.
- Anselin, L. (1988a). *Spatial Econometrics: Methods and Models*. Kluwer Academic, Dordrecht.
- Anselin, L. (1988b). A test for spatial autocorrelation in seemingly unrelated regressions. *Economics Letters* 28, 335–41.
- Anselin, L. (1988c). Lagrange multiplier test diagnostics for spatial dependence and spatial heterogeneity. *Geographical Analysis* 20, 1–17.
- Anselin, L. (1990). Some robust approaches to testing and estimation in spatial econometrics. *Regional Science and Urban Economics* 20, 141–63.
- Anselin, L. (1992a). Space and applied econometrics. Special Issue, *Regional Science and Urban Economics* 22.
- Anselin, L. (1992b). *SpaceStat, a Software Program for the Analysis of Spatial Data*. National Center for Geographic Information and Analysis, University of California, Santa Barbara, CA.
- Anselin, L. (1998). *SpaceStat Version 1.90*. <http://www.spacetstat.com>.
- Anselin, L. (1999). Spatial econometrics, An updated review. Regional Economics Applications Laboratory (REAL), University of Illinois, Urbana-Champaign.
- Anselin, L. (2000). Rao's score test in spatial econometrics. *Journal of Statistical Planning and Inference* (forthcoming).
- Anselin, L., and A. Bera (1998). Spatial dependence in linear regression models with an introduction to spatial econometrics. In A. Ullah and D.E.A. Giles (eds.) *Handbook of Applied Economic Statistics*, pp. 237–89. New York: Marcel Dekker.
- Anselin, L., and R. Florax (1995a). Introduction. In L. Anselin and R. Florax (eds.) *New Directions in Spatial Econometrics*, pp. 3–18. Berlin: Springer-Verlag.
- Anselin, L., and R. Florax (1995b). Small sample properties of tests for spatial dependence in regression models: some further results. In L. Anselin and R. Florax (eds.) *New Directions in Spatial Econometrics*, pp. 21–74. Berlin: Springer-Verlag.
- Anselin, L., and R. Florax (2000). *Advances in Spatial Econometrics*. Heidelberg: Springer-Verlag.
- Anselin, L., and S. Hudak (1992). Spatial econometrics in practice, a review of software options. *Regional Science and Urban Economics* 22, 509–36.
- Anselin, L., and H.H. Kelejian (1997). Testing for spatial error autocorrelation in the presence of endogenous regressors. *International Regional Science Review* 20, 153–82.

- Anselin, L., and S. Rey (1991). Properties of tests for spatial dependence in linear regression models. *Geographical Analysis* 23, 112–31.
- Anselin, L., and S. Rey (1997). Introduction to the special issue on spatial econometrics. *International Regional Science Review* 20, 1–7.
- Anselin, L., A. Bera, R. Florax, and M. Yoon (1996). Simple diagnostic tests for spatial dependence. *Regional Science and Urban Economics* 26, 77–104.
- Baltagi, B., and D. Li (2000a). Double length artificial regressions for testing spatial dependence. *Econometric Review*, forthcoming.
- Baltagi, B., and D. Li (2000b). Prediction in the panel data model with spatial correlation. In L. Anselin and R. Florax (eds.) *Advances in Spatial Econometrics*. Heidelberg: Springer-Verlag.
- Barry, R.P., and R.K. Pace (1999). Monte Carlo estimates of the log determinant of large sparse matrices. *Linear Algebra and its Applications* 289, 41–54.
- Basu, S., and T.G. Thibodeau (1998). Analysis of spatial autocorrelation in housing prices. *Journal of Real Estate Finance and Economics* 17, 61–85.
- Bell, K.P., and N.E. Bockstael (2000). Applying the generalized moments estimation approach to spatial problems involving micro-level data. *Review of Economics and Statistics* 82, 72–82.
- Beron, K.J., and W.P.M. Vijverberg (2000). Probit in a spatial context: a Monte Carlo approach. In L. Anselin and R. Florax (eds.) *Advances in Spatial Econometrics*. Heidelberg: Springer-Verlag.
- Besag, J. (1974). Spatial interaction and the statistical analysis of lattice systems. *Journal of the Royal Statistical Society B* 36, 192–225.
- Besag, J., and P.A.P. Moran (1975). On the estimation and testing of spatial interaction in Gaussian lattice processes. *Biometrika* 62, 555–62.
- Bolthausen, E. (1982). On the central limit theorem for stationary mixing random fields. *Annals of Probability* 10, 1047–50.
- Brett, C., and J. Pinkse (1997). Those taxes all over the map! A test for spatial independence of municipal tax rates in British Columbia. *International Regional Science Review* 20, 131–51.
- Brock, W.A., and S.N. Durlauf (1995). Discrete choice with social interactions I: Theory. NBER Working Paper No. W5291. Cambridge, MA: National Bureau of Economic Research.
- Burridge, P. (1980). On the Cliff-Ord test for spatial autocorrelation. *Journal of the Royal Statistical Society B* 42, 107–8.
- Case, A. (1992). Neighborhood influence and technological change. *Regional Science and Urban Economics* 22, 491–508.
- Case, A., H.S. Rosen, and J.R. Hines (1993). Budget spillovers and fiscal policy interdependence: evidence from the States. *Journal of Public Economics* 52, 285–307.
- Cliff, A., and J.K. Ord (1973). *Spatial Autocorrelation*. London: Pion.
- Cliff, A., and J.K. Ord (1981). *Spatial Processes: Models and Applications*. London: Pion.
- Conley, T.G. (1996). *Econometric modelling of cross-sectional dependence*. Ph.D. dissertation. Department of Economics, University of Chicago, Chicago, IL.
- Cressie, N. (1993). *Statistics for Spatial Data*. New York: Wiley.
- Davidson, J. (1994). *Stochastic Limit Theory*. Oxford: Oxford University Press.
- Davidson, R. and J.G. MacKinnon (1988). Double-length artificial regressions. *Oxford Bulletin of Economics and Statistics* 50, 203–17.
- Driscoll, J.C., and A.C. Kraay (1998). Consistent covariance matrix estimation with spatially dependent panel data. *Review of Economics and Statistics* 80, 549–60.
- Dubin, R. (1988). Estimation of regression coefficients in the presence of spatially auto-correlated error terms. *Review of Economics and Statistics* 70, 466–74.

- Dubin, R. (1992). Spatial autocorrelation and neighborhood quality. *Regional Science and Urban Economics* 22, 433–52.
- Frees, E.W. (1995). Assessing cross-sectional correlation in panel data. *Journal of Econometrics* 69, 393–414.
- Haining, R. (1990). *Spatial Data Analysis in the Social and Environmental Sciences*. Cambridge: Cambridge University Press.
- Hansen, L.P. (1982). Large sample properties of generalized method of moments estimators. *Econometrica* 50, 1029–54.
- Kelejian, H., and I. Prucha (1997). Estimation of spatial regression models with autoregressive errors by two stage least squares procedures: a serious problem. *International Regional Science Review* 20, 103–11.
- Kelejian, H., and I. Prucha (1998). A generalized spatial two stage least squares procedure for estimating a spatial autoregressive model with autoregressive disturbances. *Journal of Real Estate Finance and Economics* 17, 99–121.
- Kelejian, H., and I. Prucha (1999a). A generalized moments estimator for the autoregressive parameter in a spatial model. *International Economic Review* 40, 509–33.
- Kelejian, H.H., and I. Prucha (1999b). On the asymptotic distribution of the Moran I test statistic with applications. Working Paper, Department of Economics, University of Maryland, College Park, MD.
- Kelejian, H.H., and D.P. Robinson (1993). A suggested method of estimation for spatial interdependent models with autocorrelated errors, and an application to a county expenditure model. *Papers in Regional Science* 72, 297–312.
- Kelejian, H.H., and D.P. Robinson (1998). A suggested test for spatial autocorrelation and/or heteroskedasticity and corresponding Monte Carlo results. *Regional Science and Urban Economics* 28, 389–417.
- Kelejian, H.H., and D.P. Robinson (2000). The influence of spatially correlated heteroskedasticity on tests for spatial correlation. In L. Anselin and R. Florax (eds.) *Advances in Spatial Econometrics*. Heidelberg: Springer-Verlag (forthcoming).
- King, M. (1981). A small sample property of the Cliff-Ord test for spatial correlation. *Journal of the Royal Statistical Society B* 43, 263–4.
- Lahiri, S.N. (1996). On the inconsistency of estimators under infill asymptotics for spatial data. *Sankhya A* 58, 403–17.
- Magnus, J. (1978). Maximum likelihood estimation of the GLS model with unknown parameters in the disturbance covariance matrix. *Journal of Econometrics* 7, 281–312. (Corrigenda, *Journal of Econometrics* 10, 261).
- Mardia, K.V., and R.J. Marshall (1984). Maximum likelihood estimation of models for residual covariance in spatial regression. *Biometrika* 71, 135–46.
- MathSoft (1996). *S+SpatialStats User's Manual for Windows and Unix*. Seattle, WA: MathSoft, Inc.
- McMillen, D.P. (1992). Probit with spatial autocorrelation. *Journal of Regional Science* 32, 335–48.
- Moran, P.A.P. (1948). The interpretation of statistical maps. *Biometrika* 35, 255–60.
- Newey, W.K., and K.D. West (1987). A simple, positive semi-definite, heteroskedasticity and autocorrelation consistent covariance matrix. *Econometrica* 55, 703–8.
- Ord, J.K. (1975). Estimation methods for models of spatial interaction. *Journal of the American Statistical Association* 70, 120–6.
- Pace, R.K., and R. Barry (1997a). Sparse spatial autoregressions. *Statistics and Probability Letters* 33, 291–7.
- Pace, R.K., and R. Barry (1997b). Quick computation of spatial autoregressive estimators. *Geographical Analysis* 29, 232–46.

- Pace, R.K., and R. Barry (1998). *Spatial Statistics Toolbox 1.0*. Real Estate Research Institute, Louisiana State University, Baton Rouge, LA.
- Pace, R.K., R. Barry, C.F. Sirmans (1998). Spatial statistics and real estate. *Journal of Real Estate Finance and Economics* 17, 5–13.
- Paelinck, J., and L. Klaassen (1979). *Spatial Econometrics*. Farnborough: Saxon House.
- Pinkse, J. (1998). Asymptotic properties of the Moran and related tests and a test for spatial correlation in probit models. Working Paper, Department of Economics, University of British Columbia, Vancouver, BC.
- Pinkse, J. (2000). Moran-flavored tests with nuisance parameters. In L. Anselin and R. Florax (eds.) *Advances in Spatial Econometrics*. Heidelberg: Springer-Verlag.
- Pinkse, J., and M.E. Slade (1998). Contracting in space: an application of spatial statistics to discrete-choice models. *Journal of Econometrics* 85, 125–54.
- Poirier, D.J., and P.A. Ruud (1988). Probit with dependent observations. *Review of Economic Studies* 55, 593–614.
- Pötscher, B.M., and I.R. Prucha (1997). *Dynamic Nonlinear Econometric Models*. Berlin: Springer.
- Rao, C.R. (1973). *Linear Statistical Inference and its Applications* (2nd edn). New York: Wiley.
- Schmidt, P. (1976). *Econometrics*. New York: Marcel Dekker.
- Smirnov, O., and L. Anselin (2000). Fast maximum likelihood estimation of very large spatial autoregressive models: a characteristic polynomial approach. *Computational Statistics and Data Analysis*.
- Upton, G.J., and B. Fingleton (1985). *Spatial Data Analysis by Example. Volume 1: Point Pattern and Quantitative Data*. New York: Wiley.
- Vijverberg, W. (1997). Monte Carlo evaluation of multivariate normal probabilities. *Journal of Econometrics* 76, 281–307.
- Whittle, P. (1954). On stationary processes in the plane. *Biometrika* 41, 434–49.

CHAPTER FIFTEEN

# Essentials of Count Data Regression

*A. Colin Cameron and Pravin K. Trivedi*

## 1 INTRODUCTION

In many economic contexts the dependent or response variable of interest ( $y$ ) is a nonnegative integer or count which we wish to explain or analyze in terms of a set of covariates ( $x$ ). Unlike the classical regression model, the response variable is discrete with a distribution that places probability mass at nonnegative integer values only. Regression models for counts, like other limited or discrete dependent variable models such as the logit and probit, are nonlinear with many properties and special features intimately connected to discreteness and nonlinearity.

Let us consider some examples from microeconomics, beginning with samples of independent cross section observations. Fertility studies often model the number of live births over a specified age interval of the mother, with interest in analyzing its variation in terms of, say, mother's schooling, age, and household income (Winkelmann, 1995). Accident analysis studies model airline safety, for example, as measured by the number of accidents experienced by an airline over some period, and seek to determine its relationship to airline profitability and other measures of the financial health of the airline (Rose, 1990). Recreational demand studies seek to place a value on natural resources such as national forests by modeling the number of trips to a recreational site (Gurmu and Trivedi, 1996). Health demand studies model data on the number of times that individuals consume a health service, such as visits to a doctor or days in hospital in the past year (Cameron, Trivedi, Milne and Piggott, 1988), and estimate the impact of health status and health insurance.

Examples of count data regression based on time series and panel data are also available. A time series example is the annual number of bank failures over some period, which may be analyzed using explanatory variables such as bank

profitability, corporate profitability, and bank borrowings from the Federal Reserve Bank (Davutyan, 1989). A panel data example that has attracted much attention in the industrial organization literature on the benefits of research and development expenditures is the number of patents received annually by firms (Hausman, Hall, and Griliches, 1984).

In some cases, such as number of births, the count is the variable of ultimate interest. In other cases, such as medical demand and results of research and development expenditure, the variable of ultimate interest is continuous, often expenditures or receipts measured in dollars, but the best data available are, instead, a count.

In all cases the data are concentrated on a few small discrete values, say 0, 1, and 2; skewed to the left; and intrinsically heteroskedastic with variance increasing with the mean. In many examples, such as number of births, virtually all the data are restricted to single digits, and the mean number of events is quite low. But in other cases, such as number of patents, the tail can be very long with, say, one-quarter of the sample being awarded no patents while one firm is awarded 400 patents.

These features motivate the application of special methods and models for count regression. There are two ways to proceed. The first approach is a fully parametric one that completely specifies the distribution of the data, fully respecting the restriction of  $y$  to nonnegative integer values. The second approach is a mean-variance approach, which specifies the conditional mean to be nonnegative, and specifies the conditional variance to be a function of the conditional mean.

These approaches are presented for cross section data in Sections 2 to 4. Section 2 details the Poisson regression model. This model is often too restrictive and other, more commonly-used, fully parametric count models are presented in Section 3. Less-used alternative parametric approaches for counts, such as discrete choice models and duration models, are also presented in this section. The partially parametric approach of modeling the conditional mean and conditional variance is detailed in Section 4. Extensions to other types of data, notably time series, multivariate and panel data, are given in Section 5. In Section 6 practical recommendations are provided. For pedagogical reasons the Poisson regression model for cross section data is presented in some detail. The other models, many superior to Poisson, are presented in less detail for space reasons. For more complete treatment see Cameron and Trivedi (1998) and the guide to further reading in Section 7.

## 2 POISSON REGRESSION

The Poisson is the starting point for count data analysis, though it is often inadequate. In Sections 2.1–2.3 we present the Poisson regression model and estimation by maximum likelihood, interpretation of the estimated coefficients, and extensions to truncated and censored data. Limitations of the Poisson model, notably overdispersion, are presented in Section 2.4.

## 2.1 Poisson MLE

The natural stochastic model for counts is a Poisson point process for the occurrence of the event of interest. This implies a Poisson distribution for the number of occurrences of the event, with density

$$\Pr[Y = y] = \frac{e^{-\mu} \mu^y}{y!}, \quad y = 0, 1, 2, \dots, \quad (15.1)$$

where  $\mu$  is the intensity or rate parameter. We refer to the distribution as  $P[\mu]$ . The first two moments are

$$\begin{aligned} E[Y] &= \mu, \\ V[Y] &= \mu. \end{aligned} \quad (15.2)$$

This shows the well known equality of mean and variance property of the Poisson distribution.

By introducing the observation subscript  $i$ , attached to both  $y$  and  $\mu$ , the framework is extended to non-iid data. The *Poisson regression model* is derived from the Poisson distribution by parameterizing the relation between the mean parameter  $\mu$  and covariates (regressors)  $x$ . The standard assumption is to use the exponential mean parameterization,

$$\mu_i = \exp(x'_i \beta), \quad i = 1, \dots, n, \quad (15.3)$$

where by assumption there are  $k$  linearly independent covariates, usually including a constant. Because  $V[y_i | x_i] = \exp(x'_i \beta)$ , by (15.2) and (15.3), the Poisson regression is intrinsically heteroskedastic.

Given (15.1) and (15.3) and the assumption that the observations  $(y_i | x_i)$  are independent, the most natural estimator is maximum likelihood (ML). The loglikelihood function is

$$\ln L(\beta) = \sum_{i=1}^n \{y_i x'_i \beta - \exp(x'_i \beta) - \ln y_i!\}. \quad (15.4)$$

The Poisson MLE (maximum likelihood estimation), denoted  $\hat{\beta}_p$ , is the solution to  $k$  nonlinear equations corresponding to the first-order condition for maximum likelihood,

$$\sum_{i=1}^n (y_i - \exp(x'_i \beta)) x_i = 0. \quad (15.5)$$

If  $x_i$  includes a constant term then the residuals  $y_i - \exp(x'_i \beta)$  sum to zero by (15.5). The loglikelihood function is globally concave; hence solving these equations by

Gauss–Newton or Newton–Raphson iterative algorithm yields unique parameters estimates.

By standard maximum likelihood theory of correctly specified models, the estimator  $\hat{\beta}_P$  is consistent for  $\beta$  and asymptotically normal with the sample covariance matrix

$$V[\hat{\beta}_P] = \left( \sum_{i=1}^n \mu_i x_i x_i' \right)^{-1}, \quad (15.6)$$

in the case where  $\mu_i$  is of the exponential form (15.3). In practice an alternative more general form for the variance matrix should be used; see Section 4.1.

## 2.2 Interpretation of regression coefficients

For linear models, with  $E[y|x] = x'\beta$ , the coefficients  $\beta$  are readily interpreted as the effect of a one-unit change in regressors on the conditional mean. For nonlinear models this interpretation needs to be modified. For any model with exponential conditional mean, differentiation yields

$$\frac{\partial E[y|x]}{\partial x_j} = \beta_j \exp(x'\hat{\beta}), \quad (15.7)$$

where the scalar  $x_j$  denotes the  $j$ th regressor. For example, if  $\hat{\beta}_j = 0.25$  and  $\exp(x'_j\hat{\beta}) = 3$ , then a one-unit change in the  $j$ th regressor increases the expectation of  $y$  by 0.75 units. This partial response depends upon  $\exp(x'_j\hat{\beta})$  which is expected to vary across individuals. It is easy to see that  $\beta_j$  measures the relative change in  $E[y|x]$  induced by a unit change in  $x_j$ . If  $x_j$  is measured on log-scale,  $\beta_j$  is an elasticity.

For purposes of reporting a single response value, a good candidate is an estimate of the *average response*,  $\frac{1}{n} \sum_{i=1}^n \partial E[y_i|x_i]/\partial x_{ij} = \hat{\beta}_j \times \frac{1}{n} \sum_{i=1}^n \exp(x'_i\hat{\beta})$ . For Poisson regression models with intercept included, this can be shown to simplify to  $\hat{\beta}_j \bar{y}$ .

Another consequence of (15.7) is that if, say,  $\beta_j$  is twice as large as  $\beta_k$ , then the effect of changing the  $j$ th regressor by one unit is twice that of changing the  $k$ th regressor by one unit.

## 2.3 Truncation and censoring

In some studies, inclusion in the sample requires that sampled individuals have been engaged in the activity of interest. Then the count data are *truncated*, as the data are observed only over part of the range of the response variable. Examples of truncated counts include the number of bus trips made per week in surveys taken on buses, the number of shopping trips made by individuals sampled at a mall, and the number of unemployment spells among a pool of unemployed. In all these cases we do not observe zero counts, so the data are said to be

zero-truncated, or more generally left-truncated. Right truncation results from loss of observations greater than some specified value.

Truncation leads to inconsistent parameter estimates unless the likelihood function is suitably modified. Consider the case of zero truncation. Let  $f(y | \theta)$  denote the density function and  $F(y | \theta) = \Pr[Y \leq y]$  denote the cumulative distribution function of the discrete random variable, where  $\theta$  is a parameter vector. If realizations of  $y$  less than a positive integer 1 are omitted, the ensuing zero-truncated density is given by

$$f(y | \theta, y \geq 1) = \frac{f(y | \theta)}{1 - F(0 | \theta)}, \quad y = 1, 2, \dots. \quad (15.8)$$

This specializes in the zero-truncated Poisson case, for example, to  $f(y | \mu, y \geq 1) = e^{-\mu} \mu^y / [y!(1 - \exp(-\mu))]$ . It is straightforward to construct a loglikelihood based on this density and to obtain maximum likelihood estimates.

*Censored* counts most commonly arise from aggregation of counts greater than some value. This is often done in survey design when the total probability mass over the aggregated values is relatively small. Censoring, like truncation, leads to inconsistent parameter estimates if the uncensored likelihood is mistakenly used.

For example, the number of events greater than some known value  $c$  might be aggregated into a single category. Then some values of  $y$  are incompletely observed; the precise value is unknown but it is known to equal or exceed  $c$ . The observed data has density

$$g(y | \theta) = \begin{cases} f(y | \theta) & \text{if } y < c, \\ 1 - F(c | \theta) & \text{if } y \geq c, \end{cases} \quad (15.9)$$

where  $c$  is known. Specialization to the Poisson, for example, is straightforward.

A related complication is that of *sample selection* (Terza, 1998). Then the count  $y$  is observed only when another random variable, potentially correlated with  $y$ , crosses a threshold. For example, to see a medical specialist one may first need to see a general practitioner. Treatment of count data with sample selection is a current topic of research.

## 2.4 Overdispersion

The Poisson regression model is usually too restrictive for count data, leading to alternative models as presented in Sections 3 and 4. The fundamental problem is that the distribution is parameterized in terms of a single scalar parameter ( $\mu$ ) so that all moments of  $y$  are a function of  $\mu$ . By contrast the normal distribution has separate parameters for location ( $\mu$ ) and scale ( $\sigma^2$ ). (For the same reason the one-parameter exponential is too restrictive for duration data and more general two-parameter distributions such as the Weibull are superior. Note that this complication does not arise with binary data. Then the distribution is clearly the

one-parameter Bernoulli, as if the probability of success is  $p$  then the probability of failure must be  $1 - p$ . For binary data the issue is instead how to parameterize  $p$  in terms of regressors.)

One way this restrictiveness manifests itself is that in many applications a Poisson density predicts the probability of a zero count to be considerably less than is actually observed in the sample. This is termed the *excess zeros* problem, as there are more zeros in the data than the Poisson predicts.

A second and more obvious way that the Poisson is deficient is that for count data the variance usually exceeds the mean, a feature called *overdispersion*. The Poisson instead implies equality of variance of mean, see (15.2), a property called *equidispersion*.

Overdispersion has qualitatively similar consequences to the failure of the assumption of homoskedasticity in the linear regression model. Provided the conditional mean is correctly specified, that is (15.3) holds, the Poisson MLE is still consistent. This is clear from inspection of the first-order conditions (15.5), since the left-hand side of (15.5) will have an expected value of zero if  $E[y_i | x_i] = \exp(x_i'\beta)$ . (This consistency property applies more generally to the quasi-MLE when the specified density is in the linear exponential family (LEF). Both Poisson and normal are members of the LEF.) It is nonetheless important to control for overdispersion for two reasons. First, in more complicated settings such as with truncation and censoring, overdispersion leads to the more fundamental problem of inconsistency. Second, even in the simplest settings large overdispersion leads to grossly deflated standard errors and grossly inflated  $t$ -statistics in the usual ML output.

A statistical test of overdispersion is therefore highly desirable after running a Poisson regression. Most count models with overdispersion specify overdispersion to be of the form

$$V[y_i | x_i] = \mu_i + \alpha g(\mu_i), \quad (15.10)$$

where  $\alpha$  is an unknown parameter and  $g(\cdot)$  is a known function, most commonly  $g(\mu) = \mu^2$  or  $g(\mu) = \mu$ . It is assumed that under both null and alternative hypotheses the mean is correctly specified as, for example,  $\exp(x_i'\beta)$ , while under the null hypothesis  $\alpha = 0$  so that  $V[y_i | x_i] = \mu_i$ . A simple test statistic for  $H_0 : \alpha = 0$  versus  $H_1 : \alpha \neq 0$  or  $H_1 : \alpha > 0$  can be computed by estimating the Poisson model, constructing fitted values  $\hat{\mu}_i = \exp(x_i'\hat{\beta})$  and running the auxiliary OLS regression (without constant)

$$\frac{(y_i - \hat{\mu}_i)^2 - y_i}{\hat{\mu}_i} = \alpha \frac{g(\hat{\mu}_i)}{\hat{\mu}_i} + u_i, \quad (15.11)$$

where  $u_i$  is an error term. The reported  $t$ -statistic for  $\alpha$  is asymptotically normal under the null hypothesis of no overdispersion. This test can also be used for *underdispersion*, in which case the conditional variance is less than the conditional mean.

### 3 OTHER PARAMETRIC COUNT REGRESSION MODELS

Various models that are less restrictive than Poisson are presented in this section.

First, overdispersion in count data may be due to unobserved heterogeneity. Then counts are viewed as being generated by a Poisson process, but the researcher is unable to correctly specify the rate parameter of this process. Instead the rate parameter is itself a random variable. This mixture approach, presented in Sections 3.1–3.2, leads to the widely-used negative binomial model.

Second, overdispersion, and in some cases underdispersion, may arise because the process generating the first event may differ from that determining later events. For example, an initial doctor consultation may be solely a patient's choice, while subsequent visits are also determined by the doctor. This leads to the hurdle model, presented in Section 3.3.

Third, overdispersion in count data may be due to failure of the assumption of independence of events which is implicit in the Poisson process. One can introduce dependence so that, for example, the occurrence of one doctor visit makes subsequent doctor visits more likely. This approach has not been widely used in count data analysis. (In duration data analysis this is called true state dependence, to be contrasted with the first approach of unobserved heterogeneity.) Particular assumptions again lead to the negative binomial; see also Winkelmann (1995). A discrete choice model that progressively models  $\Pr[y = j | y \geq j - 1]$  is presented in Section 3.4, and issues of dependence also arise in Section 5 on time series.

Fourth, one can refer to the extensive and rich literature on univariate iid count distributions, which offers intriguing possibilities such as the logarithmic series and hypergeometric distribution (Johnson, Kotz, and Kemp, 1992). New regression models can be developed by letting one or more parameters be a specified function of regressors. Such models are not presented here. The approach has less motivation than the first three approaches and the resulting models may not be any better.

#### 3.1 Continuous mixture models

The negative binomial model can be obtained in many different ways. The following justification using a mixture distribution is one of the oldest and has wide appeal.

Suppose the distribution of a random count  $y$  is Poisson, conditional on the parameter  $\lambda$ , so that  $f(y | \lambda) = \exp(-\lambda)\lambda^y/y!$ . Suppose now that the parameter  $\lambda$  is random, rather than being a completely deterministic function of regressors  $x$ . In particular, let  $\lambda = \mu v$ , where  $\mu$  is a deterministic function of  $x$ , for example  $\exp(x'\beta)$ , and  $v > 0$  is iid distributed with density  $g(v | \alpha)$ . This is an example of *unobserved heterogeneity*, as different observations may have different  $\lambda$  (heterogeneity) but part of this difference is due to a random (unobserved) component  $v$ .

The marginal density of  $y$ , unconditional on the random parameter  $v$  but conditional on the deterministic parameters  $\mu$  and  $\alpha$ , is obtained by integrating out  $v$ . This yields

$$h(y | \mu, \alpha) = \int f(y | \mu, v) g(v | \alpha) dv, \quad (15.12)$$

where  $g(v | \alpha)$  is called the *mixing distribution* and  $\alpha$  denotes the unknown parameter of the mixing distribution. The integration defines an “average” distribution. For some specific choices of  $f(\cdot)$  and  $g(\cdot)$ , the integral will have an analytical or closed-form solution.

If  $f(y | \lambda)$  is the Poisson density and  $g(v)$ ,  $v > 0$ , is the gamma density with  $E[v] = 1$  and  $V[v] = \alpha$  we obtain the negative binomial density

$$h(y | \mu, \alpha) = \frac{\Gamma(\alpha^{-1} + y)}{\Gamma(\alpha^{-1})\Gamma(y+1)} \left( \frac{\alpha^{-1}}{\alpha^{-1} + \mu} \right)^{\alpha^{-1}} \left( \frac{\mu}{\mu + \alpha^{-1}} \right)^y, \quad \alpha > 0, \quad (15.13)$$

where  $\Gamma(\cdot)$  denotes the gamma integral which specializes to a factorial for an integer argument. Special cases of the negative binomial include the Poisson ( $\alpha = 0$ ) and the geometric ( $\alpha = 1$ ).

The first two moments of the negative binomial distribution are

$$\begin{aligned} E[y | \mu, \alpha] &= \mu, \\ V[y | \mu, \alpha] &= \mu(1 + \alpha\mu). \end{aligned} \quad (15.14)$$

The variance therefore exceeds the mean, since  $\alpha > 0$  and  $\mu > 0$ . Indeed it can be shown easily that overdispersion always arises if  $y | \lambda$  is Poisson and the mixing is of the form  $\lambda = \mu v$  where  $E[v] = 1$ . Note also that the overdispersion is of the form (15.10) discussed in Section 2.4.

Two standard variants of the negative binomial are used in regression applications. Both variants specify  $\mu_i = \exp(x_i \beta)$ . The most common variant lets  $\alpha$  be a parameter to be estimated, in which case the conditional variance function,  $\mu + \alpha\mu^2$  from (15.14), is quadratic in the mean. The loglikelihood is easily obtained from (15.13), and estimation is by maximum likelihood.

The other variant of the negative binomial model has a linear variance function,  $V[y | \mu, \alpha] = (1 + \delta)\mu$ , obtained by replacing  $\alpha$  by  $\delta/\mu$  throughout (15.13). Estimation by ML is again straightforward. Sometimes this variant is called negative binomial 1 (NB1) in contrast to the variant with a quadratic variance function which has been called negative binomial 2 (NB2) model (Cameron and Trivedi, 1998).

The negative binomial model with quadratic variance function has been found to be very useful in applied work. It is the standard cross section model for counts, which are usually overdispersed, along with the quasi-MLE of Section 4.1.

For mixtures other than Poisson-gamma, such as those that instead use as mixing distribution the lognormal distribution or the inverse-Gaussian distribution, the marginal distribution cannot be expressed in a closed form. Then one may have to use numerical quadrature or simulated maximum likelihood to estimate the model. These methods are entirely feasible with currently available

computing power. If one is prepared to use simulation-based estimation methods, see Gouriéroux and Monfort (1997), the scope for using mixed-Poisson models of various types is very extensive.

### 3.2 Finite mixture models

The mixture model in the previous subsection was a continuous mixture model, as the mixing random variable  $v$  was assumed to have continuous distribution. An alternative approach instead uses a *discrete* representation of unobserved heterogeneity, which generates a class of models called *finite mixture models*. This class of models is a particular subclass of *latent class models*.

In empirical work the more commonly used alternative to the continuous mixture is in the class of modified count models discussed in the next section. However, it is more natural to follow up the preceding section with a discussion of finite mixtures. Further, the subclass of modified count models can be viewed as a special case of finite mixtures.

We suppose that the density of  $y$  is a linear combination of  $m$  different densities, where the  $j$ th density is  $f_j(y | \lambda_j)$ ,  $j = 1, 2, \dots, m$ . Thus an  $m$ -component finite mixture is

$$f(y | \lambda, \pi) = \sum_{j=1}^m \pi_j f_j(y | \lambda_j), \quad 0 < \pi_j < 1, \quad \sum_{j=1}^m \pi_j = 1. \quad (15.15)$$

For example, in a study of the use of medical services with  $m = 2$ , the first density may correspond to heavy users of the service and the second to relatively low users, and the fractions of the two types in the populations are  $\pi_1$  and  $\pi_2 (= 1 - \pi_1)$  respectively.

The goal of the researcher who uses this model is to estimate the unknown parameters  $\lambda_j$ ,  $j = 1, \dots, m$ . It is easy to develop regression models based on (15.15). For example, if NB2 models are used then  $f_j(y | \lambda_j)$  is the NB2 density (15.13) with parameters  $\mu_j = \exp(x'\beta_j)$  and  $\alpha_j$ , so  $\lambda_j = (\beta_j, \alpha_j)$ . If the number of components,  $m$ , is given, then under some regularity conditions maximum likelihood estimation of the parameters  $(\pi_j, \lambda_j)$ ,  $j = 1, \dots, m$ , is possible. The details of the estimation methods, less straightforward due to the presence of the mixing parameters  $\pi_j$ , is omitted here because of space constraints. See Cameron and Trivedi (1998, ch. 4). It is possible also to probabilistically assign each case to a subpopulation (in the sense that the estimated probability of the case belonging to that subpopulation is the highest) *after* the model has been estimated.

### 3.3 Modified count models

The leading motivation for modified count models is to solve the so-called problem of excess zeros, the presence of more zeros in the data than predicted by count models such as the Poisson.

The *hurdle model* or *two-part model* relaxes the assumption that the zeros and the positives come from the same data generating process. The zeros are determined by the density  $f_1(\cdot)$ , so that  $\Pr[y = 0] = f_1(0)$ . The positive counts come from the truncated density  $f_2(y | y > 0) = f_2(y)/(1 - f_2(0))$ , which is multiplied by  $\Pr[y > 0] = 1 - f_1(0)$  to ensure that probabilities sum to unity. Thus

$$g(y) = \begin{cases} f_1(0) & \text{if } y = 0, \\ \frac{1 - f_1(0)}{1 - f_2(0)} f_2(y) & \text{if } y \geq 1. \end{cases} \quad (15.16)$$

This reduces to the standard model only if  $f_1(\cdot) = f_2(\cdot)$ . Thus in the modified model the two processes generating the zeros and the positives are not constrained to be the same. While the motivation for this model is to handle excess zeros, it is also capable of modeling too few zeros.

Maximum likelihood estimation of the hurdle model involves separate maximization of the two terms in the likelihood, one corresponding to the zeros and the other to the positives. This is straightforward.

A hurdle model has the interpretation that it reflects a two-stage decision making process. For example, a patient may initiate the first visit to a doctor, but the second and subsequent visits may be determined by a different mechanism (Pohlmeier and Ulrich, 1995).

Regression applications use hurdle versions of the Poisson or negative binomial, obtained by specifying  $f_1(\cdot)$  and  $f_2(\cdot)$  to be the Poisson or negative binomial densities given earlier. In application the covariates in the hurdle part which models the zero/one outcome need not be the same as those which appear in the truncated part, although in practice they are often the same. The hurdle model is widely used, and the hurdle negative binomial model is quite flexible. Drawbacks are that the model is not very parsimonious, typically the number of parameters is doubled, and parameter interpretation is not as easy as in the same model without hurdle.

The conditional mean in the hurdle model is the product of a probability of positives and the conditional mean of the zero-truncated density. Therefore, using a Poisson regression when the hurdle model is the correct specification implies a misspecification which will lead to inconsistent estimates.

### 3.4 Discrete choice models

Count data can be modeled by discrete choice model methods, possibly after some grouping of counts to limit the number of categories. For example, the categories may be 0, 1, 2, 3, and 4 or more if few observations exceed four. Unordered models such as multinomial logit are not parsimonious and more importantly are inappropriate. Instead, one should use a sequential discrete choice model that recognizes the ordering of the data, such as ordered logit or ordered probit.

## 4 PARTIALLY PARAMETRIC MODELS

By partially parametric models we mean that we focus on modeling the data via the conditional mean and variance, and even these may not be fully specified. In Section 4.1 we consider models based on specification of the conditional mean and variance. In Section 4.2 we consider and critique the use of least squares methods that do not explicitly model the heteroskedasticity inherent in count data. In Section 4.3 we consider models that are even more partially parametric, such as incomplete specification of the conditional mean.

### 4.1 Quasi-ML estimation

In the econometric literature *pseudo-ML* (PML) or *quasi-ML* (QML) estimation refers to estimating by ML, under the assumption that the specified density is possibly incorrect (Gouriéroux *et al.*, 1984a). PML and QML are often used interchangeably. The distribution of the estimator is obtained under weaker assumptions about the data generating process than those that led to the specified likelihood function. In the statistics literature QML often refers to nonlinear generalized least squares estimation. For the Poisson regression QML in the latter sense is equivalent to standard maximum likelihood.

From (15.5), the Poisson PML estimator,  $\hat{\beta}_p$ , has first-order conditions  $\sum_{i=1}^n (y_i - \exp(x'_i \beta)) x_i = 0$ . As already noted in Section 2.4, the summation on the left-hand side has an expectation of zero if  $E[y_i | x_i] = \exp(x'_i \beta)$ . Hence the Poisson PML is consistent under the weaker assumption of correct specification of the conditional mean – the data need not be Poisson distributed. Using standard results, the variance matrix is of the sandwich form, with

$$V_{\text{PML}}[\hat{\beta}_p] = \left( \sum_{i=1}^n \mu_i x_i x'_i \right)^{-1} \left( \sum_{i=1}^n \omega_i x_i x'_i \right) \left( \sum_{i=1}^n \mu_i x_i x'_i \right)^{-1} \quad (15.17)$$

and  $\omega_i = V[y_i | x_i]$  is the conditional variance of  $y_i$ .

Given an assumption for the functional form for  $\omega_i$ , and a consistent estimate  $\hat{\omega}_i$  of  $\omega_i$ , one can consistently estimate this covariance matrix. We could use the Poisson assumption,  $\omega_i = \mu_i$ , but as already noted the data are often overdispersed, with  $\omega_i > \mu_i$ . Common variance functions used are  $\omega_i = (1 + \alpha \mu_i) \mu_i$ , that of the NB2 model discussed in Section 3.1, and  $\omega_i = (1 + \alpha) \mu_i$ , that of the NB1 model. Note that in the latter case (15.17) simplifies to  $V_{\text{PML}}[\hat{\beta}_p] = (1 + \alpha) (\sum_{i=1}^n \mu_i x_i x'_i)^{-1}$ , so with overdispersion ( $\alpha > 0$ ) the usual ML variance matrix given in (15.6) is underestimating the true variance.

If  $\omega_i = E[(y_i - x'_i \beta)^2 | x_i]$  is instead unspecified, a consistent estimate of  $V_{\text{PML}}[\hat{\beta}_p]$  can be obtained by adapting the Eicker–White robust sandwich variance estimate formula to this case. The middle sum in (15.17) needs to be estimated. If  $\hat{\mu}_i \xrightarrow{p} \mu_i$  then  $n^{-1} \sum_{i=1}^n (y_i - \hat{\mu}_i)^2 x_i x'_i \xrightarrow{p} \lim n^{-1} \sum_{i=1}^n \omega_i x_i x'_i$ . Thus a consistent estimate of  $V_{\text{PML}}[\hat{\beta}_p]$  is given by (15.17) with  $\omega_i$  and  $\mu_i$  replaced by  $(y_i - \hat{\mu}_i)^2$  and  $\hat{\mu}_i$ .

When doubt exists about the form of the variance function, the use of the PML estimator is recommended. Computationally this is essentially the same as Poisson ML, with the qualification that the variance matrix must be recomputed. The calculation of robust variances is often an option in standard packages.

These results for Poisson PML estimation are qualitatively similar to those for PML estimation in the linear model under normality. They extend more generally to PML estimation based on densities in the linear exponential family. In all cases consistency requires only correct specification of the conditional mean (Nelder and Wedderburn, 1972; Gouriéroux *et al.*, 1984a). This has led to a vast statistical literature on *generalized linear models* (GLM), see McCullagh and Nelder (1989), which permits valid inference providing the conditional mean is correctly specified and nests many types of data as special cases – continuous (normal), count (Poisson), discrete (binomial) and positive (gamma). Many methods for complications, such as time series and panel data models, are presented in the more general GLM framework rather than specifically for count data.

Many econometricians find it more natural to use the *generalized method of moments* (GMM) framework rather than GLM. Then the starting point is the conditional moment  $E[y_i - \exp(x_i'\beta) | x_i] = 0$ . If data are independent over  $i$  and the conditional variance is a multiple of the mean it can be shown that the optimal choice of instrument is  $x_i$ , leading to the estimating equations (15.5); for more detail, see Cameron and Trivedi (1998, pp. 37–44). The GMM framework has been fruitful for panel data on counts, see Section 5.3, and for *endogenous* regressors. Fully specified simultaneous equations models for counts have not yet been developed, so instrumental variables methods are used. Given instruments  $z_i$ ,  $\dim(z) \geq \dim(x)$ , satisfying  $E[y_i - \exp(x_i'\beta) | z_i] = 0$ , a consistent estimator of  $\beta$  minimizes

$$Q(\beta) = \left( \sum_{i=1}^n (y_i - \exp(x_i'\beta)) z_i \right)' W \left( \sum_{i=1}^n (y_i - \exp(x_i'\beta)) z_i \right),$$

where  $W$  is a symmetric weighting matrix.

## 4.2 Least squares estimation

When attention is focused on modeling just the conditional mean, least squares methods are inferior to the approach of the previous subsection.

Linear least squares regression of  $y$  on  $x$  leads to consistent parameter estimates if the conditional mean is linear in  $x$ . But for count data the specification  $E[y|x] = x'\beta$  is inadequate as it permits negative values of  $E[y|x]$ . For similar reasons the linear probability model is inadequate for binary data.

Transformations of  $y$  may be considered. In particular the logarithmic transformation regresses  $\ln y$  on  $x$ . This transformation is problematic if the data contain zeros, as is often the case. One standard solution is to add a constant term, such as 0.5, and to model  $\ln(y + .5)$  by OLS. This method often produces unsatisfactory

results, and complicates the interpretation of coefficients. It is also unnecessary as software to estimate basic count models is widely available.

### 4.3 Semiparametric models

By *semiparametric models* we mean partially parametric models that have an infinite-dimensional component.

One example is optimal estimation of the regression parameters  $\beta$ , when  $\mu_i = \exp(x_i'\beta)$  is assumed but  $V[y_i | x_i] = \omega_i$  is left unspecified. The infinite-dimensional component arises because as  $n \rightarrow \infty$  there are infinitely many variance parameters  $\omega_i$ . An optimal estimator of  $\beta$ , called an *adaptive estimator*, is one that is as efficient as that when  $\omega_i$  is known. Delgado and Kniesner (1997) extend results for the linear regression model to count data with exponential conditional mean function, using kernel regression methods to estimate weights to be used in a second-stage nonlinear least squares regression. In their application the estimator shows little gain over specifying  $\omega_i = \mu_i(1 + \alpha\mu_i)$ , overdispersion of the NB2 form.

A second class of semiparametric models incompletely specifies the conditional mean. Leading examples are *single-index models* and *partially linear models*. Single-index models specify  $\mu_i = g(x_i'\beta)$  where the functional form  $g(\cdot)$  is left unspecified. Partially linear models specify  $\mu_i = \exp(x_i'\beta + g(z_i))$  where the functional form  $g(\cdot)$  is left unspecified. In both cases root- $n$  consistent asymptotically normal estimators of  $\beta$  can be obtained, without knowledge of  $g(\cdot)$ .

## 5 TIME SERIES, MULTIVARIATE AND PANEL DATA

In this section we very briefly present extension from cross section to other types of count data (see Cameron and Trivedi, 1998, for further detail). For time series and multivariate count data many models have been proposed but preferred methods have not yet been established. For panel data there is more agreement in the econometrics literature on which methods to use, though a wider range of models is considered in the statistics literature.

### 5.1 Time series data

If a time series of count data is generated by a Poisson point process then event occurrences in successive time intervals are independent. Independence is a reasonable assumption when the underlying stochastic process for events, conditional on covariates, has no memory. Then there is no need for special time series models. For example, the number of deaths (or births) in a region may be uncorrelated over time. At the same time the population, which cumulates births and deaths, will be very highly correlated over time.

The first step for time series count data is therefore to test for serial correlation. A simple test first estimates a count regression such as Poisson, obtains the residual, usually  $(y_t - \exp(x_t'\hat{\beta}))$  where  $x_t$  may include time trends, and tests for zero correlation between current and lagged residuals, allowing for the complication that the residuals will certainly be heteroskedastic.

Upon establishing the data are indeed serially correlated, there are several models to choose from. An aesthetically appealing model is the INAR(1) model (*integer autoregressive model* of order one and its generalization to the negative binomial and to higher orders of serial correlation. This model specifies  $y_t = \rho_t \circ y_{t-1} + \varepsilon_t$ , where  $\rho_t$  is a correlation parameter with  $0 < \rho_t < 1$ , for example  $\rho_t = 1/[1 + \exp(-z_t'\gamma)]$ . The symbol  $\circ$  denotes the *binomial thinning* operator, whereby  $\rho_t \circ y_{t-1}$  is the realized value of a binomial random variable with probability of success  $\rho_t$  in each of  $y_{t-1}$  trials. One may think of each event as having a replication or survival probability of  $\rho_t$  in the following period. As in a linear first-order Markov model, this probability decays geometrically. A Poisson INAR(1) model, with a Poisson marginal distribution for  $y_t$  arises when  $\varepsilon_t$  is Poisson distributed with mean, say,  $\exp(x_t'\beta)$ . A negative binomial INAR(1) model arises if  $\varepsilon_t$  is negative binomial distributed.

An *autoregressive model*, or *Markov model*, is a simple adjustment to the earlier cross section count models that directly enters lagged values of  $y$  into the formula for the conditional mean of current  $y$ . For example, we might suppose  $y_t$  conditional on current and past  $x_t$  and past  $y_t$  is Poisson distributed with mean  $\exp(x_t'\beta + \rho \ln y_{t-1}^*)$ , where  $y_{t-1}^*$  is an adjustment to ensure a nonzero lagged value, such as  $y_{t-1}^* = (y_{t-1} + 0.5)$  or  $y_{t-1}^* = \max(0.5, y_{t-1})$ .

*Seriously correlated error models* induce time series correlation by introducing unobserved heterogeneity, see Section 3.1, and allowing this to be serially correlated. For example,  $y_t$  is Poisson distributed with mean  $\exp(x_t'\beta)v_t$  where  $v_t$  is a serially correlated random variable (Zeger, 1988).

*State space models* or *time-varying parameters models* allow the conditional mean to be a random variable drawn from a distribution whose parameters evolve over time. For example,  $y_t$  is Poisson distributed with mean  $\mu_t$  where  $\mu_t$  is a draw from a gamma distribution (Harvey and Fernandes, 1989).

*Hidden Markov models* specify different parametric models in different regimes, and induce serial correlation by specifying the stochastic process determining which regime currently applies to be an unobserved Markov process (MacDonald and Zucchini, 1997).

## 5.2 Multivariate data

In some data sets more than one count is observed. For example, data on the utilization of several different types of health service, such as doctor visits and hospital days, may be available. Joint modeling will improve efficiency and provide richer models of the data if counts are correlated.

Most parametric studies have used the *bivariate Poisson*. This model, however, is too restrictive as it implies variance–mean equality for the counts and restricts the correlation to be positive. Development of better parametric models is a current area of research.

## 5.3 Panel data

One of the major and earliest applications of count data methods in econometrics is to panel data on the number of patents awarded to firms over time (Hausman

*et al.*, 1984). The starting point is the Poisson regression model with exponential conditional mean and multiplicative individual-specific term

$$y_{it} \sim P[\alpha_i \exp(x'_{it}\beta)], \quad i = 1, \dots, n, \quad t = 1, \dots, T, \quad (15.18)$$

where we consider a short panel with  $T$  small and  $n \rightarrow \infty$ . As in the linear case, both fixed effects and random effects models are possible.

The *fixed effects model* lets  $\alpha_i$  be an unknown parameter. This parameter can be eliminated by quasi-differencing and modeling the transformed random variable  $y_{it} - (\lambda_{it}/\bar{\lambda}_i)\bar{y}_{it}$ , where  $\bar{\lambda}_i$  and  $\bar{y}_{it}$  denote the individual-specific means of  $\lambda_{it}$  and  $y_{it}$ . By construction this has zero mean, conditional on  $x_{i1}, \dots, x_{iT}$ . A moments-based estimator of  $\beta$  then solves the sample moment condition  $\sum_{i=1}^n \sum_{t=1}^T x_{it}(y_{it} - (\lambda_{it}/\bar{\lambda}_i)\bar{y}_{it}) = 0$ .

An alternative to the quasi-differencing approach is the conditional likelihood approach that was followed by Hausman *et al.* (1984). In this approach the fixed effects are eliminated by conditioning the distribution of counts on  $\sum_{i=1}^T y_{it}$ .

The *random effects model* lets  $\alpha_i$  be a random variable with specified distribution that depends on parameters, say  $\delta$ . The random effects are integrated out, in a similar way to the unobserved heterogeneity in Section 3.1, and the parameters  $\beta$  and  $\delta$  are estimated by maximum likelihood. In some cases, notably when  $\alpha_i$  is gamma distributed, a closed-form solution is obtained upon integrating out  $\alpha_i$ . In other cases, such as normally distributed random effects, a closed-form solution is not obtained, but ML estimation based on numerical integration is feasible.

*Dynamic panel data models* permit the regressors  $x$  to include lagged values of  $y$ . Several studies use the fixed effects variant of (15.18), where  $x_{it}$  now includes, for example,  $y_{it-1}$ . This is an autoregressive count model, see Section 5.1, adapted to panel data. The quasi-differencing procedure for the nondynamic fixed effects case can be adapted to the dynamic case.

## 6 PRACTICAL CONSIDERATIONS

Those with experience of nonlinear least squares will find it easy to use packaged software for Poisson regression, which is a widely available option in popular econometrics packages like LIMDEP, STATA, and TSP. One should ensure, however, that reported standard errors are based on (15.17) rather than (15.6). Many econometrics packages also include negative binomial regression, also widely used for cross section count regression, and the basic panel data models. Statistics packages such as SAS and SPSS include count regression in a generalized linear models module. Standard packages also produce some goodness-of-fit statistics, such as the  $G^2$ -statistic and pseudo- $R^2$  measures, for the Poisson (see Cameron and Windmeijer, 1996).

More recently developed models, such as finite mixture models, most time series models and dynamic panel data models, require developing one's own programs. A promising route is to use matrix programming languages such as GAUSS, MATLAB, SAS/IML, or SPLUS in conjunction with software for implementing estimation based on user-defined objective functions. For simple models

packages such as LIMDEP, STATA, and TSP make it possible to implement maximum likelihood estimation and (highly desirable) robust variance estimation for user-defined functions.

In addition to reporting parameter estimates it is useful to have an indication of the magnitude of the estimated effects, as discussed in Section 2.2. And as noted in Section 2.4, care should be taken to ensure that reported standard errors and  $t$ -statistics for the Poisson regression model are based on variance estimates robust to overdispersion.

In addition to estimation it is strongly recommended that specification tests are used to assess the adequacy of the estimated model. For Poisson cross section regression overdispersion tests are easy to implement. For time series regression tests of serial correlation should be used. For any parametric model one can compare the actual and fitted frequency distribution of counts. Formal statistical specification and goodness-of-fit tests based on actual and fitted frequencies are available.

In most practical situations one is likely to face the problem of model selection. For likelihood-based models that are nonnested one can use selection criteria, such as the Akaike and Schwarz criteria, which are based on the fitted loglikelihood but with degrees of freedom penalty for models with many parameters.

## 7 FURTHER READING

All the topics dealt with in this chapter are treated at greater length and depth in Cameron and Trivedi (1998) which also provides a comprehensive bibliography. Winkelmann (1997) also provides a fairly complete treatment of the econometric literature on counts. The statistics literature generally analyzes counts in the context of generalized linear models (GLM). The standard reference is McCullagh and Nelder (1989). The econometrics literature generally fails to appreciate the contributions of the GLM literature on generalized linear models. Fahrmeir and Tutz (1994) provide a recent and more econometric exposition of GLMs.

The material in Section 2 is very standard and appears in many places. A similar observation applies to the negative binomial model in Section 3.1. Cameron and Trivedi (1986) provide an early presentation and application. For the finite mixture approach of Section 3.2 see Deb and Trivedi (1997). Applications of the hurdle model in Section 3.3 include Mullahy (1986), who first proposed the model, Pohlmeier and Ulrich (1995), and Gurmu and Trivedi (1996). The quasi-MLE of Section 4.1 is presented in detail by Gouriéroux *et al.* (1984a, 1984b) and by Cameron and Trivedi (1986).

Regression models for the types of data discussed in Section 5 are in their infancy. The notable exception is that (static) panel data count models are well established, with the standard reference being Hausman *et al.* (1984). See also Brännäs and Johansson (1996). For reviews of the various time series models see MacDonald and Zucchini (1997, ch. 2) and Cameron and Trivedi (1998, ch. 7). Developing adequate regression models for multivariate count data is currently an active area. For dynamic count panel data models there are several recent references, including Blundell *et al.* (1995).

For further discussion of diagnostic testing, only briefly mentioned in Section 6, see Cameron and Trivedi (1998, ch. 5).

## References

- Blundell, R., R. Griffith, and J. Van Reenen (1995). Dynamic count data models of technological innovation. *Economic Journal* 105, 333–44.
- Brännäs, K., and P. Johansson (1996). Panel data regression for counts. *Statistical Papers* 37, 191–213.
- Cameron, A.C., and P.K. Trivedi (1986). Econometric models based on count data: Comparisons and applications of some estimators. *Journal of Applied Econometrics* 1, 29–53.
- Cameron, A.C., and P.K. Trivedi (1998). *Regression Analysis of Count Data*. New York: Cambridge University Press.
- Cameron, A.C., P.K. Trivedi, F. Milne, and J. Piggott (1988). A microeconometric model of the demand for health care and health insurance in Australia. *Review of Economic Studies* 55, 85–106.
- Cameron, A.C., and F.A.G. Windmeijer (1996). R-squared measures for count data regression models with applications to health care utilization. *Journal of Business and Economic Statistics* 14, 209–20.
- Davutyan, N. (1989). Bank failures as Poisson variates. *Economic Letters* 29, 333–8.
- Deb, P., and P.K. Trivedi (1997). Demand for medical care by the elderly: A finite mixture approach. *Journal of Applied Econometrics* 12, 313–26.
- Delgado, M.A., and T.J. Kniesner (1997). Count data models with variance of unknown form: An application to a hedonic model of worker absenteeism. *Review of Economics and Statistics* 79, 41–9.
- Fahrmeir, L., and G.T. Tutz (1994). *Multivariate Statistical Modelling Based on Generalized Linear Models*. New York: Springer-Verlag.
- Gouriéroux, C., and A. Monfort (1997). *Simulation Based Econometric Methods*. Oxford: Oxford University Press.
- Gouriéroux, C., A. Monfort, and A. Trognon (1984a). Pseudo maximum likelihood methods: Theory. *Econometrica* 52, 681–700.
- Gouriéroux, C., A. Monfort, and A. Trognon (1984b). Pseudo maximum likelihood methods: Applications to Poisson models. *Econometrica* 52, 701–20.
- Gurmu, S., and P.K. Trivedi (1996). Excess zeros in count models for recreational trips. *Journal of Business and Economic Statistics* 14, 469–77.
- Harvey, A.C., and C. Fernandes (1989). Time series models for count or qualitative observations (with discussion). *Journal of Business and Economic Statistics* 7, 407–17.
- Hausman, J.A., B.H. Hall, and Z. Griliches (1984). Econometric models for count data with an application to the patents-R and D relationship. *Econometrica* 52, 909–38.
- Johnson, N.L., S. Kotz, and A.W. Kemp (1992). *Univariate Distributions*, 2nd edn. New York: John Wiley.
- MacDonald, I.L., and W. Zucchini (1997). *Hidden Markov and other Models for Discrete-valued Time Series*. London: Chapman and Hall.
- McCullagh, P., and J.A. Nelder (1989). *Generalized Linear Models*, 2nd edn. London: Chapman and Hall.
- Mullahy, J. (1986). Specification and testing of some modified count data models. *Journal of Econometrics* 33, 341–65.
- Nelder, J.A., and R.W.M. Wedderburn (1972). Generalized linear models. *Journal of the Royal Statistical Society A* 135, 370–84.

- Pohlmeier, W., and V. Ulrich (1995). An econometric model of the two-part decision-making process in the demand for health care. *Journal of Human Resources* 30, 339–61.
- Rose, N. (1990). Profitability and product quality: Economic determinants of airline safety performance. *Journal of Political Economy* 98, 944–64.
- Terza, J. (1998). Estimating count data models with endogenous switching: Sample selection and endogenous switching effects. *Journal of Econometrics* 84, 129–39.
- Winkelmann, R. (1995). Duration dependence and dispersion in count-data models. *Journal of Business and Economic Statistics* 13, 467–74.
- Winkelmann, R. (1997). *Econometric Analysis of Count Data*. Berlin: Springer-Verlag.
- Zeger, S.L. (1988). A regression model for time series of counts. *Biometrika* 75, 621–9.

CHAPTER SIXTEEN

# Panel Data Models

*Cheng Hsiao\**

## 1 INTRODUCTION

A panel (or longitudinal or temporal cross-sectional) data set is one which follows a number of individuals over time. By providing sequential observations for a number of individuals, panel data allow us to distinguish inter-individual differences from intra-individual differences, thus allow us to construct and test more complicated behavioral models than a single time series or cross section data set would allow. Moreover, panel data offer many more degrees of freedom, provide the possibility to control for omitted variable bias and reduce the problem of multicollinearity, hence improving the accuracy of parameter estimates and prediction (e.g. Baltagi, 1995; Chamberlain, 1984; Hsiao, 1985, 1986, 1995; Mátyás and Sevestre, 1996).

However, the emphasis of panel data is on individual outcomes. Factors affecting individual outcomes are numerous, yet a model is a simplification of the real world. It is neither feasible nor desirable to include all factors that affect the outcomes in the specification. If the specification of the relationships among variables appears proper, yet the outcomes conditional on the included explanatory variables cannot be viewed as random draws from a probability distribution, then standard statistical procedures will lead to misleading inferences. Therefore, the focus of panel data research is on controlling the impact of unobserved heterogeneity among cross-sectional units over time in order to draw inference about the population characteristics.

If heterogeneity among cross-sectional units over time cannot be captured by the explanatory variables, one can either let this heterogeneity be represented by the error term or let the coefficients vary across individuals and/or over time. For instance, for a panel of  $N$  individuals over  $T$  time periods, a linear model specification can take the form

$$y_{it} = \beta'_{it} x_{it} + u_{it}, \quad i = 1, \dots, N \\ t = 1, \dots, T, \quad (16.1)$$

where both the coefficients of  $x$  variables and the error of the equation vary across individuals and over time. However, model (16.1) only has descriptive value. One can neither estimate  $\beta_{it}$  nor use it to draw inference about the population if each individual is different and varies their behavioral patterns over time. In this chapter we will give a selected survey of panel data models.

For ease of exposition we shall assume that the unobserved heterogeneities vary across individuals but stay constant over time. We discuss linear models in Section 2, dynamic models in Section 3, nonlinear models in Section 4, sample attrition and sample selectivity in Section 5. Conclusions are in Section 6.

## 2 LINEAR MODELS

Suppose there are observations of  $1 + k_1 + k_2$  variables  $(y_{it}, \underline{x}'_{it}, \underline{z}'_{it})$  of  $N$  cross-sectional units over  $T$  time periods, where  $i = 1, \dots, N$ , and  $t = 1, \dots, T$ . Let  $y = (y'_1, \dots, y'_N)', X = \text{diag}(X_i)$ ,  $Z = \text{diag}(Z_i)$ , where  $y'_i = (y_{i1}, \dots, y_{iT})'$ ,  $X_i$  and  $Z_i$  are  $T \times k_1$  and  $T \times k_2$  matrices of  $T$  observed values of the explanatory variables  $\underline{x}'_{it}$  and  $\underline{z}'_{it}$  for  $i = 1, \dots, N$ . If all the individuals in the panel data are different, we have the unconstrained linear model,

$$\underline{y} = \underline{X}\underline{\beta} + \underline{Z}\underline{\gamma} + \underline{u}, \quad (16.2)$$

where  $\underline{\beta} = (\underline{\beta}'_1, \dots, \underline{\beta}'_N)'$  and  $\underline{\gamma} = (\underline{\gamma}'_1, \dots, \underline{\gamma}'_N)'$  are  $Nk_1 \times 1$  and  $Nk_2 \times 1$  vector of constants,  $\underline{u} = (\underline{u}'_1, \dots, \underline{u}'_N)'$  is an  $NT \times 1$  vector of the error term,  $\underline{\beta}_i$ ,  $\underline{\gamma}_i$  and  $\underline{u}_i$  denote the coefficients of  $X_i$ ,  $Z_i$  and the error of the  $i$ th individual for  $i = 1, \dots, N$ . We assume that  $\underline{u}$  is independent of  $x$  and  $z$  and is multivariately normally distributed with mean  $0$  and covariance matrix  $C_1$ ,<sup>1</sup>

$$\underline{u} \sim N(0, C_1). \quad (16.3)$$

There is no particular advantage of pooling the panel data to estimate (16.2) except for the possibility of exploiting the common shocks in the error term if  $C_1$  is not block diagonal by applying the Zellner's (1962) seemingly unrelated regression estimator. To take advantage of the panel data there must be constraints on (16.2). Two types of constraints are commonly imposed – stochastic and exact. We shall assume that the coefficients of  $\underline{x}_{it}$  are subject to stochastic constraints and the coefficients of  $\underline{z}_{it}$  are subject to exact constraints.

To postulate stochastic constraints, we let

$$\underline{\beta} = \begin{pmatrix} \underline{\beta}_1 \\ \vdots \\ \underline{\beta}_N \end{pmatrix} = A_1 \tilde{\underline{\beta}} + \underline{\xi}, \quad (16.4)$$

where  $A_1$  is an  $Nk_1 \times m$  matrix with known elements,  $\tilde{\underline{\beta}}$  is an  $m \times 1$  vector of constants, and

$$\varepsilon \sim N(\mathbf{0}, C_2). \quad (16.5)$$

The variance covariance matrix  $C_2$  is assumed to be nonsingular. Furthermore, we assume that<sup>2</sup>

$$\text{cov}(\varepsilon, \mathbf{u}) = \mathbf{0}, \text{cov}(\varepsilon, X) = \mathbf{0} \quad \text{and} \quad \text{cov}(\varepsilon, Z) = \mathbf{0}. \quad (16.6)$$

To postulate exact constraints, we let

$$\gamma = \begin{pmatrix} \gamma_1 \\ \vdots \\ \gamma_N \end{pmatrix} = A_2 \bar{\gamma}, \quad (16.7)$$

where  $A_2$  is an  $Nk_2 \times n$  matrix with known elements, and  $\bar{\gamma}$  is an  $n \times 1$  vector of constants. Because  $A_2$  is known, (16.2) is formally identical to

$$\underline{y} = X\hat{\beta} + \tilde{Z}\bar{\gamma} + \mathbf{u}, \quad (16.8)$$

where  $\tilde{Z} = ZA_2$ .

Formulation (16.4)–(16.8) encompasses various linear panel data models as special cases. These include:

1. *A common model for all cross-sectional units* by letting  $X = \mathbf{0}$ ,  $A_2 = \mathbf{e}_N \otimes I_{k_2}$ , where  $\mathbf{e}_N$  is an  $N \times 1$  vector of ones,  $I_p$  denotes a  $p \times p$  identity matrix and  $\otimes$  denotes Kronecker product.
2. *Different models for different cross-sectional units* by letting  $X = 0$  and  $A_2$  be an  $Nk_2 \times Nk_2$  identity matrix.
3. *Variable intercept models* (e.g. Kuh, 1963; Mundlak, 1978) by letting  $X = 0$ ,  $Z_i = (\mathbf{e}_T, \tilde{Z}_i)$ ,  $A_2 = (I_N \otimes i_{k_2} : \mathbf{e}_N \otimes I_{k_2-1}^*)$  where  $i_{k_2}$  is a  $k_2 \times 1$  vector of  $(1, 0, \dots, 0)'$  and  $I_{k_2-1}^*$  is a  $k_2 \times (k_2 - 1)$  matrix of  $\begin{pmatrix} \mathbf{0}' \\ I_{k_2-1} \end{pmatrix}$ .
4. *Error components model* (e.g. Balestra and Nerlove, 1966; Wallace and Hussain, 1969) by letting  $X_i = \mathbf{e}_T$ ,  $A_1 = \mathbf{e}_N$ ,  $C_2 = \sigma_\beta^2 I_N$ .
5. *Random coefficients models* (e.g. Hsiao, 1974, 1975; Swamy, 1970) by letting  $Z = \mathbf{0}$ ,  $A_1 = \mathbf{e}_N \otimes I_{k_1}$ ,  $C_2 = I_N \otimes \Delta$ , where  $\Delta = E(\hat{\beta}_i - \bar{\beta})(\hat{\beta}_i - \bar{\beta})'$ .
6. *Mixed fixed and random coefficients models* (e.g. Hsiao *et al.*, 1989; Hsiao and Mountain, 1995; Hsiao and Tahmisioglu, 1997) as postulated by (16.4)–(16.8).

Substituting (16.4) into (16.8), we have

$$\underline{y} = XA_1\hat{\beta} + \tilde{Z}\bar{\gamma} + \mathbf{u}^*, \quad (16.9)$$

with  $\mathbf{u}^* = \mathbf{u} + X\varepsilon$ . The generalized least squares (GLS) estimator of (16.9) is

$$\begin{pmatrix} \hat{\beta}_{\text{GLS}} \\ \hat{\gamma}_{\text{GLS}} \end{pmatrix} = \begin{pmatrix} A_1'X'(C_1 + XC_2X')^{-1}XA_1 & A_1'X(C_1 + XC_2X')^{-1}\tilde{Z} \\ \tilde{Z}'(C_1 + X'C_2X')^{-1}XA_1 & \tilde{Z}'(C_1 + XC_2X')^{-1}\tilde{Z} \end{pmatrix}^{-1} \begin{pmatrix} A_1'X'(C_1 + XC_2X')^{-1}\underline{y} \\ \tilde{Z}'(C_1 + XC_2X')^{-1}\underline{y} \end{pmatrix}. \quad (16.10)$$

The GLS estimator is also the Bayes estimator conditional on  $C_1$  and  $C_2$  with a diffuse prior for  $\hat{\beta}$  and  $\hat{\gamma}$  (Lindley and Smith, 1972; Hsiao, 1990). Moreover, if predicting individual  $\beta_i$  is of interest, the Bayes procedure predicts  $\hat{\beta}_i$  as a weighted average between the GLS estimator of  $\hat{\beta}$  and the individual least squares estimator of  $\hat{\beta}_i$  (Hsiao *et al.*, 1993)

$$\hat{\beta}^* = \{X'DX + C_2^{-1}\}^{-1} \{X'DX\hat{\beta} + C_2^{-1}A_1\hat{\beta}\}, \quad (16.11)$$

where  $D = [C_1^{-1} - C_1^{-1}\tilde{Z}(\tilde{Z}'C_1^{-1}\tilde{Z})^{-1}\tilde{Z}'C_1]$  and

$$\hat{\beta} = \{X'DX\}^{-1}\{X'D\underline{y}\} \quad (16.12)$$

In other words, if cross-sectional units are similar as postulated in (16.4), there is an advantage of pooling since if there is not enough information about a cross-sectional unit, one can obtain a better prediction of that individual's outcome by learning from other cross section units that behave similarly.

The above formulation presupposes that which variables are subject to stochastic constraints and which variables are subject to deterministic constraints is known. In practice, there is very little knowledge about it. In certain special cases, formal statistical testing procedures have been proposed (e.g. Breusch and Pagan, 1980; Hausman, 1978). However, a typical test postulates a simple null versus a composite alternative. The distribution of a test statistic is derived under an assumed true null hypothesis. A rejection of a null hypothesis does not automatically imply the acceptance of a particular alternative. However, most of the tests of fixed versus random effects specification are indirect in the sense that they exploit a particular implication of the random effects formulation. For instance, the rejection of a null of homoskedasticity does not automatically imply the acceptance of a particular alternative of random effects. In fact, it would be more useful to view the above different formulations as different models and treat them as an issue of model selection. Various model selection criteria (e.g. Akaike, 1973; Schwarz, 1978) can be used. Alternatively, predictive density ratio (e.g. Hsiao and Tahmisioglu, 1997; Min and Zellner, 1993) can be used to select the appropriate formulation. The predictive density ratio approach divides the time series observations into two periods, 1 to  $T_1$ , denoted by  $\underline{y}_1^*$  and  $T_1 + 1$  to  $T$ , denoted by  $\underline{y}_2^*$ . The first period observations,  $\underline{y}_1^*$ , are used to derive the posterior probability distribution of  $\theta_0$  and  $\theta_1$  given hypothesis  $H_0$  and  $H_1$ ,  $f(\theta_0 | \underline{y}_1^*)$  and  $f(\theta_1 | \underline{y}_1^*)$ . The second period observations are used to compare how  $H_0$  or  $H_1$  predicts the outcome. The predictive density ratio is then computed as

$$\frac{\int f(\underline{y}_2^* | \underline{\theta}_0, \underline{y}_1^*) f(\underline{\theta}_0 | \underline{y}_1^*) d\underline{\theta}_0}{\int f(\underline{y}_2^* | \underline{\theta}_1, \underline{y}_1^*) f(\underline{\theta}_1 | \underline{y}_1^*) d\underline{\theta}_1}, \quad (16.13)$$

where  $f(\underline{y}_2^* | \underline{\theta}_1, \underline{y}_1^*)$  and  $f(\underline{y}_2^* | \underline{\theta}_2, \underline{y}_1^*)$  are the conditional densities of  $\underline{y}_2^*$  given  $\underline{y}_1^*$  and  $\underline{\theta}_0$  or  $\underline{\theta}_1$ . If (16.13) is greater than 1, then  $H_0$  is favored. If (16.13) is less than 1, then  $H_1$  is favored. When  $T$  is small, a recursive updating scheme of the posterior probability distribution of  $\underline{\theta}_0$  and  $\underline{\theta}_1$  each with additional observations can be used to balance the sample dependence of predictive outcome and informativeness of the conditional density of  $\underline{\theta}_i$  given observables. The Monte Carlo studies appear to indicate that the predictive density ratio performs well in selecting the appropriate formulation (e.g. Hsiao *et al.*, 1995; Hsiao and Tahmisioglu, 1997).

### 3 DYNAMIC MODELS

When  $x_{it}$  and/or  $z_{it}$  contain lagged dependent variables, because a typical panel contains a large number of cross-sectional units followed over a short period of time, it turns out that how the initial value of the dependent variable is modeled plays a crucial role with regard to the consistency and efficiency of an estimator (e.g. Anderson and Hsiao, 1981, 1982; Bhargava and Sargan, 1983; Blundell and Bond, 1998). Moreover, if there exists unobserved heterogeneity and the individual specific effects is more appropriately treated as random, the random effects and the lagged dependent variables are correlated, thus (16.6) is violated. If the individual effects are treated as fixed, then the number of individual specific parameters increases with the number of cross-sectional units,  $N$ . Contrary to the static case of Section 2, the estimation of the individual specific effects in a dynamic model is not independent of the estimation of the structural parameters that are common across  $N$  and  $T$ . Since for each individual there are only a finite number of observations, there is no way the individual specific parameters can be accurately estimated. The errors in the estimation of individual specific parameters will be transmitted into the estimation of the structural parameters if the two estimators are not independent. This is the classical incidental parameters problem (Neyman and Scott, 1948).

Consider a simple dynamic model with individual specific effects appearing in the intercepts only,

$$\underline{y}_i = \underline{\epsilon}_T \alpha_i + \underline{y}_{i,-1} \gamma_1 + Z_i \underline{y}_2 + \underline{u}_i \quad (16.14)$$

where  $z_{it}$  is a  $k_2 - 1$  dimensional exogenous variables,  $\underline{y}_{i,-1} = (y_{i,0}, \dots, y_{i,T-1})$  and for ease of exposition, we assume that  $y_{i,0}$  are observable. Let  $H$  be an  $m \times T$  transformation matrix such that  $H\underline{\epsilon}_T = 0$ . Multiplying  $H$  to (16.14), we eliminate the individual specific effects  $\alpha_i$  from the specification,

$$H\underline{y}_i = H\underline{y}_{i,-1} \gamma_1 + HZ_i \underline{y}_2 + Hu_i. \quad (16.15)$$

Since (16.15) does not depend on  $\alpha_i$ , if we can find instruments that are correlated with the explanatory variables but uncorrelated with  $Hu_i$ , we can apply the

instrumental variable method to estimate  $\gamma_1$  and  $\gamma_2$  (e.g. Anderson and Hsiao, 1981, 1982). Let  $W_i$  be the  $q \times m$  matrix of instrumental variables that satisfies

$$E(W_i H \underline{u}_i) = 0. \quad (16.16)$$

The generalized method of moments (GMM) estimator takes the form

$$\begin{aligned} \begin{pmatrix} \hat{\gamma}_1 \\ \hat{\gamma}_2 \end{pmatrix} &= \left\{ \left[ \sum_{i=1}^N \begin{pmatrix} y'_{i,-1} \\ Z'_i \end{pmatrix} H' W'_i \right] \left( \sum_{i=1}^N W_i \Phi W'_i \right)^{-1} \left[ \sum_{i=1}^N W_i H(y_{i,-1}, Z_i) \right] \right\}^{-1} \\ &\quad \left\{ \left[ \sum_{i=1}^N \begin{pmatrix} y'_{i,-1} \\ Z'_i \end{pmatrix} H' W'_i \right] \left( \sum_{i=1}^N W_i \Phi W'_i \right)^{-1} \left[ \sum_{i=1}^N W_i H \underline{y}_i \right] \right\}, \end{aligned} \quad (16.17)$$

where  $\Phi = E(H \underline{u}_i \underline{u}'_i H')$ . In the case when the transformation matrix  $H$  takes the form of first differencing (16.14),  $H$  is a  $(T - 1) \times T$  matrix with all the elements in the  $t$ th row equal to zero except for the  $t$ th and  $(t + 1)$ th element that takes the value of  $-1$  and  $1$ , respectively. If  $u_{it}$  is iid, then  $W_i$  takes the form (e.g. Ahn and Schmidt, 1995; Amemiya and MacCurdy, 1986; Arellano and Bover, 1995; Blundell and Bond, 1998)

$$W'_i = \begin{pmatrix} z'_i & 0' & 0' & 0 & 0 \\ 0 & y_{i0}, z'_i & 0' & 0 & 0 \\ 0 & 0 & y_{i0}, y_{i1}, z'_i & 0 & \cdot \\ \cdot & \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot & y_{i0}, y_{i1}, \dots, y_{iT-2}, z'_i \end{pmatrix}, \quad (16.18)$$

where  $z'_i = (z'_{i1}, \dots, z'_{iT})$ .

GMM estimator (16.17) makes use of  $\frac{T(T-1)}{2} + T(k_2 - 1)$  orthogonality conditions. Where  $k_2 - 1$  denotes the dimension of  $\underline{z}_{it}$ . In most applications, this is a large number. For instance, even in the case of  $k_2 - 1 = 0$ , there are still 45 orthogonality conditions for a model with only one lagged dependent variable when  $T = 10$  which makes the implementation of the GMM estimator (16.17) nontrivial. Moreover, in finite sample, it suffers severe bias as demonstrated in a Monte Carlo study conducted by Ziliak (1997) because of the correlation between the estimated weight matrix and the sample moments and/or the weak instruments phenomena (Wansbeek and Knaap, 1999). The bias of the GMM estimator leads to poor coverage rates for confidence intervals based on asymptotic critical values. Hsiao, Pesaran and Tahmisioglu (1998b) propose a transformed maximum likelihood estimator that maximizes the likelihood function of (16.15) and the initial value function

$$\nabla y_{i1} = \delta' \nabla z_i + v_v \quad (16.19)$$

where  $\nabla z_i$  denotes the first difference of  $z_i$ . The transformed MLE is consistent and asymptotically normally distributed provided that the generating process of  $z_{it}$  is difference stationary, i.e. the data generating process of  $z_{it}$  is of the form

$$\tilde{z}_{it} = \mu_i + gt + \sum_{j=0}^{\infty} B_j \tau_{i,t-j}, \quad (16.20)$$

where  $\tau_{it}$  is iid with constant mean and covariances, and  $\Sigma \|B_j\| < \infty$ . The transformed MLE is easier to implement than the GMM and is asymptotically more efficient because the GMM requires each instrument to be orthogonal to the transformed error while the transformed MLE only requires a linear combination of the instruments to be orthogonal to the transformed errors. The Monte Carlo studies conducted by Hsiao, Pesaran and Tahmisioglu (1998) show that the transformed MLE performs very well even when  $T$  and  $N$  both are small.

Where individual heterogeneity cannot be completely captured by individual time-invariant effects, a varying parameter model is often used. However, while it may be reasonable to assume (16.5), (16.6) cannot hold if  $x_{it}$  contains lagged dependent variables. For instance, consider a simple dynamic model

$$\begin{aligned} y_{it} &= \beta_i y_{i,t-1} + \gamma z_{it} + u_{it} \\ &= \bar{\beta} y_{i,t-1} + \bar{\gamma} z_{it} + v_{it}, \end{aligned} \quad (16.21)$$

where  $v_{it} = \varepsilon_i y_{i,t-1} + u_{it}$ . By continuous substitution, it can be shown that

$$y_{i,t-1} = \bar{\gamma} \sum_{j=0}^{\infty} (\bar{\beta} + \varepsilon_i)^j z_{i,t-j-1} + \sum_{j=0}^{\infty} (\bar{\beta} + \varepsilon_i)^j u_{i,t-j-1}. \quad (16.22)$$

It follows that  $E(v_{it} | y_{i,t-1}) \neq 0$ . Therefore, the least squares estimator is inconsistent. Neither is the instrumental variable estimator feasible because the instruments that are uncorrelated with  $v_{it}$  are most likely uncorrelated with  $z_{it}$  as well.

Noting that when  $T \rightarrow \infty$ , estimating the coefficients of each cross-sectional unit using the least squares method is consistent, Pesaran and Smith (1995) propose a mean group estimator that takes an average of the individual least squares estimated  $\tilde{\beta}_i$ ,

$$\hat{\beta} = \frac{1}{N} \sum_{i=1}^N \tilde{\beta}_i. \quad (16.23)$$

When both  $T$  and  $N \rightarrow \infty$ , (16.23) is consistent. However, the Monte Carlo studies conducted by Hsiao, Pesaran and Tahmisioglu (1999) show that (16.23) does not perform well in finite sample.

Under (16.4), (16.5), conditional on  $y_{i0}$  being a fixed constant, the Bayes estimators of  $\tilde{\beta}$  and  $\tilde{\gamma}$  are identical to (16.10), conditional on  $C_1$  and  $C_2$  with diffuse priors for

$\hat{\beta}$  and  $\bar{\gamma}$  (Hsiao, Pesaran and Tahmisioglu, 1999). The Monte Carlo studies show that a hierarchical Bayesian approach (e.g. Lindley and Smith, 1972) performs fairly well even when  $T$  is small and better than other consistent estimators despite the fact that  $y_{i0}$  being a fixed constant is not justifiable. This makes Bayes' procedure particularly appealing. Moreover, the implementation of a Bayesian approach has been substantially simplified by the recent advance of Markov Chain Monte Carlo methods (e.g. Gelfand and Smith, 1990).

#### 4 NONLINEAR MODELS

The existence of individual specific effects in nonlinear models is particularly difficult to handle. First, in general there is no simple transformation of the data to eliminate the individual specific effects. Second, the estimation of the individual specific parameters and the structural (or common) parameters are not independent of each other. When a panel contains a large number of individuals but only over a short time period, the error in the estimation of the individual specific coefficients is transmitted into the estimation of the structural parameters, and hence leads to inconsistency of the structural parameter estimation. If individual specific effects are treated as random, the estimation of the structural parameters often requires integrating out the individual effects. It is often computationally unwieldy (e.g. Hsiao, 1992).

Many approaches have been suggested for estimating nonlinear panel data models. One approach is the conditional maximum likelihood approach by conditioning the likelihood function on the minimum sufficient statistics for the incidental parameters to eliminate the individual specific parameters (Anderson, 1970; Chamberlain, 1980). For instance, consider a binary logit model in which the probability of  $y_{it} = 1$  is given by

$$P(y_{it} = 1) = \frac{e^{\alpha_i + \gamma' z_{it}}}{1 + e^{\alpha_i + \gamma' z_{it}}} \quad (16.24)$$

For ease of exposition, we assume that  $T = 2$ . It can be shown that when  $z_{i1} = 0$ ,  $z_{i2} = 1$  the MLE of  $\gamma$  converges to  $2\gamma$  as  $N \rightarrow \infty$ , which is not consistent (e.g. Hsiao, 1986). However, we may condition the likelihood function on the minimum sufficient statistic of  $\alpha_i$ . We note that for those individuals that  $y_{i1} + y_{i2} = 2$  or  $y_{i1} + y_{i2} = 0$ , they provide no information about  $\gamma$  since the former leads to  $\hat{\alpha}_i = \infty$  and the latter leads to  $\hat{\alpha}_i = -\infty$ . Only those individuals that  $y_{i1} + y_{i2} = 1$  provide information about  $\gamma$ . For individuals that  $y_{i1} + y_{i2} = 1$ , let  $d_i = 1$  if  $y_{i1} = 0$  and  $y_{i2} = 1$  and  $d_i = 0$  if  $y_{i1} = 1$  and  $y_{i2} = 0$ , then

$$\text{prob}(d_i = 1 | y_{i1} + y_{i2} = 1) = \frac{e^{\gamma'(z_{i2} - z_{i1})}}{1 + e^{\gamma'(z_{i2} - z_{i1})}}. \quad (16.25)$$

Equation (16.25) no longer contains the individual effects  $\alpha_i$  and is of the form of a standard logit model. Therefore, maximizing the conditional loglikelihood function on the subset of individuals with  $y_{i1} + y_{i2} = 1$  is consistent and asymptotically normally distributed as  $N$  tends to infinity.

The problem becomes more complicated if lagged dependent variables representing state dependence also appear in  $\underline{z}_{it}$  in the specification (16.24), Chamberlain (1984) shows that we need  $T \geq 4$  for the identification of a logit model of the form

$$P(y_{it} = 1 | \alpha_i, y_{i1}, \dots, y_{i,t-1}) = \frac{e^{\alpha_i + \gamma_1 y_{i,t-1}}}{1 + e^{\alpha_i + \gamma_1 y_{i,t-1}}}. \quad (16.26)$$

When exogenous variables are also present, consider the events  $A = \{y_{i1} = 0, y_{i3} = 1, y_{i4}\}$  and  $B = \{y_{i1}, y_{i2} = 1, y_{i3} = 0, y_{i4}\}$ , where  $y_{i1}$  and  $y_{i4}$  are either 0 or 1. Then

$$\begin{aligned} P(A | \underline{z}_i, \alpha_i) &= P_1(\underline{z}_i, \alpha_i)^{y_{i1}} (1 - P_1(\underline{z}_i, \alpha_i))^{1-y_{i1}} \\ &\cdot \frac{1}{1 + \exp(\gamma_1 y_{i1} + \underline{z}'_{i2} \gamma_2 + \alpha_i)} \cdot \frac{\exp(\underline{z}'_{i3} \gamma_2 + \alpha_i)}{1 + \exp(\underline{z}'_{i3} \gamma_2 + \alpha_i)} \\ &\cdot \frac{\exp y_{i4} (\gamma_1 + \underline{z}'_{i2} z_{i4} + \alpha_i)}{1 + \exp(\gamma_1 + \underline{z}'_{i2} z_{i4} + \alpha_i)}, \end{aligned} \quad (16.27)$$

and

$$\begin{aligned} P(B | \underline{z}_i, \alpha_i) &= P_1(\underline{z}_i, \alpha_i)^{y_{i1}} (1 - P_1(\underline{z}_i, \alpha_i))^{1-y_{i1}} \\ &\cdot \frac{\exp(\gamma_1 y_{i1} + \underline{z}'_{i2} \gamma_2 + \alpha_i)}{1 + \exp(\gamma_1 y_{i1} + \underline{z}'_{i2} \gamma_2 + \alpha_i)} \cdot \frac{1}{1 + \exp(\gamma_1 + \underline{z}'_{i3} \gamma_2 + \alpha_i)} \\ &\cdot \frac{\exp y_{i4} (\underline{z}'_{i2} z_{i4} + \alpha_i)}{1 + \exp(\underline{z}'_{i2} z_{i4} + \alpha_i)}, \end{aligned} \quad (16.28)$$

where  $P_1(\underline{z}_i, \alpha_i)$  denotes the probability that  $y_{i1} = 1$  given  $\alpha_i$  and  $\underline{z}'_i = (\underline{z}_{i1}, \dots, \underline{z}_{iT})$  where  $\underline{z}_{it}$  is a  $k_2 - 1$  dimensional exogenous variables and  $T = 4$  here. In general,  $P(A | \underline{z}_i, \alpha_i, A \cup B)$  will depend on  $\alpha_i$ , hence the conditional method in general will not work with the presence of exogenous explanatory variables.

However, if  $\underline{z}_{i3} = \underline{z}_{i4}$ , then

$$P(A | \underline{z}_i, \alpha_i, A \cup B, \underline{z}_{i3} = \underline{z}_{i4}) = \frac{1}{1 + \exp[(\underline{z}_{i2} - \underline{z}_{i3})' \gamma_2 + \gamma_1(y_{i4} - y_{i1})]} \quad (16.29)$$

will not depend on  $\alpha_i$ . Thus, Honoré and Kyriazidou (1997), Chintagunta, Kyriazidou and Perktold (1998) suggest maximizing

$$\begin{aligned} &\sum_{i=1}^N \sum_{1 \leq s < t \leq T-1} I(y_{is} + y_{it} = 1) K\left(\frac{\underline{z}_{i,s+1} - \underline{z}_{i,t+1}}{h_N}\right) \\ &\cdot \ln \left[ \frac{\exp[(\underline{z}_{is} - \underline{z}_{it})' \gamma_2 + \gamma_1(y_{i,s-1} - y_{i,t+1}) + \gamma_1(y_{i,s+1} - y_{i,t-1})] \mathbf{1}(t-s \geq 3)^{y_{is}}}{1 + \exp[(\underline{z}_{is} - \underline{z}_{it})' \gamma_2 + \gamma_1(y_{i,s-1} - y_{i,t+1}) + \gamma_1(y_{i,s+1} - y_{i,t-1})] \mathbf{1}(t-s \geq 3)} \right] \end{aligned} \quad (16.30)$$

where  $I(\cdot)$  denotes the indicator function,  $K(\cdot)$  is a kernel function such that  $K(v) \rightarrow 0$  as  $|v| \rightarrow \infty$ , and  $h_N$  is a bandwidth which shrinks to zero as  $N \rightarrow \infty$ . Honoré and Kyriazidou (1997) show that the estimator is consistent and asymptotically normal, with the rate of convergence proportional to  $\sqrt{Nh_N^{k_2-1}}$ . (For additional discussion, see chapter 17 by Maddala and Flores-Lagunes in this volume.)

Another approach to estimating a nonlinear panel data model is to apply some data transformation to eliminate the individual effects if the nonlinear model is of the form of a single index model where the index possesses a linear structure. Some semiparametric methods can then be applied to the transformed data. For instance, a binary choice model can be written in the form  $y_{it} = 1$  if  $y_{it}^* > 0$  and  $y_{it} = 0$  if  $y_{it}^* \leq 0$ , where

$$y_{it}^* = \alpha_i + \gamma' z_{it} + \varepsilon_{it}. \quad (16.31)$$

Thus, if  $\varepsilon_{it}$  follows a logistic distribution, we have a logit model (16.21). If  $\varepsilon_{it}$  is normally distributed, we have a probit model (e.g. Hsiao, 1992). Then

$$\begin{aligned} \gamma' z_{it} > \gamma' z_{i,t-1} &\Leftrightarrow E(y_{it} | z_{it}) > E(y_{i,t-1} | z_{i,t-1}), \\ \gamma' z_{it} = \gamma' z_{i,t-1} &\Leftrightarrow E(y_{it} | z_{it}) = E(y_{i,t-1} | z_{i,t-1}), \\ \gamma' z_{it} < \gamma' z_{i,t-1} &\Leftrightarrow E(y_{it} | z_{it}) < E(y_{i,t-1} | z_{i,t-1}). \end{aligned} \quad (16.32)$$

Rewriting (16.28) into the equivalent first difference form, Manski (1975) proposes a maximum score estimator (MS) that maximizes the sample average function

$$\frac{1}{N} \sum_{i=1}^N \sum_{t=2}^T \text{sgn}[(z_{it} - z_{i,t-1})'\gamma](y_{it} - y_{i,t-1}), \quad (16.33)$$

where  $\text{sgn}[(z_{it} - z_{i,t-1})'\gamma] = 1$  if  $(z_{it} - z_{i,t-1})'\gamma \geq 0$  and  $\text{sgn}[(z_{it} - z_{i,t-1})'\gamma] = -1$  if  $(z_{it} - z_{i,t-1})'\gamma < 0$ . The MS is consistent but is not root- $n$  consistent, where  $n = N(T - 1)$ . Its rate of convergence is  $n^{1/3}$  and  $n^{1/3}(\hat{\gamma} - \gamma)$  converges to a nonnormal distribution.<sup>3</sup>

A third approach is to find an orthogonal reparameterization of the fixed effects for each individual, say  $\alpha_i$ , to a new fixed effects, say  $g_i$ , which is independent of the structural parameters in the information matrix sense. The  $g_i$  are then integrated out of the likelihood with respect to an uninformative, uniform prior distribution which is independent of the prior distribution of the structural parameters. Lancaster (1998) uses a two period duration model to show that the marginal posterior density of the structural parameter possesses a mode which consistently estimates the true parameter.

While all these methods are quite ingenious, unfortunately, none of these approaches can claim general applicability. For instance, the conditional maximum likelihood cannot work for the probit model because there does not exist simple minimum sufficient statistics for the incidental parameters that are independent of the structural parameters. The data transformation approach cannot work if the model does not have a latent linear structure. The orthogonal reparameterization approach only works for some particular model. The general properties

of such procedures remain unknown. In short, the method and consistency of nonlinear panel data estimators must be established case by case.

## 5 SAMPLE ATTRITION AND SAMPLE SELECTION

Missing observations occur frequently in panel data. If individuals are missing randomly, most estimation methods for the balanced panel can be extended in a straightforward manner to the unbalanced panel (e.g. Hsiao, 1986). For instance, suppose that

$$d_{it}y_{it} = d_{it}[\alpha_i + \gamma' z_{it} + u_{it}], \quad (16.34)$$

where  $d_{it}$  is an observable scalar indicator variable which denotes whether information about  $(y_{it}, z_{it}')$  for the  $i$ th individual at  $t$ th time period is available or not. The indicator variable  $d_{it}$  is assumed to depend on a  $q$ -dimensional variables,  $w_{it}$ , individual specific effects  $\lambda_i$  and an unobservable error term  $\eta_{it}$ ,

$$d_{it} = I(\lambda_i + \delta' w_{it} + \eta_{it} > 0), \quad (16.35)$$

where  $I(\cdot)$  is the indicator function that takes the value of 1 if  $\lambda_i + \delta' w_{it} + \eta_{it} > 0$  and 0 otherwise. In other words, the indicator variable  $d_{it}$  determines whether  $(y_{it}, z_{it})$  in (16.34) is observed or not (e.g. Hausman and Wise, 1979).

Without sample selectivity, that is  $d_{it} = 1$  for all  $i$  and  $t$ , (16.31) is the standard variable intercept (or fixed effects) model for panel data discussed in Section 2. With sample selection and if  $\eta_{it}$  and  $u_{it}$  are correlated,  $E(u_{it} | z_{it}, d_{it} = 1) \neq 0$ . Let  $\theta(\cdot)$  denote the conditional expectation of  $u_{it}$  conditional on  $d_{it} = 1$  and  $w_{it}$ , then (16.31) can be written as

$$y_{it} = \alpha_i + \gamma' z_{it} + \theta(\lambda_i + \delta' w_{it}) + \varepsilon_{it}, \quad (16.36)$$

where  $E(\varepsilon_{it} | z_{it}, d_{it} = 1) = 0$ . The form of the selection function is derived from the joint distribution of  $u$  and  $\eta$ . For instance, if  $u$  and  $\eta$  are bivariate normal, then we have the Heckman (1979) sample selection model with  $\theta(\lambda_i + \delta' w_{it}) = \sigma_{u\eta} \frac{\phi(\lambda_i + \delta' w_{it})}{\Phi(\lambda_i + \delta' w_{it})}$ , where  $\sigma_{u\eta}$  denotes the covariance between  $u$  and  $\eta$ ,  $\phi(\cdot)$  and  $\Phi(\cdot)$  are standard normal density and distribution, respectively, and the variance of  $\eta$  is normalized to be 1. Therefore, in the presence of sample attrition or selection, regressing  $y_{it}$  on  $z_{it}$  using only the observed information is invalidated by two problems. First, the presence of the unobserved effects  $\alpha_i$ , and second, the “selection bias” arising from the fact that  $E(u_{it} | z_{it}, d_{it} = 1) = \theta(\lambda_i + \delta' w_{it})$ .

When individual effects are random and the joint distribution function of  $(u, \eta, \gamma_i, \lambda_i)$  is known, both the maximum likelihood and two- or multi-step estimators can be derived (e.g. Heckman, 1979; and Ryu, 1998). The resulting estimators are consistent and asymptotically normally distributed. The speed of convergence is proportional to the square root of the sample size. However, if the joint distribution of  $u$  and  $\eta$  is misspecified, then even without the presence of

$\alpha_i$ , both the maximum likelihood and Heckman (1979) two-step estimators will be inconsistent. This sensitivity of parameter estimate to the exact specification of the error distribution has motivated the interest in semiparametric methods.

The presence of individual effects is easily solved by pairwise differencing those individuals that are observed for two time periods  $t$  and  $s$ , i.e. who has  $d_{it} = d_{is} = 1$ . However, the sample selectivity factors are not eliminated by pairwise differencing. The expected value of  $y_{it} - y_{is}$  given  $d_{it} = 1$  and  $d_{is} = 1$  takes the form

$$E(y_{it} - y_{is} | d_{it} = 1, d_{is} = 1) = (\bar{z}_{it} - \bar{z}_{is})'\gamma + E[u_{it} - u_{is} | d_{it} = 1, d_{is} = 1]. \quad (16.37)$$

In general,

$$\theta_{its} = E(u_{it} - u_{is} | d_{it} = 1, d_{is} = 1) \neq 0 \quad (16.38)$$

and are different from each other. If  $(u_{it}, \eta_{it})$  are independent, identically distributed (iid) and are independent of  $\alpha_i, \lambda_i, \bar{z}$  and  $w$ , then

$$\begin{aligned} \theta_{it} &= E(u_{it} | d_{it} = 1, d_{is} = 1) = E(u_{it} | d_{it} = 1) \\ &= E(u_{it} | \eta_{it} > -w_{it}'\delta - \lambda_i) = \theta(\delta'w_{it} + \lambda_i), \end{aligned} \quad (16.39)$$

where the second equality is due to the independence over time assumption of the error vector and the third equality is due to the independence of the errors to the individual effects and the explanatory variables. The function  $\theta(\cdot)$  of the single index,  $\delta'w_{it} + \lambda_i$ , is the same over  $i$  and  $t$  because of the iid assumption of  $(u_{it}, \eta_{it})$ , but in general,  $\theta(\delta'w_{it} + \lambda_i) \neq \theta(\delta'w_{is} + \lambda_i)$  because of the time variation of the scalar index  $\delta'w_{it}$ . However, for an individual  $i$  that has  $\delta'w_{it} = \delta'w_{is}$  and  $d_{it} = d_{is} = 1$ , the sample selection effect  $\theta_{it}$  will be the same in the two periods. Therefore, for this particular individual, time differencing eliminates both the unobserved individual effect and the sample selection effect,

$$y_{it} - y_{is} = \gamma'(\bar{z}_{it} - \bar{z}_{is}) + (\varepsilon_{it} - \varepsilon_{is}). \quad (16.40)$$

This suggests estimating  $\gamma$  by the least squares from a subsample that consists of those observations that satisfy  $\delta'w_{it} = \delta'w_{is}$  and  $d_{it} = d_{is} = 1$ ,

$$\hat{\gamma} = \left[ \sum_{i=1}^N \sum_{1 \leq s < t \leq T_i} (\bar{z}_{it} - \bar{z}_{is})(\bar{z}_{it} - \bar{z}_{is})' \mathbf{1}\{(\bar{w}_{it} - \bar{w}_{is})'\delta = 0\} d_{it} d_{is} \right]^{-1} \left[ \sum_{i=1}^N \sum_{1 \leq s < t \leq T_i} (\bar{z}_{it} - \bar{z}_{is})(y_{it} - y_{is}) \mathbf{1}\{(\bar{w}_{it} - \bar{w}_{is})'\delta = 0\} d_{it} d_{is} \right] \quad (16.41)$$

where  $T_i$  denotes the number of  $i$ th individual's time series observations.

The estimator (16.41) cannot be directly implemented because  $\delta$  is unknown. Moreover, the scalar index  $\delta'w_{it}$  will typically be continuous if any of the variables

in  $w_{it}$  is continuous. Ahn and Powell (1993) note that if  $\theta$  is a sufficiently "smooth" function, and  $\hat{\delta}$  is a consistent estimator of  $\delta$ , observations for which the difference  $(w_{it} - w_{is})'\hat{\delta}$  is close to zero should have  $\theta_{it} - \theta_{is} \approx 0$ . They propose a two-step procedure. In the first step, consistent semiparameter estimates of the coefficients of the "selection" equation are obtained. The result is used to obtain estimates of the "single index,  $w_{it}\hat{\delta}$ " variables characterizing the selectivity bias in the equation of index. The second step of the approach estimates the parameters of the equation of interest by a weighted instrumental variables regression of pairwise differences in dependent variables in the sample on the corresponding differences in explanatory variables; the weights put more emphasis on pairs with  $w_{it}\hat{\delta} \approx w'_{i,t-1}\hat{\delta}$ .

Kyriazidou (1997) and Honoré and Kyriazidou (1998) generalize this concept and propose to estimate the fixed effects sample selection models in two steps: In the first step, estimate  $\delta$  by either the Anderson (1970), Chamberlain (1980) conditional maximum likelihood approach or the Manski (1975) maximum score method. In the second step, the estimated  $\hat{\delta}$  is used to estimate  $\gamma$ , based on pairs of observations for which  $d_{it} = d_{is} = 1$  and for which  $(w_{it} - w_{is})'\hat{\delta}$  is "close" to zero. This last requirement is operationalized by weighting each pair of observations with a weight that depends inversely on the magnitude of  $(w_{it} - w_{is})'\hat{\delta}$ , so that pairs with larger differences in the selection effects receive less weight in the estimation. The Kyriazidou (1997) estimator takes the form:

$$\begin{aligned} \hat{\gamma} = & \left\{ \sum_{i=1}^N \sum_{1 \leq s < t \leq T_i} (\bar{z}_{it} - \bar{z}_{is})(\bar{z}_{it} - \bar{z}_{is})' K \left[ \frac{(w_{it} - w_{is})'\hat{\delta}}{h_N} \right] d_{it} d_{is} \right\}^{-1} \\ & \left\{ \sum_{i=1}^N \sum_{1 \leq s < t \leq T_i} (\bar{z}_{it} - \bar{z}_{is})(y_{it} - y_{is}) K \left[ \frac{(w_{it} - w_{is})'\hat{\delta}}{h_N} \right] d_{it} d_{is} \right\} \end{aligned} \quad (16.42)$$

where  $K$  is a kernel density function which tends to zero as the magnitude of its argument increases and  $h_N$  is a positive constant that decreases to zero as  $N \rightarrow \infty$ .

Under appropriate regularity conditions, Kyriazidou (1997) shows that  $\hat{\gamma}$  (16.42) is consistent and asymptotically normally distributed. However, the rate of convergence is slower than the standard square root of the sample size.

## 6 CONCLUSIONS

There is an explosion of techniques and procedures for the analysis of panel data (e.g. Mátyás and Sevestre, 1996). In this chapter we have discussed some popular panel data models. We did not discuss issues of duration and count data models (e.g. Cameron and Trivedi, 1998; Heckman and Singer, 1984; Lancaster, 1990; Lancaster and Intrator, 1998), simulation-based inference (e.g. Gouriéroux and Monfort, 1993), specification analysis (e.g. Baltagi and Li, 1995; Lee, 1987; Li and Hsiao, 1998; Maddala, 1995; Wooldridge, 1995), measurement errors (e.g. Biorn, 1992; Griliches and Hausman, 1984; Hsiao, 1991; Hsiao and Taylor, 1991) pseudo panels or matched samples (e.g. Deaton, 1985; Moffit, 1993; Peracchi and Welsch,

1995; Verbeek, 1992), etc. In general, there does not exist a panacea for panel data analysis. It appears more fruitful to explicitly recognize the limitations of the data and focus attention on providing solutions for a specific type of model. A specific model often contains specific structural information that can be exploited. However, the power of panel data depends on the validity of the assumptions upon which the statistical methods have been built (e.g. Griliches, 1979).

## Notes

- \* This work was supported in part by National Science Foundation grant SBR96-19330. I would like to thank two referees for helpful comments.
- 1 Normality is made for ease of relating sampling approach and Bayesian approach estimators. It is not required.
- 2 See Chamberlain (1984), Hausman and Taylor (1981) for the approaches of estimating models when  $u$  and  $\varepsilon$  are correlated.
- 3 Under smooth conditions, Horowitz (1992) proposed a smoothed maximum score estimator that has a  $n^{-2/5}$  rate of convergence. With even stronger conditions Lee (1999) is able to propose a root- $n$  consistent semiparametric estimator.

## References

- Ahn, H., and J.L. Powell (1993). Semiparametric estimation of censored selection models with a nonparametric selection mechanism. *Journal of Econometrics* 58, 3–30.
- Ahn, S.C., and P. Schmidt (1995). Efficient estimation of models for dynamic panel data. *Journal of Econometrics* 68, 29–52.
- Akaike, H. (1973). Information theory and an extension of the maximum likelihood principle. In *Prac. 2nd Int. Symp. Information Theory*, pp. 267–81.
- Amemiya, T., and T. MacCurdy (1986). Instrumental variable estimation of an error components model. *Econometrica* 54, 869–81.
- Anderson, E.B. (1970). Asymptotic properties of conditional maximum likelihood estimators. *Journal of the Royal Statistical Society series B* 32, 283–301.
- Anderson, E.B. (1973). *Conditional Inference and Models for Measuring*. Kobenhavn: Mentalhygiejinsk Forlag.
- Anderson, T.W., and C. Hsiao (1981). Estimation of dynamic models with error components. *Journal of the American Statistical Association* 76, 598–606.
- Anderson, T.W., and C. Hsiao (1982). Formulation and estimation of dynamic models using panel data. *Journal of Econometrics* 18, 47–82.
- Arellano, M., and O. Bover (1995). Another look at the instrumental variable estimation of error-components models. *Journal of Econometrics* 68, 29–52.
- Balestra, P., and M. Nerlove (1966). Pooling cross-section and time series data in the estimation of a dynamic model: The demand for natural gas. *Econometrica* 34, 585–612.
- Baltagi, B.H. (1995). *Econometric Analysis of Panel Data*. New York: Wiley.
- Baltagi, B.H., and Q. Li (1995). Testing AR(1) against MA(1) disturbances in an error components model. *Journal of Econometrics* 68, 133–52.
- Bjorn, E. (1992). Econometrics of panel data with measurement errors. In L. Mátyás and P. Sevestre (eds.) *Econometrics of Panel Data: Theory and Applications*, pp. 152–95. Kluwer.
- Bhargava A., and J.D. Sargan (1983). Estimating dynamic random effects models from panel data covering short time periods. *Econometrica* 51, 1635–59.
- Blundell, R., and S. Bond (1998). Initial conditions and moment restrictions in dynamic panel data models. *Journal of Econometrics* 87, 115–43.

- Breusch, T.S., and A.R. Pagan (1979). A simple test for heteroscedasticity and random coefficient variation. *Econometrica* 47, 1287–94.
- Cameron, A.C., and P.K. Trivedi (1998). *Regression Analysis of Count Data*. Cambridge: Cambridge University Press.
- Chamberlain, G. (1980). Analysis of covariance with qualitative data. *Review of Economic Studies* 47, 225–38.
- Chamberlain, G. (1984). Panel data. In Z. Griliches and M. Intriligator (eds.), *Handbook of Econometrics, Volume 2*. pp. 1247–1318. Amsterdam: North-Holland.
- Chintagunta, P., E. Kyriazidou, and J. Perktold (1998). Panel data analysis of household brand choices. *Journal of Econometrics*.
- Deaton, A. (1985). Panel data from time series of cross-sections. *Journal of Econometrics* 30, 109–26.
- Gelfand, A.E., and A.F.M. Smith (1990). Sampling-based approaches to calculating marginal densities. *Journal of the American Statistical Association* 85, 398–409.
- Gouriéroux, C., and A. Monfort (1993). Simulation based inference: A survey with special reference to panel data models. *Journal of Econometrics* 59, 5–34.
- Griliches, Z. (1979). Sibling models and data in economics, beginning of a survey. *Journal of Political Economy* 87, supplement 2, S37–S64.
- Griliches, Z., and J.A. Hausman (1984). Errors-in-variables in panel data. *Journal of Econometrics* 31, 93–118.
- Hausman, J.A. (1978). Specification tests in econometrics. *Econometrica* 46, 1251–71.
- Hausman, J.A., and W.E. Taylor (1981). Panel data and unobservable individual effects. *Econometrica* 49, 1377–98.
- Hausman, J.A., and D.A. Wise (1979). Attrition bias in experimental and panel data: The Gary income maintenance experiment. *Econometrica* 47, 455–73.
- Heckman, J. (1979). Sample selection bias as a specification error. *Econometrica* 47, 153–61.
- Heckman, J.J., and B. Singer (1984). Econometric duration analysis. *Journal of Econometrics* 24, 63–132.
- Honoré, B.E., and E. Kyriazidou (1997). Panel data discrete choice models with lagged dependent variables. Mimeo.
- Honoré, B.E., and E. Kyriazidou (1998). Estimation of tobit-type models with individual specific effects. Mimeo.
- Hsiao, C. (1974). Statistical inference for a model with both random cross-sectional and time effects. *International Economic Review* 15, 12–30.
- Hsiao, C. (1975). Some estimation methods for a random coefficients model. *Econometrica* 43, 305–25.
- Hsiao, C. (1985). Benefits and limitations of panel data. *Econometric Reviews* 4, 121–74.
- Hsiao, C. (1986). *Analysis of Panel Data*. Econometric Society monographs No. 11, New York: Cambridge University Press.
- Hsiao, C. (1989). Consistent estimation for some nonlinear errors-in-variables models. *Journal of Econometrics* 41, 159–85.
- Hsiao, C. (1990). A mixed fixed and random coefficients framework for pooling cross-section and time series data. Paper presented at the Third Conference on Telecommunication Demand Analysis with Dynamic Regulation, Hilton Head, S. Carolina.
- Hsiao, C. (1991). Identification and estimation of latent binary choice models using panel data. *Review of Economic Studies* 58, 717–31.
- Hsiao, C. (1992). Nonlinear latent variables models. In L. Matyas and P. Sevestre (eds.) *Econometrics of Panel Data*. pp. 242–61. Kluwer.
- Hsiao, C. (1995). Panel analysis for metric data. G. Arminger, C.C. Clogg, and M.E. Sobel *Handbook of Statistical Modelling in the Social and Behavioral Sciences*, (3rd edn). pp. 361–400. Plenum.

- Hsiao, C., and D. Mountain (1995). A framework for regional modelling and impact analysis – an analysis of demand for electricity by large municipalities in Ontario, Canada. *Journal of Regional Science* 34, 361–85.
- Hsiao, C., and A.K. Tahmisioglu (1997). A panel analysis of liquidity constraints and firm investment. *Journal of the American Statistical Association* 92, 455–65.
- Hsiao, C., and G. Taylor (1991). Some remarks on measurement errors and the identification of panel data models. *Statistical Neerlandica* 45, 187–94.
- Hsiao, C., T.W. Applebe, and C.R. Dineen (1993). A general framework for panel data models – with an application to Canadian customer-dialed long distance telephone service. *Journal of Econometrics* 59, 63–86.
- Hsiao, C., D.C. Mountain, K.Y. Tsui, and M.W. Luke Chan (1989). Modelling Ontario regional electricity system demand using a mixed fixed and random coefficients approach. *Regional Science and Urban Economics* 19, 567–87.
- Hsiao, C., M.H. Pesaran, and A.K. Tahmisioglu (1998). Maximum likelihood estimation of fixed effects dynamic panel data models covering short time periods. Mimeo, Cambridge University.
- Hsiao, C., M.H. Pesaran, and A.K. Tahmisioglu (1999). Bayes estimation of short-run coefficients in dynamic panel data models. In C. Hsiao, L.F. Lee, K. Lahiri, and M.H. Pesaran (eds.) *Analysis of Panels and Limited Dependent Variables Models*. Cambridge: Cambridge University Press, pp. 268–96.
- Hsiao, C., B.H. Sun, and J. Lightwood (1995). Fixed vs. random effects specification for panel data analysis. Paper presented in International Panel Data Conference, Paris.
- Horowitz, J. (1992). A smoothed maximum score estimator for the binary response model. *Econometrica* 60, 505–31.
- Kuh, E. (1963). *Capital Stock Growth: A Micro-Econometric Approach*. Amsterdam: North-Holland.
- Kyriazidou, E. (1997). Estimation of a panel data sample selection model. *Econometrica* 65, 1335–64.
- Lancaster, T. (1990). *The Econometric Analysis of Transition Data*. Cambridge: Cambridge University Press.
- Lancaster, T. (1998). Some econometrics of scarring. In C. Hsiao, K. Morimune, and J. Powell (eds.) *Nonlinear Statistical Inference*. Cambridge: Cambridge University Press.
- Lancaster, T., and O. Intrator (1998). Panel data with survival: Hospitalization of HIV-positive patients. *Journal of the American Statistical Association* 93, 46–53.
- Lee, L.F. (1987). Nonparametric testing of discrete panel data models. *Journal of Econometrics* 34, 147–78.
- Lee, M.J. (1999). A root-n consistent semiparametric estimator for related effect binary response panel data. *Econometrica* 67, 427–33.
- Li, Q., and C. Hsiao (1998). Testing serial correlation in semiparametric panel data models. *Journal of Econometrics* 87, 207–37.
- Lindley, D.V., and A.F.M. Smith (1972). Bayes estimates for the linear model. *Journal of the Royal Statistical Society B* 34, 1–41.
- Maddala, G.S. (1995). Specification tests in limited dependent variable models. In G.S. Maddala, P.C.B. Phillips, and T.N. Srinivasan (eds.) *Advances in Econometrics and Quantitative Economics: Essays in Honor of C.R. Rao*. Oxford: Blackwell. pp. 1–49.
- Manski, C.F. (1975). Maximum score estimation of the stochastic utility model of choice. *Journal of Econometrics* 3, 205–28.
- Mátyás, L., and P. Sevestre (1996). *The Econometrics of Panel Data – Handbook of Theory and Applications*, 2nd edn. Dordrecht: Kluwer.

- Min, C.K., and A. Zellner (1993). Bayesian and non-Bayesian methods for combining models and forecasts with applications to forecasting international growth rate. *Journal of Econometrics* 56, 89–118.
- Moffitt, R. (1993). Identification and estimation of dynamic models with a time series of repeated cross-sections. *Journal of Econometrics* 59, 99–123.
- Mundlak, Y. (1978). On the pooling of time series and cross section data. *Econometrica* 46, 69–85.
- Neyman, J., and E.L. Scott (1948). Consistent estimates based on partially consistent observations. *Econometrica* 16, 1–32.
- Peracchi, F., and F. Welsch (1995). How representative are matched cross-sections? Evidence from the current population survey. *Journal of Econometrics* 68, 153–80.
- Pesaran, M.H., and R. Smith (1995). Estimation of long-run relationships from dynamic heterogeneous panels. *Journal of Econometrics* 68, 79–114.
- Ryu, K.K. (1998). New approach to attrition problem in longitudinal studies. In C. Hsiao, K. Morimune, and J. Powell (eds.) *Nonlinear Statistical Inference*. Cambridge University Press.
- Schwarz, G. (1978). Estimating the dimension of a model. *Annals of Statistics* 6, 461–4.
- Swamy, P.A.V.B. (1970). Efficient inference in a random coefficient regression model. *Econometrica* 38, 311–23.
- Verbeek, M. (1992). The design of panel surveys and the treatment of missing observations. Ph.D. dissertation, Tilburg University.
- Wallace, T.D., and A. Hussain (1969). The use of error components models in combining cross-section with time series data. *Econometrica* 37, 55–72.
- Wansbeek, T., and T. Knaap (1999). Estimating a dynamic panel data model with heterogeneous trend. *Annales d'Economie et de Statistique* 55–6, 331–50.
- Wooldridge, J.M. (1995). Selection corrections for panel data models under conditional mean independence assumptions. *Journal of Econometrics* 68, 115–32.
- Zellner, A. (1962). An efficient method of estimating seemingly unrelated regressions and tests for aggregation bias. *Journal of the American Statistical Association* 57, 348–68.
- Ziliak, J.P. (1997). Efficient estimation with panel data when instruments are predetermined: An empirical comparison of moment-condition estimators. *Journal of Business and Economic Statistics* 15, 419–31.

CHAPTER SEVENTEEN

# Qualitative Response Models

*G.S. Maddala and A. Flores-Lagunes\**

## 1 INTRODUCTION

This chapter deals with regression models when the dependent variable is qualitative. There are many situations in economics where the dependent variable takes discrete values, due to the qualitative nature of many behavioral responses. For instance, consider a worker deciding whether or not to participate in the labor force, a consumer deciding whether or not to buy a good, etc.

There are good reviews of qualitative response models (QRM) in the literature (Amemiya, 1981; McFadden, 1981, 1984; Maddala, 1983, chs. 2–5). Here we focus on some basic concepts about QRM, some topics not covered in the above reviews, and on some recent work in this area. We discuss binary and multinomial response models, specification tests, panel data with qualitative variables, semi-parametric estimation, and simulation methods.

Section 2 introduces some basic concepts about QRM, with an emphasis on univariate QRM and, in particular, on the estimation of binary and multinomial logit and probit models. A more in-depth review of all types of QRM, both univariate and multivariate, can be found in Maddala (1983, chs. 2–5).

Some concepts about specification tests in QRM are discussed in Section 3. Pagan and Vella (1989) argued that the use of specification tests in qualitative response models is not common since they are difficult to compute, but that has changed as computationally simpler specification tests have become available.

Sections 4 to 6 review some further topics in QRM. Panel data with qualitative variables (Section 4) has been an intensive area of research, both theoretical and applied, since the early reviews by Heckman (1981) and Chamberlain (1980, 1984). The principal issue in this literature is on controlling for heterogeneity and state dependence.

The other two topics, semiparametric estimation (Section 5) and simulation methods (Section 6), correspond to two specific problems in qualitative response models: the inconsistency of estimators when the distributional assumption of the model is incorrect, and the computational problem of evaluating higher order integrals in multinomial qualitative response models.

## 2 BINARY AND MULTINOMIAL RESPONSE MODELS

In this section we present the basic analysis of models with a single explanatory variable that is observed as a dichotomous (binary or binomial) or polychotomous (multinomial) variable.

Both binary and multinomial response models are models in which the dependent variable assumes discrete values. The simplest of these models is that in which the dependent variable  $y$  is binary; for instance,  $y$  can be defined as 1 if the individual is in the labor force, 0 otherwise.

When a dependent variable  $y$  can assume more than two values it can be classified as (i) categorical variable and (ii) count (noncategorical) variable. For instance, a categorical variable  $y$  may be defined as:  $y = 1$  if the individual earns less than \$10,000;  $y = 2$  if the individual earns between \$10,000 and \$30,000; and  $y = 3$  if the individual earns more than \$30,000. Note that, as the name indicates, the variable categorizes individuals into different categories. A count variable is discrete but it does not categorize, like the number of strikes on a country in a given year. The methods of analysis are different for models with categorical and count variables (see Cameron and Trivedi, Chapter 15 in this companion).

Categorical variables can be further classified as (i) unordered, (ii) ordered, and (iii) sequential. Unordered categorical variables can be defined in any order desired, for instance:  $y = 1$  if occupation is lawyer;  $y = 2$  if occupation is teacher; and  $y = 3$  if occupation is doctor. An example of an ordered categorical variable is the one above concerning the level of earnings. Finally, a sequential categorical variable can be illustrated as:  $y = 1$  if the individual has not completed high school;  $y = 2$  if the individual has completed high school but not college;  $y = 3$  if the individual has completed college but not a higher degree; and  $y = 4$  if the individual has completed a professional degree.

Let us now turn our attention to binary response models. We motivate these models by introducing the linear probability model. This is a regression model in which the dependent variable  $y$  is a binary variable. The model can be written as:

$$y_i = \beta' x_i + u_i, \quad (17.1)$$

with  $E(u_i) = 0$ . The conditional expectation  $E(y_i/x_i)$  is equal to  $\beta' x_i$ , which is interpreted as the probability that the event will occur given the  $x_i$ . The fitted value,  $\hat{y}_i = \hat{\beta}' x_i$ , will give the estimated probability that the event will occur given the particular value of  $x$ .

Since the model is heteroskedastic (the reader can easily check it), it has to be estimated by weighted least squares. However, the linear probability model is seldom used because the least squares method is not fully efficient due to the

nonnormality of the residuals  $u_i$ , and more importantly, because in many cases  $E(y_i/x_i)$ , interpreted as a probability, can lie outside the limits (0, 1).

Two alternative models that avoid the previous two problems are widely used for the estimation of binary response models: the probit and the logit. Both models assume that there is an underlying response variable  $y_i^*$  defined by the regression relationship  $y_i^* = \beta'x_i + u_i$ . In practice,  $y_i^*$  is unobservable but we observe a dummy variable defined by:

$$y_i = 1 \quad \text{if} \quad y_i^* > 0; \quad y_i = 0 \quad \text{otherwise.} \quad (17.2)$$

Note that in this formulation,  $\beta'x_i$  is not  $E(y_i/x_i)$  as in the linear probability model; it is  $E(y_i^*/x_i)$ . From the regression relationship for  $y_i^*$  and (17.2) we get:

$$P(y_i = 1) = P(u_i > -\beta'x_i) = 1 - F(-\beta'x_i), \quad (17.3)$$

where  $F$  is the cumulative distribution function for  $u_i$ .

In this case the observed values of  $y$  are just realizations of a binomial process with probabilities given by (17.3) and varying from trial to trial (depending on  $x_i$ ). Hence, the loglikelihood function is:

$$\log L = \sum_{i=1}^n y_i \log F(\beta'x_i) + \sum_{i=1}^n (1 - y_i) \log [1 - F(\beta'x_i)]. \quad (17.4)$$

The functional form for  $F$  depends on the assumptions made about  $u_i$ . If the cumulative distribution assumed is the logistic we have the logit model, if it is assumed to be the normal distribution then we have the probit model. The logistic and the normal distributions are very close to each other, except at the tails; but even though the parameter estimates are usually close, they are not directly comparable because the logistic distribution has variance  $\pi^2/3$  rather than normalized to 1 as for the normal. Amemiya (1981) suggests that the logit estimates be multiplied by 0.625, instead of the exact  $3^{1/2}/\pi$ , arguing that it produces a closer approximation.

For the purpose of predicting effects of changes in one of the independent variables on the probability of belonging to a group, the derivative of the probability with respect to the particular independent variable needs to be computed. Letting  $x_{ik}$  and  $\beta_k$  be the  $k$ th elements of the vector of explanatory variables and parameters, respectively, the derivatives for the logit and probit models are given by:

$$\frac{\partial}{\partial x_{ik}} L(x_i'\beta) = \frac{\exp(x_i'\beta)}{[1 + \exp(x_i'\beta)]^2} \beta_k = \beta_k P(y_i = 1)P(y_i = 0) \quad \text{for the logit, and} \quad (17.5)$$

$$\frac{\partial}{\partial x_{ik}} \Phi(x_i'\beta) = \phi(x_i'\beta)\beta_k \quad \text{for the probit,} \quad (17.6)$$

where  $L$ ,  $\Phi$ , and  $\phi$  are the logistic cdf, and the standard normal cdf and pdf, respectively. In both models, we need to calculate the derivatives at different levels of the explanatory variables to get an idea of the range of variation of the resulting changes in probabilities. A common practice in empirical work is to evaluate them at the mean of the vector of independent variables.

The estimation of logit and probit models is done by maximization of the log-likelihood function (17.4) after substituting either the logistic or normal distribution for the functional form  $F$ . Since the derivatives of the loglikelihood function are nonlinear in  $\beta$ , we have to use iterative methods like the Newton–Raphson or the scoring methods. The asymptotic covariance matrix, which can be used for hypothesis testing, is obtained by inverting the corresponding information matrix. In practice, the logit and probit models are readily available in statistical software. When dealing with multiple observations (as with grouped data), a general method based on weighted least squares, known as the minimum chi-square method, can be used (see Maddala, 1983, section 2.8).

Regarding multinomial response models, we will only cover the case of unordered categorical variables. The estimation of models with ordered and sequential categorical variables follow the same rationale. The reader is referred to sections 2.13 and 2.14 of Maddala (1983) for the particulars involved. Models with count data are often estimated using Poisson regression, which is covered in section 2.15 of the same monograph and in Cameron and Trivedi's chapter in this companion.

The multinomial logit (MNL) and probit (MNP) models are often used to estimate models with unordered categorical variables. The MNP model, however, involves the computation of multidimensional integrals which for models with variables taking more than 3 or 4 values are infeasible to compute by direct means. Nonetheless, we can use simulation methods to evaluate such integrals. These methods are reviewed in Section 6 below and in Chapter 22 by Geweke, Houser, and Keane in this companion.

Both MNL and MNP models can be motivated by a random utility formulation. Let  $y_{ij}^*(i = 1, \dots, n, j = 1, \dots, m)$  be the stochastic utility associated with the  $j$ th alternative for individual  $i$ , with

$$y_{ij}^* = \beta'_j x_i + u_{ij}, \quad (17.7)$$

where  $x_i$  are explanatory variables,  $\beta_j$  are unknown parameters and  $u_{ij}$  is an unobservable random variable. We assume that the individual chooses the alternative for which the associated utility is highest. Define a set of dummy variables  $y_{ij} = 1$  if the  $i$ th individual chooses the  $j$ th alternative,  $y_{ij} = 0$  otherwise. Then, for example, the probability that alternative 1 is chosen is given by:

$$P_{i1} = P(y_{i1} = 1) = \int_{-\infty}^{x'_1 \beta_1 - x'_j \beta_2} \dots \int_{-\infty}^{x'_1 \beta_1 - x'_j \beta_m} f(\eta_{21}, \dots, \eta_{m1}) d\eta_{21} \dots d\eta_{m1}, \quad (17.8)$$

where  $\eta_{kj} = u_{ik} - u_{ij}$ . Considering the observations as arising from a multinomial distribution with probabilities given by  $P_{ij}$ , the loglikelihood function for either the MNL or MNP models can be written as:

$$\log L = \sum_{i=1}^n \sum_{j=1}^m y_{ij} \log P_{ij}. \quad (17.9)$$

For the MNL model, we assume that the  $u_{ij}$ s follow independent extreme-value distributions. McFadden (1974) showed that the probabilities in (17.9) for the MNL model are given by:

$$P_{ij} = \frac{\exp(\beta'_j x_i)}{\sum_{j=1}^{m-1} \exp(\beta'_j x_i)}. \quad (17.10)$$

This model is computationally convenient since it avoids the problem of evaluating multidimensional integrals as opposed to the MNP model (see below). The estimation of the MNL model is through maximum likelihood (ML) using an iterative method (since again the derivatives are nonlinear in  $\beta$ ). The asymptotic covariance matrix is also obtained by inverting the corresponding information matrix.

McFadden (1974) also suggested the conditional logit model. The main difference between this model and the MNL considered in (17.10) is that the former considers the effect of choice characteristics on the determinants of choice probabilities as well, whereas the MNL model makes the choice probabilities dependent on individual characteristics only. To illustrate this, let  $x_{ij}$  denote the vector of the values of the characteristics of choice  $j$  as perceived by individual  $i$ . Then, the probability that individual  $i$  chooses alternative  $j$  is

$$P_{ij} = \frac{\exp(\beta' x_{ij})}{\sum_{k=1}^m \exp(\beta' x_{ik})}. \quad (17.11)$$

Note that, as opposed to (17.10),  $P_{ij}$  does not have different coefficient vectors  $\beta_j$ . In (17.11) the vector  $\beta$  gives the vector of implicit prices for the characteristics.<sup>1</sup> The conditional logit model is similarly estimated by ML.

Both MNL and conditional logit models have the property referred to as “independence of irrelevant alternatives” (IIA). This is because the odds ratio for any two choices  $i$  and  $j$  is  $\exp(\beta' x_i)/\exp(\beta' x_j)$ , which is the same irrespective of the total number  $m$  of choices considered. If the individual is offered an expanded choice set, that does not change its odds ratio. This property is in fact a drawback in many applications. Debreu (1960) pointed out that these models predict too high a joint probability of selection for two alternatives that are in fact perceived as similar rather than independent by the individual. To see this, consider the following choices: (i) red bus, (ii) blue bus, and (iii) auto. Suppose that consumers treat the two buses as equivalent and are indifferent between auto and bus. Then, the relative odds of alternatives (i) and (iii) depend on the presence of alternative (ii). They are 1:1 if choice (ii) is not present. They are 1:2 if choice (ii) is present. However, this is inconsistent with the IIA property.

The MNP model does not have the IIA property and thus would be preferred to the MNL model when such property is inappropriate. Nonetheless, recall the

MNP model computational problems mentioned above. To illustrate this, consider only three alternatives in the formulation (17.7) (i.e.  $j = 3$ ), and assume that the residuals have a trivariate normal distribution with mean vector zero and some covariance matrix  $\Sigma$ . Using the same definitions as above, to compute the corresponding probabilities as in (17.8), the  $\eta_{kj}$ 's will have a bivariate normal distribution with covariance matrix  $\Omega_1$  (that can be derived from  $\Sigma$  by standard formulae for normal densities), requiring thus the computation of bivariate integrals. It is easy to see that we must deal with trivariate integrals when considering four alternatives, and so on.

There are other models for the analysis of polychotomous variables. The elimination-by-aspects (EBA) model assumes that each alternative is described by a set of aspects (characteristics) and that at each stage of the process an aspect is selected. The selection of the aspect eliminates alternatives that do not contain the selected aspect, and the selection continues until a single alternative remains. Aspects common to all alternatives do not affect the choice probabilities avoiding thus the IIA property. Another model is the hierarchical elimination-by-aspects (HEBA) model. When aspects have a tree structure, the EBA model reduces to the HEBA model. McFadden (1981) introduced the generalized extreme value (GEV) and the nested multinomial logit (NMNL) models, which are models based on a random utility formulation where the error distribution is a multivariate generalization of the extreme-value distribution, and are formulated in such a way that the multivariate integrals analogous to (17.8) are analytically tractable. McFadden (1984) considers the GEV model as an elimination model that can be expressed in latent variable form and the NMNL model as a hierarchical elimination model based on the GEV structure. None of these models have the IIA property. References for the EBA, HEBA, GEV, and NMNL models are in Maddala (1983, ch. 3) and McFadden (1984).

So far, we have only discussed models with univariate qualitative variables. There are also models with multivariate qualitative variables in the literature. Among such models are, for instance, simultaneous equation models with qualitative variables, simultaneous equation models with qualitative variables and structural shift, and others. Multivariate qualitative response models are reviewed in Maddala (1983, ch. 5).

### 3 SPECIFICATION TESTS

The area of specification tests in QRM (in fact all limited dependent variable models) has been surveyed in Pagan and Vella (1989), and more exhaustively in Maddala (1995). Pagan and Vella argue that the use of specification tests in limited dependent variable models is not common because they are difficult to compute. They suggest a number of specification tests based on conditional moments that they argue are easier to use. Maddala (1995) discusses these tests and shows that they are equivalent to score tests.

Most of the specification tests for qualitative response models are Rao's score tests (known as LM tests in the econometric literature) but there are others which are Hausman (1978) type tests, White's (1982) IM (information matrix) tests,

and CM (conditional moment) tests suggested by Newey (1985) and Tauchen (1985). A detailed description of all these categories of tests is in Maddala (1995, pp. 2–10).

The survey by Maddala also considers tests for a variety of specification errors as follows: tests for heteroskedasticity, tests for normality, tests for autocorrelation, tests for sample selection, tests for exogeneity, tests for omitted variables, tests for stability, and multinomial logit specification tests, which are tests for the IIA property. The reader is referred to that paper for a detailed review of all these tests and references to the literature. In the rest of this section, we will limit ourselves to provide some general ideas.

1. Most of the specification tests for qualitative response models use the concept of generalized residuals introduced by Gouriéroux, Monfort, and Trognon (1987). These are the counterpart for nonlinear models (such as probit and logit) of the usual residuals in linear regression, and are used as the latter in hypothesis testing.
2. On the different specification tests, one can use the information matrix or its sample estimates like the Hessian of the loglikelihood function for the model or the outer product gradient (OPG), to avoid evaluating expectations (which are often cumbersome). It is important to keep in mind that, whereas the OPG version is computationally simpler, it has been found to have bad small sample properties (see Davidson and MacKinnon, 1989; Orme, 1990; and the references in those papers). Tests based on the Hessian version have been found to perform only slightly better than the OPG (Taylor, 1991). Therefore, whenever possible, one should try to use the exact information matrix.<sup>2</sup>
3. Specification tests in qualitative response models are particularly useful when the unrestricted model is much more complicated to estimate than the restricted model, which is often the case. For instance, consider testing for sample selection in the following censored bivariate probit model:

$$\begin{aligned} y_{1i}^* &= \beta_1' x_{1i} + u_{1i} \\ y_{2i}^* &= \beta_2' x_{2i} + u_{2i}, \end{aligned} \tag{17.12}$$

where  $y_{1i}^*$  is censored by the second equation (selection equation). The estimation of this model under sample selection is not trivial (since it involves bivariate integrals). Hence, it is useful to have a test for the null of no sample selectivity, since univariate probit models can then be used.

The above reviews do not cover specification tests in panel data, a topic discussed in the following section, nor models estimated by semiparametric and simulation based methods, covered in Sections 5 and 6. Baltagi (1999) briefly mentions score type tests for fixed effects in logit and probit models with panel data (section 5 of his paper). Lee (1997) has an exhaustive discussion of all the specification errors discussed in Maddala's paper in the context of models estimated by simulation methods.

## 4 PANEL DATA WITH QUALITATIVE VARIABLES

Early reviews of models of panel data with qualitative variables are in Heckman (1981) and Chamberlain (1980, 1984). This work in the early 1980s is reviewed in Maddala (1987) which discusses fixed effects logit and probit models, random effects probit models, autoregressive logit models, autoregressive probit models, and the problems of serial correlation and state dependence.

The main issue in panel data models with qualitative response variables is that the presence of individual effects (heterogeneity) complicates the estimation. For instance, the fixed effects model for panel data does not give consistent estimates of the slope parameters. Chamberlain (1980) suggested a conditional ML approach in which the likelihood function of the model is conditioned on sufficient statistics of the incidental parameters  $\alpha_i$  (i.e. the fixed effects), which in this case are  $\sum_t y_{it}$ . The method is illustrated in Maddala (1987) for the logit model. The probit model cannot be used with Chamberlain's method due to computational problems.

In contrast, when using the random effects approach, the probit model is about the only choice, since on the multivariate logistic distribution the correlations between the error term and the cross section units are all constrained to be 1/2 (see Johnson and Kotz, 1972, pp. 293–4), which is implausible to assume in most applications. The random effects probit model implies serial correlation of a specific nature, the equicorrelation model, for which there is an efficient ML algorithm due to Butler and Moffitt (1982). For the case of a general type of serial correlation, the Avery, Hansen, and Hotz (1983) method of GMM was discussed in Maddala (1987).

An additional problem in panel data models with qualitative response variables occurs if there is “state dependence”, which can be interpreted as a situation in which an individual’s past state  $y_{i,t-1}$  helps predict his or her current state  $y_{i,t}$ , after allowing for individual effects.

The problem of discrete choice panel data models with lagged dependent variables was first discussed in Chamberlain (1985). The model considered is a model with heterogeneity and state dependence. Heterogeneity is captured by the inclusion of individual specific effects  $\alpha_i$ , and state dependence is captured by the inclusion of the lagged dependent variable  $y_{i,t-1}$ . The model considered by Chamberlain is:

$$\begin{aligned}
 P(y_{i0} = 1/\alpha_i) &= P_0(\alpha_i) \\
 P(y_{it} = 1/\alpha_i, y_{i0}, y_{i1}, \dots, y_{i,t-1}) \\
 &= \frac{\exp(\alpha_i + \gamma y_{i,t-1})}{1 + \exp(\alpha_i + \gamma y_{i,t-1})} \quad t = 1, 2 \dots T, \quad T \geq 3. \tag{17.13}
 \end{aligned}$$

The procedure used by Chamberlain (1985) is to derive a set of probabilities that do not depend on the individual effects. Let  $T = 3$  and consider the events:

$$A = \{y_{i0} = d_0, y_{i1} = 0, y_{i2} = 1, y_{i3} = d_3\}$$

$$B = \{y_{i0} = d_0, y_{i1} = 1, y_{i2} = 0, y_{i3} = d_3\} \quad (17.14)$$

where  $d_0$  and  $d_3$  are either 0 or 1. Then the conditional probabilities  $P(A/\alpha_i, A \cup B)$  and  $P(B/\alpha_i, A \cup B)$  will not depend on  $\alpha_i$ . However, this method will not work in the presence of explanatory variables.

Honoré and Kyriazidou (1998), to be referred to as H-K, consider a more general model with exogenous explanatory variables. Their model is:

$$\begin{aligned} P(y_{it} = 1/x_i, \alpha_i) &= P_0(x_i, \alpha_i) \\ P(y_{it} = 1/x_i, \alpha_i, y_{i0}, y_{i1}, \dots, y_{i,t-1}) &= \frac{\exp(x_{it}\beta + \gamma y_{i,t-1} + \alpha_i)}{1 + \exp(x_{it}\beta + \gamma y_{i,t-1} + \alpha_i)} \quad t = 1, 2 \dots T, \quad T \geq 3 \end{aligned} \quad (17.15)$$

H-K observe that if say  $x_{i2} = x_{i3}$  then these conditional probabilities do not depend on  $\alpha_i$ . In practice this is not a reasonable assumption. In the special case where all explanatory variables are discrete, and  $P(x_{i2} = x_{i3}) > 0$ , one may estimate  $\beta$  and  $\gamma$  by maximizing with respect to  $b$  and  $g$  the weighted likelihood function:

$$\sum_{i=1}^N I(y_{i1} + y_{i2} = 1)I(x_{i2} - x_{i3} = 0) \ln \frac{\exp\{(x_{i1} - x_{i2})b + g(y_{i0} - y_{i3})\}^{y_{i1}}}{1 + \exp\{(x_{i1} - x_{i2})b + g(y_{i0} - y_{i3})\}} \quad (17.16)$$

where  $I(\cdot)$  are indicator functions. The resulting estimator will have all the usual properties (consistency and root- $n$  asymptotic normality).

In the case where the  $xs$  are not discrete, H-K suggest replacing the indicator function  $I(x_{i2} - x_{i3} = 0)$  by a weight function  $k[(x_{i2} - x_{i3})/\sigma_n]$  with weights depending inversely on the magnitude of  $x_{i2} - x_{i3}$  giving more weight to observations for which  $x_{i2}$  is "close" to  $x_{i3}$ .  $\sigma_n$  is a bandwidth that shrinks as  $n$  increases and  $k(\cdot)$  is the kernel, chosen so that  $k(v) \rightarrow 0$  as  $v \rightarrow \infty$ .

H-K show that the resultant estimators are consistent and asymptotically normal, although their rate of convergence will be slower than  $n^{-\frac{1}{2}}$  and will depend on the number of covariates in  $x_{it}$ . Furthermore, for identification, their model requires that  $x_{i2} - x_{i3}$  be continuously distributed with support in a neighborhood of 0, and that  $x_{i1} - x_{i2}$  has sufficient variation conditional on the event  $x_{i2} - x_{i3} = 0$ . H-K also extend their model for the general case of  $M$  alternatives.

## 5 SEMIPARAMETRIC ESTIMATION

We have seen that in the estimation of QRM the assumption is made about the error term being distributed according to some known distribution (i.e. logistic or normal). The validity of this assumption may be rejected using appropriate

specification tests, implying (if correct) the inconsistency of the estimators. For this reason, several semiparametric methods of estimation have been proposed in the literature, which, based on weaker assumptions about the error distribution, estimate the relevant parameters of the model.

In this section, we will consider four semiparametric estimators: the Maximum Score (MS) estimator (Manski, 1975, 1985; Manski and Thompson, 1986), the Quasi Maximum Likelihood (QML) estimator (Klein and Spady, 1993), the Generalized Maximum Likelihood (GML) estimator (Cosslett, 1983), and the semi-nonparametric (SNP) estimator (Gallant and Nychka, 1987). This is, by no means, an exhaustive review of the many existing semiparametric estimators for QRM. The interested reader is referred to the comprehensive review of semiparametric models in the context of limited dependent variables by Powell (1994).

Our main focus is on the binary response model, however, the MS and QML can be extended to multinomial response models. Other estimators for the multinomial response model have been developed by Thompson (1989a, 1989b) and Lee (1994).

Gabler, Laisney, and Lechner (1993) use a small-scale Monte Carlo study to conclude that the bias associated with incorrect distributional assumptions in the binary probit model can be substantial both in finite samples and asymptotically. Similar results for the binary logit model are shown by Horowitz (1993).

The MS estimator for the parameter vector  $\beta$  is intuitively obtained by maximizing the number of correct predictions of  $y$  (dependent variable) by the sign of the latent regression function  $x'\beta$ . More formally, this estimator maximizes the following score function over a suitable parameter space:

$$S_n(\beta) \equiv \sum_{i=1}^N [y_i I\{x'_i \beta > 0\} + (1 - y_i) I\{x'_i \beta < 0\}]. \quad (17.17)$$

The only restriction imposed on the distribution of the error term is to have conditional median zero, which ensures consistency of the estimator. Despite the fact that the MS estimator is consistent, it is not root- $n$  consistent under standard regularity conditions, nor asymptotically normal. Its rate of convergence is  $n^{\frac{1}{3}}$ , and  $n^{\frac{1}{3}}(\hat{\beta} - \beta_0)$  converges to a nonnormal distribution (Kim and Pollard, 1990).

The QML estimator, proposed by Klein and Spady (1993), is obtained as follows.<sup>3</sup> By the use of Bayes' theorem, we can write the probability of  $y = 1$  given  $x$  as:

$$P\{y = 1/x\} = \frac{P\{y = 1\}g(x\beta/y = 1)}{g(x\beta)} \equiv p(x\beta). \quad (17.18)$$

The mean of  $y$  is a consistent estimator for  $P\{y = 1\}$  above, and we can also obtain consistent estimators for  $g(x\beta/y = 1)$  and  $g(x\beta)$  using kernel estimates (for known  $\beta$ ). Having those estimators, we can obtain consistent estimates  $\hat{p}(x\beta)$  that are used to compute the QML estimator by substituting them for  $F(\beta'x_i)$  on (17.4) and maximizing the expression.

This estimator is strongly consistent, asymptotically normal, and attains the semiparametric efficiency bound (Newey, 1990); however, it relies on the stringent assumption of independence between the error term and the regressors.

Cosslett's (1983) GML estimator uses the likelihood function for the binary choice model (17.4) and maximizes it with respect to the functional form  $F$  as well as  $\beta$ , subject to the condition that  $F$  is a distribution function. This maximization is done in two steps: first,  $\beta$  is fixed and (17.4) is maximized with respect to  $F$  to obtain  $\hat{F}$ ; then, in the second step,  $\hat{F}$  is used as the correct distribution and (17.4) is maximized with respect to  $\beta$ , obtaining  $\hat{\beta}$ . Cosslett (1983) derived the conditions under which this estimator is consistent; however, its asymptotic normality has not been proved yet, and the second step is computationally costly since the likelihood function at that stage varies in discrete steps over the parameter space.

The paper by Gabler, Laisney, and Lechner (1993), to be referred as GLL, is an application of the SNP estimator proposed by Gallant and Nychka (1987) to the binary choice model. In general, the term semi-nonparametric is used when the goal is to approximate a function of interest with a parametric approximation (in this case the distribution function of the errors). The larger the number of observations available to estimate the function, the larger the number of parameters to be used in approximating the function of interest and, as a result, the better the approximation.

Gallant and Nychka (1987) proposed the approximation of any smooth density that has a moment generating function with the Hermite form:<sup>4</sup>

$$h^*(u) = \sum_{i,j=0}^K \alpha_i \alpha_j u^{i+j} \exp\{-(u/\delta)^2\}. \quad (17.19)$$

The estimation method used by GLL is to fix  $K$  (the degree of the Hermite polynomial) in a somewhat optimal way and use the framework of pseudo maximum likelihood (PML).

For the binary choice model, we use the above approximation for the density  $F$  in the likelihood function (17.4). The likelihood function is then maximized taking into account two additional restrictions on  $\alpha$  and  $\delta$ :  $u$  has to have zero expectation, and the requirement of unit mass below the density  $F$ . Moreover, a condition for consistency in this approach is the degree  $K$  of the approximation to increase with the sample size.<sup>5</sup> For the details of the estimation method refer to GLL's paper.

The asymptotic normality of this SNP estimator follows from the asymptotic theory of PML estimation. References in this literature are White (1982, 1983) and Gouriéroux and Monfort (1995). This allows for hypothesis testing with the usual techniques. Note that the PML theory is being used for the asymptotic distribution of a potentially inconsistent estimator (due to the use of a fixed  $K$ ).<sup>6</sup> In addition, let us note that this asymptotic normality result is not a standard asymptotic result, since the asymptotic variance is only approximated as  $N \rightarrow \infty$ , holding  $K$  at the fixed value chosen a priori.<sup>7</sup>

While all four estimators discussed above relax distributional assumptions about the error term, each of them shows drawbacks that are worth considering before their use in empirical applications. The MS estimator has been used more than the other estimators due to its computational feasibility (it is even available in some software packages, although with size limitations), but its asymptotic nonnormality is a drawback when hypothesis testing beyond parameter significance is needed since additional steps are required. This last drawback is shared with GML, which is also computationally costly. On the other hand, QML is asymptotically normal and efficient, but should not be used when dependence between the error term and the regressors is suspected, limiting its application considerably. Finally, the relatively new SNP estimator does not share the drawbacks of the MS and GML estimators, and the authors offer a GAUSS program upon request. However, the estimator is inconsistent if  $K$  is not chosen carefully, and the nonglobal concavity of its objective function require good starting values, which usually implies the estimation of a probit model. Besides, it shares QML's drawback regarding the dependence between the error term and the regressors. More evidence about the relative performance of these estimators is needed.

## 6 SIMULATION METHODS

In binary QRM there is little basis to choose between the logit and probit models because of the similarity of the cumulative normal and the logistic distributions. In the multinomial situation this is not the case. The multinomial logit (MNL) model has a closed form representation and is computationally tractable but lacks flexibility due to the IIA property. The multinomial probit (MNP) model gives flexibility but is computationally burdensome because of the need to evaluate multidimensional integrals. Another problem which involves high dimensional integrals is the probit (or tobit) model with serially correlated errors.

Until a few years ago, only 3 or 4 dimensional integrals could be evaluated. However, developments during the last decade on simulation based estimation allow us to estimate otherwise intractable models by approximating high-dimensional integrals. A simple exposition of the econometric literature on simulation methods can be found in Stern (1997).

The generic problem for simulation is to evaluate expressions like  $E[g(x)] = \int g(x)f(x)dx$  where  $x$  is a random variable with density  $f(x)$  and  $g(x)$  is a function of  $x$ . The simulation method consists of drawing samples  $x_1, x_2, \dots, x_N$  from  $f(x)$  and computing  $g(x_i)$ . Then,  $\widehat{E[g(x)]} = \frac{1}{N} \sum_{i=1}^N g(x_i)$  is an unbiased estimator of  $E[g(x)]$ , and its variance is  $\text{var}(g(x_i))/N$ . As  $N \rightarrow \infty$ , the variance of the simulator goes to zero.  $f(x)$  can be a discrete distribution.

Different applications of the simulation method depend on refinements in the way the sampling of  $f(x)$  is done. These refinements in probability simulators fall into the following categories: importance sampling, GHK simulator, Gibbs sampling, and antithetic variables. Importance sampling involves oversampling some "important" parts of  $f(x)$  from a well-chosen distribution  $g(x)$  from which it is easy to sample. The GHK simulator algorithm decomposes the joint density

of the errors into a product of conditional densities. It has been found to perform very well for simulating multinomial probit probabilities. Gibbs sampling is an iterative sampling method from conditional densities. Finally, the method of antithetic variables can be used on the above probability simulators to reduce sampling costs and the variance of the simulator through the sampling of pairs of random variables which are negatively correlated.<sup>8</sup>

The estimation methods commonly used are the method of simulated moments (MSM), the method of simulated likelihood (MSL), and the method of simulated scores (MSS). A detailed review of the above estimation methods can be found in Hajivassiliou and Ruud (1994) and in Geweke, Houser, and Keane, Chapter 22, in this companion.

The MSM is based on the GMM (generalized method of moments) method of estimation. The least squares method is a simple example of the method of moments. The GMM method depends on orthogonality conditions. For example, Avery, Hansen, and Hotz (1983) suggest how to estimate the multiperiod probit model by a set of within period orthogonality conditions that allows consistent estimation of the parameters and their standard errors under serial correlation. In many problems, the orthogonality conditions cannot be evaluated analytically. This is where simulation methods help. The resulting estimator is the MSM estimator. Let us illustrate this method with the MNP model. The loglikelihood function for a sample of size  $n$  with  $m$  alternatives is given by (17.9). The MSM consists of simulating  $\tilde{P}(j/x, \beta)$  using the multivariate normal distribution and substituting it in the moment equation:

$$\sum_{i=1}^n \sum_{j=1}^m w_{ijk} [y_{ij} - \tilde{P}(j/x, \beta)] = 0, \quad (17.20)$$

for some weighting function  $w$  ( $k$  is an index taking on the values  $1, \dots, K$  where  $K$  is the dimension of  $\beta$ ). The estimator  $\hat{\beta}_{\text{MSM}}$  is obtained by solving (17.20).

The MSL is based on the maximum likelihood method of estimation and is useful when the likelihood function is not easily computed by analytical methods. In the context of multinomial choice models, MSL consists in simulating the actual choice probabilities given by  $P_{ij}(x, \beta)$  in (17.9).  $\hat{\beta}_{\text{MSL}}$  is obtained by maximizing (17.9) once the simulated choice probabilities are substituted.

The MSS is based on the idea that the score statistic (the derivative of the likelihood function) should have an expected value of zero at the true value of  $\beta$ . The potential advantage of the MSS relative to MSM is that it uses the efficiency properties of ML, but it is computationally more difficult than the MSM since the weight function itself has to be simulated, while it is analytically computed on the MSM. To see this, note that the MSS implies simulating  $\tilde{P}(j/x, \beta)$  in the following expression:

$$\sum_{i=1}^n \sum_{j=1}^m \frac{1}{P(j/x, \beta)} \frac{\partial P(j/x, \beta)}{\partial \beta} [y_{ij} - P(j/x, \beta)] = 0, \quad (17.21)$$

which makes clear that the weight function itself has to be simulated.

There are several papers using simulation methods but there are very few that compare the different methods. Geweke, Keane, and Runkle (1994) compare different methods in the estimation of the multinomial probit model, based on two Monte Carlo experiments for a seven choice model. They compare the MSL estimator using the GHK recursive probability simulator, the MSM method using the GHK recursive probability simulator and kernel-smoothed frequency simulators, and Bayesian methods based on Gibbs sampling. Overall, the Gibbs sampling algorithm had a slight edge, while the relative performance of the MSM and MSL (based on the GHK simulator) was difficult to evaluate. The MSM with the kernel-smoothed frequency simulator was clearly inferior.

In another paper, Geweke, Keane, and Runkle (1997) again compare the Bayesian methods based on Gibbs sampling and the MSM and MSL methods based on the GHK simulator. They do Monte Carlo studies with AR(1) errors in the multinomial multiperiod probit model, finding that the Gibbs sampling algorithm performs better than the MSM and MSL, especially under strong serial correlation in the disturbances (e.g. an AR(1) parameter of 0.8). They also find that to have root mean squared errors (RMSEs) for MSL and MSM within 10 percent of the RMSEs by the Gibbs sampling method one needs samples of 160 and 80 draws respectively (much higher than the 20 draws normally used). Thus, with serially correlated errors, the performance ranking is Gibbs sampling first, MSM second, and MSL last.

Keane (1993) discusses in detail the MSM estimator of limited dependent variable models with general patterns of serial correlation. He uses the recursive GHK algorithm. He argues that the equicorrelation model (implicit in the random effects probit model), for which the Butler and Moffitt algorithm works, is not empirically valid.

## Notes

- \* We would like to thank three anonymous referees, the editor, Stephen Cosslett, and Kajal Lahiri for helpful comments. Remaining errors are our own. A. Flores-Lagunes gratefully acknowledges financial support from the National Council for Science and Technology of Mexico (CONACYT).
- 1 In (17.11) we need some normalization, like setting the first element of  $\beta$  to 1.
- 2 In the case of the binary choice models (logit and probit), Davidson and MacKinnon (1989) show that the score test statistic based on the exact information matrix can be computed easily using a particular artificial regression. See also Davidson and MacKinnon Chapter 1 in this companion.
- 3 We draw here from Gouriéroux (1989).
- 4 Note that the approximating density is positive for all  $u$ .
- 5 GLL suggest, based on simulations, that using  $K = 3$  allows for considerable flexibility in the distribution.
- 6 The asymptotic bias of the estimator will become negligible as  $K$  increases. GLL suggest starting with  $K = 3$  (see previous fn.) and use score tests to determine whether such a value of  $K$  is appropriate.
- 7 The corresponding theorems can be found in Gallant and Nychka (1987).
- 8 References on probability simulators are Geweke, Keane, and Runkle (1994, 1997), Stern (1997), and Tanner (1996). For the reader interested in applying these methods,

Vassilis Hajivassiliou offers GAUSS and FORTRAN routines for the GHK simulator and other simulators at the following electronic address: <http://econ.lse.ac.uk/~vassilis/pub/simulation/gauss/>

## References

- Amemiya, T. (1981). Qualitative response models: A survey. *Journal of Economic Literature* 19, 483–536.
- Avery, R., L. Hansen, and V.J. Hotz (1983). Multiperiod probit models and orthogonality condition estimation. *International Economic Review* 24, 21–35.
- Baltagi, B.H. (1999). Specification tests in panel data models using artificial regressions. *Annales d'Economie et de Statistique* 55–6, 277–98.
- Butler, I., and R. Moffitt (1982). A computationally efficient procedure for the one factor multinomial probit model. *Econometrica* 50, 761–4.
- Chamberlain, G. (1980). Analysis of covariance with qualitative data. *Review of Economic Studies*, 47, 225–38.
- Chamberlain, G. (1984). Panel data. In Z. Griliches and M.D. Intrilligator (eds.) *Handbook of Econometrics*, Volume 2. Amsterdam and New York: North-Holland, ch. 22.
- Chamberlain, G. (1985). Heterogeneity, omitted variable bias and duration dependence. In J.J. Heckman and B. Singer (eds.) *Longitudinal Analysis of Labor Market Data*. Cambridge: Cambridge University Press.
- Cosslett, S.R. (1983). Distribution-free maximum likelihood estimator of the binary choice model. *Econometrica* 51, 765–82.
- Davidson, R., and J.G. MacKinnon (1989). Testing for consistency using artificial regressions. *Econometric Theory* 50, 363–84.
- Debreu, G. (1960). Review of R.D. Luce "Individual Choice Behavior". *American Economic Review* 50, 186–8.
- Gabler, S., F. Laisney, and M. Lechner (1993). Seminonparametric estimation of binary-choice models with application to labor-force participation. *Journal of Business and Economic Statistics* 11, 61–80.
- Gallant, R., and D.W. Nychka (1987). Seminonparametric maximum likelihood estimation. *Econometrica* 55, 363–90.
- Geweke, J.F., M.P. Keane, and D.E. Runkle (1994). Alternative computational approaches to inference in the multinomial probit model. *Review of Economics and Statistics* 76, 609–32.
- Geweke, J.F., M.P. Keane, and D.E. Runkle (1997). Statistical inference in the multinomial multiperiod probit model. *Journal of Econometrics* 80, 125–65.
- Gouriéroux, C. (1989). *Econométrie des Variables Qualitatives*, 2nd edn. Paris: Economica.
- Gouriéroux, C., and A. Monfort (1995). *Statistics and Econometrics Models*, Volume 1. Cambridge: Cambridge University Press.
- Gouriéroux, C., A. Monfort, and A. Trognon (1987). Generalised residuals. *Journal of Econometrics* 34, 5–32.
- Hajivassiliou, V.A., and P.A. Ruud (1994). Classical estimation methods for LDV models using simulation. In R.F. Engle and D.L. McFadden (eds.) *Handbook of Econometrics*, Volume 4. Amsterdam and New York: North-Holland, ch. 40.
- Hausman, J. (1978). Specification tests in econometrics. *Econometrica* 46, 1251–71.
- Heckman, J.J. (1981). Statistical models for discrete panel data. In C.F. Manski and D.L. McFadden (eds.) *Structural Analysis of Discrete Data with Econometric Applications*. Cambridge, MA: MIT Press.
- Honoré, B.E., and E. Kyriazidou (2000). Panel data discrete choice models with lagged dependent variables. *Econometrica* 68, 839–74.

- Horowitz, J.L. (1993). Semiparametric and nonparametric estimation of quantal response models. In G.S. Maddala, C.R. Rao, and H.D. Vinod (eds.) *Handbook of Statistics*, Volume 11. Amsterdam and New York: North-Holland, ch. 2.
- Johnson, N.L. and S. Kotz (1972). *Continuous Multivariate Distributions*. New York: Wiley.
- Keane, M.P. (1993). Simulation estimation for panel data models with limited dependent variables. In G.S. Maddala, C.R. Rao, and H.D. Vinod (eds.) *Handbook of Statistics*, Volume 11. Amsterdam and New York: North-Holland, ch. 20.
- Kim, J., and D. Pollard (1990). Cube root asymptotics. *The Annals of Statistics* 18, 191–219.
- Klein, R.W., and R.H. Spady (1993). An efficient semiparametric estimator for discrete choice models. *Econometrica* 61, 387–421.
- Lee, L.F. (1994). Semiparametric instrumental variables estimation of simultaneous equation sample selection models. *Journal of Econometrics* 63, 341–88.
- Lee, L.F. (1997). Some common structures of simulated specification tests in multinormal discrete and limited dependent variable models. Working Paper 97/04, Hong Kong University of Science and Technology.
- Maddala, G.S. (1983). *Limited Dependent and Qualitative Variables in Econometrics*. Cambridge: Cambridge University Press.
- Maddala, G.S. (1987). Limited dependent variable models using panel data. *Journal of Human Resources* 22, 307–38.
- Maddala, G.S. (1995). Specification tests in limited dependent variable models. In G.S. Maddala, P.C.B. Phillips and T.N. Srinivasan (eds.) *Advances in Econometrics and Quantitative Economics: Essays in Honor of C.R. Rao*. pp. 1–49. Oxford: Blackwell.
- Manski, C.F. (1975). Maximum score estimation of the stochastic utility model of choice. *Journal of Econometrics* 3, 205–28.
- Manski, C.F. (1985). Semiparametric analysis of discrete response: Asymptotic properties of the maximum score estimator. *Journal of Econometrics* 27, 313–33.
- Manski, C.F., and S. Thompson (1986). Operational characteristics of maximum score estimation. *Journal of Econometrics* 32, 85–108.
- McFadden, D.L. (1974). Conditional logit analysis of qualitative choice behavior. In P. Zarembka (ed.) *Frontiers in Econometrics*. New York: Academic Press.
- McFadden, D.L. (1981). Econometric models of probabilistic choice. In C.F. Manski and D.L. McFadden (eds.) *Structural Analysis of Discrete Data with Econometric Applications*. Cambridge, MA: MIT Press.
- McFadden, D.L. (1984). Econometric analysis of qualitative response models. In Z. Griliches and M.D. Intriligator (eds.) *Handbook of Econometrics*, Volume 2. Amsterdam and New York: North-Holland, ch. 24.
- Newey, W.K. (1985). Maximum likelihood specification testing and conditional moment tests. *Econometrica* 53, 1047–70.
- Newey, W.K. (1990). Semiparametric efficiency bounds. *Journal of Applied Econometrics* 5, 99–135.
- Orme, C. (1990). The small-sample performance of the information matrix test. *Journal of Econometrics* 46, 309–31.
- Pagan, A.R. and F. Vella (1989). Diagnostic tests for models based on unit record data: A survey. *Journal of Applied Econometrics* 4, 529–59.
- Powell, J. (1994). Estimation of semiparametric models. In R.F. Engle and D.L. McFadden (eds.) *Handbook of Econometrics*, Volume 4. Amsterdam and New York: North-Holland, ch. 41.
- Stern, S. (1997). Simulation-based estimation. *Journal of Economic Literature* 35, 2006–39.
- Tanner, M.A. (1996). *Tools for Statistical Inference: Methods for the Exploration of Posterior Distributions and Likelihood Functions*. New York: Springer-Verlag.

- Tauchen, G. (1985). Diagnostic testing and evaluation of maximum likelihood models. *Journal of Econometrics* 30, 415–43.
- Taylor, L. (1991). Testing exclusion restrictions for misspecified tobit model. *Economics Letters* 37, 411–16.
- Thompson, T.S. (1989a). Identification of semiparametric discrete choice models. Discussion Paper 249, Center for Economic Research, University of Minnesota.
- Thompson, T.S. (1989b). Least squares estimation of semiparametric discrete choice models. Manuscript, Department of Economics, University of Minnesota.
- White, H. (1982). Maximum likelihood estimation of misspecified models. *Econometrica* 50, 1–25.
- White, H. (1983). Corrigendum. *Econometrica* 51, 513.

---

CHAPTER EIGHTEEN

# Self-Selection

*Lung-fei Lee\**

## 1 INTRODUCTION

This paper provides some account on econometric models and analysis of sample selection problems. The paper is divided into three parts. The first part considers possible selection-bias issues in econometric data. Selection biases can occur as the observed outcomes are results of individual's self-selection. The second part points out some development on econometric models with sample selection. This part concentrates on the specification, estimation, and test problems for parametric models. The third part lists some of the development of semiparametric estimation of sample selection models. Related surveys on the earlier development on sample selection models are Maddala (1983) and Amemiya (1984), which concern mainly parametric specification and estimation. Recent developments of the subject concern semiparametric estimation methods. Surveys on the latter are Powell (1994), Vella (1998), and M.-J. Lee (1997). Powell's survey covers broad areas of semiparametric estimation methods for microeconometric models in addition to sample selection models. A large part of the survey in Vella (1998) concerns the recent development on sample selection panel models. Treatment effect models as compared with sample selection models have been discussed in M.-J. Lee (1997). Because of their coverage and many developments on panel and treatment effect models being currently in their development stages in working paper format and because of space limitation, we skip these topics in this survey.

## 2 SAMPLE SELECTION BIAS

### 2.1 Self-selection

The problem of selection bias in economics arises when sampling observations are generated from the population by rules other than simple random sampling. Consequently, the sample representation of a true population is distorted. This is the essence of the selection problem. Distorted sample generation may be the

outcome of sample collection by surveyors. More importantly, distorted sample observations result from self-selection decisions by the agents being studied. A sample generated by self-selection may not represent the true population distribution of characteristics no matter how big the sample size. However, self-selection biases can be corrected to produce an accurate description of the underlying population if the underlying sampling generating processes can be understood and relevant identification conditions are available. Economic theories and institutional settings can provide guidance. It is for this reason that the econometrics of self-selection is, by and large, a subject of microeconomics.

The issue of selectivity bias first arose in labor economics, namely, the determinants of occupational wages in Roy (1951) and labor supply behavior of females in Gronau (1974) and Heckman (1974). Consider the labor supply problem of females in a free society. In a population of women, each individual is characterized by her endowments of observable and unobservable characteristics. (All characteristics are, of course, known to an individual herself, but some may be unobservable to an investigator.) She has the freedom to engage in market activities. It may be observed that only a subsample of the population is engaged in market employment and reports wages. A researcher or a policy maker may be interested in identifying the determinants of wages for working women so as to understand the determinants of wages for all women. The decision to work or not to work is not random as it is made in accordance with an individual's own interest. Consequently, the working and nonworking samples may have different characteristics. Sample selection bias arises when some component of the work decision is relevant to the wage determining process or the (expected) wage is a factor in the working decision. When the relationship between the work decision and the wage is purely through the observables and those observable variables are exogenous, the sample selection bias is controlled for when all relevant exogenous variables are included in the equations. The possibility of sample selection bias arises when there are unobservable characteristics that influence both the observed outcomes and the decision process.

## 2.2 Some conventional sample selection models

The tobit model assumes that the censoring threshold is deterministic and known. A generalization of the tobit model assumes that the censoring threshold is an unobservable stochastic variable. This generalization consists of two latent regression functions defined on the population:  $y_1^* = x_1\beta_1 + u_1$  and  $y_2^* = x_2\beta_2 + u_2$ . The sample observation is  $(y_I, I)$ , where  $y_I = y_1^*$  and  $I = 1$  if  $y_1^* \geq y_2^*$ , and  $I = 0$  if  $y_1^* < y_2^*$ . An example of this model is a labor supply model (Gronau, 1974; Heckman, 1974; Nelson, 1977), where  $y_1^*$  is an offered wage and  $y_2^*$  is the reservation wage of an individual. In a labor supply model, the individual maximizes utility with respect to income and leisure time subject to income and time constraints:  $\max\{U(t, c, u) : c = y_1^*(T - t) + c_0, t \leq T\}$ , where  $T$  is the available time,  $c_0$  is nonlabor income available,  $c$  is the total income, and  $u$  represents unobserved characteristics of an individual. The reservation wage  $y_2^*$  is  $(\partial U / \partial t) / (\partial U / \partial c)|_{t=T}$ . The market wage  $y_1^*$  can be observed only for the worker. This formulation can

further be generalized to incorporate fixed costs of participation. A general framework for these models can be written as

$$y^* = x\beta + u, \quad \text{and} \quad I^* = z\gamma - \varepsilon, \quad (18.1)$$

where  $y = y^*$  can be observed only if  $I^* > 0$ . This two-equation formulation provides the prototypical sample selection model in econometrics. The sample is censored if the sign of  $I^*$  is observable in addition to  $Iy$ . It is a truncated case if only the event  $I = 1$  and its corresponding sample observations of  $y$  are available.

Sample data can be generated by individuals making choices of belonging to one or another group, i.e. by the self-selection of individuals. A prototypical choice theoretic model of self-selection is that of Roy (1951). Roy (1951) discussed the problem of individuals choosing between two professions, hunting and fishing, based on their productivity (income) in each. There is a latent population of skills. While every person can, in principle, do the work in each "occupation", self-interest drives individuals to choose the occupation that produces the highest income for them. Roy's model is special in that an individual chooses his occupation based on the highest income among occupations. A more general setting is that individuals choose between several alternatives based on their preferences and the (potential) outcomes can be factors in their utility functions (Lee, 1978; Willis and Rosen, 1979).

A self-selection model with two alternatives and a potential outcome equation for each alternative can be

$$y_1^* = x_1\beta_1 + u_1, \quad \text{and} \quad y_2^* = x_2\beta_2 + u_2, \quad (18.2)$$

with a choice equation

$$I^* = z\gamma - \varepsilon. \quad (18.3)$$

The sample observation  $(I, y)$  is  $I = 1$  and  $y = y_1^*$  if  $I^* > 0$ , and  $I = 0$  and  $y = y_2^*$  if  $I^* \leq 0$ . For cases with polychotomous choices, a self-selection model with  $m$  alternatives and  $m_1$  potential outcome equations, where  $0 < m_1 \leq m$ , is

$$y_j = x_j\beta_j + u_j, \quad j = 1, \dots, m_1, \quad (18.4)$$

and

$$U_j = z_j\gamma + v_j, \quad j = 1, \dots, m. \quad (18.5)$$

The  $U_j$  represents the utility of the alternative  $j$ . Outcomes are available for some  $m_1$  alternatives. The outcome  $y_j$  can be observed only if the alternative  $j$  is chosen by the individual. In a utility maximization framework, the alternative  $j$  will be chosen if  $U_j > U_l$  for all  $l \neq j$ ,  $l = 1, \dots, m$ .

The selection equations in the preceding models provide discrete choices. In some cases, the selection equations may provide more sample information than

discrete choices. A censored regression selection criterion is such a case. A model of female labor supply without participation cost is an example. The market wage can be observed when the hour of work of an individual is positive and the hours-of-work equation can be modeled by a tobit model. A sample selection model with a tobit (censored) selection rule can be specified as

$$y_1^* = x\gamma + u_1, \quad \text{and} \quad y_2^* = x\beta + u_2, \quad (18.6)$$

where  $(y_1, y_2)$  can be observed such that  $(y_1, y_2) = (y_1^*, y_2^*)$  when  $y_1^* > 0$ . This model provides additional information in that positive values of  $y_1^*$  can be observed instead of just the sign of  $y_1^*$ . Other models that are of interest are simultaneous equation models and panel data models.

An important feature of a sample selection model is its usage to investigate potential outcomes or opportunity costs besides observed outcomes. For sectorial wage or occupational choices with self-selection, Roy's model (Roy, 1951; Heckman and Honoré, 1990) has emphasized comparative advantage in individuals and its effect on income distribution. For example, the comparative advantage measure of Sattinger (1978) involves the computation of opportunity costs for forgone choices (Lee, 1995). Opportunity costs are counterfactual outcomes. The evaluations of counterfactual outcomes are important in social welfare programs (Heckman and Robb, 1985; Bjorklund and Moffitt, 1987) because of their policy implications.

### 3 PARAMETRIC ESTIMATION

#### 3.1 Two-stage estimation

Consider the estimation of the model (18.1). Assume that  $u$  and  $\varepsilon$  are jointly normally distributed with zero-means and the variance of  $\varepsilon$  being normalized to be a unity, i.e.  $\text{var}(\varepsilon) = 1$ . The least squares procedure applied to the observed  $y$  and  $x$  will give inconsistent estimates, if  $E(u|x, z\gamma \geq \varepsilon)$  is not zero and is correlated with  $x$ . This omitted selection-bias term needs to be corrected for consistent estimation (Heckman, 1979). With normally distributed disturbances,  $E(\varepsilon|x, z, I=1) = -\sigma_{1\varepsilon} \frac{\phi(z\gamma)}{\Phi(z\gamma)}$ , where  $\phi$  and  $\Phi$  denote, respectively, the standard normal density and distribution functions and  $\sigma_{1\varepsilon}$  is the covariance of  $u$  and  $\varepsilon$ . The bias-corrected regression equation is  $y = x\beta - \sigma_{1\varepsilon} \frac{\phi(z\gamma)}{\Phi(z\gamma)} + \eta$ , where  $E(\eta|x, z, I=1) = 0$ . A two-stage method can be applied to estimate the corrected equation for  $\beta$  (Heckman, 1979). In the first stage,  $\gamma$  is estimated by the probit maximum likelihood method. The least squares method can then be applied to estimate  $\beta$  and  $\sigma_{1\varepsilon}$  in

$$y = x\beta + \sigma_{1\varepsilon} \left( -\frac{\phi(z\hat{\gamma})}{\Phi(z\hat{\gamma})} \right) + \tilde{\eta}, \quad (18.7)$$

with the observed subsample corresponding to  $I = 1$ , where  $\hat{\gamma}$  is the probit maximum likelihood estimate of  $\gamma$ . The estimator is consistent but the asymptotic

distribution of the two-stage estimator is not the conventional one of a linear regression model. The disturbances  $\eta$  and, hence,  $\hat{\eta}$  are heteroskedastic. The estimated bias corrected term is a generated regressor. The replacement of  $\gamma$  by the estimate  $\hat{\gamma}$  introduces additional errors, which are correlated across different sample units. By taking into account the heteroskedasticity of disturbances but ignoring the randomness of  $\hat{\gamma}$  in equation (18.7), the constructed asymptotic variance matrix of the two-stage estimates  $\hat{\beta}$  and  $\hat{\sigma}_{1e}$  will underestimate the correct one unless  $\sigma_{1e} = 0$ .

For the two-sector model (18.2)–(18.3), the expected observable outcome equation for  $y_1^*$  is  $E(y_1 | x, I = 1) = x_1\beta_1 + \sigma_{1e}(-\frac{\phi(z\gamma)}{\Phi(z\gamma)})$ , and the expected observable outcome for  $y_2^*$  will be  $E(y_2 | x, I = 0) = x_2\beta_2 + \sigma_{2e}(\frac{\phi(z\gamma)}{1 - \Phi(z\gamma)})$ . For this model, each bias-corrected equation can be either separately or jointly estimated by the two-stage method (Lee, 1978). While  $\sigma_{1e}$  is the covariance of  $u_1$  and  $\varepsilon$  and  $\sigma_{2e}$  is the covariance of  $u_2$  and  $\varepsilon$ , their signs may be of special interest for some empirical studies as they determine the direction of selection bias. When  $\sigma_{1e}$  is negative, the observed outcome  $y_1$  is subject to positive selection as  $\sigma_{1e}(-\frac{\phi(z\gamma)}{\Phi(z\gamma)})$  is strictly positive. For example, for the study on the return to college education, if high school graduates with high unobserved ability are likely to go to college and that ability could increase earning, one might expect that the observed earning of college graduates would be subject to positive selection. In some situations, negative selection might also be meaningful. For example, from the view of comparative advantage, what matters is the sign of  $\sigma_{2e} - \sigma_{1e}$ . As the measure of expected unobservable comparative advantage is  $E(u_1 - u_2 | x, I = 1) + E(u_2 - u_1 | x, I = 0) = (\sigma_{2e} - \sigma_{1e}) (\frac{\phi(z\gamma)}{\Phi(z\gamma)} + \frac{\phi(z\gamma)}{1 - \Phi(z\gamma)})$ , an individual has comparative advantage in his or her chosen task when  $\sigma_{2e} - \sigma_{1e}$  is positive. The relevance of comparative advantage in self-selection has been explored in Lee (1978) and Willis and Rosen (1979). Heckman and Honoré (1990) provide an in-depth analysis of the implications of Roy's model.

The normal distribution is a common distributional assumption for sample selection models. For the simplicity of the two-stage estimation method, Olsen (1980) pointed out that the crucial property underlying the derivation of the bias correction is the linearity of the conditional expectation of  $u$  given  $\varepsilon$ . Based on that property, Olsen specified the potential outcome equation  $y^* = x\beta + \lambda(\varepsilon - \mu_\varepsilon) + \eta$ , where  $\mu_\varepsilon$  is the mean of  $\varepsilon$ , as the basic structure and suggested a linear probability modification to correct for the selection bias in observed outcomes. This modification is useful as it provides an alternative parametric specification without being restricted to normal disturbances. The correction of selection bias is now dependent on the marginal distribution of  $\varepsilon$ . However, the selectivity bias terms may be sensitive to a specific choice probability model. Lee (1982) suggested the use of nonlinear transformations to overcome possible restrictions in Olsen's approach and suggested some flexible functional form and series expansion for the selection-bias correction. The dichotomous indicator  $I$  is determined by the selection decision such that  $I = 1$  if and only if  $z\gamma > \varepsilon$ . Thus, for any strictly increasing transformation  $J$ ,  $I = 1$  if and only if  $J(z\gamma) > J(\varepsilon)$ . The Olsen specification

was generalized into  $y^* = x\beta + \lambda(J(\varepsilon) - \mu_j) + \eta$ , where  $\mu_j = E(J(\varepsilon))$ . The selection bias term for the observed  $y$  is  $E(\varepsilon^* | J(z\gamma) \geq \varepsilon^*) = \frac{\mu(J(z\gamma))}{F(J(z\gamma))}$  where  $F$  is the distribution of  $\varepsilon$ ,  $\varepsilon^* = J(\varepsilon)$  and  $\mu(J(\varepsilon)) = \int_{J(-\infty)}^{J(z\gamma)} \varepsilon^* f_J(\varepsilon^*) d\varepsilon^*$  with  $f_J$  being the implied density function of  $\varepsilon^*$ . Conditional on  $y$  being observed, the outcome equation becomes  $y = x\beta + \lambda(\frac{\mu(J(z\gamma))}{F(z\gamma)} - \mu_j) + \eta$ , which can be estimated by a simple two-stage method. This approach generates a large class of models with selectivity while the probability choice model can be chosen to be a specific popular choice model and remains unchanged. The choice of different possible  $J$ s is, in general, a regressor selection problem in a linear regression model.

The specification of the conditional expectation of  $u$  on  $J(\varepsilon)$  being linear can be further relaxed by introducing a flexible expansion of distributions (Lee, 1982). It was noted that if the bivariate distribution of  $u^*$  and  $\varepsilon$ , where  $u^* = u/\sigma_u$ , could be represented by a bivariate Edgeworth expanded distribution, the conditional expectation of  $u^*$  conditional on  $\varepsilon$  would be  $E(u^* | \varepsilon) = \{\rho\varepsilon + \sum_{r \geq 3} [\rho A_{0r} H_{r+1}(\varepsilon)/r! + A_{1,r-1} H_{r-1}(\varepsilon)/(r-1)!]\}/D(\varepsilon)$  where  $D(\varepsilon) = 1 + \sum_{r \geq 3} A_{0r} H_r(\varepsilon)/r!$ ,  $A_{rs}$  are functions of cumulants of  $u^*$  and  $\varepsilon$ , and  $H_r(\varepsilon)$  is the  $r$ th order Hermite polynomial. When the marginal distribution of  $\varepsilon$  is normal,  $D(\varepsilon) = 1$ . Bias correction can be based on the expanded conditional expectation. With a normal  $\varepsilon$  (or a normally transformed  $\varepsilon^*$ ) and expanded terms up to  $r+s=4$  (instead of an infinite series), the bias corrected outcome equation is

$$\begin{aligned} y &= x\beta + \rho\sigma_u[-\phi(z\gamma)/\Phi(z\gamma)] + \mu_{12}\sigma_u[-(z\gamma)\phi(z\gamma)/(2\Phi(z\gamma))] \\ &\quad + (\mu_{13} - 3\rho)\sigma_u[(1 - (z\gamma)^2)\phi(z\gamma)/(6\Phi(z\gamma))] + \eta. \end{aligned} \quad (18.8)$$

The additional terms generalize the selection-bias correction of a normally distributed  $u$  to a flexible distributional one. The two-stage estimation can be applied to the estimation of equation (18.8). The correct variance matrix of the two-stage estimator shall take into account both the heteroskedasticity of the disturbance  $\eta$  and the distribution of the first-stage estimator of  $\gamma$ . For the model in (18.8), the exact expression of the heteroskedastic variance of  $\eta$  would be too complicated to be useful for estimation. To overcome that complication, Lee (1982) suggests the adoption of White's robust variance formulation. White's correction will estimate the first component of the correct variance matrix. The second component will be the variance and covariances due to the first stage estimate of  $\gamma$ . Equation (18.8) can also be used for the testing of normality disturbance by checking whether the coefficients of the last two terms are zero. A test of the presence of selection bias is to see whether the coefficients of all the (three) bias-correction terms are zero. In principle, it is possible to formulate the bias correction with more terms. However, with additional terms, one might quickly run into possible increasing multicollinearity in a regression framework. The proper selection of expanded terms is an issue on the selection of regressors or a model selection problem. One may think that it is sensible to incorporate more expanded terms as sample size increases. Such a strategy can be better justified in a semiparametric estimation framework (Newey, Powell, and Walker, 1990).

For the estimation of sample selection model, multicollinearity due to the addition of the bias-correction term in the observed outcome equation remains a serious point of contention. Olsen's specification with a linear probability choice equation highlights eloquently the multicollinearity issue at an early development stage of the literature. The linear choice probability corresponds to a uniform distribution for  $\epsilon$ . With  $\epsilon$  being a uniform random variable on  $[0, 1]$ ,  $E(\epsilon | z\gamma > \epsilon) = z\gamma/2$ . If  $z = x$  or  $z$  is a subvector of  $x$ , the bias-correction term will be perfectly multicollinear with the included  $x$  of the outcome equation. Consequently, the two-stage estimation method will break down completely. This multicollinearity issue is related to model identification. With normal disturbances, the bias-corrected term is a nonlinear function of  $z$  and, because of the nonlinearity, the two-stage method would not completely break down. However, severe multicollinearity might still be an issue for some samples. Nawata (1993) and Leung and Yu (1996) showed that  $\phi(z\gamma)/\Phi(z\gamma)$  is almost linear in  $z\gamma$  on a range of approximately  $[-3, 3]$ . The two-stage estimator would not be reliable under multicollinearity. Wales and Woodland (1980), Nelson (1984), Manning, Duan, and Rogers (1987), Nawata and Nagase (1996), and Leung and Yu (1996), among others, investigate this issue by several Monte Carlo studies. The overall conclusions from these Monte Carlo studies are that the effectiveness of the two-stage estimator depends on either exclusion restrictions whereby some relevant variables in  $z$  do not appear in  $x$  or on the fact that at least one of the relevant variables in  $z$  displays sufficient variation to make the nonlinearity effective. In a distribution-free sample selection model, the exclusion restriction is a necessary condition for identification as it needs to rule out the linear probability setting of Olsen.

The possible multicollinearity of the two-stage estimation procedure created a debate in the health economics literature on whether a sample selection model is a better model than multi-part models for modeling discrete choices with outcome equations. A multi-part model essentially assumes uncorrelated disturbances among outcomes and choice equations (Manning *et al.*, 1987; Maddala, 1985; Hay and Olsen, 1984; Leung and Yu, 1966). Hay and Olsen (1984) and Maddala (1985) point out an important feature of the sample selection model is its usage to investigate potential outcomes in addition to observed outcomes. For the normal distribution model, in the presence of severe multicollinearity, one has to resort to the method of maximum likelihood for a better inference.

### 3.2 Maximum likelihood estimation

Under the assumption that  $(\epsilon_i, u_i)$ ,  $i = 1, \dots, n$ , are iid normally distributed with zero means, a unit variance for  $\epsilon$ , a variance  $\sigma^2$  for  $u$ , and a correlation coefficient  $\rho_{12}$ , the loglikelihood function for the model (18.1) is

$$\begin{aligned} \ln L(\theta) = \sum_{i=1}^n & \left\{ (1 - I_i) \ln(1 - \Phi(z_i\gamma)) - \frac{1}{2} I_i \ln(2\pi\sigma^2) - \frac{1}{2\sigma^2} I_i (y_i - x_i\beta)^2 \right. \\ & \left. + I_i \ln \Phi \left[ (z_i\gamma - \frac{\rho}{\sigma}(y_i - x_i\beta)) / \sqrt{1 - \rho^2} \right] \right\} \end{aligned} \quad (18.9)$$

where  $\theta = (\beta', \sigma^2, \gamma', \rho)'$ . The use of the maximum likelihood method can be found, for example, in Heckman (1974), Nelson (1977), and Griliches, Hall, and Hausman (1978). The logarithmic likelihood function (18.9) may not have a unique maximum (Nelson, 1977; Olsen, 1982). However, Olsen (1982) showed that, for a given value of  $\rho$ , it has a unique maximum for the remaining parameter. Despite the latter property, the implementation of the maximum likelihood method of this model remains an overlooked issue. Nawata and Nagase (1996) pointed out that the maximum likelihood method implemented in several popular econometric packages by standard optimization methods may be incomplete. They recommended the use of scanning methods.

Under some regularity conditions, the maximum likelihood estimate (MLE) is, in general, consistent, asymptotically normal and efficient (Amemiya, 1973). But in some special cases, the loglikelihood function has irregularities at some parameter subspace. At  $\rho = 0$ , if  $x$  and  $\phi(z\gamma)/\Phi(z\gamma)$  are linearly dependent, the information matrix will be singular. This follows because  $\frac{\partial \ln L}{\partial \beta} = \frac{1}{\sigma^2} \sum_{i=1}^n I_i x'_i u_i$  and  $\frac{\partial \ln L}{\partial \rho} = -\frac{1}{\sigma} \sum_{i=1}^n \frac{\phi(z_i\gamma)}{\Phi(z_i\gamma)} I_i u_i$  at  $\rho = 0$ . Singular information matrix in this sample selection model does not indicate that the parameter vector of interest is not identifiable. Lee and Chesher (1986) investigated the implications on the MLE under such an irregularity for a concrete case that  $x$  contains an intercept term and  $z$  contains only a constant term, namely,  $y = x\beta + u$  and  $y^* = \gamma - \varepsilon$ , under the normal assumption, when the true but unknown  $\rho$  is zero. The asymptotic distribution can be analyzed by the Taylor series expansion of a concentrated likelihood function. It turns out that for some components of the parameter vector their MLEs are still consistent at the usual  $1/\sqrt{n}$ -rate of convergence and their distributions are asymptotically normal, but the MLEs of remaining parameters converge at much lower rates of convergence and are nonnormally distributed. The MLE  $\hat{\rho}$  has the asymptotic property that  $\sqrt{n} \hat{\rho}^3$  is asymptotically normal with zero mean and a finite variance. Hence,  $\hat{\rho}$  converges at the rate of  $n^{-1/6}$  and  $n^{1/6}\hat{\rho}$  is asymptotically distributed as a cubic root of a normal variable. For the remaining estimates, let  $\beta = (\beta_1, \beta_2)$  where  $\beta_1$  is the unknown intercept and  $\beta_2$  is the slope vector of regressors. The MLEs of  $\beta_2$  and  $\gamma$  turn out to be  $\sqrt{n}$ -consistent and  $\sqrt{n}(\hat{\beta}_2 - \beta_2)$  and  $\sqrt{n}(\hat{\gamma} - \gamma)$  are asymptotically normal. However,  $\hat{\beta}_1$  converges at the low  $n^{-1/6}$ -rate and  $n^{1/6}(\hat{\beta}_1 - \beta_1)$  is asymptotically a cubic root of a normal variable. For  $\sigma^2$ , its MLE  $\hat{\sigma}^2$  converges at an  $n^{-1/3}$ -rate and  $n^{1/3}(\hat{\sigma}^2 - \sigma^2)$  is asymptotically the  $2/3$  power of a normal variable.

### 3.3 Polychotomous choice sample selection models

A polychotomous choice model with  $m$ -alternatives specifies the latent utility values  $U_j = z_j\gamma + v_j$ ,  $j = 1, \dots, m$ , and the alternative  $j$  is chosen if and only if  $U_j > \max\{U_k : k = 1, \dots, m; k \neq j\}$ . A specified joint distribution for  $v = (v_1, \dots, v_m)$  implies the choice probability  $G_j$  for the  $j$ th alternative, where  $G_j(x\gamma) = P(z_j\gamma - z_k\gamma > v_k - v_j, k \neq j, k = 1, \dots, m | x)$ . Familiar parametric polychotomous choice models are the conditional logit model and the nested logit model of McFadden (1973, 1978) and the multinomial probit model. For a recent survey on qualitative

response models, see Chapter 17 by Maddala and Flores-Lagunes in this volume. For a sample selection model, say,  $y = x\beta + u$  for the alternative 1, a desirable specification will allow  $u$  to correlate with the disturbances in the utility equations. A traditional approach may specify a joint distribution for  $v = (v_1, \dots, v_m)$  and  $u$ . In such an approach, the marginal distributions for  $v$  and  $u$  are determined by the joint distribution. If they are jointly normally distributed, the implied choice model will be the multinomial probit model. However, it is less obvious how to incorporate selected outcome equations with other familiar choice models. Dubin and McFadden (1984) and Lee (1983) suggest alternative specification approaches. Dubin and McFadden (1984) suggest a linear conditional expectation specification in that  $E(u|v, x)$  is a linear function of  $v$ . The distribution of  $v$  can be the one which generates the logit or nested logit choice component. They suggest two-stage estimation methods based on bias-corrected outcome equations. Lee (1983) suggests an approach based on order statistics and distributional transformations. The marginal distributions of  $v$  and  $u$  are first specified, and the model is then completed with a distribution with specified margins. From the choice equations, define a random variable  $\varepsilon_1$  as  $\varepsilon_1 = \max\{y_k^* : k = 2, \dots, m\} - v_1$ . The first alternative is chosen if and only if  $z_1\gamma > \varepsilon_1$ . In terms of  $\varepsilon_1$ , the choice inequality looks like a binary choice criterion for alternative 1. Given the distributions for  $v$  (and hence  $G_1$ ), the implied distribution of  $\varepsilon_1$  is  $F_1(c|x) = G_1(c - z_2\gamma_2, \dots, c - z_m\gamma_m)$ . When  $u_1$  is normally distributed, a normal-distribution transformation is suggested to transform  $\varepsilon_1$  to a standard normal variable  $\varepsilon_1^*$  as  $\varepsilon_1^* = \Phi^{-1}(F_1(\varepsilon_1|x))$ . The  $u$  and  $\varepsilon_1^*$  are then assumed to be jointly normally distributed with zero means and a covariance  $\sigma_{u_1\varepsilon_1^*}$ . Under such a specification, the bias-corrected outcome equation is similar to the familiar one as  $E(y_1|x, I_1 = 1) = x_1\beta_1 - \sigma_{u_1\varepsilon_1^*} \frac{\phi(\Phi^{-1}(G_1(z\gamma)))}{G_1(z\gamma)}$ . If  $u$  were not normally distributed, other transformations rather than the normal distribution might be desirable. If the marginal distribution of  $u_1$  were unknown, flexible functional specifications such as the bivariate Edgeworth expansion might be used. Under this approach, both a simple two-stage method and the method of maximum likelihood can be used. Schmertmann (1994) compares the advantages and disadvantages of the McFadden and Dubin, and Lee approaches. His conclusion is that the McFadden and Dubin specification can likely be affected by multicollinearity as several bias-correction terms may be introduced; and Lee's specification imposes restrictive covariance structures on  $u$  and  $v$  and may be sensitive to misspecification. On the latter, Lee (1995) derives some implications on comparative advantage measures.

### 3.4 Simulation estimation

The multinomial probit model has long been recognized as a useful discrete choice model. But, because its choice probability does not have a closed-form expression and its computation involves multiple integrals, it has not been a popular model for empirical studies until the 1990s. The advancement of computing technology and the recent development on simulation estimation techniques provide

effective ways to implement this model. For a general description on various simulation methods, see Chapter 22 by Geweke, Houser, and Keane in this volume. Simulation estimation methods can be developed for the estimation of the multinomial probit sample selection model (Lee, 1996). The model can be estimated by two-stage methods and/or the method of maximum likelihood via Monte Carlo simulation. A simulated two-stage method is similar to the method of simulated moments (McFadden, 1989). For a two-stage method, the choice equations are first estimated. The likelihood function of the choice model can be simulated with the GHK (Geweke–Hajivassiliou–Keane) simulator. The probabilities of  $U_j - U_l > 0$ ,  $l = 1, \dots, m$  but  $l \neq j$  are determined by the normal distribution of  $\varepsilon_j = (v_1 - v_{j'}, \dots, v_{j-1} - v_{j'}, v_{j+1} - v_{j'}, \dots, v_m - v_{j'})$ . Denote  $w_j = (z_j - z_1, \dots, z_j - z_{j-1}, z_j - z_{j+1}, \dots, z_j - z_m)$ . The  $\varepsilon_j$  can be represented as  $\varepsilon_j = H_j \eta_j$ , where  $H_j$  is a lower triangular matrix of the Cholesky decomposition of the variance of  $\varepsilon_j$  and  $\eta_j = (\eta_{j1}, \dots, \eta_{jm-1})$  is a standard normal vector. Define  $L_{j1} = w_{j1}\gamma/h_{j11}$  and  $L_{jl} = [w_{jl}\gamma - \sum_{k=1}^{l-1} h_{jlk}\eta_k]/h_{jl}$  for  $l = 2, \dots, m-1$ . It follows that

$$P(I_j = 1) = \int_{-\infty}^{L_{j,m-1}} \dots \int_{-\infty}^{L_{j1}} \prod_{l=1}^{m-1} \phi(\eta_{jl}) d\eta_{jl} = \int_{-\infty}^{\infty} \dots \int_{-\infty}^{\infty} \prod_{l=1}^{m-1} \Phi(L_{jl}) \phi_{(-\infty, L_{jl})}(\eta_{jl}) d\eta_{jl},$$

where  $\phi_{(a,b)}$  is a truncated standard normal density with support on  $(a, b)$ . The GHK sampler is to generate sequentially truncated standard normal variables from  $\prod_{l=1}^{m-1} \phi_{(-\infty, L_{jl})}(\eta_{jl})$ . With  $S$  simulation runs, the GHK likelihood simulator of the choice model is  $\hat{L}_S(\bar{I}) = \prod_{j=1}^m \left\{ \frac{1}{S} \sum_{s=1}^S \prod_{l=1}^{m-1} \Phi(L_{jl}^{(s)}) \right\}^{I_j}$ . The parameter  $\gamma$  can be estimated by maximizing this simulated likelihood function. With the first-stage estimate  $\hat{\gamma}$ , the outcome equation can be estimated by the method of simulated moments. Simulated moment equations can be derived from observed outcome equations. Consider the bias-corrected outcome equation  $y = x\beta + E(u|x, I_1 = 1) + \eta$ . As  $\varepsilon_1$  has the Cholesky decomposition  $\varepsilon_1 = H_1 \eta_1$ ,

$$E(u|x, I_1 = 1) = \frac{1}{P(I_1 = 1)} \int_{-\infty}^{\infty} \dots \int_{-\infty}^{\infty} E(u|\varepsilon_1 = H_1 \eta_1) \prod_{l=1}^{m-1} \Phi(L_{1l}) \phi_{(-\infty, L_{1l})}(\eta_{1l}) d\eta_{1l}. \quad (18.10)$$

The GHK sampler  $\prod_{l=1}^{m-1} \phi_{(-\infty, L_{1l})}(\eta_{1l})$  can be used simultaneously to simulate both the numerator and denominator in (18.10). With  $S$  simulation runs from the GHK sampler, (18.10) can be simulated as  $\hat{E}_S(u|I_1 = 1) = \sum_{s=1}^S E(u|\varepsilon_1 = H_1 \eta_1^{(s)}) \omega^{(s)}$ , where  $\omega^{(s)} = \prod_{l=1}^{m-1} \Phi(L_{1l}^{(s)}) / \sum_{r=1}^S \prod_{l=1}^{m-1} \Phi(L_{1l}^{(r)})$ .  $\hat{E}_S(u|I_1 = 1)$  is a consistent estimator of  $E(u|I_1 = 1)$  when  $S$  goes to infinity. The simulated method of moments can then be applied to estimate the equation  $y = x\beta + \hat{E}_S(u|I_1 = 1) + \tilde{\eta}$ . The method of simulated moments is a generalized method of (simulated) moments. It is essentially an instrumental variable (IV) method. It is not desirable to apply a least squares procedure to the bias-corrected equation because  $\hat{E}_S(u|I_1 = 1)$  with a finite  $S$  creates an error-in-variable (on regressors) problem.

The method of simulated maximum likelihood can be asymptotically efficient if the number of random draws  $S$  increases at a rate faster than the  $\sqrt{n}$ -rate

(Lee, 1992b). The likelihood function for an observation of the model with  $m$ -alternatives and with outcome for alternative 1 is

$$L(\bar{I}, y) = \left[ \int_{-\infty}^{L_{1,m-1}} \cdots \int_{-\infty}^{L_{11}} f(u | \varepsilon_1 = H_1 \eta_1) \prod_{l=1}^{m-1} \phi(\eta_{ll}) d\eta_{ll} \right]^{I_1} \prod_{j=2}^m P(I_j = 1)^{I_j},$$

where  $u = y - x\beta$ . Each of the likelihood components can be simulated without bias with a generalization of the GHK simulator. The likelihood function can be simulated as

$$\tilde{L}_S(\bar{I}, y) = \left\{ \frac{1}{S} \sum_{s=1}^S f(u | \varepsilon_1 = H_1 \eta_1^{(s)}) \prod_{l=1}^{m-1} \Phi(L_{1l}^{(s)}) \right\}^{I_1} \prod_{j=2}^m \left\{ \frac{1}{S} \sum_{s=1}^S \prod_{l=1}^{m-1} \Phi(L_{jl}^{(s)}) \right\}^{I_j},$$

where the random variables are drawn from  $\prod_{l=1}^{m-1} \phi_{(-\infty, L_{jl})}(\eta_{jl})$  for each  $j = 1, \dots, m$ . Lee (1996) compares these simulation methods and finds that the simulated maximum likelihood method can indeed be more efficient with a moderate amount of simulation draws. One may expect that simulation methods may play an important role in the future development of sample selection models with dynamic structures.

### 3.5 Estimation of simultaneous equation sample selection model

A two-stage estimation method can be easily generalized for the estimation of a simultaneous equation model. Consider the linear simultaneous equation  $y^* = y^*B + xC + u$ , which can be observed only if  $z\gamma > \varepsilon$ . For the estimation of structural parameters, consider the first structural equation  $y_1^* = y_{(1)}^* \beta_1 + x_1 \delta_1 + u_1$  where  $y_{(1)}^*$  consists of included endogenous variables on the right-hand side of the structural equation. The bias-corrected structural equation is  $y_1 = y_{(1)} \beta_1 + x_1 \delta_1 + \sigma_{1e} \left( -\frac{\phi(z\hat{\gamma})}{\Phi(z\hat{\gamma})} \right) + \tilde{\eta}_1$ . The system implies the reduced form equations  $y^* = x\Pi + v$ . Lee, Maddala, and Trost (1980) suggest the estimation of the reduced form parameters  $\Pi$  by the Heckman two-stage method, and used the predicted  $y_{(1)}$  to estimate the bias-corrected structural equation similar to Theil's two-stage method for a conventional simultaneous equation model.

The structural parameters can also be estimated by a general minimum distance procedure (Amemiya, 1979). Amemiya's method is a systematic procedure for estimating structural parameters directly from estimated reduced form parameters. Let  $J_1$  and  $J_2$  be the selection matrices such that  $y_{(1)}^* = y^*J_1$  and  $x_1 = xJ_2$ . As  $y_1^* = y^*J_1\beta_1 + xJ_2\delta_1 + u_1 = x(\Pi J_1\beta_1 + J_2\delta_1) + v_1$ , one has  $\pi_1 = \Pi J_1\beta_1 + J_2\delta_1$ . Let  $\hat{\Pi}$  be the reduced form estimate from Heckman's two-stage estimation. Amemiya's minimum distance procedure is to estimate  $\beta_1$  and  $\delta_1$  from the linear equation  $\hat{\pi}_1 = \hat{\Pi}J_1\beta_1 + J_2\delta_1 + \zeta_1$ , where  $\zeta_1 = (\hat{\pi}_1 - \pi_1) - (\hat{\Pi} - \Pi)J_1\beta_1$  is the disturbance, by least squares or generalized least squares. The relative efficiency of an estimator

from this minimum distance approach depends on the relative efficiency of the reduced form parameter estimates (Amemiya, 1983). Lee (1981), Amemiya (1983) and Newey (1987) compare various two-stage IV estimation methods with Amemiya's generalized minimum distance estimators. It was found that many two-stage IV estimators are special cases of Amemiya's minimum distance estimators depending on appropriate reduced form estimates. Lee (1992a) shows that the minimized generalized sum of squares residuals from Amemiya's generalized least-squares procedure also provides a test of overidentification restrictions of a linear structural equation. However, because Amemiya's approach relies on solving structural parameters from reduced form parameters, it cannot be generalized to the estimation of a nonlinear simultaneous equation system while many IV approaches can.

### 3.6 Misspecification and tests

In the sample selection model (18.1), the data on observed outcomes  $y$  are censored and least squares estimators of  $\beta$  obtained using  $y$  suffer from selectivity bias when the disturbances  $u$  and  $\varepsilon$  are correlated. But if they were independent, the model would have a simple structure and  $\beta$  and  $\sigma^2$  could be estimated by applying ordinary least squares to the observed outcome equation and  $\gamma$  of the selection equation can be estimated by the probit MLE. To test whether there is selectivity for sample under normal disturbances, one may examine the hypothesis  $H_0 : \rho = 0$ . The score test statistic for this hypothesis is derived in Melino (1982). The score test statistic has the same asymptotic distribution as a  $t$ -statistic for testing the significance of the coefficient of the bias-corrected term in a two-stage estimation procedure. The simple  $t$ -statistic is an asymptotically efficient test statistic.

For special cases where the bias-corrected term is perfectly collinear with included regressors, these test statistics break down. The score vector evaluated at the restricted MLE under such a circumstance is identically zero and the corresponding information matrix is singular. Lee and Chesher (1986) suggest a generalization of the score test to extremum tests. The intuition behind the score test statistic is to ask whether the average loglikelihood of a model evaluated at the restricted MLE has a significantly nonzero gradient. When it does, we are led to reject the null hypothesis because by moving from the parameter vector under the null hypothesis, higher values of the loglikelihood can be achieved. The score test is exploiting the first-order derivative testing for a maximum. Thus, when the score test statistic is identically zero, one should consider tests for extremum based on higher-order derivatives as in calculus. For testing the presence of sample selection bias, the extremum test statistic is asymptotically equivalent to a test of skewness constructed from the statistic  $\frac{1}{N} \sum_{i=1}^N I_i e_i^3 / (3\hat{\sigma}^3)$  where  $e_i$  is the least squares residual of the outcome equation. This statistic is intuitively appealing because, if  $\rho \neq 0$ , the disturbance of the observed outcome equation has nonzero mean and its distribution (conditional on selection) is not symmetric.

The normal distribution is a popular assumption in parametric sample selection models. Contrary to the standard linear regression model, the misspecification

of normality of disturbances will, in general, provide inconsistent estimates under the two-stage or maximum likelihood methods. Theoretical consequences of misspecification are well presented for limited dependent variables models (Goldberger, 1983). For the sample selection model, investigations of sensitivity of distributional misspecification can be found in Olsen (1982), Lee (1982), Mroz (1987), and others. Diagnostic test statistics have been developed for detecting model misspecification. The computationally simple and motivating approach is the Lagrange multiplier (efficient score) approach. For various misspecifications such as omitted variables, heteroskedasticity, serial correlation, and normal disturbances, the Lagrange multiplier statistics have simple moment structures (see the survey by Pagan and Vella, 1989). This can be seen as follows. In an econometric model with latent variables, suppose that  $g(y^*, y | \theta)$  is the joint density of latent variables  $y^*$  and observed sample  $y$ , where  $\theta$  is a vector of possible parameters. Let  $f(y | \theta)$  be the density of  $y$ , and  $g(y^* | y, \theta)$  be the conditional density of  $y^*$  given  $y$ . Since  $g(y^*, y | \theta) = g(y^* | y, \theta)f(y | \theta)$ ,  $\ln f(y | \theta) = \ln g(y^*, y | \theta) - \ln g(y^* | y, \theta)$ , and  $\frac{\partial \ln f(y | \theta)}{\partial \theta} = \frac{\partial \ln g(y^*, y | \theta)}{\partial \theta} - \frac{\partial \ln g(y^* | y, \theta)}{\partial \theta}$ . Integrating these expressions with respect to  $g(y^* | y, \theta^+)$  where  $\theta^+$  is an arbitrary value of  $\theta$ , it follows that  $\ln f(y | \theta) = \int_{-\infty}^{\infty} [\ln g(y^*, y | \theta)]g(y^* | y, \theta^+)dy^* - \int_{-\infty}^{\infty} [\ln g(y^* | y, \theta)]g(y^* | y, \theta^+)dy^*$  and

$$\frac{\partial \ln f(y | \theta)}{\partial \theta} = \int_{-\infty}^{\infty} \frac{\partial \ln g(y^*, y | \theta)}{\partial \theta} g(y^* | y, \theta^+)dy^* - \int_{-\infty}^{\infty} \frac{\partial \ln g(y^* | y, \theta)}{\partial \theta} g(y^* | y, \theta^+)dy^*. \quad (18.11)$$

At  $\theta^+ = \theta$ , the first-order derivative of the loglikelihood (18.11) becomes  $\frac{\partial \ln f(y | \theta)}{\partial \theta} = E_{\theta}(\frac{\partial \ln g(y^*, y | \theta)}{\partial \theta} | y)$ , because  $E_{\theta}(\frac{\partial \ln g(y^* | y, \theta)}{\partial \theta} | y) = 0$ . A test statistic based on the score  $\frac{\partial \ln f(y | \theta)}{\partial \theta}$  will be a conditional moment statistic of  $\frac{\partial \ln g(y^*, y | \theta)}{\partial \theta}$  conditional on sample observations. For many specification tests of the sample selection model, the score test statistics are based on some simple conditional moments. Lee (1984) considered the efficient score test of normality of the sample selection model (18.1). The test statistic is derived within the bivariate Edgeworth series of distributions. For the truncated sample selection case, the test compares some sample moments of order  $(r, s)$  for which  $r + s > 2$  with correspondingly estimated hypothesized conditional moments of disturbances. For the censored case, the test is equivalent to the testing of some sample semi-invariants for which  $r + s > 2$  are zeros.

## 4 SEMIPARAMETRIC AND NONPARAMETRIC APPROACHES

### 4.1 Semiparametric two-stage estimation

Manski (1975) showed that a parametric distribution is not necessary for consistent estimation of discrete choice models, and thus originated the semiparametric estimation literature in microeconomics. Recognition of inconsistency of the maximum likelihood and two-stage estimation methods under a misspecified

error distribution has speeded up the studies on semiparametric and nonparametric estimation methods. Cosslett (1991) initiated semiparametric estimation of the sample selection model with a binary selection equation. Based on a series approximation to an unknown density function, Gallant and Nychka (1987) have proposed a consistent semi-nonparametric maximum likelihood method. The asymptotic distributions of both the estimators of Cosslett and Gallant and Nychka are unknown. Robinson (1988) obtained a semiparametric estimator of parameters of the outcome equation and derived its asymptotic distribution. His method is based on a nonparametric kernel regression correction of the selection-bias term. Subsequent contributions have largely concentrated on index formulations for dimension reduction. Powell (1987) considered a single index case and Ichimura and Lee (1991) studied the general multiple index situation. Ahn and Powell (1993) considered a probability index formulation. The approaches in Robinson, Powell, and Ahn and Powell are single equation two-stage estimation methods with nonparametric kernel regression functions. The approach in Ichimura and Lee (1991) is a semiparametric nonlinear least squares method, which can also be used for truncated sample on outcome equations. Others (Newey, 1988; Andrews, 1991) have used series approximations for conditional expectations.

These two-stage estimation methods are motivated by the implied bias-corrected outcome equation having the form

$$y_i = x_i\beta + \psi(z_i\gamma) + \eta_i, \quad (18.12)$$

where  $E(\eta_i | I_i = 1, x_i) = 0$  for cross-sectional data. For semiparametric models,  $\psi$  is an unknown function but can be estimated by some nonparametric estimators. The various semiparametric two-stage methods differ from each other on how  $\psi$  has been estimated. As  $\psi$  may be a linear function, it is essential that there is at least one variable which is in  $z$  but not in  $x$  for the identification of  $\beta$  in a semiparametric two-stage estimation procedure. If this exclusion condition does not hold for (18.12), some linear transformation of  $\beta$ , which creates the exclusion requirement, can still be identified and estimated (Chamberlain, 1986; Lee, 1994b). With the unknown  $\psi$  replaced by a nonparametric function  $E_n(z_i, \theta)$  where  $\theta = (\beta, \gamma)$ , various suggested approaches amount to estimate unknown parameters of the equation  $y_i = x_i\beta + E_n(z_i, \theta) + \tilde{\eta}_i$ . In the index context, one only needs to know that  $\psi(z\gamma) = E(y - x\beta | z\gamma)$  is a function of  $z\gamma$ . For a kernel type estimator,  $\psi(z_i\gamma)$  can

be estimated by a nonparametric regression estimator,  $E_n(z_i, \theta) = \frac{\sum_{j \neq i}^n (y_j - x_j\beta)K(\frac{z_j\gamma - z_i\gamma}{a_n})}{\sum_{j \neq i}^n K(\frac{z_j\gamma - z_i\gamma}{a_n})}$ ,

where  $K$  is a kernel function and  $a_n$  is a bandwidth or window width. The consistency and asymptotic distribution of a derived estimator of  $\beta$  (and/or  $\gamma$ ) depend on certain essential conditions on a selected sequence of bandwidths  $\{a_n\}$ . The bandwidth sequence is required to converge to zero as  $n$  goes to infinity, but its rate of convergence cannot be too fast. The rate of convergence of a nonparametric regression will, in general, depend on the degree of smoothness of underlying densities of disturbances and regressors of the model. For series approximations, the corresponding problem refers to the number of terms included

in an approximation. For general issues on nonparametric regression, see Ullah, Chapter 20 in this volume.

The use of the kernel-type regression function (or series approximation) is valuable in that asymptotic properties of estimators can be established. The  $\beta$  can be consistently estimated and the semiparametric estimators are  $\sqrt{n}$ -consistent and asymptotic normal. For empirical applications, one has to be careful on selecting appropriate bandwidths. The bandwidth selection can be a complicated issue. In practice, one may hope that a bandwidth parameter can be automatically determined. The Cosslett two-stage approach has the latter feature. The implicit window widths in his approach are automatically determined. At the first stage, a semiparametric maximum likelihood procedure is used to estimate  $\gamma$  and the distribution  $F$  of choice equation disturbance  $\varepsilon$  under the assumption that  $\varepsilon$  and  $u$  are independent of all regressors. The estimator of  $F$  for each  $\gamma$  is  $\hat{F}(\cdot | \gamma)$  derived by maximizing the loglikelihood function  $\ln L(F, \gamma) = \sum_{i=1}^n [I_i \ln F(z_i \gamma) + (1 - I_i) \ln (1 - F(z_i \gamma))]$  with respect to  $F$ . The estimator of  $\gamma$  is then derived by maximizing  $\ln L(\hat{F}(\cdot | \gamma), \gamma)$  with respect to  $\gamma$ . The estimator  $\hat{F}$  is also used for the estimation of  $\beta$  in the second stage. The estimator  $\hat{F}$  is a step function with steps located at some  $\varepsilon_j^*$ ,  $j = 1, \dots, J$ , where  $\varepsilon_1^* < \varepsilon_2^* < \dots < \varepsilon_J^*$ . The number of steps, their locations and their heights are all determined in the first-stage estimation. The implicit bandwidths  $\varepsilon_j^* - \varepsilon_{j-1}^*$  for  $j = 1, \dots, J$  with  $\varepsilon_0^* = -\infty$  as a convention, are automatic. Under the assumption that  $u$  and  $\varepsilon$  are independent of  $x$  and  $z$ ,  $\psi(z|\hat{\gamma})$  in (18.12) is  $\psi(z|\hat{\gamma}) = \int_{-\infty}^{z|\hat{\gamma}} E(u|\varepsilon)dF(\varepsilon)/\int_{-\infty}^{z|\hat{\gamma}} dF(\varepsilon)$ . With  $\hat{F}$  replacing  $F$ , the estimated  $\hat{\psi}(z|\hat{\gamma})$  is a constant  $\lambda_j$  for all  $z|\hat{\gamma}$  in the interval  $(\varepsilon_{j-1}^*, \varepsilon_j^*)$ , where  $\lambda_j = \int_{-\infty}^{\varepsilon_{j-1}^*} E(u|\varepsilon)d\hat{F}(\varepsilon)/\int_{-\infty}^{\varepsilon_{j-1}^*} d\hat{F}(\varepsilon)$ . Define the subset of sample observations  $S_j = \{i \mid \varepsilon_{j-1}^* < z_i|\hat{\gamma} < \varepsilon_j^* \text{ and } I_i = 1\}$  and the set indicator  $I_{S_j}$ . Cosslett's approach leads to the estimation of the regression equation with added dummy regressors:  $y_i = x_i\beta + \sum_{j=1}^J \lambda_j I_{S_j}(i) + \eta_i$ . Cosslett (1991) showed that the estimator is consistent. However, its asymptotic distribution remains unknown. The automatic window width in the approach may have induced complications for asymptotic analysis.

## 4.2 Semiparametric efficiency bound and semiparametric MLE

On asymptotic efficiency for semiparametric estimation, Chamberlain (1986) derived an asymptotically lower bound for variances of  $\sqrt{n}$ -consistent regular semiparametric estimators for the sample selection model (18.1). The notion of asymptotic efficiency for parameter estimators in a semiparametric model was originated in Stein (1956). The semiparametric asymptotic variance bound  $V$  for a semiparametric model is defined as the supremum of the Cramer–Rao bounds for all regular parametric submodels. The intuition behind this efficient criterion is that a parametric MLE for any parametric submodel should be at least as efficient as any semiparametric estimator. The class of estimators is restricted to regular estimators so as to exclude superefficient estimators and estimators using information that is not contained in a semiparametric model. Newey (1990)

provides a useful characterization for a semiparametric estimator to be regular. A way to derive the semiparametric bound is to derive a relevant tangent set and its efficient score  $S$ . Let  $\delta = (\theta', \eta')'$  be the parameters of a submodel and let  $S_\delta = (S'_\theta, S'_\eta)'$  be the score vector. A tangent set  $I$  is defined as the mean-squares closure of all linear combinations of scores  $S_\eta$  for parameter models. The efficient score  $S$  is the unique vector such that  $S_\theta - S \in I$  and  $E(S't) = 0$ , for all  $t \in I$ . The semiparametric variance bound is  $V = (E[SS'])^{-1}$ .

The sample selection model considered in Chamberlain (1986) is (18.1) under the assumption that  $\varepsilon$  and  $u$  are independent of all the regressors in the model. With a parametric submodel, the scores for  $(\beta, \gamma)$  of a single observation have the form  $\frac{\partial \ln L(\beta, \gamma)}{\partial \beta} = a_1(y - x\beta, I, z\gamma)x$  and  $\frac{\partial \ln L(\beta, \gamma)}{\partial \gamma} = a_2(y - x\beta, I, z\gamma)z$ , for some functions  $a_1$  and  $a_2$ . Chamberlain (1986) showed that the effective scores of the semiparametric sample selection model are simply the preceding scores with the factors  $x$  and  $z$  replaced, respectively, by  $x - E(x|z\gamma)$  and  $z - E(z|z\gamma)$ , i.e.,

$$\begin{aligned}\frac{\partial \ln L(\beta, \gamma)}{\partial \beta} &= a_1(y - x\beta, I, z\gamma)(x - E(x|z\gamma)), \quad \text{and} \\ \frac{\partial \ln L(\beta, \gamma)}{\partial \gamma} &= a_2(y - x\beta, I, z\gamma)(z - E(z|z\gamma)).\end{aligned}\tag{18.13}$$

The information matrix of the semiparametric model is formed from these effective scores. The expressions in (18.13) provide insight into qualitative differences between parametric and semiparametric models. The information matrix of the semiparametric model is singular if there is no restriction on  $\gamma$  of the choice equation. This is so because  $(z - E(z|z\gamma))\gamma = 0$ . Furthermore, if  $z$  consists of all discrete variables, then  $E(z|z\gamma)$  generally equals  $z$ . So, in order that the information matrix is nonsingular, one component of  $z$  must have a continuously distributed variable. If  $z$  were a subvector of  $x$ , the effective scores would also be linearly dependent and the information matrix would be singular. So an exclusion restriction on  $\beta$  is also needed. Chamberlain (1986) pointed out that if the information matrix of the semiparametric model is singular, then there are parameters for which no (regular) consistent estimator can converge at rate  $n^{-1/2}$ .

Ai (1997) and Chen and Lee (1998) proposed semiparametric scoring estimation methods based on index restriction and kernel regression functions. The approaches involve the construction of efficient score functions. Ai's estimator was defined by setting an estimated sample score equal to zero and involved solving nonlinear equations. Chen and Lee's approach was a two-step efficient scoring method. Given an initial  $\sqrt{n}$ -consistent estimator, the latter estimator has a closed form. Both the estimator of Ai and that of Chen and Lee were shown to be asymptotically efficient for model (18.1) under the independence assumption. Chen and Lee (1998) also derived the efficient bound for the polychotomous choice model with index restrictions and pointed out that their estimator attained that bound. As an index model with  $L$  choice alternatives,  $P(I_l|x) = E(I_l|z\gamma)$  are functions of index  $z\gamma$  of the choice equations and the density function of  $y$

conditional on  $I_1 = 1$  and all exogenous variables in the model is the conditional density function of  $y - x\beta$  conditional on  $I_1 = 1$  and  $z\gamma$  at the true parameter vector  $(\beta_0, \gamma_0)$ , i.e.  $f(y | I_1 = 1, x, z) = f(y - x\beta_0 | I_1 = 1, z\gamma_0) = g(y - x\beta_0, x\gamma_0 | I_1 = 1)/p(x\gamma_0 | I_1 = 1)$ , where  $g(\epsilon, x\gamma_0 | I_1 = 1)$  is the conditional density of  $\epsilon$  and  $x\gamma_0$  conditional on  $I_1 = 1$ , and  $p(x\gamma_0 | I_1 = 1)$  is the conditional density of  $x\gamma_0$ . Given a random sample of size  $n$ , for any possible value  $\theta$ , the probability function  $E(I_l | z_i\gamma, \theta)$  of  $I_l$  conditional on  $z\gamma$  evaluated at point  $z_i\gamma$  can be estimated by  $P_{nl}(x_i, \theta) = A_n(I_l | x_i, \theta)/A_n(1 | x_i, \theta)$ , where  $A_n(v | x_i, \theta) = \frac{1}{n-1} \sum_{j \neq i}^n v_j \frac{1}{a_{1,n}^m} K(\frac{z_j\gamma - z_i\gamma}{a_{1,n}})$  for  $v = I_1, \dots, I_K$ , or  $1$ ,  $K(\cdot)$  is a kernel function on  $R^m$  when  $z\gamma$  is an  $m$ -dimensional vector of indices. On the other hand,  $f(y - x\beta | I_1 = 1, z\gamma, \theta)$  evaluated at point  $(y_i - x_i\beta, z_i\gamma)$  can be estimated by  $f_n(y_i - x_i\beta | I_{1i} = 1, z_i\gamma) = C_n(x_i, y_i, \theta)/A_n(I_{1i} | x_i, \theta)$ , where  $C_n(x_i, y_i, \theta) = \frac{1}{n-1} \sum_{j \neq i}^n I_{1j} \frac{1}{a_{2,n}^{m+k}} J(\frac{(y_j - x_j\beta) - (y_i - x_i\beta)}{a_{2,n}}, \frac{z_j\gamma - z_i\gamma}{a_{2,n}})$  when  $y - x\beta$  is a vector of dimension  $k$ . These nonparametric functions formulate a semiparametric loglikelihood function  $\ln L_n(\theta) = \frac{1}{n} \sum_{i=1}^n \{I_{1i} \ln f_n(y_i - x_i\beta | I_{1i} = 1, z_i\gamma) + \sum_{l=1}^L I_{li} \ln P_{nl}(x_i, \theta)\}$ . But, due to technical difficulties, this semiparametric likelihood function can hardly be used. Instead, one can work with its implied score function, i.e. the derivative of the semiparametric loglikelihood function with respect to  $\theta$ . With an initial  $\sqrt{n}$ -consistent estimate of  $\theta$ , the Chen–Lee estimator is a semiparametric two-step scoring estimator.

### 4.3 Semiparametric IV estimation and conditional moments restrictions

Simultaneous equation models with selectivity can also be estimated by semiparametric methods. Semiparametric IV methods for the estimation of sample selection models are considered in Powell (1987) and Lee (1994b). Powell (1987) has an interest in the asymptotic property of a general semiparametric IV estimator. Lee (1994b) follows the literature on classical linear simultaneous equation models by focusing on both the identification and estimation of a structural equation with sample selection. It considered possible generalizations of two-stage least squares methods and their possible optimum IV property. Consider a single linear structural equation

$$y_1^* = y_2^* \alpha + xJ\delta + u_1, \quad (18.14)$$

where  $y_1^*$  is a latent endogenous variable,  $y_2^*$  is a vector of latent endogenous variables not including  $y_1^*$ ,  $x$  is a vector of exogenous variables in the system, and  $xJ$ , where  $J$  is a selection matrix, represents the subset of exogenous variables included in this structural equation. The reduced form equation of  $y_2^*$  is  $y_2^* = x\Pi_2 + v_2$ . The sample observations  $y_1$  and  $y_2$  of  $y_1^*$  and  $y_2^*$  are subject to selection. The selection equation is  $I^* = x\gamma - \epsilon$ .  $y_1$  and  $y_2$  are observed if and only if  $I^* > 0$ . As in the index framework, the joint distribution of  $(u_1, v_2, \epsilon)$  conditional on  $x$  is assumed to be a function of the index  $x\gamma$ . In this system with sample selection, the identification of structural parameters requires stronger conditions than the usual

rank condition in the classical linear simultaneous equation model (without selectivity) and the parametric linear simultaneous equation sample selection model considered in Lee *et al.* (1980) and Amemiya (1983). Let  $y_1^* = x\pi_1 + v_1$  be the implied reduced form equation for  $y_1^*$ . Conditional on  $x$  and  $I = 1$ ,  $E(y_1 | x, I = 1) = x\pi_1 + E(v_1 | x\gamma, x\gamma > \varepsilon)$  and  $E(y_2 | x, I = 1) = x\Pi_2 + E(v_2 | x\gamma, x\gamma > \varepsilon)$ . As in the classical linear simultaneous equation model, the identification of structural parameters is directly related to the reduced form parameters. However, contrary to the classical system, the reduced form parameter vectors  $\pi_1$  and  $\Pi_2$  are not identifiable because the same  $x$  appears in the selection equation. It turns out that some linear combinations of the reduced form parameters can be identified. As the selection equation is a single index model, a conventional normalization suggested by Ichimura (1993) is to set the coefficient of a continuous and exogenous variable to be unity, i.e.  $x\gamma = x_{(1)} + x_{(2)}\zeta$ , where  $x_{(1)}$  is a relevant continuous and exogenous variable in  $x = (x_{(1)}, x_{(2)})$ . With the partition of  $x$  into  $x_{(1)}$  and  $x_{(2)}$ , the above equations imply that the parameters  $\pi_{(1)}^*$  and  $\Pi_2^*$  in  $E(y_1 | x, I = 1) = x_{(2)}\pi_{(1)}^* + E(v_1^* | x\gamma, I = 1)$  and  $E(y_2 | x, I = 1) = x_{(2)}\Pi_2^* + E(v_2^* | x\gamma, I = 1)$  are identifiable. The structural parameters are related to those identified reduced-form parameters as

$$\pi_1^* = \Pi_2^*\alpha - \delta_1\zeta + \delta_2. \quad (18.15)$$

The identification of the structural parameters follows from this relation. With exclusion restrictions as in (18.14), one can see that the order identification condition for the semiparametric model corresponds to the overidentification condition of the classical linear simultaneous equation model. The stronger condition for the identification of the semiparametric model is due to the addition of a selection bias term with an unknown form in the bias-corrected structural equation. Exogenous variables excluded from the structural equation (18.14) before bias correction reappear in the selection bias term through the index  $x\gamma$ . Such exogenous variables identify the bias term. But, the bias-correction term introduces excluded exogenous variables back into the bias-corrected structural equation. It follows that the effective number of the included exogenous variables in this equation is the number of originally included exogenous variables plus one. Therefore, the order condition for identification requires stronger exclusion restrictions than the classical model or a parametric model. For a parametric model under normal disturbances, the bias-correction term has a known nonlinear form which can help identification.

The structural parameters  $\alpha$  and  $\delta$  can be estimated via (18.15) by Amemiya's minimum distance methods. But semiparametric least squares are relatively simple and illustrative. For the semiparametric estimation of the structural equation (18.14), let  $w = (y_2, x)$  and  $\beta = (\alpha, \delta)$ . For any possible value  $(\beta, \gamma)$ ,  $E(y_1 - w\beta | x\gamma, I = 1)$  evaluated at a point  $x_i\gamma$  of  $x\gamma$  can be estimated by a nonparametric kernel estimator  $E_n(y_1 | x_i\hat{\gamma}) - E_n(w | x_i\hat{\gamma})\beta$ , where  $\hat{\gamma}$  is a consistent estimate of  $\gamma$  from the selection equation. The bias-corrected structural equation becomes

$$y_{1i} - E_n(y_1 | x_i\hat{\gamma}) = (w_i - E_n(w | x_i\hat{\gamma}))\beta + \hat{u}_{ni}. \quad (18.16)$$

If  $\mathbf{p}$  is a vector of instrumental variables for  $w$ , a semiparametric IV estimator of  $\beta$  can be

$$\hat{\beta}_p = \left( \sum_{i=1}^n t_n(x_i \hat{\gamma}) p'_i(w_i - E_n(w | x_i \hat{\gamma})) \right)^{-1} \sum_{i=1}^n t_n(x_i \hat{\gamma}) p'_i(y_{1i} - E_n(y_1 | x_i \hat{\gamma})),$$

where  $t_n(x_i \hat{\gamma})$  is a weighting or trimming function. Powell (1987) used the denominator in the nonparametric kernel regression functions  $E_n(w | x_i \hat{\gamma})$  and  $E_n(y | x_i \hat{\gamma})$  as the weight so as to cancel the denominator in those kernel regression functions. This weighting plays the role of trimming and has nothing to do with the variance of disturbances  $\hat{u}_{ni}$  in equation (18.16). Lee (1994b) suggested a semiparametric two-stage least squares estimator and a more efficient semiparametric generalized two-stage least squares estimator. The latter takes into account both the heteroskedasticity of disturbances and the variance and covariance due to  $\hat{\gamma}$ . The disturbance  $\hat{u}_{ni}$  consists of three components as  $\hat{u}_{ni} = (u_{1i} - E_n(u_1 | x_i \hat{\gamma})) = (u_{1i} - E(u_1 | x_i \gamma)) - (E_n(u_1 | x_i \hat{\gamma}) - E_n(u_1 | x_i \gamma)) - (E_n(u_1 | x_i \gamma) - E(u_1 | x_i \gamma))$ . The first component represents the disturbance in the structural equation after the correction of selection bias. The second component represents the disturbance introduced in  $E_n(u_1 | x_i \gamma)$  by replacing  $\gamma$  by the estimate  $\hat{\gamma}$ . These two components are asymptotically uncorrelated. The last component represents the error introduced by the nonparametric estimate of the conditional expectation of  $u_{1i}$ . The last component does not influence the asymptotic distribution of a semiparametric two-stage estimator due to an asymptotic orthogonality property of the index structure. As the variance of  $u_{1i} - E(u_1 | x_i \gamma)$  is a function of  $x_i \gamma$ , it can be estimated by a nonparametric kernel estimator  $\hat{\omega}_{ni}$ . Let  $\Sigma$  be the variance matrix of the vector consisting of  $\hat{u}_{ni}$ , which is determined by the first two components of  $\hat{u}_{ni}$ . It captures the heteroskedastic variances of the first component and the covariance of the second component across sample observations due to  $\hat{\gamma}$ . A feasible semiparametric generalized two-stage least-squares estimator can either be  $\hat{\beta} = [\hat{W}' \hat{X}_2 (\hat{X}_2' \hat{X}_2)^{-1} \hat{X}_2' \hat{\Sigma}^{-1} \hat{W}]^{-1} \hat{W}' \hat{X}_2 (\hat{X}_2' \hat{X}_2)^{-1} \hat{X}_2' \hat{\Sigma}^{-1} \hat{Y}$ , or  $\tilde{\beta} = [\hat{W}' \hat{\Sigma}^{-1} \hat{X}_2 (\hat{X}_2' \hat{\Sigma}^{-1} \hat{X}_2)^{-1} \hat{X}_2' \hat{\Sigma}^{-1} \hat{W}]^{-1} \hat{W}' \hat{\Sigma}^{-1} \hat{X}_2 (\hat{X}_2' \hat{\Sigma}^{-1} \hat{X}_2)^{-1} \hat{X}_2' \hat{\Sigma}^{-1} \hat{Y}$ , where the elements of  $\hat{X}_2$ ,  $\hat{W}$  and  $\hat{Y}$  are, respectively,  $t_n(x_i \hat{\gamma})(x_{(2)i} - E_n(x_{(2)} | x_i \hat{\gamma}))$ ,  $t_n(x_i \hat{\gamma})(w_i - E_n(w | x_i \hat{\gamma}))$  and  $t_n(x_i \hat{\gamma})(y_{1i} - E_n(y_1 | x_i \hat{\gamma}))$ . These two estimators are asymptotically equivalent and are asymptotically efficient semiparametric IV estimators (conditional on the choice of first-stage estimator  $\hat{\gamma}$  and the trimming function). These semiparametric methods are two-stage estimation methods in that the selection equation is separately estimated and its coefficient estimate  $\hat{\gamma}$  is used for the estimation of the outcome equations.

Instead of two-stage methods, it is possible to estimate jointly the selection and structural outcome equations so as to improve efficiency (Lee, 1998). As a generalization, consider the estimation of a nonlinear simultaneous equation sample selection model:  $g(y^*, x, \beta) = u$ , where the vector  $y^*$  can be observed only if  $x\gamma > \varepsilon$ . Under the index assumption that the joint distribution of  $u$  and  $\varepsilon$  conditional on  $x$  may depend only on  $x\gamma$ , the bias-corrected structural system is  $g(y, x, \beta) = E(g(y, x, \beta) | I = 1, x\gamma) + \eta$ , where  $\eta = u - E(u | I = 1, x\gamma)$  with its variance being a

function of  $x\gamma$ . This system implies the moment equation  $E(g(y, x, \beta) | I = 1, x) = E(g(y, x, \beta) | I = 1, x\gamma)$  at the true parameter vector. For a truncated sample selection model where only sample observations of  $y$  and the event  $I = 1$  are available, this moment equation forms the system for estimation. For a (censored) sample selection model, the events of  $I = 1$  or  $I = 0$  are observed that introduces an additional moment equation  $E(I | x) = E(I | x\gamma)$  at the true parameter vector. These moment equations can be used together for estimation. Suppose that these moment equations are combined and are written in a general format as  $E(f(z, \beta) | x) = E(f(z, \beta) | x\gamma)$ , where  $z$  includes all endogenous and exogenous variables in the model. The parameter vector  $\beta$  in the system can be estimated by semiparametric nonlinear two-stage least squares methods.  $E(f(z, \beta) | x\gamma)$  can be estimated by a nonparametric regression function  $E_n(f(z, \beta) | x\gamma)$ . The relevant variance function can be estimated by  $V_n(x\gamma) = E_n(f(z, \beta)f'(z, \beta) | x\gamma) - E_n(f(z, \beta) | x\gamma)E_n(f'(z, \beta) | x\gamma)$ . Let  $u_n(z, \theta) = f(z, \beta) - E_n(f(z, \beta) | x\gamma)$  and  $t_n$  be a proper trimming function. Let  $w$  be an IV vector. The semiparametric nonlinear weighted two-stage method with the IV  $w$  is

$$\min_{\theta} \sum_{i=1}^n t_{ni} u'_n(z_i, \theta) V_n^{-1}(x_i \hat{\gamma}) w_i \left( \sum_{i=1}^n t_{ni} w_i' V_n^{-1}(x_i \hat{\gamma}) w_i \right)^{-1} \sum_{i=1}^n t_{ni} w_i' V_n^{-1}(x_i \hat{\gamma}) u_n(z_i, \theta),$$

where  $\hat{\gamma}$  is a consistent estimate of  $\gamma$ . Lee (1998) shows that an optimal IV is any consistent estimate of  $G_\theta(x, \theta)$  where  $G_\theta(x, \theta) = [E(\frac{\partial f(z, \beta)}{\partial \theta} | x) - E(\frac{\partial f(z, \beta)}{\partial \theta} | x\gamma)] - \nabla' E(f(z, \beta) | x\gamma)[\frac{\partial \gamma'(\theta)x'}{\partial \theta'} - E(\frac{\partial \gamma'(\theta)x'}{\partial \theta'} | x\gamma)]$ , where  $\nabla E(\cdot | x\gamma)$  denotes the gradient of  $E(\cdot | x\gamma)$  with respect to the vector  $x\gamma$ . In Lee (1998), semiparametric minimum-distance methods have also been introduced. Semiparametric minimum-distance methods compare directly  $E_n(f(z, \beta) | x)$  with  $E_n(f(z, \beta) | x\gamma)$ . A semiparametric minimum-distance method with weighting is

$$\min_{\theta} \sum_{i=1}^n t_{ni} [E_n(f(z, \beta) | x_i) - E_n(f(z, \beta) | x_i \gamma)] V_n^{-1}(x_i \hat{\delta}) [E_n(f(z, \beta) | x_i) - E_n(f(z, \beta) | x_i \gamma)].$$

The semiparametric weighted minimum-distance estimator is asymptotically equivalent to the optimal IV estimator. The semiparametric minimum-distance method has a interesting feature of not emphasizing the construction of instrumental variables. As  $z$  in  $f(z, \beta)$  may or may not contain endogenous variables, semiparametric minimum-distance methods can be applied to the estimation of regression models as well as simultaneous equation models in a single framework.

The efficiency of estimating a structural equation is related to the efficiency issue for semiparametric models with conditional moment restrictions. Chamberlain (1992) investigates semiparametric efficiency bounds for semiparametric models with conditional moment restrictions. The conditional moment restriction considered has the form  $E[\rho(x, y, \beta_0, q_0(x_2)) | x] = 0$ , where  $x_2$  is a subvector of  $x$ ,  $\rho(x, y, \beta, \tau)$  is a known function, but  $q(x_2)$  is an unknown mapping. Chamberlain (1992)

derives an efficiency bound for estimators of  $\beta$  under the conditional moment restriction. Several concrete examples are provided. Among them is a sample selection model. The sample selection model considered in Chamberlain (1992) is  $y^* = x_1\beta + x_2\delta + u$  and  $I = 1$ , if  $g(x_2, \varepsilon) \geq 0$ ; 0, otherwise, where  $x = (x_1, x_2)$  and  $y = (y_1, I)$  with  $y_1 = Iy^*$ , are observed. The unknown function  $g$  depends on  $x$  only via  $x_2$  but is otherwise unrestricted. The disturbances  $u$  and  $\varepsilon$  satisfy the restrictions that  $E(u|x, \varepsilon) = E(u|\varepsilon)$  and  $\varepsilon$  is independent of  $x_1$  conditional on  $x_2$ . It is a sample selection model with indices  $x_2$ . This model implies that  $E(y_1|x, I=1) = x_1\beta_0 + q_0(x_2)$ , where  $q_0(x_2) = x_2\delta_0 + E(u|x_2, I=1)$ . Thus  $\rho(x, y, \beta, \tau) = I(y_1 - x_1\beta - \tau)$  in Chamberlain's framework for this sample selection model. Chamberlain pointed out that one might extend the  $\rho$  function to include the restriction that  $E(I|x) = E(I|x_2)$ , so that  $\rho(x, y, \beta, \tau) = [I(y_1 - x_1\beta - \tau_1), I - \tau_2]$  but the efficiency bound for  $\beta$  is the same with either one of the above  $\rho$ . Let  $\sigma^2(x)$  denote  $\text{var}(y_1|x, I=1)$ . The efficient bound for  $\beta$  is  $J = E\{E(I|x_2)[E(\frac{x_1'x_1}{\sigma^2(x)}) - E(\frac{x_1'}{\sigma^2(x)}|x_2)E(\frac{x_1}{\sigma^2(x)}|x_2)/E(\frac{1}{\sigma^2(x)}|x_2)]\}$ . For the case where  $\sigma^2(x)$  happens to depend on  $x$  only through  $x_2$ , the efficiency bound will be simplified to  $J = E\{E(I|x_2)\sigma^2(x_2)[x_1 - E(x_1|x_2)][x_1 - E(x_1|x_2)]'\}$ . The semiparametric weighted minimum-distance method can be applied to estimate the sample selection model with  $f'(z, \beta) = [I(y - x_1\beta), I]$ . Lee (1998) showed that the semiparametric weighted minimum-distance estimator attains the efficiency bound if  $\sigma^2(x)$  happens to depend only on  $x_2$ . It is of interest to note that if the moment equation  $E(I|x) = E(I|x_2)$  were ignored and only the moment equation  $E(I(y_1 - x_1\beta_0)|x) = E(I(y_1 - x_1\beta_0)|x_2)$  were used in the estimation, the resulting estimator will have a larger variance. The point is that even though the moment restriction  $E(I|x) = E(I|x_2)$  does not contain  $\beta$ , it helps to improve the efficiency in estimating  $\beta$  (as in a seemingly unrelated regression framework). On the other hand, if the conditional moment restriction  $E(y - x_1\beta_0|x, I=1) = E(y_1 - x_1\beta_0|x\gamma_0, I=1)$  is used for estimation, the resulted estimator will have the same smaller variance. This is so because  $I - E(I|x)$  is uncorrelated with the disturbance  $(y - x_1\beta_0) - E(y - x_1\beta_0|x, I=1)$ .

#### 4.4 Estimation of the intercept

The semiparametric estimation of sample selection models described has focused on the estimation of regression coefficients in outcome and selection equations. The intercept of the outcome equation has been absorbed in the unknown distribution of disturbances. For some empirical applications, one might be interested in estimating counterfactual outcomes and, hence, the intercept of an outcome equation. The semi-nonparametric likelihood approach of Gallant and Nychka (1987) based on series expansion can consistently estimate the unknown intercept but the asymptotic distribution of the estimator remains unknown. Instead of the likelihood approach, alternative approaches can be based on the observed outcome equation under the assumption that the underlying disturbances have zero-mean. One may imagine that the observed outcome equation is not likely to be subject to selection bias for individuals whose decisions of participation are almost certain, i.e. individuals with observed characteristics  $x$  such that  $P(I=1|x) = 1$ . This idea is in Olsen (1982), and has lately been picked up in

Heckman (1990) and Andrews and Schafgans (1998). Consider the sample selection model with outcome  $y = \beta_1 + x\beta_2 + u$ , which can be observed only if  $z\gamma > \varepsilon$ , where  $(u, \varepsilon)$  is independent of  $x$  and  $z$ . The  $\beta_2$  and  $\gamma$  can be estimated by various semiparametric methods as described before. Let  $\hat{\beta}_2$  and  $\hat{\gamma}$  be, respectively, consistent estimates of  $\beta_2$  and  $\gamma$ . Heckman (1990) suggests the estimator  $\hat{\beta}_1 = \sum_{i=1}^n I_i(y_i - x_i\hat{\beta}_2)I_{(b_n, \infty)}(z_i\hat{\gamma})/\sum_{i=1}^n I_iI_{(b_n, \infty)}(z_i\hat{\gamma})$ , where  $\{b_n\}$  is supposed to be a sequence of bandwidth parameters depending on sample size  $n$  such that  $b_n \rightarrow \infty$ . The latter design is necessary as only the upper tails of  $z\gamma$  would likely identify the individual with the choice probability close to one. Heckman did not provide an asymptotic analysis of his suggested estimator. Andrews and Schafgans (1998) suggest a smooth version by replacing  $I_{(b_n, \infty)}(z_i\hat{\gamma})$  with a smooth distribution function. The later modification provides relatively easy asymptotic analysis of the estimator. The rate of convergence and possible asymptotic distribution of the estimator depend on some relative behaviors of the upper tail distributions of  $z\gamma$  and  $\varepsilon$ . Andrews and Schafgans (1998) show that the intercept estimator can only achieve up to a cube-root- $n$  rate of convergence when the upper tail of the distribution of  $z\gamma$  has the same tail thickness as the upper tail of the distribution of  $\varepsilon$ . The rate of convergence can be up to square-root- $n$  only when the upper tail of the distribution of  $z\gamma$  is thicker than the upper tail of the distribution of  $\varepsilon$ . One may recall some similarity on the asymptotic behavior of order statistics for the estimation of the range of a distribution.

## 4.5 Sample selection models with a tobit selection rule

For the semiparametric estimation of a sample selection model with a discrete choice equation, the exclusion restriction of a relevant regressor in the choice equation from the outcome equation provides the crucial identification condition. Such an exclusion restriction will not be required when the selection criterion is a tobit rule. The identification and estimation of such a model are considered in Lee (1994a), Chen (1997), and Honoré, Kyriazidou, and Udry (1997). The model in Lee (1994a) assumes that the disturbances  $(u_1, u_2)$  in equation (18.6) are independent of the regressors  $x$ . With iid disturbances, model (18.6) implies two observable outcome equations:  $E(y_{2i} | y_{1i} > 0, x_i) = x_i\beta + E(u_{2i} | y_{1i} > 0, x_i)$  and

$$E(y_{2i} | u_1 > -x_i\gamma, x\gamma > x_i\gamma, x_i) = E(x | x\gamma > x_i\gamma, x_i)\beta + E(u_{2i} | u_1 > -x_i\gamma, x\gamma > x_i\gamma, x_i).$$

As the disturbances are independent of  $x$ , the conditional moment restriction  $E(u_{2i} | u_1 > -x_i\gamma, x\gamma > x_i\gamma, x_i) = E(u_{2i} | y_{1i} > 0, x_i)$  provides the identification of  $\beta$  given  $\gamma$  from the tobit equation. The intuition behind these formulations is based on the fact that the density of  $(u_{1i}, u_{2i})$  conditional on  $y_{1i} > 0$  and  $x_i$  is the same as the density of any  $(u_{1j}, u_{2j})$  conditional on  $u_{1j} > -x_i\gamma$  and  $x_j\gamma > x_i\gamma$  at the point  $x_i\gamma$ . The conditions  $x_j\gamma > x_i\gamma$  and  $u_{1j} > -x_i\gamma$  imply that  $y_{1j} = x_j\gamma + u_{1j} > 0$  and, hence, the observability of  $(y_{1j}, y_{2j})$ . The  $E(u_{2i} | u_1 > -x_i\gamma, x\gamma > x_i\gamma, x_i)$  can be estimated by  $\sum_{j=1}^n u_{2j}I(u_{1j} > -x_i\gamma, x_{1j}\gamma > x_{1i}\gamma)/\sum_{j=1}^n I(u_{1j} > -x_i\gamma, x_{1j}\gamma > x_{1i}\gamma)$ . Instead of this estimator,

Lee (1994a) suggests a kernel smoothing estimator  $E_n(y_2 - x\beta | x_{1i}\hat{\gamma})$  where  $\hat{\gamma}$  is a consistent estimator from first-stage semiparametric estimation of the tobit selection equation, and proposed a semiparametric least squares procedure:  $\min_{\beta} \frac{1}{n} \sum_{i=1}^n I_X(x_i)(y_{2i} - x_i\beta - E_n(y_2 - x\beta | x_{1i}, \hat{\gamma}))$ , where  $I_X(x_i)$  is a trimming function on  $x$ . The two-stage estimator of  $\beta$  has a closed form expression and is  $\sqrt{n}$ -consistent and asymptotically normal. Chen (1997) proposed two estimation approaches for this model. One is similar to the semiparametric least squares procedure in Lee (1994a) except that the ratio of sample indicators is used without smoothing and the trimming function is replaced by the weighting function as in Powell (1987). The second estimation approach in Chen (1997) is  $\min_{\beta, \alpha} \frac{1}{n} \sum_{i=1}^n I(y_{1i} - x_i\hat{\gamma} > 0, x_i\hat{\gamma} > 0)(y_{2i} - x_i\beta - \alpha)$ . This approach utilizes a different portion of the observable disturbances where  $E(u_{2i} | u_{1i} > 0, x_i\hat{\gamma} > 0) = \alpha$  is a constant for all  $i$ s under the independence assumption. Chen (1997) also derives the asymptotic efficiency bound for this semiparametric model. None of the estimators available in the literature (including the estimators in Honoré *et al.* (1997) discussed later) attain the efficiency bound. Honoré *et al.* (1997) propose two-stage estimators based on symmetry. They consider first the case that  $(u_1, u_2)$  is symmetrically distributed conditional on  $x$  (arbitrary heteroskedasticity is allowed), i.e.  $(-u_1, -u_2)$  is distributed like  $(u_1, u_2)$  conditional on  $x$ . The property of symmetry in disturbances was first explored for estimating the censored and truncated regression models in Powell (1986). Even though the underlying disturbances are symmetrically distributed, the observable disturbances are no longer symmetrically distributed under sample selection. Honoré *et al.* (1997) restore the symmetric property by restricting the estimation of  $\beta$  with sample observations in the region where  $-x\gamma < u_1 < x\gamma$  (equivalently,  $0 < y_1 < 2x\gamma$ ). With sample observations in the restricted region,  $u_1$  is symmetrically distributed around 0 and the proposed estimation procedures can be based on least absolute deviations or least squares, i.e.  $\min_{\beta} \frac{1}{n} \sum_{i=1}^n I(0 < y_{1i} < 2x_i\hat{\gamma}) | y_{2i} - x_i\beta |$ , or  $\min_{\beta} \frac{1}{n} \sum_{i=1}^n I(0 < y_{1i} < 2x_i\hat{\gamma})(y_{2i} - x_i\beta)^2$ , where  $\hat{\gamma}$  is a first-stage estimator, e.g. semiparametric censored or truncated regression from Powell (1986). The restoration of symmetric region demonstrates elegantly the usefulness of observed residuals of  $u_1$  from the tobit selection equation for estimation. Honoré *et al.* (1997) consider also the case that  $(u_1, u_2)$  is independent of  $x$  in the underlying equations. They suggest estimation approaches based on pairwise differences across sample observations. The pairwise difference approach is to create a possible symmetric property on the difference of disturbances. The difference of two iid random variables must be symmetrically distributed around zero if there is no sample selection. Under sample selection, one has to restore the symmetry property of the pairwise difference. Honoré *et al.* suggested the trimming of  $u_{1i}$  and  $u_{1j}$  identically so that  $u_{1i} > \max\{-x_i\gamma, -x_j\gamma\}$  and  $u_{1j} > \max\{-x_i\gamma, -x_j\gamma\}$  (equivalently,  $y_{1i} > \max\{0, (x_i - x_j)\gamma\}$  and  $y_{1j} > \max\{0, (x_j - x_i)\gamma\}$ ). On this trimmed region, the independence assumption implies that  $u_{2i} - u_{2j}$  is distributed symmetrically around 0. Their suggested pairwise difference estimators are  $\min_{\beta} \sum_{i < j} I(y_{1i} > \max\{0, (x_i - x_j)\hat{\gamma}\}, y_{1j} > \max\{0, (x_j - x_i)\hat{\gamma}\}) | y_{2i} - y_{2j} - (x_{2i} - x_{2j})\beta |$  or  $\min_{\beta} \sum_{i < j} I(y_{1i} > \max\{0, (x_i - x_j)\hat{\gamma}\}, y_{1j} > \max\{0, (x_j - x_i)\hat{\gamma}\}) | y_{2i} - y_{2j} - (x_{2i} - x_{2j})\beta |^2$ . Consistency and asymptotic normality of the estimators are derived by empirical process arguments from Pakes and Pollard (1989).

## 4.6 Identification and estimation of counterfactual outcomes

As counterfactual outcomes are important objects of inference, one may be interested in the identification and estimation of counterfactual outcomes. The possible identification of counterfactual outcomes follows from model structures and observed decisions and outcomes (Heckman, 1990; Lee, 1995). Observed outcomes and choice probabilities provide sample information. Latent variable models provide prior structural restrictions.

Professor C. Manski in a series of articles put aside the latent-variable model perspective to go back to probabilistic basics. His results are summarized in Manski (1994). The main findings provide informative bounds on some counterfactual outcomes. Without latent variable modeling, if one is not satisfied with just bounds, the identification and evaluation of a counterfactual outcome would require extra prior restrictions on some other counterfactual outcomes. Statisticians approach the selected sample as a mixture problem. A widely-used method of evaluation in statistics is the method of matching (Rubin, 1987). Heckman, Ichimura, and Todd (1998) show that the fundamental identification condition (or assumption) for the matching method is a condition imposed on a specific counterfactual outcome. Corresponding to the two-sector model in (18.2)–(18.3), this identification condition requires that  $\sigma_{2e} = 0$  (Heckman *et al.*, 1998, pp. 268–9). Professor J. Heckman and his associates in a series of forthcoming papers contrast the econometrics and statistical approaches on program evaluation. Some preliminary review can be found in M.J. Lee (1997).

### **Note**

- \* The author acknowledges research support from the Research Grants Council of Hong Kong under grant HKUST595/96H for his research.

### **References**

- Ahn, H., and J.L. Powell (1993). Semiparametric estimation of censored selection models with a nonparametric selection mechanism. *Journal of Econometrics* 58, 3–29.
- Ai, C. (1997). A semiparametric maximum likelihood estimator. *Econometrica* 65, 933–63.
- Amemiya, T. (1973). Regression analysis when the dependent variable is truncated normal. *Econometrica* 41, 997–1016.
- Amemiya, T. (1974). Multivariate regression and simultaneous equation models when the dependent variables are truncated normal. *Econometrica* 42, 999–1012.
- Amemiya, T. (1979). The estimation of a simultaneous equation tobit model. *International Economic Review* 20, 169–81.
- Amemiya, T. (1983). A comparison of the Amemiya GLS and the Lee-Maddala-Trost G2SLS in a simultaneous equations tobit model. *Journal of Econometrics* 23, 295–300.
- Amemiya, T. (1984). Tobit models: a survey. *Journal of Econometrics* 24, 3–61.
- Andrews, D.W.K. (1991). Asymptotic normality of series estimators for nonparametric and semiparametric regression models. *Econometrica* 59, 307–45.

- Andrews, D.W.K., and M.M.A. Schafgans (1998). Semiparametric estimation of the intercept of a sample selection model. *Review of Economic Studies* 65, 497–517.
- Bjorklund, A., and R. Moffitt (1987). Estimation of wage gains and welfare gains in self-selection models. *Review of Economics and Statistics* 69, 42–9.
- Chamberlain, G. (1986). Asymptotic efficiency in semiparametric models with censoring. *Journal of Econometrics* 32, 189–218.
- Chamberlain, G. (1992). Efficiency bounds for semiparametric regression. *Econometrica* 60, 567–96.
- Chen, S. (1997). Semiparametric estimation of the Type-3 tobit model. *Journal of Econometrics* 80, 1–34.
- Chen, S., and L.F. Lee (1998). Efficient semiparametric scoring estimation of sample selection models. *Econometric Theory* 14, 423–62.
- Cosslett, S.R. (1991). Semiparametric estimation of regression model with sample selectivity. In W.A. Barnett, J. Powell, and G. Tauchen (eds.) *Nonparametric and Semiparametric Methods in Econometrics and Statistics*. pp. 175–97. Cambridge: Cambridge University Press.
- Dubin, J., and D. McFadden (1984). An econometric analysis of residential electric appliance holdings and consumption. *Econometrica* 52, 345–62.
- Gallant, A.R., and D.W. Nychka (1987). Semiparametric maximum likelihood estimation. *Econometrica* 55, 363–93.
- Goldberger, A.S. (1983). Abnormal selection bias. In S. Karlin, T. Amemiya, and L.A. Goodman (eds.) *Studies in Econometrics, Time Series and Multivariate Statistics*. New York: Wiley.
- Griliches, Z., B.H. Hall, and J.A. Hausman (1978). Missing data and self-selection in large panels. *Annals de l'INSEE* 30–31, 137–76.
- Gronau, R. (1974). Wage comparisons: a selectivity bias. *Journal of Political Economy* 82, 119–43.
- Hay, J., and R.J. Olsen (1984). Let them eat cake: a note on comparing alternative models of the demand for medical care. *Journal of Business and Economic Statistics* 2, 279–82.
- Heckman, J.J. (1974). Shadow prices, market wages, and labor supply. *Econometrica* 42, 679–94.
- Heckman, J.J. (1979). Sample selection bias as specification error. *Econometrica* 47, 153–61.
- Heckman, J.J. (1990). Varieties of selection bias. *American Economic Association Papers and Proceedings* 313–18.
- Heckman, J.J., and R. Robb (1985). Alternative methods for evaluating the impact of interventions. In J. Heckman and B. Singer (eds.) *Longitudinal Analysis of Labor Market Data*. Cambridge: Cambridge University Press.
- Heckman, J.J., and B.E. Honore (1990). The empirical content of the Roy model. *Econometrica* 58, 1121–49.
- Heckman, J.J., H. Ichimura, and P. Todd (1998). Matching as an econometric evaluation estimator. *Review of Economic Studies* 65, 261–94.
- Honoré, B.E., E. Kyriazidou, and C. Udry (1997). Estimation of type 3 tobit models using symmetric trimming and pairwise comparisons. *Journal of Econometrics* 76, 107–28.
- Ichimura, H. (1993). Semiparametric least squares estimation of single index models. *Journal of Econometrics* 58, 71–120.
- Ichimura, H., and L.F. Lee (1991). Semiparametric estimation of multiple index models: single equation estimation. In W.A. Barnett, J. Powell, and G. Tauchen (eds.) *Nonparametric and Semiparametric Methods in Econometrics and Statistics*, ch. 1. New York: Cambridge University Press.
- Lee, L.F. (1978). Unionism and wage rates: a simultaneous equation model with qualitative and limited dependent variables. *International Economic Review* 19, 415–33.

- Lee, L.F. (1981). Simultaneous equations models with discrete endogenous variables. In C.F. Manski and D. McFadden (eds.) *Structural Analysis of Discrete Data and Econometric Applications*, ch. 9. Cambridge, MA: MIT Press.
- Lee, L.F. (1982). Some approaches to the correction of selectivity bias. *Review of Economic Studies* 49, 355–72.
- Lee, L.F. (1983). Generalized econometrics models with selectivity. *Econometrica* 51, 507–12.
- Lee, L.F. (1984). Tests for the bivariate normal distribution in econometric models with selectivity. *Econometrica* 52, 843–63.
- Lee, L.F. (1992a). Amemiya's generalized least squares and tests of overidentification in simultaneous equation models with qualitative or limited dependent variables. *Econometric Reviews* 11, 319–28.
- Lee, L.F. (1992b). On efficiency of methods of simulated moments and maximum simulated likelihood estimation of discrete response models. *Econometric Theory* 8, 518–52.
- Lee, L.F. (1994a). Semiparametric two-stage estimation of sample selection models subject to tobit-type selection rules. *Journal of Econometrics* 61, 305–44.
- Lee, L.F. (1994b). Semiparametric instrumental variable estimation of simultaneous equation sample selection models. *Journal of Econometrics* 63, 341–88.
- Lee, L.F. (1995). The computational of opportunity costs in polytomous choice models with selectivity. *Review of Economics and Statistics* 77, 423–35.
- Lee, L.F. (1996). Simulation estimation of sample selection models. Working Paper no. 96/97-4, Department of Economics, Hong Kong University of Science and Technology.
- Lee, L.F. (1998). Semiparametric estimation of simultaneous-equation microeconometric models with index restrictions. *Japanese Economic Review* 49, 343–80.
- Lee, L.F., and A. Chesher (1986). Specification testing when some test statistics are identically zero. *Journal of Econometrics* 31, 121–49.
- Lee, L.F., G.S. Maddala, and R.P. Trost (1980). Asymptotic covariance matrices of two-stage probit and two-stage tobit methods for simultaneous equations models with selectivity. *Econometrica* 48, 491–503.
- Lee, M.-J. (1997). Econometric methods for sample selection and treatment effect models. Manuscript, Institute of Policy and Planning Science, University of Tsukuba, Japan.
- Leung, S.F., and S. Yu (1996). On the choice between sample selection and two-part models. *Journal of Econometrics* 72, 197–229.
- Manning, W.G., N. Duan, and W.H. Rogers (1987). Monte Carlo evidence on the choice between sample selection and two-part models. *Journal of Econometrics* 35, 59–82.
- Maddala, G.S. (1983). *Limited Dependent and Qualitative Variables in Econometrics*. Cambridge: Cambridge University Press.
- Maddala, G.S. (1985). A survey of the literature on selectivity bias as it pertains to health care markets. *Advances in Health Economics and Health Services Research* 6, 3–18.
- Manski, C. (1975). Maximum score estimation of the stochastic utility model of choice. *Journal of Econometrics* 3, 205–28.
- Manski, C. (1994). The selection problem. In C. Sims (ed.) *Advances in Econometrics*, ch. 4, pp. 143–70. Cambridge: Cambridge University Press.
- Melino, A. (1982). Testing for selection bias. *Review of Economic Studies* 49, 151–3.
- McFadden, D. (1973). Conditional logit analysis of qualitative choice behavior. In P. Zarembka (ed.) *Frontiers in Econometrics*. New York: Academic Press.
- McFadden, D. (1978). Modeling the choice of residential location. In A. Karlquist *et al.* (eds.) *Spatial Interaction Theory and Residential Location*. Amsterdam: North-Holland.
- McFadden, D. (1989). A method of simulated moments for estimation of discrete response models without numerical integration, *Econometrica* 57, 995–1026.
- Mroz, T.A. (1987). The sensitivity of empirical models with sample-selection biases. *Econometrica* 55, 765–99.

- Nawata, K. (1993). A note on the estimation with sample-selection bias. *Economics Letters* 42, 15–24.
- Nawata, K., and N. Nagase (1996). Estimation of sample selection bias models. *Econometric Reviews* 15, 387–400.
- Nelson, F.D. (1977). Censored regression models with unobserved, stochastic censoring thresholds. *Journal of Econometrics* 6, 309–27.
- Nelson, F. (1984). Efficiency of the two step estimator for models with endogenous sample selection. *Journal of Econometrics* 24, 181–96.
- Newey, W.K. (1987). Efficient estimation of limited dependent variable models with endogenous explanatory variables. *Journal of Econometrics* 36, 231–50.
- Newey, W.K. (1988). Two step estimation of sample selection models. Manuscript, Department of Economics, Princeton University.
- Newey, W.K. (1990). Semiparametric efficiency bounds. *Journal of Applied Econometrics* 5, 99–135.
- Newey, W.K., J.L. Powell, and J.R. Walker (1990). Semiparametric estimation of selection models: some empirical results. *AEA Papers and Proceedings* 80, 324–8.
- Olsen, R. (1980). A least squares correction for selectivity bias. *Econometrica* 48, 1815–20.
- Olsen, R. (1982). Distributional tests for selectivity bias and a more robust likelihood estimator. *International Economic Review* 23, 223–40.
- Pagan, A., and F. Vella (1989). Diagnostic tests for models based on individual data: a survey. *Journal of Applied Econometrics* 4, S29–S59.
- Pakes, A., and D. Pollard (1989). Simulation and the asymptotic of optimization estimators. *Econometrica* 57, 1027–57.
- Powell, J.L. (1986). Symmetrically trimmed least squares estimation for tobit models. *Econometrica* 54, 1435–60.
- Powell, J.L. (1987). Semiparametric estimation of bivariate latent variable models. Discussion paper no. 8704, Social Systems Research Institute, University of Wisconsin, Madison, WI.
- Powell, J.L. (1994). Estimation of semiparametric models. In R.F. Engle and D.L. McFadden (eds.) *Handbook of Econometrics, Volume 4*, ch. 14. Amsterdam: North-Holland.
- Robinson, P.M. (1988). Root-n-consistent semiparametric regression. *Econometrica* 56, 931–54.
- Roy, A. (1951). Some thoughts on the distribution of earnings. *Oxford Economic Papers* 3, 135–46.
- Rubin, D. (1987). *Multiple Imputation for Nonresponse in Surveys*. New York: John Wiley.
- Sattiger, M. (1978). Comparative advantage in individuals. *Review of Economics and Statistics* 60, 259–67.
- Schmertmann, C.P. (1994). Selectivity bias correction methods in polychotomous sample selection models. *Journal of Econometrics* 60, 101–32.
- Stein, C. (1956). Efficient nonparametric testing and estimation. In *Proceedings of the Third Berkeley Symposium on Mathematical Statistics and Probability*, Volume 1 pp. 187–95. Berkeley: University of California Press.
- Vella, F. (1998). Estimating models with sample selection bias: a survey. *Journal of Human Resources* 33, 127–69.
- Wales, T.J., and A.D. Woodland (1980). Sample selectivity and the estimation of labour supply functions. *International Economic Review* 21, 437–68.
- Willis, R.J., and S. Rosen (1979). Education and self-selection. *Journal of Political Economy* 87, S7–S36.

CHAPTER NINETEEN

# Random Coefficient Models

*P.A.V.B. Swamy and George S. Tavlas\**

## 1 INTRODUCTION

Random coefficient models (RCMs) grew out of Zellner's path-breaking article (1969) on aggregation and have undergone considerable modification over time.<sup>1</sup> Initially, RCMs were primarily concerned with relaxing an assumption typically made by researchers who use classical models. This assumption is that there is a constant vector of coefficients relating the dependent and independent variables. Unfortunately, as Keynes (Moggridge, 1973, p. 286) long ago observed, the assumption of constant coefficients is unlikely to be a reasonable one. Recent work on RCMs has focused on also relaxing the following assumptions frequently made by researchers in econometrics: (i) the true functional forms of the systematic components of economic relationships (whether linear or nonlinear) are known; (ii) excluded variables are proxied through the use of an additive error term and, therefore, have means equal to zero and are independent of the included explanatory variables; and (iii) variables are not subject to measurement error.

The purpose of this chapter is to provide an accessible description of RCMs. The chapter is divided into six sections, including this introduction. Section 2 discusses some characteristics of what we characterize as first-generation models. Essentially, these models attempt to deal with the problem that arises because aggregate time series data and cross-section data on micro units are unlikely to have constant coefficients. Thus, the focus of this literature is on relaxing the constant coefficient assumption. Section 3 describes a generalized RCM whose origins are in work by Swamy and Mehta (1975) and Swamy and Tinsley (1980). The generalized RCM, which relaxes the constant coefficient assumption and all three of the restrictions mentioned in the preceding paragraph, is referred to as a second-generation model. It is based on the assumptions that

1. any variable that is not mismeasured is true;
2. any economic equation with the correct functional form, without any omitted explanatory variable, and without mismeasured variables is true.

Economic theories are true if and only if they deal with the true economic relationships. They cannot be tested unless we know how to estimate the true economic relationships. The generalized RCM corresponds to the underlying true economic relationship if each of its coefficients is interpreted as the sum of three parts: (i) a direct effect of the true value of an explanatory variable on the true value of an explained variable; (ii) an indirect effect (or omitted-variable bias) due to the fact that the true value of the explanatory variable affects the true values of excluded variables and the latter values, in turn, affect the true value of the explained variable; and (iii) an effect of mismeasuring the explanatory variable. A necessary condition that a specified model coincides with the underlying true economic relationship is that each of its coefficients has this interpretation. Importantly, the second-generation models satisfy the conditions for observability of stochastic laws, defined by Pratt and Schlaifer (1988), whenever they coincide with the underlying true economic relationships. In order to enhance the relevance of the following discussion, the presentation of RCMs is made in the context of a money-demand model that has been extensively applied in the literature. Also, to heighten accessibility, the section does not attempt rigorous exposition of some technical concepts (e.g. stochastic laws), but refers the interested reader to the relevant literature. Section 4 applies five criteria to validate RCMs. Section 5 uses an example of a money demand model to illustrate application of the second-generation RCMs. Section 6 concludes.

## 2 SOME FIRST-GENERATION RCMs

When considering situations in which the parameters of a regression model are thought to change – perhaps as frequently as every observation – the parameter variation must be given structure to make the problem tractable. A main identifying characteristic of first-generation RCMs is that they are concerned with providing a structure to the process generating the coefficients. In other words, this class of models seeks to account for the process generating the coefficients of the regression model, but does not address specification issues related to functional form, omitted variables, and measurement errors.<sup>2</sup>

To explain, consider the following model:

$$m_t = x_{t1}\beta_{11} + \sum_{j=2}^K x_{tj}\beta_{1j} = x'_t\beta_t \quad (t = 1, 2, \dots, T), \quad (19.1)$$

where  $m_t$  is the logarithm of real money balances (i.e. a measure of the money supply divided by a price-level variable);  $x'_t$  is a row vector of  $K$  elements having the  $j$ th explanatory variable  $x_{tj}$  as its  $j$ th element;  $x_{t1} = 1$  for all  $t$ ; the remaining  $x_{tj}$  ( $j = 2, \dots, K$ ) are the variables thought to influence the demand for real money

balances (such as the logarithms of real income and interest rates);  $\beta_t$  is a column vector of  $K$  elements having the  $j$ th coefficient  $\beta_{tj}$  as its  $j$ th element; the first coefficient  $\beta_{t1}$  combines the usual disturbance term and intercept and  $t$  indexes time series or cross section observations.

Since the model assumes that  $\beta_t$  is changing, we have to say something about how it is changing. That is, we have to introduce some structure into the process thought to be determining the coefficients. One simple specification is

$$\beta_t = \bar{\beta} + \varepsilon_t, \quad (19.2)$$

where  $\bar{\beta} = (\bar{\beta}_1, \dots, \bar{\beta}_K)'$  is the mean and the  $\varepsilon_t = (\varepsilon_{t1}, \dots, \varepsilon_{tK})'$  are the disturbance terms that are identically and independently distributed with  $E(\varepsilon_t) = 0$ . Basically, this equation says that the variations in all the coefficients of equation (19.1) are random unless shown otherwise by the real-world sources and interpretations of  $\beta_t$  and at any particular time period or for any one individual these coefficients are different from their means.<sup>3</sup> Assume further that the random components,  $\varepsilon_{t1}, \dots, \varepsilon_{tK}$ , are uncorrelated,

$$E(\varepsilon_t \varepsilon_t') = \text{diag}[\sigma_{\varepsilon_1}^2, \sigma_{\varepsilon_2}^2, \dots, \sigma_{\varepsilon_K}^2] \quad (19.3)$$

which is a  $K \times K$  diagonal matrix whose  $i$ th diagonal element is  $\sigma_{\varepsilon_i}^2$ . This assumption stipulates that the random component of one coefficient is uncorrelated with that of another. We also assume that the  $x_t$  are independent of the  $\beta_t$ .<sup>4</sup> Correlations among the elements of  $\varepsilon_t$  will be introduced at the end of the section.

Since we have postulated that all the coefficients of equation (19.1) vary over time or across individuals according to equation (19.2), an issue that arises is how to test whether equations (19.1) and (19.2) are true. In order to answer this question, first substitute equation (19.2) into equation (19.1) to obtain

$$m_t = x_t' \bar{\beta} + \sum_{j=1}^K x_{tj} \varepsilon_{tj}. \quad (19.4)$$

Next, let  $w_t$  denote the combined disturbances (i.e. the sum of the products of  $x_{tj}$  and  $\varepsilon_{tj}$ ),  $w_t = \sum_{j=1}^K x_{tj} \varepsilon_{tj}$ , so that  $m_t = x_t' \bar{\beta} + w_t$ , where  $E(w_t | x_t) = 0$ . Moreover, the conditional variance of the combined disturbance term of equation (19.4) is

$$E(w_t^2 | x_t) = x_{t1}^2 \sigma_{\varepsilon_1}^2 + \dots + x_{tK}^2 \sigma_{\varepsilon_K}^2, \quad (19.5)$$

since  $w_t$  is a linear combination of uncorrelated random variables and the  $x_t$  are independent of the  $\varepsilon_t$ . For  $t \neq s$ , the covariance between  $w_t$  and  $w_s$  is zero. Let  $v_t = w_t^2 - E(w_t^2 | x_t)$ , where  $E(w_t^2 | x_t)$  is given in (19.5). It is straightforward to show that the conditional mean of  $v_t$  given  $x_t$  is zero.

The definition of  $v_t$  can be rearranged to form a regression as

$$w_t^2 = E(w_t^2 | x_t) + v_t, \quad (19.6)$$

where  $E(v_t | x_t)$ , as pointed out, is zero. Accordingly, using equations (19.5) and (19.6),  $w_t^2$  can be expressed as  $w_t^2 = \sum_{j=1}^K x_{tj}^2 \sigma_{\epsilon_j}^2 + v_t$ , where for  $t \neq s$ ,  $v_t$  and  $v_s$  are uncorrelated if the  $\epsilon_{tj}$  are independent.

If our goal is to test whether equation (19.1) is the same as the fixed-coefficient model of the conventional type, then our null hypothesis is that the random components of all the coefficients on  $x_{t2}, \dots, x_{tK}$  are zero with probability 1. A complete statement of this hypothesis is

$H_0$ : For  $i, j = 1, \dots, K; t, s = 1, \dots, T$ ,  $E(\beta_{tj}) = \bar{\beta}_j$  and

$$E(\epsilon_{ti}\epsilon_{sj}) = \begin{cases} \sigma_{\epsilon_i}^2 > 0 & \text{if } i = j = 1 \text{ and } t = s \\ 0 & \text{otherwise} \end{cases}$$

such that  $E(\epsilon_{t1} | x_t) = E(\epsilon_{t1}) = 0$  and  $\epsilon_{t1}$  is normally distributed.

There are several alternatives to this hypothesis. One of them, provided by equations (19.1)–(19.3), is

$H_1$ : For  $i, j = 1, \dots, K; t, s = 1, \dots, T$ ,  $E(\beta_{tj}) = \bar{\beta}_j$  and

$$E(\epsilon_{ti}\epsilon_{sj}) = \begin{cases} \sigma_{\epsilon_i}^2 > 0 & \text{if } i = j \text{ and } t = s \\ 0 & \text{otherwise} \end{cases}$$

such that  $E(\epsilon_{tj} | x_t) = E(\epsilon_{tj}) = 0$  and  $\epsilon_{tj}$  is normally distributed.

The following steps lead to a test of  $H_0$  against  $H_1$ : (i) obtain the ordinary least squares (OLS) estimate of  $\bar{\beta}$ , denoted by  $\bar{b}_{OLS}$ , by running the classical least squares regression of  $m_t$  on  $x_t$ ; (ii) in equation (19.4), replace  $\bar{\beta}$  by  $\bar{b}_{OLS}$  to calculate  $\hat{w}_t = m_t - x_t' \bar{b}_{OLS}$ ; (iii) square  $\hat{w}_t$  and run the classical least squares regression of  $\hat{w}_t^2$  on  $x_{t1}^2, \dots, x_{tK}^2$ ,<sup>5</sup> the sum of squares of the residuals of this regression gives an unrestricted error sum of squares ( $ESS_U$ ); and (iv) regress  $\hat{w}_t^2$  only on  $x_{t1}^2$ ; the sum of squares of the residuals of this regression gives a restricted error sum of squares ( $ESS_R$ ) because all the  $\sigma_{\epsilon_j}^2, j = 2, \dots, K$ , are assumed to be zero (i.e. it is restrictive because it imposes the restrictions implied by  $H_0$ ). Let  $q(\tilde{w}) = (T/(K-1))(ESS_R - ESS_U)/ESS_U$  be a test statistic. Reject  $H_0$  if the value  $q(w)$  of  $q(\tilde{w})$  obtained in a particular sample is greater than some critical value  $c$  and do not reject  $H_0$  otherwise.

Although the foregoing testing methodology is characteristic of the first-generation literature, this test is misleading under the usual circumstances. Probabilities of false rejection of  $H_0$  (Type I error) and false acceptance of  $H_1$  (Type II error) associated with this test are unknown because both the exact finite sample (or asymptotic) distributions of  $q(\tilde{w})$  when  $H_0$  and  $H_1$  are true are unknown. All that is known is that these probabilities are positive and less than 1. This is all that we need to show that the following argument is valid: Rejection of  $H_0$  is not proof that  $H_0$  is false or acceptance of  $H_0$  is not proof that  $H_0$  is true (Goldberger, 1991, p. 215). However, the occurrence of a real-world event does constitute strong evidence against  $H_0$  if that event has a small probability of occurring whenever  $H_0$  is true and a high probability of occurring whenever  $H_1$  is true. This

is an attractive definition of evidence, but finding such strong evidence is no easy task. To explain, we use the above test procedure. Note that there is no guarantee that one of  $H_0$  and  $H_1$  is true. When both  $H_0$  and  $H_1$  are false, observing a value of  $q(\tilde{w})$  that lies in the critical region  $\{q(\tilde{w}) > c\}$  is not equivalent to observing a real-world event because the critical region whose probability is calculated under  $H_0$  (or  $H_1$ ) to evaluate Type I error (or  $1 - \text{Type II error}$ ) probability is not the event of observing the actual data. Both  $H_0$  and  $H_1$  are false when, for example, the  $x_t$  are correlated with the  $\beta_t$  or when  $\varepsilon_t$  is not normal. We show in the next section that we need to assume that the  $x_t$  are correlated with the  $\beta_t$  if we want to make assumptions that are consistent with the real-world interpretations of the coefficients of equation (19.1). If our assumptions are inconsistent, then both  $H_0$  and  $H_1$  are false. In that event, finding the value  $q(w) > c$  should not be taken as strong evidence against  $H_0$  and for  $H_1$ , since the probabilities of the critical region  $\{q(\tilde{w}) > c\}$  calculated under  $H_0$  and  $H_1$  are incorrect. More generally, a test of a false null hypothesis against a false alternative hypothesis either rejects the false null hypothesis in favor of the false alternative hypothesis or accepts the false null hypothesis and rejects the false alternative hypothesis. Such tests continue to trap the unwary. We can never guarantee that one of  $H_0$  and  $H_1$  is true, particularly when the assumptions under which the test statistic  $q(\tilde{w})$  is derived are inconsistent with the real-world interpretations of the coefficients of equation (19.1). We explain in the next section why such inconsistencies arise.<sup>6</sup>

At the very minimum, the above argument should indicate how difficult it is to produce strong evidence against hypotheses of our interest. For this reason, de Finetti (1974a, p. 128) says, "accept or reject is the unhappy formulation which I consider as the principal cause of theogginess widespread all over the field of statistical inference and general reasoning." It is not possible to find useful approximations to reality by testing one false hypothesis against another false hypothesis. As discussed below, second-generation RCMs stress the importance of finding sufficient and logically consistent explanations of real phenomena (see, e.g., Zellner, 1988, p. 8) because of the limits to the usefulness of hypothesis testing.

In order to estimate the foregoing model, note that the error structure embedded in equation (19.4) is heteroskedastic. Specifically, the variance of the error at each sample point is a linear combination of the squares of the explanatory variables at that point (equation (19.5)). This suggests that this RCM can be estimated using a feasible generalized least squares estimation procedure that accounts for the heteroskedastic nature of the error process (Judge *et al.*, 1985, pp. 808–9).

In the case where  $t$  indexes time, a natural extension of this class of models is to incorporate serial correlation in the process determining the coefficients as follows: For  $t = 1, 2, \dots, T$ ,

$$(a) m_t = x_t' \beta_t, (b) \beta_t = \bar{\beta} + \varepsilon_t, (c) \varepsilon_t = \Phi \varepsilon_{t-1} + a_t. \quad (19.7)$$

This model differs from the previous model because it assumes that the process generating the coefficients is autoregressive, where  $\Phi$  is a matrix with eigenvalues less than 1 in absolute value.

This discussion above has presented the basic building blocks of first-generation RCMs. These models have been extended in a number of directions. For example, the stochastic structure determining the coefficients can be made to vary as a function of some observable variables. A main difficulty in using such models, however, is that one must specify the structure explaining coefficient variations. Errors in specification will lead to familiar and unfortunate consequences. Further, the structures are usually specified in a mechanical manner. Accordingly, we now discuss the class of second-generation RCMs, which directly confront these and other specification errors, and can be shown to include the first-generation RCMs and several well-known fixed-coefficient models as special cases (Swamy and Tavlas, 1995).

### 3 SECOND-GENERATION RCMs

As previously noted, second-generation RCMs are concerned with relaxing the usual restrictions concerning the direct effects of explanatory variables on the explained variables, functional forms, measurement errors, and use of an additive error term to proxy excluded variables. If these restrictions are violated – for example, if there are measurement errors when no measurement errors are assumed or a specified functional form is incorrect – the resulting estimates are subject to specification errors. In two path-breaking papers, Pratt and Schlaifer (1984, 1988) have demonstrated that, in order to assess the plausibility of these restrictions, we need “real-world” interpretations of the coefficients in equation (19.1). In essence, two questions need to be addressed: (i) what are these real-world interpretations? (ii) are the above restrictions consistent with these interpretations? In what follows we show that any equation relating observable variables cannot coincide with the corresponding true economic relationship if each of its coefficients is not treated as the sum of three parts – one corresponding to a direct effect of the true value of a regressor on the true value of the dependent variable, a second part capturing omitted-variable biases, and a third part capturing the effect of mismeasuring the regressor. We also show that the true functional form of the equation is a member of a class of functional forms.

To explain, consider the following model of money demand:

$$m_t = \gamma_0 + \gamma_1 r_t + \gamma_2 y_t + u_t, \quad (19.8)$$

where  $r_t$  is the logarithm of an interest rate (i.e. opportunity cost) variable,  $y_t$  is the logarithm of real income, and  $u_t$  is an error term. It is assumed that the  $\gamma$ s are constant and  $u_t$  has mean zero and is mean independent of  $r_t$  and  $y_t$ . As before,  $m_t$  is the logarithm of real money balances. Note that the variables in equation (19.8) are observed values. Because of measurement errors, they are unlikely to represent true values. Also, equation (19.8) is unlikely to coincide with the “true” money demand equation, as we now show.

Consider the implications arising from a typical errors-in-the-variables model. Thus, suppose that  $m_t = m_t^* + v_{0t}$ ,  $r_t = r_t^* + v_{1t}$ , and  $y_t = y_t^* + v_{2t}$ , where  $m_t$ ,  $r_t$ , and  $y_t$  are the observed values of the variables, the variables with an asterisk represent

the true values, and the *vs* represent the errors made in the measurement of the variables. For example,  $y_t^*$  could be permanent income and  $r_t^*$  could be the opportunity cost of holding money that is implied by the definition of permanent income. One consequence of using the observed rather than the true values is that a set of random variables – the *vs* – is incorporated into the equation determining  $m_t$ . Additionally, we show below that the existence of measurement errors contradicts the assumption that the coefficients of equation (19.8) can be constants.

For the time being, suppose we are fortunate enough to know the true values of the variables. In that case, the true functional form of the money demand equation is a member of the class:

$$m_t^* = \alpha_{0t} + \alpha_{1t}r_t^* + \alpha_{2t}y_t^* + \sum_{j=3}^{n_t} \alpha_{jt}x_{jt}^* \quad (t = 1, \dots, T), \quad (19.9)$$

where the  $x_{jt}^*$  are all the determinants of  $m_t^*$  other than  $r_t^*$  and  $y_t^*$  and where the  $\alpha$ s and  $n_t$  are coefficients and the number of explanatory variables, respectively; they are time-varying, as indicated by their time subscripts. The variables  $r_t^*$  and  $y_t^*$  are called the included variables and the variables  $x_{jt}^*$  are called the excluded variables. Temporal changes in the set of excluded variables change  $n$  over time. Equation (19.9) is not necessarily linear, since the time-varying quality of the coefficients permits the equation to pass through every data point even when the number of observations on its variables exceeds  $n_t + 1$ . Thus, with time-varying coefficients the equation can be nonlinear. Equation (19.9) will have different functional forms for different sequences of its coefficients (or for different paths of variation in its coefficients) and the class of functional forms it represents is unrestricted as long as its coefficients are unrestricted. This lends support to our speculation that a member of this class is true. Equation (19.9) with unrestricted coefficients is more general than even the true money demand function,  $m_t^* = f(r_t^*, y_t^*, x_{3t}^*, \dots, x_{n_t,t}^*)$ , whose functional form is unknown. However, for a certain (unknown) pattern of variation in its coefficients, (19.9) coincides with the true equation. The essential feature of equation (19.9) is that, since it encompasses time-varying coefficients, the true values of variables, the correct functional form, and any omitted variables, it covers the true function determining money demand as a special case. The  $\alpha$  coefficients that follow the true pattern of variation would represent the true elasticities of the determinants of money demand. We call  $\alpha_{1t}$  and  $\alpha_{2t}$  (representing the true pattern of variation) the direct effects on  $m_t^*$  of  $r_t^*$  and  $y_t^*$ , respectively, since they are not contaminated by any of the four specification errors discussed above.

Two fundamental problems are involved in any attempt to estimate equation (19.9). First, we may not know much (if anything) about the  $x_{jt}^*$ . Whatever we know may not be enough to prove that they are uncorrelated with the other explanatory variables in equation (19.9) (Pratt and Schlaifer, 1984, pp. 11–12). Second, the observed (as opposed to the true) values of the variables, are likely to contain measurement errors. A way to resolve the former problem is to assume that the  $x_{jt}^*$  are correlated with the other explanatory variables as follows:

$$x_{jt}^* = \Psi_{0jt} + \Psi_{1jt}r_t^* + \Psi_{2jt}y_t^* \quad (j = 3, \dots, n). \quad (19.10)$$

The coefficient  $\psi_{0jt}$  has a straightforward interpretation. It is the portion of  $x_{jt}^*$  (i.e. an excluded variable) remaining after the effects of the variables  $r_t^*$  and  $y_t^*$  have been removed. The remaining  $\psi$  coefficients represent the partial correlations between the excluded and the included variables. As with equation (19.9), the coefficients of equation (19.10) will not be constants unless this equation is known with certainty to be linear.

In order to take account of the correlations among the  $r_t^*$  and  $y_t^*$  and the  $x_{jt}^*$ , substitute equation (19.10) into equation (19.9):

$$m_t^* = \left( \alpha_{0t} + \sum_{j=3}^{n_t} \alpha_{jt} \psi_{0jt} \right) + \left( \alpha_{1t} + \sum_{j=3}^{n_t} \alpha_{jt} \psi_{1jt} \right) r_t^* + \left( \alpha_{2t} + \sum_{j=3}^{n_t} \alpha_{jt} \psi_{2jt} \right) y_t^*. \quad (19.11)$$

Equation (19.11) expresses the time-varying relationship between the true values of the variables, where the time-varying effects of the excluded variables (i.e. the  $x_{jt}^*$ ) are included in each coefficient on the right-hand side. For example, the remaining portion of  $x_{jt}^*$  after the effects of the variables  $r_t^*$  and  $y_t^*$  have been removed is captured in the first coefficient (via  $\psi_{0jt}$ ), while the effects of  $r_t^*$  on the  $x_{jt}^*$  are captured in the second term (via  $\psi_{1jt}$ ).

Equation (19.11) involves a relationship between the true variables, which are unobservable. Accordingly, to bring us closer to estimation substitute the observable counterparts of these variables into equation (19.11) to obtain:

$$m_t = \gamma_{0t} + \gamma_{1t} r_t + \gamma_{2t} y_t, \quad (19.12)$$

where  $\gamma_{0t} = (\alpha_{0t} + \sum_{j=3}^{n_t} \alpha_{jt} \psi_{0jt} + v_{0t})$ ,  $\gamma_{1t} = (\alpha_{1t} + \sum_{j=3}^{n_t} \alpha_{jt} \psi_{1jt}) (1 - \frac{v_{1t}}{r_t})$ , and  $\gamma_{2t} = (\alpha_{2t} + \sum_{j=3}^{n_t} \alpha_{jt} \psi_{2jt}) (1 - \frac{v_{2t}}{y_t})$ .

The coefficients of equation (19.12) have straightforward real-world interpretations corresponding to the direct effect of each variable and the effects of omitted variables and measurement errors. Consider, for example, the coefficient  $\gamma_{1t}$  on  $r_t$ . It consists of three parts: a direct effect,  $\alpha_{1t}$ , of the true interest rate ( $r_t^*$ ) on the true value of real money balances ( $m_t^*$ ) given by equation (19.9); a term ( $\sum_{j=3}^{n_t} \alpha_{jt} \psi_{1jt}$ ) capturing an indirect effect or omitted-variable bias (recall,  $\alpha_{jt}$  with  $j \geq 3$  is the effect of an omitted variable on  $m_t^*$ , and  $\psi_{1jt}$  is the effect of  $r_t^*$  on that omitted variable); and a term capturing the effect of measurement error,  $-(\alpha_{1t} + \sum_{j=3}^{n_t} \alpha_{jt} \psi_{1jt})(v_{1t}/r_t)$  (recall that  $v_{1t}$  is the measurement error associated with the interest rate). The coefficient  $\gamma_{2t}$  can be interpreted analogously. The direct effects provide economic explanations. The term  $\gamma_{0t}$  also consists of three parts and these include the intercepts of equations (19.9) and (19.10), the effects of omitted variables on  $m_t^*$ , and the measurement error in  $m_t$ . It is the connection between  $\gamma_{0t}$  and the intercepts of equations (19.9) and (19.10) that demonstrates the real-world origin of  $\gamma_{0t}$ . In other words, all the coefficients in equation (19.12) have been derived on the basis of a set of realistic assumptions which directly confront the problems that arise because of omitted explanatory variables, their correlations with the included explanatory variables, measurement errors, and unknown functional

forms. When these problems are present, as they usually are, a necessary condition for equation (19.12) to coincide with the true money demand function is that its coefficients are the sums of three parts stated below equation (19.12). At least two of these parts (omitted-variable biases and measurement error effects) cannot be constant and hence it may not be reasonable to assume that the constant coefficients of equation (19.8) are the sums of these three parts. Thus, equation (19.8)'s premises are inconsistent with the real-world interpretations of the coefficients of equation (19.12), and equation (19.8) cannot coincide with the true money demand function. These results are false if (i)  $v_{1t}$  and  $v_{2t}$  are equal to zero for all  $t$  (i.e. there are no measurement errors in  $r_t$  and  $y_t$ ), (ii)  $\psi_{1jt}$  and  $\psi_{2jt}$  are equal to zero for all  $j \geq 3$  and  $t$  (i.e. the included variables are independent of excluded variables), (iii)  $\alpha_{1t}$  and  $\alpha_{2t}$  are constant (i.e. the direct effects of  $r_t^*$  and  $y_t^*$  on  $m_t^*$  are constant), and (iv)  $\gamma_{0t} = \gamma_0 + u_t$  (i.e. the intercept of equation (19.12) is equal to the intercept plus the error of equation (19.8)). Though under these conditions, equations (19.8) and (19.12) coincide with equation (19.9) and no inconsistencies arise, the difficulty is that these conditions are shown to be false by Pratt and Schlaifer (1984, pp. 11–12).

Equation (19.12) may be correct in theory, but we need to implement it empirically. Ideally, we would like to have empirical estimates of the direct effects, but as shown above, the direct effects are commingled with mismeasurement effects and omitted-variable biases. It should also be observed that equation (19.12) is more complicated than a structural equation without exogenous variables since  $\gamma_{0t}$ ,  $\gamma_{1t}$ , and  $\gamma_{2t}$  are correlated both with each other and with the variables  $r_t$  and  $y_t$ . These correlations arise because  $\gamma_{1t}$  and  $\gamma_{2t}$  are functions of  $r_t$  and  $y_t$ , respectively, and  $\gamma_{0t}$ ,  $\gamma_{1t}$ , and  $\gamma_{2t}$  have a common source of variation in  $\alpha_{jt}$ ,  $j = 3, \dots, n_t$ . Instrumental variable estimation (IVE) – intended to deal with the problem of correlations between  $\gamma_{0t}$  and  $r_t$  and  $y_t$  when  $\gamma_{1t}$  and  $\gamma_{2t}$  are constant – of equation (19.12) does not “purge” its coefficients of mismeasurement effects and omitted-variable biases and, hence, cannot be used. IVE is designed neither to decompose the  $\gamma$ s into direct, indirect, and mismeasurement effects nor to deal with the correlations between the included explanatory variables and their coefficients.

In an attempt to estimate  $\alpha_{1t}$  and  $\alpha_{2t}$ , we need to introduce some additional terminology.<sup>7</sup> To derive estimates of  $\alpha_{1t}$  and  $\alpha_{2t}$ , we will attempt to estimate the  $\gamma$ s using concomitants. A formal definition of concomitants is provided in footnote 7. Intuitively, these may be viewed as variables that are not included in the equation used to estimate money demand, but help deal with the correlations between the  $\gamma$ s and the explanatory variables (in this example, interest rates and real income). This notion can be stated more precisely in the form of the following two assumptions:

**Assumption 1.** The coefficients of equation (19.12) satisfy the stochastic equations

$$\gamma_{kt} = \pi_{k0} + \sum_{j=1}^p \pi_{kj} z_{jt} + \varepsilon_{kt} \quad (k = 0, 1, 2), \quad (19.13)$$

where the concomitants  $z_{jt}$  explain the variation in the  $\gamma_{kt}$ ,  $E(\varepsilon_{kt} | z_t) = E(\varepsilon_{kt}) = 0$  for all  $t$  and each  $k$ , and the  $\varepsilon_{kt}$  satisfy the stochastic equation

$$\varepsilon_{kt} = \varphi_{kk}\varepsilon_{k,t-1} + a_{kt}, \quad (19.14)$$

where for  $k, k' = 0, 1, 2, -1 < \varphi_{kk} < 1$ , and  $a_{kt}$  are serially uncorrelated with  $E(a_{kt}) = 0$  and  $Ea_{kt}a_{k't} = \sigma_{kk'}$  for all  $t$ .

**Assumption 2.** The explanatory variables of equation (19.12) are independent of the  $\varepsilon_{kt}$ , given any values of the concomitants  $z_{jt}$ , and condition (iii) of footnote 7 holds.

Equation (19.14) is not needed if  $t$  indexes individuals and is needed if  $t$  indexes time and if the  $\varepsilon_{kt}$  in equation (19.13) are partly predictable. It can also be assumed that  $\varepsilon_{kt}$  and  $\varepsilon_{k',t-1}$  with  $k \neq k'$  are correlated. The explanatory variables of equation (19.12) can be independent of their coefficients conditional on a given value of the  $z$ s even though they are not unconditionally independent of their coefficients. This property provides a useful procedure for consistently estimating the direct effects contained in the coefficients of equation (19.12). The criticism of Assumption 1 contained in the last paragraph of Section 2 (or the errors in the specification of equation (19.13)) can be avoided by following the criteria laid out in Section 4.

To illustrate the procedure, suppose a money demand specification includes two explanatory variables – real income and a short-term interest rate. Also, suppose two concomitants (so that  $p = 2$ ) are used to estimate the  $\gamma$ s – a long-term interest rate (denoted as  $z_{1t}$ ) and the inflation rate (denoted as  $z_{2t}$ ). A straightforward interpretation of the use of these concomitants is the following. The direct effect ( $\alpha_{1t}$ ) component of the coefficient (i.e.  $\gamma_{1t}$ ) on the short-term interest-rate variable  $r_t$  in equation (19.12) is represented by the linear function  $(\pi_{10} + \pi_{11}z_{1t})$  of the long-term rate. The indirect and mismeasurement effects are captured by using a function  $(\pi_{12}z_{2t} + \varepsilon_{1t})$  of the inflation rate and  $\varepsilon_{1t}$ . In this example, the measure of the direct effects ( $\alpha_{2t}$ ) contained in  $\gamma_{2t}$  (the coefficient on real income) is represented in  $(\pi_{20} + \pi_{21}z_{1t})$ ; the measure of indirect and mismeasurement effects contained in  $\gamma_{2t}$  is represented in  $(\pi_{22}z_{2t} + \varepsilon_{2t})$ . These definitions do not impose any zero restrictions, but may need to be extended (see Section 4).

Substituting equation (19.13) into equation (19.12) gives an equation in estimable form:

$$\begin{aligned} m_t = & \pi_{00} + \sum_{j=1}^p \pi_{0j}z_{jt} + \pi_{10}r_t + \sum_{j=1}^p \pi_{1j}z_{jt}r_t + \pi_{20}y_t \\ & + \sum_{j=1}^p \pi_{2j}z_{jt}y_t + \varepsilon_{0t} + \varepsilon_{1t}r_t + \varepsilon_{2t}y_t \quad (t = 1, 2, \dots, T). \end{aligned} \quad (19.15)$$

A computer program developed by Chang, Swamy, Hallahan and Tavlas (1999) can be used to estimate this equation. Note that equation (19.15) has three error terms, two of which are the products of  $\varepsilon$ s and the included explanatory variables

of equation (19.8). The sum of these three terms is both heteroskedastic and serially correlated. Under Assumptions 1 and 2, the right-hand side of equation (19.15) with the last three terms suppressed gives the conditional expectation of the left-hand side variable as a nonlinear function of the conditioning variables. This conditional expectation is different from the right-hand side of equation (19.8) with  $u_t$  suppressed. This result demonstrates why the addition of a single error term to a mathematical formula and the exclusion of the interaction terms on the right-hand side of equation (19.15) introduce inconsistencies in the usual situations where measurement errors and omitted-variable biases are present and the true functional forms are unknown.

In these usual situations, equation (19.8) can be freed of its inconsistencies by changing it to equation (19.12) and making Assumptions 1 and 2. A similar approach does not work for probit and logit models which are also based on assumptions that are inconsistent with the real-world interpretations of their coefficients. As regards switching regressions, Swamy and Mehta (1975) show that these regressions do not approximate the underlying true economic relationships better than random coefficient models.

Note the validity of our above remarks regarding IVE. There cannot be any instrumental variables that are uncorrelated with the error term of equation (19.15) and highly correlated with the explanatory variables of equation (19.12) because these explanatory variables also appear in the error term.

Second-generation RCMs have been applied in recent years to a wide variety of circumstances and with much success in terms of forecasting performance relative to models of the type in equation (19.8) (Akhavein, Swamy, Taubman, and Singamsetti, 1997; Leusner, Akhavein, and Swamy, 1998; Phillips and Swamy, 1998; and Hondroyiannis, Swamy, and Tavlas, 1999).

#### **4 CRITERIA FOR CHOOSING CONCOMITANTS IN RCMs**

Equations (19.9)–(19.12) incorporate in a consistent way all the prior information that is usually available about these equations. The most difficult step arises in the form of equation (19.13). Not much prior information is available about the proper concomitants that satisfy Assumptions 1 and 2. As a minimum exercise of caution, the applied econometrician who approaches the problem of estimating equation (19.15) should choose among various sets of concomitants after carefully examining their implications for the estimates of the direct effect components of the coefficients of equation (19.12). Different models of the form (19.15) are obtained by including different sets of concomitants in equation (19.13). The question we address in this section is the following: how can we validate these different models? In what follows, we briefly describe a set of validation criteria and relate them to the RCM described above.

A money demand model can be considered to be validated if (i) it fits within-sample values well; (ii) it fits out-of-sample values well; (iii) it has high explanatory power; (iv) it is derived from equation (19.12) by making assumptions that are consistent with the real-world interpretations of the coefficients of equation (19.12); (v) the signs and statistical significance of the estimates of direct effects remain

virtually unchanged as one set of concomitants (other than the determinants of direct effects) after another is introduced into equation (19.13).

*Condition (i)* is used in almost all econometric work as a measure of fitted-model adequacy. The definition of the coefficient of determination ( $R^2$ ) that is appropriate to a regression equation with nonspherical disturbances can be applied to equation (19.15). Such a definition is given in Judge *et al.* (1985, p. 32). The coefficient is a measure of the proportion of weighted variation in  $m_t$ ,  $t = 1, 2, \dots, T$ , explained by the estimated equation (19.15). Within the RCM framework, a low  $R^2$  implies that the set of concomitants included in equation (19.13) together with the explanatory variables of equation (19.12) do not adequately explain the weighted variation in  $m_t$ ,  $t = 1, 2, \dots, T$ . The problem with  $R^2$ , however, is that a high value can result by arbitrarily increasing the number of concomitants in equation (19.13), even if all these concomitants are not relevant for explaining the coefficients of equation (19.12).

*Condition (ii)* is based on cross validation, which is used to assess the ability of equation (19.15) to predict out-of-sample values of  $m_t$ . In this procedure, the data sample is divided into two subsamples. The choice of a model with a set of concomitants, including any necessary estimation, is based on one subsample and then its performance is assessed by measuring its prediction against the other subsample. This method is related to Stone's (1974) cross-validatory choice of statistical predictions.

The premise of this approach is that the validity of statistical estimates should be judged by data different from those used to derive the estimates (Mosteller and Tukey, 1977, pp. 36–40; Friedman and Schwartz, 1991, p. 47). Underlying this approach is the view that formal hypothesis tests of a model on the data that are used to choose its numerical coefficients are almost certain to overestimate performance. Also, statistical tests lead to false models with probability 1 if both the null and alternative hypotheses considered for these tests are false, as we have already shown in Section 2. This problem can arise in the present case because of the lack of any guarantee that either a null or an alternative hypothesis will be true if inconsistent restrictions are imposed on equation (19.9).

Predictive testing – extrapolation to data outside the sample – also has its limitations. All forecasts and forecast comparisons should take into account the result, due to Oakes (1985), that there is no universal algorithm to guarantee accurate forecasts forever. This result implies that equation (19.15) with a single set of concomitants cannot predict  $m_t$  well in all future periods. This is especially true when the set of  $x_{jt}^*$ s in equation (19.9) changes over time. Also, past success does not guarantee future success. That is, if all we knew about equation (19.15) was that it had produced accurate forecasts in the past, there would be no way we could guarantee that future forecasts of equation (19.15) would be sufficiently accurate, since there are some sets of concomitants (e.g. dummy (or shift) variables that are appropriate for a past period) for which past values do not control future values. Even false models based on contradictory premises can sometimes predict their respective dependent variables well. To satisfy a necessary condition under which models are true, de Finetti (1974b) sets up minimal coherence criteria that forecasts should satisfy based on data currently available. By these

criteria, different forecasts are equally valid *now* if they all satisfy the requirements for coherence, given currently available knowledge. Thus, a forecast from equation (19.15) can at best represent a measure of the confidence with which one expects that equation to predict an event in the future, based on currently available evidence and not on information yet to be observed, provided that the forecast satisfies the requirements for coherence.<sup>8</sup>

To choose models that satisfy de Finetti's criteria of coherence, we impose the additional conditions (iii)–(v) on RCMs. As with de Finetti's concept of coherence, condition (iv) also explicitly prohibits the use of contradictory premises.<sup>9</sup> Together, conditions (i)–(v) provide an improved method of model validation.

*Condition (iii)* has also been advocated by Zellner (1988). If prediction were the only criterion of interest, there would be no need to separate direct effects from indirect and mismeasurement effects. But if we are interested in economic explanations – for example, a transmission mechanism of a particular policy action – we need to separate these effects. Equation (19.9) will have the highest explanatory power whenever it coincides with the true money demand function. It would be fortunate if  $\sum_{j=3}^{n_t} \alpha_{jt} \Psi_{0jt}$  were offset exactly by  $v_{0t}$  and if  $\sum_{j=3}^{n_t} \alpha_{jt} \Psi_{1jt}$  (or  $\sum_{j=3}^{n_t} \alpha_{jt} \Psi_{2jt}$ ) and  $-(\alpha_{1t} + \sum_{j=3}^{n_t} \alpha_{jt} \Psi_{1jt}) \frac{v_{1t}}{r_t}$  (or  $-(\alpha_{2t} + \sum_{j=3}^{n_t} \alpha_{jt} \Psi_{2jt}) \frac{v_{2t}}{y_t}$ ) canceled each other. In this case,  $\gamma_{0t} = \alpha_{0t}$ ,  $\gamma_{1t} = \alpha_{1t}$ ,  $\gamma_{2t} = \alpha_{2t}$ , and equation (19.12) has the same explanatory power as equation (19.9). Alternatively, when  $\gamma_{1t} \neq \alpha_{1t}$  and  $\gamma_{2t} \neq \alpha_{2t}$ , equation (19.12) explains well if it is closer to the true money demand function than to any other equation and cannot provide the proper explanations otherwise. To discern which one of these cases is true given the data on  $m_t$ ,  $r_t$ , and  $y_t$ , an accurate means for separating  $\alpha_{1t}$  and  $\alpha_{2t}$  from the other terms of  $\gamma_{1t}$  and  $\gamma_{2t}$  is needed. Equation (19.15) attempts to make such a separation.

*Condition (iv)* conforms to de Finetti's (1974b) requirement of coherence – namely, that statistical analysis applied to data should not violate probability laws. We apply this requirement in a somewhat different manner. To explain, consider the following example. Equation (19.12) represents a particular economic (i.e. money demand) relationship that we have in mind but cannot estimate. What we can estimate is equation (19.15). Underlying equation (19.15) are equation (19.12) and Assumptions 1 and 2. Thus, estimation requires that Assumptions 1 and 2 are consistent with the real-world interpretations of the coefficients of equation (19.12) so that de Finetti's condition is satisfied. Assumptions 1 and 2 are consistent with the real-world interpretations of the coefficients of equation (19.12) if the concomitants included in equation (19.13) satisfy Assumptions 1 and 2. For example, Assumptions 1 and 2 are satisfied if the correlations between  $(\gamma_{0t}, \gamma_{1t}, \gamma_{2t})$  and  $(r_t, y_t)$  arise because of their dependence on a common third set of variables,  $z_{jt}$ ,  $j = 1, \dots, p$ , and if these  $z$ s together with  $\epsilon_{k,t-1}$ ,  $k = 0, 1, 2$ , capture all the variation in  $\gamma_{1t}$  and  $\gamma_{2t}$  and almost all the variation in  $\gamma_{0t}$ . An example of forecasts that do not satisfy de Finetti's requirement of coherence is a forecast of  $m_t$  from equation (19.8), since the premises of this equation are inconsistent with the real-world interpretations of the coefficients of equation (19.12).

*Condition (v)* concerns the sensitivity of the signs and magnitudes of direct effects to changes in the set of concomitants. Following Pratt and Schlaifer (1988,

p. 45), we state that the only convincing evidence that equation (19.12) under Assumptions 1 and 2 coincides with the true money demand function is of the following kind. It is found that the signs and statistical significance of the estimates of  $\alpha_{1t}$  and  $\alpha_{2t}$  remain virtually unchanged as one set of concomitants (other than those determining  $\alpha_{1t}$  and  $\alpha_{2t}$ ) after another is introduced into equation (19.13), until finally it is easier to believe that  $r_t^*$  and  $y_t^*$  have the effects they seem to have on  $m_t^*$  than to believe that they are merely the proxies for some other, as yet undiscovered, variable or variables.

## 5 AN EMPIRICAL EXAMPLE

In this section, we use annual UK data to estimate the demand for money (i.e. equation (19.12) extended to include one additional explanatory variable), with and without concomitants, over the long period 1881–1990. The algebraic forms of these two models are given by the sets (19.12)–(19.14) and (19.7) of equations, respectively. Post-sample forecasts are generated over the period 1991–95. The dependent variable is the log of M3 (currency held by the public plus gross deposits at London and country joint stock and private banks divided by the implicit price deflator for net national product). One of the regressors is the log of per capita net national income, deflated by the implicit price deflator. The antilog of  $r_t$  is a short-term rate – the rate on three month bank bills. Following Friedman and Schwartz (1982), and others, we use the rate of change of nominal income as a regressor to proxy the nominal yield on physical assets – that is, as an additional opportunity cost variable. Two concomitants are used: (i) a long-term interest rate – the annual yield on consols; and (ii) the inflation rate, as measured by the rate of change of the implicit price deflator. All data are from Friedman and Schwartz (1982) and have been updated by the present authors.<sup>10</sup>

Table 19.1 presents the results.<sup>11</sup> RCM1 and RCM2 denote the above extended equation (19.12) without and with concomitants, respectively. The coefficient estimates are the average values of the individual time-varying coefficients. Point estimates of the elasticities of income and the interest rate in Table 19.1, but not their  $t$ -ratios, are within the range of those yielded in previous empirical studies of UK money demand (see, e.g. Hondroyannis *et al.*, 1999). Also, both RCM1 and RCM2 produce low root mean square errors (RMSEs). In this example, the equation without concomitants yields a lower RMSE over the post-sample period than does the equation with concomitants. An explanation of this result is that over the range of the values of its dependent and independent variables for the period 1991–95 the money demand function without concomitants seems to approximate the true money demand function better than the money demand function with concomitants. The specifications were also used to provide forecasts over various decades beginning with the 1930s. For this purpose, each specification was re-estimated using data prior to the decade for which it was used to forecast. The equation with concomitants produced lower RMSEs in four out of the six decades. For the sake of brevity, these results are not reported but are available from the authors.

**Table 19.1** Long-run elasticities

	RCM1	RCM2
Intercept	-3.69 (-4.97)	-3.19 (-4.03)
Short-term interest rate	-0.04 (-2.86)	-0.01 (-0.51)
Real per capita income	0.74 (6.03)	0.67 (4.98)
Nominal income growth	-0.41 (-4.15)	-0.39 (-3.09)
RMSE	0.020	0.041

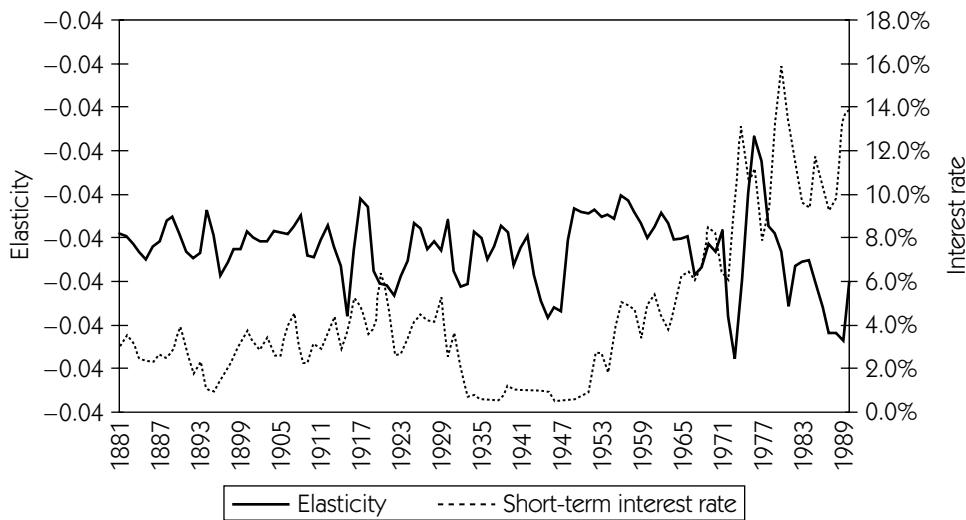
Estimation period is 1881–1990. Forecast period is 1991–95. Figures in parentheses are *t*-ratios.

**Table 19.2** Long-run elasticities and direct effects from RCM2

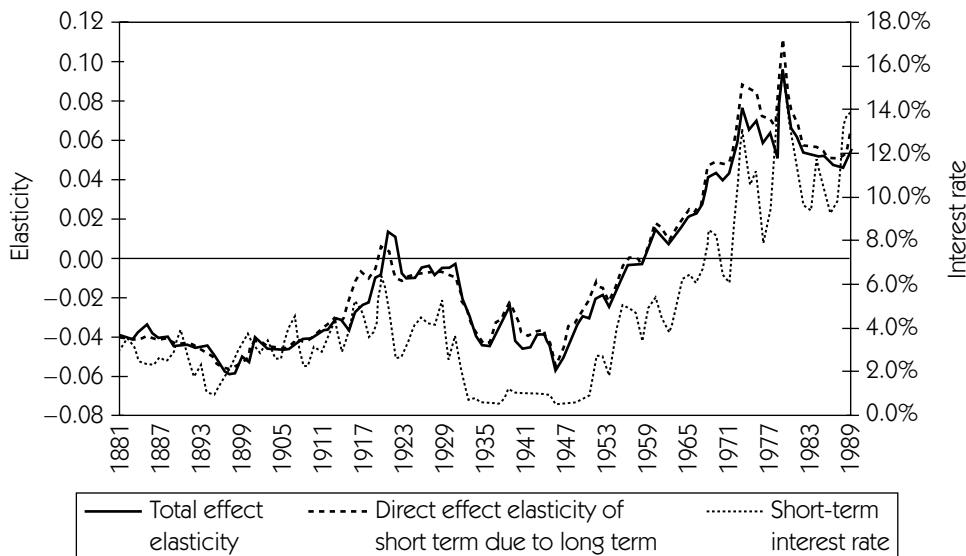
<i>Intercept</i>	<i>Interest rate</i>		<i>Real per capita income</i>		<i>Nominal income growth</i>	
	$\gamma_{0t}$	$\gamma_{1t}$	<i>Direct effect</i>	$\gamma_{2t}$	<i>Direct effect</i>	$\gamma_{3t}$
-3.19 (-4.0)	-0.01 (-0.5)	-0.007 (-0.4)	0.67 (5.0)	0.68 (5.1)	-0.39 (-3.1)	-0.42 (-3.1)

Figures in parentheses are *t*-ratios.

Table 19.2 reports the averages of the total and direct effect components of the coefficients from the equation with concomitants (i.e. RCM2). Recall, since this specification includes two concomitants, it is possible to extract the direct effects from the total effects, as shown in Section 3. (For the other specification, the indirect and mismeasurement effects in each coefficient are captured in the corresponding error term.) As shown in the table, the  $\gamma$  (i.e. total) coefficients and the direct-effect coefficients are very close to each other for all variables, as are the corresponding *t*-ratios. Figures 19.1 and 19.2 show the time profiles of the elasticity of the short-term interest rate in the absence of concomitants and with concomitants, respectively. The figures also include the time profile of the short-term interest rate. Without concomitants, the interest rate coefficients vary within extremely narrow ranges around their average values. To be exact, the elasticity of the short rate varies between -0.0422 (in 1976) and -0.0431 (in 1973). The narrow range of the interest rate elasticities is due to the specification of the coefficients



**Figure 19.1** Short-term interest rate elasticity for RCM1 (without concomitants)



**Figure 19.2** Short-term interest rate elasticity for RCM2 (with concomitants)

in the absence of concomitants. Without concomitants the coefficients are equal to a constant mean plus an error term. If the error term has a small variance and does not exhibit serial correlation, the coefficient itself will not vary very much. As shown in Figure 19.2, the coefficients with concomitants exhibit a wider variation than is the case in the specification estimated without concomitants.

The RCM procedure also provides the coefficients of the other regressors. To save space, we report only the time profiles of the interest rate elasticity to give a

flavor of the results obtainable from the procedure. Obviously, a richer specification of concomitants might well have provided different results. Examples of the use of varying combinations of concomitants can be found in the papers cited at the end of Section 3.

## 6 CONCLUSIONS

This chapter has attempted to provide a basic introduction to the rationale underlying RCMs. The focus has been on what we characterized as second-generation RCMs, which have been developed to deal with four main problems frequently faced by researchers in applied econometrics. These second-generation RCMs aim to satisfy the conditions for observability of stochastic laws. The use of concomitants – variables that are not included as regressors in the economic relationship to be estimated, but which help deal with correlations between the explanatory variables and their coefficients – allows estimation of both direct and total effects. A set of model validation criteria have also been presented that can be used to discriminate among models and these criteria have been applied to RCMs.

### Notes

- \* Views expressed in this chapter are those of the authors and do not necessarily reflect those of the Office of the Comptroller of the Currency, the Department of the Treasury, International Monetary Fund, or the Bank of Greece. We are grateful to Badi Baltagi for encouragement and guidance, and to four anonymous referees for helpful comments.
- 1 The suggestion that the coefficients of regression could be random was also made by Klein (1953).
- 2 An exposition of first-generation RCMs at text-book level has been provided by Judge, Griffiths, Hill, Lütkepohl and Lee (1985, chs 11, 13, and 19). See, also, Chow (1984) and Nicholls and Pagan (1985).
- 3 The discussion of the real-world sources and interpretations of  $\beta_t$  and their implications for its distribution is postponed until the next section.
- 4 We start with this assumption and, in the next section, detect departures from it that are warranted by the real-world sources and interpretations of the coefficients of equation (19.1).
- 5 For a derivation of  $E(\hat{w}_t^2 | x_t)$ , see Judge *et al.* (1985, p. 435).
- 6 The above argument is valid even when  $q(\hat{w})$  is replaced by the Breusch and Pagan (1979) or Judge *et al.* (1985, p. 436) test statistic.
- 7 We can write  $p(m_t, r_t, y_t | z_t, \theta) = p(m_t | r_t, y_t, z_t, \theta_1)p(r_t, y_t | z_t, \theta_2)$ , where  $p(\cdot)$  is a probability density function,  $z_t$  is a vector of concomitants, and  $\theta$ ,  $\theta_1$ , and  $\theta_2$  are the vectors of fixed parameters. Since  $\gamma_{0t}$ ,  $\gamma_{1t}$ ,  $\gamma_{2t}$ ,  $r_t$ , and  $y_t$  are correlated with one another, the inferences about  $\alpha_{1t}$  and  $\alpha_{2t}$  can be drawn without violating probability laws by using the density  $p(\gamma_{0t} + \gamma_{1t}r_t + \gamma_{2t}y_t | r_t, y_t, z_t, \theta_1)$  if the following three conditions are satisfied: (i)  $\theta_1$  and  $\theta_2$  are independent – a good discussion of parameter independence is given in Basu (1977); (ii)  $(\gamma_{0t}, \gamma_{1t}, \gamma_{2t})$  are independent of  $(r_t, y_t)$ , given a value of  $z_t$  – a good discussion of conditional independence is given in Dawid (1979, pp. 3–4); (iii)  $\text{pr}(r_t \in S_r, y_t \in S_y | z_t, m_t) = \text{pr}(r_t \in S_r, y_t \in S_y | z_t)$ , where the symbol  $\text{pr}$  is short-hand for probability, and  $S_r$  and  $S_y$  are the intervals containing the realized values of

$r_t$  and  $y_t$ , respectively, to which the observed values of  $m_t$  are connected by a law – a definition of stochastic law is given in Pratt and Schlaifer (1988). When condition (ii) holds,  $p(\gamma_{0t}, \gamma_{1t}, \gamma_{2t}, r_t, y_t | z_t, \pi) = p(\gamma_{0t}, \gamma_{1t}, \gamma_{2t} | z_t, \pi_1)p(r_t, y_t | z_t, \theta_2)$ . This equation provides a formal definition of concomitants.

- 8 For further discussion of these points, see Schervish (1985).
- 9 With an inconsistently formulated model, even use of Bayesian methods will probably lead to incoherent forecasts (Pratt and Schlaifer, 1988, p. 49).
- 10 The data and their sources are available from the authors.
- 11 The  $t$ -ratios were computed by taking account of the correlations among the coefficients.

## References

- Akhavein, J.D., P.A.V.B. Swamy, S.B. Taubman, and R.N. Singamsetti (1997). A general method of deriving the inefficiencies of banks from a profit function. *Journal of Productivity Analysis* 8, 71–94.
- Basu, D. (1977). On the elimination of nuisance parameters. *Journal of the American Statistical Association* 72, 355–66.
- Breusch, T.S., and A.R. Pagan (1979). A simple test for heteroscedasticity and random coefficient variation. *Econometrica* 47, 1287–94.
- Chang, I., P.A.V.B. Swamy, C. Hallahan, and G.S. Tavlas (1999). A computational approach to finding causal economic laws. *Computational Economics* forthcoming.
- Chow, G.C. (1984). Random and changing coefficient models. In Z. Griliches and M.D. Intrilligator (eds.) *Handbook of Econometrics*, Volume 2, Amsterdam: North-Holland Publishing Company.
- Dawid, A.P. (1979). Conditional independence in statistical theory. *Journal of the Royal Statistical Society B*, 41, 1–31.
- de Finetti, B. (1974a). Bayesianism. *International Statistical Review* 42, 117–30.
- de Finetti, B. (1974b). *The Theory of Probability*, Volume 1. New York: John Wiley & Sons.
- Friedman, M., and A.J. Schwartz (1982). *Monetary Trends in the United States and the United Kingdom: Their Relation to Income, Prices, and Interest Rates, 1867–1975*. Chicago: University of Chicago Press.
- Friedman, M., and A.J. Schwartz (1991). Alternative approaches to analyzing economic data. *American Economic Review* 81, 39–49.
- Goldberger, A.S. (1991). *A Course in Econometrics*. Cambridge, MA: Harvard University Press.
- Hondroyannis, G., P.A.V.B. Swamy, and G.S. Tavlas (1999). Modelling the long-run demand for money in the United Kingdom: A random coefficient analysis. *Economic Modelling* forthcoming.
- Judge, G.G., W.E. Griffiths, R. Carter Hill, H. Lütkepohl, and T. Lee (1985). *The Theory and Practice of Econometrics*, 2nd edn. New York: John Wiley and Sons.
- Klein, L.R. (1953). *A Textbook of Econometrics*. Evanston: Row Peterson and Company.
- Leusner, J., J.D. Akhavein, and P.A.V.B. Swamy (1998). Solving an empirical puzzle in the capital asset pricing model. In A.H. Chen (ed.) *Research in Finance*, Volume 16, Stamford, CT: JAI Press, Inc.
- Moggridge, D. (1973). *The General Theory and After*. New York: Macmillan.
- Mosteller, F., and J.W. Tukey (1977). *Data Analysis and Regression*. Reading, MA: Addison-Wesley Publishing Company.
- Nicholls, D.F., and A.R. Pagan (1985). Varying coefficient regression. In E.J. Hannan, P.R. Krishnaiah, and M.M. Rao (eds.) *Handbook of Statistics*, Volume 5, New York: Elsevier Science Publishers.

- Oakes, D. (1985). Self-calibrating priors do not exist. *Journal of the American Statistical Association* 80, 339.
- Phillips, R.J., and P.A.V.B. Swamy (1998). Par clearance in the domestic exchanges: The impact of national bank notes. In A.J. Field, G. Clark, and W.A. Sundstrom (eds.) *Research in Economic History*. pp. 121–44. Stamford, CT: JAI Press, Inc.
- Pratt, J.W., and R. Schlaifer (1984). On the nature and discovery of structure. *Journal of the American Statistical Association* 79, 9–21, 29–33.
- Pratt, J.W., and R. Schlaifer (1988). On the interpretation and observation of laws. *Journal of Econometrics* 39, 23–52.
- Schervish, M.J. (1985). Discussion of “Calibration-based empirical probability” by A.P. Dawid. *Annals of Statistics* 13, 1274–82.
- Stone, M. (1974). Cross-validatory choice and assessment of statistical predictions. *Journal of the Royal Statistical Society B*, 36, 111–33.
- Swamy, P.A.V.B., and J.S. Mehta (1975). Bayesian and non-Bayesian analysis of switching regressions and of random coefficient regression models. *Journal of the American Statistical Association* 70, 593–602.
- Swamy, P.A.V.B., and G.S. Tavlas (1995). Random coefficient models: Theory and applications. *Journal of Economic Surveys* 9, 165–82.
- Swamy, P.A.V.B., and P.A. Tinsley (1980). Linear prediction and estimation methods for regression models with stationary stochastic coefficients. *Journal of Econometrics* 12, 103–42.
- Zellner, A. (1969). On the aggregation problem. In K.A. Fox, J.K. Sengupta, and G.V.L. Narasimham (eds.) *Economic Models, Estimation and Risk Programming: Essays in Honor of Gerhard Tintner*. pp. 365–74. New York: Springer-Verlag.
- Zellner, A. (1988). Causality and causal laws in economics. *Journal of Econometrics* 39, 7–22.

CHAPTER TWENTY

# Nonparametric Kernel Methods of Estimation and Hypothesis Testing

*Aman Ullah\**

## 1 INTRODUCTION

Over the last five decades much research in empirical and theoretical econometrics has been centered around the estimation and testing of various econometric functions. For example the regression functions studying the consumption and production functions, the heteroskedasticity functions studying the variability or volatility of financial returns, the autocorrelation function exploring the nature of the time series, and the density functions analyzing the shape of the residuals or any economic variable. A traditional approach to studying these functions has been to first impose a parametric functional form and then proceed with the estimation and testing of interest. A major disadvantage of this traditional approach is that the econometric analysis may not be robust to the slight data inconsistency with the particular parametric specification. Indeed any misspecification in the functional form may lead to erroneous conclusions. In view of these problems, recently a vast amount of literature has appeared on the nonparametric and semiparametric approaches to econometrics, see the books by Prakasa Rao (1983), Silverman (1986), Härdle (1990), Fan and Gijbels (1996) and Pagan and Ullah (1999). In fact a large number of papers continue to pour in to various journals of statistics and econometrics.

The basic point in the nonparametric approach to econometrics is to realize that, in many instances, one is attempting to estimate an expectation of one variable,  $y$ , conditional upon others,  $x$ . This identification directs attention to the need to be able to estimate the conditional mean of  $y$  given  $x$  from the data  $y_i$  and  $x_i$ ,  $i = 1, \dots, n$ . A nonparametric estimate of this conditional mean simply follows as a weighted average  $\sum w(x_i, x)y_i$ , where  $w(x_i, x)$  are a set of weights that depend upon the distance of  $x_i$  from the point  $x$  at which the conditional expectation is to be evaluated. A kernel weight is considered and it is the subject of discussion in Section 2. This section also indicates how the procedures extend to the estimation of any higher order moments and the estimation of the derivatives of the function linking  $y$  and  $x$ . Finally, a detailed discussion of the existing and some new goodness-of-fit procedures for the nonparametric regression are presented; and their applications for determining the window width in the kernel weight, and the variables selection are discussed.

A problem with the a priori specified parametric function is that when it is misspecified, even in the small regions of the data, the parametric fit may be poor (biased) though it may be smooth (low variance). On the other hand the nonparametric functional estimation techniques, which totally depend on the data and have no a priori specified functional form may trace the irregular pattern in the data well (less bias) but may be more variable (high variance). A solution discussed in Section 3 is to use a combination of parametric and nonparametric regressions which can improve upon, in the mean squared error (MSE) sense, the drawbacks of each when used individually.

Perhaps the major complications in a purely nonparametric approach to estimation is the “curse of dimensionality”, which implies that, if an accurate measurement of the function is to be made, the size of sample should increase rapidly with the number of variables involved in any relation. This problem has lead to the development of additive nonparametric regressions which estimate the regressions with the large numbers of  $x$  with a similar accuracy as the regression with one variable. This is discussed in Section 4. Another solution is to consider a linear relationship for some variables while allowing a much smaller number to have an unknown nonlinear relation. Accordingly, Section 4.1 deals with these and other related models which are referred to as the semiparametric models.

While the major developments in nonparametric and semiparametric research have been in the area of estimation, only recently have papers started appearing which deal with the hypothesis testing issues. The general question is how to deal with the traditional hypothesis testing problems – such as the test of functional forms, restrictions, heteroskedasticity – in the nonparametric and semiparametric models. This is explored in Section 5.

The plan of the paper is as follows. In Section 2 we present the estimation of nonparametric regression. Then in Section 3 we discuss the combined regressions. Section 4 deals with the additive regressions and the semiparametric models. Finally, in Section 5 we explore the issues in the nonparametric hypothesis testing.

## 2 NONPARAMETRIC REGRESSION

Consider the regression model

$$y_i = m(x_i) + u_i,$$

where  $i = 1, \dots, n$ ,  $y_i$  is the dependent variable,  $x_i = (x_{i1}, \dots, x_{iq})$  are  $q$  regressors,  $m(x_i) = E(y_i | x_i)$  is the true but unknown regression function, and  $u_i$  is the error term such that  $E(u_i | x_i) = 0$  and  $V(u_i | x_i) = \sigma^2(x_i)$ .

If  $m(x_i) = f(\beta, x_i)$  is a correctly specified family of parametric regression functions then  $y_i = f(\beta, x_i) + u_i$  is a correct model and, in this case, one can construct a consistent least squares (LS) estimator of  $m(x_i)$  given by  $\hat{f}(\hat{\beta}, x_i)$ , where  $\hat{\beta}$  is the LS estimator of the parameter vector  $\beta$ . This  $\hat{\beta}$  is obtained by minimizing  $\sum u_i^2 = \sum (y_i - f(\beta, x_i))^2$  with respect to  $\beta$ . For example, if  $f(\beta, x_i) = \alpha + x_i\beta = X_i\delta$ ,  $\delta = (\alpha, \beta')'$ , is linear we can obtain the LS estimator of  $\delta$  as  $\hat{\delta} = (X'X)^{-1}X'y$ , where  $X$  is a  $n \times (q+1)$  matrix with the  $i$ th row given by  $X_i = (1, x_i)$ . Further the predicted (fitted) values are  $\hat{y}_i = X_i\hat{\delta} = X_i(X'X)^{-1}X'y$ . In general, if the parametric regression  $f(\beta, x)$  is incorrect or the form of  $m(x)$  is unknown then the  $\hat{f}(\hat{\beta}, x)$  may not be a consistent estimate of  $m(x)$ .

An alternative approach is to use the nonparametric regression estimation of the unknown  $m(x)$  by using techniques such as kernel, series, and spline, among others, see Newey (1997), Härdle (1990) and Pagan and Ullah (1999) for details. Here we consider the kernel estimation since it is simple to implement and its asymptotic properties are well established. Essentially the kernel estimator is a local LS (LLS) estimator obtained by minimizing  $\sum u_i^2 K(\frac{x_i - x}{h})$  where  $u_i = y_i - f(\beta, x_i)$ ,  $K_{i,x} = K(\frac{x_i - x}{h})$  are a decreasing function of the distances of the regressor vector  $x_i$  from the point  $x = (x_1, \dots, x_q)$ , and  $h > 0$  is the window width (smoothing parameter) which determines how rapidly the weights decrease as the distance of  $x_i$  from  $x$  increases. When  $h = \infty$ ,  $K_{i,x} = K(0)$  is a constant so that the minimization of  $K(0) \sum u_i^2$  is the same as the minimization of  $\sum u_i^2$ , that is the LLS estimator becomes the global LS estimator described above. In general, while the nonparametric LLS estimation fits  $f(\beta, x_i)$  to the points in the interval of length  $h$  around  $x$ , the parametric LS estimator fits  $f(\beta, x_i)$  globally to the entire scattering of the data.

When one treats  $m(x_i) = f(\beta, x_i)$  locally (around  $x$ ) as  $X_i\delta(x)$ , where  $\delta(x) = (\alpha(x), \beta(x)')$ , an explicit expression of the LLS estimator of  $\delta(x)$  is

$$\hat{\delta}(x) = (X'K(x)X)^{-1}X'K(x)y,$$

and the predicted values of  $m(x)$  are

$$\hat{y}_i = \hat{m}(x_i) = X_i\hat{\delta}(x_i) = X_i(X'K(x_i)X)^{-1}X'K(x_i)y = w_i y$$

or  $\hat{y} = Wy = \hat{m}$  where  $w_i = X_i(X'K(x_i)X)^{-1}X'K(x_i)$  is an  $n \times n$   $i$ th row of  $W$ ,  $K(x)$  is the diagonal matrix with the diagonal elements  $(K(\frac{x_j - x}{h}))$ ,  $\dots$ , and  $\hat{m} = [\hat{m}(x_1), \dots,$

$\hat{m}(x_n)'$ . The estimator  $\delta(x)$  is the local linear LS (LLLS) or simply the local linear (LL) estimator. One can consider  $f(\beta, x_i)$  to be the polynomials in  $x_i$  of degree  $d$ , in which case the matrix  $X$  will contain polynomials and the estimator  $\delta(x)$  will be the local polynomial LS (LPLS) estimator. For more details, see Fan and Gijbels (1996).

When one treats  $m(x_i)$  locally (around  $x$ ) as a scalar constant  $\alpha(x)$ , the LLS estimator of  $\alpha(x)$  is

$$\hat{\alpha}(x) = (\mathbf{1}'K(x)\mathbf{1})^{-1}\mathbf{1}'K(x)y = \frac{\sum_i y_i K_{i,x}}{\sum_i K_{i,x}},$$

and the predicted values are  $\hat{y}_i = \hat{m}(x_i) = \hat{\alpha}(x_i) = w_i y = \frac{\sum_j y_j K_{ji}}{\sum_j K_{ji}} = \sum_j y_j w_{ji}$ , where  $\mathbf{1}$  is an  $n \times 1$  vector of unit elements,  $K_{ji} = K(\frac{x_j - x_i}{h})$ ,  $w_i = (\mathbf{1}'K(x_i)\mathbf{1})^{-1}\mathbf{1}'K(x_i)$  is the  $i$ th row of  $W$ , and  $w_{ji} = K_{ji}/\sum_j K_{ji}$ . The estimator  $\hat{\alpha}(x)$  is the local constant LS (LCLS) estimator, and it was first introduced by Nadaraya (1964) and Watson (1964) (N-W).

The traditional approach to LLLS estimator (Fan and Gijbels, 1996) is to take a first-order Taylor series expansion of  $m(x_i)$  around  $x$  so that  $y_i = m(x_i) + u_i = m(x) + (x_i - x)m^{(1)}(x) + v_i = \alpha(x) + x_i\beta(x) + v_i = X_i\delta(x) + v_i$ ; where  $m^{(1)}(x) = \beta(x) = \partial m(x)/\partial x$  is the first derivative of  $m(x)$ ,  $\alpha(x) = m(x) - x\beta(x)$  and  $X_i$  and  $\delta(x)$  are as given above. The LLLS estimator  $\hat{\delta}(x)$  is then obtained by minimizing  $\sum v_i^2 K(\frac{x_i - x}{h})$ , and it is equivalent to  $\hat{\delta}(x)$  given above. Furthermore  $\hat{m}(x_i) = \hat{\alpha}(x_i) + x_i\hat{\beta}(x_i) = X_i\hat{\delta}(x_i)$  is an estimator of  $m(x_i) = \alpha(x_i) + X_i\beta(x_i) = X_i\delta(x_i)$ . We note that while LLLS provides the estimates of the unknown function  $m(x)$  and its derivative  $\beta(x)$  simultaneously, LCLS estimator of N-W provides the estimator of  $m(x)$  only. The first analytical derivative of  $\hat{m}(x)$  is then taken to obtain  $\hat{\beta}(x)$ , see Pagan and Ullah (1999, ch. 4).

The LLS results provide the point-wise estimates of  $\beta$  which vary with  $x$ . In some situations one may be interested in knowing how the parameters change with respect to a vector of variables  $z_i$  which is not necessarily in the model. That is, the model to be estimated is, say,  $y_i = f(\beta(z_i), x_i) + u_i$  or in the linear case  $y_i = x_i\beta(z_i) + u_i$ . This can be estimated by minimizing  $\sum u_i^2 K(\frac{z_i - z}{h}) = \sum [y_i - x_i\beta]^2 K(\frac{z_i - z}{n})$  which gives  $\hat{\beta}(z) = (X'K(z)X)^{-1}X'K(z)y$ . For examples and applications of these models, see Cai, Fan, and Yao (1998), Robinson (1989) and Li, Huang, and Fu (1998).

The above results extend to the estimation of  $E(g(y_i) | x_i)$  where  $g(y_i)$  is a function of  $y_i$  such that  $E|g(y_i)| < \infty$ , for example,  $E(y_i^2 | x_i)$ , where  $g(y_i) = y_i^2$ .

The asymptotic normality results of LLS and N-W (LCLS) estimators are similar and they are well established in the literature. But their finite sample approximate bias expressions to  $O(h^2)$  are different while the variance expressions are the same. These are now well known, see Pagan and Ullah (1999, chs. 3 and 4) for details. These results accompanied by several simulation studies (see Fan and Gijbels, 1996) indicate that the MSE performance of the LLLS is much better than that of N-W estimator especially in the tails where data are fewer. In particular, the bias of the N-W estimator is much worse compared to LLLS in the tails, and

while the LLLS is unbiased when the true regression function  $m(x)$  is linear, the N-W estimator is biased. Intuitively this makes sense since while the N-W estimator fits a constant to the data around each point  $x$  the LLLS estimator fits a linear line around  $x$ . These properties and the fact that the LLLS estimator provides derivatives (elasticities) and the regression estimators simultaneously, and that it is simple to calculate, make LLLS an appealing estimation technique. In the future, however, more research is needed to compare the performances of LPLS and LLLS.

An important implication of the asymptotic results of LLS and N-W estimators of  $m(x)$  and  $\beta(x)$  is that the rate of convergence of  $\hat{m}(x)$  is  $(nh^q)^{1/2}$  and that of  $\hat{\beta}(x)$  is  $(nh^{q+2})^{1/2}$  which are slower than the parametric rate of  $n^{1/2}$ . In fact as the dimension of regressors  $q$  increases the rates become worse. This is the well known problem of the “curse of dimensionality” in the purely nonparametric regression. In recent years there have been several attempts to resolve this problem. One idea is to calculate the average regression coefficients, e.g.  $\sum \hat{\beta}(x_i)/n$  or weighted average coefficients which give  $n^{1/2}$  convergence rate. This, however, may not have much economic meaning in general, except in the case of single index models used in labor econometrics, see Powell, Stock, and Stoker (1989). Another solution is to use the additive regression models which improve the  $(nh^q)^{1/2}$  rate to a univariate rate of  $(nh)^{1/2}$ . This is described in Section 4.

The asymptotic results described above are well established for the independently and identically distributed (iid) observations, and for the weakly dependent time series data. For the extensions of the results to nonparametric kernel estimation with nonstationary data, see Phillips and Park (1998), and for the case of a purely nonparametric single index model  $y_i = m(x_i\beta) + u_i$ , see Lewbel and Linton (1998).

## 2.1 Goodness of fit measures and choices of kernel and bandwidth

The LLS estimators are easy to implement. Once a window width and kernel are chosen,  $K_{i,x} = K(\frac{x_j - x}{h})$  can be computed for each value of the  $x = x_j$ ,  $j = 1, \dots, n$ , in the sample and then substituted in the LLS, LLLS, and N-W (LCLS) estimators given above. Confidence intervals for the LLLS and N-W estimators can then be obtained by using the asymptotic normality results. The key issues about the LLS estimators therefore involve the selection of kernel and window width. Regarding the choice of kernel we merely remind readers that for the large data sets it is now believed that the choice of smoothing kernel is not crucial, and that data should be transformed to a standardized from before entry into kernels. Also, in practice the product kernels are easy to use and perform well, that is  $K(\psi_i) = \prod_s K(\psi_{si})$  where  $s = 1, \dots, q$ ; and  $K(\psi_{si})$  can be taken as the univariate normal with unbounded support or the Epanechnikov kernel with bounded support  $K(\psi_{si}) = \frac{3}{4}(1 - \psi_{si}^2)$ ,  $|\psi_{si}| \leq 1$ . These kernels are second-order kernels, implying their first moments are zero but the second are finite. Another class of kernels, known as higher order kernels with higher order moments as zero, are

used in order to reduce the asymptotic bias problem in the LLS estimators. However, in practice the gains from these higher order kernels are not significant. For details on the choice of kernels see Silverman (1986).

The window width  $h$  controls the smoothness of the LLS estimate of  $m(x)$  and in practice is crucial in obtaining a good estimate that controls the balance between the variance, which is high when  $h$  is too small, and the squared bias which is high when  $h$  is too large. With this in mind several window width selection procedures have tried to choose  $h$  to minimize a number of mean squared error (MSE) criteria, for example  $\int(\hat{m}(x) - m(x))^2 dx$ ,  $E\int((\hat{m}(x) - m(x))^2 dx) = \int \text{MSE}(\hat{m}(x))dx$  = integrated MSE = IMSE and the average IMSE = AIMSE =  $E\int(\hat{m}(x) - m(x))^2 f(x)dx$ . The minimization of IMSE is known to provide the optimal  $h$  to be  $c n^{-1/(q+4)}$  where  $c$  is a constant of proportionality which depends on unknown density and  $m(x)$  and their derivatives. An initial estimate of  $c$  can be constructed giving "plug-in" estimators of  $h$  but this has not been a very popular procedure in practice. A recent proposal from Härdle and Bowman (1988) is to estimate AIMSE by bootstrapping and then minimizing the simulated AIMSE with respect to  $h$ . An advantage of this approach is that it also provides a confidence interval for  $\hat{m}$ .

Cross validation is a popular alternative procedure, which chooses  $h$  by minimizing the sum of squares of the estimated prediction error (EPE) or residual sum of squares (RSS),  $EPE = RSS = \frac{1}{n} \sum_i^n (y_i - \hat{y}_{-i})^2 = \frac{1}{n} \sum_i^n \hat{u}_{-i}^2 = \frac{\hat{u}' \hat{u}}{n} = \frac{y' M' My}{n}$  where  $\hat{y}_{-i} = \hat{m}_{-i}(x_i) = w_{-i} y$ ,  $\hat{u}_{-i} = y_i - \hat{y}_{-i}$ ,  $\hat{u} = y - W_{-i} y = My$  and  $M = I - W_{-i}$ ; subscript  $-i$  indicates the "leave-one-out" estimator, deleting  $i$ th observation in the sums is used. An alternative is to consider  $EPE^* = \frac{y' M^* y}{tr(M^*)}$  where  $M^* = M'M$  and  $tr(M^*)$  can be treated as the degrees of freedom in the nonparametric regression, as an analogue to the linear parametric regression case.

One drawback of the "goodness-of-fit function" EPE is that there is no penalty for large or small  $h$ . In view of this, many authors have recently considered the penalized goodness of fit function to choose  $h$  see Rice (1984) and Härdle, Hall, and Marron (1992). The principle is the same as in the case of penalty functions used in the parametric regression for the choice of number of parameters (variables), for example the Akaike criterion. The idea of a penalty function is, in general, an attractive one and it opens up the possibility of future research.

A related way to obtain  $h$  is to choose a value of  $h$  for which the square of correlation between  $y_i$  and  $\hat{y}_i(p_{y,\hat{y}}^2)$  is maximum, that is  $0 \leq R^2 = \hat{p}_{y,\hat{y}}^2 \leq 1$  is maximum. One could also use the correlation between  $y$  and leave-one-out estimator  $\hat{y}_{-i}$ .

When  $V(u_i | x_i) = \sigma^2(x_i)$ , one can choose  $h$  such that an estimate of unconditional variance of  $u_i$ ,  $Eu_i^2 = E[E(u_i^2 | x_i)] = E[\sigma^2(x_i)] = \int \sigma^2(x) f(x) dx$  is minimum. That is, choose  $h$  such that  $EPE_1 = \hat{E}u_i^2 = \int \hat{\sigma}^2(x) d\hat{F}(x)$  where  $\hat{\sigma}^2(x_i) = \hat{E}(\tilde{u}_i^2 | x_i)$  is obtained by the LPLS regression of  $\tilde{u}_i^2$  on  $x_i$ ;  $\tilde{u}_i = y_i - \hat{m}(x_i)$  is the nonparametric residual, and  $\hat{f}(x) = \sum_i^n w_i(x)$  is a nonparametric density estimator for some weight function  $w_i(x)$  such that  $\int w_i(x) dx = 1$ . For the kernel density estimator  $w_i(x) = K(\frac{x_i - x}{h})/nh^q$ , see Silverman (1986) and Pagan and Ullah (1999). It is better to use  $EPE_1$  than EPE when there is a heteroskedasticity of unknown form. Also, since  $\hat{E}(\tilde{u}_i^2)$  can be shown to be a consistent estimator of  $Eu_i^2$  the nonparametric version of  $R^2$ ,

$R_1^2 = 1 - \frac{\text{EPE}_1}{\frac{1}{n} \sum_i^n (y_i - \hat{y})^2}$  lies between 0 and 1, and it is an estimator of  $\rho^2 = 1 - \frac{Eh_i^2}{V(y_i)} = 1 - \frac{E(y_i - m(x_i))^2}{V(y_i)}$ .

Thus, an alternative is to choose  $h$  such that  $R_1^2$  is maximum. One can also use  $\text{EPE}_1^*$  in  $R_1^2$ . This will correspond to  $\bar{R}^2$  in the parametric regression. A simple way to calculate  $\text{EPE}_1$  is to consider the empirical distribution function so that  $\text{EPE}_1 = \frac{1}{n} \sum_i^n \hat{\sigma}^2(x_i)$ .

In general the move from independent to dependent observations should not change the way window width selection is done. However, care has to be taken since, as indicated by Robinson (1986), a large window width might be needed for the dependent observations case due to the positive serial correlations (see Herrman, Gasser, and Kneip, 1992). Faraway (1990) considers the choice of varying window width. For details on the choices of  $h$  and their usefulness see Pagan and Ullah (1999).

### 3 COMBINED REGRESSION

Both the parametric and nonparametric regressions, when used individually, have certain drawbacks. For example, when the a priori specified parametric regression  $m(x) = f(\beta, x)$  is misspecified even in the small regions of the data, the parametric fit may be poor (biased) though it may be smooth (low variance). On the other hand, the nonparametric regression techniques, which totally depend on the data and have no a priori specified functional form may trace the irregular pattern in the data well (less bias) but may be more variable (high variance). Thus, when the functional form of  $m(x)$  is unknown, a parametric model may not adequately describe the data in its entire range, whereas a nonparametric analysis would ignore the important a priori information about the underlying model. A solution considered in the literature is to use a combination of parametric and nonparametric regressions which can improve upon the drawbacks of each when used individually, see Eubank and Spiegelman (1990), Fan and Ullah (1998), and Glad (1998). Essentially the combined regression estimator controls both the bias and variance and hence improves the MSE of the fit. To see the idea behind the combined estimation let us start with a parametric model ( $m(x) = f(\beta, x)$ ) which can be written as

$$y_i = m(x_i) + u_i = f(\beta, x_i) + g(x_i) + \varepsilon_i,$$

where  $g(x_i) = E(u_i | x_i) = m(x_i) - E(f(\beta, x_i) | x_i)$  and  $\varepsilon_i = u_i - E(u_i | x_i)$  such that  $E(\varepsilon_i | x_i) = 0$ . Note that  $f(\beta, x_i)$  may not be a correctly specified model so  $g(x_i) \neq 0$ . If it is indeed a correct specification  $g(x_i) = 0$ . The combined estimation of  $m(x)$  can be written as

$$\hat{m}_c(x_i) = f(\hat{\beta}, x_i) + \hat{g}(x_i),$$

where  $\hat{g}(x_i) = \hat{E}(\hat{u}_i | x_i)$  is obtained by the LLS estimation technique and  $\hat{u}_i = y_i - f(\hat{\beta}, x_i)$  is the parametric residual.

An alternative way to combine the two models is to introduce a weight parameter  $\lambda$  and write  $y_i = f(\beta, x_i) + \lambda g(x_i) + \varepsilon_i$ . If the parametric model is correct,  $\lambda = 0$ . Thus, the parameter  $\lambda$  measures the degree of accuracy of the parametric model. A value of  $\lambda$  in the range 0 to 1 can be obtained by using the goodness of fit measures described in Section 2.1, especially  $R^2$  and EPE. Alternatively, an LS estimator  $\hat{\lambda}$  can be obtained by doing a density weighted regression of  $y_i - f(\beta, x_i) = \hat{u}_i$  on  $\hat{g}(x_i)$ , see Fan and Ullah (1998). The combined estimator of  $m(x)$  can now be given by

$$\hat{m}_c^*(x_i) = f(\hat{\beta}, x_i) + \hat{\lambda} \hat{g}(x_i).$$

This estimation indicates that a parametric start model  $f(\hat{\beta}, x)$  be adjusted by  $\hat{\lambda}$  times  $\hat{g}(x)$  to get a more accurate fit of the unknown  $m(x)$ .

Instead of additive adjustments to the parametric start in  $\hat{m}_c(x)$  and  $\hat{m}_{c*}(x)$ , Glad (1998) proposed a multiplicative adjustment as given below. Write

$$m(x_i) = f(\beta, x_i) \frac{m(x_i)}{f(\beta, x_i)} = f(\beta, x_i) E(y_i^* | x_i),$$

where  $y_i^* = y_i/f(\beta, x_i)$ . Then  $\hat{m}_g(x_i) = f(\hat{\beta}, x_i) \hat{E}(\hat{y}_i^* | x_i)$  is the estimator proposed by Glad (1998), where  $\hat{y}_i^* = y_i/f(\hat{\beta}, x_i)$ , and  $\hat{E}(\cdot)$  is obtained by the LLS estimator described above.

The asymptotic convergence rates of  $\hat{m}_c(x)$  and its asymptotic normality are given in Fan and Ullah (1998). In small samples, the simulation results of Rahman and Ullah (1999) suggest that the combined estimators perform, in the MSE sense, as well as the parametric estimator if the parametric model is correct and perform better than both the parametric and nonparametric LLS estimators if the parametric model is incorrect.

## 4 ADDITIVE REGRESSIONS

In recent years several researchers have attempted to estimate  $m(x_i)$  by imposing some structure upon the nature of the conditional mean  $m(x_i)$ . One popular solution is the generalized additive models of Hastie and Tibshirani (1990), which is

$$y_i = m(x_i) + u_i = m_1(x_{i1}) + m_2(x_{i2}) + \dots + m_q(x_{iq}) + u_i,$$

where  $m_s$ ,  $s = 1, \dots, q$ , are functions of single variables with  $E m_s(x_{is}) = 0$ ,  $s = 2, \dots, q$ , for identification. Each of  $m_s$  and hence  $m(x_i)$  is then estimated by one dimensional convergence rate of  $(nh)^{1/2}$  which is faster than the convergence rate  $(nh)^{1/2}$  achieved by direct nonparametric estimation of  $m(x_i)$ . The statistical properties of Hastie and Tibshirani (1990) estimation algorithm is complicated. For practical implementations, simpler estimation techniques are proposed in Linton and Nielson (1995) and Chen *et al.* (1996). The basic idea behind this is as follows. At

the first stage estimate  $\hat{m}(x_i) = \hat{m}(x_{i1}, \dots, x_{iq}) = \hat{m}(x_{i1}, x_{i\underline{l}})$  by the nonparametric LLS procedure, where  $x_{i\underline{l}}$  is a vector of  $x_{i2}, \dots, x_{iq}$ . Then, using  $E m_s(x_{is}) = 0$ , we note that

$$m_1(x_{i1}) = \int m(x_{i1}, x_{i\underline{l}}) dF(x_{i\underline{l}})$$

and hence  $\hat{m}_1(x_{i1}) = \int \hat{m}(x_{i1}, x_{i\underline{l}}) d\hat{F}(x_{i\underline{l}})$ . Using the empirical distribution one can calculate  $\hat{m}_1(x_{i1}) = \frac{1}{n} \sum_{j=1}^n \hat{m}(x_{i1}, x_{j\underline{l}})$ .  $\hat{m}_s(x_{is})$  for any  $s$  can be similarly calculated. Under the assumptions that  $[y_i, x_i]$  are iid,  $nh^3 \rightarrow \infty$  and  $nh^5 \rightarrow 0$  as  $n \rightarrow \infty$ , Linton and Nielson show the  $(nh)^{1/2}$  convergence to normality for  $\hat{m}_1$ . For the test of additivity of  $m(x_i)$  see Linton and Gozalo (1996), and for the application to estimating a production function see Chen *et al.* (1996).

Alternative useful approaches which impose structure on  $m(x_i)$  are the projection pursuit regression and the neural networks procedures. For details on them, see Friedman and Tukey (1974), Breiman and Friedman (1985), Kuan and White (1994), Härdle (1990) and Pagan and Ullah (1999).

## 4.1 Semiparametric models

A partial solution to the dimensionality problem was also explored in Speckman (1988) and Robinson (1988). They considered the case where  $x_i = (x_{i1}, x_{i2})$  and  $m(x_i) = m(x_{i1}, x_{i2}) = m_1(x_{i1}) + m_2(x_{i2})$ , but the researcher knows the functional form of  $m_1(x_{i1})$  as  $x_{i1}\beta$ , where  $x_{i1}$  is a  $q_1$  dimensional and  $x_{i2}$  is  $q_2$  dimensional with no common elements. That is, they considered the partial linear models or semi-parametric (SP) model of the following form

$$y_i = x_{i1}\beta + m(x_{i2}) + u_i,$$

where  $E(u_i | x_i) = 0$ . For example, in the earning functions log earning ( $y_i$ ) may be an unknown function of age ( $x_{i2}$ ) but a linear function of education ( $x_{i1}$ ). The estimation of  $\beta$  can be carried out by first eliminating  $m(x_{i2})$ , and then using the following procedure of taking conditional expectations so that

$$E(y_i | x_{i2}) = E(x_{i1} | x_{i2})\beta + m(x_{i2})$$

and  $y_i - E(y_i | x_{i2}) = (x_{i1} - E(x_{i1} | x_{i2}))\beta + u_i$  or  $y_i^* = x_{i1}^*\beta + u_i$ . This can then be estimated by the LS procedure as

$$\hat{\beta}_{SP} = \left( \sum_i^n x_{i1}^* x_{i1}^{*\prime} \right)^{-1} \sum_i^n x_{i1}^* y_i^*.$$

For the implementation we need to know  $y_i^*$  and  $x_{i1}^*$ , which can be obtained by estimating  $E(y_i | x_{i2})$  and  $E(x_{i1} | x_{i2})$  using the LLS procedures in Section 2. After obtaining  $\hat{\beta}_{SP}$  one can proceed for the estimation of  $m(x_{i2})$  by writing

$$y_i - x_{i1}\hat{\beta}_{SP} = y_i^{**} = m(x_{i2}) + u_i = E(y_i^{**} | x_{i2}) + u_i,$$

and then again doing the LLS regression of  $y_i^{**}$  on  $x_{i2}$ .

While the estimator of  $m(x_{i1})$  achieves the nonparametric slow rate of convergence of  $(nh^{q/2})^{1/2}$ , the remarkable point is that the  $\hat{\beta}_{SP}$  achieves the parametric rate of convergence of  $n^{1/2}$ . It is in this respect that this procedure is better than the univariate nonparametric convergence rates of generalized additive models above. However, in the partial linear model we need to be sure of the linearity of  $x_{i1}\beta$ . If  $m(x_{i1}, \beta)$  is a nonlinear function of  $x_{i1}$  and  $\beta$ , then it is not clear how one proceeds with the above estimation technique, though it seems that a nonlinear semiparametric LS procedure might be helpful. For the empirical applications of the above models, see Engle *et al.* (1986) for an electricity expenditure estimation, Anglin and Gencay (1996) for a hedonic price estimation of Canadian housing.

There are various extensions of the idea of the partial linear models. Fan and Li (1997) combine the partial linearity with the generalized additive models to consider

$$y_i = x_{i1}\beta + m_2(x_{i2}) + m_3(x_{i3}) + \dots + m_q(x_{iq}) + u_i$$

and suggest the  $\sqrt{n}$  convergent estimate of  $\beta$  and  $(nh)^{1/2}$  convergent estimates of  $m_s(x_{is})$  for  $s = 2, \dots, q$ . This improves upon the  $(nh^q)^{1/2}$  state of convergence of  $m(x_{i2})$  above.

The partially linear models have been extensively studied in the labor econometric literature on the selection models where  $m(x_{i2}) = m(x_{i2}\delta)$  is an unknown function of single index  $x_{i2}\delta$  and  $x_{i2}$  and  $x_{i1}$  may have some common variables. For details on this literature, see Pagan and Ullah (1999, chs. 7–9). For the maximum likelihood estimation of the purely parametric model,  $y_i = x_i\beta + u_i$ , partial linear, and selection models without the assumption about the form of the density of  $u_i$ , see the excellent work of Ai (1997). The estimation of panel data based partially linear models has been developed in Ullah and Roy (1998), Li and Ullah (1998) and Li and Stengos (1996), among others.

## 5 HYPOTHESIS TESTING

An obvious question is how to carry out various diagnostic tests done in the parametric econometrics within the nonparametric and semiparametric models. Several papers have appeared in the recent literature which deal with this issue. We present them here and show their links.

First consider the problem of testing a specified parametric model against a nonparametric alternative,  $H_0 : f(\beta, x_i) = E(y_i | x_i)$  against  $H_1 : m(x_i) = E(y_i | x_i)$ . The idea behind the Ullah (1985) test statistic is to compare the parametric RSS (PRSS)  $\sum \hat{u}_i^2$ ,  $\hat{u}_i = y_i - f(\hat{\beta}, x_i)$  with the nonparametric RSS (NPRSS),  $\sum \tilde{u}_i^2$ , where  $\tilde{u}_i = y_i - \hat{m}(x_i)$ . His test statistic is

$$T_1 = \frac{(PRSS - NPRSS)}{NPRSS} = \frac{PRSS}{NPRSS} - 1 = \frac{\sum \hat{u}_i^2 - \sum \tilde{u}_i^2}{\sum \tilde{u}_i^2},$$

or simply  $T_1^* = (PRSS - NPRSS)$ , and reject the null hypothesis when  $T_1$  is large.  $\sqrt{n} T_1$  has a degenerate distribution under  $H_0$ . Lee (1994) uses density weighted

residuals and compares  $\sum w_i \hat{u}_i^2$  with  $\sum \hat{u}_i^2$  to avoid degeneracy, for other procedures see Pagan and Ullah (1999). Pagan and Ullah (1999) also indicate the normalizing factor needed for the asymptotic normality of  $T_1$ . An alternative suggested here is the following nonparametric bootstrap method:

1. Generate the bootstrap residuals  $u_i^*$  from the centered residuals  $(\tilde{u}_i - \bar{\tilde{u}})$  where  $\bar{\tilde{u}}$  is the average of  $\tilde{u}_i$ .
2. Generate  $y_i^* = f(\hat{\beta}, x_i) + u_i^*$  from the null model.
3. Using the bootstrap sample  $x_i, y_i^*, i = 1, \dots, n$ , estimate  $m(x_i)$  nonparametrically, say and  $\hat{m}^*(x_i)$ , and obtain the bootstrap residual  $\tilde{u}_i^* = y_i^* - \hat{m}^*(x_i)$ .
4. Calculate the bootstrap test statistic  $T_1^* = (\sum \hat{u}_i^2 - \sum \tilde{u}_i^*)^2 / \sum \tilde{u}_i^{*2}$ .
5. Repeat steps (1) to (4) B times and use the empirical distribution of  $T_1^*$  as the null distribution of  $T_1^*$ .

An alternative is to use wild bootstrap method or pivotal bootstrap which will preserve the conditional heteroskedasticity in the original residuals. Another alternative is to use the block bootstrap method (Bühlman and Künsch, 1995).

An alternative test statistic is based on comparing the parametric fit with the nonparametric fit. Defining  $a(x)$  as a smooth weight function, this test statistic is

$$T_2 = \frac{1}{n} \sum_i^n (f(\hat{\beta}, x_i) - \hat{m}(x_i))^2 a(x_i) = \int (f(\hat{\beta}, x_i) - \hat{m}(x_i))^2 a(x_i) d\hat{F}(x),$$

where  $\hat{F}$  is the empirical distribution function, see Ullah (1985) and Gozalo (1995), and Aït-Sahalia *et al.* (1998) who also indicate that  $f(\hat{\beta}, x_i)$  and  $\hat{m}(x_i)$  can also be replaced by  $f(\hat{\beta}, x_i) - \int f(\hat{\beta}, x_i) \hat{f}(x_i) dx_i$  and  $\hat{m}(x_i)$  by  $\hat{m}(x_i) - \int \hat{m}(x_i) \hat{f}(x_i) dx_i$  without affecting the results in practice and provide the asymptotic normality of  $nh^{q/2} T_2$ .

Härdle and Mammen (1993) suggest a weighted integrated square difference between the nonparametric estimator and the kernel smoothed parametric estimator  $\hat{f}(\hat{\beta}, x_i) = \hat{E}(f(\hat{\beta}, x_i) | x_i)$  which can be calculated by the LLS procedures with  $y_i$  replaced by  $f(\hat{\beta}, x_i)$ . This is

$$T_3 = \int_x (\hat{f}(\hat{\beta}, x) - \hat{m}(x))^2 a(x) dx = \int_x [\hat{E}(\hat{u} | x)]^2 a(x) dx,$$

where  $\hat{E}(\hat{u} | x) = \hat{E}(y | x) - \hat{E}(f(\hat{\beta}, x) | x) = \hat{m}(x) - \hat{f}(\hat{\beta}, x)$ . It has been shown in Rahman and Ullah (1999) that the use of the kernel smoothed estimator  $\hat{f}(\hat{\beta}, x)$  gives better size and power performances of the tests compared to the case of using  $f(\hat{\beta}, x)$ .  $T_3$  is similar to  $T_2$  if we write  $a(x) = a(x) \hat{f}^{-1}(x) d\hat{F}(x)$  and use the empirical distribution. This test statistic is computationally involved. In view of this, Li and Wang (1998) and Zheng (1996) proposed a conditional moment test (CMT) which is easy to calculate, and has a better power performance. Its form is

$$T_4 = \frac{1}{n} \sum_i^n \hat{u}_i \hat{E}(\hat{u}_i | x_i) \hat{f}(x_i) = \frac{1}{n} \sum_i^n \hat{u}_i (\hat{m}(x_i) - \hat{f}(\hat{\beta}, x_i)) \hat{f}(x_i).$$

This statistic is based on the idea that under the null  $E(u_i | x_i) = E[u_i E(u_i | x_i)] = E[(Eu_i | x_i)^2] = E[u_i E(u_i | x_i)a(x_i)] = 0$  for any positive  $a(x_i)$ .  $T_4$  is an estimate of  $E[u_i E(u_i | x_i)a(x_i)]$  for  $a(x_i) = f(x_i)$ . Eubank and Spiegelman (1990), however, tests for  $E[(Eu_i | x_i)^2] = 0$  using a series type estimator of  $E(u_i | x_i)$ . The test statistic  $T_4$  has  $nh^{q/2}$  rate of convergence to normality.

An intuitive and simple test of the parametric specification follows from the combined regression  $y_i = f(\beta, x_i) + \lambda g(x_i) + \varepsilon_i$  or  $u_i = y_i - f(\beta, x_i) = \lambda E(u_i | x_i) + \varepsilon_i$  given in Section 3. The estimator of  $\lambda$  is then

$$\hat{\lambda} = \frac{\frac{1}{n} \sum_i^n \hat{u}_i \hat{E}(\hat{u}_i | x_i) a(x_i)}{\frac{1}{n} \sum_i^n (\hat{E}(\hat{u}_i | x_i))^2 a(x_i)} = \frac{\frac{1}{n} \sum_i^n \hat{u}_i (\hat{m}(x_i) - \hat{f}(\hat{\beta}, x_i)) a(x_i)}{\frac{1}{n} \sum_i^n (\hat{m}(x_i) - \hat{f}(\hat{\beta}, x_i))^2 a(x_i)} = \frac{\hat{\lambda}_N}{\hat{\lambda}_D},$$

which is the weighted LS of  $\hat{u}_i$  on  $\hat{E}(\hat{u}_i | x_i)$ . Fan and Ullah (1998) considered the case where  $a(x_i) = \hat{f}^2(x_i)$  and  $\hat{f}(\hat{\beta}, x_i) = f(\hat{\beta}, x_i)$ , and established the asymptotic normality of  $h^{-q/2}\hat{\lambda}$  and  $nh^{q/2}\hat{\lambda}_N$ . Their test statistics for parametric specification,  $H_0 : \lambda = 0$  are

$$T_5 = \frac{\hat{\lambda}}{\sqrt{V(\hat{\lambda})}} \text{ and } T_6 = \frac{\hat{\lambda}_N}{\sqrt{V(\hat{\lambda}_N)}},$$

and they indicate the better performances of size and power of  $T_6$  compared to  $T_5$ . It is interesting to see the links between the test statistics  $T_1$  to  $T_6$ . First, since  $\hat{u}_i = y_i - f(\hat{\beta}, x_i)$  and  $\tilde{u}_i = y_i - \hat{m}(x_i)$ , it follows that

$$\frac{1}{n} \sum_i^n (\hat{u}_i^2 - \tilde{u}_i^2) a(x_i) = -\frac{1}{n} \sum_i^n (f(\hat{\beta}, x_i) - \hat{m}(x_i))^2 a(x_i) - \frac{2}{n} \sum_i^n \hat{u}_i (f(\hat{\beta}, x_i) - \hat{m}(x_i)) a(x_i)$$

or  $T_1 = -T_2 + 2T_4$ , except that  $T_4$  has  $\hat{f}(\hat{\beta}, x_i)$ . Thus under the null hypothesis  $T_4 \approx 0$  and  $T_2 \approx 0$  may imply  $T_1 \approx 0$ . We also note that  $T_6$ , with  $a(x_i) = \hat{f}(x_i)$ , is the Li-Wang and Zheng tests  $T_4$ .

All the above nonparametric tests are generally calculated with the leave-one-out estimators of  $\hat{m}(x_i) = \hat{m}_{-i}(x_i)$  and the weight  $a(x_i) = \hat{f}(x_i) = \hat{f}_{-i}(x_i)$ . Theoretically, the use of leave-one-out estimators helps to get asymptotic normality centered at zero. The tests are consistent model specification tests in the sense that their power goes to one as  $n \rightarrow \infty$  against all the alternatives. The usual parametric specification tests are however consistent against a specified alternative. An approach to developing consistent model specification test, without using any nonparametric estimator of  $m(x)$ , is the CMT due to Bierens-Newey-Tauchen, see Pagan and Ullah (1999) for details. An important difference between Bierens-type tests and the tests  $T_1$  to  $T_6$  is the treatment of  $h$ . While  $T_1$  to  $T_6$  tests consider  $h \rightarrow 0$  as  $n \rightarrow \infty$ , Bierens (1982) type tests treat  $h$  to be fixed which makes the asymptotic distribution of their tests to be nonnormal but can detect the Pitman's

local alternative that approach the null at the rate  $O(n^{-1/2})$  compared to the slower rate of  $O((nh^{q/2})^{-1/2})$  of  $T_1$  to  $T_6$ . However, Fan and Li (1996) indicate that under high frequency type local alternatives the tests with vanishing  $h$  may be more powerful than tests based on fixed  $h$ . For asymptotic normality of the tests  $T_1$  to  $T_6$  for the dependent observations, see Li (1997).

The test statistics  $T_1$  to  $T_6$  can also be used for the problem of variable selections. For example, testing  $H_0 : m(x_i) = m(x_{i1}, x_{i2}) = m(x_{i1})$  against  $H_1 : m(x_i) \neq m(x_{i1})$  can be carried out by calculating the difference between the NPRSS due to  $\hat{m}(x_{i1}, x_{i2})$  and the NPRSS due to  $\hat{m}(x_{i1})$ , or using  $T_2 = n^{-1}\sum(\hat{m}(x_{i1}, x_{i2}) - \hat{m}(x_{i1}))^2 a(x_i)$  test, see Ait-Sahalia *et al.* (1998) for asymptotic normality. An alternative is suggested in Racine (1997). We can also do the diagnostics for variable selection by using the goodness of fit measures described in Section 2.1; in addition see Vien (1994) where the cross-validation method has been used.

The tests  $T_1$  to  $T_6$  can also be extended to do nonnested testing (Delgado and Mora, 1998), testing for parametric and semiparametric models  $y_i = f(\beta, x_i) + \lambda m(x_i) + u_i$ , (Li, 1997; Fan and Li, 1996) and single index models, Ait-Sahalia *et al.*, 1998). Finally, for testing the restrictions on the parameters, testing heteroskedasticity, and testing serial correlation in the parametric model  $y_i = f(\beta, x_i) + u_i$  with the unknown form of density, see Gonzalez-Rivera and Ullah (1999) where they develop the semiparametric Rao-score test (Lagrange multiplier) with the unknown density replaced by its kernel estimator. Also see Li and Hsiao (1998) for a semiparametric test of serial correlation.

## Note

- \* The author is thankful to two referees for their constructive comments and suggestions.  
The research support from the Academic Senate, UCR, is gratefully acknowledged.

## References

- Ai, C. (1997). A semiparametric maximum likelihood estimator. *Econometrica* 65, 933–63.
- Ait-Sahalia, Y., P.J. Bickel, and T.M. Stoker *et al.* (1998). Goodness-of-fit regression using kernel methods. Manuscript, University of Chicago.
- Anglin, P., and R. Gencay (1996). Semiparametric estimation of a hedonic price function. *Journal of Applied Econometrics* 11, 633–48.
- Bierens, H.J. (1982). Consistent model specification tests. *Journal of Econometrics* 20, 105–34.
- Breiman, L., and J. Friedman (1985). Estimating optimal transformations for multiple regression and correlation. *Journal of the American Statistical Association* 80, 580–619.
- Bühlman, P., and H.R. Künsch (1995). The blockwise bootstrap for general parameters of a stationary time series. *Scandinavian Journal of Statistics* 22, 35–54.
- Chen, R., W. Härdle, O.B. Linton, and E. Sevarance-Lossin (1996). Nonparametric estimation of additive separable regression model. *Statistical Theory and Computational Aspect of Smoothing*; Physica-Verlag 247–65.
- Cai, Z., J. Fan, and Q. Yao (1998). Functional-coefficient regression models for non-linear time series. Manuscript, University of North Carolina.
- Delgado, M.A., and J. Mora (1998). Testing non-nested semiparametric models: An application to Engel curve specification. *Journal of Applied Econometrics* 13, 145–62.

- Engel, R.F., C.W.J. Granger, J. Rice, and A. Weiss (1986). Semiparametric estimates of the relation between weather and electricity sales. *Journal of the American Statistical Association* 81, 310–20.
- Eubank, R.L., and C.H. Spiegelman (1990). Testing the goodness-of-fit of the linear models via nonparametric regression techniques. *Journal of the American Statistical Association* 85, 387–92.
- Fan, J., and I. Gijbels (1996). *Local Polynomial Modelling and Its Applications*. London: Chapman and Hall.
- Fan, Y., and Q. Li (1996). Consistent model specification tests: Omitted variables and semiparametric functional forms. *Econometrica* 64, 865–90.
- Fan, Y., and Q. Li (1997). On estimating additive partially linear models. Manuscript, University of Windsor.
- Fan, Y., and A. Ullah (1998). Asymptotic normality of a combined regression estimator. *Journal of Multivariate Analysis* 71, 191–240.
- Fan, Y., and A. Ullah (1999). On goodness-of-fit tests for weakly dependent processes using kernel method. *Journal of Nonparametric Statistics* 11, 337–60.
- Faraway, J. (1990). Bootstrap selection for bandwidth and confidence bands for nonparametric regression. *Journal of Statistics Computational Simulations* 37, 37–44.
- Friedman J.H., and J.W. Tukey (1974). A projection pursuit algorithm for exploratory data analysis. *IEEE Transactions on Computers*, C-23, 881–90.
- Glad, I.K. (1998). Parametrically guided nonparametric regression. *Scandinavian Journal of Statistics* 25, 649–68.
- Gozalo, P.L. (1995). Nonparametric specification testing with  $\sqrt{n}$ -local power and bootstrap critical values. Working Paper no. 95–21R; Brown University.
- Gonzalez-Rivera, G., and A. Ullah (1999). Rao's score test with nonparametric density estimators. *Journal of Statistical Planning and Inference*.
- Härdle, W. (1990). *Applied Nonparametric Regression*. New York: Cambridge University Press.
- Härdle, W., and A. Bowman (1988). Bootstrapping in nonparametric regression: Local adaptive smoothing and confidence bounds. *Journal of the American Statistical Association* 83, 102–10.
- Härdle, W., P. Hall, and J.S. Marron (1992). Regression smoothing parameters that are not far from their optimum. *Journal of the American Statistical Association* 87, 277–33.
- Härdle, W., and E. Mammen (1993). Comparing nonparametric versus parametric regression fits. *Annals of Statistics* 21, 1926–47.
- Hastie, T., and R. Tibshirani (1990). *General Additive Models*. New York: Chapman and Hall.
- Herrman, E., T. Gasser, and A. Kneip (1992). Choice of bandwidth for kernel regression when residuals are correlated. *Biometrika* 79, 783–95.
- Kuan, C.M., and H. White (1994). Artificial neural networks: An econometric perspective. *Econometric Reviews* 13, 1–91.
- Lee, B.J. (1994). Asymptotic distribution of the Ullah-type against the nonparametric alternative. *Journal of Quantitative Economics* 10, 73–92.
- Lewbel, A., and O. Linton (1998). Nonparametric censored regression. Manuscript, no. 1186, Yale University.
- Li, Q. (1997). Consistent model specification tests for time series models. Manuscript, University of Guelph.
- Li, Q., and C. Hsiao (1998). Testing serial correlation in semiparametric panel data models. *Journal of Econometrics* 87, 207–37.
- Li, Q., and T. Stengos (1996). Semiparametric estimation of partially linear panel data models. *Journal of Econometrics* 71, 389–97.

- Li, Q., and A. Ullah (1998). Estimating partially linear panel data models with one way error components. *Econometric Reviews* 17, 145–66.
- Li, Q., and S. Wang (1998). A simple consistent bootstrap test for a parametric regression function. *Journal of Econometrics* 87, 145–65.
- Li, Q., C. Huang, and T.T. Fu (1998). Semiparametric smooth coefficient stochastic frontier models. Manuscript, Institute of Economics, Taiwan.
- Linton, O., and D. Nielson (1995). Estimating structured nonparametric regression by the kernel method. *Biometrika* 82, 93–101.
- Linton, O., and P.L. Gozalo (1996). Testing additivity in generalized nonparametric regression models. Working Paper, Yale University and Brown University.
- Nadaraya, É.A. (1964). On estimating regression. *Theory of Probability and its Applications* 9, 141–2.
- Newey, W.K. (1985). Maximum likelihood specification testing and conditional moment tests. *Econometrica* 53, 1047–70.
- Newey, W.K. (1997). Convergence rates and asymptotic normality of series estimators. *Journal of Econometrics* 29, 147–68.
- Pagan, A.R., and A. Ullah (1999). *Nonparametric Econometrics*. Cambridge: Cambridge University Press.
- Phillips, P.C.B., and J.Y. Park (1998). Nonstationary density estimation and kernel autoregression. Manuscript, no. 1181, Yale University.
- Powell, J.L., H. Stock, and T.M. Stoker (1989). Semiparametric estimation of index coefficients. *Econometrica* 57, 1403–30.
- Prakasa Rao, B.L.S. (1983). *Nonparametric Functional Estimation*. New York: Academic Press.
- Racine, J. (1997). Consistent significance testing for nonparametric regression. *Journal of Business and Economic Statistics* 15, 369–78.
- Rahman, M., and A. Ullah (1999). Improved combined parametric and nonparametric regression: Estimation and hypothesis testing. Manuscript, University of California, Riverside.
- Rice, J. (1984). Bandwidth choice for nonparametric regression. *Annals of Statistics* 12, 1215–30.
- Robinson, P.M. (1988). Root-n-consistent semiparametric regression. *Econometrica* 56, 931–54.
- Robinson, P.M. (1986). On the consistency and finite sample properties of nonparametric kernel time series regression, autoregression and density estimation. *Annals of the Institute of Statistical Mathematics* 38, 539–49.
- Robinson, P.M. (1989). Nonparametric estimation of time varying parameters. In P. Hack (ed.) *Statistical Analysis and Forecasting of Economic Structural Change*. Springer-Verlag.
- Silverman, B.W. (1986). *Density Estimation for Statistics and Data Analysis*. New York: Chapman and Hall.
- Speckman, P. (1988). Kernel smoothing in a partial linear model. *Journal of Royal Statistical Society Series B* 50, 413–46.
- Ullah, A. (1985). Specification analysis of econometric models. *Journal of Quantitative Economics* 1, 187–209.
- Ullah, A., and N. Roy (1998). Nonparametric and semiparametric econometrics of panel data. In A. Ullah and D.E.A. Giles (eds.) *Handbook of Applied Economic Statistics*. ch. 17, pp. 579–604. Marcel Dekker.
- Vien, P. (1994). Choice of regressors in nonparametric estimation. *Computational Statistics and Data Analysis* 17, 575–94.
- Watson, G.S. (1964). Smooth regression analysis. *Sankhya Series A* 26, 359–72.
- Zheng, J.X. (1996). Consistent test of functional form via nonparametric estimation techniques. *Journal of Econometrics* 75, 263–90.

---

CHAPTER TWENTY-ONE

# Durations

*Christian Gouriéroux and Joann Jasiak*

## 1 INTRODUCTION

Duration data represent times elapsed between random arrivals of events. They play an important role in many areas of science such as engineering, management, physics, economics, and operational research. In economics, duration data frequently appear in labor and health studies, insurance analysis, and finance. For example, a commonly used duration-based statistic is the average individual lifetime, called the life expectancy, provided yearly by national surveys. It serves a variety of purposes. The macroeconomists quote it as an indicator of the level of development and welfare of the society, while applied microeconomists consider it implicitly in designing and pricing contracts such as life insurances. For specific projects, duration data are collected from a number of different sources. Various longitudinal studies are conducted on the national level to record durations of individual unemployment spells for job search studies. Data on durations of hospital treatments provide information on the anticipated expenses of the health care system. In academic research, the business cycle analysis and macroeconomic forecasting require studies of durations of recessions and expansions measuring times elapsed between subsequent turning points of the economy. Finally, in the private sector, businesses collect their own duration data. For example, insurance agencies record the times between reported car accidents to determine individual insurance premia or bonus–malus schemes, and learn about the attitude of their customers with respect to risk.

The probability theory often defines the distributional properties of durations with respect to the distribution of delimiting random events. In particular, the arrival frequency of these events has some major implications for research. For illustration, let us compare the durations of job searches to durations between consecutive transactions on a computerized stock market, like the New York Stock Exchange (NYSE). While a job search may last from a few days up to several months, a duration between trades may only amount to a fraction of a minute. We also expect that although the number of unemployment durations experienced by one person is theoretically unlimited, their total length cannot

exceed the maximum time during which the individual is able to actively participate in the labor force. In practice we do not observe more than a few unemployment spells per person, on average. For this reason researchers are mainly interested in cross section studies of unemployment durations based on a large number of individuals, rather than in investigating personal duration patterns. This is not the case of intertrade durations where each stock generates a series of durations consisting of thousands of observations per month. Such duration data are primarily interesting from the point of view of their dynamics and fall into a distinct category of duration time series. Therefore it is important to distinguish between duration models applied to panel and time series data.

The dynamics of durations is often related to transitions of a stochastic process between different admissible states. In this framework, a duration may be viewed as the sojourn time spent in a given state (unemployment) before exiting this state to enter into another state (employment). Besides the randomness related to the exit time, its destination may also be stochastic. There exist durations which may be terminated by events admitting several various states. An example of such a duration is the length of a hospital stay, which may end up in a recovery or a death of the patient. These durations data are called *transition data*.

An important issue in duration analysis concerns the measurement, or more precisely the choice of the time scale of reference. Several economic applications require time scales different from the conventional calendar time. The change of the time scale is called time deformation. The *operational* time unit is often selected with respect to some exogenous variables which may effect the speed of the time flow. Intuitively, an individual has a different perception of the time flow during busy working hours, and quiet periods of leisure. For objects like machines and instruments a typical determinant of the time speed is the depreciation rate. Accordingly, the time measuring the lifetime of a car flows at a different speed for a new car which leaves the assembly line, from an old car which has accumulated 100,000 km on the odometer. For this reason, a natural time scale of reference seems to be in this example the calendar time discounted by the mileage. As another example illustrating the economic sense of operational time, imagine an efficient stock trader who instead of measuring his time spent on the market floor in minutes is using instead time units necessary to trade, say, 1,000 shares or to make transactions worth 1,000 dollars. Obviously, deformed time does not have equal, unitary increments, but it resembles the calendar time in that it cannot stop or reverse its direction.

Finally, note that in everyday life we often encounter durations arising as conditions specified by various contracts, such as lease agreements, rentals, or credit terms. As such these predetermined durations are not of interest to analysts, who examine durations between events which are intrinsically random. However the randomness reappears if a side of the contract is allowed to quit by early termination, i.e. when, for example, borrowers have the option to prepay the outstanding credit balances. Given that not all individuals display the same behavior, this population is considered by a duration analyst as a heterogenous one.

This chapter is organized as follows. In Section 2, we discuss the standard characterizations of duration variables and present the main duration distribution

families. In Section 3, we introduce individual heterogeneity in parametric duration models. This heterogeneity is partly observed through individual explanatory variables and partly unobserved. We discuss the effect of unobserved heterogeneity in terms of negative duration dependence. Section 4 covers semiparametric models with a parametric effect of observed explanatory variables, and an unspecified baseline duration distribution. Finally, we introduce in Section 5 dynamic models for the analysis of time series of durations which are especially useful for applications to financial transactions data.

## 2 DURATION VARIABLES

In this section we introduce basic concepts in duration analysis and present the commonly used duration distributions.

### 2.1 Survivor and hazard functions

Let us consider a continuous duration variable  $Y$  measuring the time spent in a given state, taking values in  $R^+$ . The probabilistic properties of  $Y$  can be defined either by:

the probability density (pdf) function  $f(y)$ , assumed strictly positive,  
or the cumulative distribution (cdf) function  $F(y) = \int_0^y f(u)du$ ,  
or the survivor function  $S(y) = 1 - F(y) = \int_y^\infty f(u)du$ .

The *survivor function* gives the probability of survival to  $y$ , or otherwise, the chance of remaining in the present state for at least  $y$  time units. Essentially, the survivor function concerns the future.

In many applications the exit time has an economic meaning and may signify a transition into a desired or undesired state. Let us pursue the example of individual life expectancy. A related important indicator is the instantaneous mortality rate at age  $y$ . It is given by:

$$\lambda(y) = \lim_{dy \rightarrow 0} \frac{1}{dy} P[y \leq Y < y + dy | Y \geq y]. \quad (21.1)$$

In this formula  $\lambda(y)$  defines the probability per unit of time that a person dies within a short interval of  $dy$  (seconds) given that he/she is still alive at age  $y$ . It can easily be written in terms of the survivor function. Indeed we get:

$$\begin{aligned} \lambda(y) &= \lim_{dy \rightarrow 0} \frac{1}{dy} \frac{P[y \leq Y < y + dy]}{P[Y \geq y]} \\ &= \lim_{dy \rightarrow 0} \frac{1}{dy} \frac{S(y) - S(y + dy)}{S(y)} \\ &= -\frac{1}{S(y)} \frac{dS(y)}{dy}, \end{aligned}$$

$$\lambda(y) = \frac{f(y)}{S(y)}. \quad (21.2)$$

The *hazard function* is:

$$\lambda(y) = \frac{f(y)}{S(y)} = \lim_{dy \rightarrow 0} \frac{1}{dy} P[y < Y < y + dy | Y \geq y], \quad \forall y \in R^+. \quad (21.3)$$

It gives the instantaneous exit rate per unit of time evaluated at  $y$ . Among often encountered exit rates are, besides the aforementioned mortality rate, the bankruptcy rate, and the failure rate of instruments.

The duration variable can equivalently be defined by  $S, f$  or  $\lambda$ , in reason of the following relationship between the survivor function and the hazard function:

$$S(y) = \exp - \int_0^y \lambda(u) du. \quad (21.4)$$

This means that once we know the hazard function we can always find the survivor function.

## 2.2 Duration dependence

The duration dependence describes the relationship between the exit rate and the time already spent in the state. Technically it is determined by the hazard function, which may be a decreasing, increasing, or constant function of  $y$ . Accordingly, we distinguish (i) negative duration dependence; (ii) positive duration dependence; and (iii) absence of duration dependence.

### NEGATIVE DURATION DEPENDENCE

The longer the time spent in a given state, the lower the probability of leaving it soon. This negative relationship is found for example in the job search analysis. The longer the job search lasts, the less chance an unemployed person has of finding a job.

### POSITIVE DURATION DEPENDENCE

The longer the time spent in a given state, the higher the probability of leaving it soon. Positive duration dependence is observed in the failure rate of instruments which are getting used up in time, or depreciate gradually. For example, the longer a lightbulb works, the higher the probability that it fails within the next hour (say).

### ABSENCE OF DURATION DEPENDENCE

The hazard function is constant. In this case there is no relationship between the duration spent in the state and the probability of exit.

The absence of duration dependence is often imposed as a simplifying although very restrictive assumption. It implies that items, such as instruments or

machines, do not deteriorate: for example, an item, which has been in use for ten hours, is as good as a new item with regard to the amount of time remaining until the item fails. This effect is usually not supported by the data and durations observed in empirical research usually belong to the first or second category or else display a nonmonotone hazard function. For this reason empirical hazards often need to be studied case by case. Let us consider, for instance, a typical hazard function representing the rate of bankruptcies. A newly created firm has a low probability of failure; however six months to two years later the failure rate increases sharply. The bankruptcy rate usually shows a tendency to diminish for companies which operate for a fairly long time, acquire more experience, and become better known to their customers and suppliers.

### 2.3 Basic duration distributions

In this section we introduce some parametric families of duration distributions.

#### EXPONENTIAL FAMILY

The exponentially distributed durations feature no duration dependence. In consequence of the time-independent durations, the hazard function is constant,  $\lambda(y) = \lambda$ . The cdf is given by the expression  $F(y) = 1 - \exp(-\lambda y)$ , and the survivor function is  $S(y) = \exp(-\lambda y)$ .

The density is given by:

$$f(y) = \lambda \exp(-\lambda y), \quad y > 0. \quad (21.5)$$

This family is parametrized by the parameter  $\lambda$  taking strictly positive values.

An important characteristic of the exponential distributions is that the mean and standard deviation are equal, as implied by  $EY = \frac{1}{\lambda}$ ,  $VY = \frac{1}{\lambda^2}$ . In empirical research the data violating this condition are called over- or under-dispersed depending on whether the standard deviation exceeds the mean or is less than the mean.

#### GAMMA FAMILY

This family of distributions depends on two positively valued parameters  $a$  and  $v$ . The density is given by:

$$f(y) = [a^v y^{v-1} \exp(-ay)] / \Gamma(v), \quad (21.6)$$

where  $\Gamma(v) = \int_0^\infty \exp(-y)y^{v-1}dy$ . When  $v = n$  is integer valued, this distribution may be obtained by summing  $n$  independent exponentially distributed durations with parameter  $\lambda = a$ . In such a case  $\Gamma(n) = (n - 1)!$ .

The form of the hazard function depends on the parameter  $v$ .

1. If  $v > 1$ , the hazard function is decreasing and approaching asymptotically  $a$ .
2. If  $v = 1$ , the hazard function is a constant, and the model reduces to the exponential model.
3. If  $v < 1$ , the hazard function is decreasing from  $+\infty$  and approaches an asymptote at  $a$ .

The gamma model is quite often employed in practice. The mean and variance of *gamma* distributed durations are  $EY = \frac{v}{a}$ ,  $VY = \frac{v}{a^2}$ .

### WEIBULL MODEL

This family of distributions also depends on two positive parameters  $a$  and  $b$ . The density is:

$$f(y) = aby^{b-1} \exp(-ay^b). \quad (21.7)$$

The formula of the survivor function is  $S(y) = \exp(-ay^b)$ . The behavior of the hazard function  $\lambda(y) = aby^{b-1}$  is determined by  $b$ . It is increasing for  $b > 1$  at either a growing or diminishing rate, and it is decreasing for values of  $b < 1$ .

### LOGNORMAL FAMILY

These distributions, contrary to those discussed above, have a nonmonotone hazard function which is first increasing, and next decreasing in  $y$ . Therefore they can be used for the analysis of bankruptcy rates. The lognormal duration distribution is such that  $\log Y$  follows a normal distribution with mean  $m$  and variance  $\sigma^2$ . The density is a function of the normal density denoted by  $\phi$ :

$$f(y) = \frac{1}{\sigma y} \phi\left(\frac{\log y - m}{\sigma}\right). \quad (21.8)$$

The survivor function is  $S(y) = 1 - \Phi(\frac{\log y - m}{\sigma})$ , where  $\Phi$  denotes the cdf of a standard normal. The hazard function can be written as the ratio:

$$\lambda(y) = \frac{1}{y} \frac{[1/\sigma\phi(x)]}{1 - \Phi(x)},$$

where  $x = \frac{\log y - m}{\sigma}$ .

## 3 PARAMETRIC MODELS

In empirical research we may wish to investigate the dependence of individual hazard functions on exogenous variables. These variables, called the control variates, depict in general various individual characteristics. Let us point out a few examples. In the job search analysis, a typical control variate is the amount of unemployment benefits, which influences the effort of unemployed individuals devoted to the job search and consequently the duration of unemployment. Empirical findings also suggest that family support provided by the state influences the birth rate, and that the expected increase of the insurance premium has an effect on the frequency of declared car accidents. As well, there is evidence indicating that the lengths of hospital stays depend on the cost incurred by patients, or else, the duration of an outstanding balance on a credit card is in part determined by the interest paid by the cardholder. Some explanatory variables

differ across individuals, and are invariant in time (e.g. gender), while others (e.g. age) are individual and time dependent. Such variables need to be doubly indexed by individual and time (see Hsiao, Chapter 16, in this volume). Other variables may have a common impact on all individuals in the sample and vary in time, like the rate of inflation or the global rate of unemployment.

Parametric duration models can accommodate the effect of observable or unobservable individual characteristics on durations. Let us denote by  $x_i$  the observable explanatory variables and by  $\mu_i$  a latent heterogeneity factor. We proceed in two steps to define the extended duration model. First, we consider the conditional distribution of the duration variable  $Y_i$  given the observable covariates and heterogeneity. It is characterized by either the conditional pdf  $f(y_i | x_i, \mu_i)$ , or the conditional hazard function  $\lambda(y_i | x_i, \mu_i)$ . Next, we introduce a heterogeneity distribution  $\pi(\mu_i)$  (say), which is used to derive the conditional distribution of the duration variable given the observable covariates only. This latter distribution is characterized by either the conditional pdf  $f(y_i | x_i)$ , or the conditional hazard function  $\lambda(y_i | x_i)$ .

In the first subsection we describe the exponential duration model without heterogeneity and its estimation by the maximum likelihood. In the following subsection we introduce a gamma distributed heterogeneity factor, which leads us to the Pareto regression model. The effect of unobservable heterogeneity and its relationship with the negative duration dependence are covered in the third part of this section. Finally, we discuss the problem of partial observability of duration variables due to truncation or censoring effects.

### 3.1 Exponential regression model

Recall that the exponential duration model depends on the parameter  $\lambda$ , which is the constant hazard rate. We now assume an exponential distribution for each individual duration, with a rate  $\lambda_i$  depending on the observable characteristics of this individual represented by explanatory variables. The positive sign of  $\lambda$  is ensured by assuming that:

$$\lambda_i = \exp(x_i \theta),$$

where  $\theta$  is a vector of unknown parameters. The survivor function is given by:

$$S_i(y | x_i; \theta) = \exp[-(\exp x_i \theta)y],$$

whereas the conditional pdf of the duration variable given the covariates is:

$$\begin{aligned} f(y_i | x_i; \theta) &= \lambda_i \exp(-\lambda_i y_i) \\ &= \exp(x_i \theta) \exp[-y_i \exp(x_i \theta)]. \end{aligned} \tag{21.9}$$

The parameter  $\theta$  can be estimated by the maximum likelihood from a random sample of  $N$  observations on  $(x_i, y_i)$ ,  $i = 1, \dots, N$ . The conditional loglikelihood function is:

$$\begin{aligned}\log l(y | x; \theta) &= \sum_{i=1}^N \log f(y_i | x_i; \theta) \\ &= \sum_{i=1}^N [x_i \theta - y_i \exp(x_i \theta)],\end{aligned}$$

and the maximum likelihood estimator  $\hat{\theta} = \text{Argmax}_{\theta} \log l(y | x; \theta)$  solves the optimization problem. The first order conditions are:

$$\begin{aligned}\frac{\partial \log l(y | x, \hat{\theta})}{\partial \theta} &= 0 \\ \Leftrightarrow \sum_{i=1}^N [1 - y_i \exp(x_i \hat{\theta})] x'_i &= 0 \\ \Leftrightarrow \sum_{i=1}^N \exp(x_i \hat{\theta}) [y_i - \exp(-x_i \hat{\theta})] x'_i &= 0.\end{aligned}\tag{21.10}$$

Since the conditional expectation of the duration variable is  $E[Y_i | x_i] = \exp(-x_i \hat{\theta})$ , the first-order equations are equivalent to orthogonality conditions between the explanatory variables and the residuals:  $\hat{u}_i = y_i - \exp(-x_i \hat{\theta})$ , with weights  $\exp(x_i \hat{\theta})$  due to the individual heteroskedasticity.

### 3.2 The exponential model with heterogeneity

The exponential regression model can easily be extended by introducing unobservable variables. We express the individual hazard rate as:

$$\lambda_i = \mu_i \exp(x_i \theta),\tag{21.11}$$

where  $\mu_i$  is a latent variable representing the heterogeneity of individuals in the sample, called the heterogeneity factor. We assume that the heterogeneity factor is gamma distributed  $\gamma(a, a)$ , with two identical parameters to ensure  $E\mu_i = 1$ . The conditional duration distribution given the observable covariates is found by integrating out the unobservable heterogeneity.

$$\begin{aligned}f(y_i | x_i; \theta, a) &= \int_0^\infty f(y_i | x_i, \mu; \theta) \pi(\mu; a) d\mu \\ &= \int_0^\infty \mu \exp(x_i \theta) \exp[-y_i \mu \exp(x_i \theta)] \frac{a^a \mu^{a-1} \exp(-a\mu)}{\Gamma(a)} d\mu \\ &= \frac{a^a \exp(x_i \theta)}{[a + y_i \exp(x_i \theta)]^{a+1}} \frac{\Gamma(a+1)}{\Gamma(a)} \\ f(y_i | x_i; \theta, a) &= \frac{a^{a+1} \exp(x_i \theta)}{[a + y_i \exp(x_i \theta)]^{a+1}}.\end{aligned}\tag{21.12}$$

We find that the conditional duration distribution is Pareto translated. The associated conditional survivor function is:

$$\begin{aligned} S(y_i | x_i; \theta, a) &= \int_{y_i}^{\infty} \frac{a^{a+1} \exp(x_i \theta)}{[a + \mu \exp(x_i \theta)]^{a+1}} du \\ &= \frac{a^a}{[a + y_i \exp(x_i \theta)]^a}, \end{aligned}$$

whereas the hazard function is:

$$\lambda(y_i | x_i; \theta, a) = \frac{a \exp(x_i \theta)}{a + y_i \exp(x_i \theta)}.$$

The hazard function of the Pareto distribution with drift is a decreasing function of  $y$ , and features negative duration dependence at the level of a representative individual. Hence, by aggregating exponentially distributed durations with constant hazards across infinitely many different individuals with a gamma distributed heterogeneity, we obtain a decreasing aggregate hazard function. The heterogeneity parameter  $a$  provides a natural measure of the negative duration dependence: the smaller  $a$ , the stronger the negative duration dependence. In the limiting case  $a = +\infty$ , we get  $\mu_i = 1$ , and  $\lambda(y_i | x_i; \theta, a) = \exp(x_i \theta)$ ; there is no duration dependence and the Pareto regression model reduces to the exponential regression model.

For the Pareto regression model, the loglikelihood function is:

$$\begin{aligned} \log l(y | x; \theta, a) &= \sum_{i=1}^N \log f(y_i | x_i; \theta, a) \\ &= \sum_{i=1}^N \{(a+1)\log a + x_i \theta - (a+1)\log [a + y_i \exp(x_i \theta)]\}. \end{aligned}$$

### 3.3 Heterogeneity and negative duration dependence

The effect of unobservable covariates can be measured by comparing models with and without heterogeneity. In this section we perform such a comparison using the exponential model. For simplicity we do not include observable covariates in the model. The conditional distribution of the duration variable given the heterogeneity factor  $\mu_i$  is exponential with parameter  $\lambda_i = \mu_i$  whereas the marginal distribution of the heterogeneity is  $\pi$ . Therefore, the conditional and marginal survivor functions are:

$$\begin{aligned} S(y_i | \mu_i) &= \exp(-\mu_i y_i), \\ S(y_i) &= \int_0^{\infty} \exp(-\mu y_i) \pi(\mu) d\mu. \end{aligned}$$

The corresponding hazard functions are:

$$\lambda(y_i | \mu_i) = \mu_i,$$

$$\begin{aligned}\lambda(y_i) &= -\frac{d \log S(y_i)}{dy} = -\frac{1}{S(y_i)} \frac{dS(y_i)}{dy} \\ &= \frac{\int_0^\infty \exp(-\mu y_i) \mu \pi(\mu) d\mu}{\int_0^\infty \exp(-\mu y_i) \pi(\mu) d\mu}.\end{aligned}$$

The marginal hazard rate is an average of the individual hazard rates  $\mu_i$  with respect to a modified probability distribution with pdf:

$$\pi_{y_i}(\mu_i) = \exp(-\mu y_i) \pi(\mu) / \int_0^\infty \exp(-\mu y_i) \pi(\mu) d\mu. \quad (21.13)$$

We also get:

$$\lambda(y_i) = E_{\pi_{y_i}}[\lambda(y_i | \mu_i)] = E_{\pi_{y_i}}(\mu_i). \quad (21.14)$$

This marginal hazard function features negative duration dependence. Indeed, by taking the first-order derivative we find:

$$\begin{aligned}\frac{d\lambda(y_i)}{dy} &= \frac{-\int_0^\infty \mu^2 \exp(-\mu y_i) \pi(\mu) d\mu}{\int_0^\infty \exp(-\mu y_i) \pi(\mu) d\mu} + \frac{[\int_0^\infty \exp(-\mu y_i) \mu \pi(\mu) d\mu]^2}{[\int_0^\infty \exp(-\mu y_i) \pi(\mu) d\mu]^2} \\ &= -E_{\pi_{y_i}} \mu_i^2 + [E_{\pi_{y_i}}(\mu_i)]^2 \\ &= -\text{var}_{\pi_{y_i}} \mu_i \leq 0.\end{aligned}$$

We note that the negative duration dependence at level  $y_i$  is related to the magnitude of heterogeneity with respect to a modified probability.

To illustrate previous results let us consider a sample of individuals belonging to two categories with respective exit rates  $\mu_1 > \mu_2$ . The individuals in the first category with a high exit rate are called movers, whereas we call stayers the individuals belonging to the second category. The structure of the whole population at date 0 is  $\pi_1 = \pi$ ,  $\pi_2 = 1 - \pi$ . The marginal hazard rate derived in the previous section becomes:

$$\lambda(y) = \frac{\pi_1 S_1(y) \mu_1 + \pi_2 S_2(y) \mu_2}{\pi_1 S_1(y) + \pi_2 S_2(y)}. \quad (21.15)$$

Between 0 and  $y$  some individuals exit from the population. The proportions of those who leave differ in the two subpopulations; they are given by  $S_1(y) = \exp(-\mu_1 y) < S_2(y) = \exp(-\mu_2 y)$ , which implies a modified structure of remaining individuals at date  $y$ . This modified structure is:

$$\pi_1(y) = \pi_1 S_1(y) / [\pi_1 S_1(y) + \pi_2 S_2(y)], \quad \pi_2(y) = 1 - \pi_1(y). \quad (21.16)$$

Since  $S_1(y) < S_2(y)$ , the proportion of movers is lower at date  $y$  than at date 0, which implies  $\lambda(y) < \lambda(0) = \pi_1 \mu_1 + \pi_2 \mu_2$ . Finally, we note that, for large  $y$ ,  $\pi_2(y)$  tends to one and the remaining population becomes homogenous including stayers only.

### 3.4 Truncation and censoring

Econometric data used in duration analysis are often panel data comprising a number of individuals observed over a fixed interval of time. Let us suppose that the survey concerns unemployment durations; the sampling period is January 2000–December 2000 and the individuals also provided information on their job history prior to January 2000. We can consider two different sampling schemes, which imply truncation and censoring.

#### CENSORING

Let us first consider a sample drawn from the population including both employed and unemployed people, and assume at most one unemployment spell per individual. Within this sample we find persons, who:

1. are unemployed in January and remain unemployed in December too;
2. are unemployed in January and find a job before December;
3. are employed in January, lose their job before December and are still unemployed at this date;
4. are employed in January, next lose their job and find new employment before December.

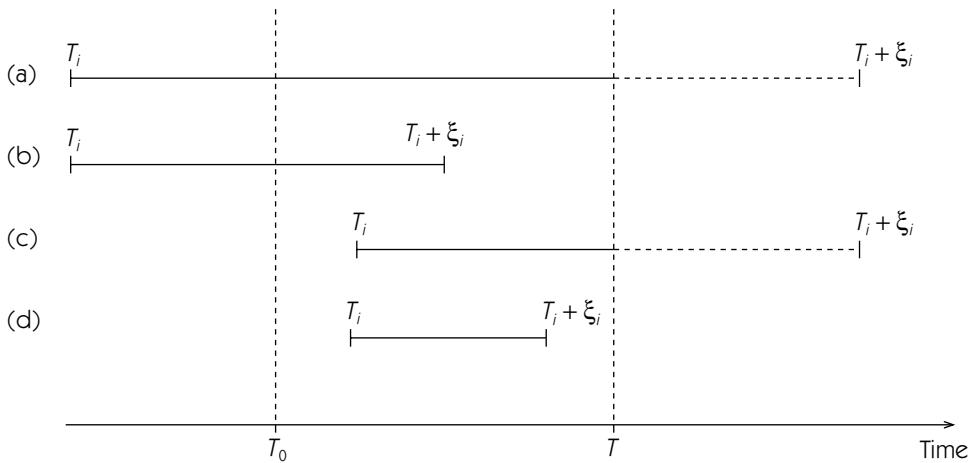
Due to the labor force dynamics, unemployment durations of some individuals are only partially observed. For groups (2) and (4) the unemployment spells are complete, whereas they are right censored for groups (1) and (3).

To identify the right censored durations we can introduce an indicator variable  $d_i$ . It takes value 1 if the observed duration spell for individual  $i$  is complete, and 0 if this observation is right censored. We also denote by  $T_i$  the date of the entry into the unemployment state, by  $\xi_i$  the total unemployment duration and by  $y_i$  the *observed* unemployment duration knowing that the sampling period ends at  $T$ .

The model involves two latent variables  $T_i$  and  $\xi_i$ . The observed variables  $d_i$  and  $y_i$  are related to the latent variables by:

$$\begin{cases} d_i = 1 \\ y_i = \xi_i \end{cases}, \quad \text{if } T_i + \xi_i < T,$$

$$\begin{cases} d_i = 0 \\ y_i = T - T_i \end{cases}, \quad \text{if } T_i + \xi_i > T,$$



**Figure 21.1** Censoring scheme: unemployment spells

Conditional on  $T_i$  the density of the observed pair  $(y_i, d_i)$  is:

$$l_i(y_i, d_i) = f_i(y_i)^{d_i} S_i(y_i)^{1-d_i}, \quad (21.17)$$

or, otherwise:

$$l_i(y_i, d_i) = \lambda_i(y_i)^{d_i} S_i(y_i), \quad (21.18)$$

by substituting the hazard expression into equation (21.17). The loglikelihood function for this model can be written by assuming that individual durations are independent conditional on explanatory variables:

$$\begin{aligned} \log L(y; d) &= \sum_{i=1}^N \log l_i(y_i; d_i) \\ &= \sum_{i=1}^N d_i \log \lambda_i(y_i) + \sum_{i=1}^N \log S_i(y_i). \end{aligned}$$

Note that the duration distributions are conditioned on the date  $T_i$ . This information has generally to be introduced among the explanatory variables to correct for the so-called cohort effect.

## TRUNCATION

We can also draw the sample in the subpopulation of people who are unemployed in January 2000 (date  $T_0$ , say). Within this sample we find persons, who:

1. are unemployed in January and remain unemployed in December too;
2. are unemployed in January and find a job before December.

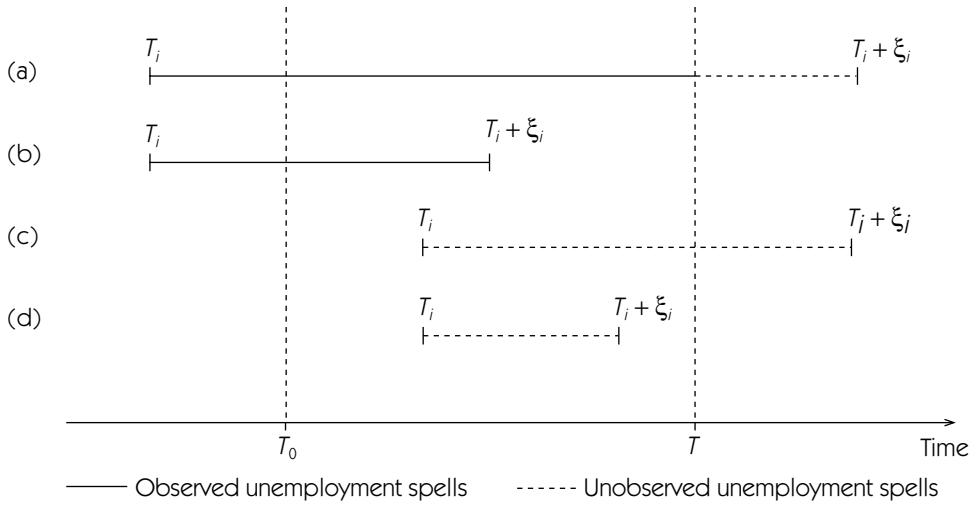


Figure 21.2 Truncation scheme

However, we now need to take into account the endogenous selection of the sample, which only contains unemployed people at  $T_0$  (see Lung-Fei Lee, Chapter 16, in this volume). This sampling scheme is called left truncated, since compared to the previous scheme we have only retained the individuals with unemployment duration larger than  $T_0 - T_i$ . Conditional on  $T_i$ , the density of the pair  $(y_i, d_i)$  becomes:

$$l_i(y_i, d_i) = f_i(y_i)^d S_i(y_i)^{1-d} / S_i(T_0 - T_i).$$

## 4 SEMIPARAMETRIC MODELS

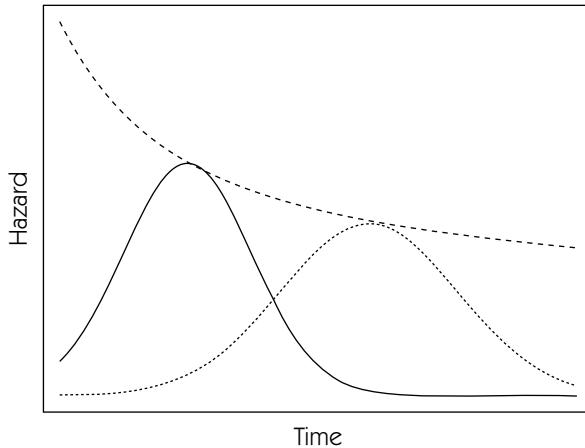
It is common to consider semiparametric specifications of duration models which distinguish a parametric scoring function and an unconstrained baseline distribution. We describe below the accelerated and proportional hazard models and introduce the estimation methods for the finite dimensional and functional parameters. For convenience we select an exponential specification of the score.

### 4.1 Accelerated hazard model

Accelerated hazard models rely on the assumption that individual durations follow the same distribution up to an individual change of the time scale (time deformation). For example, let us rescale the time by  $\exp(x_i \theta)$ , we get:

$$Y_i \exp(x_i \theta) \sim f_0, \quad (21.19)$$

where  $f_0$  is the unconstrained baseline distribution and  $\exp(x_i \theta)$  defines the change of the time unit. We deduce that the conditional hazard function given the observable covariates is:



**Figure 21.3** Hazard functions for accelerated hazard models

$$\lambda(y_i | x_i; \theta, f_0) = \lambda_0[y_i \exp(x_i \theta)] \exp(x_i \theta), \quad (21.20)$$

where the baseline hazard  $\lambda_0$  corresponds to the  $f_0$  distribution.

A consistent estimation method of the two types of parameters  $\theta$  and  $f_0$  is easily derived. Indeed from (21.20) we derive:

$$\log Y_i = -x_i \theta + u_i, \quad (21.21)$$

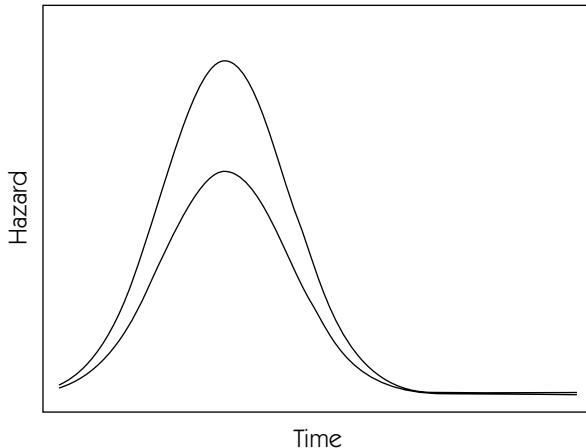
where the variables  $\exp u_i$  are iid with the unknown distribution  $f_0$ . By introducing a constant term among the regressors, it is always possible to constrain the distribution  $f_0$  so that  $E u_i = 0$ . The finite dimensional parameter  $\theta$  can be consistently estimated by ordinary least squares from a regression of the log-durations on the covariates  $x_i$ . Let us denote by  $\hat{\theta}$  the OLS estimator. In the next step we can construct the corresponding residuals  $\hat{u}_i = \log y_i + x_i \hat{\theta}$  and approximate the baseline distribution by the smoothed distribution of the exponential residuals  $\exp \hat{u}_i = y_i \exp x_i \hat{\theta}$ ,  $i = 1, \dots, N$ .

## 4.2 Proportional hazard model

In this model the conditional hazard functions are assumed homothetic and the parametric term  $\exp(x_i \theta)$  is introduced as the coefficient of proportionality:

$$\lambda(y_i | x_i; \theta, f_0) = \exp(x_i \theta) \lambda_0(y_i), \quad (21.22)$$

where  $\lambda_0$  is an unconstrained baseline hazard function. It is defined up to a multiplicative scalar. The term proportional hazard indicates that the hazards for two individuals with regressor vectors  $x_1$  and  $x_2$  are in the same ratio. For



**Figure 21.4** Hazard functions for proportional hazard models

example, the exponential model with  $\lambda(y|x_i; \theta) = \exp(x_i\theta)$  is a proportional hazard model with  $\lambda_0 = 1$ .

The parameter  $\theta$  can be consistently estimated by the partial maximum likelihood introduced by Cox (1975). The approach is the following. Let us first rank the duration data by increasing values  $y_{(1)} < y_{(2)} < \dots < y_{(N)}$ , where we implicitly assume that all observed durations are different. Then we consider the subpopulation at risk just before the exit of the  $i$ th individual:

$$\mathcal{R}_{(i)} = \{j : y_{(j)} \geq y_{(i)}\}.$$

The probability that the first individual escaping from this subpopulation  $\mathcal{R}_{(i)}$  is individual  $(i)$  is given by:

$$\begin{aligned} p_{(i)}(\theta, \lambda_0) &= \frac{\lambda(y_{(i)}|x_{(i)}; \theta, \lambda_0)}{\sum_{j \in \mathcal{R}_{(i)}} \lambda[y_{(j)}|x_{(j)}; \theta, \lambda_0]} \\ &= \frac{\exp(x_{(i)}\theta)}{\sum_{j \in \mathcal{R}_{(i)}} \exp(x_{(j)}\theta)}. \end{aligned}$$

It no longer depends on the baseline distribution. The partial maximum likelihood estimation of  $\theta$  is defined by:

$$\begin{aligned} \hat{\theta} &= \underset{\theta}{\operatorname{argmax}} \sum_{i=1}^N \log p_{(i)}(\theta, \lambda_0) \\ &= \underset{\theta}{\operatorname{argmax}} \sum_{i=1}^N \log \frac{\exp(x_{(i)}\theta)}{\sum_{j \in \mathcal{R}_{(i)}} \exp(x_{(j)}\theta)}. \end{aligned} \tag{21.23}$$

## 5 DURATION TIME SERIES

In this section we focus our attention on duration time series, i.e. sequences of random durations, indexed by their successive numbers in the sequence and possibly featuring temporal dependence. In practice these data are generated, for example, by randomly occurring transactions on credit cards, by claims randomly submitted to insurance agencies at unequal intervals, or by assets traded at a time varying rate on stock markets. According to the traditional time series analysis the ultimate purpose of our study is to model and estimate the dynamics of these stochastic duration processes.

There are two major characteristics which account for the distinct character of duration time series. Unlike the familiar time series data, duration sequences are not indexed by time, but as mentioned earlier, by numbers indicating their position in the sequence. Such indices are necessarily integer valued, and in this respect dynamic durations belong to traditional discrete time series observed at a fixed frequency. Since the duration indices correspond to arrivals of some random events, researchers often employ the notion of an operational time scale with unitary increments set by the event arrivals.

Unlike the duration data discussed earlier in the text, duration time series do not represent patterns exhibited by a sample of individuals over a fixed span of calendar time. For example, in a sample of unemployed people, we may encounter individuals who, during the sampling period, experienced not one, but several unemployment spells. Yet, the study aims at finding the probabilistic structure of durations common to all individuals. At the individual level too few consecutive durations are available for inference on the dynamics, anyway. In contrast, the time series of durations represent times between many outcomes of the same repeated experiment (for example trading a stock), and always concern the same statistical individual (for example the IBM stock).

While the traditional duration analysis is essentially applied to cross section data, the analysis of stock trading dates, claim arrivals, or transactions on a bank account require a time series approach adapted to the specific features of durations. This field of research is quite recent. It originated from a growing interest in quote-by-quote data provided by electronic systems implemented on financial markets and has been developed in parallel to progressing computer capacities allowing for treatment of large data sets. The number of transactions on a particular stock in a stock market concluded during one day may be very large indeed and easily exceed several hundred thousands.

We begin this section with insights into the dynamics of durations in the simple case of the Poisson process. Next we cover some recent developments in this field including the Autoregressive Conditional Duration (ACD) model and the Stochastic Volatility Duration (SVD) model.

### 5.1 The Poisson process

There exist two alternative ways to study a sequence of event arrivals. First we can consider the sequence of arrival dates or equivalently the sequence of durations

$Y_1, \dots, Y_n, \dots$  between consecutive events. Secondly, we can introduce the counting process  $[N(t), t \text{ varying}]$ , which counts the number of events observed between 0 and  $t$ . The counting process is a jump process, whose jumps of unitary size occur at each arrival date. It is equivalent to know the sequence of durations or the path of the counting process.

The Poisson process is obtained by imposing the following two conditions:

1. The counting process has independent increments, i.e.  $N(t_n) - N(t_{n-1}), N(t_{n-1}) - N(t_{n-2}), \dots, N(t_1) - N(t_0)$  are independent for any  $t_0 < t_1 < t_2 < \dots < t_n$ , and any  $n$ .
2. The rate of arrival of an event (a jump) is constant and two events cannot occur in a small time interval:

$$P[N(t + dt) - N(t) = 1] = \lambda dt + o(dt), \quad (21.24)$$

$$P[N(t + dt) - N(t) = 0] = 1 - \lambda dt + o(dt), \quad (21.25)$$

where the term  $o(dt)$  is a function of  $dt$  such that  $\lim_{dt \rightarrow 0} o(dt)/dt = 0$ .

Under these assumptions it is possible to deduce the distributions of the counting process and of the sequence of durations.

1. For a Poisson process, the durations  $Y_i, i = 1, \dots, n$  are independent, identically exponentially distributed with parameter  $\lambda$ .
2. For a Poisson process, the increments  $N(t_2) - N(t_1)$ , with  $t_2 > t_1$ , follow Poisson distributions, with parameters  $\lambda(t_2 - t_1)$ .

This result explains why basic duration models are based on exponential distributions, whereas basic models for count data are based on Poisson distributions, and how these specifications are related (see Cameron and Trivedi, Chapter 15, in this volume).

Similarly, a more complex dynamics can be obtained by relaxing the assumption that either the successive durations, or the increments of the counting process are independent. In the following subsections we introduce the temporal dependence duration sequences.

## 5.2 The ACD model

This model was introduced by Engle and Russell (1998) to represent the dynamics of durations between trades on stock or exchange rate markets. Typically, intertrade durations are generated by a computerized order matching system which automatically selects trading partners who satisfy elementary matching criteria. Therefore, the timing of such automatically triggered transactions is a priori unknown and adds a significant element of randomness to the trading process. From the economic point of view, research on intertrade durations is motivated by the relevance of the time varying speed of trading for purposes

such as strategic market interventions. Typically the market displays episodes of accelerated activity and slowdowns which reflect the varying *liquidity* of the asset. This concept is similar to the notion of velocity used in monetary macroeconomics to describe the rate of money circulation (Gouriéroux, Jasiak, and Le Fol, 1999). In the context of stock markets, periods of intense trading are usually associated with short intertrade durations accompanied by high price volatilities and large traded volumes.

The intertrade durations plotted against time display dynamic patterns similar to stock price volatilities, termed in the literature the *clustering* effect. This means that long durations have a tendency to be followed by long durations, while short durations tend to be followed by short durations. To capture this behavior Engle and Russell (1998) proposed the ACD model. It accommodates the duration clustering in terms of temporal dependence of conditional means of durations. From the time series perspectives it is an analog of the GARCH model representing serial correlation in conditional variances of stock prices (see Engle, 1982) and from the duration analysis point of view it is an accelerated hazard model.

Let  $N$  be the number of events observed at random times. The  $N$  events are indexed by  $i = 1, \dots, N$  from the first observed event to the last. The  $i$ th duration is the time between the  $i$ th event and the  $(i - 1)$ th event. The distribution of the sequence of durations is characterized by the form of the conditional distribution of the duration  $Y_i$  given the lagged durations  $\underline{Y}_{i-1} = \{Y_{i-1}, Y_{i-2}, \dots\}$ . The ACD( $p, q$ ) model is an accelerated hazard model, where the effect of the past is summarized by the conditional expectation  $\psi_i = E(Y_i | \underline{Y}_{i-1})$ :

$$f(y_i | \underline{y}_{i-1}) = \frac{1}{\psi_i} f_0\left(\frac{y_i}{\psi_i}\right), \quad (21.26)$$

where  $f_0$  is a baseline distribution with unitary mean, and the conditional mean  $\psi_i$  satisfies:

$$\psi_i = w + \alpha(L)Y_i + \beta(L)\psi_i, \quad (21.27)$$

where  $L$  denotes the lag operator,  $\alpha(L) = \alpha_1 L + \alpha_2 L^2 + \dots + \alpha_q L^q$  and  $\beta(L) = \beta_1 L + \beta_2 L^2 + \dots + \beta_p L^p$  are lag polynomials of degrees  $p$  and  $q$  respectively. The coefficients  $\alpha_j, j = 1, \dots, q$ ,  $\beta_j, j = 1, \dots, p$  are assumed to be nonnegative to ensure the positivity of  $\psi_i$ . They are unknown and have to be estimated jointly with the baseline distribution. This specification implies that the effect of past durations on the current conditional expected value decays exponentially with the lag length. Indeed, the ACD( $p, q$ ) process may be rewritten as an ARMA( $m, p$ ) process in  $Y_i$ :

$$[1 - \alpha(L) - \beta(L)]Y_i = w + [1 - \beta(L)]v_i, \quad (21.28)$$

where  $m = \max(p, q)$ , and  $v_i = Y_i - \psi_i = Y_i - E(Y_i | \underline{Y}_{i-1})$  is the innovation of the duration process. The stationarity condition requires that the roots of  $[1 - \alpha(L) - \beta(L)]$  and  $[1 - \beta(L)]$  lie outside the unit circle, or equivalently since  $\alpha_j, \beta_j$  are nonnegative ( $\sum_j \alpha_j + \sum_j \beta_j < 1$ ).

The baseline distribution can be left unspecified or constrained to belong to a parametric family, such as the Weibull family.

The ACD model was a pioneering specification in the domain of duration dynamics. Further research has focused on providing refinements of this model in order to improve the fit. Empirical results show for example that many duration data display autocorrelation functions decaying at a slow, hyperbolic rate. Indeed, the range of temporal dependence in durations may be extremely large suggesting that duration processes possess long memory. This empirical finding is at odds with the exponential decay rate assumed by construction in the ACD model. To accommodate the long persistence, a straightforward improvement consists in accounting for fractional integration in the ACD model (Jasiak, 1999). The corresponding fractionally integrated process is obtained by introducing a fractional differencing operator:

$$\phi(L)(1 - L)^d Y_i = w + [1 - \beta(L)]v_i, \quad (21.29)$$

where the fractional differencing operator  $(1 - L)^d$  is defined by its expansion with respect to the lag operator:

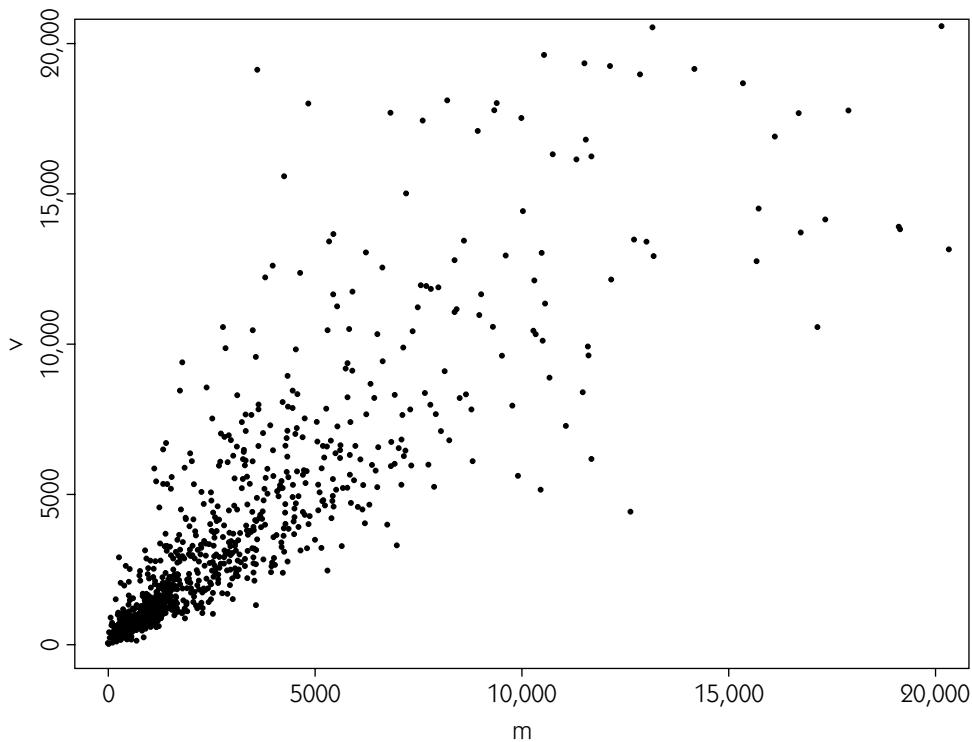
$$(1 - L)^d = \sum_{k=0}^{\infty} \Gamma(k - d)\Gamma(k + 1)^{-1}\Gamma(-d)^{-1}L^k = \sum_{k=0}^{\infty} \pi_k L^k, \quad \text{say}, \quad (21.30)$$

where  $\Gamma$  denotes the gamma function and  $0 < d < 1$ .

Although this extension successfully captures serial correlation at long lags, it fails to solve the major drawback of the basic ACD model, which consists of tying together the movements of conditional mean and conditional variance by supposing that  $E(Y_t | Y_{t-1}, Y_{t-2}, \dots) = \psi_t$  and  $V(Y_t | Y_{t-1}, Y_{t-2}, \dots) = k_0\psi_t^2$ , where the value  $k_0$  depends on the baseline distribution. We see that even though overdispersion arises whenever  $k_0 > 1$ , its magnitude is supposed to be path independent. This is a stringent assumption which in practice is often violated by the data. Empirical results based on intertrade durations on stock markets suggest on the contrary, the presence of path dependent (under)overdispersion as well as the existence of distinct dynamic patterns of the conditional mean and dispersion. As an illustration we display in Figure 21.5 a scatterplot of squared means and variances of intertrade durations of the Alcatel stock traded on the Paris bourse. We observe that the cluster is relatively dispersed with a significant number of observations featuring (conditional) underdispersion. Therefore, the data provide evidence supporting the co-existence of (conditional) under- and overdispersion, generating marginal overdispersion.

### 5.3 The SVD model

This model represents dynamics of both the conditional mean and variance in duration data. In this way it allows for the presence of both conditional under- and overdispersion in the data. Technically, it shares some similarities with the stochastic volatility models used in finance. The main difference in SVD



**Figure 21.5** (Under) Overdispersion of intertrade durations

specification compared to ACD is that it relies on two latent factor variables which are assumed to follow autoregressive stochastic processes. Note that despite the fact that the conditional variance in the ACD model is stochastic it is entirely determined by past durations. The introduction of additional random terms enhances the structure of the model and improves significantly the fit. On the other hand it makes the model more complicated and requires more advanced estimation techniques.

The approach is based on an extension of the exponential duration model with gamma heterogeneity. In this model the duration variable  $Y$  is exponentially distributed with the pdf :  $\lambda \exp(-\lambda y)$ , conditional on the hazard rate  $\lambda$ . Therefore, the duration variable may be written as:

$$Y = U/\lambda, \quad (21.31)$$

where  $U \sim \gamma(1, 1)$ . The hazard rate depends on some heterogeneity component  $V$ :

$$\lambda = aV, \quad (21.32)$$

where  $V \sim \gamma(b, b)$  is independent of  $U$ . The marginal distribution of this heterogeneity component is such that:  $EV = 1$  and  $\text{Var}(V) = 1/b$ , while the parameter  $a$  is a positive real number equal to the expected hazard rate.

Equations (21.31) and (21.32) yield the exponential model with gamma heterogeneity, namely:

$$Y = \frac{U}{aV}, \quad (21.33)$$

where  $U, V$  are independent,  $U \sim \gamma(1, 1)$  and  $V \sim \gamma(b, b)$ . This equation may be considered as a two factor model formula, where  $Y$  is a function of  $U$  and  $V$ . Some suitable nonlinear transformations can yield normally distributed factors. More explicitly, we get:

$$Y = \frac{G(1, \Phi(F_1))}{aG(b, \Phi(F_2))} = \frac{H(1, F_1)}{aH(b, F_2)}, \quad (21.34)$$

where  $F_1, F_2$  are iid standard normal variables,  $\Phi$  is the cdf of the standard normal distribution and  $G(b, \cdot)$  the quantile function of the  $\gamma(b, b)$  distribution. We have:  $H(1, F_1) = -\log[1 - \Phi(F_1)]$ . On the contrary, the function  $H(b, F_2)$  has no simple analytical expression in the general case, but admits a simple approximation in the neighborhood of the homogeneity hypothesis; namely if  $b \approx \infty : H(b, F_2) \approx 1 + (1/\sqrt{b})F_2 \approx \exp(F_2/\sqrt{b})$ , where the latter follows by the Central Limit Theorem.

The dynamics is introduced into the model through the two underlying Gaussian factors which follow a bivariate VAR process  $F_t = (F_{1t}, F_{2t})'$ , where the marginal distribution of  $F_t$  is constrained to be  $N(0, Id)$  to ensure that the marginal distribution of durations belongs to the class of exponential distributions with gamma heterogeneity.

This approach yields the class of Stochastic Volatility Duration (SVD) models (Ghysels, Gouriéroux, and Jasiak, 1997). They are defined by the following specification:

$$Y_t = \frac{1}{a} \frac{H(1, F_{1t})}{H(b, F_{2t})} = \frac{1}{a} \bar{H}(b, F_t), \quad (\text{say}) \quad (21.35)$$

where

$$F_t = \sum_{j=1}^p \Psi_j F_{t-j} + \varepsilon_t, \quad (21.36)$$

and  $\varepsilon_t$  is a Gaussian white noise random variable with variance–covariance matrix  $\Sigma(\Psi)$  such that  $\text{Var}(F_t) = Id$ .

## References

- Cox, D.R. (1975). Partial likelihood. *Biometrika* 62, 269–76.
- Cox, D.R., and D. Oakes (1984). Analysis of survival data. Chapman and Hall.
- Engle, R.F. (1982). Autoregressive conditional heteroskedasticity with estimates of the variance of United Kingdom inflation. *Econometrica* 50, 987–1007.

- Engle, R., and J.R. Russell (1998). The autoregressive conditional duration model. *Econometrica* 66, 1127–63.
- Ghysels, E., C. Gouriéroux, and J. Jasiak (1997). The stochastic volatility duration model. D.P. CREST.
- Gouriéroux, C. (1989). Econometrie des variables qualitatives. *Economica*.
- Gouriéroux, C., J. Jasiak, and G. Le Fol (1999). Intra-day market activity. *Journal of Financial Markets* 2:3, 193–226.
- Heckman, J. (1981). Heterogeneity and state dependence. In S. Rosen (ed.) *Studies in Labor Markets*. University of Chicago Press.
- Heckman, J., and B. Singer (1974). Econometric duration analysis. *Journal of Econometrics* 24, 63–132.
- Heckman, J., and B. Singer (1984). The identifiability of the proportional hazard model. *Review of Economic Studies* 231–45.
- Jasiak, J. (1999). Persistence in intertrade durations. *Finance* 19, 166–95.
- Kalbfleisch, J., and R. Prentice (1980). *The Statistical Analysis of Failure Time Data*. Wiley.
- Lancaster, T. (1985). Generalized residuals and heterogenous duration models, with application to the Weibull model. *Journal of Econometrics* 28, 155–69.
- Lancaster, T. (1990). *The Econometric Analysis of Transition Data*. Cambridge: Cambridge University Press.
- Lee, L. (1981). Maximum likelihood estimation and specification test for normal distributional assumption for accelerated failure time models. *Journal of Econometrics* 24, 159–79.

---

CHAPTER TWENTY-TWO

# Simulation Based Inference for Dynamic Multinomial Choice Models

*John Geweke, Daniel Houser,  
and Michael Keane*

## 1 INTRODUCTION

Over the last decade econometric inference based on simulation techniques has become increasingly common, particularly for latent variable models. The reason is that such models often generate econometric objective functions that embed high-order integrals, and which, consequently, can be most easily evaluated using simulation techniques.<sup>1</sup> There are several well known classical techniques for inference by simulation. Perhaps most common are the Method of Simulated Moments (McFadden, 1989) and Simulated Maximum Likelihood or SML (Lerman and Manski, 1981). In practice, both methods require that reasonably accurate simulators be used to evaluate the integrals that enter the objective function (see Geweke, Keane, and Runkle, 1994). Bayesian techniques are also becoming quite popular. These techniques typically entail Markov Chain Monte Carlo (MCMC) simulation to evaluate the integrals that define the posterior densities of a model's parameters (see Geweke and Keane (1999b) for an overview of MCMC methods).

Our goal in this chapter is to explain concretely how to implement simulation methods in a very general class of models that are extremely useful in applied work: dynamic discrete choice models where one has available a panel of

multinomial choice histories and partially observed payoffs. Many general surveys of simulation methods are now available (see Geweke, 1996; Monfort, Van Dijk, and Brown, 1995; and Gilks, Richardson, and Spiegelhalter, 1996), so in our view a detailed illustration of how to implement such methods in a specific case has greater marginal value than an additional broad survey. Moreover, the techniques we describe are directly applicable to a general class of models that includes static discrete choice models, the Heckman (1976) selection model, and all of the Heckman (1981) models (such as static and dynamic Bernoulli models, Markov models, and renewal processes). The particular procedure that we describe derives from a suggestion by Geweke and Keane (1999a), and has the advantages that it does not require the econometrician to solve the agents' dynamic optimization problem, or to make strong assumptions about the way individuals form expectations.

This chapter focuses on Bayesian inference for dynamic multinomial choice models via the MCMC method. Originally, we also hoped to discuss classical estimation of such models, so that readers could compare the two approaches. But, when we attempted to estimate the model developed below using SML it proved infeasible. The high dimension of the parameter vector caused an iterative search for the maximum of the simulated likelihood function via standard gradient based methods to fail rather dismally. In fact, unless the initial parameter values were set very close to the true values, the search algorithm would quickly stall. In contrast, the MCMC procedure was computationally feasible and robust to initial conditions. We concluded that Bayesian inference via MCMC has an important advantage over SML for high dimensional problems because it does not require a search for the optimum of the likelihood.

We consider dynamic, stochastic, parametric models with intertemporally additively separable preferences and a finite time horizon. Suppose that in each period  $t = 1, \dots, T$  ( $T < \infty$ ) each agent chooses among a finite set  $A_t$  of mutually exclusive alternatives. Let  $\mathcal{R}^{k_t}$  be the date- $t$  state space, where  $k_t$  is a positive integer. Choosing alternative  $a_t \in A_t$  in state  $I_t \in \mathcal{R}^{k_t}$  leads to period payoff  $R(I_t, a_t; \theta)$ , where  $\theta$  is a finite-vector denoting the model's structural parameters.

The value to choosing alternative  $a_t$  in state  $I_t$ , denoted by  $V_t(I_t, a_t)$ , depends on the period payoff and on the way agents expect that choice to affect future payoffs. For instance, in the familiar case when agents have rational expectations, alternative specific values can be expressed:

$$V_t(I_t, a_t) = R(I_t, a_t; \theta) + \delta E_t \max_{a_{t+1} \in A_{t+1}} V_{t+1}(I_{t+1}, a_{t+1} | I_t, a_t) \quad (t = 1, \dots, T) \quad (22.1)$$

$$V_{T+1}(\cdot) \equiv 0 \quad (22.2)$$

$$I_{t+1} = H(I_t, a_t; \theta) \quad (22.3)$$

where  $\delta$  is the constant rate of time preference,  $H(I_t, a_t; \theta)$  is a stochastic law of motion that provides an intertemporal link between choices and states, and  $E_t$  is the date- $t$  mathematical expectations operator so that expectations are taken with respect to the true distribution of the state variables  $P_H(I_{t+1} | I_t, a_t; \theta)$  as generated

by  $H(\cdot)$ . Individuals choose alternative  $a_t^*$  if and only if  $V_t(I_t, a_t^*) > V_t(I_t, a_t) \forall a_t \in A_t, a_t \neq a_t^*$ . See Eckstein and Wolpin (1989) for a description of many alternative structural models that fit into this framework.

The econometrician is interested in drawing inferences about  $\theta$ , the vector of structural parameters. One econometric procedure to accomplish this (see Rust, 1987 or Wolpin, 1984) requires using dynamic programming to solve system (22.1)–(22.3) at many trial parameter vectors. At each parameter vector, the solution to the system is used as input to evaluate a prespecified econometric objective function. The parameter space is systematically searched until a vector that “optimizes” the objective function is found. A potential drawback of this procedure is that, in general, solving system (22.1)–(22.3) with dynamic programming is extremely computationally burdensome. The reason is that the mathematical expectations that appear on the right-hand side of (22.1) are often impossible to compute analytically, and very time consuming to approximate well numerically. Hence, as a practical matter, this estimation procedure is useful only under very special circumstances (for instance, when there are a small number of state variables). Consequently, a literature has arisen that suggests alternative approaches to inference in dynamic multinomial choice models.

Some recently developed techniques for estimation of the system (22.1)–(22.3) focus on circumventing the need for dynamic programming. Several good surveys of this literature already exist, and we will not attempt one here (see Rust, 1994). Instead, we simply note that the idea underlying the more well known of these approaches, i.e., Hotz and Miller (1993) and Manski (1993), is to use choice and payoff data to draw inferences about the values of the expectations on the right-hand side of (22.1). A key limitation of these procedures is that, in order to learn about expectations, each requires the data to satisfy a strict form of stationarity in order to rule out cohort effects.

The technique proposed by Geweke and Keane (1999a) for structural inference in dynamic multinomial choice models also circumvents the need for dynamic programming. A unique advantage of their method is that it does not require the econometrician to make strong assumptions about the way people form expectations. Moreover, their procedure is not hampered by strong data requirements. It can be implemented when the data include only partially observed payoffs from a single cohort of agents observed over only part of their lifecycle.

To develop the Geweke–Keane approach, it is useful to express the value function (22.1) as:

$$V_t(I_t, a_t) = R(I_t, a_t; \theta) + F^H(I_t, a_t), \quad (22.4)$$

where  $F^H(I_t, a_t) \equiv \delta E_t \max_{a_{t+1} \in A_{t+1}} V_{t+1}(a_{t+1}, H(I_t, a_t))$ . Geweke and Keane (1999a) observed that the definition of  $F^H(\cdot)$ , henceforth referred to as the “future component”, makes sense independent of the meaning of  $E_t$ . If, as assumed above,  $E_t$  is the mathematical expectations operator then  $F^H(\cdot)$  is the rational expectations future component. On the other hand, if  $E_t$  is the zero operator, then future payoffs do not enter the individuals’ decision rules, and  $F^H(\cdot)$  is identically zero. In general, the functional form of the future component  $F^H(\cdot)$  will vary with the

way people form expectations. Unfortunately, in most circumstances the way people form expectations is unknown. Accordingly, the correct specification of the future component  $F^H(\cdot)$  is also unknown.

There are, therefore, two important reasons why an econometrician may prefer not to impose strong assumptions about the way people form expectations, or, equivalently, on the admissible forms of the future component. First, such assumptions may lead to an intractable econometric model. Second, the econometrician may see some advantage to taking a less dogmatic stance with respect to behaviors about which very little, if any, *a priori* information is available.

When the econometrician is either unwilling or unable to make strong assumptions about the way people form expectations, Geweke and Keane (1999a) suggest that the future component  $F^H(\cdot)$  be represented by a parameterized flexible functional form such as a high-order polynomial. The resulting value function can be written

$$V_t(I_t, a_t) = R(I_t, a_t; \theta) + F^H(I_t, a_t | \pi) \quad (22.5)$$

where  $\pi$  is a vector of polynomial coefficients that characterize expectation formation. Given functional forms for the contemporaneous payoff functions, and under the condition that  $\theta$  and  $\pi$  are jointly identified, it is possible to draw inferences both about the parameters of the payoff functions and the structure of expectations.

This chapter focuses on an important case in which key structural and expectations parameters are jointly identified. We consider a model where an alternative's payoff is partially observed if and only if that alternative is chosen. In this case, after substitution of a flexible polynomial function for the future component as in (22.5), the model takes on a form similar to a static Roy (1951) model augmented to include influences on choice other than the current payoffs, as in Heckman and Sedlacek (1986). The key difference is that  $F^H(\cdot)$  incorporates overidentifying restrictions on the non-payoff component of the value function that are implied by (22.1)–(22.3) and that are not typically invoked in the estimation of static selection models. Specifically, the parameters of the non-payoff component of the value function are constant across alternatives, and the arguments of the non-payoff component vary in a systematic way across alternatives that is determined by the law of motion  $H(\cdot)$  for the state variables.

The structural model (22.1)–(22.3) also implies restrictions on the nature of the future component's arguments. For instance, if  $H(\cdot)$  and  $R(\cdot)$  jointly imply that the model's payoffs are path-independent, then the future component should be specified so that path-dependent expectation formation is ruled out.<sup>2</sup> Similarly, contemporaneous realizations of serially independent stochastic variables contain no information relevant for forecasting future outcomes, so they should not enter the arguments of the flexible functional form. Without such coherency conditions one might obtain results inconsistent with the logic of the model's specification.

A finite order polynomial will in general provide only an approximation to the true future component. Hence, it is important to investigate the extent to which

misspecification of the future component may affect inference for the model's structural parameters. Below we report the outcome of some Monte Carlo experiments that shed light on this issue. The experiments are conducted under both correctly and incorrectly specified future components. We find that the Geweke–Keane approach performs extremely well when  $F^H(\cdot)$  is correctly specified, and still very well under a misspecified future component. In particular, we find that assuming the future component is a polynomial when it is actually generated by rational expectations leads to only "second order" difficulties in two senses. First, it has a small effect on inferences with regard to the structural parameters of the payoff functions.<sup>3</sup> Second, the decision rules inferred from the data in the misspecified model are very close to the optimal rule in the sense that agents using the suboptimal rule incur "small" lifetime payoff losses.

The remainder of this chapter is organized as follows. Section 2 describes the application, and Section 3 details the Gibbs sampling algorithm. Section 4 reviews our experimental design and results, and Section 5 concludes.

## 2 THE DYNAMIC MULTINOMIAL CHOICE MODEL

In this section we present an example of Bayesian inference for dynamic discrete choice models using the Geweke–Keane method of replacing the future component of the value function with a flexible polynomial function. The discussion is based on a model that is very similar to ones analyzed by Keane and Wolpin (1994, 1997).

In the model we consider,  $i = 1, \dots, N$  agents choose among  $j = 1, \dots, 4$  mutually exclusive alternatives in each of  $t = 1, \dots, 40$  periods. One can think of the first two alternatives as work in one of two occupations, the third as attending school and the fourth alternative as remaining at home. One component of the current period payoff in each of the two occupational alternatives is the associated wage,  $w_{ijt}$  ( $j = 1, 2$ ). The log-wage equation is:

$$\begin{aligned} \ln w_{ijt} &= \beta_{0j} + \beta_{1j}X_{i1t} + \beta_{2j}X_{i2t} + \beta_{3j}S_{it} + \beta_{4j}X_{ijt}^2 + \varepsilon_{ijt} \quad (j = 1, 2) \\ &= \mathbf{Y}'_{ijt}\boldsymbol{\beta}_j + \varepsilon_{ijt} \quad (j = 1, 2), \end{aligned} \quad (22.6)$$

where  $\mathbf{Y}_{ijt}$  is the obvious vector,  $\boldsymbol{\beta}_j = (\beta_{0j}, \dots, \beta_{4j})'$ ,  $S_{it}$  is the periods of school completed,  $(X_{ijt})_{j=1,2}$  is the periods of experience in each occupation  $j$ , and the  $\varepsilon_{ijt}$  are serially independent productivity shocks, with  $(\varepsilon_{i1t}, \varepsilon_{i2t})' \sim N(0, \Sigma_\varepsilon)$ . Each occupational alternative also has a stochastic nonpecuniary payoff,  $v_{ijt}$ , so the complete current period payoffs are

$$u_{ijt} = w_{ijt} + v_{ijt} \quad (j = 1, 2). \quad (22.7)$$

The schooling payoffs include tuition costs. Agents begin with a tenth-grade education, and may complete two additional grades without cost. We assume there is a fixed undergraduate tuition rate  $\alpha_1$  for attending grades 13 through 16, and a fixed graduate tuition rate  $\alpha_2$  for each year of schooling beyond 16. We

assume a “return to school” cost  $\alpha_3$  that agents face if they did not choose school the previous period. Finally, school has a nonstochastic, nonpecuniary benefit  $\alpha_0$  and a mean zero stochastic nonpecuniary payoff  $v_{i3t}$ . Thus we have

$$u_{i3t} = \alpha_0 + \alpha_1 \chi(12 \leq S_{it} \leq 15) + \alpha_2 \chi(S_{it} \geq 16) + \alpha_3 \chi(d_{i,t-1} \neq 3) + v_{i3t} \equiv \Lambda_{it} \alpha + v_{i3t}, \quad (22.8)$$

where  $\chi$  is an indicator function that takes value one if the stated condition is true and is zero otherwise,  $\Lambda_{it}$  is a vector of zeros and ones corresponding to the values of the indicator functions,  $\alpha = (\alpha_0, \dots, \alpha_3)'$ ,  $d_{it} \in \{1, 2, 3, 4\}$  denotes the choice of  $i$  at  $t$ . Lastly, we assume that option four, home, has both a nonstochastic nonpecuniary payoff  $\phi$  and a stochastic nonpecuniary payoff  $v_{ijt}$ , so

$$u_{i4t} = \phi + v_{i4t}. \quad (22.9)$$

We will set  $u_{ijt} = \bar{u}_{ijt} + v_{ijt}$ ,  $(j = 1, \dots, 4)$ . The nonpecuniary payoffs  $(v_{ijt})_{j=1,4}$  are assumed serially independent.

The state of the agent at the time of each decision is

$$I_{it} = \{(X_{ijt})_{j=1,2}, S_{it}, t, d_{i,t-1}, (\varepsilon_{ijt})_{j=1,2}, (v_{ijt})_{j=1, \dots, 4}\}. \quad (22.10)$$

We assume  $d_{i0} = 3$ . The laws of motion for experience in the occupational alternatives and school are:  $X_{ij,t+1} = X_{ijt} + \chi(d_{it} = j)$ ,  $j = 1, 2$ ,  $S_{i,t+1} = S_{it} + \chi(d_{it} = 3)$ . The number of “home” choices is excluded from the state-space as it is linearly dependent on the level of education, the period, and experience in the two occupations.

It is convenient to have notation for the elements of the state vector whose value in period  $t + 1$  depends nontrivially on their value in period  $t$  or on the current choice. The reason, as we note below, is that these elements are the natural arguments of the future component. We define

$$I_{it}^* = \{(X_{ijt})_{j=1,2}, S_{it}, t, d_{i,t-1}\}.$$

The value of each alternative is the sum of its current period payoff, the stochastic nonpecuniary payoff and the future component:

$$\begin{aligned} V_{ijt}(I_{it}) &= \bar{u}_{ijt}(I_{it}) + v_{ijt} + F(X_{i1t} + \chi(j = 1), X_{i2t} + \chi(j = 2), \\ &\quad S_{it} + \chi(j = 3), t + 1, \chi(j = 3)) \quad (j = 1, \dots, 4) \quad (t = 1, \dots, 40) \\ &\equiv \bar{u}_{ijt}(I_{it}) + v_{ijt} + F(I_{it}^*, j) \end{aligned} \quad (22.11)$$

The function  $F$  represents agents’ forecasts about the effects of their current state and choice on their future payoff stream. The function is fixed across alternatives, implying that forecasts vary across alternatives only because different choices lead to different future states, and it depends only on the choice and the state variables in  $I_{t+1}^*$ .<sup>4</sup>

Since choices depend only on relative alternative values, rather than their levels, we define for  $j \in \{1, 2, 3\}$ :

$$\begin{aligned} Z_{ijt} &\equiv V_{ijt} - V_{i4t} \\ &= \bar{u}_{ijt} + v_{ijt} + F(I_{it}^*, j) - \bar{u}_{i4t} - v_{i4t} - F(I_{it}^*, 4) \\ &= \bar{u}_{ijt} + f(I_{it}^*, j) + \eta_{ijt}, \end{aligned} \quad (22.12)$$

where  $\bar{u}_{ijt} \equiv \bar{u}_{ijt} - \bar{u}_{i4t}$ ,  $\{\eta_{ijt}\}_{j=1,2,3} \equiv (v_{ijt} - v_{i4t})_{j=1,2,3} \sim N(0, \Sigma_\eta)$  and  $f(I_{it}^*, j) = F(I_{it}^*, j) - F(I_{it}^*, 4)$ . Importantly, after differencing, the value  $\phi$  of the home payoff is subsumed in  $f$  the relative future component. Clearly, if an alternative's future component has an intercept (as each of ours does) then it and the period return to home cannot be separately identified.

The value function differences  $Z_{it}$  are latent variables unobserved by the econometrician. The econometrician only observes the agents' choices  $\{d_{it}\}$  for  $t = 1, \dots, 40$ , and, in the periods when the agent works, the wage for the chosen alternative. Thus, payoffs are never completely observed, both because wages are censored and because the nonpecuniary components of the payoffs ( $v_{ijt}$ ) are never observed. Nevertheless, given observed choices and partially observed wages, along with the functional form assumptions about the payoff functions, it is possible to learn both about the future component  $F(\cdot)$  and the structural parameters of the payoff functions without making strong assumptions about how agents form expectations. Rather, we simply assume that the future component lies along a fourth-order polynomial in the state variables. After differencing to obtain  $\{f(I_{it}^*, j)\}_{j=1,2,3}$ , the polynomial contained 53 terms of order three and lower (see Appendix A). We express the future component as

$$f(I_{it}^*, j) = \psi'_{ijt} \pi \quad (j = 1, 2, 3), \quad (22.13)$$

where  $\psi_{ijt}$  is a vector of functions of state-variables that appear in the equation for  $f(I_{it}^*, j)$  and  $\pi$  is a vector of coefficients common to each choice. Cross-equation restrictions of this type are a consequence of using the same future component function  $F$  for each alternative and reflect the consistency restrictions discussed earlier.

### 3 IMPLEMENTING THE GIBBS SAMPLING ALGORITHM

Bayesian analysis of this model entails deriving the joint posterior distribution of the model's parameters and unobserved variables. Recall that the value function differences  $Z = \{(Z_{ijt})_{j=1,2,3; i=1,N; t=1,40}\}$  are never observed, and that wages  $W = \{(w_{ijt})_{j=1,2,3; i=1,N; t=1,40}\}$  are only partially observed. Let  $W_1$  and  $W_2$  denote the set of observed and unobserved wages, respectively, and let  $Y = \{Y_{ijt}\}_{i=1,N; j=1,2; t=1,40}$  denote the log-wage equation regressors. Then the joint posterior density is  $p(W_2, Z, \beta_1, \beta_2, \alpha, \pi, \Sigma_\epsilon^{-1}, \Sigma_\eta^{-1} | W_1, Y, \Lambda)$ . By Bayes' law, this density is proportional to

$$p(W, Z | Y, \Lambda, \psi, \beta_1, \beta_2, \alpha, \pi, \Sigma_\epsilon^{-1}, \Sigma_\eta^{-1}) \cdot p(\beta_1, \beta_2, \alpha, \pi, \Sigma_\epsilon^{-1}, \Sigma_\eta^{-1}). \quad (22.14)$$

The first term in (22.14) is the so-called “complete data” likelihood function. It is the likelihood function that could be formed in the hypothetical case that we had data on  $N$  individuals observed over 40 periods each, and we observed all of the value function differences  $Z$  and the complete set of wages  $W$  for all alternatives. This is:

$$\begin{aligned}
 p(W, Z | Y, \Lambda, \psi, \beta_1, \beta_2, \alpha, \pi, \Sigma_{\epsilon}^{-1}, \Sigma_{\eta}^{-1}) &\propto \\
 \prod_{i,t} |\Sigma_{\epsilon}^{-1}|^{1/2} (w_{i1t} w_{i2t})^{-1} &\exp \left\{ -\frac{1}{2} \left( \ln w_{i1t} - Y'_{i1t} \beta_1 \right) \Sigma_{\epsilon}^{-1} \left( \ln w_{i1t} - Y'_{i1t} \beta_1 \right) \right\} \\
 \cdot |\Sigma_{\eta}^{-1}|^{1/2} &\exp \left\{ -\frac{1}{2} \left( \begin{array}{l} Z_{i1t} - w_{i1t} - \Psi'_{i1t} \pi \\ Z_{i2t} - w_{i2t} - \Psi'_{i2t} \pi \\ Z_{i3t} - \Lambda'_{it} \alpha - \Psi'_{i3t} \pi \end{array} \right) \Sigma_{\eta}^{-1} \left( \begin{array}{l} Z_{i1t} - w_{i1t} - \Psi'_{i1t} \pi \\ Z_{i2t} - w_{i2t} - \Psi'_{i2t} \pi \\ Z_{i3t} - \Lambda'_{it} \alpha - \Psi'_{i3t} \pi \end{array} \right) \right\} \\
 \cdot \chi(Z_{ijt} > 0, Z_{ikt} (k \neq j) < 0 \text{ if } d_{it} = j \text{ and } j \in \{1, 2, 3\}, \{Z_{ijt}\}_{j=1,2,3} < 0 \text{ otherwise}) & \\
 (22.15)
 \end{aligned}$$

The second term in (22.15) is the joint prior distribution. We assume flat priors on all parameters except the two precision matrices, for which we assume the standard noninformative priors (see Zellner, 1971, section 8.1):

$$p(\Sigma_{\epsilon}^{-1}) \propto |\Sigma_{\epsilon}^{-1}|^{-3/2}, p(\Sigma_{\eta}^{-1}) \propto |\Sigma_{\eta}^{-1}|^{-2} \quad (22.16)$$

The Gibbs sampler draws from a density that is proportional to the product of (22.15) and the two densities in (22.16).

The Gibbs sampling algorithm is used to form numerical approximations of the parameters’ marginal posterior distributions. It is not feasible to construct these marginal posteriors analytically, since doing so requires high dimensional integrations over unobserved wages and value function differences. Implementing the Gibbs sampling algorithm requires us to factor the joint posterior defined by (22.14)–(22.16) into a set of conditional posterior densities, in such a way that each can be drawn from easily. Then, we cycle through these conditionals, drawing a block of parameters from each in turn. As the number of cycles grows large, the parameter draws so obtained converge in distribution to their respective marginal posteriors, given certain mild regularity conditions (see Tierney (1994) for a discussion of these conditions). An important condition is that the posterior distribution be finitely integrable, which we verify for this model in Appendix B. Given the posterior distribution of the parameters, conditional on the data, the investigator can draw exact finite sample inferences.

Our Gibbs sampling-data augmentation algorithm consists of six steps or “blocks.” These steps, which we now briefly describe, are cycled through repeatedly until convergence is achieved.

- Step 1. Draw value function differences  $\{Z_{ijt}, i = 1, N; j = 1, 3; t = 1, 40\}$
- Step 2. Draw unobserved wages  $\{w_{ijt} \text{ when } d_{it} \neq j, (j = 1, 2)\}$

- Step 3. Draw the log-wage equation coefficients  $\beta_j$ .
- Step 4. Draw the log-wage equation error-covariance matrix  $\Sigma_\epsilon$ .
- Step 5. Draw the parameters of the future component  $\pi$  and school payoff parameters  $\alpha$ .
- Step 6. Draw the nonpecuniary payoff covariance matrix  $\Sigma_\eta$ .

## STEP 1

We chose to draw the  $\{Z_{ijt}, i = 1, N; j = 1, 3; t = 1, 40\}$  one by one. Taking everything else in the model as given, it is evident from (22.14)–(22.16) that the conditional distribution of a single  $Z_{ijt}$  is truncated Gaussian. Dealing with the truncation is straightforward. There are three ways in which the Gaussian distribution might be truncated.

*Case 1:*  $Z_{ijt}$  is the value function difference for the chosen alternative. Thus, we draw  $Z_{ijt} > \max\{0, (Z_{ikt})_{k \in \{1, 2, 3\} \setminus j}\}$ .

*Case 2:*  $Z_{ijt}$  is not associated with the chosen alternative, and “home” was not chosen. Thus, we draw  $Z_{ijt} < Z_{id_{it}}$ .

*Case 3:* “Home” was chosen. In this case, we draw  $Z_{ijt} < 0$ .

We draw from the appropriate univariate, truncated Gaussian distributions using standard inverse CDF methods.

## STEP 2

We chose to draw the unobserved wages  $\{w_{ijt} \text{ when } d_{it} \neq j, (j = 1, 2, 3)\}$  one by one. Suppose  $w_{11t}$  is unobserved. Its density, conditional on every other wage, future component difference and parameter being known, is from (22.14), (22.15) and (22.16) evidently given by:

$$g(w_{11t} | \cdot) \propto \frac{1}{\tilde{w}_{11t}} \exp \left\{ -\frac{1}{2} \left( \ln w_{11t} - \mathbf{Y}'_{11t} \boldsymbol{\beta}_1 \right) \Sigma_\epsilon^{-1} \left( \ln w_{11t} - \mathbf{Y}'_{11t} \boldsymbol{\beta}_1 \right) \right\} \\ \exp \left\{ -\frac{1}{2} \begin{pmatrix} Z_{11t} - w_{11t} - \Psi'_{11t} \boldsymbol{\pi} \\ Z_{12t} - w_{12t} - \Psi'_{12t} \boldsymbol{\pi} \\ Z_{13t} - \Lambda'_{1t} \boldsymbol{\alpha} - \Psi'_{13t} \boldsymbol{\pi} \end{pmatrix}' \Sigma_\eta^{-1} \begin{pmatrix} Z_{11t} - w_{11t} - \Psi'_{11t} \boldsymbol{\pi} \\ Z_{12t} - w_{12t} - \Psi'_{12t} \boldsymbol{\pi} \\ Z_{13t} - \Lambda'_{1t} \boldsymbol{\alpha} - \Psi'_{13t} \boldsymbol{\pi} \end{pmatrix} \right\}. \quad (22.17)$$

This distribution is nonstandard as wages enter in both logs and levels. Nevertheless, it is straightforward to sample from this distribution using rejection methods (see Geweke (1995) for a discussion of efficient rejection sampling). In brief, we first drew a candidate wage  $w^c$  from the distribution implied by the first exponential of (22.17), so that  $\ln w^c \sim N(\mathbf{Y}'_{11t} \boldsymbol{\beta}_1 + \lambda_{1t}, \sigma_*^2)$ , where  $\lambda_{it} \equiv \Sigma_\epsilon(1, 2) \epsilon_{i2t} / \Sigma(2, 2)$  and  $\sigma_*^2 \equiv \Sigma_\epsilon(1, 1)(1 - (\Sigma_\epsilon(1, 2))^2) / (\Sigma_\epsilon(1, 1)\Sigma_\epsilon(2, 2))$ . This draw is easily accomplished, and  $w^c$  is found by exponentiating. The probability with which this draw is accepted is found by dividing the second exponential in

(22.17) by its conditional maximum over  $w_{it}$  and evaluating the resulting expression at  $w_{it} = w^c$ . If the draw is accepted then the unobserved  $w_{it}$  is set to  $w^c$ . Otherwise, the process is repeated until a draw is accepted.

### STEP 3

Given all wages, value function differences, and other parameters, the density of  $(\beta_1, \beta_2)$  is:

$$g(\beta_1, \beta_2) \propto \exp \left\{ -\frac{1}{2} \left( \ln w_{it} - Y'_{it} \beta_1 \right)' \Sigma_{\epsilon}^{-1} \left( \ln w_{it} - Y'_{it} \beta_1 \right) \right\} \left( \ln w_{it} - Y'_{it} \beta_2 \right)' \Sigma_{\epsilon}^{-1} \left( \ln w_{it} - Y'_{it} \beta_2 \right). \quad (22.18)$$

So that  $(\beta_1, \beta_2)$  is distributed according to a multivariate normal. In particular, it is easy to show that

$$\beta \sim N[(Y'\Sigma^{-1}Y)^{-1}Y'\Sigma^{-1} \ln W, (Y'\Sigma^{-1}Y)^{-1}],$$

where  $\beta \equiv (\beta'_1, \beta'_2)', \Sigma = \Sigma_{\epsilon} \otimes I_{NT}, Y = \begin{bmatrix} Y_1 & 0 \\ 0 & Y_2 \end{bmatrix}$  and  $\ln W = [\ln W'_1, \ln W'_2]'$ , where  $Y_1$

is the regressor matrix for the first log-wage equation naturally ordered through all individuals and periods, and similarly for  $Y_2, W_1$  and  $W_2$ . It is straightforward to draw  $\beta$  from this multivariate normal density.

### STEP 4

With everything else known  $\Sigma_{\epsilon}^{-1}$  has a Wishart distribution. Specifically, it is immediate from the joint posterior that  $p(\Sigma_{\epsilon}^{-1}) \propto |\Sigma_{\epsilon}^{-1}|^{\frac{NT-3}{2}} \exp\{-\frac{1}{2} \text{tr}(S(\beta)\Sigma_{\epsilon}^{-1})\}$ , so that

$$\Sigma_{\epsilon}^{-1} \sim W(S(\beta), NT),$$

$$\text{where } S(\beta) = (\ln W_1 - Y_1 \beta_1, \ln W_2 - Y_2 \beta_2)' (\ln W_1 - Y_1 \beta_1, \ln W_2 - Y_2 \beta_2). \quad (22.19)$$

It is easy to draw from the Wishart and then invert the  $2 \times 2$  matrix to obtain  $\Sigma_{\epsilon}$ .

### STEP 5

It is convenient to draw both the future component  $\pi$  parameters and the parameters  $\alpha$  of the school payoff jointly. Since the future component for school contains an intercept, it and the constant in  $\Lambda$  cannot be separately identified. Hence, we omit  $\alpha_0$  as well as the first row from each  $\Lambda_{it}$ . Define the vector  $\pi^* \equiv [\pi', \alpha']'$ , where  $\alpha = (\alpha_1, \alpha_2, \alpha_3)'$  and define  $\Psi_{ijt}^* \equiv [\Psi'_{ijt}, 0_3']'$  ( $j = 1, 2$ ), and  $\Psi_{i3t}^* = [\Psi'_{i3t}, \Lambda'_{it}]'$ . Note that  $\pi^*$  and the  $\Psi_{ijt}^*$  are 56-vectors. Then define  $\Psi_k = [\Psi_{1k}^* \Psi_{1k}^* \dots \Psi_{NkT-1}^* \Psi_{NkT}^*]$ , and set  $\Psi = [\Psi_1 \Psi_2 \Psi_3]'$ , so that  $\Psi$  is a  $(3 \cdot NT \times 56)$  stacked-regressor matrix. Similarly, define the corresponding  $3 \cdot NT$ -vector  $\Gamma$  by  $\Gamma = (\{Z_{it} - w_{it}\}'_{it}, \{Z_{i2t} - w_{i2t}\}'_{it}, \{Z_{i3t} - w_{i3t}\}'_{it})'$ . It is immediate from (22.15), in which  $\pi^*$  enters only through

the second exponential expressions that, conditional on everything else known,  $\pi^*$  has a multivariate normal density given by:

$$\pi^* \sim N[(\Psi'\Omega^{-1}\Psi)^{-1}\Psi'\Omega^{-1}\Gamma, (\Psi'\Omega^{-1}\Psi)^{-1}] \quad (22.20)$$

where  $\Omega = \Sigma_\eta \otimes I_{NT}$ . We draw from this using a standard, multivariate normal random number generator.

### STEP 6

With everything else known the distribution of  $\Sigma_\eta^{-1}$  is Wishart;  $\Sigma_\eta^{-1} \sim W(SST_\eta, NT)$ , where  $SST_\eta = \sum_{i,t} (\eta_{ilt} \eta_{i2t} \eta_{i3t})'(\eta_{ilt} \eta_{i2t} \eta_{i3t})$ , and with the  $\eta_{ijt}$  defined by (22.12). It is easy to draw from this distribution and then invert the  $3 \times 3$  matrix to obtain  $\Sigma_\eta$ .

## 4 EXPERIMENTAL DESIGN AND RESULTS

This section details the design and results of a Monte Carlo experiment that we conducted to shed light on the performance of the Gibbs sampling algorithm discussed in Section 2. We generated data according to equations (22.6)–(22.12) using the true parameter values that are listed in column two of Table 22.3. The table does not list the discount rate and the intercepts in the school and home payoff functions, which were set to 0.95, 11,000, and 17,000 respectively, since these are not identified. In all of our experiments we set the number of people,  $N$ , to 2,000.

Data from this model were generated using two different assumptions about the way people formed expectations. First, we assumed that people had rational expectations. This required us to solve the resulting dynamic optimization problem once to generate the optimal decision rules. Since the choice set includes only four discrete alternatives it is feasible to do this. Then, to simulate choice and wage paths requires only that we generate realizations of the appropriate stochastic variables. It is important to note that the polynomial future component used in the estimation procedure does not provide a perfect fit to the rational expectations future component. Hence, analysis of this data sheds light on the effect that misspecification of the future component may have on inference.

Next, we assumed that agents used a future component that was actually a polynomial in the state variables to form decisions. Analysis of this data set sheds light on how the algorithm performs when the model is correctly specified. To ensure close comparability with the rational expectations case, we constructed this polynomial by regressing the rational expectations future components on a fourth-order polynomial in the state variables, constructed as described in the discussion preceding (22.13). We used the point estimates from this regression as the coefficients of our polynomial future component (see Appendix A for the specific form of the polynomial).

We found that a fourth-order polynomial provided a good approximation to the future component in the sense that if agents used the approximate instead of optimal decision rule they suffered rather small lifetime earnings losses. Evidence of this is given in Table 22.1, where we report the results of simulations

**Table 22.1** Quality of the polynomial approximation to the true future component

Error set	1	2	3	4	5
Mean present value of payoffs with true future component*	356,796	356,327	355,797	355,803	355,661
Mean present value of payoffs with polynomial approximation*	356,306	355,978	355,337	355,515	355,263
Mean dollar equivalent loss*	491	349	460	287	398
Mean percent loss*	0.14%	0.10%	0.13%	0.08%	0.11%
Percent choice agreement					
Aggregate	91.81%	91.71%	91.66%	92.30%	91.80%
By period					
1	95.80%	95.55%	96.30%	96.10%	96.15%
2	95.35%	94.90%	95.85%	95.95%	95.30%
3	91.30%	90.45%	90.25%	91.15%	89.90%
4	88.00%	87.75%	89.00%	88.90%	88.45%
5	87.00%	88.30%	87.00%	89.20%	87.60%
10	92.70%	92.60%	92.70%	92.30%	92.30%
20	92.20%	93.00%	92.70%	93.05%	93.10%
30	91.55%	90.90%	90.55%	91.85%	90.85%
40	92.80%	92.15%	91.70%	92.75%	92.10%

\* The mean present value of payoffs is the equally-weighted average discounted sum of ex-post lifetime payoffs over 2,000, 40 period lived agents. The values are dollar equivalents.

under optimal and suboptimal decision rules. The simulations were conducted as follows. First, for  $N = 2,000$  people we drew five sets of lifetime ( $T = 40$ ) realizations of the model's stochastic components  $\{\varepsilon_{i1}, \varepsilon_{i2}, (\eta_{ij})_{j=1,4}\}$ . In Table 22.1 these are referred to as error sets one to five. For each of the five error sets we simulated lifetime choice histories for each of the 2,000 people under the optimal and approximate decision rules. We refer to the 10 data sets constructed in this way as 1-EMAX through 5-EMAX and 1-POLY through 5-POLY, respectively. We then calculated the mean of the present value of lifetime payoffs (pecuniary plus nonpecuniary) for each of the 2,000 people under the optimal and approximate decision rules, respectively, for each of the five error sets. These are reported in the second and third rows of Table 22.1. Holding the error set fixed, the source of any difference in the mean present value of lifetime payoffs lies in the use of different decision rules. The mean present values of dollar equivalent

losses from using the suboptimal polynomial rules are small, ranging from 287 to 491. The percentage loss ranges from 8 hundredths of 1 percent to 14 hundredths of 1 percent. These findings are similar to those reported by Geweke and Keane (1999a) and Krusell and Smith (1995).

Table 22.2 reports the mean accepted wages and choice frequencies for the data generated from error-set two. The first set of columns report statistics for data generated according to the polynomial approximation (data set 2-POLY) while the second set of columns report results from the optimal decision rule (data set 2-EMAX). Under our parameterization, occupation one can be thought of as "unskilled" labor, while occupation two can be understood as "skilled" labor. The reason is the mean of the wage offer distribution is lower in occupation two early in life, but it rises more quickly with experience. The choice patterns and mean accepted wages are similar under the two decision rules. School is chosen somewhat more often under the optimal decision rule, which helps to generate slightly higher lifetime earnings. Finally, note that selection effects leave the mean accepted wage in occupation two higher than that in occupation one throughout the lifecycle under both decision rules.

Next, we ran the Gibbs algorithm described in section two for 40,000 cycles on each data set. We achieved about three cycles per minute on a Sun ultra-2 workstation.<sup>5</sup> Thus, while time requirements were substantial, they were minor compared to what estimation of such a model using a full solution of the dynamic programming problem would entail. Visual inspection of graphs of the draw sequences, as well as application of the split sequence diagnostic suggested by Gelman (1996) – which compares variability of the draws across subsequences – suggests that the algorithm converged for all 10 artificial data sets. In all cases, the final 15,000 draws from each run were used to simulate the parameters' marginal posterior distributions.

Table 22.3 reports the results of the Gibbs sampling algorithm when applied to the data generated with a polynomial future component. In this case, the econometric model is correctly specified. The first column of Table 22.3 is the parameter label, the second column is the true value, and the remaining columns report the structural parameters' posterior means and standard deviations for each of the five data sets.<sup>6</sup> The results are extremely encouraging. Across all runs, there was only one instance in which the posterior mean of a parameter for the first wage equation was more than two posterior standard deviations away from its true value: the intercept in data set one. In data sets four and five, all of the structural parameters' posterior means are within two posterior standard deviations of their true values. In the second data set, only the second wage equation's own experience term is slightly more than two posterior standard deviations from its true value. In the third data set the mean of the wage equation's error correlation is slightly more than two posterior standard deviations from the true value, as are a few of the second wage equation's parameters.

Careful examination of Table 22.3 reveals that the standard deviation of the nonpecuniary payoff was the most difficult parameter to pin down. In particular, the first two moments of the marginal posteriors of these parameters vary considerably across experiments, in relation to the variability of the other structural

parameters' marginal posteriors. This result reflects earlier findings reported by Geweke and Keane (1999a). In the earlier work they found that relatively large changes in the value of the nonpecuniary component's standard deviation had only a small effect on choices. It appears that this is the case in the current experiment as well.

It is interesting to note that an OLS regression of accepted (observed) log-wages on the log-wage equation's regressors yields point estimates that differ sharply from the results of the Gibbs sampling algorithm. Table 22.4 contains point estimates and standard errors from such an accepted wage regression. Selection bias is apparent in the estimates of the log-wage equation's parameters in all data sets. This highlights the fact that the Bayesian simulation algorithm is doing an impressive job of implementing the appropriate dynamic selection correction.

Perhaps more interesting is the performance of the algorithm when taken to data that were generated using optimal decision rules. Table 22.5 reports the results of this analysis on data sets 1-EMAX to 5-EMAX. Again, the first column labels the parameter, and the second contains its data generating value. The performance of the algorithm is quite impressive. In almost all cases, the posterior means of the wage function parameters deviate only slightly from the true values in percentage terms. Also, the posterior standard deviations are in most cases quite small, suggesting that the data contain a great deal of information about these structural parameters – even without imposing the assumption that agents form the future component “optimally.” Finally, despite the fact that the posterior standard deviations are quite small, the posterior means are rarely more than two posterior deviations away from the true values.<sup>7</sup> As with the polynomial data, the standard deviation of the nonpecuniary component seems difficult to pin down. Unlike the polynomial data, the school payoff parameters are not pinned down as well as the wage equation parameters. This is perhaps not surprising since school payoffs are never observed.

Figure 22.1 contains the simulated posterior densities for a subset of the structural parameters based on data set 3-EMAX. Each figure includes three triangles on its horizontal axis. The middle triangle defines the posterior mean, and the two flanking triangles mark the points two posterior standard deviations above and below the mean. A vertical line is positioned at the parameters' data generating (true) value. These distributions emphasize the quality of the algorithm's performance in that the true parameter values are typically close to the posterior means. The figures also make clear that not all the parameters have approximately normal distributions. For instance, the posterior density of the wage equations' error correlation is multi-modal.

The results of Table 22.5 indicate that in a case where agents form the future component optimally, we can still obtain reliable and precise inferences about structural parameters of the current payoff functions using a simplified and misspecified model that says the future component is a simple fourth-order polynomial in the state variables. But we are also interested in how well our method approximates the decision rule used by the agents. In Table 22.6 we consider an experiment in which we use the posterior means for the parameters  $\pi$  that

**Table 22.2** Choice distributions and mean accepted wages in the data generated with true and OLS polynomial future components

Period	Data Set 2 – POLY						Data Set 2 – EMAX					
	Percent in occ. 1	Percent in occ. 2	Percent in school	Percent at home	Mean accepted wage		Percent in occ. 1	Percent in occ. 2	Percent in school	Percent at home	Mean accepted wage	
					Occ. 1	Occ. 2					Occ. 1	Occ. 2
1	0.10	0.00	0.75	0.15	13,762.19	17,971.88	0.09	0.00	0.79	0.12	13,837.93	19,955.02
2	0.23	0.01	0.58	0.17	11,822.16	19,032.16	0.23	0.01	0.60	0.15	11,941.31	18,502.51
3	0.41	0.04	0.34	0.21	11,249.67	16,521.61	0.39	0.03	0.39	0.18	11,275.37	16,786.46
4	0.52	0.06	0.20	0.22	11,167.03	16,209.88	0.50	0.04	0.27	0.19	11,208.15	17,778.61
5	0.60	0.06	0.14	0.20	11,417.94	16,141.39	0.57	0.05	0.20	0.18	11,598.75	16,955.70
6	0.63	0.08	0.10	0.19	11,802.61	16,427.58	0.63	0.06	0.14	0.16	11,897.28	17,100.50
7	0.65	0.09	0.07	0.18	12,257.30	16,987.46	0.68	0.08	0.08	0.16	12,286.26	17,624.68
8	0.69	0.10	0.05	0.16	12,701.01	17,067.03	0.72	0.09	0.05	0.13	12,751.92	17,300.99
9	0.69	0.10	0.05	0.16	13,167.06	18,442.74	0.72	0.10	0.05	0.13	13,159.23	19,498.18
10	0.70	0.11	0.05	0.14	13,709.21	18,274.23	0.74	0.09	0.05	0.12	13,790.83	19,125.16
11	0.72	0.12	0.05	0.11	14,409.81	19,391.23	0.75	0.11	0.04	0.10	14,546.63	19,867.95
12	0.71	0.14	0.04	0.12	14,511.54	19,730.21	0.74	0.12	0.04	0.10	14,650.45	20,320.53
13	0.72	0.14	0.05	0.10	15,216.89	21,641.41	0.75	0.14	0.03	0.08	15,439.04	21,723.49
14	0.74	0.13	0.03	0.09	15,943.12	21,866.44	0.76	0.14	0.02	0.08	16,150.59	22,096.07
15	0.73	0.16	0.03	0.07	16,507.05	22,177.61	0.75	0.16	0.03	0.06	16,773.15	22,764.69
16	0.74	0.16	0.03	0.07	17,129.96	22,624.51	0.75	0.16	0.03	0.06	17,437.26	22,786.18

17	0.75	0.17	0.02	0.06	17,886.20	24,194.23	0.75	0.18	0.02	0.05	18,276.74	24,804.59
18	0.73	0.17	0.02	0.07	18,408.75	24,318.34	0.74	0.18	0.01	0.06	18,786.43	24,476.18
19	0.72	0.19	0.02	0.06	19,590.88	25,385.99	0.73	0.20	0.01	0.05	19,961.17	25,719.67
20	0.74	0.19	0.02	0.05	20,186.07	25,161.39	0.75	0.20	0.01	0.04	20,571.89	25,422.50
21	0.71	0.23	0.01	0.05	21,113.74	26,409.20	0.71	0.24	0.01	0.04	21,613.91	26,613.20
22	0.70	0.25	0.01	0.04	22,002.82	26,935.39	0.70	0.25	0.00	0.04	22,488.94	27,566.90
23	0.67	0.29	0.01	0.03	23,259.72	28,191.41	0.67	0.29	0.01	0.03	23,655.61	28,952.46
24	0.66	0.30	0.00	0.03	23,119.46	28,634.21	0.66	0.31	0.00	0.03	23,706.23	29,491.67
25	0.66	0.30	0.00	0.04	24,085.78	30,826.10	0.66	0.31	0.00	0.03	24,535.54	31,403.33
26	0.62	0.34	0.01	0.04	25,399.34	30,707.99	0.63	0.34	0.00	0.03	26,003.31	31,157.72
27	0.62	0.34	0.00	0.04	26,971.71	32,251.61	0.62	0.35	0.00	0.03	27,482.83	33,112.86
28	0.60	0.37	0.00	0.03	27,074.62	32,024.07	0.60	0.37	0.00	0.02	27,805.46	32,743.79
29	0.57	0.40	0.00	0.03	29,049.11	32,411.14	0.58	0.39	0.00	0.03	29,596.82	33,872.97
30	0.55	0.42	0.00	0.04	30,492.25	34,513.76	0.56	0.41	0.00	0.03	31,216.48	35,462.35
31	0.52	0.45	0.00	0.03	30,745.54	35,672.21	0.52	0.45	0.00	0.03	31,744.63	36,763.93
32	0.50	0.48	0.00	0.03	32,078.16	36,076.17	0.51	0.47	0.00	0.03	33,016.52	37,028.17
33	0.46	0.50	0.00	0.04	34,202.82	37,460.57	0.47	0.51	0.00	0.03	34,905.34	38,435.23
34	0.43	0.54	0.00	0.02	34,578.60	38,293.38	0.44	0.54	0.00	0.02	35,656.54	39,212.14
35	0.42	0.55	0.00	0.04	37,084.91	39,690.50	0.43	0.54	0.00	0.03	38,195.57	40,767.33
36	0.39	0.58	0.00	0.03	37,580.47	40,970.75	0.40	0.57	0.00	0.03	39,119.51	41,740.41
37	0.36	0.60	0.00	0.04	40,129.34	41,885.28	0.37	0.60	0.00	0.03	41,228.89	42,901.62
38	0.33	0.64	0.00	0.03	40,101.57	43,929.61	0.34	0.63	0.00	0.03	41,477.02	45,076.14
39	0.28	0.67	0.00	0.05	43,282.44	44,724.22	0.30	0.66	0.00	0.04	44,266.27	46,039.50
40	0.26	0.70	0.00	0.03	44,462.69	45,703.45	0.28	0.69	0.00	0.03	45,668.36	46,847.68

**Table 22.3** Descriptive statistics for posterior distributions of the model's structural parameters for several different data sets generated using polynomial future component

Parameter	True	Data Set 1 - POLY		Data Set 2 - POLY		Data Set 3 - POLY		Data Set 4 - POLY		Data Set 5 - POLY	
		Mean	SD								
Occ. 1 intercept	9.00000	9.01300	0.00643	9.00125	0.00797	9.00845	0.00600	9.00256	0.00661	9.00309	0.00598
Occ. 1 own experience	0.05500	0.05440	0.00080	0.05540	0.00086	0.05429	0.00076	0.05462	0.00080	0.05496	0.00060
Occ. 2 experience	0.00000	0.00093	0.00095	-0.00121	0.00103	-0.00084	0.00092	0.00114	0.00112	-0.00129	0.00115
Education	0.05000	0.04806	0.00130	0.04881	0.00132	0.04850	0.00137	0.04938	0.00140	0.04924	0.00139
Occ. 1 exp. squared	-0.00025	-0.00024	0.00003	-0.00026	0.00003	-0.00025	0.00003	-0.00024	0.00003	-0.00026	0.00002
Occ. 1 error SD	0.40000	0.398	0.002	0.402	0.002	0.401	0.002	0.400	0.002	0.400	0.002
Occ. 2 intercept	8.95000	8.91712	0.01625	8.97693	0.01582	8.91699	0.01583	8.96316	0.01551	8.92501	0.01590
Occ. 2 own experience	0.04000	0.04049	0.00039	0.03918	0.00034	0.03946	0.00037	0.03968	0.00037	0.04055	0.00042
Occ. 1 experience	0.06000	0.06103	0.00175	0.06016	0.00178	0.06461	0.00173	0.05946	0.00180	0.06009	0.00178
Education	0.07500	0.07730	0.00187	0.07245	0.00161	0.07782	0.00178	0.07619	0.00167	0.07650	0.00171
Occ. 2 exp. squared	-0.00090	-0.00088	0.00008	-0.00093	0.00008	-0.00109	0.00008	-0.00092	0.00008	-0.00087	0.00008
Occ. 2 error SD	0.40000	0.409	0.003	0.399	0.003	0.407	0.003	0.397	0.003	0.399	0.003
Error correlation	0.50000	0.481	0.031	0.512	0.025	0.420	0.033	0.528	0.037	0.438	0.042
Undergraduate tuition	-5,000	-4,629	363	-5,212	464	-5,514	404	-4,908	464	-4,512	399
Graduate tuition	-15,000	-18,006	2,085	-16,711	1,829	-16,973	1,610	-16,817	1,692	-15,091	1,972
Return cost	-15,000	-14,063	531	-16,235	894	-15,809	554	-14,895	822	-15,448	679
Preference shock SD											
Occ. 1	9,082.95	10,121.05	255.49	9,397.38	577.26	10,578.31	537.01	9,253.22	615.27	9,494.74	354.07
Occ. 2	9,082.95	8,686.12	456.77	11,613.38	346.03	10,807.31	519.13	9,610.23	457.45	9,158.09	228.32
Occ. 3	11,821.59	11,569.93	281.18	12,683.67	803.09	13,418.79	358.98	12,019.38	417.75	12,247.80	401.15
Preference shock Corr.											
Occ. 1 with Occ. 2	0.89	0.90	0.01	0.96	0.01	0.93	0.01	0.91	0.01	0.90	0.02
Occ. 1 with Occ. 3	0.88	0.89	0.01	0.88	0.01	0.90	0.01	0.88	0.01	0.88	0.01
Occ. 2 with Occ. 3	0.88	0.89	0.01	0.89	0.01	0.89	0.01	0.89	0.01	0.89	0.01

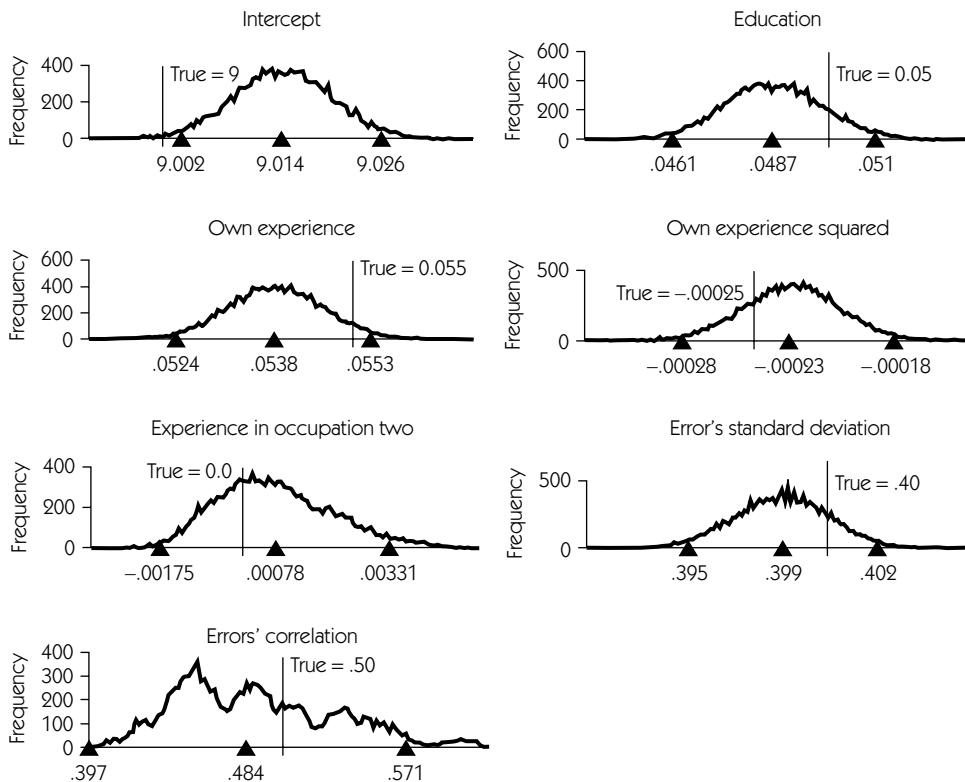
**Table 22.4** Log-wage equation estimates from OLS on observed wages generated under the polynomial future component\*

Data Set	Intercept	Occupation One				Occupation Two				Wage Error SDs		
		Occ. 1 Experience	Occ. 2 Experience	Education	Exp. Squared	Intercept	Occ. 1 Experience	Occ. 2 Experience	Education	Exp. Squared	Occ. 1	Occ. 2
TRUE	9.00000	0.05500	0.00000	0.05000	-0.00025	8.95000	0.04000	0.06000	0.07500	-0.00090	0.40000	0.40000
1 - POLY	9.15261	0.04236	0.01708	0.04247	0.00012	9.46735	0.03516	0.01953	0.05589	0.00038	0.38845	0.36574
0.000520	0.00076	0.00074	0.00133	0.00003	0.00906	0.00037	0.00146	0.00179	0.00008	0.00008		
2 - POLY	9.14715	0.04320	0.01586	0.04309	0.00010	9.47924	0.03446	0.02261	0.05311	0.00017	0.38940	0.36300
0.00528	0.00076	0.00073	0.00134	0.00003	0.00888	0.00036	0.00136	0.00170	0.00007	0.00007		
3 - POLY	9.14895	0.04230	0.01732	0.04420	0.00011	9.45851	0.03482	0.02335	0.05665	0.00016	0.38935	0.36495
0.00528	0.00076	0.00072	0.00135	0.00003	0.00914	0.00036	0.00140	0.00177	0.00007	0.00007		
4 - POLY	9.15157	0.04220	0.01734	0.04261	0.00012	9.46413	0.03480	0.02346	0.05469	0.00015	0.38900	0.36217
0.00527	0.00076	0.00074	0.00136	0.00003	0.00896	0.00037	0.00138	0.00174	0.00007	0.00007		
5 - POLY	9.14838	0.04274	0.01695	0.04408	0.00011	9.45131	0.03570	0.02021	0.05671	0.00035	0.38772	0.35781
0.00521	0.00076	0.00073	0.00135	0.00003	0.00880	0.00036	0.00139	0.00173	0.00007	0.00007		

\* Standard errors in italics.

**Table 22.5** Descriptive statistics for posterior distributions of the model's structural parameters for several different data sets generated using true future component

Parameter	True	Data Set 1 – EMAX		Data Set 2 – EMAX		Data Set 3 – EMAX		Data Set 4 – EMAX		Data Set 5 – EMAX	
		Mean	SD								
Occ. 1 intercept	9.00000	9.01342	0.00602	9.00471	0.00527	9.01436	0.00584	9.01028	0.00593	9.00929	0.00550
Occ. 1 own experience	0.05500	0.05427	0.00073	0.05489	0.00071	0.05384	0.00072	0.05394	0.00072	0.05410	0.00071
Occ. 2 experience	0.00000	0.00111	0.00093	0.00092	0.00114	0.00078	0.00126	0.00107	0.00100	0.00051	0.00093
Education	0.05000	0.04881	0.00118	0.05173	0.00126	0.04869	0.00129	0.04961	0.00123	0.05067	0.00124
Occ. 1 exp. squared	-0.00025	-0.00023	0.00002	-0.00025	0.00002	-0.00023	0.00002	-0.00022	0.00002	-0.00023	0.00002
Occ. 1 error SD	0.40000	0.397	0.002	0.399	0.002	0.399	0.002	0.397	0.002	0.397	0.002
Occ. 2 intercept	8.95000	8.90720	0.01704	8.98989	0.01970	8.93943	0.01850	8.93174	0.01649	8.94097	0.01410
Occ. 2 own experience	0.04000	0.04093	0.00037	0.03967	0.00037	0.03955	0.00038	0.04001	0.00037	0.04060	0.00039
Occ. 1 experience	0.06000	0.06087	0.00178	0.05716	0.00190	0.06200	0.00201	0.06211	0.00179	0.05880	0.00157
Education	0.07500	0.07822	0.00166	0.07338	0.00171	0.07579	0.00165	0.07743	0.00167	0.07613	0.00159
Occ. 2 exp. squared	-0.00090	-0.00087	0.00008	-0.00081	0.00008	-0.00098	0.00008	-0.00101	0.00008	-0.00084	0.00007
Occ. 2 error SD	0.40000	0.409	0.003	0.397	0.003	0.404	0.003	0.402	0.003	0.397	0.003
Error correlation	0.50000	0.517	0.023	0.607	0.029	0.484	0.044	0.521	0.035	0.488	0.028
Undergraduate tuition	-5,000	-2,261	313	-2,937	358	-3,407	371	-3,851	426	-3,286	448
Graduate tuition	-15,000	-10,092	1,046	-10,788	1,141	-11,983	1,188	-10,119	1,380	-11,958	1,823
Return cost	-15,000	-14,032	482	-16,014	431	-16,577	500	-16,168	662	-18,863	1,065
Preference shock SD											
Occ. 1	9,082.95	10,634.90	423.85	10,177.24	165.11	11,438.63	438.72	9,973.32	371.64	9,071.29	509.80
Occ. 2	9,082.95	9,436.10	372.86	12,741.02	405.25	11,432.19	287.69	9,310.37	718.15	7,770.66	555.39
Occ. 3	11,821.59	11,450.65	338.28	12,470.12	259.81	13,999.95	351.33	13,183.33	471.47	13,897.62	533.67
Preference shock corr.											
Occ. 1 with Occ. 2	0.89	0.93	0.01	0.98	0.00	0.94	0.01	0.91	0.02	0.86	0.03
Occ. 1 with Occ. 3	0.88	0.89	0.01	0.88	0.01	0.90	0.01	0.88	0.01	0.88	0.01
Occ. 2 with Occ. 3	0.88	0.87	0.01	0.90	0.01	0.90	0.01	0.89	0.02	0.89	0.02



**Figure 22.1** Marginal posterior densities of first log-wage equation's parameters from data set 3-EMAX\*

\* On each graph, the vertical line indicates the data generating parameter value. The middle triangle indicates the empirical mean, and the two flanking triangles are located two standard deviations from the mean

characterize how agents form expectations to form an estimate of agents' decision rules. We then simulate five new artificial data sets, using the exact same draws for the current period payoffs as were used to generate the original five artificial data sets. The only difference is that the estimated future component is substituted for the true future component in forming the decision rule. The results in Table 22.6 indicate that the mean wealth losses from using the estimated decision rule range from five-hundredths to three-tenths of 1 percent. The percentage of choices that agree between agents who use the optimal versus the approximate rules ranges from 89.8 to 93.5 percent. These results suggest that our estimated polynomial approximations to the optimal decision rules are reasonably accurate.

Figure 22.2 provides an alternative way to examine the quality of the polynomial approximation to the future component. This figure plots the value of the approximate and the true EMAX future components when evaluated at the mean

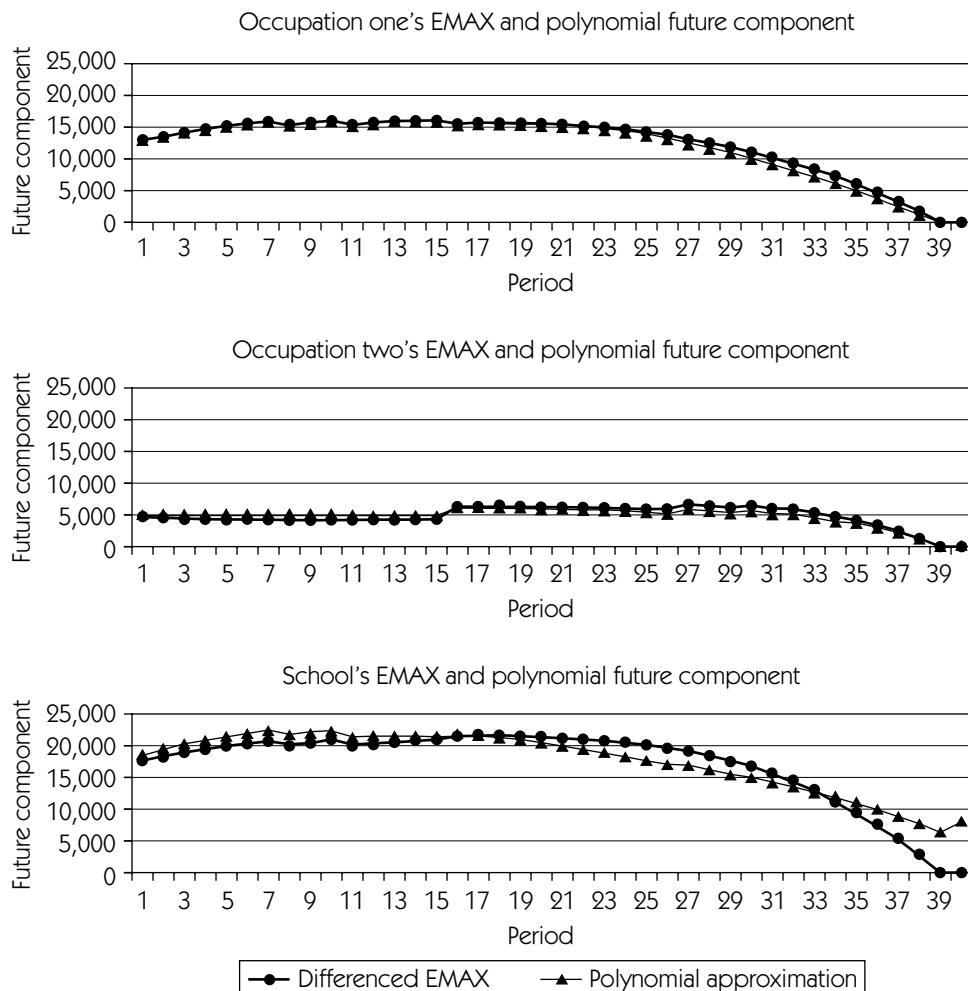
**Table 22.6** Wealth loss when posterior polynomial approximation is used in place of true future component\*

Data set	Using		Using			
	True EMAX**	Posterior EMAX**	Mean	Aggregate choice	Percent with 0–35 agreement (%)	Percent with 36–39 agreements (%)
	Mean present value of payoffs***	Mean dollar present value equivalent loss	Mean percent loss (%)	agreement (%)	agreement (%)	Percent choosing same path (%)
1-EMAX	356,796	356,134	663	0.19	90.80	34.25
2-EMAX	356,327	355,836	491	0.14	91.34	33.00
3-EMAX	355,797	354,746	1,051	0.30	89.79	39.00
4-EMAX	355,803	355,450	353	0.10	93.48	24.60
5-EMAX	355,661	355,485	176	0.05	93.18	24.95
					30.50	30.50

\* Polynomial parameter values are set to the mean of their respective empirical posterior distributions.

\*\* Each simulation includes 2,000 agents that live for exactly 40 periods.

\*\*\* The mean present value of payoffs is the equal-weight sample average of the discounted streams of ex-post lifetime payoffs.



**Figure 22.2** EMAX and polynomial future components evaluated at mean values of state variables at each period

of each period's state vector.<sup>8</sup> Each vertical axis corresponds to the value of the future component, and the horizontal axis is the period. Clearly, the approximation reflects the true EMAX future component's main features. The fit of the polynomial seems relatively strong for each occupational alternative throughout the lifecycle. The fit is good for school in early periods, but begins to deteriorate later. One reason is that school is chosen very infrequently after the first five periods, so there is increasingly less information about its future component. A second reason is that the contemporaneous return begins to dominate the future component in alternative valuations. Consequently, each data point in later periods contains relatively less information about the future component's value.

Overall, however, these figures furnish additional evidence that the polynomial approximation does a reasonable job of capturing the key characteristics of the true future component.

## 5 CONCLUSION

This chapter described how to implement a simulation based method for inference that is applicable to a wide class of dynamic multinomial choice models. The results of a Monte Carlo analysis demonstrated that the method works very well in relatively large state-space models with only partially observed payoffs where very high dimensional integrations are required. Although our discussion focused on models with discrete choices and independent and identically distributed stochastic terms, the method can also be applied to models with mixed continuous/discrete choice sets and serially correlated shocks (see Houser, 1999).

## APPENDIX A THE FUTURE COMPONENT

The future component we used was a fourth-order polynomial in the state variables. Below, in the interest of space and clarity, we will develop that polynomial only up to its third-order terms. The extension to the higher-order terms is obvious. From equation (22.12), the future component is the flexible functional form

$$F(X_{i1t} + \chi(j=1), X_{i2t} + \chi(j=2), S_{it} + \chi(j=3), t+1, \chi(j=3))$$

Define  $\iota_k \equiv \chi(j=k)$ . Then, to third-order terms, we used the following polynomial to represent this function.

$$\begin{aligned} F(X_1 + \iota_1, X_2 + \iota_2, S + \iota_3, t+1, \iota_3) = & P_1 + P_2(X_1 + \iota_1) + P_3(X_2 + \iota_2) + P_4(S + \iota_3) \\ & + P_5(t+1) + P_6(X_1 + \iota_1)^2 + P_7(X_2 + \iota_2)^2 + P_8(S + \iota_3)^2 + P_9(t+1)^2 + P_{10}(X_1 + \iota_1)^3 \\ & + P_{11}(X_2 + \iota_2)^3 + P_{12}(S + \iota_3)^3 + P_{13}(t+1)^3 + P_{14}(X_1 + \iota_1)^2(X_2 + \iota_2) + P_{15}(X_1 + \iota_1)^2(S + \iota_3) \\ & + P_{16}(X_1 + \iota_1)^2(t+1) + P_{17}(X_2 + \iota_2)^2(X_1 + \iota_1) + P_{18}(X_2 + \iota_2)^2(S + \iota_3) + P_{19}(X_2 + \iota_2)^2(t+1) \\ & + P_{20}(S + \iota_3)^2(X_1 + \iota_1) + P_{21}(S + \iota_3)^2(X_2 + \iota_2) + P_{22}(S + \iota_3)^2(t+1) + P_{23}(t+1)^2(X_1 + \iota_1) \\ & + P_{24}(t+1)^2(X_2 + \iota_2) + P_{25}(t+1)^2(S + \iota_3) + P_{26}\iota_3 + P_{27}\iota_3(X_1 + \iota_1) + P_{28}\iota_3(X_2 + \iota_2) \\ & + P_{29}\iota_3(S + \iota_3) + P_{30}\iota_3(t+1) + P_{31}\iota_3(X_1 + \iota_1)^2 + P_{32}\iota_3(X_2 + \iota_2)^2 + P_{33}\iota_3(S + \iota_3)^2 \\ & + P_{34}\iota_3(t+1)^2 \end{aligned}$$

The differenced future components used above are defined by  $f(I_{it}^*, j) = F(I_{it}^*, j) - F(I_{it}^*, 4)$ . Several of the parameters of the level future component drop out due to differencing. For instance, the intercept  $P_1$  and all coefficients of terms involving only  $(t+1)$  vanish. Simple algebra reveals the differenced future components have the following forms.

$$\begin{aligned} f(I_{it}^*, 1) &= \pi_1 + \pi_2 g(X_1) + \pi_3 h(X_1) + \pi_4 X_2 g(X_1) + \pi_5 S g(X_1) + \pi_6(t+1)g(X_1) \\ &\quad + \pi_7 X_2^2 + \pi_8 S_2^2 + \pi_9(t+1)^2. \end{aligned}$$

$$\begin{aligned} f(I_{it}^*, 2) &= \pi_4 X_1^2 + \pi_7 X_1 g(X_2) + \pi_{10} + \pi_{11} g(X_2) + \pi_{12} h(X_2) + \pi_{13} S g(X_2) + \pi_{14}(t+1)g(X_2) \\ &\quad + \pi_{15} S^2 + \pi_{16}(t+1)^2. \end{aligned}$$

$$\begin{aligned} f(I_{it}^*, 3) &= \pi_5 X_1^2 + \pi_8 X_1 g(S) + \pi_{13} X_2^2 + \pi_{15} X_2 g(S) + \pi_{17} + \pi_{18} g(S) + \pi_{19} h(S) + \pi_{20} X_1^2 \\ &\quad + \pi_{21} X_2^2 + \pi_{22}(t+1)g(S) + \pi_{23}(t+1)^2 + \pi_{24} X_1 + \pi_{25} X_2 + \pi_{26}(t+1). \end{aligned}$$

where  $g(x) = 2x + 1$ , and  $h(x) = 3x^2 + 3x + 1$ . Several of the parameters appear in multiple equations. Such cross equation restrictions reflect the specification's logical consistency. The future components' asymmetry arises since choosing school both augments school experience and removes the cost of returning to school that one would otherwise face. In contrast, choosing alternative one or two only augments experience within that alternative.

## APPENDIX B EXISTENCE OF JOINT POSTERIOR DISTRIBUTION

Let  $\omega$  denote the number of missing wage observations, and let  $\Omega \equiv [A_\Omega, B_\Omega]^\omega \in \Re_{++}^\omega$  be the domain of unobserved wages, where  $0 < A_\Omega < B_\Omega < \infty$ . Also, let  $\Delta \equiv [A_\Delta, B_\Delta]^{3NT} \in \Re^{3NT}$  be the domain of latent relative utilities, where  $-\infty < A_\Delta < B_\Delta < \infty$ . We want to show that:

$$\int_{\Omega, \Delta, \beta, \pi, \Sigma_\epsilon^{-1}, \Sigma_\eta^{-1}} \left( \prod_{i,t} (w_{i1t} w_{i2t})^{-1} \right) g(V, \beta, \Sigma_\epsilon^{-1}) h(Z, W, \pi, \Sigma_\eta^{-1}) < \infty. \quad (22.B1)$$

Here, we have subsumed  $\alpha$  and  $\Lambda$  into  $\pi$  and  $\Psi$ , respectively (as we did in Step 5 of Section 3) and defined

$$g(V, \beta, \Sigma_\epsilon^{-1}) = |\Sigma_\epsilon^{-1}|^{\frac{NT-3}{2}} \exp\{-\frac{1}{2}(V - Y\beta)'(\Sigma_\epsilon^{-1} \otimes I_{NT})(V - Y\beta)\}, \quad (22.B2)$$

where

$$V = \begin{pmatrix} V_1 \\ V_2 \end{pmatrix}, \quad V_i = \begin{pmatrix} \ln w_{1it} \\ \vdots \\ \ln w_{Nit} \end{pmatrix}, \quad Y = \begin{bmatrix} Y_1 & 0 \\ 0 & Y_2 \end{bmatrix} \quad \text{and} \quad \beta = \begin{bmatrix} \beta_1 \\ \beta_2 \end{bmatrix}. \quad (22.B3)$$

We have also defined

$$\begin{aligned} h(Z, W, \pi, \Sigma_\eta^{-1}) &= |\Sigma_\eta^{-1}|^{\frac{NT-4}{2}} \exp\{-\frac{1}{2}(Z - W - \Psi\pi)'(\Sigma_\eta^{-1} \otimes I_{NT})(Z - W - \Psi\pi)\} \\ &\cdot I(Z_{ijt} > 0, Z_{ikt}(k \neq j) < 0 \text{ if } d_{it} = j \text{ and } j \in \{1, 2, 3\}, \{Z_{ijt}\}_{j=1,2,3} < 0 \text{ otherwise}), \end{aligned} \quad (22.B4)$$

where

$$Z = \begin{pmatrix} Z_1 \\ Z_2 \\ Z_3 \end{pmatrix}, \quad Z_i = \begin{pmatrix} Z_{i1l} \\ \vdots \\ Z_{NiT} \end{pmatrix}, \quad W = \begin{pmatrix} W_1 \\ W_2 \\ W_3 \end{pmatrix}, \quad W_i = \begin{pmatrix} w_{i1l} \\ \vdots \\ w_{NiT} \end{pmatrix} (i = 1, 2), \quad \text{and} \quad W_3 = (0).$$

Since the only arguments common to both  $g$  and  $h$  are those that include functions of wages, we can express (22.B1) as

$$\int_{\Omega} \left\{ \left( \prod_{i,t} (w_{i1t} w_{i2t})^{-1} \right) \int_{\beta, \Sigma_{\epsilon}^{-1}} g(V, \beta, \Sigma_{\epsilon}^{-1}) \int_{\Delta, \pi, \Sigma_{\eta}^{-1}} h(Z, W, \pi, \Sigma_{\eta}^{-1}) \right\}. \quad (22.B5)$$

We first observe that for any configuration of unobserved wages in  $\Omega$ ,

$$g(V) \equiv \int_{\beta, \Sigma_{\epsilon}^{-1}} g(V, \beta, \Sigma_{\epsilon}^{-1}) < \infty. \quad (22.B6)$$

To see this, note that we can express  $g(V, \beta, \Sigma_{\epsilon}^{-1})$  as

$$g(V, \beta, \Sigma_{\epsilon}) = |\Sigma_{\epsilon}^{-1}|^{\frac{NT-3}{2}} \exp\left\{-\frac{1}{2}\text{tr}(S(\hat{\beta})\Sigma_{\epsilon}^{-1}) - \frac{1}{2}(\beta - \hat{\beta})'Y'(\Sigma_{\epsilon}^{-1} \otimes I_{NT})Y(\beta - \hat{\beta})\right\}, \quad (22.B7)$$

where

$$\begin{aligned} \hat{\beta} &= (Y'(\Sigma_{\epsilon}^{-1} \otimes I_{NT})Y)^{-1}Y'(\Sigma_{\epsilon}^{-1} \otimes I_{NT})V \\ S(\hat{\beta}) &= (V_1 - Y_1\hat{\beta}_1, V_2 - Y_2\hat{\beta}_2)'(V_1 - Y_1\hat{\beta}_1, V_2 - Y_2\hat{\beta}_2) \end{aligned} \quad (22.B8)$$

Hence,  $g(V, \beta, \Sigma_{\epsilon})$  is proportional to a normal-Wishart density (Bernardo and Smith, 1994, p. 140), hence finitely integrable to a value  $g(V)$  that, in general, will depend on the configuration of unobserved wages. Since  $g(V)$  is finite over the compact set  $\Omega$ , it follows that  $g(V)$  is bounded over  $\Omega$ .

Turn next to  $h(Z, W, \pi, \Sigma_{\eta}^{-1})$ . Since (22.B4) has the same form as (22.B2), just as (22.B7) we can write:

$$\begin{aligned} h(Z, W, \pi, \Sigma_{\eta}^{-1}) &= |\Sigma_{\eta}^{-1}|^{\frac{NT-4}{2}} \exp\left\{-\frac{1}{2}\text{tr}(S(\hat{\pi})\Sigma_{\eta}^{-1}) - \frac{1}{2}(\pi - \hat{\pi})'\Psi'(\Sigma_{\eta}^{-1} \otimes I_{NT})\Psi(\pi - \hat{\pi})\right\} \\ &\cdot I(Z_{ijt} > 0, Z_{iRt}(k \neq j) < 0 \text{ if } d_{it} = j \text{ and } j \in \{1, 2, 3\}, \{Z_{ijt}\}_{j=1,2,3} < 0 \text{ otherwise}) \end{aligned} \quad (22.B9)$$

where  $\hat{\pi}$  and  $S(\hat{\pi})$  are defined in a way that is exactly analogous to (22.B8).

Hence,

$$h(Z, W) \equiv \int_{\Delta, \pi, \Sigma_{\eta}^{-1}} h(Z, W, \pi, \Sigma_{\eta}^{-1}) < \infty \quad (22.B10)$$

for any configuration of unobserved wages in  $\Omega$  and latent utilities in  $\Delta$ . It follows that  $h(Z, W)$  is bounded over  $\Omega \times \Delta$ . Thus, the integral (B1) reduces to

$$\int_{\Omega, \Delta} \left\{ \left( \prod_{i,t} (w_{i1t} w_{i2t})^{-1} \right) g(V) h(Z, W) \right\} \quad (22.B11)$$

which is finite since each element of the integrand is bounded over the compact domain of integration.

## Notes

- 1 Currently, approaches to numerical integration such as quadrature and series expansion are not useful if the dimension of the integration is greater than three or four.
- 2 Restrictions of this type can be tested easily by estimating versions of the model with different but nested future components.
- 3 These findings are related to those of Lancaster (1997), who considered Bayesian inference in the stationary job search model. He found that if the future component is treated as a free parameter (rather than being set “optimally” as dictated by the offer wage function, offer arrival rate, unemployment benefit and discount rate) there is little loss of information about the structural parameters of the offer wage functions. (As in our example, however, identification of the discount factor is lost.) The stationary job search model considered by Lancaster (1997) has the feature that the future component is a constant (i.e. it is not a function of state variables). Our procedure of treating the future component as a polynomial in state variables can be viewed as extending Lancaster’s approach to a much more general class of models.
- 4 As noted earlier, the future component’s arguments reflect restrictions implied by the model. For instance, because the productivity and preference shocks are serially independent, they contain no information useful for forecasting future payoffs and do not appear in the future component’s arguments. Also, given total experience in each occupation, the order in which occupations one and two were chosen in the past does not bear on current or future payoffs. Accordingly, only total experience in each occupation enters the future component.
- 5 To begin the Gibbs algorithm we needed an initial guess for the model’s parameters (although the asymptotic behavior of the Gibbs sampler as the number of cycles grows large is independent of starting values). We chose to set the log-wage equation  $\beta_3$  equal to the value from an OLS regression on observed wages. The diagonal elements of  $\Sigma_\epsilon$  were set to the variance of observed log-wages, while the off-diagonal elements were set to zero. The school payoff parameters were all initialized at zero. All of the future component’s  $\pi$  values were also started at zero, with the exception of the alternative-specific intercepts. The intercepts for alternatives one, two, and three were initialized at -5,000, 10,000, and 20,000, respectively. These values were chosen with an eye towards matching aggregate choice frequencies in each alternative. We initialized the  $\Sigma_\eta$  covariance matrix by setting all off-diagonal elements to zero, and each diagonal element to  $5 \times 10^8$ . We used large initial variances because doing so increases the size of the initial Gibbs steps, and seems to improve the rate of convergence of the algorithm.
- 6 Space considerations prevent us from reporting results for individual expectations parameters. Instead, below we will graphically compare the form of the estimated future component to that which was used to generate the data.

- 7 We also ran OLS accepted log-wage regressions for the 1-EMAX through 5-EMAX data sets. The results are very similar to those in Table 22.4, so we do not report them here. The estimates again show substantial biases for all the wage equation parameters. Thus, the Gibbs sampling algorithm continues to do an impressive job of implementing a dynamic selection correction despite the fact that the agents' decision rules are misspecified.
- 8 The mean state vectors were derived from the choices in data set 5-EMAX, and the coefficients of the polynomial were valued at the posterior means derived from the analysis of data set 5-EMAX.

## References

- Bernardo, J., and A.F.M. Smith (1994). *Bayesian Theory*. New York: John Wiley and Sons, Ltd.
- Eckstein, Z., and K. Wolpin (1989). The specification and estimation of dynamic stochastic discrete choice models. *Journal of Human Resources* 24, 562–98.
- Gelman, A. (1996). Inference and monitoring convergence. In W.R. Gilks, S. Richardson, and D.J. Spiegelhalter (eds.) *Markov Chain Monte Carlo in Practice*. pp. 131–43. London: Chapman & Hall.
- Geweke, J. (1995). Priors for macroeconomic time series and their application. Working paper, Federal Reserve Bank of Minneapolis.
- Geweke, J. (1996). Monte Carlo simulation and numerical integration. In H.M. Amman, D.A. Kendrick, and J. Rust (eds.) *Handbook of Computational Economics*, Volume 1, pp. 731–800.
- Geweke, J., and M. Keane (1999a). Bayesian inference for dynamic discrete choice models without the need for dynamic programming. In Mariano, Schuermann, and Weeks (eds.) *Simulation Based Inference and Econometrics: Methods and Applications*. Cambridge: Cambridge University Press. (Also available as Federal Reserve Bank of Minneapolis working paper #564, January, 1996.)
- Geweke, J., and M. Keane (1999b). Computationally intensive methods for integration in econometrics. In J. Heckman and E. Leamer (eds.) *Handbook of Econometrics*, Volume 5. Amsterdam: North-Holland.
- Geweke, J., M. Keane, and D. Runkle (1994). Alternative computational approaches to inference in the multinomial probit model. *Review of Economics and Statistics* 76, 609–32.
- Gilks, W.R., S. Richardson, and D.J. Spiegelhalter (1996). Introducing Markov Chain Monte Carlo. In W.R. Gilks, S. Richardson, and D.J. Spiegelhalter (eds.) *Markov Chain Monte Carlo in Practice*. pp. 1–20. London: Chapman & Hall.
- Heckman, J. (1976). The common structure of statistical models of truncation, sample selection and limited dependent variables and a simple estimator for such models. *Annals of Economic and Social Measurement* 5, 475–92.
- Heckman, J. (1981). Statistical models for discrete panel data. In C. Manski and D. McFadden (eds.) *Structural Analysis of Discrete Data with Econometric Applications*. pp. 115–78. Cambridge, MA: MIT Press.
- Heckman, J., and G. Sedlacek (1985). Heterogeneity, aggregation and market wage functions: an empirical model of self-selection in the labor market. *Journal of Political Economy* 93, 1077–125.
- Hotz, V.J., and R.A. Miller (1993). Conditional choice probabilities and the estimation of dynamic programming models. *Review of Economic Studies* 60, 497–530.
- Houser, D. (1999). Bayesian analysis of a dynamic, stochastic model of labor supply and saving. Manuscript, University of Arizona.

- Keane, M., and K. Wolpin (1994). Solution and estimation of discrete choice dynamic programming models by simulation and interpolation: Monte Carlo Evidence. *Review of Economics and Statistics* 76, 648–72.
- Keane, M., and K. Wolpin (1997). The career decisions of young men. *Journal of Political Economy* 105, 473–522.
- Krusell, P., and A.A. Smith (1995). Rules of thumb in macroeconomic equilibrium: a quantitative analysis. *Journal of Economic Dynamics and Control* 20, 527–58.
- Lancaster, T. (1997). Exact structural inference in optimal job search models. *Journal of Business and Economic Statistics* 15, 165–79.
- Lerman, S., and C. Manski (1981). On the use of simulated frequencies to approximate choice probabilities. In C. Manski and D. McFadden (eds.) *Structural Analysis of Discrete Data with Econometric Applications*. pp. 305–19. Cambridge, MA: MIT Press.
- Manski, C. (1993). Dynamic choice in social settings. *Journal of Econometrics* 58, 121–36.
- McFadden, D. (1989). A method of simulated moments for estimation of discrete response models without numerical integration. *Econometrica* 57, 995–1026.
- Monfort, A., Van Dijk, H., and B. Brown (eds.) (1995). *Econometric Inference Using Simulation Techniques*. New York: John Wiley and Sons, Ltd.
- Roy, A.D. (1951). Some thoughts on the distribution of earnings. *Oxford Economic Papers* 3, 135–46.
- Rust, J. (1987). Optimal replacement of GMC bus engines: an empirical model of Harold Zurcher. *Econometrica* 55, 999–1033.
- Rust, J. (1994). Estimation of dynamic structural models, problems and prospects: discrete decision processes. In C. Sims (ed.) *Advances in Econometrics, Sixth World Congress Vol. II*. pp. 119–70. Cambridge: Cambridge University Press.
- Tierney, L. (1994). Markov chains for exploring posterior distributions. *Annals of Statistics* 22, 1701–32.
- Wolpin, K. (1984). An estimable dynamic stochastic model of fertility and child mortality. *Journal of Political Economy* 92, 852–74.
- Zellner, A. (1971). *An Introduction to Bayesian Inference in Econometrics*. New York: John Wiley and Sons.

CHAPTER TWENTY-THREE

# Monte Carlo Test Methods in Econometrics

*Jean-Marie Dufour and Lynda Khalaf\**

## 1 INTRODUCTION

During the last 20 years, computer-based simulation methods have revolutionized the way we approach statistical analysis. This has been made possible by the rapid development of increasingly quick and inexpensive computers. Important innovations in this field include the bootstrap methods for improving standard asymptotic approximations (for reviews, see Efron, 1982; Efron and Tibshirani, 1993; Hall, 1992; Jeong and Maddala, 1993; Vinod, 1993; Shao and Tu, 1995; Davison and Hinkley, 1997; Horowitz, 1997) and techniques where estimators and forecasts are obtained from criteria evaluated by simulation (see Mariano and Brown, 1993; Hajivassiliou, 1993; Keane, 1993; Gouriéroux, Monfort, and Renault, 1993; Gallant and Tauchen, 1996). An area of statistical analysis where such techniques can make an important difference is hypothesis testing which often raises difficult distributional problems, especially in view of determining appropriate critical values.

This paper has two major objectives. First, we review some basic notions on hypothesis testing from a finite-sample perspective, emphasizing in particular the specific role of hypothesis testing in statistical analysis, the distinction between the level and the size of a test, the notions of exact and conservative tests, as well as randomized and non-randomized procedures. Second, we present a relatively informal overview of the possibilities of Monte Carlo test techniques, whose original idea originates in the early work of Dwass (1957), Barnard (1963) and Birnbaum (1974), in econometrics. This technique has the great attraction of providing provably *exact* (randomized) tests based on any statistic whose finite

sample distribution may be intractable but can be simulated. Further, the validity of the tests so obtained does not depend on the number of replications employed (which can be small). These features may be contrasted with the bootstrap, which only provides asymptotically justified (although hopefully improved) large-sample approximations.

In our presentation, we will try to address the fundamental issues that will allow the practitioners to use Monte Carlo *test* techniques. The emphasis will be on concepts rather than technical detail, and the exposition aims at being intuitive. The ideas will be illustrated using practical econometric problems. Examples discussed include: specification tests in linear regressions contexts (normality, independence, heteroskedasticity and conditional heteroskedasticity), nonlinear hypotheses in univariate and SURE models, and tests on structural parameters in instrumental regressions. More precisely, we will discuss the following themes.

In Section 2, we identify the important statistical issues motivating this econometric methodology, as an alternative to standard procedures. The issues raised have their roots in practical test problems and include:

- an *exact* test strategy: what is it, and why should we care?;
- the *nuisance-parameter* problem: what does it mean to practitioners?;
- understanding the *size/level* control problem;
- pivotal and *boundedly-pivotal* test criteria: why is this property important?;
- identification and near non-identification: a challenging setting.

Further, the relevance and severity of the problem will be demonstrated using simulation studies and/or empirical examples.

Sections 3 and 4 describe the Monte Carlo (MC) test method along with various econometric applications of it. Among other things, the procedure is compared and contrasted with the bootstrap. Whereas bootstrap tests are asymptotically valid (as both the numbers of observations and simulated samples go to  $\infty$ ), a formal demonstration is provided to emphasize the size control property of MC tests. Monte Carlo tests are typically discussed in parametric contexts. Extensions to nonparametric problems are also discussed. The theory is applied to a broad spectrum of examples that illustrate the usefulness of the procedure. We conclude with a word of caution on inference problems that cannot be solved by simulation. For convenience, the concepts and themes covered may be outlined as follows.

- MC tests based on pivotal statistics: an exact randomized test procedure;
- MC tests in the presence of nuisance parameters:
  - (a) local MC  $p$ -value,
  - (b) bounds MC  $p$ -value,
  - (c) maximized MC  $p$ -value;
- MC tests versus the bootstrap:
  - (a) fundamental differences/similarities,
  - (b) the number of simulated samples: theory and guidelines;

- MC tests: breakthrough improvements and “success stories”:
  - (a) The intractable null distributions problem (e.g. tests for normality, uniform linear hypothesis in multi-equation models, tests for ARCH),
  - (b) MC tests or Bartlett corrections?,
  - (c) The case of unidentified nuisance parameters (test for structural jumps, test for ARCH-M);
- MC tests may fail: where and why? a word of caution.

We conclude in Section 5.

## 2 STATISTICAL ISSUES: A PRACTICAL APPROACH TO CORE QUESTIONS

The hypothesis testing problem is often presented as one of deciding between two hypotheses: the hypothesis of interest (the *null*  $H_0$ ) and its complement (the *alternative*  $H_A$ ). For the purpose of the exposition, consider a test problem pertaining to a *parametric* model  $(\mathcal{Y}, P_\theta)$ , i.e. the case where the data generating process (DGP) is determined up to a *finite* number of unknown real parameters  $\theta \in \Theta$ , where  $\Theta$  refers to the parameter space (usually a vector space),  $\mathcal{Y}$  is the sample space,  $P_\theta$  is the family of probability distributions on  $\mathcal{Y}$ . Furthermore, let  $Y$  denote the observations, and  $\Theta_0$  the subspace of  $\Theta$  compatible with  $H_0$ .

A statistical test partitions the sample space into two subsets: a set consistent with  $H_0$  (the acceptance region), and its complements whose elements are viewed as inconsistent with  $H_0$  (the rejection region, or the *critical region*). This may be translated into a decision rule based on a *test statistic*  $S(Y)$ : the rejection region is defined as the numerical values of the test statistic for which the null will be rejected.

Without loss of generality, we suppose the critical region has the form  $S(Y) \geq c$ . To obtain a test of level  $\alpha$ ,  $c$  must be chosen so that the probability of rejecting the null hypothesis  $P_\theta[S(Y) \geq c]$  when  $H_0$  is true (the probability of a *type I error*) is not greater than  $\alpha$ , i.e. we must have:

$$\sup_{\theta \in \Theta_0} P_\theta[S(Y) \geq c] \leq \alpha. \quad (23.1)$$

Further, the test has *size*  $\alpha$  if and only if

$$\sup_{\theta \in \Theta_0} P_\theta[S(Y) \geq c] = \alpha. \quad (23.2)$$

To solve for  $c$  in (23.1) or (23.2), it is necessary to extract the finite-sample distribution of  $S(Y)$  when the null is true. Typically,  $S(Y)$  is a complicated function of the observations and the statistical problem involved is often intractable. More importantly, it is evident from the definitions (23.1)–(23.2) that, in many cases of practical interest, the distribution of  $S(Y)$  may be different for different parameter values. When the null hypothesis completely fixes the value of  $\theta$  (i.e.  $\Theta_0$  is

a point), the hypothesis is called a *simple hypothesis*. Most hypotheses encountered in practice are *composite*, i.e. the set  $\Theta_0$  contains more than one element. The null may uniquely define some parameters, but almost invariably some other parameters are not restricted to a point-set. In the context of composite hypotheses, some unknown parameters may appear in the distribution of  $S(Y)$ . Such parameters are called *nuisance parameters*.

When we talk about an *exact test*, it must be understood that attention is restricted to level-correct critical regions, where (23.1) must hold for a given finite sample size, for all values of the parameter  $\theta$  compatible with the null. Consequently, in carrying out an exact test, one may encounter two problems. The first one is to extract the analytic form of the distribution of  $S(Y)$ . The second one is to maximize the rejection probability over the relevant nuisance parameter space, subject to the level constraint. We will see below that the first problem can easily be solved when Monte Carlo test techniques are applicable. The second one is usually more difficult to tackle, and its importance is not fully recognized in econometric practice.

A reasonable solution to both problems often exists when one is dealing with large samples. Whereas the null distribution of  $S(Y)$  may be complicated and/or may involve unknown parameters, its asymptotic null distribution in many common cases has a known form and is nuisance-parameter-free (e.g., a normal or chi-square distribution). The critical point may conveniently be obtained using asymptotic arguments. The term *approximate critical point* is more appropriate here, since we are dealing with asymptotic levels: the critical values which yield the desired size  $\alpha$  for a given sample size can be very different from these approximate values obtained through an asymptotic argument. For sufficiently large sample sizes, the standard asymptotic approximations are expected to work well. The question is, and will remain, *how large is large?* To illustrate this issue, we next consider several examples involving commonly used econometric methods. We will demonstrate, by simulation, that asymptotic procedures may yield highly unreliable decisions, with empirically relevant sample sizes. The problem, and our main point, is that *finite sample accuracy is not merely a small sample problem*.

## 2.1 Instrumental regressions

Consider the *limited information* (LI) structural regression model:

$$y = Y\beta + X_1\gamma_1 + u = Z\delta + u, \quad (23.3)$$

$$Y = X_1\Pi_1 + X_2\Pi_2 + V, \quad (23.4)$$

where  $Y$  and  $X_1$  are  $n \times m$  and  $n \times k$  matrices which respectively contain the observations on the included endogenous and exogenous variables,  $Z = [Y, X_1]$ ,  $\delta = (\beta', \gamma_1')'$  and  $X_2$  refers to the excluded exogenous variables. If more than  $m$  variables are excluded from the structural equation, the system is said to be *overidentified*. The associated LI reduced form is:

$$[y \quad Y] = X\Pi + [v \quad V], \quad \Pi = \begin{bmatrix} \pi_1 & \Pi_1 \\ \pi_2 & \Pi_2 \end{bmatrix}, \quad (23.5)$$

$$\pi_1 = \Pi_1\beta + \gamma_1, \quad \pi_2 = \Pi_2\beta. \quad (23.6)$$

The necessary and sufficient condition for identification follows from the relation  $\pi_2 = \Pi_2\beta$ . Indeed  $\beta$  is recoverable if and only if

$$\text{rank}(\Pi_2) = m. \quad (23.7)$$

To test the general linear hypothesis  $R\delta = r$ , where  $R$  is a full row rank  $q \times (m+k)$  matrix, the well-known IV analog of the Wald test is frequently applied on grounds of computational ease. For instance, consider the two-stage least squares (2SLS) estimator

$$\hat{\delta} = [Z'P(P'P)^{-1}P'Z]^{-1}Z'P(P'P)^{-1}P'y, \quad (23.8)$$

where  $P$  is the following matrix of instruments  $P = [X, X(X'X)^{-1}X'Y]$ . Application of the Wald principle yields the following criterion

$$\tau_w = \frac{1}{s^2} (r - R\hat{\delta})'[R'(Z'P(P'P)^{-1}P'Z)^{-1}R](r - R\hat{\delta}), \quad (23.9)$$

where  $s^2 = \frac{1}{n}(y - Z\hat{\delta})'(y - Z\hat{\delta})'$ . Under usual regularity conditions and imposing identification,  $\tau_w$  is distributed like a  $\chi^2(q)$  variable, where  $q = \text{rank}(R)$ .

Bartlett (1948) and Anderson and Rubin (1949, henceforth AR) suggested an exact test that can be applied only if the null takes the form  $\beta = \beta^0$ . The idea behind the test is quite simple. Define  $y^* = y - Y\beta^0$ . Under the null, the model can be written as  $y^* = X_1\gamma_1 + u$ . On the other hand, if the hypothesis is not true,  $y^*$  will be a linear function of all the exogenous variables. Thus, the null may be assessed by the F-statistic for testing whether the coefficients of the regressors  $X_2$  "excluded" from (23.3) are zero in the regression of  $y^*$  on all the exogenous variables, i.e. we simply test  $\gamma_2 = 0$  in the extended linear regression  $y^* = X_1\gamma_1 + X_2\gamma_2 + u$ .

We first consider a simple experiment based on the work of Nelson and Startz (1990a, 1990b) and Staiger and Stock (1997). The model considered is a special case of (23.3) with two endogenous variables ( $p = 2$ ) and  $k = 1$  exogenous variables. The structural equation includes only the endogenous variable. The restrictions tested are of the form  $H_{01} : \beta = \beta^0$ . The sample sizes are set to  $n = 25, 100, 250$ . The exogenous regressors are independently drawn from the standard normal distribution. These are drawn only once. The errors are generated according to a multinormal distribution with mean zero and covariance matrix

$$\Sigma = \begin{bmatrix} 1 & .95 \\ .95 & 1 \end{bmatrix}. \quad (23.10)$$

**Table 23.1** IV-based Wald/Anderson–Rubin tests: empirical type I errors

$\Pi_2$	$n = 25$		$n = 100$		$n = 250$	
	<i>Wald</i>	<i>AR</i>	<i>Wald</i>	<i>AR</i>	<i>Wald</i>	<i>AR</i>
1	0.061	0.059	0.046	0.046	0.049	0.057
0.9	0.063	0.059	0.045	0.046	0.049	0.057
0.7	0.071	0.059	0.046	0.046	0.052	0.057
0.5	0.081	0.059	0.060	0.046	0.049	0.057
0.2	0.160	0.059	0.106	0.046	0.076	0.057
0.1	0.260	0.059	0.168	0.046	0.121	0.057
0.05	0.332	0.059	0.284	0.046	0.203	0.057
0.01	0.359	0.059	0.389	0.046	0.419	0.057

The other coefficients are:

$$\beta = \beta^0 = 0; \quad \Pi_2 = 1, .9, .7, .5, .2, .1, .05, .01. \quad (23.11)$$

In this case, the 2SLS-based test corresponds to the standard *t*-test (see Nelson and Startz (1990b) for the relevant formulae). 1,000 replications are performed. Table 23.1 reports probabilities of type I error [ $P(\text{type I error})$ ] associated with the two-tailed 2SLS *t*-test for the significance of  $\beta$  and the corresponding Anderson–Rubin test. In this context, the identification condition reduces to  $\Pi_2 \neq 0$ ; this condition can be tested using a standard F-test in the first stage regression.<sup>1</sup> It is evident that IV-based Wald tests perform very poorly in terms of size control. Identification problems severely distort test sizes. While the evidence of size distortions is notable even in identified models, the problem is far more severe in near-unidentified situations. More importantly, increasing the sample size does not correct the problem. In this regard, Bound, Jaeger, and Baker (1995) report severe bias problems associated with IV-based estimators, despite very large sample sizes. In contrast, the Anderson–Rubin test, when available, is immune to such problems: the test is exact, in the sense that the null distribution of the AR criterion does not depend on the parameters controlling identification. Indeed, the AR test statistic follows an  $F(m, n - k)$  distribution, regardless of the identification status. The AR test has recently received renewed attention; see, for example, Dufour and Jasiak (1996) and Staiger and Stock (1997). Recall, however, that the test is not applicable unless the null sets the values of the coefficients of all the endogenous variables. On general linear structural restrictions, see Dufour and Khalaf (1998b).

Despite the recognition of the need for caution in the application of IV-based tests, standard econometric software packages typically implement IV-based Wald tests. In particular, the *t*-tests on individual parameters are routinely computed in the context of 2SLS or 3SLS procedures. Unfortunately, the Monte Carlo

experiments we have analyzed confirm that IV-based Wald tests realize computational savings at the risk of very poor reliability.

## 2.2 Normality tests

Let us now consider the fundamental problem of testing disturbance normality in the context of the linear regression model:

$$Y = X\beta + u, \quad (23.12)$$

where  $Y = (y_1, \dots, y_n)'$  is a vector of observations on the dependent variable,  $X$  is the matrix of  $n$  observations on  $k$  regressors,  $\beta$  is a vector of unknown coefficients and  $u = (u_1, \dots, u_n)'$  is an  $n$ -dimensional vector of iid disturbances. The problem consists in testing:

$$H_0 : f(u) = \varphi(u; 0, \sigma), \quad \sigma > 0, \quad (23.13)$$

where  $f(u)$  is the probability density function (pdf) of  $u_i$ , and  $\varphi(u; \mu, \sigma)$  is the normal pdf with mean  $\mu$  and standard deviation  $\sigma$ . In this context, normality tests are typically based on the least squares residual vector

$$\hat{u} = y - X\hat{\beta} = M_X u, \quad (23.14)$$

where  $\hat{\beta} = (X'X)^{-1} X'y$  and  $M_X = I_n - X(X'X)^{-1}X'$ . Let  $\hat{u}_{1n} \leq \hat{u}_{2n} \leq \dots \leq \hat{u}_{nn}$  denote the order statistics of the residual, and

$$s^2 = (n - k)^{-1} \sum_{i=1}^n \hat{u}_{in}^2, \quad \hat{\sigma}^2 = n^{-1} \sum_{i=1}^n \hat{u}_{in}^2. \quad (23.15)$$

Here we focus on two tests: the Kolmogorov–Smirnov (KS) test (Kolmogorov, 1933; Smirnov, 1939), and the Jarque and Bera (1980, 1987; henceforth JB) test.

The KS test is based on a measure of discrepancy between the empirical and hypothesized distributions:

$$KS = \max(D^+, D^-), \quad (23.16)$$

where  $D^+ = \max_{1 \leq i \leq n} [(i/n) - \hat{z}_i]$  and  $D^- = \max_{1 \leq i \leq n} [\hat{z}_i - (i - 1)/n]$ ,  $\hat{z}_i = \Phi(\hat{u}_{in}/s)$ ,  $i = 1, \dots, n$ , and  $\Phi(\cdot)$  denotes the cumulative  $N(0, 1)$  distribution function. The exact and limiting distributions of the KS statistic are non-standard and even asymptotic critical points must be estimated. We have used significance points from D'Agostino and Stephens (1986, Table 4.7), although these were formally derived for the location-scale model. The JB test combines the skewness ( $Sk$ ) and kurtosis ( $Ku$ ) coefficients:

$$JB = n \left[ \frac{1}{6}(Sk)^2 + \frac{1}{24}(Ku - 3)^2 \right], \quad (23.17)$$

**Table 23.2** Kolmogorov–Smirnov/Jarque–Bera residuals based tests: empirical type I errors

$k_1$		$n = 25$		$n = 50$		$n = 100$	
		KS	JB	KS	JB	KS	JB
0	<i>STD</i>	0.050	0.029	0.055	0.039	0.055	0.041
	<i>MC</i>	0.052	0.052	0.052	0.050	0.047	0.048
2, ( $n = 25$ )	<i>STD</i>	0.114	0.048	0.163	0.064	0.131	0.131
4, ( $n > 25$ )	<i>MC</i>	0.053	0.052	0.050	0.050	0.050	0.050
$k - 1$ ( $n \leq 50$ )	<i>STD</i>	0.282	0.067	0.301	0.084	0.322	0.322
8, ( $n = 100$ )	<i>MC</i>	0.052	0.048	0.050	0.047	0.047	0.047

*STD* refers to the standard normality test and *MC* denotes the (corresponding) Monte Carlo test.

where  $Sk = n^{-1} \sum_{i=1}^n \hat{u}_{in}^3 / (\hat{\sigma}^2)^{3/2}$  and  $Ku = n^{-1} \sum_{i=1}^n \hat{u}_{in}^4 / (\hat{\sigma}^2)^2$ . Under the null and appropriate regularity conditions, the JB statistic is asymptotically distributed as  $\chi^2(2)$ ; the statistic's exact distribution is intractable.

We next summarize relevant results from the simulation experiment reported in Dufour *et al.* (1998). The experiment based on (23.12) was performed as follows. For each disturbance distribution, the tests were applied to the residual vector, obtained as  $\hat{u} = M_X u$ . Hence, there was no need to specify the coefficients vector  $\beta$ . The matrix  $X$  included a constant term,  $k_1$  dummy variables, and a set of independent standard normal variates. Table 23.2 reports rejection percentages (from 10,000 replications) at the nominal size of 5 percent under the null hypothesis, with  $n = 25, 50, 100$ ,  $k =$  the largest integer less than or equal to  $\sqrt{n}$  and  $k_1 = 0, 2, 4, \dots, k - 1$ . Our conclusions may be summarized as follows. Although the tests appear adequate when the explanatory variables are generated as standard normal, the sizes of all tests vary substantially from the nominal 5 percent for all other designs, irrespective of the sample size. More specifically, (i) the KS test consistently overrejects, and (ii) the JB test based on  $\hat{\sigma}$  underrejects when the number of dummy variables relative to normal regressors is small and overreject otherwise. We will discuss the MC tests results in Section 4.

### 2.3 Uniform linear hypothesis in multivariate regression models

Multivariate linear regression (MLR) models involve a set of  $p$  regression equations with cross-correlated errors. When regressors may differ across equations, the model is known as the seemingly unrelated regression model (SUR or SURE; Zellner, 1962). The MLR model can be expressed as follows:

$$\mathbf{Y} = \mathbf{X}\mathbf{B} + \mathbf{U}, \quad (23.18)$$

where  $\mathbf{Y} = (Y_1, \dots, Y_p)$  is an  $n \times p$  matrix of observation on  $p$  dependent variables,  $X$  is an  $n \times k$  full-column rank matrix of fixed regressors,  $B = [\beta_1, \dots, \beta_p]$  is a  $k \times p$  matrix of unknown coefficients and  $U = [U_1, \dots, U_p] = [\tilde{U}_1, \dots, \tilde{U}_n]'$  is an  $n \times p$  matrix of random disturbances with covariance matrix  $\Sigma$  where  $\det(\Sigma) \neq 0$ . To derive the distribution of the relevant test statistics, we also assume the following:

$$\tilde{U}_i = JW_i, \quad i = 1, \dots, n, \quad (23.19)$$

where the vector  $\mathbf{w} = \text{vec}(W_1, \dots, W_n)$  has a known distribution and  $J$  is an unknown, nonsingular matrix; for further reference, let  $W = [W_1, \dots, W_n]' = UG'$ , where  $G = J^{-1}$ . In particular, this condition will be satisfied when the normality assumption is imposed. An alternative representation of the model is

$$Y_{ij} = \alpha_j + \sum_{k=1}^p \beta_{jk} X_{ik}, \quad i = 1, \dots, n, \quad j = 1, \dots, p. \quad (23.20)$$

Uniform linear (UL) constraints take the special form

$$H_0 : RBC = D, \quad (23.21)$$

where  $R$  is a known  $r \times k$  matrix of rank  $r \leq k$ ,  $C$  is a known  $p \times c$  matrix of rank  $c \leq p$ , and  $D$  is a known  $r \times c$  matrix. An example is the case where the same hypothesis is tested for all equations

$$H_{01} : R\beta_i = \delta_i, \quad i = 1, \dots, p, \quad (23.22)$$

which corresponds to  $C = I_p$ . Here we shall focus on hypotheses of the form (23.22) for ease of exposition; see Dufour and Khalaf (1998c) for the general case.

Stewart (1997) discusses several econometric applications where the problem can be stated in terms of UL hypotheses. A prominent example includes the multivariate test of the capital asset pricing model (CAPM). Let  $r_{jt}$ ,  $j = 1, \dots, p$ , be security returns for period  $t$ ,  $t = 1, \dots, T$ . If it is assumed that a riskless asset  $r_F$  exists, then efficiency can be tested based on the following MLR-based CAPM model:

$$r_{jt} - r_{Ft} = \alpha_j + \beta_j (r_{Mt} - r_{Ft}) + \varepsilon_{jt}, \quad j = 1, \dots, p, \quad t = 1, \dots, T,$$

where  $r_{Mt}$  are the returns on the market benchmark. The hypothesis of efficiency implies that the intercepts  $\alpha_j$  are jointly equal to zero. The latter hypothesis is a special case of (23.22) where  $R$  is the  $1 \times p$  vector  $(1, 0, \dots, 0)$ . Another example concerns demand analysis. It can be shown (see, e.g., Berndt, 1991, ch. 9) that the translog demand specification yields a model of the form (23.20) where the hypothesis of linear homogeneity corresponds to

$$H_0 : \sum_{k=1}^p \beta_{jk} = 0, \quad j = 1, \dots, p. \quad (23.23)$$

**Table 23.3** Empirical type I errors of multivariate tests: uniform linear hypotheses

Sample size	5 equations			7 equations			8 equations		
	LR	LR <sub>c</sub>	LR <sub>MC</sub>	LR	LR <sub>c</sub>	LR <sub>MC</sub>	LR	LR <sub>c</sub>	LR <sub>MC</sub>
20	0.295	0.100	0.051	0.599	0.250	0.047	0.760	0.404	0.046
25	0.174	0.075	0.049	0.384	0.145	0.036	0.492	0.190	0.042
40	0.130	0.066	0.056	0.191	0.068	0.051	0.230	0.087	0.051
50	0.097	0.058	0.055	0.138	0.066	0.050	0.191	0.073	0.053
100	0.070	0.052	0.042	0.078	0.051	0.041	0.096	0.052	0.049

$LR$ ,  $LR_c$ ,  $LR_{MC}$  denote (respectively) the standard  $LR$  test, the Bartlett corrected test and the (corresponding)  $MC$  test.

In this context, the likelihood ratio (LR) criterion is:

$$LR = n \ln(\Lambda), \quad \Lambda = |\hat{U}'_0 \hat{U}_0| / |\hat{U}' \hat{U}|, \quad (23.24)$$

where  $\hat{U}'_0 \hat{U}_0$  and  $\hat{U}' \hat{U}$  are respectively the constrained and unconstrained SSE (sum of square error) matrices. On observing that, under the null hypothesis,

$$\hat{U}' \hat{U} = G^{-1} W' M W (G^{-1})', \quad (23.25)$$

$$\hat{U}'_0 \hat{U}_0 = G^{-1} W' M_0 W (G^{-1})', \quad (23.26)$$

where  $M_0 = I - X(X'X)^{-1}(X'X - R'(R(X'X)^{-1}R')^{-1}R)(X'X)^{-1}X'$  and  $M = I - X(X'X)^{-1}X'$ , we can then rewrite  $\Lambda$  in the form

$$\Lambda = |W' M_0 W| / |W' M W|, \quad (23.27)$$

where the matrix  $W = UG'$  has a distribution which does not involve nuisance parameters. As shown in Dufour and Khalaf (1998c), decomposition (23.27) obtains only in the case of UL constraints. In Section 4 we will exploit the latter result to obtain exact MC tests based on the LR statistic.

To illustrate the performance of the various relevant tests, we consider a simulation experiment modeled after demand homogeneity tests, i.e. (23.20) and (23.23) with  $p = 5, 7, 8$ ,  $n = 20, 25, 40, 50, 100$ . The regressors are independently drawn from the normal distribution; the errors are independently generated as iid  $N(0, \Sigma)$  with  $\Sigma = GG'$  and the elements of  $G$  drawn (once) from a normal distribution. The coefficients for all experiments are available from Dufour and Khalaf (1998c). The statistics examined are the relevant LR criteria defined by (23.24) and the Bartlett-corrected LR test (Attfield, 1995, section 3.3). The results are summarized in Table 23.3. We report the tests' empirical size, based on a nominal size of 5 percent and 1,000 replications. It is evident that the asymptotic LR test overrejects

substantially. Second, the Bartlett correction, though providing some improvement, fails in larger systems. In this regard, it is worth noting that Attfield (1995, section 3.3) had conducted a similar Monte Carlo study to demonstrate the effectiveness of Bartlett adjustments in this framework, however the example analyzed was restricted to a two-equations model. We will discuss the MC test results in Section 4.

To conclude this section, it is worth noting that an exact test is available for hypotheses of the form  $H_0 : RBC = D$ , where  $\min(r, c) \leq 2$ . Indeed, Laitinen (1978) in the context of the tests of demand homogeneity and Gibbons, Ross, and Shanken (1989), for the problem of testing the CAPM efficiency hypothesis, independently show that a transformation of the relevant LR criterion has an exact F-distribution given normality of asset returns.<sup>2</sup>

## 2.4 Econometric applications: discussion

In many empirical problems, it is quite possible that the exact null distribution of the relevant test statistic  $S(Y)$  will not be easy to compute analytically, even though it is nuisance-parameter-free. In this case,  $S(Y)$  is called a pivotal statistic, i.e. the null distribution of  $S(Y)$  is uniquely determined under the null hypothesis. In such cases, we will show that the MC test easily solves the size control problem, regardless of the distributional complexities involved. The above examples on normality tests and the UL hypotheses tests, all involve pivotal statistics. The problem is more complicated in the presence of nuisance parameters. We will first discuss a property related to nuisance-parameter-dependent test statistics which will prove to be fundamental in finite sample contexts.<sup>3</sup>

In the context of a right-tailed test problem, consider a statistic  $S(Y)$  whose null distribution depends on nuisance parameters and suppose it is possible to find another statistic  $S^*(Y)$  such that

$$S(Y) \leq S^*(Y), \quad \forall \theta \in \Theta_0, \tag{23.28}$$

and  $S^*(Y)$  is pivotal under the null. Then  $S(Y)$  is said to be *boundedly pivotal*. The implications of this property are as follows. From (23.28), we obtain

$$P_\theta[S(Y) \geq c] \leq P[S^*(Y) \geq c], \quad \forall \theta \in \Theta_0.$$

Then if we calculate  $c$  such that

$$P[S^*(Y) \geq c] = \alpha, \tag{23.29}$$

we solve the level constraint for the test based on  $S(Y)$ . It is clear that (23.28) and (23.29) imply

$$P_\theta[S(Y) \geq c] \leq \alpha, \quad \forall \theta \in \Theta_0.$$

As emphasized earlier, the size control constraint is easier to deal with in the case of  $S^*(Y)$  because it is pivotal. Consequently, the maximization problem

$$\sup_{\theta \in \Theta_0} P_\theta[S(Y) \geq c]$$

has a non-trivial solution (less than 1) in the case of *boundedly pivotal statistics*. If this property fails to hold, the latter optimization problem may admit only the trivial solution, so that it becomes mathematically impossible to control the significance level of the test.

It is tempting to dismiss such considerations assuming they will occur only in “textbook” cases. Yet it can be shown (we will consider this issue in the next section) that similar considerations explain the poor performance of the Wald tests and confidence intervals discussed in Sections 2.1 and 2.3 above. *These are problems of empirical relevance in econometric practice.* In the next session, we will show that the bootstrap will also fail for such problems! For further discussion of the basic notions of statistical testing mentioned in this section, the reader may consult Lehmann (1986, ch. 3), Gouriéroux and Monfort (1995), and Dufour (1990, 1997).

### 3 THE MONTE CARLO TEST TECHNIQUE: AN EXACT RANDOMIZED TEST PROCEDURE

*If there were a machine that could check 10 permutations a second, the job would run something on the order of 1,000 years. The point is, then, that an impossible test can be made possible, if not always practical.*

Dwass (1957)

The Monte Carlo (MC) test procedure was first proposed by Dwass (1957) in the following context. Consider two independent samples  $X_1, \dots, X_m$  and  $Y_1, \dots, Y_n$ , where  $X_1, \dots, X_m \stackrel{\text{iid}}{\sim} F(x)$ ,  $Y_1, \dots, Y_n \stackrel{\text{iid}}{\sim} F(x - \delta)$  and the cdf  $F(\cdot)$  is continuous. No further distributional assumptions are imposed. To test  $H_0 : \delta = 0$ , the following procedure may be applied.

- Let  $z = (X_1, \dots, X_m, Y_1, \dots, Y_n)$  and  $s = \frac{1}{m} \sum_{i=1}^m X_i - \frac{1}{n} \sum_{i=1}^n Y_i$ .
- Obtain all possible  $Q = (n + m)!$  permutations of  $z$ ,  $z^{(1)}, \dots, z^{(Q)}$ , and calculate the associated “permuted analogs” of  $s$

$$s^{(j)} = \frac{1}{m} \sum_{i=1}^m z_i^{(j)} - \frac{1}{n} \sum_{i=m+1}^{m+n} z_i^{(j)}, \quad j = 1, \dots, Q.$$

- Let  $r$  denote the number of  $s^{(j)}$ s for which  $s \leq s^{(j)}$ . Reject the null (e.g. against  $H_A : \delta > 0$ ) if  $r \leq k$ , where  $k$  is a predetermined integer.

It is easy to see that  $P(r \leq k) = k/Q$  under the null because the  $X$ s and the  $Y$ s are exchangeable. In other words, the test just described is exactly of size  $k/Q$ .

The procedure is intuitively appealing, yet there are  $(n + m)!$  permutations to examine. To circumvent this problem, Dwass (1957) proposed to apply the same principle to a sample of  $P$  permutations  $\tilde{s}^{(1)}, \dots, \tilde{s}^{(P)}$ , in a way that will preserve the size of the test. The modified test may be applied as follows.

- Let  $\tilde{r}$  denote the number of  $\tilde{s}^{(j)}$ s for which  $s \leq \tilde{s}^{(j)}$ . Reject the null (against  $\delta > 0$ ) if  $\tilde{r} \leq d$ , where  $d$  is chosen such that

$$\frac{d+1}{P+1} = \frac{k}{Q}.$$

Dwass formally shows that with this choice for  $d$ , the size of the modified test is exactly  $k/Q$  = the size of the test based on all permutations. This means that, if we wish to get a 5 percent-level permutation test, and 99 random permutations can be generated, then  $d + 1$  should be set to 5. The latter decision rule may be restated as follows: reject the null if the rank of  $s$  in the series  $s, \tilde{s}^{(1)}, \dots, \tilde{s}^{(P)}$  is less than or equal to 5. Since each  $\tilde{s}^{(j)}$  is “weighted” by the probability that it is sampled from all possible permutations, the modification due to Dwass yields a randomized test procedure.

The principles underlying the MC test procedure are highly related to the randomized permutation test just described. Indeed, this technique is based on the above test strategy where the sample of permutations is replaced by *simulated samples*. Note that Barnard (1963) later proposed a similar idea.<sup>4</sup>

### 3.1 Monte Carlo tests based on pivotal statistics

In the following, we briefly outline the MC test methodology as it applies to the pivotal statistic context and a right-tailed test; for a more detailed discussion, see Dufour (1995) and Dufour and Kiviet (1998).

Let  $S_0$  denote the observed test statistic  $S$ , where  $S$  is the test criterion. We assume  $S$  has a unique continuous distribution under the null hypothesis ( $S$  is a *continuous pivotal statistic*). Suppose we can generate  $N$  iid replications,  $S_j$ ,  $j = 1, \dots, N$ , of this test statistic under the null hypothesis. Compute

$$\hat{G}_N(S_0) = \frac{1}{N} \sum_{j=1}^N I_{[0, \infty]}(S_j - S_0), \quad I_A(z) = \begin{cases} 1, & \text{if } z \in A \\ 0, & \text{if } z \notin A \end{cases}.$$

In other words,  $N\hat{G}_N(S_0)$  is the number of simulated statistics which are greater or equal to  $S_0$ , and provided none of the simulated values  $S_j$ ,  $j = 1, \dots, N$ , is equal to  $S_0$ ,  $\hat{R}_N(S_0) = N - N\hat{G}_N(S_0) + 1$  gives the rank of  $S_0$  among the variables  $S_0, S_1, \dots, S_N$ .<sup>5</sup> Then the critical region of a test with level  $\alpha$  is:

$$\hat{p}_N(S_0) \leq \alpha, \tag{23.30}$$

where  $0 < \alpha < 1$  and

$$\hat{p}_N(x) = \frac{N\hat{G}_N(x) + 1}{N + 1}. \tag{23.31}$$

The latter expression gives the *empirical probability* that a value as extreme or more extreme than  $S_0$  is realized if the null is true. Hence  $\hat{p}_N(S_0)$  may be viewed as a MC  $p$ -value.

Note that the MC decision rule may also be expressed in terms of  $\hat{R}_N(S_0)$ . Indeed the critical region

$$\frac{N\hat{G}_N(S_0) + 1}{N + 1} \leq \alpha$$

is equivalent to

$$\hat{R}_N(S_0) \geq (N + 1)(1 - \alpha) + 1. \quad (23.32)$$

In other words, for 99 replications a 5 percent MC test is significant if the rank of  $S_0$  in the series  $S_0, S_1, \dots, S_N$  is at least 96, or informally, if  $S_0$  lies in the series top 5 percent percentile. We are now faced with the immediate question: does the MC test just defined achieve size control?

If the null distribution of  $S$  is nuisance-parameter-free and  $\alpha(N + 1)$  is an integer, the critical region (23.30) is provably exact, in the sense that

$$P_{(H_0)}[\hat{p}_N(S_0) \leq \alpha] = \alpha$$

or alternatively

$$P_{(H_0)}[\hat{R}_N(S_0) \geq (N + 1)(1 - \alpha) + 1] = \alpha.$$

The proof is based on the following theorem concerning the distribution of the ranks associated with a finite dimensional array of exchangeable variables; see Dufour (1995) for a more formal statement of the theorem and related references.

**Theorem 23.1** Consider an  $M \times 1$  vector of exchangeable real random variables  $(Y_1, \dots, Y_M)$  such that  $P[Y_i = Y_j] = 0$  for  $i \neq j$ , and let  $R_j$  denote the rank of  $Y_j$  in the series  $Y_1, \dots, Y_M$ . Then

$$P\left[\frac{R_j}{M} \geq z\right] = \frac{I[(1 - z)M] + 1}{M}, \quad 0 < z \leq 1. \quad (23.33)$$

where  $I(x)$  is the largest integer less than or equal to  $x$ .

If  $S$  is a continuous pivotal statistic, it follows from the latter result that

$$P_{(H_0)}[\hat{R}_N(S_0) \geq (N + 1)(1 - \alpha) + 1].$$

Indeed, in this case, the observed test statistic and the simulated statistic are exchangeable if the null is true. Here it is worth recalling that the  $S_j$ s must be simulated imposing the null. Now using (23.33), it is easy to show that  $P_{(H_0)}[\hat{R}_N(S_0) \geq (N + 1)(1 - \alpha) + 1] = \alpha$ , provided  $N$  is chosen so that  $\alpha(N + 1)$  is an integer.

We emphasize that the sample size and the number of replications are explicitly taken into consideration in the above arguments. No asymptotic theory has been used so far to justify the procedure just described.

It will be useful at this stage to focus on a simple illustrative example. Consider the Jarque and Bera normality test statistic,

$$JB = n \left[ \frac{1}{6} (Sk)^2 + \frac{1}{24} (Ku - 3)^2 \right],$$

in the context of the linear regression model  $Y = X\beta + u$ .<sup>6</sup> The MC test based on JB and  $N$  replications may be obtained as follows.

- Calculate the constrained OLS estimates  $\hat{\beta}$ ,  $s$  and the associated residual  $\hat{u}$ .
- Obtain the Jarque–Bera statistic based on  $s$  and  $\hat{u}$  and denote it  $JB^{(0)}$ .
- Treating  $s$  as fixed, repeat the following steps for  $j = 1, \dots, N$ :
  - (a) draw an  $(n \times 1)$  vector  $\tilde{u}^{(j)}$  as iid  $N(0, s^2)$ ;
  - (b) obtain the simulated independent variable  $\tilde{Y}^{(j)} = X\hat{\beta} + \tilde{u}^{(j)}$ ;
  - (c) regress  $\tilde{Y}^{(j)}$  on  $X$ ;
  - (d) derive the Jarque–Bera statistic  $JB^{(j)}$  associated with the regression of  $\tilde{Y}^{(j)}$  on  $X$ .
- Obtain the rank  $\hat{R}_N(JB^{(0)})$  in the series  $JB^{(0)}, JB^{(1)}, \dots, JB^{(N)}$ .
- Reject the null if  $\hat{R}_N(JB^{(0)}) \geq (N + 1)(1 - \alpha) + 1$ .

Furthermore, an MC  $p$ -value may be obtained as  $\hat{p}_N(S_0) = [N + 1 - \hat{R}_N(S_0)]/(N + 1)$ . Dufour *et al.* (1998) show that the JB statistics can be computed from the standardized residual vector  $\hat{u}/s$ . Using (23.14), we see that

$$\hat{u}/s = \frac{\hat{u}}{(\hat{u}'\hat{u}/(n-k))^{1/2}} = (n-k)^{1/2} \frac{M_X u}{(u' M_X u)^{1/2}} = (n-k)^{1/2} \frac{M_X w}{(w' M_X w)^{1/2}}, \quad (23.34)$$

where  $w = u/\sigma \stackrel{\text{iid}}{\sim} N(0, 1)$  when  $u \sim N(0, \sigma^2 I_n)$ . It follows that the simulated statistics  $JB^{(j)}$  may be obtained using draws from a nuisance-parameter-free (standard normal) null distribution.

### 3.2 Monte Carlo tests in the presence of nuisance parameters

In Dufour (1995), we discuss extensions of MC tests when nuisance parameters are present. We now briefly outline the underlying methodology. In this section,  $n$  refers to the sample size and  $N$  the number of MC replications.

Consider a test statistic  $S$  for a null hypothesis  $H_0$ , and suppose the null distribution of  $S$  depends on an unknown parameter vector  $\theta$ .

- From the observed data, compute:
  - (a) the test statistic  $S_0$ , and
  - (b) a restricted consistent estimator  $\hat{\theta}_n^0$  of  $\theta$ .

- Using  $\hat{\theta}_n^0$ , generate  $N$  simulated samples and, from them,  $N$  simulated values of the test statistic. Then compute  $\hat{p}_N(S_0 | \hat{\theta}_n^0)$ , where  $\hat{p}_N(x | \bar{\theta})$  refers to  $\hat{p}_N(x)$  based on realizations of  $S$  generated given  $\theta = \bar{\theta}$  and  $\hat{p}_N(x)$  is defined in (23.31).
- An MC test may be based on the critical region

$$\hat{p}_N(S_0 | \hat{\theta}_n^0) \leq \alpha, \quad \alpha \leq 0 \leq 1.$$

For further reference, we denote the latter procedure a *local Monte Carlo* (LMC) test. Under general conditions, this LMC test has the correct level asymptotically (as  $n \rightarrow \infty$ ), i.e. under  $H_0$ ,

$$\lim_{n \rightarrow \infty} \{P[\hat{p}_N(S_0 | \hat{\theta}_n^0) \leq \alpha] - P[\hat{p}_N(S_0 | \theta) \leq \alpha]\} = 0. \quad (23.35)$$

In particular, these conditions are usually met whenever the test criterion involved is asymptotically pivotal. We emphasize that no asymptotics on the number of replications is required to obtain (23.35).

- To obtain an exact critical region, the MC  $p$ -value ought to be maximized with respect to the intervening parameters. Specifically, in Dufour (1995), it is shown that the test (henceforth called a *maximized Monte Carlo* (MMC) test) based on the critical region

$$\sup_{\theta \in M_0} [\hat{p}_N(S_0 | \theta)] \leq \alpha \quad (23.36)$$

where  $M_0$  is the subset of the parameter space compatible with the null hypothesis (i.e. the nuisance parameter space) is exact at level  $\alpha$ .

The LMC test procedure is closely related to a parametric bootstrap, with however a fundamental difference. Whereas bootstrap tests are valid as  $N \rightarrow \infty$ , the number of simulated samples used in MC tests is explicitly taken into account. Further the LMC  $p$ -value may be viewed as exact in a *liberal* sense, i.e. if the LMC fails to reject, we can be sure that the exact test involving the maximum  $p$ -value is not significant at level  $\alpha$ .

In practical applications of exact MMC tests, a global optimization procedure is needed to obtain the maximal randomized  $p$ -value in (23.36). We use the simulated annealing (SA) algorithm (Corana, Marchesi, Martini, and Ridella, 1987; Goffe, Ferrier, and Rogers, 1994). SA starts from an initial point (it is natural to use  $\hat{\theta}_n^0$  here) and sweeps the parameter space (user defined) at random. An *uphill* step is always accepted while a *downhill* step may be accepted; the decision is made using the Metropolis criterion. The direction of all moves is determined by probabilistic criteria. As it progresses, SA constantly adjusts the step length so that *downhill* moves are less and less likely to be accepted. In this manner, the algorithm escapes local optima and gradually converges towards the most probable area for optimizing. SA is robust with respect to nonquadratic and

even noncontinuous surfaces and typically escapes local optima. The procedure is known not to depend on starting values. Most importantly, SA readily handles problems involving a fairly large number of parameters.<sup>7</sup>

To conclude this section, we consider another application of MC tests which is useful in the context of boundedly pivotal statistics. Using the above notation, the statistic at hand  $S$  is boundedly pivotal if it is possible to find another statistic  $S^*$  such that

$$S \leq S^*, \quad \forall \theta \in \Theta_0, \tag{23.37}$$

and  $S^*$  is pivotal under the null. Let  $c$  and  $c^*$  refer to the  $\alpha$  size-correct cut-off points associated with  $S$  and  $S^*$ . As emphasized earlier, inequality (23.37) entails that  $c^*$  may be used to define a critical region for  $S$ . The resulting test will have the correct level and may be viewed as *conservative* in the following sense: if the test based on  $c^*$  is significant, we can be sure that the exact test involving the (unknown!)  $c$  is significant at level  $\alpha$ . The main point here is that it is easier to calculate  $c^*$ , because  $S^*$  is pivotal, whereas  $S$  is nuisance-parameter dependent. Of course, this presumes that the null exact distribution of  $S^*$  is known and tractable; see Dufour (1989, 1990) for the underlying theory and several illustrative examples. Here we argue that the MC test technique may be used to produce simulation-based conservative  $p$ -values based on  $S^*$  even if the analytic null distribution of  $S^*$  is unknown or complicated (but may be simulated). The procedure involved is the same as above, except that the  $S^*$  rather than  $S$  is evaluated from the simulated samples. We denote the latter procedure a bound MC (BMC) test.

A sound test strategy would be to perform the bounds tests first and, on failure to reject, to apply randomized tests. We recommend the following computationally attractive exact  $\alpha$  test procedure:

1. compute the test statistic from the data;
2. if a bounding criterion is available, compute a BMC  $p$ -value; reject the null if: BMC  $p$ -value  $\leq \alpha$ ;
3. if the observed value of the test statistic falls in the BMC acceptance region, obtain a LMC  $p$ -value; declare the test not significant if: LMC  $p$ -value  $> \alpha$ ;
4. if the LMC  $p$ -value  $\leq \alpha <$  BMC  $p$ -value, obtain the MMC  $p$ -value and reject the null if the latter is less than or equal to  $\alpha$ .

## 4 MONTE CARLO TESTS: ECONOMETRIC APPLICATIONS

### 4.1 Pivotal statistics

In Dufour and Kiviet (1996, 1998), Kiviet and Dufour (1997), Dufour *et al.* (1998), Dufour and Khalaf (1998a, 1998c), Bernard, Dufour, Khalaf, and Genest (1998), Saphores, Khalaf, and Pelletier (1998), several applications of MC tests based on pivotal statistics are presented. The problems considered include: normality tests, heteroskedasticity tests including tests for (G)ARCH and tests for break in variance at unknown points, independence tests and tests based on autocorrelations.<sup>8</sup>

The reader will find in the above papers simulation results which show clearly that the technique of Monte Carlo tests completely corrects often important size distortions due to poor large sample approximations; power studies are also reported on a case by case basis to assess the performance of MC size corrected tests.

Relevant results pertaining to the examples considered above are included in Tables 23.2 and 23.3. It is evident from Table 23.2 that the size of the JB and KS tests is perfectly controlled for all designs considered.<sup>9</sup> Table 23.3 includes the empirical size of the MC LR test for linear restrictions. From (23.27), we see that under the distributional assumption (23.19), the simulated statistics may be obtained using draws from a nuisance-parameter free null distribution, namely the hypothesized distribution of the vector  $w$ . Consequently, application of the MC test procedure yields exact  $p$ -values. Indeed, it is shown in Table 23.3 that the MC LR test achieves perfect size control.<sup>10</sup>

Now to illustrate the feasibility of MMC tests and the usefulness of BMC tests, we will focus on examples involving nuisance parameters.

## 4.2 Monte Carlo tests in the presence of nuisance parameters: examples from the multivariate regression model

In this section, we provide examples from Dufour and Khalaf (1998a, 1998b) pertaining to LR test criteria in the MLR (reduced form) model. The model was introduced in Section 2.3. Consider the three equations system

$$\begin{aligned} Y_1 &= \beta_{10} + \beta_{11}X_1 + U_1, \\ Y_2 &= \beta_{20} + \beta_{22}X_2 + U_2, \\ Y_3 &= \beta_{30} + \beta_{33}X_3 + U_3, \end{aligned} \tag{23.38}$$

imposing normality, and the hypothesis  $H_0 : \beta_{11} = \beta_{22} = \beta_{33}$ . First restate  $H_0$  in terms of the MLR model which includes the SURE system as a special case, so that it incorporates the SURE exclusion restrictions. Formally, in the framework of the MLR model

$$\begin{aligned} Y_1 &= \beta_{10} + \beta_{11}X_1 + \beta_{12}X_2 + \beta_{13}X_3 + U_1, \\ Y_2 &= \beta_{20} + \beta_{21}X_1 + \beta_{22}X_2 + \beta_{23}X_3 + U_2, \\ Y_3 &= \beta_{30} + \beta_{31}X_1 + \beta_{32}X_2 + \beta_{33}X_3 + U_3, \end{aligned} \tag{23.39}$$

$H_0$  is equivalent to the joint hypothesis

$$H_0^* : \beta_{11} = \beta_{22} = \beta_{33} \text{ and } \beta_{12} = \beta_{13} = \beta_{21} = \beta_{23} = \beta_{31} = \beta_{32} = 0. \tag{23.40}$$

The associated LR statistic is

$$LR = n \ln(\Lambda), \quad \Lambda = |\hat{\Sigma}_0| / |\hat{\Sigma}|, \quad (23.41)$$

where  $\hat{\Sigma}_0$  and  $\hat{\Sigma}$  are the restricted and unrestricted SURE MLE. We also consider

$$LR^* = n \ln(\Lambda^*), \quad \Lambda^* = |\hat{\Sigma}_0| / |\hat{\Sigma}_u|, \quad (23.42)$$

where  $\hat{\Sigma}_u$  is the unconstrained estimate of  $\Sigma$  in the “nesting” MLR model. Since the restricted model is the same in both LR and  $LR^*$ , while the unrestricted model used in  $LR^*$  includes as a special case the one in LR, it is straightforward to see that  $LR \leq LR^*$ , so that the distribution function of  $LR^*$  provides an upper bound on the distribution function of LR.

In order to apply a BMC test, we need to construct a set of UL restrictions that satisfy (23.40) so that the corresponding LRc criterion conforming with these UL restrictions yields a valid bound on the distribution of LR. Indeed, as emphasized above, the LR test statistic for UL restrictions is pivotal. Furthermore, by considering UL restrictions obtained as a special case of  $H_0^*$ , we can be sure that the associated statistic is always  $\geq LR$ . Here, it is easy to see that the constraints setting the coefficients  $\beta_{ij}$ ,  $i, j = 1, \dots, 3$ , to specific values meet this criterion. Note that the statistic just derived serves to bound both LR and  $LR^*$ .

Define  $\theta \equiv C(\Sigma)$  as the vector of the parameters on or below the diagonal of the Cholesky factor  $T(\Sigma)$  of the covariance matrix  $\Sigma$  (i.e.  $T(\Sigma)$  is the lower triangular matrix such that  $T(\Sigma)T(\Sigma)' = \Sigma$ ). The algorithm for performing MC tests based on  $LR^*$ , at the 5 percent level with 99 replications, can be described as follows.

- Compute  $\hat{\Sigma}_0$  and  $\hat{\Sigma}$ , the restricted and unrestricted SURE (iterative) MLE.
- Compute  $\hat{\Sigma}_u$  as the unconstrained (OLS) estimate of  $\Sigma$  in the “nesting” MLR model.
- Compute  $\Lambda^* = |\hat{\Sigma}_0| / |\hat{\Sigma}_u|$  and  $LR^* = n \ln(\Lambda^*)$ .
- Draw 99 realizations from a multivariate  $(n, 3, I)$  normal distribution:  $U^{(1)}, U^{(2)}, \dots, U^{(p)}$  and store.
- Consider the linear constraints

$$H_{02} : \begin{bmatrix} 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} \beta_{10} & \beta_{20} & \beta_{30} \\ \beta_{11} & \beta_{21} & \beta_{31} \\ \beta_{12} & \beta_{22} & \beta_{32} \\ \beta_{13} & \beta_{23} & \beta_{33} \end{bmatrix} = \begin{bmatrix} \hat{\beta}_{11} & 0 & 0 \\ 0 & \hat{\beta}_{22} & 0 \\ 0 & 0 & \hat{\beta}_{33} \end{bmatrix}$$

where  $\hat{\beta}_{11} = \hat{\beta}_{22} = \hat{\beta}_{33}$  are the constrained SURE estimates calculated in the first step.

- Call the bound MC procedure  $BMC(\theta)$ , described below, for  $\theta \equiv C(\hat{\Sigma}_0)$ . The Cholesky decomposition is used to impose positive definiteness and avoid redundant parameters. The output is the BMC  $p$ -value. Reject the null if the latter is  $\leq .05$  and STOP.

- Otherwise, call the procedure  $MC(\theta)$ , also described below, for  $\theta \equiv C(\hat{\Sigma}_0)$ . It is important to note here that  $\Sigma$  is the only relevant nuisance parameter, for the example considered involves linear constraints (see Breusch, 1980). The output is the LMC  $p$ -value. Declare the test not significant if the latter exceeds .05 and STOP.
- Otherwise, call the maximization algorithm (for example, SA) for the function  $MC(\theta)$  using  $\theta \equiv C(\hat{\Sigma}_0)$  as a starting value. Obtain the MMC  $p$ -value and reject the null if the latter is  $\leq .05$ . Note: if only a decision is required, the maximization algorithm may be instructed to exit as soon as a value larger than .05 is attained. This may save considerable computation time.

Description of the procedure  $MC(\theta)$ :

- Construct a triangular  $\Omega$  from  $\theta$  (this gives the Cholesky decomposition of the variance which will be used to generate the simulated model).
- Do for  $j = 1, \dots, N$  (independently)
  - (a) Generate the random vectors  $Y_1^{(j)} Y_2^{(j)} Y_3^{(j)}$  conformably with the nesting MLR model, using the restricted SURE coefficient estimates,  $U^{(j)}$ , the observed regressors, and  $\Omega$ .
  - (b) Estimate the MLR model with the observed regressors as dependent variable, and  $Y_1^{(j)} Y_2^{(j)} Y_3^{(j)}$  as independent variables: obtain the unrestricted estimates and the estimates imposing  $H_0$ .
  - (c) From these estimates, form the statistics  $LR^{(j)}$  and store.
- Obtain the rank of  $LR^*$  in the series  $LR^*, LR^{*(1)}, \dots, LR^{*(99)}$ .
- This yields a MC  $p$ -value as described above which is the output of the procedure.
- The BMC( $\theta$ ) procedure may be obtained as just described, replacing  $LR^{*(j)}$  by  $LR_c^{(j)}$ . Alternatively, the BMC procedure may be rewritten following the methodology relating to MC tests of UL hypotheses so that no (unknown) parameters intervene in the generation of the simulated (bounding) statistics. Indeed, the bounding statistic satisfies (23.27) under (23.19). Thus  $LR_c^{(j)}$  may be obtained using draws from, e.g., the multivariate independent normal distribution.

In Dufour and Khalaf (1998c), we report the results of a simulation experiment designed according to this example. In particular, we examine the performance of LMC and BMC tests. We show that the MC test procedure achieves perfect size control and has good power. The same methodology may also be applied in simultaneous equations models such as (23.3). In Dufour and Khalaf (1998b), we present simulations which illustrate the performance of limited-information LR-based MC tests in this context. We have attempted to apply the MC test procedure to the IV-based Wald-type test for linear restrictions on structural parameters. In this case, the performance of the standard bootstrap was disappointing. The LMC Wald tests failed completely in near-unidentified conditions. Furthermore, in all cases examined, *the Wald tests maximal randomized p-values were always one*. This is a case (refer to Section 2.3) where the MC procedure does not (and cannot) correct the performance of the test.

In other words, Wald statistics do not constitute valid pivotal functions in such models and it is even impossible to bound their distributions over the parameter space (except by the trivial bound 0 and 1). (Dufour, 1997)

These results are also related to the non-invariance problems associated with Wald tests in nonlinear contexts (see, e.g. Dufour, 1997; Dagenais and Dufour, 1991). Indeed, it is evident from (23.3)–(23.5) that *seemingly linear* constraints on structural coefficients in instrumental regressions often involve nonlinear hypotheses implied by the structure. Of course, not all Wald tests will suffer from such problems. For instance, Wald tests for linear restrictions in linear regression models yield exact F-tests.

We conclude this section with a specific problem where the MC test strategy conveniently solves a difficult and non-standard distributional problem: the problem of unidentified nuisance parameters.

### 4.3 Non-identified nuisance parameters

The example we discuss here is the problem of testing for the significance of jumps in the context of a jump-diffusion model. For econometric applications and references, see Saphores *et al.* (1998). Formally, consider the following model written, for convenience, in discrete time:

$$S_t - S_{t-1} = \mu + \sigma \xi_t + \sum_{i=1}^{n_t} \ln(Y_i), \quad t = 1, \dots, T,$$

where  $\xi \stackrel{\text{iid}}{\sim} N(0, 1)$  and  $\ln(Y) \stackrel{\text{iid}}{\sim} N(\theta, \delta^2)$  and  $n_t$  is the number of jumps which occur in the interval  $[t-1, t]$ ; the arrival of jumps is assumed to follow a *Poisson* process with parameter  $\lambda$ . The associated likelihood function is as follows:

$$L_1 = -T \ln(\lambda) - \frac{T}{2} \ln(2\pi) + \sum_{t=1}^T \ln \left[ \sum_{j=0}^{\infty} \frac{\lambda^j}{j!} \frac{1}{\sqrt{\sigma^2 + \delta^2 j}} \exp \left( \frac{-(x_t - \mu - \theta j)^2}{2(\sigma^2 + \delta^2 j)} \right) \right].$$

The hypothesis of no jumps corresponds to  $\lambda = 0$ . It is clear that in this case, the parameters  $\theta, \delta^2$  are not identified under the null, and hence, following the results of Davies (1977, 1987), the distribution of the associated LR statistic is non-standard and quite complicated. Although this problem is well recognized by now, a  $\chi^2(3)$  asymptotic distribution is often (inappropriately) used in empirical applications of the latter LR test. See Diebold and Chen (1996) for related arguments dealing with structural change tests.

Let  $\hat{\mu}, \hat{\sigma}^2$  denote the MLE under the null, i.e. imposing a Geometric Brownian Motion. Here we argue that in this case, the MC  $p$ -value calculated as described above, drawing iid  $N(\hat{\mu}, \hat{\sigma}^2)$  disturbances (with  $\hat{\mu}$  and  $\hat{\sigma}^2$  taken as given) will not depend on  $\theta$  and  $\delta^2$ . This follows immediately from the implications of non-identification. Furthermore, the invariance to location and scale ( $\mu$  and  $\sigma$ ) is straightforward to see. Consequently, the MC test described in the context of pivotal statistics will yield exact  $p$ -values.

The problem of unidentified nuisance parameters is prevalent in econometrics. Bernard *et al.* (1998) consider another illustrative example: testing for ARCH-in-mean effects, and show that the MC method works very well in terms of size and power.

## 5 CONCLUSION

In this chapter, we have demonstrated that finite sample concerns may arise in several empirically pertinent test problems. But, in many cases of interest, the MC test technique produces valid inference procedures no matter how small your sample is.

We have also emphasized that the problem of constructing a good test – although simplified – cannot be solved *just* using simulations. Yet in most examples we have reviewed, MC test techniques emerge as indispensable tools.

Beyond the cases covered above, it is worthwhile noting that the MC test technique may be applied to many other problems of interest. These include, for example, models where the estimators themselves are also simulation-based, e.g. estimators based on indirect inference or involving simulated maximum likelihood. Furthermore, the MC test technique is by no means restricted to nested hypotheses. It is therefore possible to compare nonnested models using MC LR-type tests; assessing the success of this strategy in practical problems is an interesting research avenue.

Of course, the first purpose of the MC test technique is to control the probability of type I errors (below a given *level*) so that rejections can properly be interpreted as showing that the null hypothesis is “incompatible” with the data. However, once level is controlled, we can (and should) devote more attention to finding procedures with good *power* properties. Indeed, by helping to put the problem of level control out of the way, we think the technique of MC tests should help econometricians devote research to power issues as opposed to level. So an indirect consequence of the implementation of the technique may well be an increased emphasis on the design of more powerful tests.

Your data are valuable, and the statistical analysis you perform is often policy oriented. Why tolerate questionable *p*-values and confidence intervals, when exact or improved approximations are available?

## Notes

- \* The authors thank three anonymous referees and the Editor, Badi Baltagi, for several useful comments. This work was supported by the Bank of Canada and by grants from the Canadian Network of Centres of Excellence (program on *Mathematics of Information Technology and Complex Systems* (MITACS)), the Social Sciences and Humanities Research Council of Canada, the Natural Sciences and Engineering Council of Canada, and the Government of Québec (Fonds FCAR).
- 1 The problem is more complicated when the structural equation includes more than one endogenous variable. See Dufour and Khalaf (1998b) for a detailed discussion of this case.
- 2 The underlying distributional result is due to Wilks (1932).

- 3 For a formal treatment see Dufour (1997).
- 4 Bera and Jarque (1982), Breusch and Pagan (1979, 1980) have also proposed related simulation-based techniques. However, these authors do not provide finite-sample theoretical justification for the proposed procedures. In particular, in contrast with Dwass (1957) and Barnard (1963) (and similarly to many other later authors who have proposed exploiting Monte Carlo techniques), they do not observe that appropriately randomized tests allow one to exactly control the level of a test in finite samples.
- 5 The subscript  $N$  in the notation adopted here may be misleading. We emphasize that  $\hat{R}_N(T_0)$  gives the rank of  $S_0$  in the  $N + 1$  dimensional array  $S_0, S_1, \dots, S_N$ . Throughout this section  $N$  refers to the number of MC replications.
- 6 See Section 2.2 for a formal presentation of the model and test statistics. Some equations are redefined here for convenience.
- 7 Global optimization is generally considered to be (relatively) computationally demanding. We have experimented (see Dufour and Khalaf, 1998c, 1998b) with several MMC tests where the number of nuisance parameters referred to the simulated annealing algorithm was up to 20. Our simulations show that the method works well. Convergence was slow in some cases (less than 5 per 1,000). Recall, however, that for the problem at hand, one just practically needs to check whether the maximized function exceeds  $\alpha$ , which clearly reduces the computational burdens.
- 8 In connection, it is worth mentioning that the MC test procedure applied to the Durbin–Watson test for AR(1) disturbances solves the inconclusive region problem.
- 9 See Dufour *et al.* (1998) for the power study.
- 10 See Dufour and Khalaf (1998c) for the power study.

## References

- Anderson, T.W., and H. Rubin (1949). Estimation of the parameters of a single equation in a complete system of stochastic equations. *Annals of Mathematical Statistics* 20, 46–63.
- Attfield, C.L.F. (1995). A Bartlett adjustment to the likelihood ratio test for a system of equations. *Journal of Econometrics* 66, 207–23.
- Barnard, G.A. (1963). Comment on “The Spectral Analysis of Point Processes” by M.S. Bartlett. *Journal of the Royal Statistical Society, Series B* 25, 294.
- Bartlett, M.S. (1948). A note on the statistical estimation of supply and demand relations from time series. *Econometrica* 16, 323–9.
- Bera, A.K., and C.M. Jarque (1982). Model specification tests: A simultaneous approach. *Journal of Econometrics* 20, 59–82.
- Bernard, J.-T., J.-M. Dufour, L. Khalaf, and I. Genest (1998). Monte Carlo tests for heteroskedasticity. Discussion paper, Département d'économique, Université Laval and CRDE, Université de Montréal.
- Berndt, E.R. (1991). *The Practice of Econometrics: Classic and Contemporary*. Reading (MA): Addison-Wesley.
- Birnbaum, Z.W. (1974). Computers and unconventional test-statistics. In F. Proschan, and R.J. Serfling (eds.) *Reliability and Biometry*, pp. 441–58. Philadelphia, PA: SIAM.
- Bound, J., D.A. Jaeger, and R.M. Baker (1995). Problems with instrumental variables estimation when the correlation between the instruments and the endogenous explanatory variable is weak. *Journal of the American Statistical Association* 90, 443–50.
- Breusch, T.S. (1980). Useful invariance results for generalized regression models. *Journal of Econometrics* 13, 327–40.
- Breusch, T.S., and A.R. Pagan (1979). A simple test for heteroscedasticity and random coefficient variation. *Econometrica* 47, 1287–94.

- Breusch, T.S., and A.R. Pagan (1980). The Lagrange multiplier test and its applications to model specification in econometrics. *Review of Economic Studies* 47, 239–54.
- Corana, A., M. Marchesi, C. Martini, and S. Ridella (1987). Minimizing multimodal functions of continuous variables with the “Simulated Annealing” algorithm. *ACM Transactions on Mathematical Software* 13, 262–80.
- Dagenais, M.G., and J.-M. Dufour (1991). Invariance, nonlinear models and asymptotic tests. *Econometrica* 59, 1601–15.
- D’Agostino, R.B., and M.A. Stephens (eds.) (1986). *Goodness-of-Fit Techniques*. New York: Marcel Dekker.
- Davies, R.B. (1977). Hypothesis testing when a nuisance parameter is present only under the alternative. *Biometrika* 64, 247–54.
- Davies, R.B. (1987). Hypothesis testing when a nuisance parameter is present only under the alternative. *Biometrika* 74, 33–43.
- Davison, A., and D. Hinkley (1997). *Bootstrap Methods and Their Application*. Cambridge: Cambridge University Press.
- Diebold, F.X., and C. Chen (1996). Testing structural stability with endogenous breakpoint: A size comparison of analytic and bootstrap procedures. *Journal of Econometrics* 70, 221–41.
- Dufour, J.-M. (1989). Nonlinear hypotheses, inequality restrictions, and nonnested hypotheses: Exact simultaneous tests in linear regressions. *Econometrica* 57, 335–55.
- Dufour, J.-M. (1990). Exact tests and confidence sets in linear regressions with autocorrelated errors. *Econometrica* 58, 475–94.
- Dufour, J.-M. (1995). Monte Carlo tests with nuisance parameters: A general approach to finite-sample inference and nonstandard asymptotics in econometrics. Discussion paper, C.R.D.E., Université de Montréal.
- Dufour, J.-M. (1997). Some impossibility theorems in econometrics, with applications to structural and dynamic models. *Econometrica* 65, 1365–89.
- Dufour, J.-M., A. Farhat, L. Gardiol, and L. Khalaf (1998). Simulation-based finite sample normality tests in linear regressions. *Econometrics Journal* 1, 154–73.
- Dufour, J.M., and J. Jasiak (1996). Finite sample inference methods for simultaneous equations and models with unobserved and generated regressors. Discussion paper, C.R.D.E., Université de Montréal.
- Dufour, J.-M., and L. Khalaf (1998a). Monte Carlo tests for contemporaneous correlation of disturbances in multiequation SURE models. Discussion paper, C.R.D.E., Université de Montréal.
- Dufour, J.-M., and L. Khalaf (1998c). Simulation based finite and large sample inference methods in multivariate regressions and seemingly unrelated regressions. Discussion paper, C.R.D.E., Université de Montréal.
- Dufour, J.-M., and L. Khalaf (1998b). Simulation-based finite and large sample inference methods in simultaneous equations. Discussion paper, C.R.D.E., Université de Montréal.
- Dufour, J.-M., and J.F. Kiviet (1996). Exact tests for structural change in first-order dynamic models. *Journal of Econometrics* 70, 39–68.
- Dufour, J.-M., and J.F. Kiviet (1998). Exact inference methods for first-order autoregressive distributed lag models. *Econometrica* 66, 79–104.
- Dwass, M. (1957). Modified randomization tests for nonparametric hypotheses. *Annals of Mathematical Statistics* 28, 181–7.
- Efron, B. (1982). *The Jackknife, the Bootstrap and Other Resampling Plans*, CBS-NSF Regional Conference Series in Applied Mathematics, Monograph No. 38. Society for Industrial and Applied Mathematics, Philadelphia, PA.

- Efron, B., and R.J. Tibshirani (1993). *An Introduction to the Bootstrap*, vol. 57 of *Monographs on Statistics and Applied Probability*. New York: Chapman & Hall.
- Gallant, A.R., and G. Tauchen (1996). Which moments to match? *Econometric Theory* 12, 657–81.
- Gibbons, M.R., S.A. Ross, and J. Shanken (1989). A test of the efficiency of a given portfolio. *Econometrica* 57, 1121–52.
- Goffe, W.L., G.D. Ferrier, and J. Rogers (1994). Global optimization of statistical functions with simulated annealing. *Journal of Econometrics* 60, 65–99.
- Gouriéroux, C., and A. Monfort (1995). *Statistics and Econometric Models, Volumes One and Two*. Cambridge: Cambridge University Press.
- Gouriéroux, C., A. Monfort, and E. Renault (1993). Indirect inference. *Journal of Applied Econometrics* 8S, 85–118.
- Hajivassiliou, V.A. (1993). Simulation estimation methods for limited dependent variables. In G.S. Maddala, C.R. Rao, and H.D. Vinod (eds.) *Handbook of Statistics, Volume 11, Econometrics*, pp. 519–43. Amsterdam: North-Holland.
- Hall, P. (1992). *The Bootstrap and Edgeworth Expansion*. New York: Springer-Verlag.
- Horowitz, J.L. (1997). Bootstrap methods in econometrics: Theory and numerical performance. In D. Kreps, and K.W. Wallis (eds.) *Advances in Economics and Econometrics*, vol. 3, pp. 188–222. Cambridge: Cambridge University Press.
- Jarque, C.M., and A.K. Bera (1980). Efficient tests for normality, heteroscedasticity and serial independence of regression residuals. *Economics Letters* 6, 255–9.
- Jarque, C.M., and A.K. Bera (1987). A test for normality of observations and regression residuals. *International Statistical Review* 55, 163–72.
- Jeong, J., and G.S. Maddala (1993). A perspective on application of bootstrap methods in econometrics. In G.S. Maddala, C.R. Rao, and H.D. Vinod (eds.) *Handbook of Statistics, Volume 11, Econometrics*, pp. 573–610. Amsterdam: North-Holland.
- Keane, M.P. (1993). Simulation estimation for panel data models with limited dependent variables. In Maddala, Rao, and Vinod (1993), pp. 545–571.
- Kiviet, J.F., and J.-M. Dufour (1997). Exact tests in single equation autoregressive distributed lag models. *Journal of Econometrics* 80, 325–53.
- Kolmogorov, A.N. (1933). Sulla determinazione empirica di una legge di distribuzione. *Giorn. Ist. Attuari* 4, 83–91.
- Laitinen, K. (1978). Why is demand homogeneity so often rejected? *Economics Letters* 1, 187–91.
- Lehmann, E.L. (1986). *Testing Statistical Hypotheses*, 2nd edn. New York: John Wiley & Sons.
- Maddala, G.S., C.R. Rao, and H.D. Vinod (eds.) (1993). *Handbook of Statistics, Volume 11, Econometrics*. Amsterdam: North-Holland.
- Mariano, R.S., and B.W. Brown (1993). Stochastic simulation for inference in nonlinear errors-in-variables models. In Maddala, Rao, and Vinod (1993), pp. 611–27.
- Nelson, C.R., and R. Startz (1990a). The distribution of the instrumental variable estimator and its *t*-ratio when the instrument is a poor one. *Journal of Business* 63, 125–40.
- Nelson, C.R., and R. Startz (1990b). Some further results on the exact small properties of the instrumental variable estimator. *Econometrica* 58, 967–76.
- Saphores, J.-D., L. Khalaf, and D. Pelletier (1998). Modelling unexpected changes in stumpage prices: an application to Pacific Northwest National Forests. Discussion paper, GREEN, Université Laval, Québec.
- Shao, S., and D. Tu (1995). *The Jackknife and Bootstrap*. New York: Springer-Verlag.
- Smirnov, N.V. (1939). Sur les écarts de la courbe de distribution empirique (Russian/French Summary). *Matematicheskiĭ Sbornik N.S.* 6, 3–26.

- Staiger, D., and J.H. Stock (1997). Instrumental variables regression with weak instruments. *Econometrica* 65, 557–86.
- Stewart, K.G. (1997). Exact testing in multivariate regression. *Econometric Reviews* 16, 321–52.
- Vinod, H.D. (1993). Bootstrap methods: Applications in econometrics. in Maddala, Rao, and Vinod (1993), pp. 629–61.
- Wilks, S.S. (1932). Certain generalizations in the analysis of variance. *Biometrika* 24, 471–94.
- Zellner, A. (1962). An efficient method for estimating seemingly unrelated regressions and tests for aggregate bias. *Journal of the American Statistical Association* 57, 348–68.

CHAPTER TWENTY-FOUR

# Bayesian Analysis of Stochastic Frontier Models

*Gary Koop and Mark F.J. Steel*

## 1 INTRODUCTION

Stochastic frontier models are commonly used in the empirical study of firm<sup>1</sup> efficiency and productivity. The seminal papers in the field are Aigner, Lovell, and Schmidt (1977) and Meeusen and van den Broeck (1977), while a recent survey is provided in Bauer (1990). The ideas underlying this class of models can be demonstrated using a simple production model<sup>2</sup> where output of firm  $i$ ,  $Y_i$ , is produced using a vector of inputs,  $X_i$ , ( $i = 1 \dots N$ ). The best practice technology for turning inputs into output depends on a vector of unknown parameters,  $\beta$ , and is given by:

$$Y_i = f(X_i; \beta). \quad (24.1)$$

This so-called production frontier captures the maximum amount of output that can be obtained from a given level of inputs. In practice, actual output of a firm may fall below the maximum possible. The deviation of actual from maximum output is a measure of inefficiency and is the focus of interest in many applications. Formally, equation (24.1) can be extended to:

$$Y_i = f(X_i; \beta)\tau_i, \quad (24.2)$$

where  $0 \leq \tau_i \leq 1$  is a measure of firm-specific efficiency and  $\tau_i = 1$  indicates firm  $i$  is fully efficient.

In this chapter we will discuss Bayesian inference in such models. We will draw on our previous work in the area: van den Broeck, Koop, Osiewalski, and Steel (1994), Koop, Osiewalski, and Steel (1994, 1997, 1999, 2000) and Koop, Steel, and Osiewalski (1995), whereas theoretical foundations can be found in Fernández, Osiewalski, and Steel (1997). It is worthwhile to digress briefly as to why we think these models are worthy of serious study. Efficiency measurement is very important in many areas of economics<sup>3</sup> and, hence, worthy of study in and of itself. However, stochastic frontier models are also close to other classes of models and can be used to illustrate ideas relating to the linear and nonlinear regression models; models for panel data, variance components, random coefficients, and, generally, models with unobserved heterogeneity. Thus, stochastic frontier models can be used to illustrate Bayesian methods in many areas of econometrics. To justify our adoption of the Bayesian paradigm, the reader is referred to our work in the area. Suffice it to note here that the competitors to the Bayesian approach advocated here are the classical econometric stochastic frontier approach (see Bauer, 1990 for a survey) and the deterministic or nonparametric data envelopment analysis (DEA) approach (see, e.g., Färe, Grosskopf, and Lovell, 1994). Each of the three approaches has strengths and weaknesses, some of which will be noted in this chapter.

This chapter is intended to be reasonably self-contained. However, we do assume that the reader has a basic knowledge of Bayesian methods as applied to the linear regression model (e.g. Judge, Griffiths, Hill, Lütkepohl, and Lee, 1985, ch. 4 or Poirier, 1995, pp. 288–309 and 524–50). Furthermore, we assume some knowledge of simulation methods. Koop (1994, pp. 12–26) provides a simple survey of some of these methods. Osiewalski and Steel (1998) focuses on simulation methods in the context of stochastic frontier models. Casella and George (1992) and Chib and Greenberg (1995) are good expository sources for Gibbs sampling and Metropolis–Hastings algorithms, respectively. Geweke (1999) is a complete survey of both Bayesian methods and computation.

The remainder of the chapter is organized as follows. The second section considers the stochastic frontier model with cross-sectional data beginning with a simple loglinear model then considering a nonlinear extension and one where explanatory variables enter the efficiency distribution. The third section discusses the issues raised by the availability of panel data.

## 2 THE STOCHASTIC FRONTIER MODEL WITH CROSS-SECTIONAL DATA

### 2.1 Introduction and notation

The model given in equation (24.2) implicitly assumes that all deviations from the frontier are due to inefficiency. This assumption is also typically made in the DEA approach. However, following standard econometric practice, we add a random error to the model,  $\zeta_i$ , to capture measurement (or specification) error,<sup>4</sup> resulting in:

$$Y_i = f(X_i; \beta)\tau_i\zeta_i. \quad (24.3)$$

The addition of measurement error makes the frontier stochastic, hence the term "stochastic frontier models". We assume that data for  $i = 1 \dots N$  firms is available and that the production frontier,  $f(\cdot)$ , is log-linear (e.g. Cobb–Douglas or translog). We define  $X_i$  as a  $1 \times (k + 1)$  vector (e.g.  $X_i = (1 \ L_i \ K_i)$  in the case of a Cobb–Douglas frontier with two inputs,  $L$  and  $K$ ) and, hence, (24.3) can be written as:

$$y_i = x_i\beta + v_i - z_i, \quad (24.4)$$

where  $\beta = (\beta_0 \dots \beta_k)'$ ,  $y_i = \ln(Y_i)$ ,  $v_i = \ln(\zeta_i)$ ,  $z_i = -\ln(\tau_i)$  and  $x_i$  is the counterpart of  $X_i$  with the inputs transformed to logarithms.  $z_i$  is referred to as inefficiency and, since  $0 \leq \tau_i \leq 1$ , it is a nonnegative random variable. We assume that the model contains an intercept with coefficient  $\beta_0$ . Equation (24.4) looks like the standard linear regression model, except that the "error" is composed of two parts. This gives rise to another name for these models, viz. "composed error models".

For future reference, we define  $y = (y_1 \dots y_N)'$ ,  $v = (v_1 \dots v_N)'$ ,  $z = (z_1 \dots z_N)'$  and the  $N \times (k + 1)$  matrix  $x = (x_1' \dots x_N')'$ . Also, let  $f_G(a|b, c)$  denote the density function of a Gamma distribution with shape parameter  $b$  and scale  $c$  so that  $a$  has mean  $b/c$  and variance  $b/c^2$ .  $p(d) = f_N^r(d|g, F)$  indicates that  $d$  is  $r$ -variate normal with mean  $g$  and covariance matrix  $F$ . We will use  $I(\cdot)$  to denote the indicator function; i.e.  $I(G) = 1$  if event  $G$  occurs and is otherwise 0. Furthermore,  $I_N$  will indicate the  $N \times N$  identity matrix and  $\mathbf{1}_N$  and  $N \times 1$  vector of ones. Sample means will be indicated with a bar, e.g.  $\bar{y} = \frac{1}{N}\mathbf{1}'_Ny$ .

## 2.2 Bayesian inference

In order to define the sampling model,<sup>5</sup> we make the following assumptions about  $v_i$  and  $z_i$  for  $i = 1 \dots N$ :

1.  $p(v_i|h^{-1}) = f_N^1(v_i|0, h^{-1})$  and the  $v_i$ s are independent;
2.  $v_i$  and  $z_l$  are independent of one another for all  $i$  and  $l$ ;
3.  $p(z_i|\lambda^{-1}) = f_G(z_i|1, \lambda^{-1})$  and the  $z_i$ s are independent.

The first assumption is commonly made in cross-sectional analysis, but the last two require some justification. Assumption 2 says that measurement error and inefficiency are independent of one another. Assumption 3 is a common choice for the nonnegative random variable,  $z_i$ , although others (e.g. the half-normal) are possible. Ritter and Simar (1997) show that the use of very flexible one-sided distributions for  $z_i$  such as the unrestricted gamma may result in a problem of weak identification. Intuitively, if  $z_i$  is left too flexible, then the intercept minus  $z_i$  can come to look too much like  $v_i$  and it may become virtually impossible to distinguish between these two components with small data sets. The gamma with shape parameter 1 is the exponential distribution, which is sufficiently different from the normal to avoid this weak identification problem.<sup>6</sup> In addition, van den Broeck *et al.* (1994) found the exponential model the least sensitive to changes in prior assumptions in a study of the most commonly used models. Note that  $\lambda$  is the mean of the inefficiency distribution and let  $\theta = (\beta', h, \lambda)'$  denote the parameters of the model.

The likelihood function is defined as:

$$L(y; \theta) = \prod_{i=1}^N p(y_i | x_i, \theta),$$

which requires the derivation of  $p(y_i | x_i, \theta) = \int p(y_i | x_i, z_i, \theta)p(z_i | \theta)dz_i$ . This is done in Jondrow, Lovell, Materov, and Schmidt (1982) for the exponential model and in van den Broeck *et al.* (1994) for a wider class of inefficiency distributions. However, we do not repeat the derivation here, since we do not need to know the explicit form of the likelihood function. To understand why isolating the likelihood function is not required, it is necessary to explain the computational methods that we recommend for Bayesian inference in stochastic frontier models.

Bayesian inference can be carried out using a posterior simulator which generates draws from the posterior,  $p(\theta | y, x)$ . In this case, Gibbs sampling with data augmentation is a natural choice for a posterior simulator. This algorithm relies on the fact that sequential draws,  $\theta^{(s)}$  and  $z^{(s)}$ , from the conditional posteriors  $p(\theta | y, x, z^{(s-1)})$  and  $p(z | y, x, \theta^{(s)})$ , respectively, will converge to draws from  $p(\theta, z | y, x)$  from which inference on the marginal posteriors of  $\theta$  or of functions of  $z$  (such as efficiencies) can immediately be derived. In other words, we do not need to have an analytical formula for  $p(\theta | y, x)$  (and, hence, the likelihood function), but rather we can suffice with working out the full conditional distributions  $p(\theta | y, x, z)$  and  $p(z | y, x, \theta)$ . Intuitively, the former is very easy to work with since, conditional on  $z$ , the stochastic frontier model reduces to the standard linear regression model.<sup>7</sup> If  $p(\theta | y, x, z)$  as a whole is not analytically tractable, we can split up  $\theta$  into, say,  $\beta$  and  $(h, \lambda)$  and draw sequentially from the full conditionals  $p(\beta | h, \lambda, y, x, z)$ ,  $p(h, \lambda | \beta, y, x, z)$ , and  $p(z | y, x, \beta, h, \lambda)$ . However, before we can derive the Gibbs sampler, we must complete the Bayesian model by specifying a prior for the parameters.

The researcher can, of course, use any prior in an attempt to reflect his/her prior beliefs. However, a proper prior for  $h$  and  $\lambda^{-1}$  is advisable: Fernández *et al.* (1997) show that Bayesian inference is not feasible (in the sense that the posterior distribution is not well-defined) under the usual improper priors for  $h$  and  $\lambda^{-1}$ . Here, we will assume a prior of the product form:  $p(\theta) = p(\beta)p(h)p(\lambda^{-1})$ . In stochastic frontier models, prior information exists in the form of economic regularity conditions. It is extremely important to ensure that the production frontier satisfies these, since it is highly questionable to interpret deviations from a non-regular frontier as representing inefficiency. In an extreme case, if the researcher is using a highly flexible (or nonparametric) functional form for  $f(\cdot)$  it might be possible for the frontier to fit the data nearly perfectly. It is only the imposition of economic regularity conditions that prevent this overfitting. The exact form of the economic regularity conditions depend on the specification of the frontier. For instance, in the Cobb–Douglas case,  $\beta_i \geq 0, i = 1 \dots k$  ensures global regularity of the production frontier. For the translog specification things are more complicated and we may wish only to impose local regularity. This requires checking certain conditions at each data point (see Koop *et al.*, 1999). In either case, we can choose a prior for  $\beta$  which imposes economic regularity. As emphasized by

Fernández *et al.* (1997), a proper or bounded prior is sufficient for  $\beta$ . Thus, it is acceptable to use a uniform (flat) prior:

$$p(\beta) \propto I(E), \quad (24.5)$$

where  $I(E)$  is the indicator function for the economic regularity conditions. Alternatively, a normal prior for  $\beta$  is proper and computationally convenient. In this chapter, we will use  $p(\beta)$  as a general notation, but assume it is either truncated uniform or truncated normal. Both choices will easily combine with a normal distribution to produce a truncated normal posterior distribution.

For the other parameters, we assume gamma priors:

$$p(h) = f_G(h | a_h, b_h) \quad (24.6)$$

and

$$p(\lambda^{-1}) = f_G(\lambda^{-1} | a_\lambda, b_\lambda). \quad (24.7)$$

Note that, by setting  $a_h = 0$  and  $b_h = 0$  we obtain  $p(h) \propto h^{-1}$ , the usual noninformative prior for the error precision in the normal linear regression model. Here, the use of this improper prior is precluded (see Theorem 1 (ii) of Fernández *et al.*, 1997), but small values of these hyperparameters will allow for Bayesian inference (see Proposition 2 of Fernández *et al.*, 1997) while the prior is still dominated by the likelihood function. The hyperparameters  $a_\lambda$  and  $b_\lambda$  can often be elicited through consideration of the efficiency distribution. That is, researchers may often have prior information about the shape or location of the efficiency distribution. As discussed in van den Broeck *et al.* (1994), setting  $a_\lambda = 1$  and  $b_\lambda = -\ln(\tau^*)$  yields a relatively noninformative prior which implies the prior median of the efficiency distribution is  $\tau^*$ . These are the values for  $a_\lambda$  and  $b_\lambda$  used in the following discussion.

The Gibbs sampler can be developed in a straightforward manner by noting that, if  $z$  were known, then we could write the model as  $y + z = x\beta + v$  and standard results for the normal linear regression model can be used. In particular, we can obtain

$$p(\beta | y, x, z, h, \lambda^{-1}) = f_N^{k+1}(\beta | \hat{\beta}, h^{-1}(x'x)^{-1})p(\beta), \quad (24.8)$$

where

$$\hat{\beta} = (x'x)^{-1}x'(y + z).$$

Furthermore,

$$p(h | y, x, z, \beta, \lambda^{-1}) = f_G\left(h \left| a_h + \frac{N}{2}, b_h + \frac{(y - x\beta + z)'(y - x\beta + z)}{2}\right.\right). \quad (24.9)$$

Also, given  $z$ , the full conditional posterior for  $\lambda^{-1}$  can easily be derived:

$$p(\lambda^{-1} | y, x, z, \beta, h) = f_G(\lambda^{-1} | N + 1, z' \mathbf{1}_N - \ln(\tau^*)). \quad (24.10)$$

Equations (24.8), (24.9), and (24.10) are the full conditional posteriors necessary for setting up the Gibbs sampler *conditional on*  $z$ . To complete the posterior simulator, it is necessary to derive the posterior distribution of  $z$  conditional on  $\theta$ . Noting that we can write  $z = x\beta - y + v$ , where  $v$  has pdf  $f_N^N(v | 0, h^{-1}I_N)$  and  $z_i$  is a priori assumed to be iid  $f_G(z_i | 1, \lambda^{-1})$ ,<sup>8</sup> we obtain:

$$p(z | y, x, \beta, h, \lambda^{-1}) \propto f_N^N(z | x\beta - y - h^{-1}\lambda^{-1}\mathbf{1}_N, h^{-1}I_N) \prod_{i=1}^N I(z_i \geq 0). \quad (24.11)$$

A Gibbs sampler with data augmentation on  $(\beta, h, \lambda^{-1}, z)$  can be set up by sequentially drawing from (24.8), (24.9), (24.10), and (24.11), where  $(\beta, h)$  and  $\lambda^{-1}$  are independent given  $z$ , so that (24.10) can be combined with either (24.8) or (24.9) and there are only three steps in the Gibbs. Note that all that is required is random number generation from well known distributions, where drawing from the high-dimensional vector  $z$  is greatly simplified as (24.11) can be written as the product of  $N$  univariate truncated normals.

Given posterior simulator output, posterior properties of any of the parameters or of the individual  $\tau_i$ s can be obtained.<sup>9</sup> The latter can be calculated using simulated draws from (24.11) and transforming according to  $\tau_i = \exp(-z_i)$ . It is worth stressing that the Bayesian approach provides a finite sample distribution of the efficiency of each firm. This allows us to obtain both point and interval estimates, or even e.g.  $P(\tau_i > \tau_j | y, x)$ . The latter is potentially crucial since important policy consequences often hinge on one firm being labeled as more efficient in a statistically significant sense. Both DEA and classical econometric approaches typically only report point estimates. The DEA approach is nonparametric and, hence, confidence intervals for the efficiency measures obtained are very hard to derive.<sup>10</sup> Distributional theory for the classical econometric approach is discussed in Jondrow *et al.* (1982) and Horrace and Schmidt (1996). These papers point out that, although point estimates and confidence intervals for  $\tau_i$  can be calculated, the theoretical justification is not that strong. For example, the maximum likelihood estimator for  $\tau_i$  is inconsistent and the methods for constructing confidence intervals assume unknown parameters are equal to their point estimates. For this reason, it is common in classical econometric work to present some characteristics of the efficiency distribution as a whole (e.g. estimates of  $\lambda$ ) rather than discuss firm specific efficiency. However, firm specific efficiencies are often of fundamental policy importance and, hence, we would argue that an important advantage of the Bayesian approach is its development of finite sample distributions for the  $\tau_i$ s.

### 2.3 Extensions

There are many ways of extending the previous model. For instance, we could allow for different distributions for  $z_i$  (see Koop *et al.*, 1995) or for many outputs

to exist (see Fernández, Koop and Steel, 2000). Here we focus on two other extensions which are interesting in and of themselves, but also allow us to discuss some useful Bayesian techniques.

### EXPLANATORY VARIABLES IN THE EFFICIENCY DISTRIBUTION

Consider, for instance, a case where data are available for many firms, but some are private companies and others are state owned. Interest centers on investigating whether private companies tend to be more efficient than state owned ones. This type of question can be formally handled by stochastic frontier models if we extend them to allow for explanatory variables in the efficiency distribution. Let us suppose that data exist on  $m$  variables which may affect the efficiency of firms (i.e.  $w_{ij}$ , for  $i = 1 \dots N$  and  $j = 1 \dots m$ ). We assume  $w_{i1} = 1$  is an intercept and  $w_{ij}$  are 0–1 dummy variables for  $j = 2 \dots m$ . The latter assumption could be relaxed at the cost of increasing the complexity of the computational methods. Since  $\lambda$ , the mean of the inefficiency distribution, is a positive random variable, a logical extension of the previous model is to allow it to vary over firms in the following manner:

$$\lambda_i^{-1} = \prod_{j=1}^m \phi_j^{w_{ij}}, \quad (24.12)$$

where the  $\phi_j > 0$  are unknown parameters. Note that if  $\phi_j = 1$  for  $j = 2 \dots m$  then this model reduces to the previous one. To aid in interpretation, observe how this specification allows, for instance, for private and state owned firms to have different inefficiency distributions. If  $w_{i2} = 1$  indicates that firm  $i$  is private, then  $\phi_2 > 1$  implies that the mean of the inefficiency distribution is lower for private firms and, hence, that private firms tend to be more efficient than state owned ones. We stress that such a finding would not imply that every private firm is more efficient than every state owned one, but rather that the former are drawing their efficiencies from a distribution with a higher mean. Such a specification seems very suitable for many sorts of policy issues and immediately allows for out-of-sample predictions.

For the new parameters,  $\phi = (\phi_1 \dots \phi_m)'$ , we assume independent gamma priors:  $p(\phi) = p(\phi_1) \dots p(\phi_m)$  with  $p(\phi_j) = f_G(\phi_j | a_j, b_j)$  for  $j = 1 \dots m$ . If the explanatory variables have no role to play (i.e.  $\phi_2 = \dots = \phi_m = 1$ ), then  $\phi_1$  is equivalent to  $\lambda^{-1}$  in the previous model. This suggests one may want to follow the prior elicitation rule discussed above and set  $a_1 = 1$  and  $b_1 = -\ln(\tau^*)$ . The other prior hyperparameters,  $a_j$  and  $b_j$  for  $j = 2 \dots m$ , can be selected in the context of particular applications with moderate values for these parameters yielding a relatively noninformative prior. See Koop *et al.* (1997) for details.

A posterior simulator using Gibbs sampling with data augmentation can be set up as a straightforward extension of the one considered above. In fact, the posterior conditionals for  $\beta$  and  $h$  (i.e. equations (24.8) and (24.9)) are completely unaffected and the conditional for  $z$  in (24.11) is only affected in that  $\lambda^{-1}\mathbf{1}_N$  must be replaced by the vector  $\eta = (\lambda_1^{-1} \dots \lambda_N^{-1})'$ , where  $\lambda_i^{-1}$  is given in equation (24.12). It can also be verified that for  $j = 1 \dots m$ :<sup>11</sup>

$$p(\phi_j | y, x, z, \beta, h, w, \phi^{(-j)}) = f_G\left(\phi_j \middle| a_j + \sum_{i=1}^N w_{ij}, b_j + \sum_{i=1}^N w_{ij} z_i, \prod_{s \neq j} \phi_s^{w_{is}}\right), \quad (24.13)$$

where  $\phi^{(-j)} = (\phi_1 \dots \phi_{j-1}, \phi_{j+1} \dots \phi_m)'$ . Hence, Bayesian inference in this model can again be conducted through sequential drawing from tractable distributions.

So far, we have focused on posterior inference. This stochastic frontier model with varying efficiency distribution can be used to illustrate Bayesian model comparison. Suppose  $m = 2$  and we are interested in calculating the Bayes factor comparing model  $M_1$  where  $\phi_2 = 1$  (e.g. there is no tendency for state owned and private firms to differ in their efficiency distributions) against model  $M_2$  with  $\phi_2 \neq 1$ . The prior for  $M_2$  is given above. Define  $\psi = (\beta, h, \phi^{(-2)})'$  as the parameters in the model  $M_1$  and let  $p_l(\cdot)$  indicate a density under  $M_l$  for  $l = 1, 2$ . If we make the reasonable assumption that  $p_2(\psi | \phi_2 = 1) = p_1(\psi)$ , then the Bayes factor in favor of  $M_1$  can be written as the Savage–Dickey density ratio (see Verdinelli and Wasserman, 1995):

$$B_{12} = \frac{p_2(\phi_2 = 1 | y, x, w)}{p_2(\phi_2 = 1)}, \quad (24.14)$$

the ratio of posterior to prior density values at the point being tested. Note that the denominator of (24.14) is trivial to calculate since it is merely the gamma prior for  $\phi_2$  evaluated at a point. The numerator is also easy to calculate using (24.13). As Verdinelli and Wasserman (1995) stress, a good estimator of  $p(\phi_2 = 1 | y, x, w)$  on the basis of  $R$  Gibbs replications is:

$$\frac{1}{R} \sum_{r=1}^R p(\phi_2 = 1 | y, x, z^{(r)}, \beta^{(r)}, h^{(r)}, w, \phi^{(-2)(r)}), \quad (24.15)$$

where superscript  $(r)$  denotes the  $r$ th draw in the Gibbs sampling algorithm. That is, we can just evaluate (24.13) at  $\phi_2 = 1$  for each draw and average. Bayes factors for hypotheses such as this can be easily calculated without recourse to evaluating the likelihood function or adding steps to the simulation algorithm (as in the more general methods of Gelfand and Dey, 1994 and Chib, 1995, respectively).

## NONLINEAR PRODUCTION FRONTIERS

The previous models both assumed that the production frontier was log-linear. However, many common production functions are inherently nonlinear in the parameters (e.g. the constant elasticity of substitution or CES or the asymptotically ideal model or AIM, see Koop *et al.*, 1994). However, the techniques outlined above can be extended to allow for an arbitrary production function. Here we assume a model identical to the stochastic frontier model with common efficiency distribution (i.e.  $m = 1$ ) except that the production frontier is of the form:<sup>12</sup>

$$y_i = f(x_i; \beta) + v_i - z_i. \quad (24.16)$$

The posterior simulator for everything except  $\beta$  is almost identical to the one given above. Equation (24.10) is completely unaffected, and (24.9) and (24.11) are slightly altered by replacing  $x\beta$  by  $f(x, \beta) = (f(x_1; \beta) \dots f(x_N; \beta))'$ .

However, the conditional posterior for  $\beta$  is more complicated, having the form:

$$p(\beta | y, x, z, h, \lambda^{-1}) \propto \exp\left(-\frac{h}{2} \sum_{i=1}^N (y_i - f(x_i; \beta) + z_i)^2\right) p(\beta). \quad (24.17)$$

Equation (24.17) does not take the form of any well known density and the computational algorithm selected will depend on the exact form of  $f(x; \beta)$ . For the sake of brevity, here we will only point the reader in the direction of possible algorithms that may be used for drawing from (24.17). Two major cases are worth mentioning. First, in many cases, it might be possible to find a convenient density which approximates (24.17) well. For instance, in the case of the AIM model a multivariate- $t$  density worked well (see Koop *et al.*, 1994). In this case, importance sampling (Geweke, 1989) or an independence chain Metropolis–Hastings algorithm (Chib and Greenberg, 1995) should work well. On the other hand, if no convenient approximating density can be found, a random walk chain Metropolis–Hastings algorithm might prove a good choice (see Chib and Greenberg, 1995). The precise choice of algorithm will be case-specific and, hence, we do not discuss this issue in any more detail here.

### 3 THE STOCHASTIC FRONTIER MODEL WITH PANEL DATA

#### 3.1 Time-invariant efficiency

It is increasingly common to use panel data<sup>13</sup> in the classical econometric analysis of the stochastic frontier model. Some of the statistical problems (e.g. inconsistency of point estimates of firm specific efficiency) of classical analysis are alleviated with panel data and the assumption of a particular distributional form for the inefficiency distribution can be dispensed with at the cost of assuming time-invariant efficiencies (i.e. treating them as “individual effects”). Schmidt and Sickles (1984) is an early influential paper which develops a relative efficiency measure based on a fixed effects specification and an absolute efficiency measure based on a random effects specification. In this paper, we describe a Bayesian alternative to this classical analysis and relate the random/fixed effects distinction to different prior structures for the efficiency distribution.

Accordingly, assume that data is available for  $i = 1 \dots N$  firms for  $t = 1 \dots T$  time periods. We will extend the notation of the previous section so that  $y_i$  and  $v_i$  are now  $T \times 1$  vectors and  $x_i$  a  $T \times k$  matrix containing the  $T$  observations for firm  $i$ . Note, however, that the assumption of constant efficiency over time implies that  $z_i$  is still a scalar and  $z$  an  $N \times 1$  vector. For future reference, we now define  $y = (y'_1 \dots y'_N)'$  and  $v = (v'_1 \dots v'_N)'$  as  $NT \times 1$  vectors and  $x = (x'_1 \dots x'_N)'$  as an  $NT \times k$  matrix. In contrast to previous notation,  $x_i$  does not contain an intercept. We assume that the stochastic frontier model can be written as:

$$y_i = \beta_0 \mathbf{1}_T + x_i \boldsymbol{\delta} + v_i - z_i \mathbf{1}_T, \quad (24.18)$$

where  $\beta_0$  is the intercept coefficient and  $v_i$  is iid with pdf  $f_N^T(v_i | 0, h^{-1}I_T)$ . As discussed in Fernández *et al.* (1997), it is acceptable to use an improper noninformative prior for  $h$  when  $T > 1$  and, hence, we assume  $p(h) \propto h^{-1}$ . We discuss different choices of priors for  $\beta_0$  and  $z_i$  in the following material.

### BAYESIAN FIXED EFFECTS MODEL

Equation (24.18) looks like a standard panel data model (see, e.g., Judge *et al.*, 1985, ch. 13). The individual effect in the model can be written as:

$$\alpha_i = \beta_0 - z_i,$$

and the model rewritten as:

$$y_i = \alpha_i \mathbf{1}_T + x_i \boldsymbol{\delta} + v_i. \quad (24.19)$$

Classical fixed effects estimation of (24.19) proceeds by making no distributional assumption for  $\alpha_i$ , but rather using firm-specific dummy variables. The Bayesian analog to this is to use flat, noninformative priors for the  $\alpha_i$ 's.<sup>14</sup> Formally, defining  $\alpha = (\alpha_1 \dots \alpha_N)'$ , we then adopt the prior:

$$p(\alpha, \boldsymbol{\delta}, h) \propto h^{-1} p(\boldsymbol{\delta}). \quad (24.20)$$

The trouble with this specification is that we cannot make direct inference about  $z_i$  (since  $\beta_0$  is not separately identified) and, hence, the absolute efficiency of firm  $i$ :  $\tau_i = \exp(-z_i)$ . However, following Schmidt and Sickles (1984), we define relative inefficiency as:

$$z_i^{rel} = z_i - \min_j(z_j) = \max_j(\alpha_j) - \alpha_i. \quad (24.21)$$

In other words, we are measuring inefficiency relative to the most efficient firm (i.e. the firm with the highest  $\alpha_i$ ).<sup>15</sup> Relative efficiency is defined as  $r_i^{rel} = \exp(-z_i^{rel})$  and we assume that the most efficient firm has  $r_i^{rel} = 1$ .

It is worth noting that this prior seems like an innocuous noninformative prior, but this initial impression is false since it implies a rather unusual prior for  $r_i^{rel}$ . In particular, as shown in Koop *et al.* (1997),  $p(r_i^{rel})$  has a point mass of  $N^{-1}$  at full efficiency and is  $p(r_i^{rel}) \propto 1/r_i^{rel}$  for  $r_i^{rel} \in (0, 1)$ . The latter is an L-shaped improper prior density which, for an arbitrary small  $a \in (0, 1)$  puts an infinite mass in  $(0, a)$  but only a finite mass in  $(a, 1)$ . In other words, this “noninformative” prior strongly favors low efficiency.

Bayesian inference in the fixed effects model can be carried out in a straightforward manner, by noting that for uniform  $p(\boldsymbol{\delta})$ , (24.19)–(24.20) is precisely a normal linear regression model with Jeffreys’ prior. The vector of regression coefficients  $(\alpha' \ \boldsymbol{\delta}')'$  in such a model has a  $(N + k)$ -variate student- $t$  posterior with  $N(T - 1) - k$  degrees of freedom (where we have assumed that  $N(T - 1) > k$ , which

implies  $T > 1$ ). For typical values of  $N$ ,  $T$ , and  $k$  the degrees of freedom are enormous and the student- $t$  will be virtually identical to the normal distribution. Hence, throughout this subsection we present results in terms of this normal approximation.

Using standard Bayesian results for the normal linear regression model (e.g. Judge *et al.*, 1985, ch. 4), it follows that the marginal posterior for  $\delta$  is given by (for general  $p(\delta)$ ):

$$p(\delta | y, x) = f_N^k(\delta | \hat{\delta}, \hat{h}^{-1}S^{-1})p(\delta), \quad (24.22)$$

where

$$\hat{\delta} = S^{-1} \sum_{i=1}^N (x_i - \mathbf{1}_T \bar{x}_i)'(y_i - \bar{y}_i \mathbf{1}_T), \quad (24.23)$$

$$S = \sum_{i=1}^N S_i, \quad \bar{x}_i = \frac{1}{T} \mathbf{1}'_T x_i$$

and

$$S_i = (x_i - \mathbf{1}_T \bar{x}_i)'(x_i - \mathbf{1}_T \bar{x}_i).$$

Note that (24.23) is the standard “within estimator” from the panel data literature. Finally,

$$\hat{h}^{-1} = \frac{1}{N(T-1)-k} \sum_{i=1}^N (y_i - \hat{\alpha}_i \mathbf{1}_T - x_i \hat{\delta})' (y_i - \hat{\alpha}_i \mathbf{1}_T - x_i \hat{\delta}),$$

where  $\hat{\alpha}_i$  is the posterior mean of  $\alpha_i$  defined below.

The marginal posterior of  $\alpha$  is the  $N$ -variate normal with means

$$\hat{\alpha}_i = \bar{y}_i - \bar{x}_i \hat{\delta},$$

and covariances

$$\text{cov}(\alpha_i, \alpha_j) = \hat{h}^{-1} \left( \frac{\Delta(i,j)}{T} + \bar{x}_i S^{-1} \bar{x}_j' \right),$$

where  $\Delta(i, j) = 1$  if  $i = j$  and 0 otherwise. Thus, analytical formulae for posterior means and standard deviations are available and, if interest centers on these, posterior simulation methods are not required. However, typically interest centers on the relative efficiencies which are a complicated nonlinear function of  $\alpha$ , viz.,

$$r_i^{rel} = \exp(\alpha_i - \max_j(\alpha_j)), \quad (24.24)$$

and, hence, posterior simulation methods are required. However, direct Monte Carlo integration is possible since the posterior for  $\alpha$  is multivariate normal and can easily be simulated. These simulated draws of  $\alpha$  can be transformed using (24.24) to yield posterior draws of  $r_i^{rel}$ . However, this procedure is complicated by the fact that we do not know which firm is most efficient (i.e. which firm has largest  $\alpha_j$ ) and, hence, is worth describing in detail.

We begin by calculating the probability that a given firm,  $i$ , is the most efficient:

$$P(r_i^{rel} = 1 | y, x) = P(\alpha_i = \max_j (\alpha_j) | y, x), \quad (24.25)$$

which can be easily calculated using Monte Carlo integration. That is, (24.25) can simply be estimated by the proportion of the draws of  $\alpha$  which have  $\alpha_i$  being the largest.

Now consider the posterior for  $r_i^{rel}$  over the interval  $(0, 1)$  (i.e. assuming it is *not* the most efficient):

$$p(r_i^{rel} | y, x) = \sum_{j=1, j \neq i}^N p(r_i^{rel} | y, x, r_j^{rel} = 1) P(r_j^{rel} = 1 | y, x). \quad (24.26)$$

Here  $P(r_j^{rel} = 1 | y, x)$  can be calculated as discussed in the previous paragraph. In addition,  $p(r_i^{rel} | y, x, r_j^{rel} = 1)$  can be calculated using the same posterior simulator output. That is, assuming firm  $j$  is most efficient, then  $r_i^{rel} = \exp(\alpha_i - \alpha_j)$  which can be evaluated from those draws of  $\alpha$  that correspond to  $\alpha_j = \max_l (\alpha_l)$ . Hence, posterior analysis of the relative efficiencies in a Bayesian fixed effects framework can be calculated in a straightforward manner.<sup>16</sup>

## BAYESIAN RANDOM EFFECTS MODEL

The Bayesian fixed effects model described above might initially appeal to researchers who do not want to make distributional assumptions about the inefficiency distribution. However, as we have shown above, this model is implicitly making strong and possibly unreasonable prior assumptions. Furthermore, we can only calculate relative, as opposed to absolute, efficiencies. For these reasons, it is desirable to develop a model which makes an explicit distributional assumption for the inefficiencies. With such a model, absolute efficiencies can be calculated in the spirit of the cross-sectional stochastic frontier model of Section 2, since the distribution assumed for the  $z_i$ s allows us to separately identify  $z_i$  and  $\beta_0$ . In addition, the resulting prior efficiency distributions will typically be more in line with our prior beliefs. Another important issue is the sensitivity of the posterior results on efficiency to the prior specification chosen. Since  $T$  is usually quite small, it makes sense to “borrow strength” from the observations for the other firms by linking the inefficiencies. Due to Assumption 3 in subsection 2.2, this is not done through the sampling model. Thus, Koop *et al.* (1997) define the difference between Bayesian fixed and random effects models through the prior for  $z_i$ . In particular, what matters are the prior links that are assumed between the  $z_i$ s. Fixed effects models assume, *a priori*, that the  $z_i$ s are fully separated.

Random effects models introduce links between the  $z_i$ s, typically by assuming they are all drawn from distributions that share some common unknown parameter(s). In Bayesian language, the random effects model then implies a hierarchical prior for the individual effects.

Formally, we define a Bayesian random effects model by combining (24.18) with the prior:

$$p(\beta_0, \delta, h, z, \lambda^{-1}) \propto h^{-1} p(\delta) f_G(\lambda^{-1} | 1, -\ln(\tau^*)) \prod_{i=1}^N f_G(z_i | 1, \lambda^{-1}). \quad (24.27)$$

That is, we assume noninformative priors for  $h$  and  $\beta_0$ , whereas the inefficiencies are again assumed to be drawn from the exponential distribution with mean  $\lambda$ . Note that the  $z_i$ s are now linked through this common parameter  $\lambda$ , for which we choose the same prior as in Section 2.

Bayesian analysis of this model proceeds along similar lines to the cross-sectional stochastic frontier model presented in Section 2. In particular, a Gibbs sampler with data augmentation can be set up. Defining  $\beta = (\beta_0 \ \delta)'$  and  $X = (\mathbf{1}_{NT} : x)$  the posterior conditional for the measurement error precision can be written as:

$$p(h | y, x, z, \beta, \lambda^{-1}) = f_G\left(h \middle| \frac{NT}{2}, \frac{1}{2}[y - X\beta + (I_N \otimes \mathbf{1}_T)z]'[y - X\beta + (I_N \otimes \mathbf{1}_T)z]\right). \quad (24.28)$$

Next we obtain:

$$p(\beta | y, x, z, h, \lambda^{-1}) = f_N^{k+1}(\beta | \bar{\beta}, h^{-1}(X'X)^{-1})p(\delta), \quad (24.29)$$

where

$$\bar{\beta} = (X'X)^{-1} [y + (I_N \otimes \mathbf{1}_T)z].$$

The posterior conditional for the inefficiencies takes the form:

$$p(z | y, x, \beta, h, \lambda^{-1}) \propto f_N^N(z | (\mathbf{1}_N : \tilde{x})\beta - \tilde{y} - (Th\lambda)^{-1}\mathbf{1}_N, (Th)^{-1}I_N) \prod_{i=1}^N I(z_i \geq 0), \quad (24.30)$$

where  $\tilde{y} = (\bar{y}_1 \dots \bar{y}_N)'$  and  $\tilde{x} = (\tilde{x}'_1 \dots \tilde{x}'_N)'$ .

Furthermore, the posterior conditional for  $\lambda^{-1}$ ,  $p(\lambda^{-1} | y, x, z, \beta, h)$ , is the same as for the cross-sectional case (i.e. equation 24.10).

Using these results, Bayesian inference can be carried out using a Gibbs sampling algorithm based on (24.10), (24.28), (24.29), and (24.30). Although the formulas look somewhat complicated, it is worth stressing that all conditionals are either gamma or truncated normal.

### 3.2 Extensions

Extending the random effects stochastic frontier model to allow for a nonlinear production function or explanatory variables in the efficiency distribution can easily be done in a similar fashion as for the cross-sectional model (see subsection 2.3 and Koop *et al.*, 1997). Furthermore, different efficiency distributions can be allowed for in a straightforward manner and multiple outputs can be handled as in Fernández *et al.* (2000). Here we concentrate on extending the model in a different direction. In particular, we free up the assumption that each firm's efficiency,  $\tau_{it}$ , is constant over time. Let us use the definitions of  $X$  and  $\beta$  introduced in the previous subsection and write the stochastic frontier model with panel data as:

$$y = X\beta - \gamma + v, \quad (24.31)$$

where  $\gamma$  is a  $TN \times 1$  vector containing inefficiencies for each individual observation and  $y$  and  $v$  are defined as in subsection 3.1. In practice, we may want to put some structure on  $\gamma$  and, thus, Fernández *et al.* (1997) propose to rewrite it in terms of an  $M$ -dimensional vector  $u$  as:

$$\gamma = Du, \quad (24.32)$$

where  $M \leq TN$  and  $D$  is a known  $TN \times M$  matrix. Above, we implicitly assumed  $D = I_N \otimes \mathbf{1}_T$  which implies  $M = N$  and  $u_i = \gamma_{it} = z_i$ . That is, firm-specific inefficiency was constant over time. However, a myriad of other possibilities exist. For instance,  $D$  can correspond to cases where clusters of firms or time periods share common efficiencies, or parametric time dependence exists in firm-specific efficiency. Also note the case  $D = I_{TN}$ , which allows each firm in each period to have a different inefficiency (i.e.  $\gamma = u$ ). Thus, we are then effectively back in the cross section framework without exploiting the panel structure of the data. This case is considered in Koop *et al.* (1999, 2000), where interest is centered on the change in efficiency over time.<sup>17</sup> With all such specifications, it is possible to conduct Bayesian inference by slightly altering the posterior conditionals presented above in an obvious manner.

However, as discussed in Fernández *et al.* (1997), it is very important to be careful when using improper priors on any of the parameters. In some cases, improper priors imply that the posterior does not exist and, hence, valid Bayesian inference cannot be carried out. Intuitively, the inefficiencies can be interpreted as unknown parameters. If there are too many of these, prior information becomes necessary. As an example of the types of results proved in Fernández *et al.* (1997), we state one of their main theorems:

**Theorem 24.1 (Fernández *et al.*, 1997, Theorem 1).** Consider the general model given in (24.31) and (24.32) and assume the standard noninformative prior for  $h : p(h) \propto h^{-1}$ . If  $\text{rank}(X : D) < TN$  then the posterior distribution exists for

any bounded or proper  $p(\beta)$  and any proper  $p(u)$ . However, if  $\text{rank}(X : D) = TN$ , then the posterior does not exist.

The Bayesian random effects model discussed above has  $\text{rank}(X : D) < TN$ , so the posterior does exist even though we have used an improper prior for  $h$ . However, for the case where efficiency varies over time and across firms (i.e.  $D = I_{TN}$ ), more informative priors are required in order to carry out valid Bayesian inference. Fernández *et al.* (1997, Proposition 2) show that a weakly informative (not necessarily proper) prior on  $h$  that penalizes large values of the precision is sufficient.

## 4 SUMMARY

In this chapter, we have described a Bayesian approach to efficiency analysis using stochastic frontier models. With cross-sectional data and a log-linear frontier, a simple Gibbs sampler can be used to carry out Bayesian inference. In the case of a nonlinear frontier, more complicated posterior simulation methods are necessary. Bayesian efficiency measurement with panel data is then discussed. We show how a Bayesian analog of the classical fixed effects panel data model can be used to calculate the efficiency of each firm relative to the most efficient firm. However, absolute efficiency calculations are precluded in this model and inference on efficiencies can be quite sensitive to prior assumptions. Accordingly, we describe a Bayesian analog of the classical random effects panel data model which can be used for robust inference on absolute efficiencies. Throughout we emphasize the computational methods necessary to carry out Bayesian inference. We show how random number generation from well known distributions is sufficient to develop posterior simulators for a wide variety of models.

### Notes

- 1 Throughout this chapter, we will use the term “firm” to refer to the cross-sectional unit of analysis. In practice, it could also be the individual or country, etc.
- 2 In this chapter we focus on production frontiers. However, by suitably redefining  $Y$  and  $X$ , the methods can be applied to cost frontiers.
- 3 In addition to standard microeconomic studies of firm efficiencies, stochastic frontier models have been applied to e.g. environmental issues and macroeconomic growth studies.
- 4 This error reflects the stochastic nature of the frontier and we shall conveniently denote it by “measurement error”. The treatment of measurement error is a crucial distinction between econometric and DEA methods. Most economic data sets are quite noisy and, hence, we feel including measurement error is important. DEA methods can be quite sensitive to outliers since they ignore measurement error. Furthermore, since the statistical framework for DEA methods is nonparametric, confidence intervals for parameter and efficiency estimates are very difficult to derive. However, econometric methods require the researcher to make more assumptions (e.g. about the error distribution) than do DEA methods. Recently, there has been some promising work on using the bootstrap with DEA methods which should lessen some of the

criticisms of DEA (see Simar and Wilson, 1998a, 1998b, and the references contained therein).

- 5 We use the terminology “sampling model” to denote the joint distribution of  $(y, z)$  given the parameters and shall base the likelihood function on the marginal distribution of  $y$  given the parameters.
- 6 In van den Broeck *et al.* (1994) and Koop *et al.* (1995), the Erlang distribution (i.e. the gamma distribution with fixed shape parameter, here chosen to be 1, 2, or 3) was used for inefficiency. The computational techniques necessary to work with Erlang distributions are simple extensions of those given in this section.
- 7 As shown in Fernández *et al.* (1997), the use of the full model with data augmentation also allows for the derivation of crucial theoretical results on the existence of the posterior distribution and moments.
- 8 The assumption that the inefficiencies are drawn from the exponential distribution with unknown common mean  $\lambda$  can be interpreted as a hierarchical prior for  $z_i$ . Alternatively, a classical econometrician would interpret this distributional assumption as part of the sampling model. This difference in interpretation highlights the fact that the division into prior and sampling model is to some extent arbitrary. See Fernández *et al.* (1997) for more discussion of this issue.
- 9 Note that we have not formally proven that the posterior mean and variance of  $\theta$  exist (although numerical evidence suggests that they do). Hence, we recommend using the posterior median and interquartile range of  $\theta$  to summarize properties of the posterior. Since  $0 \leq \tau_i \leq 1$ , we know that all posterior moments exist for the firm specific efficiencies.
- 10 Recent work on bootstrapping DEA frontiers is promising to surmount this problem and this procedure seems to be gaining some acceptance.
- 11 This is where the assumption that the  $w_{ij}$ s are 0–1 dummies is crucial.
- 12 The extension to a varying efficiency distribution as in (24.12) is trivial and proceeds along the lines of the previous model.
- 13 Of course, many of the issues which arise in the stochastic frontier model with panel data also arise in traditional panel data models. It is beyond the scope of the present chapter to attempt to summarize the huge literature on panel data. The reader is referred to Mátyás and Sevestre (1996) or Baltagi (1995) for an introduction to the broader panel data literature.
- 14 Note that this implies we now deviate from Assumption 3 in subsection 2.2.
- 15 It is worth noting that the classical econometric analysis assigns the status of most efficient firm to one particular firm and measures efficiency relative to this. The present Bayesian analysis also measures efficiency relative to the most efficient firm, but allows for uncertainty as to which that firm is.
- 16 This procedure can be computationally demanding since  $P(r_j^{\text{rel}} = 1 | y, x)$  and  $p(r_i^{\text{rel}} | y, x, r_j^{\text{rel}} = 1)$  must be calculated for every possible  $i$  and  $j$ . However, typically,  $P(r_j^{\text{rel}} = 1 | y, x)$  is appreciable (e.g.  $> 0.001$ ) for only a few firms and the rest can be ignored (see Koop *et al.*, 1997, p. 82).
- 17 Koop *et al.* (1999, 2000) also allow the frontier to shift over time and interpret such shifts as technical change. In such a framework, it is possible to decompose changes in output growth into components reflecting input change, technical change, and efficiency change. The ability of stochastic frontier models with panel data to calculate such decompositions is quite important in many practical applications. Also of interest are Baltagi and Griffin (1988) and Baltagi, Griffin, and Rich (1995), which develop a more general framework relating changes in the production function with technical change in a nonstochastic frontier panel data model.

## References

- Aigner, D., C.A.K. Lovell, and P. Schmidt (1977). Formulation and estimation of stochastic frontier production function models. *Journal of Econometrics* 6, 21–37.
- Baltagi, B. (1995). *Econometric Analysis of Panel Data*. New York: John Wiley and Sons.
- Baltagi, B., and J. Griffin (1988). A general index of technical change. *Journal of Political Economy* 90, 20–41.
- Baltagi, B., J. Griffin, and D. Rich (1995). The measurement of firm-specific indexes of technical change. *Review of Economics and Statistics* 77, 654–63.
- Bauer, P. (1990). Recent developments in the econometric estimation of frontiers. *Journal of Econometrics* 46, 39–56.
- van den Broeck, J., G. Koop, J. Osiewalski, and M.F.J. Steel (1994). Stochastic frontier models: A Bayesian perspective. *Journal of Econometrics* 61, 273–303.
- Casella, G., and E. George (1992). Explaining the Gibbs sampler. *The American Statistician* 46, 167–74.
- Chib, S. (1995). Marginal likelihood from the Gibbs output. *Journal of the American Statistical Association* 90, 1313–21.
- Chib, S., and E. Greenberg (1995). Understanding the Metropolis–Hastings algorithm. *The American Statistician* 49, 327–35.
- Färe, R., S. Grosskopf, and C.A.K. Lovell (1994). *Production Frontiers*. Cambridge: Cambridge University Press.
- Fernández, C., G. Koop, and M.F.J. Steel (2000). A Bayesian analysis of multiple output production frontiers. *Journal of Econometrics* 98, 47–9.
- Fernández, C., J. Osiewalski, and M.F.J. Steel (1997). On the use of panel data in stochastic frontier models with improper priors. *Journal of Econometrics* 79, 169–93.
- Gelfand, A., and D.K. Dey (1994). Bayesian model choice: Asymptotics and exact calculations. *Journal of the Royal Statistical Society, Series B* 56, 501–14.
- Geweke, J. (1989). Bayesian inference in econometric models using Monte Carlo integration. *Econometrica* 57, 1317–40.
- Geweke, J. (1999). Using simulation methods for Bayesian econometric models: Inference, development and communication (with discussion). *Econometric Reviews* 18, 1–126.
- Horrace, W., and P. Schmidt (1996). Confidence statements for efficiency estimates from stochastic frontiers. *Journal of Productivity Analysis* 7, 257–82.
- Jondrow, J., C.A.K. Lovell, I.S. Materov, and P. Schmidt (1982). On the estimation of technical inefficiency in the stochastic frontier production function model. *Journal of Econometrics* 19, 233–8.
- Judge, G., W. Griffiths, R.C. Hill, H. Lütkepohl, and T.-C. Lee (1985). *The Theory and Practice of Econometrics*, 2nd edn. New York: John Wiley and Sons, Ltd.
- Koop, G. (1994). Recent progress in applied Bayesian econometrics. *Journal of Economic Surveys* 8, 1–34.
- Koop, G., J. Osiewalski, and M.F.J. Steel (1994). Bayesian efficiency analysis with a flexible form: The AIM cost function. *Journal of Business and Economic Statistics* 12, 93–106.
- Koop, G., J. Osiewalski, and M.F.J. Steel (1997). Bayesian efficiency analysis through individual effects: Hospital cost frontiers. *Journal of Econometrics* 76, 77–105.
- Koop, G., J. Osiewalski, and M.F.J. Steel (1999). The components of output growth: A stochastic frontier analysis. *Oxford Bulletin of Economics and Statistics* 61, 455–87.
- Koop, G., J. Osiewalski, and M.F.J. Steel (2000). Modeling the sources of output growth in a panel of countries. *Journal of Business and Economic Statistics* 18, 284–99.
- Koop, G., M.F.J. Steel, and J. Osiewalski (1995). Posterior analysis of stochastic frontier models using Gibbs sampling. *Computational Statistics* 10, 353–73.

- Mátyás, L., and P. Sevestre (eds.) (1996). *The Econometrics of Panel Data*. Dordrecht: Kluwer Academic Publishers.
- Meeusen, W., and J. van den Broeck (1977). Efficiency estimation from Cobb–Douglas production functions with composed errors. *International Economic Review* 8, 435–44.
- Osiewalski, J., and M.F.J. Steel (1998). Numerical tools for the Bayesian analysis of stochastic frontier models. *Journal of Productivity Analysis* 10, 103–17.
- Poirier, D. (1995). *Intermediate Statistics and Econometrics: A Comparative Approach*. Cambridge, MA: The MIT Press.
- Ritter, C., and L. Simar (1997). Pitfalls of Normal–Gamma stochastic frontier models. *Journal of Productivity Analysis* 8, 167–82.
- Schmidt, P., and R. Sickles (1984). Production frontiers and panel data. *Journal of Business and Economic Statistics* 2, 367–74.
- Simar, L., and P.W. Wilson (1998a). A general methodology for bootstrapping in nonparametric frontier models. Manuscript.
- Simar, L., and P.W. Wilson (1998b). Sensitivity analysis of efficiency scores: How to bootstrap in nonparametric frontier models. *Management Science* 44, 49–61.
- Verdinelli, I., and L. Wasserman (1995). Computing Bayes factors using a generalization of the Savage–Dickey Density Ratio. *Journal of the American Statistical Association* 90, 614–18.

CHAPTER TWENTY-FIVE

# Parametric and Nonparametric Tests of Limited Domain and Ordered Hypotheses in Economics

*Esfandiar Maasoumi\**

## 1 INTRODUCTION

In this survey, technical and conceptual advances in testing multivariate linear and nonlinear inequality hypotheses in econometrics are summarized. This is discussed for economic applications in which either the null, or the alternative, or both hypotheses define more limited domains than the two-sided alternatives typically tested. The desired goal is increased power which is laudable given the endemic power problems of most of the classical asymptotic tests. The impediments are a lack of familiarity with implementation procedures, and characterization problems of distributions under some composite hypotheses.

Several empirically important cases are identified in which practical “one-sided” tests can be conducted by either the  $\bar{\chi}^2$ -distribution, or the union intersection mechanisms based on the Gaussian variate, or the increasingly feasible and popular resampling/simulation techniques. Point optimal testing and its derivatives find a natural medium here whenever unique characterization of the null distributions for the “least favorable” cases is not possible.

Most of the recent econometric literature in this area is parametric deriving from the multivariate extensions of the classical Gaussian means test with

ordered alternatives. Tests for variance components, random coefficients, overdispersion, heteroskedasticity, regime change, ARCH effects, curvature regularity conditions on flexible supply, demand, and other economic functions, are examples. But nonparametric tests for ordered relations between distributions, or their quantiles, or curvature regularity conditions on nonparametric economic relations, have witnessed rapid development and applications in economics and finance. We detail tests for Stochastic Dominance which indicate a major departure in the practice of empirical decision making in, so far, the areas of welfare and optimal financial strategy.

The additional information available when hypotheses can restrict attention to subspaces of the usual two-sided (unrestricted) hypotheses, can enhance the power of tests. Since good power is a rare commodity the interest in inequality restricted hypothesis tests has increased dramatically. In addition, the two-sided formulation is occasionally too vague to be of help when more sharply ordered alternatives are of interest. An example is the test of order relations (e.g. stochastic dominance) amongst investment strategies, or among income/welfare distributions. The two-sided formulation fails to distinguish between "equivalent" and "unrankable" cases.

In statistics, D.J. Bartholomew, H. Chernoff, V.J. Chacko, A. Kudo, and P.E. Nuesch are among the first to refine and extend the Neyman–Pearson testing procedure for one-sided alternatives, first in the one and then in the multivariate settings. For example, see Bartholomew (1959a, 1959b) and Kudô (1963). Later refinements and advances were obtained by Nuesch, Feder, Perlman, and others. At least in low dimensional cases, the power gains over the two-sided counterparts have been shown to be substantial, see Bartholomew (1959b), and Barlow *et al.* (1972). While Chernoff and Feder clarified the local nature of tests and gave some solutions when the true parameter value is "near" the boundaries of the hypotheses regions (see Wolak, 1989), Kudo, Nuesch, Perlman, and Shorack were among the first to develop the elements of the  $\bar{\chi}^2$ -distribution theory for the likelihood ratio and other classical tests. See Barlow *et al.* (1972) for references.

In econometrics, Gouriéroux, Holly, and Monfort (1980, 1982), heretofore GHM, are seminal contributions which introduced and extended this literature to linear and nonlinear econometric/regression models. The focus in GHM (1982) is on the following testing situation:

$$y = X\beta + u \quad (25.1)$$

$$R\beta \geq r, R : q \times K, q \leq K, \text{ the dimension of } \beta.$$

We wish to test

$$H_0 : R\beta = r, \dots \text{ vs } \dots H_1 : R\beta \geq r. \quad (25.2)$$

$u$  is assumed to be a Gaussian variate with zero-mean and finite variance  $\Omega$ . Gouriéroux *et al.* (1982) derive the Lagrange multiplier (LM)/Kuhn–Tucker (KT) test, as well as the likelihood ratio (LR) and the Wald (W) tests with known and

unknown covariance matrix,  $\Omega$ , of the regression errors. With known covariance the three tests are equivalent and distributed exactly as a  $\bar{\chi}^2$ -distribution. They note that the problem considered here is essentially equivalent to the following in the earlier statistical literature: Let there be  $T$  independent observations from a  $p$ -dimensional  $N(\mu, \Sigma)$ . Test

$$H_0 : \mu = 0, \text{ against the alternative}$$

$$H_1 : \mu_i \geq 0, \text{ all } i, \text{ with at least one strict inequality} \quad (25.3)$$

The LR test of this hypothesis has the  $\bar{\chi}^2$ -distribution which is a mixture of chi-squared distributions given by:

$$\sum_{j=0}^p w(p, j) \chi_{(j)}^2, \quad (25.4)$$

with  $\chi_{(0)}^2 = 1$  at the origin. The weights  $w(\cdot)$  are probabilities to be computed in a multivariate setting over the space of alternatives. This is one of the practical impediments in this area, inviting a variety of solutions which we shall touch upon. These include obtaining bounds, exact tests for low dimension cases, and resampling/Monte Carlo techniques.

When  $\Omega$  is unknown but depends on a finite set of parameters, GHM (1982) and others have shown that the same distribution theory applies asymptotically. GHM show that these tests are asymptotically equivalent and satisfy the usual inequality:  $W \geq LR \geq LM(KT)$ . We'll give the detailed form of these test statistics. In particular the LM version may be desirable as it can avoid the quadratic programming (QP) routine needed to obtain estimators under the inequality restrictions. We also point to routines that are readily available in Fortran and GAUSS (but alas not yet in the standard econometric software packages).

It should be noted that this simultaneous procedure competes with another approach based on the union intersection technique. In the latter, each univariate hypothesis is tested, with the decision being a rejection of the joint null if the least significant statistic is greater than the  $\alpha$ -critical level of a standard Gaussian variate. Consistency of such tests has been established. We will discuss examples of these alternatives. Also, the nonexistence generally of an optimal test in the multivariable case has led to consideration of point optimal testing, and tests that attempt to maximize power in the least favorable case, or on suitable "averages". This is similar to recent attempts to deal with power computation when alternatives depend on nuisance parameters. For example see King and Wu (1997) and their references.

In the case of nonlinear models and/or nonlinear inequality restrictions, GHM (1980) and Wolak (1989, 1991) discuss the distribution of the same  $\bar{\chi}^2$  tests, while Dufour considers modified classical tests. In this setting, however, there is another problem, as pointed out by Wolak (1989, 1991). When  $q \leq K$  there is generally no unique solution for the "true  $\beta$ " from  $R\beta = r$  (or its nonlinear counterpart). But

convention dictates that in this-type case of composite hypotheses, power be computed for the “least favorable” case which arises at the boundary  $R\beta = r$ . It then follows that the asymptotic distribution (when  $\Omega$  is consistently estimated in customary ways) cannot, in general, be uniquely characterized for the least favorable case. Sufficient conditions for a unique distribution are given in Wolak (1991) and will be discussed here. In the absence of these conditions, a “localized” version of the hypothesis is testable with the same  $\bar{\chi}^2$ -distribution.

All of the above developments are parametric. There is at least an old tradition for the nonparametric “two sample” testing of homogeneity between two distributions, often assumed to belong to the same family. Pearson type and Kolmogorov–Smirnov (KS) tests are prominent, as well as the Wilcoxon rank test. In the case of inequality or ordered hypotheses regarding relations between two unknown distributions, Anderson (1996) is an example of the modified Pearson tests based on relative cell frequencies, and Xu, Fisher, and Wilson (1995) is an example of quantile-based tests which incorporate the inequality information in the hypotheses and, hence lead to the use of  $\bar{\chi}^2$ -distribution theory. The multivariate versions of the KS test have been studied by McFadden (1989), Klecan, McFadden, and McFadden (1991), Kaur, Rao, and Singh (1994), and Maasoumi, Mills, and Zandvakili (1997). The union intersection alternative is also fully discussed in Davidson and Duclos (1998), representing a culmination of this line of development. The union intersection techniques do not exploit the inequality information and are expected to be less powerful. We discuss the main features of these alternatives.

In Section 2 we introduce the classical multivariate means problem and a general variant of it that makes it amenable to immediate application to very general econometric models in which an asymptotically normal estimator can be obtained. At this level of generality, one can treat very wide classes of processes, as well as linear and non-linear models, as described in Potscher and Prucha (1991a, 1991b). The linear model is given as an example, and the asymptotic distribution of the classical tests is described.

The next section describes the nonlinear models and the local nature of the hypothesis than can be tested. Section 4 is devoted to the nonparametric setting. Examples from economics and finance are cited throughout the chapter. Section 5 concludes.

## 2 THE GENERAL MULTIVARIATE PARAMETRIC PROBLEM

Consider the setting in (25.3) when  $\hat{\mu} = \mu + v$ , and  $v \sim N(0, \Omega)$ , is an available unrestricted estimator. Consider the restricted estimator  $\tilde{\mu}$  as the solution to the following quadratic programming (QP) problem:

$$\min_{\mu} (\hat{\mu} - \mu)' \Omega^{-1} (\hat{\mu} - \mu), \quad \text{subject to } \mu \geq 0. \quad (25.5)$$

Then the likelihood ratio (LR) test of the hypothesis in (25.3) is:

$$LR = \tilde{\mu}' \Omega^{-1} \tilde{\mu}. \quad (25.6)$$

Several researchers, for instance Kudô (1963) and Perlman (1969), have established the distribution of the LR statistic under the null as:

$$\begin{aligned} \text{Sup}_{\mu \geq 0} \cdot \text{pr}_{\mu, \Omega}(LR \geq c_\alpha) &= \text{pr}_{0, \Omega}(LR \geq c_\alpha) \\ &= \sum_{i=0}^p w(i, p, \Omega) \times \text{pr}(\chi_{(i)}^2 \geq c_\alpha) \end{aligned} \quad (25.7)$$

a weighted sum of chi-squared variates for an exact test of size  $\alpha$ . The weights  $w(i, \cdot)$  sum to unity and each is the probability of  $\hat{\mu}$  having  $i$  positive elements.

If the *null* hypothesis is one of *inequality* restrictions, a similar distribution theory applies. To see this, consider:

$$H_0 : \mu \geq 0 \quad \text{vs.} \quad H_1 : \mu \in R^p, \quad (25.8)$$

where  $\hat{\mu} = \mu + v$ , and  $v \sim N(0, \Omega)$ . Let  $\tilde{\mu}$  be the restricted estimator from the following QP problem:

$$D = \min_{\mu} (\hat{\mu} - \mu) \Omega^{-1} (\hat{\mu} - \mu) \quad \text{subject to } \mu \geq 0. \quad (25.9)$$

$D$  is the LR statistic for (25.8). Perlman (1969) showed that the power function is monotonic in this case. In view of this result, taking  $C_\alpha$  as the critical level of a test of size  $\alpha$ , we may use the same distribution theory as in (25.7) above except that the weight  $w(i, \cdot)$  will be the probability of  $\tilde{\mu}$  having exactly  $p - i$  positive elements.

There is a relatively extensive literature dealing with the computation of the weights  $w(\cdot)$ . Their computation requires evaluation of multivariate integrals which become tedious for  $p \geq 8$ . For example, Kudô (1963) provides exact expressions for  $p \leq 4$ , and Bohrer and Chow (1978) provide computational algorithms for  $p \leq 10$ . But these can be slow for large  $p$ . Kodde and Palm (1986) suggest an attractive bounds test solution which requires obtaining lower and upper bounds,  $c_l$  and  $c_u$ , to the critical value, as follows:

$$\alpha_l = \frac{1}{2} \text{pr}(\chi_{(1)}^2 \geq c_l), \text{ and}$$

$$\alpha_u = \frac{1}{2} \text{pr}(\chi_{(p-1)}^2 \geq c_u) + \frac{1}{2} \text{pr}(\chi_{(p)}^2 \geq c_u) \quad (25.10)$$

The null in (25.8) is rejected if  $D \geq c_u$ , but is inconclusive when  $c_l \leq D \leq c_u$ . Advances in Monte Carlo integration suggest resampling techniques may be used for large  $p$ , especially if the bounds test is inconclusive.

In the case of a single hypothesis ( $\mu_1 = 0$ ), the above test is the one-sided UMP test. In this situation:

$$\text{pr}(LR \geq c_\alpha) = \text{pr}(\frac{1}{2} \chi_{(0)}^2 + \frac{1}{2} \chi_{(1)}^2 \geq c_\alpha) = \alpha \quad (25.11)$$

The standard two-sided test would be based on the critical values  $c'_\alpha$  from a  $\chi^2_{(1)}$  distribution. But  $\text{pr}(\chi^2_{(1)} \geq c'_\alpha) = \alpha$  makes clear that  $c'_\alpha \geq c_\alpha$ , indicating the substantial power loss which was demonstrated by Bartholomew (1959a, 1959b) and others.

In the two dimension ( $p = 2$ ) case, under the null we have:

$$\text{pr}(LR \geq c_\alpha) = w(2, 0)\chi^2_{(0)} + w(2, 1)\chi^2_{(1)} + w(2, 2)\chi^2_{(2)} \quad (25.12)$$

where  $w(2, 0) = \text{pr}[LR = 0] = \text{pr}[\hat{\mu}_1 \leq 0, \hat{\mu}_2 \leq 0]$ ,  $w(2, 1) = \frac{1}{2}$ ,  $w(2, 2) = \frac{1}{2} - w(2, 0)$ . While difficult to establish analytically, the power gains over the standard case can be substantial in higher dimensions where UMP tests do not generally exist. See Kudô (1963) and GHM (1982).

## 2.1 LR, W, and LM tests

We give an account of the three classical tests in the context of the general linear regression model introduced in (25.1) above. We take  $R$  to be a  $(p \times K)$  known matrix of rank  $p \leq K$ . Consider three estimators of  $\beta$  under the exact linear restrictions, under inequality restrictions, and when  $\beta \in R^p$  (no restrictions). Denote these by  $\tilde{\beta}$ ,  $\tilde{\beta}$ , and  $\hat{\beta}$ , respectively. We note that  $\hat{\beta} = (X'\Omega^{-1}X)^{-1}(X'\Omega^{-1}y)$  is the ML (GLS) estimator here. Let  $(X'\Omega^{-1}X) = G$ , and consider the following optimization programs:

$$\max - (y - X\beta)' \Omega^{-1} (y - X\beta), \quad \text{subject to } R\beta \geq r, \quad (25.13)$$

and the same objective function but with equality restrictions. Denote by  $\tilde{\lambda}$  and  $\bar{\lambda}$  the Lagrange multipliers, respectively, of these two programs (conventionally,  $\lambda = 0$  for  $\hat{\beta}$ ). Then:

$$\tilde{\beta} = \hat{\beta} + G^{-1}R'\tilde{\lambda}/2, \quad \text{and} \quad \bar{\beta} = \hat{\beta} + G^{-1}R'\bar{\lambda}/2. \quad (25.14)$$

See GHM (1982). Employing these relations it is straightforward to show that the following three classical tests are identical:

$$\xi_{LR} = -2 \log LR = 2(\tilde{L} - \bar{L}), \quad (25.15)$$

where  $\tilde{L}$  and  $\bar{L}$  are the logarithms of the maxima of the respective likelihood functions;

$$\xi_{LM} = \min (\lambda - \bar{\lambda})' RG^{-1} R' (\lambda - \bar{\lambda})/4, \quad \text{subject to } \lambda \leq 0 \quad (25.16)$$

is the Kuhn-Tucker/Lagrange multiplier (LM) test computed at  $\tilde{\lambda}$ , and,

$$\xi_W = (R\tilde{\beta} - r)' [RG^{-1}R']^{-1} (R\tilde{\beta} - r) \quad (25.17)$$

is the Wald test. In order to utilize the classical results stated above for problems in (25.3), or (25.8), it is customary to note that the LR test in (25.15) above is identical to the LR test of the following problem:

$$\hat{\beta} = \beta + v$$

$$R\beta \geq r$$

$$v \sim N(0, G^{-1}) \quad (25.18)$$

For this problem  $\xi_{LR}$  is the optimum of the following QP problem:

$$\begin{aligned} & \max - (\beta - \hat{\beta})'G(\beta - \hat{\beta}) + (\bar{\beta} - \hat{\beta})'G(\bar{\beta} - \hat{\beta}) \\ & \text{subject to } R\beta \geq r. \end{aligned} \quad (25.19)$$

This is identical to the one-sided multivariate problem in (25.3). It also suggests that the context for applications can be very general indeed. All that is needed is normally distributed estimators,  $\hat{\beta}$ , which are then projected on to the cone defined by the inequality restrictions in order to obtain the restricted estimator  $\tilde{\beta}$ .

## 2.2 Asymptotically normal estimators

The assumption of normality can be relaxed for the situations that allow asymptotically normal estimators for  $\beta$ . This is because the inference theory developed for problems (25.3) or (25.8) is asymptotically valid for much broader classes of models and hypotheses. In fact, when consistent estimators of  $\Omega$  are available and, in (25.18),  $v$  has the stated distribution *asymptotically*, an asymptotically exact test of size  $\alpha$  is based on the same  $\bar{\chi}^2$ -distribution given above. To obtain this result one needs to replace  $\Omega$  in the optimization problems with its corresponding consistent estimator. This is routinely possible when  $\Omega$  is a continuous function of a finite set of parameters other than  $\beta$ .

The three tests are not identical in this situation, of course, but have the same asymptotic distribution. Furthermore, the usual inequality, viz.  $\xi_W \geq \xi_{LR} \geq \xi_{LM}$  is still valid, see GHM (1982). Often the test which can avoid the QP problem is preferred, which means the LM test for the null of equality of the restrictions, and the Wald test when the null is one of inequality and the alternative is unrestricted. But much recent evidence, as well as invariance arguments, suggest that the LR test be used.

In the general linear regression models with linear and/or nonlinear inequality restrictions, other approaches are available. Kodde and Palm (1986, 1987), Dufour (1989), Dufour and Khalaf (1995), and Stewart (1997) are examples of theoretical and empirical applications in economics and finance. Dufour (1989) is an alternative “conservative” bounds test for the following type situation:

$$H_0 : R\beta \in \Gamma_0 \text{ against } H_1 : \beta \in \Gamma_1, \quad (25.20)$$

where  $\Gamma_0$  and  $\Gamma_1$  are non-empty subsets, respectively of  $\mathbf{R}^p$  and  $\mathbf{R}^K$ . This also allows a consideration of such cases as  $h(\mathbf{R}\beta) = 0$ , or  $h(\mathbf{R}\beta) \geq 0$ . Dufour (1989) suggests a generalization of the well known, two-sided F-test in this situation as follows:

$$\text{pr} \left[ \frac{SS_0 - SS_1}{SS_1} \geq \frac{p}{T - K} F_\alpha(p, T - K) \right] \leq \alpha, \quad (25.21)$$

where there are  $T$  observations from which  $SS_i$ ,  $i = 0, 1$ , are calculated as residual sums of squares under the null and the alternative, respectively. Thus the traditional  $p$ -values will be upper-bounds for the true values and offer a conservative bounds testing strategy. Dufour (1989), Dufour and Khalaf (1993) and Stewart (1997), *inter alia*, consider "liberal bounds", and extensions to multivariate/simultaneous equations models and nonlinear inequality restrictions. Applications to demand functions and negativity constraints on the substitution matrix, as well as tests of nonlinear nulls in the CAPM models show size and power improvements over the traditional asymptotic tests. The latter are known for their tendency to overreject in any case. Stewart (1997) considers the performance of the standard LR, the Kodde and Palm (1986) bounds for the  $\bar{\chi}^2$ -distribution, and the Dufour-type bound test of negativity of the substitution matrix for the demand data for Germany and Holland. Stewart looks at, among other things, the hypothesis of negativity against an unrestricted alternative, and the null of negativity when symmetry and homogeneity are maintained. It appears that, while the Dufour test did well in most cases, certainly reversing the conclusions of the traditional LR test (which rejects everything!), the Kodde and Palm bounds test does consistently well when the conservative bounds test was not informative (with  $\alpha = 1$ ). Both the lower and upper bounds for the  $\bar{\chi}^2$ -squared distribution are available, while the "liberal"/lower bounds for the Dufour adjustment are not in this case.

### 3 NONLINEAR MODELS AND NONLINEAR INEQUALITY RESTRICTIONS

Wolak (1989, 1991) gives a general account of this topic. He considers the general formulation in (25.18) with nonlinear restrictions. Specifically, consider the following problem:

$$\begin{aligned} \hat{\beta} &= \beta + v \\ h(\beta) &\geq 0 \\ v &\sim_a N(0, \Psi) \end{aligned} \quad (25.22)$$

where  $h(\cdot)$  is a smooth vector function of dimension  $p$  with a derivative matrix denoted by  $H(\cdot)$ . We wish to test

$$H_0 : h(\beta) \geq 0, \quad \text{vs.} \quad H_1 : \beta \in \mathbf{R}^K. \quad (25.23)$$

This is very general since model classes that allow for estimation results given in (25.22) are very broad indeed. As the results in Potscher and Prucha (1991a, 1991b) indicate, many nonlinear dynamic processes in econometrics permit consistent and asymptotically normal estimators under regularity conditions.

In general an asymptotically exact size test of the null in (25.23) is not possible without a localization to some suitable neighborhood of the parameter space. To see this, let  $h(\beta^0) = 0$  define  $\beta^0$ , and  $H(\beta^0)$ , and  $I(\beta^0)$  the evaluations of  $\partial h / \partial \beta = H(\beta)$ , and the information matrix, respectively. Let

$$C = \{\beta \mid h(\beta) \geq 0, \beta \in \mathbb{R}^K\} \quad (25.24)$$

and  $N_{\delta_r}(\beta^0)$  as a  $\delta_r$ -neighborhood with  $\delta_r = O(T^{-\frac{1}{2}})$ . It is known that the global hypotheses of the type in (25.23)–(25.24) do not generally permit large sample approximations to the power function for fixed alternative hypotheses for nonlinear multivariate equality restrictions. See Wolak (1989). If we localize the null in (25.23) to only  $\beta \in N_{\delta_r}(\beta^0)$ , then *asymptotically* exact size tests are available and are as given by the appropriately defined chi-bar distribution. In fact, we would be testing whether  $\beta \in \{\text{cone of tangents of } C \text{ at } \beta^0\}$ , where  $\Psi = H(\beta^0)I(\beta^0)^{-1}H(\beta^0)$ , in (25.22).

In order to appreciate the issues, we note that the asymptotic distribution of the test depends on  $\Psi$ , which in turn depends on  $H(\cdot)$  and  $I(\cdot)$ . But the latter generally vary with  $\beta^0$ . Also, we note that  $h(\beta^0) = 0$  does not have a unique solution unless it is linear and of rank  $K = p$ . Thus the case of  $H_0 : \beta \geq 0$  does not have a problem. The case  $R\beta \geq 0$ , will present a problem in nonlinear models when  $\text{rank}(R) < K$  since  $I(\beta)$  will generally depend on the  $K - p$  free parameters in  $\beta$ . It must be appreciated that this is specially serious since inequalities define composite hypotheses which force a consideration of power over regions. Optimal tests do not exist in multivariate situations. Thus other conventions must be developed for test selection. One method is to consider power in the least favorable case. Another is to maximize power at given points that are known to be “desirable”, leading to point optimal testing. Closely related, and since such points may be difficult to select a priori, are tests that maximize mean power over sets of desirable points. See King and Wu (1997) for discussion. In the instant case, the “least favorable” cases are all those defined by  $h(\beta^0) = 0$ . Hence the indeterminacy of the asymptotic power function.

From this point on our discussion pertains to the “local” test whenever the estimates of  $\Psi$  cannot converge to a unique  $H(\beta^0)I(\beta^0)^{-1}H(\beta^0) = \Psi$ .

Let  $x_t = (x_{t1}, x_{t2}, \dots, x_{tn})'$ ,  $t = 1, \dots, T$ , be a realization of a random variable in  $\mathbb{R}^n$ , with a density function  $f(x_t, \beta)$  which is continuous in  $\beta$  for all  $x_t$ . We assume a compact subspace of  $\mathbb{R}^n$  contains  $\beta$ ,  $h(\cdot)$  is continuous with continuous partial derivatives  $\partial h_i(\beta) / \partial \beta_j = H_{ij}$ , defining the  $p \times K$  matrix  $H(\beta)$  that is assumed to have full rank  $p \leq K$  at an interior point  $\beta^0$  such that  $h(\beta^0) = 0$ . Finally, let  $\beta_T^0$  denote the “true” value under the local hypothesis, then,

$$\beta_T^0 \in C_T = \{\beta \mid h(\beta) \geq 0, \beta \in N_{\delta_r}(\beta^0)\} \quad \text{for all } T$$

and  $(\beta_T^0 - \beta^0) = o(1)$  and  $T^{\frac{1}{2}}(\beta_T^0 - \beta^0) = O(1)$ . Let  $x$  represent  $T$  random observations from  $f(x_t, \beta)$ , and the loglikelihood function given below:

$$L(\beta) = L(x, \beta) = \sum_{t=1}^T \ln(f(x_t, \beta)). \quad (25.25)$$

Following GHM (1982), again we consider the three estimators of  $(\beta, \lambda)$ , obtained under the inequality constraints, equality constraints, and no constraints as  $(\tilde{\beta}, \tilde{\lambda})$ ,  $(\bar{\beta}, \bar{\lambda})$ , and  $(\hat{\beta}, \hat{\lambda} = 0)$ , respectively. It can be verified that (see Wolak, 1989) the three tests LR, Wald, and LM are asymptotically equivalent and have the distribution given earlier. They are computed as follows:

$$\xi_{LR} = 2[L(\hat{\beta}) - L(\tilde{\beta})] \quad (25.26)$$

$$\xi_W = T(h(\tilde{\beta}) - h(\hat{\beta}))'[H(\hat{\beta})I(\beta^0)^{-1}H(\hat{\beta})'](h(\tilde{\beta}) - h(\hat{\beta})) \quad (25.27)$$

$$\xi'_W = T(\tilde{\beta} - \hat{\beta})'I(\beta^0)(\tilde{\beta} - \hat{\beta}) \quad (25.28)$$

$$\xi_{LM} = T\tilde{\lambda}'H(\tilde{\beta})I(\beta^0)^{-1}H(\tilde{\beta})'\tilde{\lambda} \quad (25.29)$$

where  $I(\beta^0)$  is the value of the information matrix,  $\lim_{T \rightarrow \infty} T^{-1} E_{\beta^0}[-\partial^2 L / \partial \beta \partial \beta']$ , at  $\beta^0$ , and (25.27)–(25.28) are two asymptotically equivalent ways of computing the Wald test. This testifies to its lack of invariance which has been widely appreciated in econometrics. The above results also benefit from the well known asymptotic approximations:

$$h(\beta^0) \simeq H(\beta^0)(\beta - \beta^0), \quad \text{and} \quad H(\beta^0) - H(\beta) \simeq 0,$$

which hold for all of the three estimators of  $\beta$ . As Wolak (1989) shows, these statistics are asymptotically equivalent to the generalized distance statistic  $D$  introduced in Kodde and Palm (1986):

$$D = \min_{\beta} T(\hat{\beta} - \beta)'I(\beta^0)(\hat{\beta} - \beta), \quad (25.30)$$

$$\text{subject to } H(\beta^0)(\beta - \beta^0) \geq 0, \quad \text{and} \quad \beta \in N_{\delta_r}(\beta^0).$$

For local  $\beta$  defined above, all these statistics have the same  $\chi^2$ -distribution given earlier. Kodde and Palm (1987) employ this statistic for an empirical test of the negativity of the substitution matrix of demand systems. They find that it outperforms the two-sided asymptotic LR test. Their bounds also appear to deal with the related problem of overrejection when nominal significance levels are used with other classical tests against the two-sided alternatives. Gouriéroux *et al.* (1982) give the popular artificial regression method of deriving the LM test. In the same general context, Wolak (1989) specializes the above tests to a test of joint nonlinear inequality and equality restrictions.

With the advent of cheap computing and Monte Carlo integration in high dimensions, the above tests are quite accessible. Certainly, the critical values from the bounds procedures deserve to be incorporated in standard econometric routines, as well as the exact bounds for low dimensional cases ( $p \leq 8$ ). The power gains justify the extra effort.

## 4 NONPARAMETRIC TESTS OF INEQUALITY RESTRICTIONS

All of the above models and hypotheses were concerned with comparing means and/or variance parameters of either known or asymptotically normal distributions. We may not know the distributions and/or be interested in comparing more general characteristics than the first few moments, and the distributions being compared may not be from the same family. All of these situations require a nonparametric development that can also deal with ordered hypotheses.

Order relations between distributions present one of the most important and exciting areas of development in economics and finance. These include stochastic dominance relations which in turn include Lorenz dominance, and such others as likelihood and uniform orders. Below we focus on the example of stochastic dominance (SD) of various orders. An account of the definitions and tests is first given, followed by some applications.

### 4.1 Tests for stochastic dominance

In the area of income distributions and tax analysis, it is important to look at Lorenz curves and similar comparisons. In practice, a finite number of ordinates of the desired curves or functions are compared. These ordinates are typically represented by quantiles and/or conditional interval means. Thus, the distribution theory of the proposed tests are typically derived from the existing asymptotic theory for ordered statistics or conditional means and variances. A most up-to-date outline of the required asymptotic theory is Davidson and Duclos (1998). To control for the size of a sequence of tests at several points the union intersection (UI) and Studentized Maximum Modulus technique for multiple comparisons is generally favored in this area. In this line of inquiry the inequality nature of the order relations is not explicitly utilized in the manner described above for parametric tests. Therefore, procedures that do so may offer power gain. Some alternatives to these multiple comparison techniques have been suggested, which are typically based on Wald-type joint tests of *equality* of the same ordinates; e.g. see Anderson (1996). These alternatives are somewhat problematic since their implicit null and alternative hypotheses are typically not a satisfactory representation of the *inequality* (order) relations that need to be tested. For instance, Xu *et al.* (1995) take proper account of the inequality nature of such hypotheses and adapt econometric tests for inequality restrictions to testing for FSD and SSD, and to GL dominance, respectively. Their tests follow the  $\bar{\chi}^2$  theory outlined earlier.

McFadden (1989) and Klecan *et al.* (1991) have proposed tests of first- and second-order “maximality” for stochastic dominance which are extensions of the

Kolmogorov–Smirnov statistic. McFadden (1989) assumes iid observations and independent variates, allowing him to derive the asymptotic distribution of his test, in general, and its exact distribution in some cases. He provides a Fortran and a GAUSS program for computing his tests. Klecan *et al.* generalize this earlier test by allowing for weak dependence in the processes and replace independence with exchangeability. They demonstrate with an application for ranking investment portfolios. The asymptotic distribution of these tests cannot be fully characterized, however, prompting Klecan *et al.* to propose Monte Carlo methods for evaluating critical levels. Similarly, Maasoumi *et al.* (1997) propose bootstrap-KS tests with several empirical applications. In the following subsections some definitions and results are summarized which help to describe these tests.

## 4.2 Definitions and tests

Let  $X$  and  $Y$  be two income variables at either two different points in time, before and after taxes, or for different regions or countries. Let  $X_1, X_2, \dots, X_n$  be  $n$  not necessarily iid observations on  $X$ , and  $Y_1, Y_2, \dots, Y_m$  be similar observations on  $Y$ . Let  $U_1$  denote the class of all utility functions  $u$  such that  $u' \geq 0$ , (increasing). Also, let  $U_2$  denote the subset of all utility functions in  $U_1$  for which  $u'' \leq 0$  (strict concavity), and  $U_3$  denote a subset of  $U_2$  for which  $u''' \geq 0$ . Let  $X_{(i)}$  and  $Y_{(i)}$  denote the  $i$ th order statistics, and assume  $F(x)$  and  $G(x)$  are continuous and monotonic cumulative distribution functions (cdfs) of  $X$  and  $Y$ , respectively. Let the quantile functions  $X(p)$  and  $Y(p)$  be defined by, for example,  $Y(p) = \inf\{y : F(y) \geq p\}$ .

**Proposition 1.**  $X$  first-order stochastic dominates  $Y$ , denoted  $X \text{ FSD } Y$ , if and only if any one of the following equivalent conditions holds:

1.  $E[u(X)] \geq E[u(Y)]$  for all  $u \in U_1$ , with strict inequality for some  $u$ . This is the classical definition.
2.  $F(x) \leq G(x)$  for all  $x$  in the support of  $X$ , with strict inequality for some  $x$  (e.g. see McFadden, 1989).
3.  $X(p) \geq Y(p)$  for all  $0 \leq p \leq 1$ , with strict inequality for some  $p$  (e.g. see Xu *et al.*, 1995).

**Proposition 2.**  $X$  second-order stochastic dominates  $Y$ , denoted  $X \text{ SSD } Y$ , if and only if any of the following equivalent conditions holds:

1.  $E[u(X)] \geq E[u(Y)]$  for all  $u \in U_2$ , with strict inequality for some  $u$ .
2.  $\int_{-\infty}^x F(t)dt \leq \int_{-\infty}^x G(t)dt$  for all  $x$  in the support of  $X$  and  $Y$ , with strict inequality for some  $x$ .
3.  $\Phi_X(p) = \int_0^p X(t)dt \geq \Phi_Y(p) = \int_0^p Y(t)dt$ , for all  $0 \leq p \leq 1$ , with strict inequality for some value(s)  $p$ .

Weaker versions of these relations drop the requirement of strict inequality at some point. When either Lorenz or Generalized Lorenz Curves of two distributions

cross, unambiguous ranking by FSD and SSD may not be possible. Shorrocks and Foster (1987) show that the addition of a “transfer sensitivity” requirement leads to third-order stochastic dominance (TSD) ranking of income distributions. This requirement is stronger than the Pigou–Dalton principle of transfers since it makes regressive transfers less desirable at lower income levels. TSD is defined as follows:

**Proposition 3.**  $X$  third-order stochastic dominates  $Y$ , denoted  $X \text{TSD } Y$ , if any of the following equivalent conditions holds:

1.  $E[u(X)] \geq E[u(Y)]$  for all  $u \in U_3$ , with strict inequality for some  $u$ .
2.  $\int_{-\infty}^x \int_{-\infty}^v [F(t) - G(t)] dt dv \leq 0$ , for all  $x$  in the support, with strict inequality for some  $x$ , with the end-point condition:

$$\int_{-\infty}^{+\infty} [F(t) - G(t)] dt \leq 0.$$

3. When  $E[X] = E[Y]$ ,  $X \text{TSD } Y$  iff  $\bar{\sigma}_x^2(q_i) \leq \bar{\sigma}_y^2(q_i)$ , for all Lorenz curve crossing points  $q_i$ ,  $i = 1, 2, \dots, (n+1)$ ; where  $\bar{\sigma}_x^2(q_i)$  denotes the “cumulative variance” for incomes upto the  $i$ th crossing point. See Davies and Hoy (1995).

When  $n = 1$ , Shorrocks and Foster (1987) show that  $X \text{TSD } Y$  if (i) the Lorenz curve of  $X$  cuts that of  $Y$  from above, and (ii)  $\text{var}(X) \leq \text{var}(Y)$ . This situation seemingly revives the coefficient of variation as a useful statistical index for ranking distributions. But a distinction is needed between the well known (unconditional) coefficient of variation for a distribution, on the one hand, and the sequence of several conditional coefficients of variation involved in the TSD.

The tests of FSD and SSD are based on empirical evaluations of conditions (2) or (3) in the above definitions. Mounting tests on conditions (3) typically relies on the fact that quantiles are consistently estimated by the corresponding order statistics at a finite number of sample points. Mounting tests on conditions (2) requires empirical cdfs and comparisons at a finite number of observed ordinates. Also, from Shorrocks (1983) it is clear that condition (3) of SSD is equivalent to the requirement of generalized Lorenz (GL) dominance. FSD implies SSD.

The Lorenz and the generalized Lorenz curves are, respectively, defined by:

$L(p) = (1/\mu) \int_0^p Y(u) du$ , and  $GL(p) = \mu L(p) = \int Y(u) du$ , with  $GL(0) = 0$ , and  $GL(1) = \mu$ ; see Shorrocks (1983).

It is customary to consider  $K$  points on the  $L$  (or  $GL$  or the support) curves for empirical evaluation with  $0 < p_1 < p_2 < \dots < p_K = 1$ , and  $p_i = i/K$ . Denote the corresponding quantiles by  $Y(p_i)$ , and the conditional moments  $\gamma_i = E(Y | Y \leq Y(p_i))$ , and  $\bar{\sigma}_i^2 = E\{(Y - \gamma_i)^2 | Y \leq Y(p_i)\}$ . The vector of  $GL$  ordinates is given by  $\eta = (p_1 \bar{\sigma}_1^2, p_2 \bar{\sigma}_2^2, \dots, p_K \bar{\sigma}_K^2)'$ . See Xu *et al.* (1995) who adopt the  $\bar{\chi}^2$  approach described above to test quantile conditions (3) of FSD and SSD. A short description follows:

Consider the random sequence  $\{Z_t\} = \{X_t, Y_t\}'$ , a stationary  $\phi$ -mixing sequence of random vectors on a probability space  $(\Omega, \mathcal{R}, P)$ . Similarly, denote the stacked

vector of GL ordinates for the two variables as  $\eta^Z = (\eta^X, \eta^Y)',$  and the stacked vector of quantiles of the two variables by  $q^Z = (q^X, q^Y)',$  where  $q^X = (X(p_1), X(p_2), \dots, X(p_K))'$ , and similarly for Y. In order to utilize the general theory given for the  $\bar{\chi}^2$ -distribution, three ingredients are required. One is to show that the various hypotheses of interest in this context are representable as in (25.23) above. This is possible and simple. The second is to verify if and when the unrestricted estimators of the  $\eta$  and  $q$  functions satisfy the asymptotic representation given in (25.22). This is possible under conditions on the processes and their relationships, as we will summarize shortly. The third is to be able to empirically implement the  $\bar{\chi}$  statistics that ensue. In this last step, resampling techniques are and will become even more prominent.

To see that hypotheses of interest are suitably representable, we note that for the case of conditions (3) of FSD and SSD, the testing problem is the following:

$H_0 : h_K(q^Z) \geq 0$  against  $H_1 : h_K(q^Z) \not\geq 0,$  where  $h_K(q^Z) = [I_K : -I_K]q^Z = I^*q^Z,$  say, for FSD, and  $h_K(q^Z) = BI^* \times q^Z,$  for the test of SSD, where,  $B = (B_{ij}), B_{ij} = 1, i \geq j, B_{ij} = 0,$  otherwise, is the "summation" matrix which obtains the successive cumulated quantile ( $\Phi$ ) and other functions.

Tests for GL dominance (SSD) which are based on the ordinate vector  $\eta^Z$  are also of the "linear inequality" form and require  $h(\eta^Z) = I^*\eta^Z.$

Sen (1972) gives a good account of the conditions under which sample quantiles are asymptotically normally distributed. Davidson and Duclos (1998) provide the most general treatment of the asymptotic normality of the nonparametric sample estimators of the ordinates in  $\eta.$  In both cases the asymptotic variance matrix,  $\Psi,$  noted in the general setup (25.22) is derived. What is needed is to appropriately replace  $R$  in the formulations of Kodde and Palm (1986), or Gouriéroux *et al.* (1982), and to implement the procedure with consistent estimates of  $\Omega$  in  $\Psi = R\Omega R'.$

For sample order statistics,  $\hat{q}_T^Z$ , it is well known that, if X and Y are independent,

$$\begin{aligned}\sqrt{T}(\hat{q}^Z - q^Z) &\xrightarrow{d} N(0, \Omega) \\ \Omega &= G^{-1}VG'^{-1} \\ G &= \text{diag}[f_x(X(p_i)), \dots; f_y(Y(p_i)), \dots], i = 1, \dots, K \\ V &= \lim_{T \rightarrow \infty} E(gg'), g = T^{-1}(\mathfrak{F}_x \mathfrak{F}_y'), \\ \mathfrak{F}_x &= [F(X(p_1)) - p_1, \dots, F(X(p_K)) - p_K], \mathfrak{F}_y \text{ similarly defined.}\end{aligned}$$

As is generally appreciated, these density components are notoriously difficult to estimate. Kernel density methods can be used, as can Newey-West type robust estimators. But it is desirable to obtain bootstrap estimates based on block bootstrap and/or iterated bootstrap techniques. These are equally accessible computationally, but may perform much better in smaller samples and for larger numbers of ordinates  $K.$  Xu *et al.* (1995) demonstrate with an application to the hypothesis of term premia based on one- and two-month US Treasury bills. This application was based on the Kodde and Palm (1986) critical bounds and encountered some

realizations in the inconclusive region. Xu *et al.* (1995) employ Monte Carlo simulations to obtain the exact critical levels in those cases.

Sample analogs of  $\eta$  and similar functions for testing any stochastic order also have asymptotically normal distributions. Davidson and Duclos (1998) exploit the following interesting result which translates conditions (3) of the FSD and SSD into inequality restrictions among the members of the  $\eta$  functions defined above:

Let  $D_X^1(x) = F_X(x)$ , and  $D_Y(y) = F_Y(y)$ ; then,

$$D_i^s(x) = \int_0^x D^{s-1}(u)du = \frac{1}{(s-1)!} \int_0^x (x-u)^{s-1}dF(u), \text{ for any } s \geq 2.$$

This last equality clearly shows that tests of any order stochastic dominance can be based on the conditional moments estimated at a suitable finite number of  $K$  ordinates as defined above. For instance, third order SD ( $s = 3$ ) is seen to depend on the conditional/cumulative variance. Also, since poverty measures are often defined over lower subsets of the domain such that  $x \leq$  poverty line, dominance relations over poverty measures can also be tested in the same fashion. Using empirical distribution functions, Davidson and Duclos (1998) demonstrate with an example from the panels for six countries in the Luxembourg study. It should be appreciated, however, that these tests do not exploit the inequality nature of the alternative hypotheses. The union intersection method determines the critical level of the inference process here. The cases of unrankable distributions include both "equivalence" and crossing (non-dominant) distributions. A usual asymptotic  $\chi^2$  test will have power in both directions. In order to improve upon this, therefore, one must employ the  $\bar{\chi}^2$ -distribution technique.

Similarly, Kaur *et al.* (1994) propose a test for condition (2) of SSD when iid observations are assumed for independent prospects  $X$  and  $Y$ . Their null hypothesis is condition (2) of SSD for each  $x$  against the alternative of strict violation of the same condition for all  $x$ . The test of SSD then requires an appeal to a union intersection technique which results in a test procedure with maximum asymptotic size of  $\alpha$  if the test statistic at each  $x$  is compared with the critical value  $Z_\alpha$  of the standard normal distribution. They showed their test is consistent. One rejects the null of dominance if any negative distances at the  $K$  ordinates is significant.

In contrast, McFadden (1989), and Klecan *et al.* (1991) test for dominance jointly for all  $x$ . McFadden's analysis of the multivariate Kolmogorov–Smirnov type test is developed for a set of variables and requires a definition of "maximal" sets, as follows:

**Definition 1.** Let  $\mathcal{A}E = \{X_1, X_2, \dots, X_K\}$  denote a set of  $K$  distinct random variables. Let  $F_k$  denote the cdf of the  $k$ th variable. The set  $\mathcal{A}E$  is first- (second-)order maximal if no variable in  $\mathcal{A}E$  is first- (second-)order weakly dominated by another.

Let  $X_{\cdot n} = (x_{1n}, x_{2n}, \dots, x_{Kn})$ ,  $n = 1, 2, \dots, N$ , be the observed data. We assume  $X_{\cdot n}$  is strictly stationary and  $\alpha$ -mixing. As in Klecan *et al.*, we also assume  $F_i(X_i)$ ,  $i = 1, 2, \dots, K$  are exchangeable random variables, so that our resampling estimates of the test statistics converge appropriately. This is less demanding than the assumption of independence which is not realistic in many applications (as in before and after tax scenarios). We also assume  $F_k$  is unknown and estimated by the empirical distribution function  $F_{kN}(X_k)$ . Finally, we adopt Klecan *et al.*'s mathematical regularity conditions pertaining to von Neumann–Morgenstern (VNM) utility functions that generally underlie the expected utility maximization paradigm. The following theorem defines the tests and the hypotheses being tested:

**Lemma** Given the mathematical regularity conditions;

1. The variables in  $\mathcal{A}$  are first-order stochastically maximal; i.e.

$$d = \min_{i \neq j} \max_x [F_i(x) - F_j(x)] > 0,$$

if and only if for each  $i$  and  $j$ , there exists a continuous increasing function  $u$  such that  $Eu(X_i) > Eu(X_j)$ .

2. The variables in  $\mathcal{A}$  are second-order stochastically maximal; i.e.

$$S = \min_{i \neq j} \max_x \int_{-\infty}^x [F_i(\mu) - F_j(\mu)] d\mu > 0,$$

if and only if for each  $i$  and  $j$ , there exists a continuous increasing and strictly concave function  $u$  such that  $Eu(X_i) > Eu(X_j)$ .

3. Assuming, (i) the stochastic process  $X_{\cdot n}$ ,  $n = 1, 2, \dots$ , to be strictly stationary and  $\alpha$ -mixing with  $\alpha(j) = O(j^{-\delta})$ , for some  $\delta > 1$ , and (ii) the variables in the set are exchangeable (relaxing independence in McFadden, 1989):  $d_{2N} \rightarrow d$ , and  $S_{2N} \rightarrow S$ , where  $d_{2N}$  and  $S_{2N}$  are the empirical test statistics defined as:

$$d_{2N} = \min_{i \neq j} \max_x [F_{iN}(x) - F_{jN}(x)]$$

and

$$S_{2N} = \min_{i \neq j} \max_x \int_0^x [F_{iN}(\mu) - F_{jN}(\mu)] d\mu$$

**Proof 1.** See Theorems 1 and 5 of Klecan *et al.* (1991).

The null hypothesis tested by these two statistics is that, respectively,  $\mathcal{A}$  is *not* first- (second-)order maximal – i.e.  $X_i$  FSD(SSD)  $X_j$  for some  $i$  and  $j$ . We reject the null when the statistics are positive and large. Since the null hypothesis in each case is composite, power is conventionally determined in the least favorable case of identical marginals  $F_i = F_j$ . As is shown in Kaur *et al.* (1994) and Klecan *et al.*

(1991), when  $X$  and  $Y$  are independent, tests based on  $d_{2N}$  and  $S_{2N}$  are consistent. Furthermore, the asymptotic distribution of these statistics are non-degenerate in the least favorable case, being Gaussian (see Klecan *et al.*, 1991, Theorems 6–7).

As is pointed out by Klecan *et al.* (1991), for non-independent variables, the statistic  $S_{2N}$  has, in general, neither a tractable distribution, nor an asymptotic distribution for which there are convenient computational approximations. The situation for  $d_{2N}$  is similar except for some special cases – see Durbin (1973, 1985), and McFadden (1989) who assume iid observations (not crucial), and independent variables in  $\mathcal{A}E$  (consequential). Unequal sample sizes may be handled as in Kaur *et al.* (1994).

Klecan *et al.* (1991) suggest Monte Carlo procedures for computing the significance levels of these tests. This forces a dependence on an assumed parametric distribution for generating MC iterations, but is otherwise quite appealing for very large iterations. Maasoumi *et al.* (1997) employ the bootstrap method to obtain the empirical distributions of the test statistics and of  $p$ -values. Pilot studies show that their computations obtain similar results to the algorithm proposed in Klecan *et al.* (1991).

In the bootstrap procedure we compute  $d_{2N}$  and  $S_{2N}$  for a finite number  $K$  of the income ordinates. This requires a computation of sample frequencies, cdfs and sums of cdfs, as well as the differences of the last two quantities at all the  $K$  points. Bootstrap samples are generated from which empirical distributions of the differences, of the  $d_{2N}$  and  $S_{2N}$  statistics, and their bootstrap confidence intervals are determined. The bootstrap probability of these statistics being positive and/or falling outside these intervals leads to rejection of the hypotheses. Maasoumi *et al.* (1997) demonstrate by several applications to the US income distributions based on the Current Population Survey (CPS) and the panel data from the Michigan study. In contrast to the sometimes confusing picture drawn by comparisons based on inequality indices, they find frequent SSD relations, including between population subgroups, that suggest a “welfare” deterioration in the 1980s compared to the previous two decades.

## 5 CONCLUSION

Taking the one-sided nature of some linear and nonlinear hypotheses is both desirable and practical. It can improve power and lead to the improved computation of the critical levels. A  $\bar{\chi}^2$  and a multivariate KS testing strategy were described and contrasted with some alternatives, either the less powerful two-sided methods, or the union intersection procedures. The latter deserves to be studied further in comparison to the methods that are expected to have better power. Computational issues involve having to solve QP problems to obtain inequality restricted estimators, and numerical techniques for computation of the weights in the  $\bar{\chi}$  statistic. Bounds tests for the latter are available and may be sufficient in many cases.

Applications in the parametric/semiparametric, and the nonparametric testing area have been cited. They tend to occur in substantive attempts at empirical evaluation and incorporation of economic theories.

## Note

- \* Comments from the editor and an anonymous referee helped improve this chapter. A more extensive bibliography and discussion is contained in the preliminary version of this chapter which is available upon request.

## References

- Anderson, G.J. (1996). Nonparametric tests of stochastic dominance in income distributions. *Econometrica* 64, 1183–93.
- Barlow, R.E., D.J. Bartholomew, J.N. Bremner, and H.D. Brunk (1972). *Statistical Inference under Order Restrictions: The Theory and Applications of Isotonic Regression*. New York: John Wiley and Sons.
- Bartholomew, D.J. (1959a). A test of homogeneity for ordered alternatives. *Biometrika* 46, 36–48.
- Bartholomew, D.J. (1959b). A test of homogeneity for ordered alternatives. *Biometrika* 46, 328–35.
- Bohrer, R., and W. Chow (1978). Weights of one-sided multivariate inference. *Applied Statistics* 27, 100–4.
- Davidson, R., and J.-Y. Duclos (1998). Statistical inference for stochastic dominance and for the measurement of poverty and inequality. GREQAM, Doc. de Travail, no. 98A14.
- Davies, J., and M. Hoy (1995). Making inequality comparisons when Lorenz curves intersect. *American Economic Review* 85, 980–6.
- Dufour, J.-M. (1989). Nonlinear hypotheses, inequality restrictions and nonnested hypotheses: Exact simultaneous tests in linear regression. *Econometrica* 57, 335–55.
- Dufour, J.-M., and L. Khalaf (1995). Finite sample inference methods in seemingly unrelated regressions and simultaneous equations. Technical report, University of Montreal.
- Durbin, J. (1973). Distribution theory for tests based on the sample distribution function. Philadelphia, SIAM.
- Durbin, J. (1985). The first passage density of a continuous Gaussian process to a general boundary. *Journal of Applied Probability* 22, 99–122.
- Gouriéroux, C., A. Holly, and A. Monfort (1980). Kuhn–Tucker, likelihood ratio and Wald tests for nonlinear models with inequality constraints on the parameters. Harvard Institute of Economic Research, Mimeographed paper no. 770.
- Gouriéroux, C., A. Holly, and A. Monfort (1982). Likelihood ratio test, Wald test and Kuhn–Tucker test in linear models with inequality constraints in the regression parameters. *Econometrica* 50, 63–80.
- King, M.L., and Ping X. Wu (1997). Locally optimal one-sided tests for multiparameter hypotheses. *Econometric Reviews* 16, 131–56.
- Kaur, A., B.L.S. Prakasa Rao, and H. Singh (1994). Testing for second-order stochastic dominance of two distributions. *Econometric Theory* 10, 849–66.
- Klecan, L., R. McFadden, and D. McFadden (1991). A robust test for stochastic dominance. Working paper, Economics Dept., MIT.
- Kodde, D.A., and F.C. Palm (1986). Wald criteria for jointly testing equality and inequality restrictions. *Econometrica* 50, 1243–8.
- Kodde, D.A., and F. Palm (1987). A parametric test of the negativity of the substitution matrix. *Journal of Applied Econometrics* 2, 227–35.
- Kudô, A. (1963). A multivariate analogue of the one-sided test. *Biometrika* 50, 403–18.
- Maasoumi, E., J. Mills, and S. Zandvakili (1997). Consensus ranking of US income distributions: A bootstrap application of stochastic dominance tests. SMU, Economics.

- McFadden, D. (1989). Testing for stochastic dominance. In Part II of T. Fomby and T.K. Seo (eds.) *Studies in the Economics of Uncertainty* (in honor of J. Hadar). Springer-Verlag.
- Perlman, M.D. (1969). One-sided testing problems in multivariate analysis. *Annals of Mathematics and Statistics* 40, 549–62.
- Potscher, B.M., and I.R. Prucha (1991a). Basic structure of the asymptotic theory in dynamic nonlinear econometric models: I. Consistency and approximation concepts. *Econometric Reviews* 10, 125–216.
- Potscher, B.M., and I.R. Prucha (1991b). Basic structure of the asymptotic theory in dynamic nonlinear econometric models: II. Asymptotic Normality. *Econometric Reviews* 10, 253–326 (with comments).
- Sen, P.K. (1972). On the Bahadur representation of sample quantiles for sequences of  $\varphi$ -mixing random variables. *Journal of Multivariate Analysis* 2, 77–95.
- Shorrocks, A.F. (1983). Ranking income distributions. *Economica* 50, 3–17.
- Shorrocks, A., and J. Foster (1987). Transfer sensitive inequality measures. *Review of Economic Studies* 54, 485–97.
- Stewart, K.G. (1997). Exact testing in multivariate regression. *Econometric Reviews* 16, 321–52.
- Wolak, F. (1989). Local and global testing of linear and nonlinear inequality constraints in nonlinear econometric models. *Econometric Theory* 5, 1–35.
- Wolak, F. (1991). The local nature of hypothesis tests involving inequality constraints in nonlinear models. *Econometrica* 981–95.
- Xu, K., G. Fisher, and D. Wilson (1995). New distribution-free tests for stochastic dominance. Working paper No. 95-02, February, Dept. of Economics, Dalhousie University, Halifax, Nova Scotia.

CHAPTER TWENTY-SIX

# Spurious Regressions in Econometrics

*Clive W.J. Granger*

## 1 INTRODUCTION, HISTORY, AND DEFINITIONS

If  $X_t$ ,  $Y_t$  are a pair of time series, independent of each other and one runs the simple ordinary least squares regression

$$Y_t = a + bX_t + e_t, \quad (26.1)$$

then one should expect to find no evidence of a relationship, so that the estimate of  $b$  is near zero and its associated  $t$ -statistics is insignificant. However, when the individual series have strong autocorrelations, it had been realized by the early 1970s by time series analysis that the situation may not be so simple; that apparent relationships may often be observed using standard interpretations of such regressions. Because a relationship appears to be found between independent series, they have been called "spurious". To appreciate part of the problem, note that if  $b = 0$ , then  $e_t$  must have the same time series properties as  $Y_t$ , that is will be strongly autocorrelated, and so the assumptions of the classical OLS regression will not be obeyed, as discussed in virtually any statistics or econometrics textbook. The possibility of getting incorrect results from regressions was originally pointed out by Yule (1926) in a much cited but insufficiently read paper that discussed "nonsense correlations." Kendall (1954) also pointed out if  $X_t$ ,  $Y_t$  both obeyed the same autoregressive model of order one (AR (1))

$$\left. \begin{aligned} X_t &= a_1 X_{t-1} + \varepsilon_{xt} \\ Y_t &= a_2 Y_{t-1} + \varepsilon_{yt} \end{aligned} \right\} \quad (26.2)$$

with  $a_1 = a_2 = a$ , where  $\varepsilon_{xt}$ ,  $\varepsilon_{yt}$  are a pair of zero-mean, white noise (zero autocorrelated) series independent of each other at all pairs of times, then the sample correlation ( $R$ ) between  $X_t$ ,  $Y_t$  has

$$\text{var}(R) = n^{-1}(1 + a^2)/(1 - a^2),$$

where  $n$  is the sample size. Remember that  $R$ , being a correlation must be between  $-1$  and  $1$ , but if  $a$  is near one and  $n$  not very large, then  $\text{var}(R)$  will be quite big, which can only be achieved if the distribution of  $R$  values has large weights near the extreme values of  $-1$  and  $1$ , which will correspond to "significant"  $b$  values in (26.1).

## 2 SIMULATIONS

The obvious way to find evidence of spurious regressions is by using simulations. The first simulation on the topic was by Granger and Newbold (1974) who generated pairs of independent random walks, from (26.2) with  $a_1 = a_2 = 1$ . Each series had 50 terms and 100 repetitions were used. If the regression (26.1) is run, using series that are temporarily uncorrelated, one would expect that roughly 95 percent of values of  $|t|$  on  $b$  would be less than 2. This original simulation using random walks found  $|t| \leq 2$  on only 23 occasions, out of the 100,  $|t|$  was between 2 and 4 on 24 times, between 4 and 7 on 34 times, and over 7 on the other 19 occasions.

The reaction to these results was to re-assess many of the previously obtained empirical results in applied time series econometrics, which undoubtedly involved highly autocorrelated series but had not previously been concerned by this fact. Just having a high  $R^2$  value and an apparently significant value of  $b$  was no longer sufficient for a regression to be satisfactory or its interpretations relevant. The immediate questions were how one could easily detect a spurious regression and then correct for it. Granger and Newbold (1974) concentrated on the value of the Durbin–Watson statistic; if the value is too low, it suggests that the regressions results cannot be trusted. Quick fix methods such as using a Cochrane–Orcutt technique to correct autocorrelations in the residuals, or differencing the series used in a regression were inclined to introduce further difficulties and so cannot be recommended. The problem arises because equation (26.1) is misspecified, the proper reaction to having a possible spurious relationship is to add lagged dependent and independent variables, until the errors appear to be white noise, according to the Durbin–Watson statistic. A random walk is an example of an I(1) process, that is a process that needs to be differenced to become stationary. Such processes seem to be common in parts of econometrics, especially in macroeconomics and finance. One approach that is widely recommended is to test if  $X_t, Y_t$  are I(1) and, if yes, to difference before performing the regression (26.1). There are many tests available, a popular one is due to Dickey–Fuller (1979). However, as will be explained below, even this approach is not without its practical difficulties.

## 3 THEORY

A theoretical investigation of the basic unit root, ordinary least squares, spurious regression case was undertaken by Phillips (1986). He considered the asymptotic

properties of the coefficients and statistics of equation (26.1),  $\hat{a}$ ,  $\hat{b}$ , the  $t$ -statistic for  $b$ ,  $R^2$  and the Durbin–Watson statistics  $\hat{\rho}$ . To do this he introduced the link between normed sums of functions of unit root processes and integrals of Wiener processes. For example if a sample  $X_t$  of size  $T$  is generated from a driftless random walk then

$$T^{-3/2} \sum_1^T X_t \rightarrow \sigma_\epsilon \int_0^1 W(t) dt$$

where  $\sigma_\epsilon^2$  is the variance of the shock,

$$T^{-2} \sum_1^T X_t^2 \rightarrow \sigma_\epsilon^2 \int_0^1 W^2(t) dt$$

and if  $X_t$ ,  $Y_t$  are an independent pair of such random walks, then

$$T^{-2} \sum_1^T X_t Y_t \rightarrow \sigma_\epsilon \sigma_\eta \int_0^1 V(t) W(t) dt$$

where  $V(t)$ ,  $W(t)$  are independent Wiener processes. As a Wiener process is a continuous time random process on the real line  $[0, 1]$ , the various sums are converging and can thus be replaced by integrals of a stochastic process. This transformation makes the mathematics of the investigation much easier, once one becomes familiar with the new tools. Phillips is able to show that

1. the distributions of the  $t$ -statistics for  $\hat{a}$  and  $\hat{b}$  from (26.1) diverge as  $t$  becomes large, so there is no asymptotically correct critical values for these conventional tests.
2.  $\hat{b}$  converges to some random variable whose value changes from sample to sample.
3. Durbin–Watson statistics tend to zero.
4.  $R^2$  does not tend to zero but to some random variable.

What is particularly interesting is that not only do these theoretical results completely explain the simulations but also that the theory deals with asymptotics,  $T \rightarrow \infty$ , whereas the original simulations had only  $T = 50$ . It seems that spurious regression occurs at all sample sizes.

Haldrup (1994) has extended Phillips' result to the case for two independent I(2) variables and obtained similar results. (An I(2) variable is one that needs differencing twice to get to stationarity, or here, difference once to get to random walk.) Marmol (1998) has further extended these results to fractionally integrated, I(d), processes. Unpublished simulation results also exist for various other long-memory processes, including explosive autoregressive processes, (26.2) with  $a_1 = a_2 = a > 1$ . Durlauf and Phillips (1988) regress an I(1) process on deterministic

polynomials in time, thus polynomial trends, and found spurious relationships. Phillips (1998) has recently discussed how all such results can be interpreted by considering a decomposition of an I(1) series in terms of deterministic trends multiplied by stationary series.

#### 4 SPURIOUS REGRESSIONS WITH STATIONARY PROCESSES

Spurious regressions in econometrics are usually associated with I(1) processes, which was explored in Phillips' well known theory and in the best known simulations. What is less appreciated is that the problem can just also occur, although less clearly, with stationary processes. Table 26.1, taken from Granger, Hyung, and Jeon (1998), shows simulation results from independent series generated by (26.2) with  $0 < a_1 = a_2 = a \leq 1$  and  $e_{xt}, e_{yt}$  both Gaussian white noise series, using regression (26.1) estimated using OLS with sample sizes varying between 100 and 10,000.

It is seen that sample size has little impact on the percentage of spurious regressions found (apparent significance of the  $b$  coefficient in (26.1)). Fluctuations down columns do not change significantly with the number of iterations used. Thus, the spurious regression problem is not a small sample property. It is also seen to be a serious problem with pairs of autoregressive series which are not unit root processes. If  $a = 0.75$  for example, then 30 percent of regressions will give spurious implications. Further results are available in the original paper but will not be reported in detail. The Gaussian error assumption can be replaced by other distributions with little or no change in the simulation results, except for an exceptional distribution such as the Cauchy. Spurious regressions also occur if  $a_1 \neq a_2$ , although less frequently, and particularly if the smaller of the two  $a$  values is at least 0.5 in magnitude.

The obvious implications of these results is that applied econometricians should not worry about spurious regressions only when dealing with I(1), unit root, processes. Thus, a strategy of first testing if a series contains a unit root before entering into a regression is not relevant. The results suggest that many more simple regressions need to be interpreted with care, when the series involved are strongly serially correlated. Again, the correct response is to move to a better specification, using lags of all variables.

**Table 26.1** Regression between independent AR(1) series

Sample series	$a = 0$	$a = 0.25$	$a = 0.5$	$a = 0.75$	$a = 0.9$	$a = 1.0$
100	4.9	6.8	13.0	29.9	51.9	89.1
500	5.5	7.5	16.1	31.6	51.1	93.7
2,000	5.6	7.1	13.6	29.1	52.9	96.2
10,000	4.1	6.4	12.3	30.5	52.0	98.3

$a_1 = a_2 = a$  percentage of  $|t| > 2$

## 5 RELATED PROCESSES

The final generalization would take variables generated by (26.2) but now allowing  $e_{xt}$ ,  $e_{yt}$  to be correlated, say  $\rho = \text{corr}(e_{xt}, e_{yt})$ . Now the series are related and any relationship found in (26.2) will not be spurious, although the extent of the relationship is over-emphasized if the residual achieved is not white noise. The natural generalization is a bivariate vector autoregression or, if  $\rho$  is quite high, and if  $a_1 = a_2 = 1$ , the series will be cointegrated (as described in Chapter 30 on Cointegration), in which case an error-correction model is appropriate. In all these models, spurious regressions should not be a problem.

## References

- Dickey, D.A., and W.A. Fuller (1979). Distribution of the estimates for autoregressive time series with a unit root. *Journal of the American Statistical Association* 74, 427–31.
- Durlauf, S.N., and P.C.B. Phillips (1988). Trends versus random walks in time series analysis. *Econometrica* 56, 1333–54.
- Granger, C.W.J., N. Hyung, and Y. Jeon (1998). Spurious regression with stationary series. Working Paper, UCSD Department of Economics.
- Granger, C.W.J., and P. Newbold (1974). Spurious regressions in econometrics. *Journal of Econometrics* 2, 111–20.
- Haldrup, N. (1994). The asymptotics of single-equation cointegration regressions with I(1) and I(2) variables. *Journal of Econometrics* 63, 153–81.
- Kendall, M.G. (1954). *Exercises in Theoretical Statistics*. London, Griffin.
- Marmol, F. (1998). Spurious regression theory with non-stationary fractionally integrated processes. *Journal of Econometrics* 84, 233–50.
- Phillips, P.C.B. (1986). Understanding spurious regressions in econometrics. *Journal of Econometrics* 33, 311–40.
- Phillips, P.C.B. (1998). New tools for understanding spurious regressions. *Econometrica* 66, 1299–325.
- Yule, G.U. (1926). Why do we sometimes get nonsense correlations between time series? *Journal of the Royal Statistical Society* 89, 1–64.

---

CHAPTER TWENTY-SEVEN

# Forecasting Economic Time Series

*James H. Stock\**

## 1 INTRODUCTION

The construction and interpretation of economic forecasts is one of the most publicly visible activities of professional economists. Over the past two decades, increased computer power has made increasingly sophisticated forecasting methods routinely available and the role of economic forecasting has expanded. Economic forecasts now enter into many aspects of economic life, including business planning, state and local budgeting, financial management, financial engineering, and monetary and fiscal policy. Yet, with this widening scope comes greater opportunities for the production of poor forecasts and the misinterpretation of good forecasts. Responsible production and interpretation of economic forecasts requires a clear understanding of the associated econometric tools, their limits, and an awareness of common pitfalls in their application.

This chapter provides an introduction to the main methods used for forecasting economic time series. The field of economic forecasting is large, and, because of space limitations, this chapter covers only the most salient topics. The focus here will be on point forecasts, that is, forecasts of future values of the time series. It is assumed that the historical series is relatively "clean," in the sense of having no omitted observations, being observed at a consistent sampling frequency (e.g. monthly), and either having no seasonal component or having been seasonally adjusted. It is assumed that the forecaster has quadratic (i.e. mean squared error) loss. Finally, it is assumed that the time series is sufficiently long, relative to the forecast horizon, that the history of the time series will be informative for making the forecast and for estimating parametric models.

This chapter has four substantive sections. Section 2 provides a theoretical framework for considering some of the tradeoffs in the construction of economic forecasts and for the comparison of forecasting methods. Section 3 provides a

glimpse at some of the relevant empirical features of macroeconomic time series data. Section 4 discusses univariate forecasts, that is, forecasts of a series made using only past values of that series. Section 5 provides an overview of multivariate forecasting, in which forecasts are made using historical information on multiple time series.

There are many interesting and important aspects of economic forecasting that are not covered in this chapter. In some applications, it is of interest to estimate the entire distribution of future values of the variable of interest, conditional on current information, or certain functions of that conditional distribution. An example that arises in macroeconomics is predicting the probability of a recession, an event often modeled as two consecutive declines in real gross domestic product. Other functions of conditional distributions arise in finance; for examples, see Diebold, Gunther, and Tay (1998).

In some cases, time varying conditional densities might be adequately summarized by time varying conditional first and second moments, that is, by modeling conditional heteroskedasticity. For example, conditional estimates of future second moments of the returns on an asset can be used to price options written on that asset. Although there are various frameworks for estimating conditional heteroskedasticity, the premier tool for modeling conditional heteroskedasticity is Engle's (1982) so-called autoregressive conditional heteroskedasticity (ARCH) framework and variants, as discussed in Bollerslev, Engle, and Nelson (1994).

Another topic not explored in this chapter is nonquadratic loss. Quadratic loss is a natural starting point for many forecasting problems, both because of its tractability and because, in many applications, it is plausible that loss is symmetric and that the marginal cost of a forecast error increases linearly with its magnitude. However, in some circumstances other loss functions are appropriate. For example, loss might be asymmetric (would you rather be held responsible for a surprise government surplus or deficit?); see Granger and Newbold (1986, ch. 4.2) and West, Edison, and Cho (1993) for examples. Handling nonquadratic loss can be computationally challenging. The classic paper in this literature is Granger (1969), and a recent contribution is Christoffersen and Diebold (1997).

Another important set of problems encountered in practice but not addressed here involve data irregularities, such as missing or irregularly spaced observations. Methods for handling these irregularities tend to be model-dependent. Within univariate linear models and low-dimensional multivariate linear models, these are typically well handled using state space representations and the Kalman filter, as is detailed by Harvey (1989). A somewhat different set of issues arise with series that have large seasonal components. Issues of seasonal adjustment and handling seasonal data are discussed in Chapter 31 in this volume by Ghysels, Osborn, and Rodrigues.

Different issues also arise if the forecast horizon is long relative to the sample size (say, at least one-fifth the sample size) and the data exhibit strong serial correlation. Then the long-run forecast is dominated by estimates of the long-run correlation structure. Inference about the long-run correlation structure is typically non-standard and, in some formulations, is related to the presence of large, possibly unit autoregressive roots and (in the multivariate setting) to possible cointegration

among the series. Unit roots and cointegration are respectively discussed in this volume in Chapter 29 by Bierens and Chapter 30 by Dolado, Gonzalo, and Marmol. The construction of point forecasts and forecast intervals at long horizons entails considerable difficulties because of the sensitivity to the long-run dependence parameters, and methods for handling this are examined in Stock (1996).

A final area not addressed here is the combination of competing forecasts. When a variable is forecasted by two different methods that draw on different information sets and neither model is true, typically a combination of the two forecasts is theoretically preferred to either individual forecast (Bates and Granger, 1969). For an introduction to this literature, see Granger (1989), Diebold and Lopez (1995), and Chan, Stock, and Watson (1998).

This chapter makes use of concepts and methods associated with unit autoregressive roots, cointegration, vector autoregressions (VARs), and structural breaks. These are all topics of separate chapters in this volume, and the reader is referred to those chapters for background details.

## 2 ECONOMIC FORECASTING: A THEORETICAL FRAMEWORK

### 2.1 Optimal forecasts, feasible forecasts, and forecast errors

Let  $y_t$  denote the scalar time series variable that the forecaster wishes to forecast, let  $h$  denote the horizon of the forecast, and let  $F_t$  denote the set of data used at time  $t$  for making the forecast ( $F_t$  is sometimes referred to as the information set available to the forecaster). If the forecaster has squared error loss, the point forecast  $\hat{y}_{t+h|t}$  is the function of  $F_t$  that minimizes the expected squared forecast error, that is,  $E[(y_{t+h} - \hat{y}_{t+h|t})^2 | F_t]$ . This expected loss is minimized when the forecast is the conditional expectation,  $E(y_{t+h} | F_t)$ . In general, this conditional expectation might be a time varying function of  $F_t$ . However, in this chapter we will assume that the data are drawn from a stationary distribution, that is, the distribution of  $(y_s, \dots, y_{s+T})$  does not depend on  $s$  (although some mention of structural breaks will be made later); then  $E(y_{t+h} | F_t)$  is a time invariant function of  $F_t$ .

In practice,  $E(y_{t+h} | F_t)$  is unknown and is in general nonlinear. Forecasts are constructed by approximating this unknown conditional expectation by a parametric function. This parametric function, or model, is denoted  $\mu_h(F_t, \theta)$ , where  $\theta$  is a parameter vector which is assumed to lie in the parameter space  $\Theta$ . The “best” value of this parameter,  $\theta_0$ , is the value that minimizes the mean squared approximation error,  $E[\mu_h(F_t, \theta) - E(y_{t+h} | F_t)]^2$ .

Because  $\theta_0$  is unknown, it is typically estimated from historical data, and the estimate is denoted  $\hat{\theta}$ . To be concrete, suppose that  $F_t$  consists of observations on  $X_s$ ,  $1 \leq s \leq T$ , where  $X_s$  is a vector time series (which typically includes  $y_s$ ). Further suppose that only the first  $p$  lags of  $X_t$  are included in the forecast. Then  $\theta$  could be estimated by least squares, that is, by solving,

$$\min_{\theta \in \Theta} \sum_{t=p+1}^{T-h} [y_{t+h} - \mu_h(F_t, \theta)]^2. \quad (27.1)$$

There are alternative methods for the estimation of  $\theta$ . The minimization problem (27.1) uses an  $h$ -step ahead (nonlinear) least squares regression. Often an available alternative is to estimate a one-step ahead model ( $h = 1$ ) by nonlinear least squares or maximum likelihood, and to iterate that model forward. The formulation (27.1) has the advantage of computational simplicity, especially for nonlinear models. Depending on the true model and the approximate model, approximation bias can be reduced by estimating the  $h$ -step ahead model (27.1). On the other hand, if the estimated model is correct, then iterating one-step ahead forecasts will be more efficient in the statistical sense. In general, the decision of whether to estimate parameters by  $h$ -step ahead or one-step ahead methods depends on the model being estimated and the type of misspecification that might be present. See Clements and Hendry (1996) for references to this literature and for simulation results comparing the two approaches.

It is useful to consider a decomposition of the forecast error, based on the various sources of that error. Let  $\hat{e}_{t+h,t}$  denote the forecast error from the  $h$ -step ahead forecast of  $y_{t+h}$  using  $\hat{y}_{t+h|t}$ . Then,

$$\begin{aligned}\hat{e}_{t+h,t} &= y_{t+h} - \hat{y}_{t+h|t} \\ &= [y_{t+h} - E(y_{t+h} | F_t)] + [E(y_{t+h} | F_t) - \mu_h(F_t, \theta_0)] + [\mu_h(F_t, \theta_0) - \mu_h(F_t, \hat{\theta})].\end{aligned}\tag{27.2}$$

The first term in brackets is the deviation of  $y_{t+h}$  from its conditional expectation, a source of forecast error that cannot be eliminated. The second term in brackets is the contribution of model misspecification, and is the error arising from using the best parameter value for the approximate conditional expectations function. The final term arises because this best parameter value is unknown, and instead  $\theta$  is estimated from the data.

The decomposition (27.2) illustrates two facts. First, all forecasts, no matter how good, will have forecast error because of future, unknowable random events. Second, the quality of a forecasting method is therefore determined by its model approximation error and by its estimation error. These two sources of error generally entail a tradeoff. Using a flexible model with many parameters for  $\mu_h$  can reduce model approximation error, but because there are many parameters estimation error increases.

## 2.2 Model selection using information criteria

Because the object of point forecasting is to minimize expected loss out-of-sample, it is not desirable to minimize approximation error (bias) when this entails adding considerable parameter estimation uncertainty. Thus, for example, model selection based on minimizing the sum of squared residuals, or maximizing the  $R^2$ , can lead to small bias and good in-sample fit, but very poor out-of-sample forecast performance.

A formal way to make this tradeoff between approximation error and estimation error is to use information criteria to select among a few competing models. When  $h = 1$ , information criteria (IC) have the form,

$$\text{IC}(p) = \ln \hat{\sigma}^2(p) + pg(T) \quad (27.3)$$

where  $p$  is the dimension of  $\theta$ ,  $T$  is the sample size used for estimation,  $g(T)$  is a function of  $T$  with  $g(T) > 0$  and  $Tg(T) \rightarrow \infty$  and  $g(T) \rightarrow 0$  as  $T \rightarrow \infty$ , and  $\hat{\sigma}^2(p) = \text{SSR}/T$ , where SSR is the sum of squared residuals from the (in-sample) estimation. Comparing two models using the information criterion (27.3) is the same as comparing two models by their sum of squared residuals, except that the model with more parameters receives a penalty. Under suitable conditions on this penalty and on the class of models being considered, it can be shown that a model selected by the information criterion is the best in the sense of the trade-off between approximation error and sampling uncertainty about  $\theta$ . A precise statement of such conditions in AR models, when only the maximum order is known, can be found in Geweke and Meese (1981), and extensions to infinite order autoregressive models are discussed in Brockwell and Davis (1987) and, in the context of unit root tests, Ng and Perron (1995). The two most common information criteria are the Akaike information criterion (AIC), for which  $g(T) = 2/T$ , and Schwarz's (1978) Bayes information criterion (BIC), for which  $g(T) = \ln T/T$ .

## 2.3 Prediction intervals

In some cases, the object of forecasting is not to produce a point forecast but rather to produce a range within which  $y_{t+h}$  has a prespecified probability of falling. Even if within the context of point forecasting, it is useful to provide users of forecasts with a measure of the uncertainty of the forecast. Both ends can be accomplished by reporting prediction intervals.

In general, the form of the prediction interval depends on the underlying distribution of the data. The simplest prediction interval is obtained by assuming that the data are conditionally homoskedastic and normal. Under these assumptions and regularity conditions, a prediction interval with asymptotic 67 percent coverage is given by  $\hat{y}_{t+h|t} \pm \hat{\sigma}_h$ , where  $\hat{\sigma}_h = \text{SSR}_h/(T - p)$ , where  $\text{SSR}_h$  is the sum of squared residuals from the  $h$ -step ahead regression (27.1) and  $T - p$  are the degrees of freedom of that regression.

If the series is conditionally normal but is conditionally heteroskedastic, this simple prediction error formula must be modified and the conditional variance can be computed using, for example, an ARCH model. If the series is conditionally nonnormally distributed, other methods, such as the bootstrap, can be used to construct asymptotically valid prediction intervals.

## 2.4 Forecast comparison and evaluation

The most reliable way to evaluate a forecast or to compare forecasting methods is by examining out of sample performance. To evaluate the forecasting performance of a single model or expert, one looks for signs of internal consistency. If the

forecasts were made under squared error loss, the forecast errors should have mean zero and should be uncorrelated with any variable used to produce the forecast. For example,  $\hat{e}_{t+h,t}$  should be uncorrelated with  $\hat{e}_{t,t-h}$ , although  $\hat{e}_{t+h,t}$  will in general have an  $MA(h-1)$  correlation structure. Failure of out-of-sample forecasts to have mean zero and to be uncorrelated with  $F_t$  indicates a structural break, a deficiency of the forecasting model, or both.

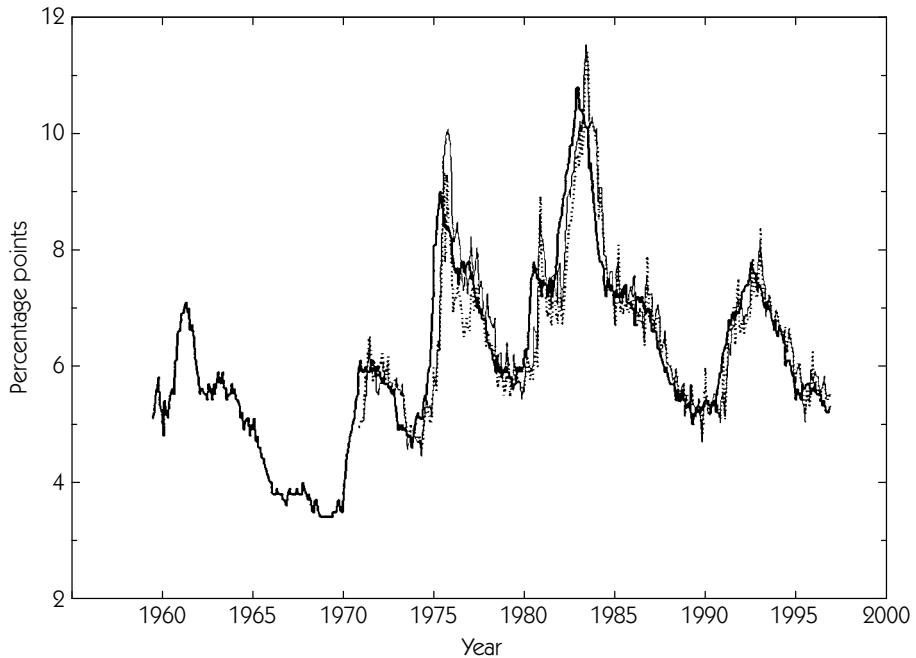
Additional insights are obtained by comparing the out-of-sample forecasts of competing models or experts. Under mean squared error loss, the relative performance of two time series of point forecasts of the same variable can be compared by computing their mean squared forecast errors (MSFE). Of course, in a finite sample, a smaller MSFE might simply be an artifact of sampling error, so formal tests of whether the MSFEs are statistically significantly different are in order when comparing two forecasts. Such tests have been developed by Diebold and Mariano (1995) and further refined by West (1996), who built on earlier work by Nelson (1972), Fair (1980), and others.

Out of sample performance can be measured either by using true out-of-sample forecasts, or by a simulated out-of-sample forecasting exercise. While both approaches have similar objectives, the practical issues and interpretation of results is quite different. Because real time published forecasts usually involve expert opinion, a comparison of true out-of-sample forecasts typically entails an evaluation of both models and the expertise of those who use the models. Good examples of comparisons of real time forecasts, and of the lessons that can be drawn from such comparisons, are McNees (1990) and Zarnowitz and Braun (1993).

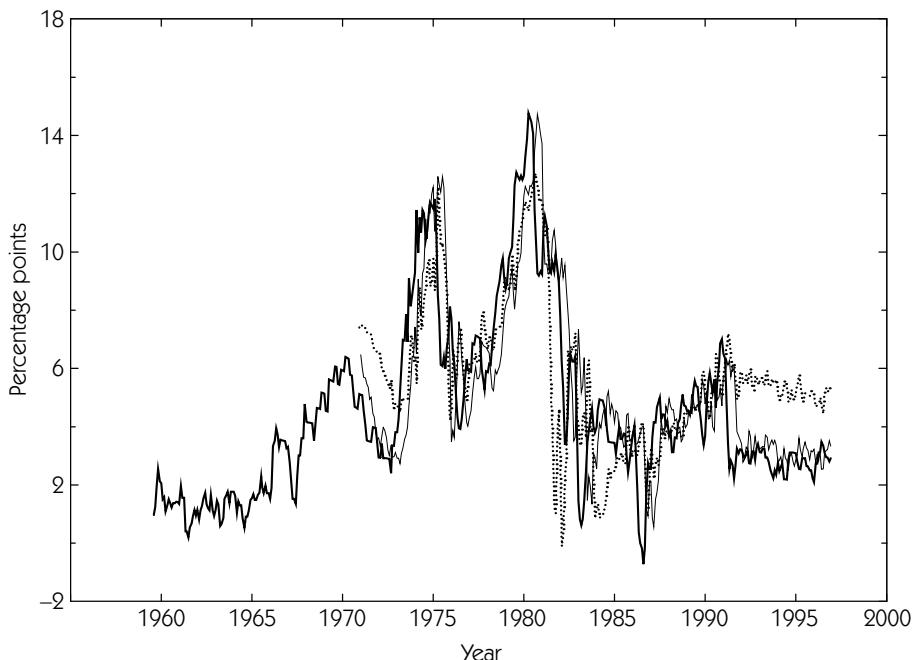
Simulated real time forecasting can be done in the course of model development and provides a useful check on the in-sample comparison measures discussed above. The essence of a simulated real time forecasting experiment is that all forecasts  $\hat{y}_{t+h|t}$ ,  $t = T_0, \dots, T_1$ , are functions only of data up through date  $t$ , so that all parameter estimation, model selection, etc., is done only using data through date  $t$ . This is often referred to as a recursive methodology (for linear models, the simulated out-of-sample forecasts can be computed using a recursion). In general this entails many re-estimations of the model, which for nonlinear models can be computationally demanding. For an example of simulated out-of-sample forecast comparisons, see Stock and Watson (1999a).

### 3 SALIENT FEATURES OF US MACROECONOMIC TIME SERIES DATA

The methods discussed in this chapter will be illustrated by application to five monthly economic time series for the US macroeconomy: inflation, as measured by the annual percentage change in the consumer price index (CPI); output growth, as measured by the growth rate of the index of industrial production; the unemployment rate; a short-term interest rate, as measured by the rate on 90-day US Treasury bills; and total real manufacturing and trade inventories, in logarithms.<sup>1</sup> Time series plots of these five series are presented as the heavy solid lines in Figures 27.1–27.5.



**Figure 27.1** US unemployment rate (heavy solid line), recursive AR(BIC)/unit root pretest forecast (light solid line), and neural network forecast (dotted line)



**Figure 27.2** Six-month US CPI inflation at an annual rate (heavy solid line), recursive AR(BIC)/unit root pretest forecast (light solid line), and neural network forecast (dotted line)

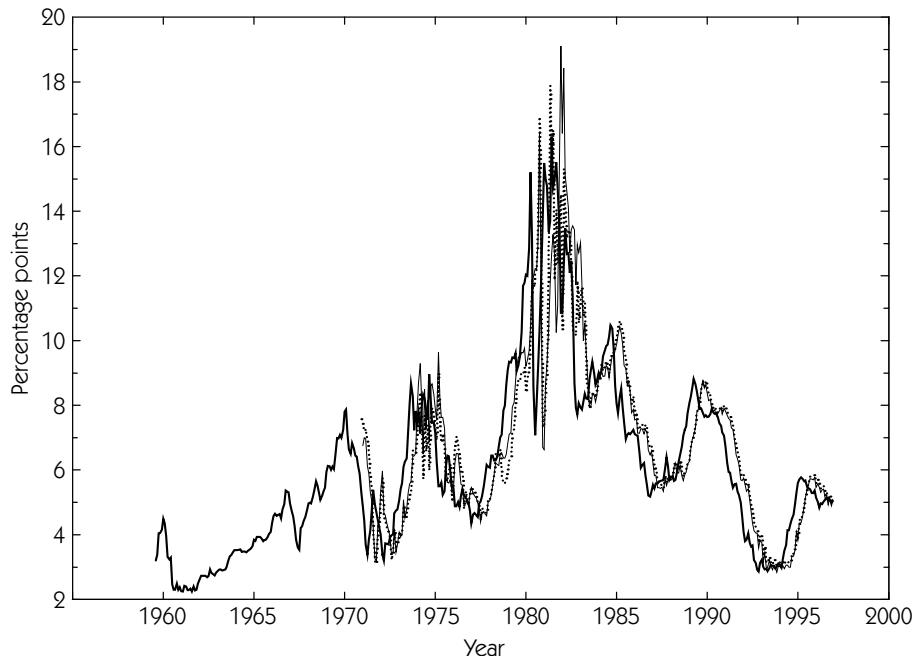


Figure 27.3 90-day Treasury bill at an annual rate (heavy solid line), recursive AR(BIC)/unit root pretest forecast (light solid line), and neural network forecast (dotted line)

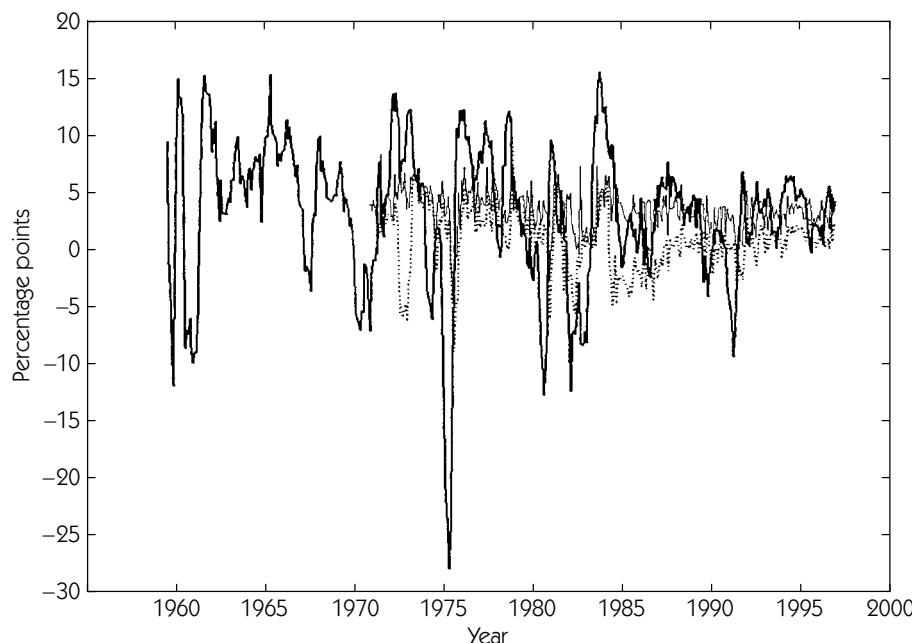
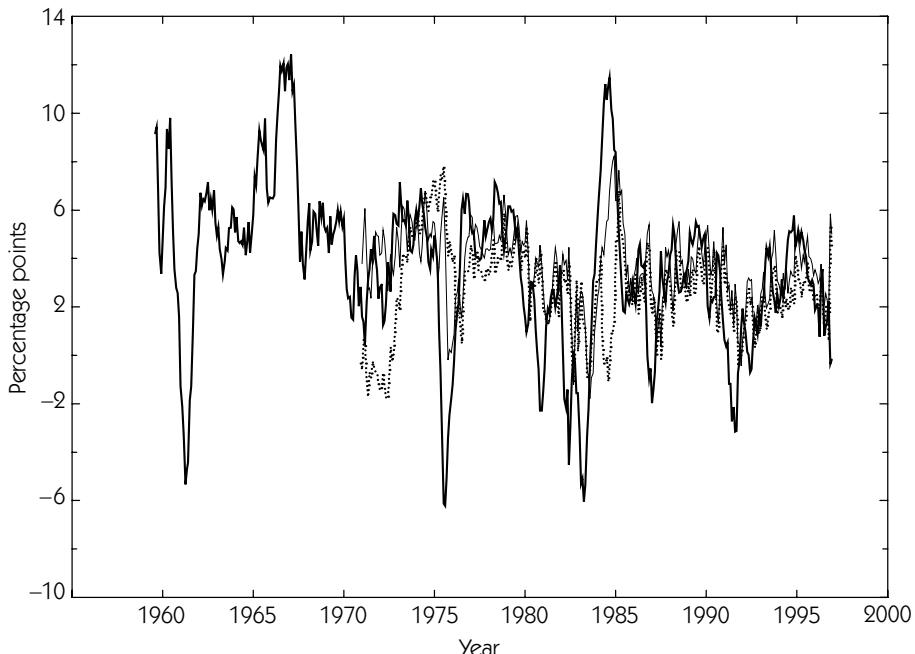


Figure 27.4 Six-month growth of US industrial production at an annual rate (heavy solid line), recursive AR(BIC)/unit root pretest forecast (light solid line), and neural network forecast (dotted line)



**Figure 27.5** Six-month growth of total real US manufacturing and trade inventories at an annual rate (heavy solid line), recursive AR(BIC)/unit root pretest forecast (light solid line), and neural network forecast (dotted line)

In addition to being of interest in their own right, these series reflect some of the main statistical features present in many macroeconomic time series from developed economies. The 90-day Treasury bill rate, unemployment, inflation, and inventories all exhibit high persistence in the form of smooth long-run trends. These trends are clearly nonlinear, however, and follow no evident deterministic form, rather, the long-run component of these series can be thought of as a highly persistent stochastic trend. There has been much debate over whether this persistence is well modeled as arising from an autoregressive unit root in these series, and the issue of whether to impose a unit root (to first difference these data) is an important forecasting decision discussed below.

Two other features are evident in these series. All five series exhibit co-movements, especially over the two to four year horizons. The twin recessions of the early 1980s, the long expansions of the mid-1980s and the 1990s, and the recession in 1990 are reflected in each series (although the IP (industrial production) growth rate series might require some smoothing to see this). Such movements over the business cycle are typical for macroeconomic time series data; for further discussion of business cycle properties of economic time series data, see Stock and Watson (1999b). Finally, to varying degrees the series contain high frequency noise. This is most evident in inflation and IP growth. This high frequency noise arises from short-term, essentially random fluctuations in economic activity and from measurement error.

## 4 UNIVARIATE FORECASTS

Univariate forecasts are made solely using past observations on the series being forecast. Even if economic theory suggests additional variables that should be useful in forecasting a particular variable, univariate forecasts provide a simple and often reliable benchmark against which to assess the performance of those multivariate methods. In this section, some linear and nonlinear univariate forecasting methods are briefly presented. The performance of these methods is then illustrated for the macroeconomic time series in Figures 27.1–27.5.

### 4.1 Linear models

One of the simplest forecasting methods is the *exponential smoothing* or exponentially weighted moving average (EWMA) method. The EWMA forecast is,

$$\hat{y}_{t+h|t} = \alpha \hat{y}_{t+h-1|t-1} + (1 - \alpha)y_t, \quad (27.4)$$

where  $\alpha$  is a parameter chosen by the forecaster or estimated by nonlinear least squares from historical data.

*Autoregressive moving average* (ARMA) models are a mainstay of univariate forecasting. The ARMA( $p, q$ ) model is,

$$a(L)y_t = \mu_t + b(L)\varepsilon_t, \quad (27.5)$$

where  $\varepsilon_t$  is a serially uncorrelated disturbance and  $a(L)$  and  $b(L)$  are lag polynomials of orders  $p$  and  $q$ , respectively. For  $y_t$  to be stationary, the roots of  $a(L)$  lie outside the unit circle, and for  $b(L)$  to be invertible, the roots of  $b(L)$  also lie outside the unit circle. The term  $\mu_t$  summarizes the deterministic component of the series. For example, if  $\mu_t$  is a constant, the series is stationary around a constant mean. If  $\mu_t = \mu_0 + \mu_1 t$ , the series is stationary around a linear time trend. If  $q > 0$ , estimation of the unknown parameters of  $a(L)$  and  $b(L)$  entails nonlinear maximization. Asymptotic Gaussian maximum likelihood estimates of these parameters are a staple of time series forecasting computer packages. Multistep forecasts are computed by iterating forward the one-step forecasts. A deficiency of ARMA models is estimator bias introduced when the MA roots are large, the so-called unit MA root pileup problem (see Davis and Dunsmuir, 1996; and, for a general discussion and references, Stock, 1994).

An important special case of ARMA models are pure *autoregressive* models with lag order  $p$  (AR( $p$ )). Meese and Geweke (1984) performed a large simulated out of sample forecasting comparison that examined a variety of linear forecasts, and found that long autoregressions and autoregressions with lags selected by information criteria performed well, and on average outperformed forecasts from ARMA models. The parameters can be estimated by ordinary least squares (OLS) and the order of the autoregression can be consistently estimated by, for example, the BIC.

Harvey (1989) has proposed a different framework for univariate forecasting, based on a decomposition of a series into various components: trend, cycle, seasonal, and irregular. Conceptually, this framework draws on an old concept in economic time series analysis in which the series is thought of as having different properties at different horizons, so that for example one might talk about the cyclical properties of a time series separately from its trend properties; he therefore calls these *structural time series models*. Harvey models these components as statistically uncorrelated at all leads and lags, and he parameterizes the components to reflect their role, for example, the trend can be modeled as a random walk with drift or a doubly integrated random walk, possibly with drift. Estimation is by asymptotic Gaussian maximum likelihood. The resulting forecasts are linear in historical data (although nonlinear in the parameters of the model) so these too are linear forecasts. Harvey (1989) argues that this formulation produces forecasts that avoid some of the undesirable properties of ARMA models. As with ARMA models, user judgment is required to select the models. One interesting application of these models is for trend estimation, see for example Stock and Watson (1998).

## 4.2 Nonlinear models

Outside of the normal distribution, conditional expectations are typically nonlinear, and in general one would imagine that these infeasible optimal forecasts would be nonlinear functions of past data. The main difficulty that arises with nonlinear forecasts is choosing a feasible forecasting method that performs well with the fairly short historical time series available for macroeconomic forecasting. With many parameters, approximation error in (27.2) is reduced, but estimation error can be increased. Many nonlinear forecasting methods also pose technical problems, such as having objective functions with many local minima, having parameters that are not globally identified, and difficulties with generating internally consistent  $h$ -step ahead forecasts from one-step ahead models.

Recognition of these issues has led to the development of a vast array of methods for nonlinear forecasting, and a comprehensive survey of these methods is beyond the limited scope of this chapter. Rather, here I provide a brief introduction to only two particular nonlinear models, smooth transition autoregressions (STAR) and artificial neural networks (NN). These models are interesting methodologically because they represent, respectively, parametric and nonparametric approaches to nonlinear forecasting, and they are interesting from a practical point of view because they have been fairly widely applied to economic data.

A third class of models that has received considerable attention in economics are the Markov switching models, in which an unobserved discrete state switches stochastically between regimes in which the process evolves in an otherwise linear fashion. Markov switching models were introduced in econometrics by Hamilton (1989) and are also known as hidden Markov models. However, space limitations preclude presenting these models here; for a textbook treatment, see Hamilton (1994). Kim and Nelson (1998, 1999) provide important extensions of this framework to multivariate models with unobserved components. The reader

interested in further discussions of and additional references to other nonlinear time series forecasting methods should see the recent surveys and/or textbook treatments of nonlinear models by Granger and Teräsvirta (1993), Priestly (1989), and Samorodnitsky and Taqqu (1994).

An *artificial neural network* (NN) model relates inputs (lagged values) to outputs (future values) using an index model formulation with nonlinear transformations. There is considerable terminology and interpretation of these formulations which we will not go into here but which are addressed in a number of textbook treatments of these models; see in particular Swanson and White (1995, 1997) for discussions and applications of NN models to economic data. Here, we consider the simplest version, a feedforward NN with a single hidden layer and  $n$  hidden units. This has the form:

$$y_{t+h} = \beta_0(L)y_t + \sum_{i=1}^n \gamma_i g(\beta_i(L)y_t) + u_{t+h}, \quad (27.6)$$

where  $\beta_i(L)$ ,  $i = 0, \dots, n$  are lag polynomials,  $\gamma_i$  are unknown coefficients, and  $g(z)$  is a function that maps  $\Re \rightarrow [0, 1]$ . Possible choices of  $g(z)$  include the indicator function, sigmoids, and the logistic function. A variety of methods are available for the estimation of the unknown parameters of NNs, some specially designed for this problem; a natural estimation method is nonlinear least squares. NNs have a nonparametric interpretation when the number of hidden units ( $n$ ) is increased as the sample size tends to infinity.

*Smooth transition autoregressions* are piecewise linear models and have the form:

$$y_{t+h} = \alpha(L)y_t + d_t \beta(L)y_t + u_{t+h}, \quad (27.7)$$

where the mean is suppressed,  $\alpha(L)$  and  $\beta(L)$  are lag polynomials, and  $d_t$  is a nonlinear function of past data that switches between the “regimes”  $\alpha(L)$  and  $\beta(L)$ . Various functions are available for  $d_t$ . For example, if  $d_t$  is the logistic function so  $d_t = 1/(1 + \exp[\gamma_0 + \gamma_1 \zeta_t])$ , then the model is referred to as the logistic smooth transition autoregression (LSTAR) model. The switching variable  $d_t$  determines the “threshold” at which the series switches, and depends on the data through  $\zeta_t$ . For example,  $\zeta_t$  might equal  $y_{t-k}$ , where  $k$  is some lag for the switch. The parameters of the model can be estimated by nonlinear least squares. Details about formulation, estimation and forecasting for TAR and STAR models can be found in Granger and Teräsvirta (1993) and in Granger, Teräsvirta, and Anderson (1993). For an application of TAR (and other models) to forecasting U.S. unemployment, see Montgomery, Zarnowitz, Tsay, and Tiao (1998).

### 4.3 Differencing the data

A question that arises in practice is whether to difference the data prior to construction of a forecasting model. This arises in all the models discussed above, but for simplicity it is discussed here in the context of a pure AR model. If one knows a priori that there is in fact a unit autoregressive root, then it is efficient to impose this information and to estimate the model in first differences. Of course,

in practice this is not known. If there is a unit autoregressive root, then estimates of this root (or the coefficients associated with this root) are generally biased towards zero, and conditionally biased forecasts can obtain. However, the order of this bias is  $1/T$ , so for short horizon forecasts ( $h$  fixed) and  $T$  sufficiently large, this bias is negligible, so arguably the decision of whether to difference or not is unimportant to first-order asymptotically.

The issue of whether or not to difference the data, or more generally of how to treat the long-term dependence in the series, becomes important when the forecast horizon is long relative to the sample size. Computations in Stock (1996) suggest that these issues can arise even if the ratio,  $h/T$ , is small, .1 or greater. Conventional practice is to use a unit root pretest to make the decision about whether to difference or not, and the asymptotic results in Stock (1996) suggest that this approach has some merit when viewed from the perspective of minimizing either the maximum or integrated asymptotic risk, in a sense made precise in that paper. Although Dickey–Fuller (1979) unit root pretests are most common, other unit root tests have greater power, and tests with greater power produce lower risk for the pretest estimator. Unit root tests are surveyed in Stock (1994) and in Chapter 29 in this volume by Bierens.

#### 4.4 Empirical examples

We now turn to applications of some of these forecasting methods to the five US macroeconomic time series in Figures 27.1–27.5.<sup>2</sup> In the previous notation, the series to be forecast,  $y_t$ , is the series plotted in those figures, for example, for industrial production  $y_t = 200 \ln(\text{IP}_t/\text{IP}_{t-6})$ , while for the interest rate  $y_t$  is the untransformed interest rate in levels (at an annual rate). The exercise reported here is a simulated out of sample comparison of six different forecasting models. All series are observed monthly with no missing observations. For each series, the initial observation date is 1959:1. Six-month ahead ( $h = 6$ ) recursive forecasts of  $y_{t+6}$  are computed for  $t = 1971:3, \dots, 1996:6$ ; because a simulated out of sample methodology was used, all models were re-estimated at each such date  $t$ .

Eight different forecasts are computed: (a) EWMA, where the parameter is estimated by NLS; (b) AR(4) with a constant; (c) AR(4) with a constant and a time trend; (d) AR where the lag length is chosen by BIC ( $0 \leq p \leq 12$ ) and the decision to difference or not is made using the Elliott–Rothenberg–Stock (1996) unit root pretest; (e) NN with a single hidden layer and two hidden units; (f) NN with two hidden layers, two hidden units in the first layer, and one hidden unit in the second layer; (g) LSTAR in levels with three lags and  $\zeta_t = y_t - y_{t-6}$ ; and (h) LSTAR in differences with three lags and  $\zeta_t = y_t - y_{t-6}$ .

For each series, the simulated out of sample forecasts (b) and (e) are plotted in Figures 27.1–27.5. The root MSFEs for the different methods, relative to method (b), are presented in Table 27.1; thus method (b) has a relative root MSFE of 1.00 for all series. The final row of Table 27.1 presents the root mean squared forecast error in the native units of the series.

Several findings are evident. First, among the linear models, the AR(4) in levels with a constant performs well. This model dominates the AR(4) in levels with a

**Table 27.1** Comparison of simulated out-of-sample linear and nonlinear forecasts for five US macroeconomic time series

Forecasting model	Relative root mean squared forecast errors				
	Unem.	Infl.	Int.	IP	Invent.
(a) EWMA	1.11	2.55	0.95	1.09	1.50
(b) AR(4), levels, constant	1.00	1.00	1.00	1.00	1.00
(c) AR(4), levels, constant and time trend	1.09	1.00	1.05	1.06	1.15
(d) AR(BIC), unit root pretest	1.05	0.84	1.07	0.99	1.01
(e) NN, levels, 1 hidden layer, 2 hidden units	1.07	1.76	9.72	1.05	1.54
(f) NN, levels, 2 hidden layers, 2(1) hidden units	1.07	0.99	0.99	1.07	1.25
(g) LSTAR, levels, 3 lags, $\zeta_t = y_t - y_{t-6}$	1.04	1.34	3.35	1.02	1.05
(h) LSTAR, differences, 3 lags, $\zeta_t = y_t - y_{t-6}$	1.04	0.89	1.17	1.01	1.03
Root mean squared forecast error for (b), AR(4), levels, constant	0.61	2.44	1.74	6.22	2.78

Sample period: monthly, 1959:1–1996:12; forecast period: 1971:3–1996:6; forecast horizon: six months.

Entries in the upper row are the root mean squared forecast error of the forecasting model in the indicated row, relative to that of model (b), so the relative root MSFE of model (b) is 1.00. Smaller relative root MSFEs indicate more accurate forecasts in this simulated out-of-sample experiment. The entries in the final row are the root mean squared forecast errors of model (b) in the native units of the series. The series are: the unemployment rate, the six-month rate of CPI inflation, the 90-day US Treasury bill annualized rate of interest, the six-month growth of IP, and the six-month growth of real manufacturing and trade inventories.

**Table 27.2** Root mean squared forecast errors of VARs, relative to AR(4)

Forecasting model	Relative mean squared forecast errors				
	Unem.	Infl.	Int.	IP	Invent.
(i) VAR: unemp <sub>t</sub> , int.rate <sub>t</sub> , ΔlnIP <sub>t</sub>	0.98	—	1.05	0.93	—
(j) VAR: unemp <sub>t</sub> , ΔlnCPI <sub>t</sub> , int.rate <sub>t</sub>	1.03	0.95	1.03	—	—
(k) VAR: all five variables	1.03	1.10	1.04	0.94	0.96

Sample period: monthly, 1959:1–1996:12; forecast period: 1971:3–1996:6; forecast horizon: six months.

Entries are relative root MSFEs, relative to the root MSFE of model (b) in Table 27.1 (AR(4) with constant in levels). The VAR specifications have lag lengths selected by BIC; the six-month ahead forecasts were computed by iterating one-month ahead forecasts. All forecasts are simulated out of sample. See the notes to Table 27.1.

constant and time trend, in the sense that for all series the AR(4) with a constant and time trend has an RMSFE that is no less than the AR(4) with a constant. Evidently, fitting a linear time trend leads to poor out-of-sample performance, a result that would be expected if the trend is stochastic rather than deterministic. Using BIC lag length selection and a unit root pretest improves upon the AR(4) with a constant for inflation, has essentially the same performance for industrial production and inventories, and exhibits worse performance for the unemployment rate and the interest rate; averaged across series, the RMSFE is 0.99, indicating a slight edge over the AR(4) with a constant on average.

None of the nonlinear models uniformly improve upon the AR(4) with a constant. In fact, two of the nonlinear models ((e) and (g)) are dominated by the AR(4) with a constant. The greatest improvement is by model (h) for inflation; however, this relative RMSFE is still greater than the AR(BIC) forecast for inflation. Interestingly, the very simple EWMA forecast is the best of all forecasts, linear and nonlinear, for the interest rate. For the other series, however, it does not get the correct long-run trend, and the EWMA forecasts are worse than the AR(4).

The final row gives a sense of the performance of these forecasts in absolute terms. The RMSFE of the unemployment rate, six months hence, is only 0.6 percentage points, and the RMSFE for the 90-day Treasury bill rate is 1.7 percentage points. CPI inflation is harder to predict, with a six-month ahead RMSFE of 2.4 percentage points. Inspection of the graph of IP growth reveals that this series is highly volatile, and in absolute terms the forecast error is large, with a six-month ahead RMSFE of 6.2 percentage points.

Some of these points can be verified by inspection of the forecasts plotted in Figures 27.1–27.5. Clearly these forecasts track well the low frequency movements in the unemployment rate, the interest rate, and inflation (although the NN forecast does quite poorly in the 1990s for inflation). Industrial production and inventory growth has a larger high frequency component, which all these models have difficulty predicting (some of this high frequency component is just unpredictable forecast error).

These findings are consistent with the conclusions of the larger forecasting model comparison study in Stock and Watson (1999a). They found that, on average across 215 macroeconomic time series, autoregressive models with BIC lag length determination and a unit root pretest performed well, indeed, outperformed a range of NN and LSTAR models for six-month ahead forecasts. The autoregressive model typically improved significantly on no-change or EWMA models. Thus there is considerable ability to predict many U.S. macroeconomic time series, but much of this predictability is captured by relatively simple linear models with data-dependent determination of the specification.

## 5 MULTIVARIATE FORECASTS

The motivation for multivariate forecasting is that there is information in multiple economic time series that can be used to improve forecasts of the variable or variables of interest. Economic theory, formal and informal, suggests a

large number of such relations. Multivariate forecasting methods in econometrics are usefully divided into four broad categories: structural econometric models; small linear time series models; small nonlinear time series models; and forecasts based on leading economic indicators.

Structural econometric models attempt to exploit parametric relationships suggested by economic theory to provide a priori restrictions. These models can be hundred-plus equation simultaneous systems, or very simple relations such as an empirical Phillips curve relating changes of inflation to the unemployment rate and supply shocks. Because simultaneous equations are the topic of Chapter 6 by Mariano in this volume, forecasts from simultaneous equations systems will be discussed no further here. Neither will we discuss further nonlinear multivariate models; although the intuitive motivation for these is sound, these typically have many parameters to be estimated and as such often exhibit poor out-of-sample performance (for a study of multivariate NNs, see Swanson and White (1995, 1997); for some positive results, see Montgomery *et al.*, 1998). This chapter therefore briefly reviews multivariate forecasting with small linear time series models, in particular, using VARs, and forecasting with leading indicators. For additional background on VARs, see the chapter in this volume by Lütkepohl.

## 5.1 Vector autoregressions

Vector autoregressions, which were introduced to econometrics by Sims (1980), have the form:

$$Y_t = \mu_t + A(L)Y_{t-1} + \varepsilon_t, \quad (27.8)$$

where  $Y_t$  is a  $n \times 1$  vector time series,  $\varepsilon_t$  is a  $n \times 1$  serially uncorrelated disturbance,  $A(L)$  is a  $p$ th order lag polynomial matrix, and  $\mu_t$  is a  $n \times 1$  vector of deterministic terms (for example, a constant or a constant plus linear time trend). If there are no restrictions on the parameters, the parameters can be estimated asymptotically efficiently (under Gaussianity) by OLS equation by equation. Multistep forecasts can be made either by replacing the left-hand side of (27.8) by  $Y_{t+h}$ , or by  $h$ -fold iteration of the one-step forecast.

Two important practical questions are the selection of the series to include in  $Y_t$  (the choice of  $n$ ) and the choice of the lag order  $p$  in the VAR( $p$ ). Given the choice of series, the order  $p$  is typically unknown. As in the univariate case, it can be estimated by information criteria. This proceeds as discussed following (27.3), except that  $\hat{\sigma}^2$  is replaced by the determinant of  $\hat{\Sigma}$  (the MLE of the variance–covariance matrix of  $\varepsilon_t$ ), and the relevant number of parameters is the total free parameters of the VAR; thus, if there are no deterministic terms,  $IC(p) = \ln \det(\hat{\Sigma}) + n^2 p g(T)$ . The choice of series is typically guided by economic theory, although the predictive least squares (PLS) criterion (which is similar to an information criterion) can be useful in guiding this choice, cf. Wei (1992).

The issue of whether to difference the series is further complicated in the multivariate context by the possible presence of cointegration among two or more of the  $n$  variables. The multiple time series  $Y_t$  is said to be cointegrated if

each element of  $Y_t$  is integrated of order 1 (is  $I(1)$ ; that is, has an autoregressive unit root) but there are  $k \geq 1$  linear combination,  $\alpha'Y_t$ , that are  $I(0)$  (that is, which do not have a unit AR root) (Engle and Granger, 1987). It has been conjectured that long-run forecasts are improved by imposing cointegration when it is present. However, even if cointegration is correctly imposed, it remains to estimate the parameters of the cointegrating vector, which are, to first-order, estimated consistently (and at the same rate) if cointegration is not imposed. If cointegration is imposed incorrectly, however, asymptotically biased forecasts with large risks can be produced. At short horizons, these issues are unimportant to first-order asymptotically. By extension of the univariate results that are known for long-horizon forecasting, one might suspect that pretesting for cointegration could improve forecast performance, at least as measured by the asymptotic risk. However, tests for cointegration have very poor finite sample performance (cf. Haug, 1996), so it is far from clear that in practice pretesting for cointegration will improve forecast performance. Although much of the theory in this area has been worked out, work remains on assessing the practical benefits of imposing cointegration for forecasting. For additional discussions of cointegration, see Watson (1994), Hatanaka (1996) and Chapter 30 in this volume by Dolado, Gonzalo, and Marmol.

It should be noted that there are numerous subtle issues involved in the interpretation of and statistical inference for VARs. Watson (1994) surveys these issues, and two excellent advanced references on VARs and related small linear time series models are Lütkepohl (1993) and Reinsel (1993). Also, VARs provide only one framework for multivariate forecasting; for a different perspective to the construction of small linear forecasting models, see Hendry (1995).

## 5.2 Forecasting with leading economic indicators

Forecasting with leading economic indicators entails drawing upon a large number of time series variables that, by various means, have been ascertained to lead the variable of interest, typically taken to be aggregate output (the business cycle). The first set of leading economic indicators was developed as part of the business cycle research program at the National Bureau of Economic Research, and was published by Mitchell and Burns (1938). More recent works using this general approach include Stock and Watson (1989) and the papers in Moore and Lahiri (1991).

The use of many variables and little theory has the exciting potential to exploit relations not captured in small multivariate time series models. It is, however, particularly susceptible to overfitting within sample. For example, Diebold and Rudebusch (1991) found that although historical values of the Index of Leading Economic Indicators (then maintained by the US Department of Commerce) fits the growth in economic activity well, the real time, unrevised index has limited predictive content for economic activity. This seeming contradiction arises primarily from periodic redefinitions of the index. Their sobering finding underscores the importance of properly understanding the statistical properties of each stage of a model selection exercise. The development of methods for exploiting

large sets of leading indicators without overfitting is an exciting area of ongoing research.

### 5.3 Empirical examples

We now turn to an illustration of the performance of VARs as forecasting models. Like the experiment reported in Table 27.1, this experiment is simulated out of sample. Three families of VARs were specified. Using the numbering in Table 27.2, model (i) is a three-variable VAR with the unemployment rate, the interest rate, and the growth rate of industrial production. Model (j) is a three-variable VAR with the unemployment rate, CPI inflation, and the interest rate. Model (k) is a VAR with all five variables. The lags in all three models were kept the same in each equation of the VAR and were chosen recursively (at each forecast date, using only data through that date) by BIC, where  $1 \leq p \leq 6$  for models (i) and (j), and  $1 \leq p \leq 2$  for model (k). The VARs were estimated by OLS, equation by equation, with a one-step ahead specification, and six-month ahead forecasts were computed by iterating the one-month ahead forecasts.

The results are summarized in Table 27.2. In some cases, the VAR forecasts improve upon the AR forecasts. For example, for output growth, the VAR forecasts in (i) and (k) are respectively best and second-best of all the output growth forecasts in both tables. In contrast, for the interest rate, the VAR forecast is worse than the AR(4), and indeed the best forecast for the interest rate remains the EWMA forecast. However, the most notable feature of these forecasts is that eight of the eleven VAR forecasts in Table 27.2 have RMSFEs within 5 percent of the RMSFE of the AR(4), and all eleven have RMSFEs within 10 percent of the AR(4). For these specifications and these series, using additional information via a VAR results in forecasts that are essentially the same as those from an AR(4).

This finding, that forecasts from multivariate models often provide only modest improvements (or no improvement at all) over univariate forecasts, is not new.<sup>3</sup> One way to interpret this result is that additional macroeconomic series have little relationship to one another. This interpretation would, however, be incorrect, indeed, among the relationships in these VARs is the relation between the unemployment rate and inflation (the Phillips curve) and between interest rates and output (a channel of monetary policy), two links that have been studied in great detail and which are robust over the postwar period (cf. Stock and Watson, 1999b). An interpretation more in keeping with this latter evidence is that while these variables are related, there are sufficiently many parameters, which might not be stable over time, that these relations are not particularly useful for multivariate forecasting.

These negative results require some caveats. Supporters of VARs might suggest that the comparison in Table 27.2 is unfair because no attempt has been made to fine tune the VAR, to use additional variables, or to impose prior restrictions or prior information on the lag structure. This criticism has some merit, and methods which impose such structure, in particular Bayesian VARs, have a better track record than the unconstrained VARs reported here; see McNees (1990) and Sims (1993). Alternatively, others would argue that time series models

developed specifically for some variables, such as an empirical Phillips curve (as in Gordon, 1998), would be expected to work better than unfocused application of a VAR. This too might be valid, but in evaluating such claims one must take great care to distinguish between in-sample fit and the much more difficult task of forecasting well out-of-sample, either in real time or in a simulated out-of-sample experiment. Finally, it should be emphasized that these conclusions are for macroeconomic time series. For example, in industry applications one can find series with more pronounced nonlinearities.

## 6 DISCUSSION AND CONCLUSION

One of the few truly safe predictions is that economic forecasters will remain the target of jokes in public discourse. In part this arises from a lack of understanding that all forecasts must in the end be wrong, and that forecast error is inevitable. Economic forecasters can, however, bolster their credibility by providing information about the possible range of forecast errors. Some consumers are uncomfortable with forecast uncertainty: when his advisors presented a forecast interval for economic growth, President Lyndon Johnson is said to have replied, "ranges are for cattle." Yet communication of forecast uncertainty to those who rely on forecasts helps them to create better, more flexible plans and supports the credibility of forecasters more generally.

A theme of this chapter has been the tradeoff between complex models, which either use more information to forecast or allow subtle nonlinear formulations of the conditional mean, and simple models, which require fitting a small number of parameters and which thereby reduce parameter estimation uncertainty. The empirical results in Tables 27.1 and 27.2 provide a clear illustration of this tradeoff. The short-term interest rate is influenced by expected inflation, monetary policy, and the general supply and demand for funds, and, because the nominal rate must be positive, the "true" model for the interest rate must be nonlinear. Yet, of the autoregressions, neural nets, LSTAR models, and VARs considered in Tables 27.1 and 27.2, the best forecast was generated by a simple exponentially weighted moving average of past values of the interest rate. No attempt has been made to uncover the source of the relatively poor performance of the more sophisticated forecasts of the interest rate, but presumably it arises from a combination of parameter estimation error and temporal instability in the more complicated models.

An important practical question is how to resolve this tradeoff in practice. Two methods have been discussed here. At a formal level, this tradeoff is captured by the use of information criteria. Information criteria can be misleading, however, when many models are being compared and/or when the forecasting environment changes over time. The other method is to perform a simulated out-of-sample forecast comparison of a small number of models. This is in fact closely related to information criteria (Wei, 1992) and shares some of their disadvantages. When applied to at most a few candidate models, however, this has the advantage of providing evidence on recent forecasting performance and how the forecasting performance of a model has evolved over the simulated forecast

period. These observations, along with those above about reporting forecast uncertainty, suggest a simple rule: even if your main interest is in more complicated models, it pays to maintain benchmark forecasts using a simple model with honest forecast standard errors evaluated using a simulated real time experiment, and to convey the forecast uncertainty to the consumer of the forecast.

Finally, an important topic not addressed in this chapter is model instability. All forecasting models, no matter how sophisticated, are stylized and simplified ways to capture the complex and rich relations among economic time series variables. There is no particular reason to believe that these underlying relations are stable – technology, global trade, and macroeconomic policy have all evolved greatly over the past three decades – and even if they were, the implied parameters of the forecasting relations need not be stable. One therefore would expect estimated forecasting models to have parameters that vary over time, and in fact this appears to be the case empirically (Stock and Watson, 1996). Indeed, Clements and Hendry (1999) argue that most if not all major economic forecast failures arise because of unforeseen events that lead to a breakdown of the forecasting model; they survey existing methods and suggest some new techniques for detecting and adjusting to such structural shifts. The question of how best to forecast in a time-varying environment remains an important area of econometric research.

## Notes

- \* The author thanks Lewis Chan for research assistance and four anonymous referees for useful suggestions.
- 1 All series were obtained from the Basic Economics Database maintained by DRI/McGraw Hill. The series mnemonics are: PUNEW (the CPI); IP (industrial production); LHUR (the unemployment rate); FYGM3 (the 90 day U.S. Treasury bill rate); and IVMTQ (real manufacturing and trade inventories).
- 2 These results are drawn from the much larger model comparison exercise in Stock and Watson (1999a), to which the reader is referred for additional details on estimation method, model definitions, data sources, etc.
- 3 In influential work, Cooper (1972) and Nelson (1972) showed this in a particularly dramatic way. They found that simple ARMA models typically produced better forecasts of the major macroeconomic aggregates than did the main large structural macroeconomic models of the time. For a discussion of these papers and the ensuing literature, see Granger and Newbold (1986, ch. 9.4).

## References

- Bates, J.M., and C.W.J. Granger (1969). The combination of forecasts. *Operations Research Quarterly* 20, 451–68.
- Bollerslev, T., R.F. Engle, and D.B. Nelson (1994). ARCH models. In R. Engle and D. McFadden (eds.), *Handbook of Econometrics*, Volume 4, pp. 2959–3038. Amsterdam: Elsevier.
- Brockwell, P.J., and R.A. Davis (1987). *Time Series: Theory and Methods*. New York: Springer-Verlag.

- Chan, Y.L., J.H. Stock, and M.W. Watson (1998). A dynamic factor model framework for forecast combination. *Spanish Economic Review* 1, 91–121.
- Christoffersen, P.F., and F.X. Diebold (1997). Optimal prediction under asymmetric loss. *Econometric Theory* 13, 808–17.
- Clements, M., and D.F. Hendry (1996). Multi-step estimation for forecasting. *Oxford Bulletin of Economics and Statistics* 58, 657–84.
- Clements, M., and D.F. Hendry (1999). *Forecasting Non-Stationary Economic Time Series*. Cambridge: MIT Press.
- Cooper, R.L. (1972). The predictive performance of quarterly econometric models of the United States. In B.G. Hickman (ed.) *Econometric Models of Cyclical Behavior*. New York: Columbia University Press.
- Davis, R.A., and W.T.M. Dunsmuir (1996). Maximum likelihood estimation for MA(1) processes with a root on or near the unit circle. *Econometric Theory* 12, 1–29.
- Diebold, F.X., T. Gunther, and A.S. Tay (1998). Evaluating density forecasts with applications to financial risk management. *International Economic Review* 39, 868–83.
- Diebold, F.X., and J.A. Lopez (1995). Forecast evaluation and combination. In G.S. Maddala and C.R. Rao (eds.) *Handbook of Statistics* 14, 241–68.
- Diebold, F.X., and R. Mariano (1995). Comparing predictive accuracy. *Journal of Business and Economic Statistics* 13, 253–63.
- Diebold, F.X., and G. Rudebusch (1991). Forecasting output with the composite leading index: a real time analysis. *Journal of the American Statistical Association* 86, 603–10.
- Elliott, G., T.J. Rothenberg, and J.H. Stock (1996). Efficient tests for an autoregressive unit root. *Econometrica* 64, 813–36.
- Engle, R.F. (1982). Autoregressive conditional heteroskedasticity with estimates of the variance of U.K. inflation. *Econometrica* 50, 987–1008.
- Engle, R.F., and C.W.J. Granger (1987). Co-integration and error correction: representation, estimation and testing. *Econometrica* 55, 251–76.
- Fair, R.C. (1980). Evaluating the predictive accuracy of econometric models. *International Economic Review* 21, 355–78.
- Geweke, J., and R. Meese (1981). Estimating regression models of finite but unknown order. *International Economic Review* 22, 55–70.
- Gordon, R.J. (1998). Foundations of the Goldilocks economy: supply shocks and the time-varying NAIRU. *Brookings Papers on Economic Activity* 2, 297–333.
- Granger, C.W.J. (1969). Prediction with a generalized cost of error function. *Operational Research Quarterly* 20, 199–207.
- Granger, C.W.J. (1989). Combining forecasts – twenty years later. *Journal of Forecasting* 8, 167–73.
- Granger, C.W.J., and P. Newbold (1986). *Forecasting Economic Time Series*, 2nd edn. Orlando: Academic Press.
- Granger, C.W.J., and T. Teräsvirta (1993). *Modelling Non-linear Economic Relationships*. Oxford: Oxford University Press.
- Granger, C.W.J., T. Teräsvirta, and H.M. Anderson (1993). Modeling nonlinearity over the business cycle. In J.H. Stock and M.W. Watson (eds.) *Business Cycles, Indicators, and Forecasting*, pp. 311–27. Chicago: University of Chicago Press for the NBER.
- Hamilton, J.D. (1994). *Time Series Analysis*. Princeton: Princeton University Press.
- Hamilton, J.D. (1989). A new approach to the economic analysis of nonstationary time series and the business cycle. *Econometrica* 57, 357–84.
- Harvey, A.C. (1989). *Forecasting, Structural Time Series Models and the Kalman Filter*. Cambridge: Cambridge University Press.
- Hatanaka, M. (1996). *Time-Series-Based Econometrics: Unit Roots and Cointegration*. Oxford: Oxford University Press.

- Haug, A.A. (1996). Tests for cointegration: a Monte Carlo comparison. *Journal of Econometrics* 71, 89–115.
- Hendry, D.F. (1995). *Dynamic Econometrics*. Oxford: Oxford University Press.
- Kim, C.-J., and C.R. Nelson (1998). Business cycle turning points, a new coincident index, and tests of duration dependence based on a dynamic factor model with regime switching. *Review of Economics and Statistics* 80, 188–201.
- Kim, C.-J., and C.R. Nelson (1999). *State-Space Models with Regime Switching: Classical and Gibbs Sampling Approaches with Applications*. Cambridge: MIT Press.
- Lütkepohl, H. (1993). *Introduction to Multiple Time Series Analysis*, 2nd edn. New York: Springer-Verlag.
- McNees, S.K. (1990). The role of judgment in macroeconomic forecasting accuracy. *International Journal of Forecasting* 6, 287–99.
- Meese, R., and J. Geweke (1984). A comparison of autoregressive univariate forecasting procedures for macroeconomic time series. *Journal of Business and Economic Statistics* 2, 191–200.
- Mitchell, W.C., and A.F. Burns (1938). *Statistical Indicators of Cyclical Revivals*. NBER Bulletin 69, New York. Reprinted in G.H. Moore (ed.) *Business Cycle Indicators*. Princeton: Princeton University Press, 1961.
- Moore, G., and K. Lahiri (eds.) (1991). *The Leading Economic Indicators: New Approaches and Forecasting Records*. Cambridge: Cambridge University Press.
- Montgomery, A.L., V. Zarnowitz, R.S. Tsay, and G.C. Tiao (1998). Forecasting the U.S. unemployment rate. *Journal of the American Statistical Association* 93, 478–93.
- Nelson, C.R. (1972). The prediction performance of the FRB-MIT-PENN model of the U.S. economy. *American Economic Review* 62, 902–17.
- Ng, S., and P. Perron (1995). Unit root tests in ARMA models with data dependent methods for the truncation lag. *Journal of the American Statistical Association* 90, 268–81.
- Priestly, M.B. (1989). *Non-linear and Non-stationary Time Series Analysis*. London: Academic Press.
- Reinsel, G.C. (1993). *Elements of Multivariate Time Series Analysis*. New York: Springer-Verlag.
- Samorodnitsky, G., and M.S. Taqqu (1994). *Stable Non-Gaussian Random Processes*. New York: Chapman & Hall.
- Schwarz, G. (1978). Estimating the dimension of a model. *Annals of Statistics* 6, 461–64.
- Sims, C.A. (1980). Macroeconomics and reality. *Econometrica* 48, 1–48.
- Sims, C.A. (1993). A nine-variable probabilistic macroeconomic forecasting model. In J.H. Stock and M.W. Watson (eds.) *Business Cycles, Indicators and Forecasting*. Chicago: University of Chicago Press for the NBER.
- Stock, J.H. (1994). Unit roots, structural breaks, and trends. In R. Engle and D. McFadden (eds.) *Handbook of Econometrics*, Volume 4, pp. 2740–843. Amsterdam: Elsevier.
- Stock, J.H. (1996). VAR, error correction and pretest forecasts at long horizons, *Oxford Bulletin of Economics and Statistics* 58, 685–701. Reprinted in A. Banerjee and D.F. Hendry (eds.) *The Econometrics of Economic Policy*, pp. 115–32. Oxford: Basil Blackwell, 1997.
- Stock, J.H., and M.W. Watson (1989). New indexes of coincident and leading economic indicators. *NBER Macroeconomics Annual* 351–93.
- Stock, J.H., and M.W. Watson (1996). Evidence on structural instability in macroeconomic time series relations. *Journal of Business and Economic Statistics*, 14, 11–30.
- Stock, J.H., and M.W. Watson (1998). Median unbiased estimation of coefficient variance in a time varying parameter model. *Journal of the American Statistical Association* 93, 349–58.
- Stock, J.H., and M.W. Watson (1999a). A comparison of linear and nonlinear univariate models for forecasting macroeconomic time series. Chapter 1 in R. Engle and H. White

- (eds.) *Cointegration, Causality and Forecasting: A Festschrift for C.W.J. Granger*. Oxford: Oxford University Press, 1–44.
- Stock, J.H., and M.W. Watson (1999b). Business cycle fluctuations in U.S. macroeconomic time series. Chapter 1 in J. Taylor and M. Woodford (eds.) *Handbook of Macroeconomics*. Amsterdam: Elsevier, 3–64.
- Swanson, N.R., and H. White (1995). A model selection approach to assessing the information in the term structure using linear models and artificial neural networks. *Journal of Business and Economic Statistics* 13, 265–75.
- Swanson, N.R., and H. White (1997). A model selection approach to real-time macroeconomic forecasting using linear models and artificial neural networks. *Review of Economics and Statistics* 79, 540–50.
- Watson, M.W. (1994). Vector autoregressions and cointegration. In R. Engle and D. McFadden (eds.) *Handbook of Econometrics*, Volume 4, pp. 2844–915. Amsterdam: Elsevier.
- Wei, C.Z. (1992). On predictive least squares principles. *The Annals of Statistics* 20, 1–42.
- West, K. (1996). Asymptotic inference about predictive ability. *Econometrica* 64, 1067–84.
- West, K., H.J. Edison, and D. Cho (1993). A utility based evaluation of some models of exchange rate variability. *Journal of International Economics* 35, 23–46.
- Zarnowitz, V., and Braun (1993). Twenty-two years of the NBER-ASA quarterly economic outlook surveys: aspects and comparisons of forecasting performance. In J.H. Stock and M.W. Watson (eds.) *Business Cycles, Indicators and Forecasting*. Chicago: University of Chicago Press for the NBER.

---

CHAPTER TWENTY-EIGHT

# Time Series and Dynamic Models

*Aris Spanos*

## 1 INTRODUCTION

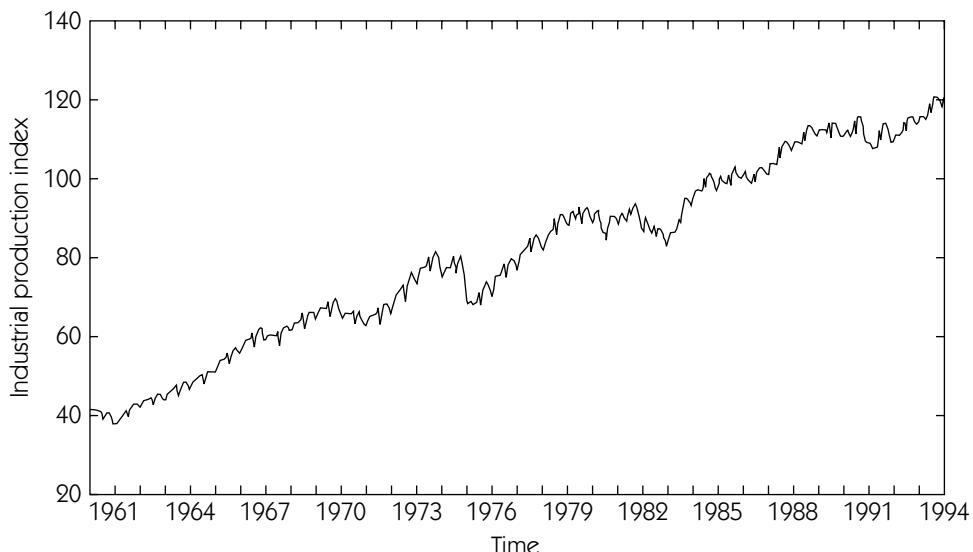
This chapter discusses certain dynamic statistical models of interest in modeling time series data. Particular emphasis is placed on the problem of *statistical adequacy*: the postulated model does not exhibit departures from the underlying assumptions. Statistical models are specified in terms of probabilistic assumptions on the observable stochastic processes involved, which can often be assessed a priori using graphical techniques. The primary objective is to render empirical modeling of time series an informed systematic procedure that gives rise to reliable empirical evidence.

### 1.1 Time series: a brief historical introduction

Time series data have been used since the dawn of empirical analysis in the mid-seventeenth century. In the “Bills of Mortality” John Graunt compared data on births and deaths over the period 1604–60 and across regions (parishes); see Stigler (1986). The time dimension of such data, however, was not properly understood during the early stages of empirical analysis. Indeed, it can be argued that the time dimension continued to bedevil statistical analysis for the next three centuries before it could be tamed in the context of proper statistical models for time series data.

#### THE DESCRIPTIVE STATISTICS PERIOD: 1665–1926

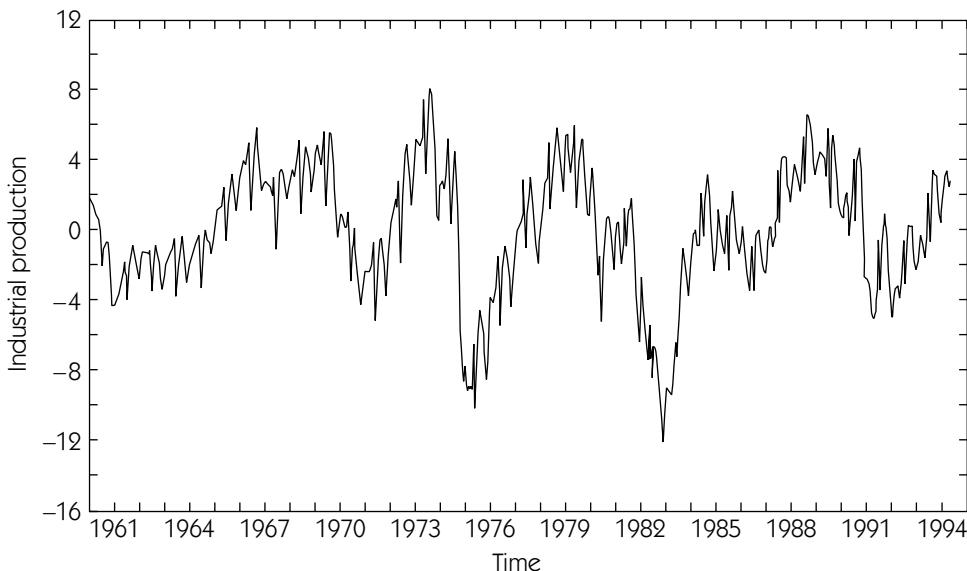
Up until the last quarter of the nineteenth century the time dimension of observed data and what that entails was not apparent to the descriptive statistics literature which concentrated almost exclusively in looking at histograms and



**Figure 28.1** US industrial production index

certain associated numerical characteristics such as the mean and variance. By its very nature, histogram analysis and the associated descriptive statistics suppress the time dimension and concentrate on a single aspect of the data generating process, the *distribution*. Two other aspects raised by the time dimension, the *dependence* and *heterogeneity* with respect to the time index, were largely ignored, because implicit in this literature is the assumption that data exhibit independence and complete homogeneity. Questions concerning the *temporal independence/homogeneity* of time series data were first explicitly raised in the last quarter of the nineteenth century by Lexis and Bienayme (see Heyde and Seneta, 1977, ch. 3). By the mid-nineteenth century it became apparent that comparisons over time required a certain stability (temporal independence/homogeneity) of the measurements being compared; births, deaths, accidents, suicides, and murders. The proposed tests for stability took the form of comparing the sample variance of the time series in question with that of a stable (independent and homogeneous) binomially distributed process. The results on the basis of such a comparison were very discouraging, however, because, with the exception of the ratio of male to female births, it suggested that all other observed time series appeared to exhibit some form of instability.

This apparent instability was, at the time, associated with the cycles and trends exhibited by time series  $\{y_1, y_2, \dots, y_T\}$  when plotted over time ( $t$ -plot); in Figure 28.1 we can see a typical economic time series, the monthly US industrial production index for the period 1960–94. The two chance regularity patterns one can see in Figure 28.1 is the secular trend (increasing function of  $t$ ) and the cycles around this trend. These cycles become more apparent when the data are de-trended as shown in Figure 28.2. By the end of the nineteenth century a time



**Figure 28.2** De-trended industrial production index

series was perceived as made up of three different components. As summarized by Davis:

The problem of single time series, . . . , is concerned with three things: first, the determination of a trend; second, the discovery and interpretation of cyclical movements in the residuals; third, the determination of the magnitude of the erratic element in the data. (Davis, 1941, p. 59)

In view of this, it was only natural that time series focused on discovering the presence and capturing this instability (trends and cycles). In the early twentieth century the attempt to model observed cycles took two alternative (but related) forms. The first form attempted to capture the apparent cycles using sinusoidal functions (see Schuster, 1906). The objective was to discover any hidden periodicities using a technique appropriately called the *periodogram*. The second way to capture cycles came in the form of the temporal correlation, the *autocorrelation* coefficients, and their graph the *correlogram*; see Granger and Newbold (1977). This was a simple adaptation of Galton's (contemporaneous) correlation coefficient.

The early empirical studies indicated that the periodogram appeared to be somewhat unrealistic for economic time series because the harmonic scheme assumes strict periodicity. The correlogram was also partly unsuccessful because the sample correlogram gave rise to *spurious* correlations when it was applied to economic time series. Various techniques were suggested at the time in an attempt to deal with the spurious correlation problem, the most widely used

being the differencing of the series and evaluating the correlogram using the differenced series (see Norton, 1902, Hooker, 1905). The first important use of these techniques with economic data was by Moore (1914) who attempted to discover the temporal interdependence among economic time series using both the periodogram and the temporal correlations. It was felt at the time that the problem of spurious correlation arises from the fact that time series are functions of time but there was no apparent functional form to be used in order to capture that effect; see Hendry and Morgan (1995). This literature culminated with the classic papers of Yule (1921, 1926) where the *spurious correlation* (and regression) problem was diagnosed as due to the apparent departures (exhibited by economic time series) from the assumptions required to render correlation analysis valid. Yule (1921) is a particularly important paper because it constitutes the first systematic attempt to relate misleading results of statistical analysis to the invalidity of the underlying probabilistic assumptions: *misspecification*. To this day, the problem of spurious correlation as due to the departures of probabilistic assumptions, necessary to render correlation analysis valid, is insufficiently understood. This is because the modeler is often unaware of the underlying probabilistic assumptions whose validity renders the analysis reliable. Such a situation arises when the statistical model is not explicitly specified and thus no assessment of the underlying assumptions can be conducted in order to ensure their validity. For instance, the sample (contemporaneous) correlation coefficient between two different series  $\{(x_t, y_t), t = 1, 2, \dots, T\}$  is a meaningful measure of first-order dependence only in cases where the means of both processes underlying the data are constant over  $t$ :  $E(X_t) = \mu_x, E(Y_t) = \mu_y$ , for all  $t \in \mathbb{T}$ ; otherwise the measure is likely to be misleading.

### THE STATISTICAL MODELING (PROPER) PERIOD: 1927–PRESENT

The formulation of explicit statistical models for time series began with the classic papers of Yule (1927) (Autoregressive (AR( $p$ )) scheme):

$$y_t = \alpha_0 + \sum_{k=1}^p \alpha_k y_{t-k} + \varepsilon_t, \quad \varepsilon_t \sim NI(0, \sigma^2), \quad t = 1, 2, \dots,$$

where “NI” stands for “Normal, Independent” and Slutsky (1927) (Moving Average (MA( $q$ ))) scheme:

$$y_t = \gamma_0 + \sum_{k=1}^q \gamma_k \varepsilon_{t-k} + \varepsilon_t, \quad \varepsilon_t \sim NI(0, \sigma^2), \quad t = 1, 2, \dots$$

Viewing these formulations from today’s vantage point, it is clear that, at the time, they were proposed as nothing more than convenient descriptive models for time series data. Their justification was based exclusively on the fact that when simulated these schemes gave rise to data series which appear to exhibit cycles similar to those observed in actual time series data. It should be noted that, at the time, the difference between regular cycles due to seasonality and irregular cycles due to positive dependence was insufficiently understood.

The first attempt to provide a proper probabilistic foundation for these schemes is undoubtedly that of Wold (1938) who successfully fused these schemes with the appropriate probabilistic concepts necessary to model the chance regularity patterns exhibited by certain time series data. The appropriate probabilistic concepts were developed by Kolmogorov (1933) and Khinchine (1932). H. Cramer (1937) was instrumental in providing the missing link between the empirical literature on time series and the mathematical literature on *stochastic processes*; the probabilistic framework of modeling time series. Wold (1938), in his Ph.D. under Cramer, proposed the first proper statistical framework for modeling *stationary* time series. The lasting effect of Wold's work comes in the form of (i) his celebrated decomposition theorem (see Section 4), where under certain regularity restrictions a stationary process can be represented in the form of a  $\text{MA}(\infty)$ :

$$y_t = \gamma_0 + \sum_{k=1}^{\infty} \gamma_k \varepsilon_{t-k} + \varepsilon_t, \quad \sum_{k=1}^{\infty} |\gamma_k| < \infty, \quad \varepsilon_t \sim \text{NI}(0, \sigma^2), \quad t = 1, 2, \dots,$$

and (ii) the  $\text{ARMA}(p, q)$  model:

$$y_t = \alpha_0 + \sum_{k=1}^p \alpha_k y_{t-k} + \sum_{k=1}^q \gamma_k \varepsilon_{t-k} + \varepsilon_t, \quad \varepsilon_t \sim \text{NI}(0, \sigma^2), \quad t = 1, 2, \dots,$$

appropriate for time series exhibiting stationarity and weak dependence.

Wold's results provided the proper probabilistic foundations for the empirical analysis based on both the periodogram and the correlogram; the autocorrelations are directly related to the coefficients  $(\alpha_0, \alpha_1, \dots, \alpha_p, \gamma_1, \gamma_2, \dots, \gamma_q)$  and the periodogram can be directly related to the spectral representation of stationary stochastic processes; see Anderson (1971) for an excellent discussion.

The mathematical foundations of stationary stochastic processes were strengthened as well as delineated further by Kolmogorov (1941) but the  $\text{ARMA}(p, q)$  formulation did not become an empirical success for the next three decades because the overwhelming majority of times series appear to exhibit temporal heterogeneity (nonstationarity), rendering this model inappropriate. The state of the art, as it relates to the  $\text{AR}(p)$  family of models, was described at the time in the classic paper by Mann and Wald (1943).

The only indirect influence of the  $\text{ARMA}(p, q)$  family of models to econometric modeling came in the form of an extension of the linear regression model by adding an  $\text{AR}(1)$  model for the error term:

$$y_t = \beta^\top x_t + u_t, \quad u_t = \rho u_{t-1} + \varepsilon_t, \quad |\rho| < 1, \quad \varepsilon_t \sim \text{NI}(0, \sigma^2), \quad t = 1, 2, \dots, T. \quad (28.1)$$

This model provided the basis of the well known Durbin–Watson bounds test (see Durbin and Watson, 1950, 1951), which enabled econometricians to test for the presence of temporal dependence in the context of the linear regression model. This result, in conjunction with Cochrane and Orcutt (1949) who suggested a

way to estimate the hybrid model (28.1), offered applied econometricians a way to use time series data in the context of the linear regression model without having to worry about *spurious regressions* (against which Yule (1926) cautioned time series analysts).

The important breakthrough in time series analysis came with Box and Jenkins (1970) who re-invented and popularized differencing as a way to deal with the apparent non-stationarity of time series:

$$\Delta^d y_t, \text{ where } \Delta := (1 - L), \quad d \text{ is a positive integer}, \quad t = 1, 2, \dots,$$

in order to achieve (trend) stationarity and seasonal differencing of period  $s$ :

$$\Delta_s y_t := (1 - L^s)y_t = (y_t - y_{t-s}),$$

to achieve (seasonal) stationarity. They proposed the ARMA( $p, q$ ) model for the differenced series  $\Delta_s^d y_t := y_t^* \equiv y_t^*$ , giving rise to the ARIMA( $p, d, q$ ) model:

$$y_t^* = \alpha_0 + \sum_{k=1}^p \alpha_k y_{t-k}^* + \sum_{k=1}^q \gamma_k \varepsilon_{t-k} + \varepsilon_t, \quad \varepsilon_t \sim NI(0, \sigma^2), \quad t = 1, 2, \dots$$

Box and Jenkins (1970) did not just propose a statistical model but a modeling strategy revolving around the ARIMA( $p, d, q$ ) model. This modeling procedure involved three stages. The first was *identification*: the choice of  $(p, d, q)$  using graphical techniques, the autocorrelations (correlogram) and the partial autocorrelations. The second stage was the *diagnostic checking* in order to assess the validity of the assumptions underlying the error term. The third stage was a formal *forecasting* procedure based on the estimated model. It must be noted that up until the 1970s the empirical forecasting schemes were based on *ad hoc* moving averages and exponential smoothing. The Box–Jenkins modeling strategy had a lasting effect on econometric modeling because it brought out an important weakness in econometric modeling: the insufficient attention paid to the temporal dependence/heterogeneity exhibited by economic time series; see Spanos (1986, 1987) for further discussion. This weakness was instrumental in giving rise to the LSE modeling methodology (see Hendry, 1993, for a collection of papers) and the popularization of the vector autoregressive (VAR( $p$ )) model by Sims (1980).

The next important development in time series modeling came in the form of *unit root testing* in the context of the AR( $p$ ) model proposed by Dickey and Fuller (1979, 1981). The proposed tests provided the modeler with a way to decide whether the time series in question was stationary, trend nonstationary or unit root nonstationary. Since the 1930s it has been generally accepted that most economic time series can be viewed as stationary around a deterministic trend. Using the Dickey–Fuller testing procedures, Nelson and Plosser (1982) showed that, in contrast to conventional wisdom, most economic time series can be better described by the AR( $p$ ) model with a unit root. These results set off an explosion of time series research which is constantly revisiting and reconsidering the initial results by developing new testing procedures and techniques.

Phillips (1986, 1987) dealt effectively with the *spurious correlation* (regression) problem which was revisited in Granger and Newbold (1974), by utilizing and extending the Dickey and Fuller (1979, 1981) asymptotic distribution results. The same results were instrumental in formulating the notion of *cointegration* among time series with unit roots (see Granger, 1983; Engle and Granger, 1987; Johansen, 1991; Phillips, 1991) and explaining the empirical success of the error-correction models proposed by the LSE tradition (see Hendry, 1993).

## 2 A PROBABILISTIC FRAMEWORK FOR TIME SERIES

A time series is defined as a finite sequence of observed data  $\{y_1, y_2, \dots, y_T\}$  where the index  $t = 1, 2, \dots, T$ , denotes time. The probabilistic concept which corresponds to this series is that of a real stochastic process:  $\{Y_t, t \in \mathbb{T}\}$ , defined on the probability space  $(S, \mathfrak{F}, \mathbb{P}(\cdot)) : Y(\cdot, \cdot) : \{S \times \mathbb{T}\} \rightarrow \mathbb{R}_Y$ , where  $S$  denotes the outcomes set,  $\mathfrak{F}$  is the relevant  $\sigma$ -field,  $\mathbb{P}(\cdot) : \mathfrak{F} \rightarrow [0, 1]$ ,  $\mathbb{T}$  denotes the relevant index set (often  $\mathbb{T} := \{1, 2, \dots, T, \dots\}$ ) and  $\mathbb{R}_Y$  denotes a subset of the real line.

The probabilistic foundation of stochastic processes comes in the form of the finite joint distribution of the process  $\{Y_t, t \in \mathbb{T}\}$  as formulated by Kolmogorov (1933) in the form of *Kolmogorov's extension theorem*. According to this result, the probabilistic structure of a stochastic process (under certain mild conditions) can be fully described by a finite dimensional joint distribution of the form:

$$f(y_1, y_2, \dots, y_T; \psi), \quad \text{for all } (y_1, y_2, \dots, y_T) \in \mathbb{R}_Y^T. \quad (28.2)$$

This joint distribution provides the starting point for the *probabilistic reduction (PR) approach* to modeling. The probabilistic structure of a stochastic process can be conveniently defined in the context of the joint distribution by imposing certain probabilistic assumptions from the following three categories:

*Distribution:* normal, student's  $t$ , gamma, beta, logistic, exponential, etc.

*Dependence:* Markov( $p$ ), ergodicity,  $m$ -dependence, martingale, mixing, etc.

*Heterogeneity:* identically distributed, stationarity (strict,  $k$ th order), etc.

Time series modeling can be viewed as choosing an appropriate statistical model which captures all the systematic information in the observed data series. Systematic statistical information comes to the modeler in the form of chance regularity patterns exhibited by the time series data. For example, the cycles exhibited by the time series data in Figures 28.1–28.2 constitute a chance regularity pattern associated with positive autocorrelation because they are not deterministic cycles that would indicate seasonality. For an extensive discussion on numerous chance regularity patterns and how they can be detected using a variety of graphical techniques including  $t$ -plots, scatter-plots, P-P and Q-Q plots, see Spanos (1999, chs 5–6).

The success for empirical modeling depends crucially on both recognizing these regularity patterns and then choosing the appropriate probabilistic concepts (in the form of assumptions) in order to model this information. The choice

of these probabilistic assumptions amounts to specifying an appropriate statistical model. In the context of the PR approach all possible models ( $\mathcal{P}$ ) are viewed as reductions from the joint distribution (28.2). That is, the chosen model  $P_0 \in \mathcal{P}$ , constitutes an element which arises by imposing certain reduction assumptions from the above three categories on the process  $\{Y_t, t \in \mathbb{T}\}$ . This methodological perspective, introduced by Spanos (1986), differs from the traditional view in so far as it does not view these models as just stochastic equations (linear, difference, differential, integral, etc.). In the next section we discuss the two alternative perspectives using the AR(1) model.

### 3 AUTOREGRESSIVE MODELS: UNIVARIATE

The objective of this section is to provide a brief overview of the most commonly used time series model, the AR(1), from both, the traditional and the probabilistic reduction (PR) perspectives.

#### 3.1 AR(1): the traditional perspective

The traditional economic theory perspective for an AR(1) time series model was largely determined by the highly influential paper by Frisch (1933) as a linear (constant coefficient) stochastic difference equation:

$$y_t = \alpha_0 + \alpha_1 y_{t-1} + \varepsilon_t, \quad |\alpha_1| < 1, \quad \varepsilon_t \sim NI(0, \sigma^2), \quad t = 1, 2, \dots, \quad (28.3)$$

with the probabilistic assumptions specified via the error process  $\{\varepsilon_t, t \in \mathbb{T}\}$ :

$$\left. \begin{array}{ll} 1^a & \text{zero mean:} \\ 2^a & \text{constant variance:} \\ 3^a & \text{no autocorrelation:} \\ 4^a & \text{Normality:} \end{array} \right\} \begin{array}{l} E(\varepsilon_t) = 0, \\ E(\varepsilon_t^2) = \sigma^2, \\ E(\varepsilon_t \varepsilon_{t-\tau}) = 0, \tau \neq 0, \\ \varepsilon_t \sim N(\cdot, \cdot), \end{array} \quad t \in \mathbb{T}. \quad (28.4)$$

These assumptions define the error  $\{\varepsilon_t, t \in \mathbb{T}\}$  to be a Normal, white noise process. The formulation (28.3) is then viewed as a data generating mechanism (DGM) from right to left, the input being the error process (and the initial condition  $y_0$ ) and the output the observable process  $\{y_t, t \in \mathbb{T}\}$ . The probabilistic structure of the latter process is generated from that of the error process via (28.3) by recursive substitution:

$$y_t = \alpha_0 + \alpha_1 y_{t-1} + \varepsilon_t = \alpha_1^t y_0 + \alpha_0 \left( \sum_{i=0}^{t-1} \alpha_1^i \right) + \left( \sum_{i=0}^{t-1} \alpha_1^i \varepsilon_{t-i} \right), \quad (28.5)$$

yielding the first two moments:

$$E(y_t) = \alpha_1^t E(y_0) + \alpha_0 \left( \sum_{i=0}^{t-1} \alpha_1^i \right), \quad \text{cov}(y_t, y_{t+\tau}) = \sigma^2 \alpha_1^{\tau} \left( \sum_{i=0}^{t-1} \alpha_1^{2i} \right), \quad \tau \geq 0.$$

Using the restriction  $|\alpha_1| < 1$  we can simplify these to:

$$E(y_t) = \alpha_1^t E(y_0) + \alpha_0 \left( \frac{1 - \alpha_1^t}{1 - \alpha_1} \right), \quad \text{cov}(y_t, y_{t+\tau}) = \sigma^2 \alpha_1^\tau \left( \frac{1 - \alpha_1^{2(t-\tau)}}{1 - \alpha_1^2} \right), \quad \tau \geq 0. \quad (28.6)$$

Viewed in terms of its first two moments, the stochastic process  $\{y_t, t \in \mathbb{T}\}$  is both normal and Markov but second-order time heterogeneous. Traditionally, however, the time heterogeneity is sidestepped by focusing on the "steady-state":

$$\lim_{t \rightarrow \infty} E(y_t) = \frac{\alpha_0}{(1 - \alpha_1)}, \quad \lim_{t \rightarrow \infty} \text{var}(y_t) = \frac{\sigma^2}{(1 - \alpha_1^2)}, \quad \lim_{t \rightarrow \infty} \text{cov}(y_t, y_{t+\tau}) = \alpha_1^{|\tau|} \left( \frac{\sigma^2}{1 - \alpha_1^2} \right). \quad (28.7)$$

Hence, the (indirect) probabilistic assumptions underlying the observable process  $\{y_t, t \in \mathbb{T}\}$ , generated via the AR(1) (28.3), are:

$$\left. \begin{array}{ll} 1^b \text{ constant mean:} & E(y_t) := \mu = \frac{\alpha_0}{(1 - \alpha_1)}, \\ 2^b \text{ constant variance:} & \text{var}(y_t) := \sigma_0 = \frac{\sigma^2}{(1 - \alpha_1^2)}, \\ 3^b \text{ Markov autocorrelation:} & \text{cov}(y_t, y_{t-\tau}) := \sigma_{|\tau|} = \alpha_1^{|\tau|} \left( \frac{\sigma^2}{1 - \alpha_1^2} \right), \tau \neq 0, \\ 4^b \text{ Normality:} & y_t \sim N(\cdot, \cdot), \end{array} \right\} t \in \mathbb{T}. \quad (28.8)$$

As argued in the next subsection, the probabilistic reduction perspective reverses this viewpoint and contemplates (28.3) in terms of probabilistic assumptions regarding the process  $\{y_t, t \in \mathbb{T}\}$  and not the error process  $\{\epsilon_t, t \in \mathbb{T}\}$ . It must be emphasized that the probabilistic perspective provides an alternative viewpoint for statistical models which has certain advantages over the traditional theory viewpoint when the statistical aspects of modeling are of interest. In contrast, the traditional theory viewpoint has certain advantages when other aspects of modeling, such as the system properties, are of interest. Hence, the two viewpoints are considered as complimentary.

### 3.2 AR(1): the probabilistic reduction perspective

The probabilistic reduction perspective has been developed in Spanos (1986). This perspective begins with the observable process  $\{y_t, t \in \mathbb{T}\}$  and specifies the statistical model exclusively in terms of this process. In particular, it contemplates the DGM (28.3) from left to right as an orthogonal decomposition of the form:

$$y_t = E(y_t | \sigma(Y_{t-1}^0)) + u_t, t \in \mathbb{T}, \quad (28.9)$$

where  $Y_{t-1}^0 := (y_{t-1}, y_{t-2}, \dots, y_0)$  and  $u_t = y_t - E(y_t | \sigma(Y_{t-1}^0))$ , with the underlying statistical model viewed as a reduction from the joint distribution of the underlying process  $\{y_t, t \in \mathbb{T}\}$ . The form of the autoregressive function depends on:

$$f(y_0, y_1, y_2, \dots, y_T; \psi), \text{ for all } (y_0, y_1, y_2, \dots, y_T) \in \mathbb{R}^{T+1},$$

in the sense of Kolmogorov (1933). In the present case, the *reduction assumptions* on the joint distribution of the process  $\{y_t, t \in \mathbb{T}\}$ , that would yield (28.3) are: (i) normal, (ii) Markov, and (iii) stationary.

Let us consider this in some detail. The assumption of Markovness for the underlying process enables one to concentrate on bivariate distributions since:

$$f(y_0, y_1, y_2, \dots, y_T; \psi) = f(y_0; \phi_0) \prod_{t=1}^T f(y_t | y_{t-1}; \phi), (y_0, y_1, y_2, \dots, y_T) \in \mathbb{R}^{T+1}. \quad (28.10)$$

The underlying bivariate distribution for model (28.3) is:

$$\begin{bmatrix} y_t \\ y_{t-1} \end{bmatrix} \sim N\left(\begin{bmatrix} \mu \\ \mu \end{bmatrix}, \begin{bmatrix} \sigma_0 & \sigma_1 \\ \sigma_1 & \sigma_0 \end{bmatrix}\right), \quad t \in \mathbb{T}, \quad (28.11)$$

which, via the orthogonal decomposition (28.9) gives rise to:

$$y_t = \alpha_0 + \alpha_1 y_{t-1} + u_t, \quad t \in \mathbb{T}. \quad (28.12)$$

The statistical parameters  $\phi := (\alpha_0, \alpha_1, \sigma^2)$  are related to the primary parameters  $\psi := (\mu, \sigma_0, \sigma_1)$  via:

$$\alpha_0 = (1 - \frac{\sigma_1}{\sigma_0})\mu \in \mathbb{R}, \quad \alpha_1 = \frac{\sigma_1}{\sigma_0} \in (-1, 1), \quad \sigma^2 = \sigma_0 - \frac{\sigma_1^2}{\sigma_0} \in \mathbb{R}_+, \quad (28.13)$$

and thus the two parameter spaces take the form:

$$\psi := (\mu, \sigma_0, \sigma_1) \in \mathbb{R}^2 \times \mathbb{R}_+, \quad \phi := (\alpha_0, \alpha_1, \sigma^2) \in \mathbb{R} \times (-1, 1) \times \mathbb{R}_+.$$

The inverse mapping from  $\phi$  to  $\psi$ :

$$\mu = \frac{\alpha_0}{(1 - \alpha_1)}, \quad \sigma_0 = \frac{\sigma^2}{(1 - \alpha_1^2)}, \quad \sigma_1 = \frac{\alpha_1 \sigma^2}{(1 - \alpha_1^2)}, \quad (28.14)$$

reveals that the admissible range of values of  $\alpha_1$  is  $(-1, 1)$ , excluding unity. Note that the parameterization (28.13) can be derived directly from (28.12) by utilizing the assumptions  $E(u_t | \sigma(y_{t-1})) = 0$ ,  $E(u_t^2 | \sigma(y_{t-1})) = \sigma^2$ ; see Spanos (1995).

The probabilistic reduction approach views the AR(1) model specified in terms of (28.12) as comprising the following *model assumptions* concerning the conditional process  $\{(y_t | Y_{t-1}^0), t \in \mathbb{T}\}$ :

- 1<sup>c</sup> normality:  $f(y_t | Y_{t-1}^0; \psi)$  is normal  
 2<sup>c</sup> linearity:  $E(y_t | \sigma(Y_{t-1}^0)) = \alpha_0 + \alpha_1 y_{t-1}$ , linear in  $y_{t-1}$ ,  
 3<sup>c</sup> homoskedasticity:  $\text{var}(y_t | \sigma(Y_{t-1}^0)) = \sigma^2$ , free of  $Y_{t-1}^0$ ,  
 4<sup>c</sup>  $t$ -homogeneity:  $(\alpha_0, \alpha_1, \sigma^2)$  are not functions of  $t \in \mathbb{T}$ ,  
 5<sup>c</sup> martingale difference:  $\{(u_t | Y_{t-1}^0), t \in \mathbb{T}\}$  is a martingale difference process.

Note that the temporal dependence assumption underlying the observable process  $\{y_t, t \in \mathbb{T}\}$  is Markov autocorrelation, whose general form (see (28.8)) is:

$$\text{cov}(y_t, y_{t-\tau}) := \sigma_{|\tau|} \leq c\lambda^{|\tau|}, \quad c > 0, \quad 0 < \lambda < 1, \quad \tau \neq 0, \quad t = 1, 2, \dots$$

The question that naturally arises at this stage is what kind of advantages the probabilistic reduction (PR) perspective offers (if any) when compared with the traditional DGM view. For a more systematic answer to this question we will consider the advantages at the different stages of empirical modeling: (i) specification, (ii) estimation, (iii) misspecification testing and (iv) respecification.

## SPECIFICATION

This refers to the initial stage of choosing a statistical model in view of the observed data and the theoretical question(s) of interest. The postulated statistical model purports to provide an adequate description of the observable stochastic phenomenon of interest; model all the statistical systematic information exhibited by the observed data (see Spanos, 1999, ch. 1). The PR perspective of the AR(1) model (defined in terms of assumptions 1<sup>c</sup>–5<sup>c</sup>) has a distinct advantage over that of the traditional approach based on assumptions 1<sup>a</sup>–4<sup>a</sup> in so far as the latter assumptions are not a priori assessable because they are defined in terms of the unobservable error term process  $\{\varepsilon_t, t \in \mathbb{T}\}$ . In contrast, assumptions 1<sup>c</sup>–5<sup>c</sup> are specified directly in terms of the process  $\{(y_t | Y_{t-1}^0), t \in \mathbb{T}\}$  and their validity can be assessed a priori via the reduction assumptions (i)–(iii) relating to  $\{y_t, t \in \mathbb{T}\}$  using graphical techniques such as  $t$ -plots and scatter plots. The relationship between the model assumptions 1<sup>c</sup>–5<sup>c</sup> and the reduction assumptions is given by the following theorem.

**Theorem 1.** Let  $\{y_t, t \in \mathbb{T}\}$  by a stochastic process with bounded moments of order two. The process  $\{y_t, t \in \mathbb{T}\}$  is normal, Markov and stationary if and only if the conditional process  $\{(y_t | Y_{t-1}^0, t \in \mathbb{T}\}$  satisfies the model assumptions 1<sup>c</sup>–5<sup>c</sup>.

**Proof.** The *if part*, (normality–Markovness–stationarity)  $\Rightarrow 1^c$ – $5^c$ , is trivial since:

$$\begin{aligned} f(y_0, y_1, y_2, \dots, y_T; \psi) &= f(y_0; \phi_0) \prod_{t=1}^T f(y_t | y_{t-1}, y_{t-2}, \dots, y_0; \phi) \\ &= f(y_0; \phi_0) \prod_{t=1}^T f_t(y_t | y_{t-1}; \phi_t) \quad - \text{Markovness} \\ &= f(y_0; \phi_0) \prod_{t=1}^T f(y_t | y_{t-1}; \phi) \quad - \text{stationarity}. \end{aligned}$$

These combined with normality implies assumptions 1<sup>c</sup>–5<sup>c</sup>. The *only if* part follows directly from Theorem 1 in Spanos (1995). ■

## ESTIMATION

Given that the likelihood function is defined in terms of the joint distribution of the observable process  $\{y_t, t \in \mathbb{T}\}$ , the PR approach enjoys a minor advantage over the traditional approach because the need to transfer the probabilistic structure from the error process  $\{\varepsilon_t, t \in \mathbb{T}\}$  onto the observable process does not arise. The primary advantage of the PR approach, however, arises from the implicit parameterization (28.13) which relates the model parameters  $\phi := (\alpha_0, \alpha_1, \sigma^2)$  and primary parameters  $\psi := (\mu, \sigma_0, \sigma_1)$ . This parameterization plays an important role in bringing out the interrelationships among the model parameters as well as determining their admissible range. For instance, the PR statistical parameterization in (28.13) brings out two important points relating to the model parameters which are ignored by the traditional time series literature. The first point is that the admissible range of values of the model parameter  $\alpha_1$  is  $(-1, 1)$  which excludes the values  $|\alpha_1| = 1$ . This has very important implications for the unit root testing literature. The second point is that the implicit restriction  $\sigma^2 = \sigma_0(1 - \alpha_1^2)$  does not involve the initial condition (as traditionally assumed) but all observations. This has important implications for the MLEs of  $\phi$  for  $\alpha_1$  near the unit root because the likelihood function based on (28.10); see Spanos and McGuirk (1999) for further details.

The PR approach can also help shed some light on the finite sample distribution of the OLS estimators of  $(\alpha_0, \alpha_1)$ . In view of the similarity between the conditioning information set  $\sigma(Y_{t-1}^0)$  of the AR(1) model and that of the stochastic normal/linear regression model  $\sigma(X_t)$  (see Spanos, 1986, ch. 20), one can conjecture that the finite sampling distributions of  $(\hat{\alpha}_0, \hat{\alpha}_1)$  are closer to the student's-t than the normal.

## MISSPECIFICATION TESTING

This refers to the testing of the model assumptions using misspecification tests which are probing beyond the boundaries of the postulated model; this should be contrasted with Neyman-Pearson testing which is viewed as testing within the boundaries (see Spanos, 1999, chs 14–15). The PR perspective has again certain distinct advantages over the traditional approach. First, the assumptions 1<sup>c</sup>–5<sup>c</sup> are specified explicitly in terms of the observable and not the error stochastic process (assumptions 1<sup>a</sup>–4<sup>a</sup>). This makes it easier to develop misspecification tests for these assumptions. In addition, the various misspecification tests developed in the context of the normal/linear regression model (see Spanos, 1986, chs 21–23) can be easily adapted to apply to the case of the AR(1) model with assumptions 1<sup>c</sup>–5<sup>c</sup>. Second, in the context of the PR approach, the connection between the reduction and model assumptions, utilized at the specification stage, can shed light on the likely directions of departures from 1<sup>c</sup>–5<sup>c</sup>, which can be useful in the choice of appropriate misspecification tests. Third, the same relationship can also be used to device joint misspecification tests (see Spanos, 1999, ch. 15).

## RESPECIFICATION

This refers to the choice of an alternative statistical model when the original model is found to be statistically inadequate. In the context of the PR approach, respecification can be viewed as the choice of an alternative statistical model which can be devised by changing the reduction (not the model) assumptions in view of the misspecification testing results. For instance, if the misspecification testing has shown departures from assumptions 1<sup>c</sup> and 3<sup>c</sup>, the correspondence between reduction and model assumptions suggests changing the normality reduction assumption to another joint distribution with a linear autoregression and a heteroskedastic conditional variance; a member of the elliptically symmetric family of joint distributions, such as the Student's-*t*, suggests itself in this case. Changing the normal to the Student's-*t* distribution will give rise to a different AR(1) model with assumption 1<sup>c</sup> replaced by the Student's-*t* and assumption 3<sup>c</sup> to a particular dynamic heteroskedasticity formulation as suggested by the Student's-*t* distribution (see Spanos, 1994). In contrast, in the context of the traditional approach respecification takes the form of changing the model assumptions without worrying about the potential internal inconsistency among these assumptions. As shown in Spanos (1995), postulating some arbitrary dynamic heteroskedasticity formulation might give rise to internal inconsistencies.

### 3.3 Extending the autoregressive AR(1) model

The probabilistic reduction (PR) perspective, as it relates to respecification, provides a systematic way to extend AR(1) in several directions. It must be emphasized, however, the these extensions constitute alternative models. Let us consider a sample of such statistical models.

#### AR(*p*) MODEL

The extension of the AR(1) to the AR(*p*) model amounts to replacing the Markov (reduction) assumption with that of *Markov of order p*, yields:

$$y_t = \alpha_0 + \sum_{k=1}^p \alpha_k y_{t-k} + u_t, \quad t \in \mathbb{T},$$

with the model assumptions 1<sup>c</sup>–5<sup>c</sup> modified accordingly.

#### AR(1) MODEL WITH A TRENDING MEAN

Extending the AR(1) model in order to include a trend, amounts to replacing the reduction assumption of mean stationarity  $E(y_t) = \mu$ , for all  $t \in \mathbb{T}$ , with a particular form of mean-heterogeneity, say  $E(y_t) = \mu t$ , for all  $t \in \mathbb{T}$ . The implied changes in the bivariate distribution (28.11), via the orthogonal decomposition (28.9), give rise to:

$$y_t = \delta_0 + \delta_1 t + \alpha_1 y_{t-1} + u_t, \quad t \in \mathbb{T}, \tag{28.15}$$

where the statistical parameters  $\phi := (\delta_0, \delta_1, \alpha_1, \sigma^2) \in \mathbb{R}^2 \times (-1, 1) \times \mathbb{R}_+$  and  $\psi := (\mu, \sigma_0, \sigma_1) \in \mathbb{R}^2 \times \mathbb{R}_+$ , are interrelated via:

$$\delta_0 = \left( \frac{\sigma_1}{\sigma_0} \right) \mu \in \mathbb{R}, \quad \delta_1 = \left( 1 - \left( \frac{\sigma_1}{\sigma_0} \right) \right) \mu, \quad \alpha_1 = \frac{\sigma_1}{\sigma_0} \in (-1, 1), \quad \sigma^2 = \sigma_0 - \frac{\sigma_1^2}{\sigma_0} \in \mathbb{R}_+, \quad (28.16)$$

$$\mu = \frac{\delta_0}{\alpha_1}, \quad \mu = \frac{\delta_1}{(1 - \alpha_1)}, \quad \sigma_0 = \frac{\sigma^2}{(1 - \alpha_1^2)}, \quad \sigma_1 = \frac{\alpha_1 \sigma^2}{(1 - \alpha_1^2)}. \quad (28.17)$$

This is an important extension because it provides the backbone of Dickey–Fuller unit root testing (see Dickey and Fuller, 1979, 1981) based on  $H_0 : \alpha_1 = 1$ . A closer look at the above implicit parameterizations, however, suggests that when  $\alpha_1 = 1 \Rightarrow (\delta_0 = \mu, \delta_1 = 0, \sigma^2 = 0)$ ; this raises important methodological issues concerning various aspects of these tests (see Spanos and McGuirk, 1999 for further details).

The extension of the AR(1) model to include higher-order trend terms and seasonal effects can be achieved by postulating mean-heterogeneity of the form

$$E(y_t) = \sum_{i=1}^m a_i D_{it} + \sum_{k=1}^l \mu_k t^k, \quad t \in \mathbb{T}, \text{ where } (D_{it}, i = 1, 2, \dots, m)$$

can be either sinusoidal functions or/and dummy variables purporting to model the seasonal effects. We conclude this subsection by noting that following an analogous procedure one can specify AR(1) models with a trending variance; see Spanos (1990).

### NON-NORMAL AUTOREGRESSIVE MODELS

By retaining the reduction assumptions of Markovness and stationarity and replacing normality with an alternative joint distribution one can specify numerous nonnormal autoregressive models; see Spanos (1999, ch. 8). The normal autoregressive model can also be extended in the direction of nonlinear models; see Granger and Teräsvirta (1993) for several important nonlinear time series models.

## 4 MOVING AVERAGE MODELS

### 4.1 The traditional approach

A *moving average model* of order  $q$ , denoted by  $\text{MA}(q)$ :

$$y_t = a_0 + \sum_{k=1}^q a_k \varepsilon_{t-k} + \varepsilon_t, \quad \varepsilon_t \sim \text{NI}(0, \sigma^2), \quad t \in \mathbb{T}, \quad (28.18)$$

is traditionally viewed as a DGM with a normal white noise process  $\{\varepsilon_t, t \in \mathbb{T}\}$  (see (28.4)) as the input and  $\{y_t, t \in \mathbb{T}\}$ , as the output process.

The question which naturally arises at this stage is “how does the DGM (28.18) fit into the orthogonal decomposition given in (28.9)?” A naïve answer will be  $y_t = E(y_t | \sigma(\varepsilon_{t-1}, \varepsilon_{t-2}, \dots, \varepsilon_{t-q})) + \varepsilon_t$ ,  $t \in \mathbb{T}$ . However, such an answer is misleading because operational conditioning cannot be defined in terms of an unobserved stochastic process  $\{\varepsilon_t, t \in \mathbb{T}\}$ . In view of this, the next question is “how is the formulation (28.18) justified as a statistical Generating Mechanism?” The answer lies with the following theorem.

**Wold decomposition theorem.** Let  $\{y_t, t \in \mathbb{T}\}$ , be a normal stationary process and define the unobservable derived process  $\{\varepsilon_t, t \in \mathbb{T}\}$ , by:

$$\varepsilon_t = y_t - E(y_t | \sigma(Y_{t-1}^0)), \text{ with } E(\varepsilon_t) = 0, \quad E(\varepsilon_t^2) = \sigma^2 > 0. \quad (28.19)$$

Then, the process  $\{y_t, t \in \mathbb{T}\}$ , can be expressed in the form:

$$y_t - \mu = w_t + \sum_{k=0}^{\infty} a_k \varepsilon_{t-k}, \quad t \in \mathbb{T}. \quad (28.20)$$

- (i)  $\varepsilon_t \sim NI(0, \sigma^2)$ ,  $t = 1, 2, \dots$ ,
- (ii)  $\sum_{k=0}^{\infty} a_k^2 < \infty$ , for  $a_k = \frac{\text{cov}(y_t, \varepsilon_{t-k})}{\text{var}(\varepsilon_{t-k})}$ ,  $k = 0, 1, 2, \dots$ ,
- (iii) for  $w_t = \sum_{k=1}^{\infty} \gamma_k w_{t-k}$ ,  $E(w_t \varepsilon_s) = 0$ , for all  $t, s = 1, 2, \dots$

It is important to note that the process  $\{w_t, t \in \mathbb{T}\}$ , is *deterministic* in the sense that it's perfectly predictable from its own past; since  $\sigma(w_{t-1}, w_{t-2}, \dots) = \bigcap_{t=-\infty}^{\infty} \sigma(Y_{t-1}^0)$ , the right-hand side being the remote past of the process  $\{y_t, t \in \mathbb{T}\}$  (see Wold, 1938). In view of this, the MA( $\infty$ ) decomposition often excludes the remote past:

$$y_t - \mu = \sum_{k=1}^{\infty} a_k \varepsilon_{t-k} + \varepsilon_t, \quad t \in \mathbb{T}. \quad (28.21)$$

As it stands, the MA( $\infty$ ) formulation is non-operational because it involves an infinite number of unknown parameters. The question arises whether one can truncate the MA( $\infty$ ) at some finite value  $q < T$ , in order to get an operational model. As the Wold decomposition theorem stands, no such truncation is justifiable. For such a truncation to be formally justifiable we need to impose certain temporal dependence restrictions on the covariances  $\sigma_{|\tau|} = \text{cov}(y_t, y_{t-\tau})$ . The most natural dependence restriction in the case of stationary processes is that of ergodicity; see Hamilton (1994) and Phillips (1987).

## 4.2 MA( $q$ ): the probabilistic reduction perspective

At this point it is important to emphasize that the above discussion relating to the convergence of certain partial sums of the MA( $\infty$ ) coefficients is not helpful

from the empirical modeling viewpoint because the restrictions cannot be assessed a priori. Alternatively, one can consider restrictions on the temporal covariances of the observable process  $\{y_t, t \in \mathbb{T}\}$  which we can assess a priori:

$$\begin{aligned} 1^d & \text{ constant mean: } E(y_t) := \mu, t \in \mathbb{T}, \\ 2^d & \text{ constant variance: } \text{var}(y_t) := \sigma_0, t \in \mathbb{T}, \\ 3^d & \text{ } m\text{-autocorrelation: } \text{cov}(y_t, y_{t-\tau}) := \begin{cases} \sigma_{|\tau|}, & \tau = 1, 2, \dots, q, \\ 0, & \tau > q, \end{cases} \\ 4^d & \text{ normality: } y_t \sim N(\cdot, \cdot), t \in \mathbb{T}. \end{aligned} \quad (28.22)$$

where the first two moments in terms of the statistical parameterization  $\phi := (a_0, a_1, \dots, a_q, \sigma^2)$  take the form:

$$\mu = a_0, \quad \sigma_0 = \sigma^2 \left( 1 + \sum_{k=1}^q a_k^2 \right), \quad \sigma_{|\tau|} = \sigma^2 (a_\tau + a_1 a_{\tau+1} + \dots + a_{q-\tau} a_q). \quad (28.23)$$

In view of the fact that the likelihood function is defined in terms of the joint distribution of the observable process  $\{y_t, t \in \mathbb{T}\}$ , it is apparent that:

$$L(\phi) \propto (2\pi)^{-\frac{T}{2}} (\det \Omega(\phi))^{-\frac{1}{2}} \exp \left\{ -\frac{1}{2} (y - 1a_0)^\top \Omega(\phi)^{-1} (y - 1a_0) \right\},$$

where the  $T \times T$  temporal variance–covariance  $\Omega(\phi)$  takes the banded Toeplitz form with all elements along the diagonal and off-diagonals up to  $q$  coincide and are nonzero but after the  $q$ th off-diagonal the covariances are zero. This gives rise to a loglikelihood function whose first-order conditions with respect to  $\phi$  are nonlinear and the estimation requires numerical optimization; see Anderson (1971).

Returning to the Wold decomposition theorem we note that the probabilistic structure of the observable process  $\{y_t, t \in \mathbb{T}\}$  involves only normality and stationarity which imply that the variance–covariance matrix is Toeplitz, which, when compared with  $\Omega(\phi)$  the result becomes apparent; the banded Toeplitz covariance matrix in  $\Omega(\phi)$  as  $T \rightarrow \infty$  gives rise to a MA( $q$ ) formulation and the unrestricted Toeplitz covariance matrix as  $T \rightarrow \infty$  gives rise to a MA( $\infty$ ) formulation. Does this mean that to get an operational model we need to truncate the temporal covariance matrix, i.e. assume that  $\sigma_\tau = 0$  for all  $\tau > q$ , for some  $q > 1$ ? This assumption will give rise to the MA( $q$ ) model but there are more general models we can contemplate that do not impose such a strong restriction. Instead, we need some restrictions which ensure that  $\sigma_\tau \rightarrow 0$  as  $\tau \rightarrow \infty$  at a “reasonable” rate such as:

$$|\sigma_\tau| \leq c\lambda^\tau, \quad c > 0, \quad 0 < \lambda < 1, \quad \tau = 1, 2, 3, \dots \quad (28.24)$$

This enables us to approximate the non-operational MA( $\infty$ ) representation with operational models from the broader ARMA( $p, q$ ) family. This should be contrasted with stochastic processes with long memory (see Granger, 1980) where:

$$|\sigma_\tau| \leq c\tau^{(2d-1)}, \quad c > 0, \quad 0 < d < .5, \quad \tau = 1, 2, 3, \dots \quad (28.25)$$

In cases where, in addition to the normality and stationarity, we assume that the process  $\{y_t, t \in \mathbb{T}\}$  satisfies the dependence restriction (28.24), we can proceed to approximate the infinite polynomial in the lag operator  $L$ ,  $a_\infty(L) = 1 + a_1L + \dots + a_kL^k + \dots$ , of the MA( $\infty$ ) representation:

$$y_t = \mu + \sum_{k=1}^{\infty} a_k \varepsilon_{t-k} + \varepsilon_t = \mu + a_\infty(L) \cdot \varepsilon_t, \quad t \in \mathbb{T}, \quad (28.26)$$

by a ratio of two finite order polynomials  $a_\infty(L) = \frac{\gamma_q(L)}{\delta_p(L)} := \frac{(1 + \gamma_1L + \gamma_2L^2 + \dots + \gamma_qL^q)}{(1 + \delta_1L + \delta_2L^2 + \dots + \delta_pL^p)}$ ,  $p \geq q \geq 0$ ; (see Dhrymes, 1971). After re-arranging the two polynomials:

$$y_t = \mu + \frac{\gamma_q(L)}{\alpha_p(L)} \varepsilon_t \Rightarrow \alpha_p(L)y_t = \mu + \gamma_q(L)\varepsilon_t, \quad t \in \mathbb{T},$$

yields the autoregressive-moving average model ARMA( $p, q$ ) popularized by Box and Jenkins (1970):

$$y_t + \sum_{k=1}^p \alpha_k y_{t-k} = \mu + \sum_{k=1}^q \gamma_k \varepsilon_{t-k} + \varepsilon_t, \quad t \in \mathbb{T}.$$

Such models proved very efficient in capturing the temporal dependence in time series data in a parsimonious way but failed to capture the imagination of economic modelers because it's very difficult to relate such models to economic theory.

The question that arises in justifying this representation is why define the statistical GM in terms of the errors? The only effective justification is when the modeler has a priori evidence that the dependence exhibited by the time series data is of the  $q$ -autocorrelation form and  $q$  is reasonably small. On the other hand, if the dependence is better described by (28.24), the AR( $p$ ) representation provides a much more effective description. The relationship between the MA( $q$ ) representation (28.18) and the autoregressive AR( $\infty$ ) representation takes the form:

$$y_t = \sum_{k=1}^{\infty} b_k y_{t-k} + \varepsilon_t, \quad t \in \mathbb{T},$$

where the coefficients are related (by equating the coefficients) via:

$$b_1 = a_1, \quad b_2 = a_2 + a_1b_1, \quad b_3 = a_3 + a_1b_2 + a_2b_1, \dots, \dots,$$

$$b_q = a_q + a_1b_{q-1} + a_2b_{q-2} + \dots + a_{q-1}b_1, \quad b_\tau = \sum_{k=1}^q a_k b_{\tau-k}, \quad \tau > q.$$

Given that  $b_\tau \xrightarrow{\tau \rightarrow \infty} 0$ , the modeler can assume that the latter representation can be approximated by a finite AR( $p$ ) model; which is often preferred for forecasting.

## 5 ARMA TYPE MODELS: MULTIVARIATE

The above discussion of the AR( $p$ ) and MA( $q$ ) models can be extended to the case where the observable process is a vector  $\{Z_t, t \in \mathbb{T}\}$ ,  $Z_t : (m \times 1)$ . This stochastic vector process is said to be second-order stationary if:

$$E(Z_t) = \mu, \quad \text{cov}(Z_t, Z_{t-\tau}) = E((Z_t - \mu)(Z_{t-\tau} - \mu)^\top) = \Sigma(\tau).$$

Note that  $\Sigma(\tau)$  is not symmetric since  $\sigma_{ij}(\tau) = \sigma_{ji}(-\tau)$ ; see Hamilton (1994).

The ARMA representations for the vector process  $\{Z_t, t \in \mathbb{T}\}$  take the form:

$$\text{VAR}(p): \quad Z_t = \alpha_0 + A_1 Z_{t-1} + A_2 Z_{t-2} + \dots + A_p Z_{t-p} + \varepsilon_t,$$

$$\text{VMA}(q): \quad Z_t = \mu + \Phi_1 \varepsilon_{t-1} + \Phi_2 \varepsilon_{t-2} + \dots + \Phi_q \varepsilon_{t-q} + \varepsilon_t,$$

$$\text{VARMA}(p, q): \quad Z_t = \gamma_0 + A_1 Z_{t-1} + \dots + A_p Z_{t-p} + \Theta_1 \varepsilon_{t-1} + \dots + \Theta_q \varepsilon_{t-q} + \varepsilon_t,$$

where the vector error process is of the form:  $\varepsilon_t \sim NI(0, \Omega)$ .

In direct analogy to the univariate case, the probabilistic assumptions are:

$$\left. \begin{array}{ll} 1^e & \text{zero mean:} \quad E(\varepsilon_t) = 0, \\ 2^e & \text{constant variance:} \quad E(\varepsilon_t \varepsilon_t^\top) = \Omega, \\ 3^e & \text{no autocorrelation:} \quad E(\varepsilon_t \varepsilon_{t-\tau}^\top) = 0, \tau \neq 0, \\ 4^e & \text{normality:} \quad \varepsilon_t \sim N(\cdot, \cdot), \end{array} \right\} t \in \mathbb{T}. \quad (28.27)$$

Looking at the above representations from the PR perspective we need to translate 1<sup>e</sup>-4<sup>e</sup> into assumptions in terms of the observable vector  $\{Z_t, t \in \mathbb{T}\}$ .

### 5.1 The probabilistic reduction perspective

The probabilistic reduction perspective contemplates the DGM of the Vector Autoregressive representation from left to right as an orthogonal decomposition:

$$Z_t = E(Z_t | \sigma(Z_{t-1}^0)) + u_t, \quad t \in \mathbb{T}, \quad (28.28)$$

where  $Z_{t-1}^0 := (Z_{t-1}, Z_{t-2}, \dots, Z_0)$  and  $u_t = Z_t - E(Z_t | \sigma(Z_{t-1}^0))$ , with the underlying statistical model viewed as reduction from the joint distribution of the underlying process  $\{Z_t, t \in \mathbb{T}\}$ . In direct analogy to the univariate case the *reduction assumptions* on the joint distribution of the process  $\{Z_t, t \in \mathbb{T}\}$ , that would yield VAR(1) model are: (i) normal, (ii) Markov, and (iii) stationary.

Let us consider the question of probabilistic reduction in some detail by imposing the reduction assumptions in a certain sequence in order to reduce the joint distribution to an operational model:

$$\begin{aligned}
D(\mathbf{Z}_0, \mathbf{Z}_1, \dots, \mathbf{Z}_T; \psi) &= D(\mathbf{Z}_0; \phi_0) \prod_{t=1}^T D_t(\mathbf{Z}_t | \mathbf{Z}_{t-1}^0; \phi_t), \\
&\stackrel{\text{M\&S}}{=} D(\mathbf{Z}_0; \phi_0) \prod_{t=1}^T D(\mathbf{Z}_t | \mathbf{Z}_{t-1}; \phi), \\
&\stackrel{\text{M\&S}}{=} D(\mathbf{Z}_0; \phi_0) \prod_{t=1}^T D(\mathbf{Z}_t | \mathbf{Z}_{t-1}; \phi), (\mathbf{Z}_0, \dots, \mathbf{Z}_T) \in \mathbb{R}^{m(T+1)}.
\end{aligned} \tag{28.29}$$

The first equality does not entail any assumptions, but the second follows from the Markovness (M) and the third from the stationarity (S) assumption. In order to see what happens to  $D(\mathbf{Z}_t | \mathbf{Z}_{t-1}; \phi)$  when the normality assumption:

$$\begin{bmatrix} \mathbf{Z}_t \\ \mathbf{Z}_{t-1} \end{bmatrix} \sim N\left(\begin{bmatrix} \boldsymbol{\mu} \\ \boldsymbol{\mu} \end{bmatrix}, \begin{bmatrix} \Sigma(0) & \Sigma(1) \\ \Sigma(1)^\top & \Sigma(0) \end{bmatrix}\right), \quad t \in \mathbb{T}, \tag{28.30}$$

is imposed, the orthogonal decomposition (28.28) gives rise to:

$$\mathbf{Z}_t = \boldsymbol{\alpha}_0 + \mathbf{A}_1 \mathbf{Z}_{t-1} + \mathbf{u}_t, \quad t \in \mathbb{T}. \tag{28.31}$$

The statistical parameters  $\phi := (\boldsymbol{\alpha}_0, \mathbf{A}_1, \Omega)$  are related to the primary parameters  $\psi = (\boldsymbol{\mu}, \Sigma(0), \Sigma(1))$  via:

$$\boldsymbol{\alpha}_0 = (\mathbf{I} - \mathbf{A}_1)\boldsymbol{\mu}, \quad \mathbf{A}_1 = \Sigma(1)^\top \Sigma(0)^{-1}, \quad \Omega = \Sigma(0) - \Sigma(1)^\top \Sigma(0)^{-1} \Sigma(1), \tag{28.32}$$

and the discussion concerning the interrelationships between two parameter spaces is analogous to the univariate case discussed in the previous section and will not be pursued any further; see Spanos (1986, chs 22–23).

The probabilistic reduction approach views the VAR(1) model specified in terms of (28.12) as comprising the following *model assumptions* concerning the conditional process  $\{(\mathbf{Z}_t | \mathbf{Z}_{t-1}^0), t \in \mathbb{T}\}$ :

- 1<sup>f</sup> normality:  $D(\mathbf{Z}_t | \mathbf{Z}_{t-1}^0; \psi)$  is normal,
- 2<sup>f</sup> linearity:  $E(\mathbf{Z}_t | \sigma(\mathbf{Z}_{t-1}^0)) = \boldsymbol{\alpha}_0 + \mathbf{A}_1 \mathbf{Z}_{t-1}$ , linear in  $\mathbf{Z}_{t-1}$ ,
- 3<sup>f</sup> homoskedasticity:  $\text{cov}(\mathbf{Z}_t | \sigma(\mathbf{Z}_{t-1}^0)) = \Omega$ , free of  $\mathbf{Z}_{t-1}^0$ ,
- 4<sup>f</sup>  $t$ -homogeneity:  $(\boldsymbol{\alpha}_0, \mathbf{A}_1, \Omega)$  are not functions of  $t \in \mathbb{T}$ ,
- 5<sup>f</sup> martingale difference:  $\{(\mathbf{u}_t | \mathbf{Z}_{t-1}^0), t \in \mathbb{T}\}$  is a vector martingale difference process.

Continuing with the analogies between the vector and univariate cases, the temporal dependence assumption for the process  $\{\mathbf{Z}_t, t \in \mathbb{T}\}$  is Markov autocorrelation:

$$\text{cov}(Z_{it}, Z_{j(t-\tau)}) = \sigma_{ij\tau} \leq c\lambda^{|\tau|}, \quad c > 0, 0 < \lambda < 1, \tau \neq 0, i, j = 1, 2, \dots, t = 1, 2, \dots$$

For the VAR(1) model as specified by (28.31) and assumptions 1<sup>f</sup>–5<sup>f</sup> above to be statistically adequate, the modeler should test the underlying assumptions, with

the misspecification tests being modifications of the ones for the univariate case (see Spanos, 1986, ch. 24).

The VAR(1) is much richer than the univariate ARMA( $p, q$ ) specification because, in addition to the self-temporal structure, it enables the modeler to consider the cross-temporal structure, fulfilling Moore's basic objective in the classic (1914) study. This can be seen in the simplest case of a VAR(1) model with  $m = 2$ :

$$\begin{pmatrix} Z_{1t} \\ Z_{2t} \end{pmatrix} = \begin{pmatrix} a_{10} \\ a_{20} \end{pmatrix} + \begin{pmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{pmatrix} \begin{pmatrix} Z_{1(t-1)} \\ Z_{2(t-1)} \end{pmatrix} + \begin{pmatrix} u_{1t} \\ u_{2t} \end{pmatrix}, \quad \Omega = \begin{pmatrix} \omega_{11} & \omega_{12} \\ \omega_{21} & \omega_{22} \end{pmatrix}.$$

The coefficients ( $a_{12}, a_{21}$ ) measure the cross-temporal dependence between the two processes. In the case where  $a_{12} = 0$ ,  $Z_{2t}$  does not *Granger cause*  $Z_{1t}$ , and vice versa in the case where  $a_{21} = 0$ . This is an important concept in the context of forecasting. The covariance  $\omega_{12}$  constitutes a measure of the contemporaneous ( $\text{cov}(Z_{1t}, Z_{2t} | Z_{t-1}^0)$ ) dependence. As argued in the next subsection, the dynamic linear regression model can be viewed as a further reduction of the VAR model which purports to model this contemporaneous dependence. For further discussion of the VAR model see Hamilton (1994).

## 6 TIME SERIES AND LINEAR REGRESSION MODELS

### 6.1 Error autocorrelation vs. temporal dependence

The classic paper of Yule (1926) had a lasting effect on econometric modeling in so far as his cautionary note that using time series data in the context of linear regression can often lead to spurious results, resulted in reducing the number of empirical studies using such data. As mentioned above, the results of Cochrane and Orcutt (1949) and Durbin and Watson (1950, 1951) were interpreted by the econometric literature as a way to sidestep the spurious regression problem raised by Yule, and legitimize the use of time series data in the context of the linear regression model. *Stage 1:* estimate the linear regression model:

$$y_t = \beta^\top x_t + u_t, \quad u_t \sim NI(0, \sigma^2), \quad t \in \mathbb{T}, \quad (28.33)$$

and test for error autocorrelation using the Durbin–Watson test based on:

$$DW(y) = \frac{\sum_{t=2}^T (\hat{u}_t - \hat{u}_{t-1})^2}{\sum_{t=1}^T \hat{u}_t^2}, \quad \hat{u}_t = y_t - \hat{\beta}^\top x_t, \quad \hat{\beta} = (X^\top X)^{-1} X^\top y.$$

The alternative model that gives rise to this test is the modified model:

$$y_t = \beta^\top x_t + u_t, \quad u_t = \rho u_{t-1} + \varepsilon_t, \quad \varepsilon_t \sim NI(0, \sigma_\varepsilon^2), \quad |\rho| < 1, \quad t \in \mathbb{T}, \quad (28.34)$$

and the test is based on the hypothesis  $H_0 : \rho = 0$ ,  $H_1 : \rho \neq 0$ . *Stage 2*: if the null hypothesis is not rejected, the modeler assumes that the original model (28.33) does not indicate the presence of error autocorrelation, otherwise the modeler can “cure” the problem by adopting the alternative model (28.34), which can be estimated using the Cochrane–Orcutt procedure. In the context of the traditional Gauss–Markov specification in matrix notation:

$$\begin{aligned} y &= X\beta + u, \quad y : T \times 1, \quad X : T \times k, \\ 1^g E(u) &= 0, \quad 2^g-3^g E(uu^\top) = \sigma^2 I_T \quad 4^g x_t \text{ is fixed}, \quad 5^g \text{ Rank}(X) = k, \quad (T > k), \end{aligned}$$

the presence of error autocorrelation takes the form  $3^g E(uu^\top) = V_T \neq \sigma^2 I_T$ . Under  $E(uu^\top) = V_T$ , (i.e.  $E(u_t u_s) \neq 0$ ,  $t \neq s$ ,  $t, s = 1, 2, \dots, T$ ) the OLS estimator  $\hat{\beta} = (X^\top X)^{-1} X^\top y$  is no longer best, linear unbiased estimator (BLUE); it is said to retain its *unbiasedness* and *consistency* but forfeit its *relative efficiency* since:

$$\text{cov}(\hat{\beta}) = (X^\top X)^{-1} (X^\top V_T X) (X^\top X)^{-1} \geq (X^\top V_T^{-1} X) = \text{cov}(\tilde{\beta}),$$

where  $\tilde{\beta} = (X^\top V_T^{-1} X)^{-1} X^\top V_T^{-1} y$  is the generalized least squares (GLS) estimator. This traditional discussion has often encouraged applied econometricians to argue that the use of time series data in the context of the linear regression model with assumptions 1<sup>g</sup>–5<sup>g</sup> would forfeit only the relative efficiency of the OLS estimators. As argued in Spanos (1986, chs 22–23), the above suggestion is very misleading because the traditional textbook scenario is highly unlikely in empirical modeling. In order to see this we need to consider the above discussion from the probabilistic reduction (PR) perspective. Viewing (28.33) from this perspective reveals that the implicit parameterization of  $\theta := (\beta, \sigma^2)$  is:

$$\beta = \text{cov}(X_t)^{-1} \text{cov}(X_t, y_t), \quad \sigma^2 = \text{var}(y_t) - \text{cov}(y_t, X_t) \text{cov}(X_t)^{-1} \text{cov}(X_t, y_t).$$

In the context of the PR specification, assumption 3<sup>g</sup> corresponds to  $(y_1, y_2, \dots, y_T)$  being “temporally” independent (see Spanos, 1986, p. 373). In the context of the PR approach, respecification amounts to replacing the original reduction assumptions on the vector process  $\{Z_t, t \in \mathbb{T}\}$  where  $Z_t := (y_t \ X_t^\top)^\top$ , i.e. (i) normality; (ii) (temporal) independence; and (iii) identically distributed, with assumptions such as (i)' normality; (ii)' Markov; and (iii)' stationarity. The PR approach replaces the original conditioning information set  $\mathcal{D}_t = \{X_t = x_t\}$  with  $\mathcal{D}_t^* = \{X_t = x_t, Z_{t-1}^0\}$ ,  $Z_{t-1}^0 = \{Z_{t-1}, \dots, Z_0\}$ , which (under (i)'–(iii)') gives rise to the *dynamic linear regression*:

$$y_t = \beta_0^\top x_t + \alpha_1^\top Z_{t-1} + \varepsilon_t, \quad t \in \mathbb{T}. \quad (28.35)$$

where the implicit statistical parameterization is:

$$\begin{aligned} \beta_0 &= \text{cov}(X_t | Z_{t-1})^{-1} [\text{cov}(X_t, y_t) - \text{cov}(X_t, Z_{t-1}) [\text{cov}(Z_{t-1})]^{-1} \text{cov}(Z_{t-1}, y_t)], \\ \alpha_1 &= \text{cov}(Z_{t-1} | X_t)^{-1} [\text{cov}(Z_{t-1}, y_t) - \text{cov}(Z_{t-1}, X_t) [\text{cov}(X_t)]^{-1} \text{cov}(X_t, y_t)], \\ \sigma_\varepsilon^2 &= \text{var}(y_t) - (\text{cov}(y_t, X_t^*)) (\text{cov}(X_t^*))^{-1} (\text{cov}(X_t^*, y_t)), \quad X_t^* = (X_t^\top, Z_{t-1}^\top)^\top. \end{aligned}$$

The important point to emphasize is that the statistical parameterization of  $\beta_0$  is very different from that of  $\beta$  in the context of the linear regression model since:

$$\beta_0 \neq \text{cov}(X_t)^{-1} \text{cov}(X_t, y_t),$$

unless  $\text{cov}(X_t, Z_{t-1}) = 0$ ; there is no temporal correlation between  $X_t$  and  $Z_{t-1}$ . Given that the latter case is excluded by assumption, the only other case when the two parameterizations coincide arises in the case where the appropriate model is (28.34). In order to understand what the latter model entails let us consider how it can be nested within the dynamic linear regression model (28.35). Substituting out the error autocorrelation in the context of (28.34) yields:

$$y_t = \beta^\top x_t - \rho \beta^\top x_{t-1} + \rho y_{t-1} + \varepsilon_t, \quad |\rho| < 1, \quad t \in \mathbb{T},$$

which is a special case of the dynamic linear regression model (28.35):

$$y_t = \beta_0^\top x_t + \beta_1^\top x_{t-1} + \alpha_1 y_{t-1} + u_t, \quad t \in \mathbb{T},$$

when one imposes the so-called *common factor restrictions* (see Hendry and Mizon, 1978)  $\beta_0 \alpha_1 + \beta_1 = 0$ . As shown in Spanos (1987), these restrictions are highly unlikely in empirical modeling because their validity presupposes that all the individual components of the vector process  $\{Z_t, t \in \mathbb{T}\}$  are Granger noncausal and their temporal structure is "almost identical". Under this PR scenario the OLS estimator  $\hat{\beta} = (X^\top X)^{-1} X^\top y$  under the assumption that the vector process  $\{Z_t, t \in \mathbb{T}\}$  is "temporally" dependent is both *biased* and *inconsistent*!

## 6.2 Dynamic linear regression models

The dynamic linear regression model (28.35) discussed above, constitutes a reduction of the VAR(1) model (28.31), given that we can decompose  $D(Z_t | Z_{t-1}; \phi)$  further, based on the separation  $Z_t = (y_t^\top, X_t^\top)^\top$ ,  $y_t : m_1 \times 1$ , to yield:

$$\begin{aligned} D(Z_0, Z_1, \dots, Z_T; \psi) &\stackrel{\text{M\&s}}{=} D(Z_0; \phi_0) \prod_{t=1}^T D(Z_t | Z_{t-1}; \phi) \\ &= D(Z_0; \phi_0) \prod_{t=1}^T D(y_t | X_t, Z_{t-1}; \phi) D(X_t | Z_{t-1}; \phi). \end{aligned} \tag{28.36}$$

Under the normality reduction assumption this gives rise to the multivariate dynamic linear regression model (MDLR):

$$y_t = B_0^\top x_t + B_1^\top x_{t-1} + A_1^\top y_{t-1} + u_t, \quad t \in \mathbb{T}, \tag{28.37}$$

where  $A = \begin{pmatrix} A_{11} & A_{12} \\ A_{21} & A_{22} \end{pmatrix}$ ,  $\Omega := \begin{pmatrix} \Omega_{11} & \Omega_{12} \\ \Omega_{21} & \Omega_{22} \end{pmatrix}$ , are reparameterized into:

$$\begin{aligned} \mathbf{B}_0^\top &= \Omega_{22}^{-1} \Omega_{21}, & \mathbf{B}_1^\top &= \mathbf{A}_{12} - \Omega_{12} \Omega_{22}^{-1} \mathbf{A}_{22}, & \mathbf{A}_1^\top &= \mathbf{A}_{11} - \Omega_{12} \Omega_{22}^{-1} \mathbf{A}_{21}, \\ \text{cov}(y_t | X_t, Z_{t-1}) &= \Omega_{11} - \Omega_{12} \Omega_{22}^{-1} \Omega_{21}. \end{aligned}$$

As we can see, the statistical parameters of the MDLR(1) (28.37) differ from those of the VAR(1) model (28.31) in so far as the former goes one step further purporting to model (in terms of the extra conditioning) the “contemporaneous dependence” captured in  $\Omega$  in the context of the former model. This model is particularly interesting in econometrics because it provides a direct link to the simultaneous equations model; see Spanos (1986), p. 645.

## 7 CONCLUSION

Statistical models, such as AR( $p$ ), MA( $q$ ), ARMA( $p, q$ ) and the linear regression model with error autocorrelation, often used for modeling time series data, have been considered from a particular viewing angle we called the probabilistic reduction (PR) approach. The emphasis of this approach is placed on specifying statistical models in terms of a consistent set of probabilistic assumptions regarding the observable stochastic process underlying the data, as opposed to the error term. Although the discussion did not cover the more recent developments in time series econometrics, it is important to conclude with certain remarks regarding these developments. The recent literature on unit roots and cointegration, when viewed from the PR viewpoint, can be criticized on two grounds. The *first* criticism is that the literature has largely ignored the statistical adequacy issue. When the estimated AR( $p$ ) models are misspecified, however, the unit root test inference results will often be misleading. The *second* criticism concerns the inadequate attention paid by the recent literature on the implicit parameterizations (discussed above) and what they entail (see Spanos and McGuirk, 1999).

## References

- Anderson, T.W. (1971). *The Statistical Analysis of Time Series*. Wiley, New York.
- Box, G.E.P., and G.M. Jenkins (1970). *Time Series Analysis: Forecasting and Control*, revised edn 1976. Holden-Day, San Francisco.
- Cochrane, D., and G.H. Orcutt (1949). Application of least-squares regression to relationships containing autocorrelated error terms. *Journal of the American Statistical Association* 44, 32–61.
- Cramer, H. (1937). *Random Variables and Probability Distributions*. Cambridge: Cambridge University Press.
- Davis, T.H. (1941). *The Analysis of Economic Time Series*. Cowles Commission Monograph No. 6, The Principia Press, Indiana.
- Dhrymes, P.J. (1971). *Distributed Lags: Problems of Estimation and Formulation*. Edinburgh: Oliver and Boyd.
- Dickey, D.A., and W.A. Fuller (1979). Distribution of the estimators for autoregressive time series with a unit root. *Journal of the American Statistical Association* 74, 427–31.
- Dickey, D.A., and W.A. Fuller (1981). Likelihood ratio statistics for autoregressive time series with a unit root. *Econometrica* 49, 1057–72.

- Durbin, J., and G.S. Watson (1950). Testing for serial correlation in least squares regression I. *Biometrika* 37, 409–28.
- Durbin, J., and G.S. Watson (1951). Testing for serial correlation in least squares regression II. *Biometrika* 38, 159–78.
- Engle, R.F., and C.W.J. Granger (1987). Cointegration and error-correction representation: estimation and testing. *Econometrica* 55, 251–76.
- Frisch, R. (1933). Propagation problems and impulse problems in dynamic economics. In *Economic Essays in Honor of Gustav Cassel*. London: Macmillan.
- Granger, C.W.J. (1980). Long memory relationships and the aggregation of dynamic models. *Journal of Econometrics* 14, 227–38.
- Granger, C.W.J. (1983). Cointegrated variables and error-correcting models. UCSD discussion paper 83–13.
- Granger, C.W.J. (ed.) (1990). *Modelling Economic Series: Readings on the Methodology of Econometric Modeling*. Oxford: Oxford University Press.
- Granger, C.W.J., and P. Newbold (1974). Spurious regressions in econometrics. *Journal of Econometrics* 2, 111–20.
- Granger, C.W.J., and P. Newbold (1977). *Forecasting Economic Time Series*. London: Academic Press.
- Granger, C.W.J., and T. Teräsvirta (1993). *Modelling Nonlinear Economic Relationships*. Oxford: Oxford University Press.
- Hamilton, J.D. (1994). *Time Series Analysis*. New Jersey: Princeton University Press.
- Hendry, D.F. (1993). *Econometrics: Alchemy or Science?*. Oxford: Blackwell.
- Hendry, D.F., and G.E. Mizon (1978). Serial correlation as a convenient simplification not a nuisance: a comment on a study of the demand for money by the Bank of England. *Economic Journal* 88, 549–63.
- Hendry, D.F., and M.S. Morgan (1995). *The Foundations of Economic Analysis: An Introduction*. New York: Cambridge University Press.
- Heyde, C.C., and E. Seneta (1977). *I.J. Bieyname: Statistical Theory Anticipated*. New York: Springer-Verlag.
- Hooker, R. (1901). Correlation of the marriage rate with trade. *Journal of the Royal Statistical Society* 64, 485–603.
- Hooker, R. (1905). On the correlation of successive observations: illustrated by corn prices. *Journal of the Royal Statistical Society* 68, 696–703.
- Johansen, S. (1991). Estimation and hypothesis testing of cointegrating vectors in Gaussian vector autoregressive models. *Econometrica* 59, 1551–81.
- Khinchine, A.Y. (1932). Selle successioni stazioarie di eventi. *Giorn. Ist. Ital. Attuari* 3, 267–74.
- Kolmogorov, A.N. (1933). *Grundbegriffe der Wahrscheinlichkeitrechnung*, Berlin. *Foundations of the Theory of Probability*, 2nd English edn. New York: Chelsea Publishing Co.
- Kolmogorov, A.N. (1941). Stationary sequences in Hilbert space. *Byull. Moskov. Gos. Univ. Mat.* 2, 1–40. English translation reprinted in Shirayev (1992), pp. 228–71.
- Mann, H.B. and A. Wald (1943). On the statistical treatment of linear stochastic difference equations. *Econometrica* 11, 173–220.
- Moore, H.L. (1914). *Economic Cycles: Their Law and Cause*. New York: Macmillan.
- Nelson, C.R., and C.I. Plosser (1982). Trends and random walks in macro-economic time series: some evidence and implications. *Journal of Monetary Economics* 10, 139–62.
- Norton, J. (1902). *Statistical Studies in the New York Money Market*. New York: Macmillan.
- Phillips, P.C.B. (1986). Understanding spurious regression in econometrics. *Journal of Econometrics* 33, 311–40.
- Phillips, P.C.B. (1987). Time series regressions with a unit root. *Econometrica* 55, 227–301.
- Phillips, P.C.B. (1991). Optimal inference in cointegrating systems. *Econometrica* 59, 283–306.

- Schuster, A. (1906). On the periodicities of sunspots. *Philosophical Transactions of Royal Society of London A*, 206, 69–100.
- Shiryayev, A.N. (ed.) (1992). *Selected Works of A.N. Kolmogorov, vol. II: Probability Theory and Mathematical Statistics*. Dordrecht: Kluwer.
- Sims, C.A. (1980). Macroeconomics and reality. *Econometrica* 48, 1–48.
- Slutsky, E. (1927). The summation of random causes as the source of cyclic processes (in Russian); English translation in *Econometrica* 5, (1937).
- Spanos, A. (1986). *Statistical Foundations of Econometric Modelling*. Cambridge: Cambridge University Press.
- Spanos, A. (1987). Error autocorrelation revisited: the AR(1) case. *Econometric Reviews* 6, 285–94.
- Spanos, A. (1990). Unit roots and their dependence of the conditioning information set. *Advances in Econometrics* 8, 271–92.
- Spanos, A. (1995). On normality and the linear regression model. *Econometric Reviews* 14, 195–203.
- Spanos, A. (1999). *Probability Theory and Statistical Inference: Econometric Modeling with Observational Data*. Cambridge: Cambridge University Press.
- Spanos, A., and A. McGuirk (1999). The power of unit root tests revisited. Mimeo, Virginia Polytechnic Institute and State University.
- Stigler, S.M. (1986). *The History of Statistics: the Measurement of Uncertainty before 1900*, Cambridge, MA: Harvard University Press.
- Wold, H.O. (1938). *A Study in the Analysis of Stationary Time Series* (revised 1954) Uppsala: Almquist and Wicksell.
- Yule, G.U. (1921). On the time-correlation problem. *Journal of the Royal Statistical Society* 84, 497–526.
- Yule, G.U. (1926). Why do we sometimes get nonsense correlations between time series – a study in sampling and the nature of time series. *Journal of the Royal Statistical Society* 89, 1–64.
- Yule, G.U. (1927). On a method of investigating periodicities in disturbed series, with special reference to Wolfer's sunspot numbers. *Philosophical Transactions of the Royal Society series A*, 226, 267–98.

CHAPTER TWENTY-NINE

# Unit Roots

*Herman J. Bierens\**

## 1 INTRODUCTION

In this chapter I will explain the two most frequently applied types of unit root tests, namely the Augmented Dickey–Fuller tests (see Fuller, 1996; Dickey and Fuller, 1979, 1981), and the Phillips–Perron tests (see Phillips, 1987; Phillips and Perron, 1988). The statistics and econometrics levels required for understanding the material below are Hogg and Craig (1978) or a similar level for statistics, and Green (1997) or a similar level for econometrics. The functional central limit theorem (see Billingsley, 1968), which plays a key role in the derivations involved, will be explained in this chapter by showing its analogy with the concept of convergence in distribution of random variables, and by confining the discussion to Gaussian unit root processes.

This chapter is not a review of the vast literature on unit roots. Such a review would entail a long list of descriptions of the many different recipes for unit root testing proposed in the literature, and would leave no space for motivation, let alone proofs. I have chosen for depth rather than breadth, by focusing on the most influential papers on unit root testing, and discussing them in detail, without assuming that the reader has any previous knowledge about this topic.

As an introduction to the concept of a unit root and its consequences, consider the Gaussian AR(1) process  $y_t = \beta_0 + \beta_1 y_{t-1} + u_t$ , or equivalently  $(1 - \beta_1 L)y_t = \beta_0 + u_t$ , where  $L$  is the lag operator:  $Ly_t = y_{t-1}$ , and the  $u_t$ s are iid  $N(0, \sigma^2)$ . The lag polynomial  $1 - \beta_1 L$  has root equal to  $1/\beta_1$ . If  $|\beta_1| < 1$ , then by backwards substitution we can write  $y_t = \beta_0/(1 - \beta_1) + \sum_{j=0}^{\infty} \beta_1^j u_{t-j}$ , so that  $y_t$  is strictly stationary, i.e. for arbitrary natural numbers  $m_1 < m_2 < \dots < m_{k-1}$  the joint distribution of  $y_t, y_{t-m_1}, y_{t-m_2}, \dots, y_{t-m_{k-1}}$  does not depend on  $t$ , but only on the lags or leads  $m_1, m_2, \dots, m_{k-1}$ . Moreover, the distribution of  $y_t$ ,  $t > 0$ , conditional on  $y_0, y_{-1}, y_{-2}, \dots$ , then converges to the marginal distribution of  $y_t$  if  $t \rightarrow \infty$ . In other words,  $y_t$  has a vanishing memory:  $y_t$  becomes independent of its past,  $y_0, y_{-1}, y_{-2}, \dots$ , if  $t \rightarrow \infty$ .

If  $\beta_1 = 1$ , so that the lag polynomial  $1 - \beta_1 L$  has a unit root, then  $y_t$  is called a unit root process. In this case the AR(1) process under review becomes  $y_t = y_{t-1} +$

$\beta_0 + u_t$ , which by backwards substitution yields for  $t > 0$ ,  $y_t = y_0 + \beta_0 t + \sum_{j=1}^t u_j$ . Thus now the distribution of  $y_t$ ,  $t > 0$ , conditional on  $y_0, y_{-1}, y_{-2}, \dots$ , is  $N(y_0 + \beta_0 t, \sigma^2 t)$ , so that  $y_t$  has no longer a vanishing memory: a shock in  $y_0$  will have a persistent effect on  $y_t$ . The former intercept  $\beta_0$  now becomes the *drift* parameter of the unit root process involved.

It is important to distinguish stationary processes from unit root processes, for the following reasons.

1. Regressions involving unit root processes may give spurious results. If  $y_t$  and  $x_t$  are mutually independent unit root processes, i.e.  $y_t$  is independent of  $x_{t-j}$  for all  $t$  and  $j$ , then the OLS regression of  $y_t$  on  $x_t$  for  $t = 1, \dots, n$ , with or without an intercept, will yield a significant estimate of the slope parameter if  $n$  is large: the absolute value of the  $t$ -value of the slope converges in probability to  $\infty$  if  $n \rightarrow \infty$ . We then might conclude that  $y_t$  depends on  $x_t$ , while in reality the  $y_t$ s are independent of the  $x_t$ s. This phenomenon is called *spurious regression*.<sup>1</sup> One should therefore be very cautious when conducting standard econometric analysis using time series. If the time series involved are unit root processes, naive application of regression analysis may yield nonsense results.

2. For two or more unit root processes there may exist linear combinations which are stationary, and these linear combinations may be interpreted as long-run relationships. This phenomenon is called *cointegration*,<sup>2</sup> and plays a dominant role in modern empirical macroeconomic research.

3. Tests of parameter restrictions in (auto)regressions involving unit root processes have in general different null distributions than in the case of stationary processes. In particular, if one would test the null hypothesis  $\beta_1 = 1$  in the above AR(1) model using the usual  $t$ -test, the null distribution involved is nonnormal. Therefore, naive application of classical inference may give incorrect results. We will demonstrate the latter first, and in the process derive the Dickey–Fuller test (see Fuller, 1996; Dickey and Fuller, 1979, 1981), by rewriting the AR(1) model as

$$\Delta y_t = y_t - y_{t-1} = \beta_0 + (\beta_1 - 1)y_{t-1} + u_t = \alpha_0 + \alpha_1 y_{t-1} + u_t, \quad (29.1)$$

say, estimating the parameter  $\alpha_1$  by OLS on the basis of observations  $y_0, y_1, \dots, y_n$ , and then testing the unit root hypothesis  $\alpha_1 = 0$  against the stationarity hypothesis  $-2 < \alpha_1 < 0$ , using the  $t$ -value of  $\alpha_1$ . In Section 2 we consider the case where  $\alpha_0 = 0$  under both the unit root hypothesis and the stationarity hypothesis. In Section 3 we consider the case where  $\alpha_0 = 0$  under the unit root hypothesis but not under the stationarity hypothesis.

The assumption that the error process  $u_t$  is independent is quite unrealistic for macroeconomic time series. Therefore, in Sections 4 and 5 this assumption will be relaxed, and two types of appropriate unit root tests will be discussed: the augmented Dickey–Fuller (ADF) tests, and the Phillips–Perron (PP) tests.

In Section 6 we consider the unit root *with drift* case, and we discuss the ADF and PP tests of the unit root with drift hypothesis, against the alternative of trend stationarity.

Finally, Section 7 contains some concluding remarks.

## 2 THE GAUSSIAN AR(1) CASE WITHOUT INTERCEPT: PART 1

### 2.1 Introduction

Consider the AR(1) model without intercept, rewritten as<sup>3</sup>

$$\Delta y_t = \alpha_0 y_{t-1} + u_t, \text{ where } u_t \text{ is iid } N(0, \sigma^2), \quad (29.2)$$

and  $y_t$  is observed for  $t = 1, 2, \dots, n$ . For convenience I will assume that

$$y_t = 0 \text{ for } t \leq 0. \quad (29.3)$$

This assumption is, of course, quite unrealistic, but is made for the sake of transparency of the argument, and will appear to be innocent.

The OLS estimator of  $\alpha_0$  is:

$$\hat{\alpha}_0 = \frac{\sum_{t=1}^n y_{t-1} \Delta y_t}{\sum_{t=1}^n y_{t-1}^2} = \alpha_0 + \frac{\sum_{t=1}^n y_{t-1} u_t}{\sum_{t=1}^n y_{t-1}^2}. \quad (29.4)$$

If  $-2 < \alpha_0 < 0$ , so that  $y_t$  is stationary, then it is a standard exercise to verify that  $\sqrt{n}(\hat{\alpha}_0 - \alpha_0) \rightarrow N(0, 1 - (1 + \alpha_0)^2)$  in distribution. On the other hand, if  $\alpha_0 = 0$ , so that  $y_t$  is a unit root process, this result reads:  $\sqrt{n}\hat{\alpha}_0 \rightarrow N(0, 0)$  in distribution, hence  $\text{plim}_{n \rightarrow \infty} \sqrt{n}\hat{\alpha}_0 = 0$ . However, we show now that a much stronger result holds, namely that  $\hat{\rho}_0 \equiv n\hat{\alpha}_0$  converges in distribution, but the limiting distribution involved is nonnormal. Thus, the presence of a unit root is actually advantageous for the efficiency of the OLS estimator  $\hat{\alpha}_0$ . The main problem is that the  $t$ -test of the null hypothesis that  $\alpha_0 = 0$  has no longer a standard normal asymptotic null distribution, so that we cannot test for a unit root using standard methods. The same applies to more general unit root processes.

In the unit root case under review we have  $y_t = y_{t-1} + u_t = y_0 + \sum_{j=1}^t u_j = \sum_{j=1}^t u_j$  for  $t > 0$ , where the last equality involved is due to assumption (29.3). Denoting

$$S_t = 0 \text{ for } t \leq 0, S_t = \sum_{j=1}^t u_j \text{ for } t \geq 1. \quad (29.5)$$

and  $\hat{\sigma}^2 = (1/n)\sum_{t=1}^n u_t^2$ , it follows that

$$\begin{aligned} \frac{1}{n} \sum_{t=1}^n u_t y_{t-1} &= \frac{1}{2n} \sum_{t=1}^n ((u_t + y_{t-1})^2 - y_{t-1}^2 - u_t^2) = \frac{1}{2} \left( \frac{1}{n} \sum_{t=1}^n y_t^2 - \frac{1}{n} \sum_{t=1}^n y_{t-1}^2 - \frac{1}{n} \sum_{t=1}^n u_t^2 \right) \\ &= \frac{1}{2} (y_n^2/n - y_0^2/n - \hat{\sigma}^2) = \frac{1}{2} (S_n^2/n - \hat{\sigma}^2), \end{aligned} \quad (29.6)$$

and similarly,

$$\frac{1}{n^2} \sum_{t=1}^n y_{t-1}^2 = \frac{1}{n} \sum_{t=1}^n (S_{t-1}/\sqrt{n})^2. \quad (29.7)$$

Next, let

$$W_n(x) = S_{[nx]}/(\sigma\sqrt{n}) \quad \text{for } x \in [0, 1], \quad (29.8)$$

where  $[z]$  means truncation to the nearest integer  $\leq z$ . Then we have:<sup>4</sup>

$$\begin{aligned} \frac{1}{n} \sum_{t=1}^n u_t y_{t-1} &= \frac{1}{2} (\sigma^2 W_n(1)^2 - \hat{\sigma}^2) \\ &= \frac{1}{2} (\sigma^2 W_n(1)^2 - \sigma^2 - O_p(1/\sqrt{n})) = \sigma^2 \frac{1}{2} (W_n(1)^2 - 1) + o_p(1), \end{aligned} \quad (29.9)$$

and

$$\frac{1}{n^2} \sum_{t=1}^n y_{t-1}^2 = \frac{1}{n} \sum_{t=1}^n \sigma^2 W_n((t-1)/n)^2 = \int W_n(x)^2 dx, \quad (29.10)$$

where the integral in (29.10) and below, unless otherwise indicated, is taken over the unit interval  $[0, 1]$ . The last equality in (29.9) follows from the law of large numbers, by which  $\hat{\sigma}^2 = \sigma^2 + O_p(1/\sqrt{n})$ . The last equality in (29.10) follows from the fact that for any power  $m$ ,

$$\begin{aligned} \int W_n(x)^m dx &= \int_0^1 W_n(x)^m dx = \frac{1}{n} \int_0^n W_n(z/n)^m dz = \frac{1}{n} \sum_{t=1}^n \int_{t-1}^t W_n(z/n)^m dz \\ &= \frac{1}{n^{1+m/2}} \sum_{t=1}^n \int_{t-1}^t (S_{[z]}/\sigma)^m dz = \frac{1}{n^{1+m/2}} \sum_{t=1}^n (S_{t-1}/\sigma)^m. \end{aligned} \quad (29.11)$$

Moreover, observe from (29.11), with  $m = 1$ , that  $\int W_n(x) dx$  is a linear combination of iid standard normal random variables, and therefore normal itself, with zero mean and variance

$$\begin{aligned} E \left( \int W_n(x) dx \right)^2 &= \iint E(W_n(x) W_n(y)) dx dy \\ &= \iint \frac{\min([nx], [ny])}{n} dx dy \rightarrow \iint \min(x, y) dx dy = \frac{1}{3}. \end{aligned} \quad (29.12)$$

Thus,  $\int W_n(x)dx \rightarrow N(0, 1/3)$  in distribution. Since  $\int W_n(x)^2dx \geq (\int W_n(x)dx)^2$ , it follows therefore that  $\int W_n(x)^2dx$  is bounded away from zero:

$$\left( \int W_n(x)^2 dx \right)^{-1} = O_p(1). \quad (29.13)$$

Combining (29.9), (29.10), and (29.13), we now have:

$$\hat{\rho}_0 \equiv n\hat{\alpha}_0 = \frac{(1/n)\sum_{t=1}^n u_t y_{t-1}}{(1/n^2)\sum_{t=1}^n y_{t-1}^2} = \frac{(1/2)(W_n(1)^2 - 1) + o_p(1)}{\int W_n(x)^2 dx} = \frac{1}{2} \left( \frac{W_n(1)^2 - 1}{\int W_n(x)^2 dx} \right) + o_p(1). \quad (29.14)$$

This result does not depend on assumption (29.3).

## 2.2 Weak convergence of random functions

In order to establish the limiting distribution of (29.14), and other asymptotic results, we need to extend the well known concept of convergence in distribution of random variables to convergence in distribution of a sequence of random functions. Recall that for random variables  $X_n$ ,  $X$ ,  $X_n \rightarrow X$  in distribution if the distribution function  $F_n(x)$  of  $X_n$  converges pointwise to the distribution function  $F(x)$  of  $X$  in the continuity points of  $F(x)$ . Moreover, recall that distribution functions are uniquely associated to probability measures on the Borel sets,<sup>5</sup> i.e. there exists one and only one probability measure  $\mu_n(B)$  on the Borel sets  $B$  such that  $F_n(x) = \mu_n((-\infty, x])$ , and similarly,  $F(x)$  is uniquely associated to a probability measure  $\mu$  on the Borel sets, such that  $F(x) = \mu((-\infty, x])$ . The statement  $X_n \rightarrow X$  in distribution can now be expressed in terms of the probability measures  $\mu_n$  and  $\mu : \mu_n(B) \rightarrow \mu(B)$  for all Borel sets  $B$  with boundary  $\delta B$  satisfying  $\mu(\delta B) = 0$ .

In order to extend the latter to random functions, we need to define Borel sets of functions. For our purpose it suffices to define Borel sets of continuous functions on  $[0, 1]$ . Let  $C[0, 1]$  be the set of all continuous functions on the unit interval  $[0, 1]$ . Define the distance between two functions  $f$  and  $g$  in  $C[0, 1]$  by the sup-norm:  $\rho(f, g) = \sup_{0 \leq x \leq 1} |f(x) - g(x)|$ . Endowed with this norm, the set  $C[0, 1]$  becomes a metric space, for which we can define open subsets, similarly to the concept of an open subset of  $\mathbb{R}$ : A set  $B$  in  $C[0, 1]$  is open if for each function  $f$  in  $B$  we can find an  $\varepsilon > 0$  such that  $\{g \in C[0, 1] : \rho(g, f) < \varepsilon\} \subset B$ . Now the smallest  $\sigma$ -algebra of subsets of  $C[0, 1]$  containing the collection of all open subsets of  $C[0, 1]$  is just the collection of Borel sets of functions in  $C[0, 1]$ .

A random element of  $C[0, 1]$  is a random function  $W(x)$ , say, on  $[0, 1]$ , which is continuous with probability 1. For such a random element  $W$ , say, we can define a probability measure  $\mu$  on the Borel sets  $B$  in  $C[0, 1]$  by  $\mu(B) = P(W \in B)$ . Now a sequence  $W_n^*$  of random elements of  $C[0, 1]$ , with corresponding probability

measures  $\mu_n$ , is said to converge weakly to a random element  $W$  of  $C[0, 1]$ , with corresponding probability measure  $\mu$ , if for each Borel set  $B$  in  $C[0, 1]$  with boundary  $\delta B$  satisfying  $\mu(\delta B) = 0$ , we have  $\mu_n(B) \rightarrow \mu(B)$ . This is usually denoted by:  $W_n^* \Rightarrow W$  (on  $[0, 1]$ ). Thus, weak convergence is the extension to random functions of the concept of convergence in distribution.

In order to verify that  $W_n^* \Rightarrow W$  on  $[0, 1]$ , we have to verify two conditions. See Billingsley (1963). First, we have to verify that the finite distributions of  $W_n^*$  converge to the corresponding finite distributions of  $W$ , i.e. for arbitrary points  $x_1, \dots, x_m$  in  $[0, 1]$ ,  $(W_n^*(x_1), \dots, W_n^*(x_m)) \Rightarrow (W(x_1), \dots, W(x_m))$  in distribution. Second, we have to verify that  $W_n^*$  is tight. Tightness is the extension of the concept of stochastic boundedness to random functions: for each  $\epsilon$  in  $[0, 1]$  there exists a compact (Borel) set  $K$  in  $C[0, 1]$  such that  $\mu_n(K) > 1 - \epsilon$  for  $n = 1, 2, \dots$ . Since convergence in distribution implies stochastic boundedness, we cannot have convergence in distribution without stochastic boundedness, and the same applies to weak convergence: tightness is a necessary condition for weak convergence.

As is well known, if  $X_n \rightarrow X$  in distribution, and  $\Phi$  is a continuous mapping from the support of  $X$  into a Euclidean space, then by Slutsky's theorem,  $\Phi(X_n) \rightarrow \Phi(X)$  in distribution. A similar result holds for weak convergence, which is known as the continuous mapping theorem: if  $\Phi$  is a continuous mapping from  $C[0, 1]$  into a Euclidean space, then  $W_n^* \Rightarrow W$  implies  $\Phi(W_n^*) \rightarrow \Phi(W)$  in distribution. For example, the integral  $\Phi(f) = \int f(x)^2 dx$  with  $f \in C[0, 1]$  is a continuous mapping from  $C[0, 1]$  into the real line, hence  $W_n^* \Rightarrow W$  implies that  $\int W_n^*(x)^2 dx \rightarrow \int W(x)^2 dx$  in distribution.

The random function  $W_n$  defined by (29.8) is a step function on  $[0, 1]$ , and therefore not a random element of  $C[0, 1]$ . However, the steps involved can be smoothed by piecewise linear interpolation, yielding a random element  $W_n^*$  of  $C[0, 1]$  such that  $\sup_{0 \leq x \leq 1} |W_n^*(x) - W_n(x)| = o_p(1)$ . The finite distributions of  $W_n^*$  are therefore asymptotically the same as the finite distributions of  $W_n$ . In order to analyze the latter, redefine  $W_n$  as

$$W_n(x) = \frac{1}{\sqrt{n}} \sum_{t=1}^{[nx]} e_t \text{ for } x \in [n^{-1}, 1], \quad W_n(x) = 0 \text{ for } x \in [0, n^{-1}], \quad e_t \text{ is iid } N(0, 1). \quad (29.15)$$

(Thus,  $e_t = u_t/\sigma$ , and let

$$\begin{aligned} W_n^*(x) &= W_n\left(\frac{t-1}{n}\right) + (nx - (t-1)) \left( W_n\left(\frac{t}{n}\right) - W_n\left(\frac{t-1}{n}\right) \right) \\ &= W_n(x) + (nx - (t-1)) \frac{e_t}{\sqrt{n}} \text{ for } x \in \left[\frac{t-1}{n}, \frac{t}{n}\right], t = 1, \dots, n, W_n^*(0) = 0. \end{aligned} \quad (29.16)$$

Then

$$\sup_{0 \leq x \leq 1} |W_n^*(x) - W_n(x)| \leq \frac{\max_{1 \leq t \leq n} |e_t|}{\sqrt{n}} = o_p(1). \quad (29.17)$$

The latter conclusion is not too hard an exercise.<sup>6</sup>

It is easy to verify that for *fixed*  $0 \leq x < y \leq 1$  we have

$$\begin{aligned} \begin{pmatrix} W_n(x) \\ W_n(y) - W_n(x) \end{pmatrix} &= \frac{1}{\sqrt{n}} \begin{pmatrix} \sum_{t=1}^{[nx]} e_t \\ \sum_{t=[nx]+1}^{[ny]} e_t \end{pmatrix} \sim N_2 \left( \begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} \frac{[nx]}{n} & 0 \\ 0 & \frac{[ny] - [nx]}{n} \end{pmatrix} \right) \\ &\rightarrow \begin{pmatrix} W(x) \\ W(y) - W(x) \end{pmatrix} \text{ in distribution,} \end{aligned} \quad (29.18)$$

where  $W(x)$  is a random function on  $[0, 1]$  such that for  $0 \leq x < y \leq 1$ ,

$$\begin{pmatrix} W(x) \\ W(y) - W(x) \end{pmatrix} \sim N_2 \left( \begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} x & 0 \\ 0 & y-x \end{pmatrix} \right). \quad (29.19)$$

This random function  $W(x)$  is called a standard Wiener process, or Brownian motion. Similarly, for arbitrary fixed  $x, y$  in  $[0, 1]$ ,

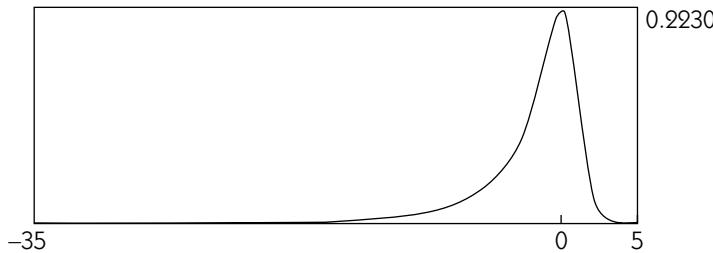
$$\begin{pmatrix} W_n(x) \\ W_n(y) \end{pmatrix} \rightarrow \begin{pmatrix} W(x) \\ W(y) \end{pmatrix} \sim N_2 \left( \begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} x & \min(x, y) \\ \min(x, y) & y \end{pmatrix} \right) \text{ in distribution} \quad (29.20)$$

and it follows from (29.17) that the same applies to  $W_n^*$ . Therefore, the finite distributions of  $W_n^*$  converge to the corresponding finite distributions of  $W$ . Also, it can be shown that  $W_n^*$  is tight (see Billingsley, 1963). Hence,  $W_n^* \Rightarrow W$ , and by the continuous mapping theorem,

$$\begin{aligned} (W_n^*(1), \int W_n^*(x) dx, \int W_n^*(x)^2 dx, \int x W_n^*(x) dx)^T &\rightarrow \\ (W(1), \int W(x) dx, \int W(x)^2 dx, \int x W(x) dx)^T \end{aligned} \quad (29.21)$$

in distribution. This result, together with (29.17), implies that:

**Lemma 1.** For  $W_n$  defined by (29.15),  $(W_n(1), \int W_n(x) dx, \int W_n(x)^2 dx, \int x W_n(x) dx)^T$  converges jointly in distribution to  $(W(1), \int W(x) dx, \int W(x)^2 dx, \int x W(x) dx)^T$ .

Figure 29.1 Density of  $p_0$ 

### 2.3 Asymptotic null distributions

Using Lemma 1, it follows now straightforwardly from (29.14) that:

$$\hat{p}_0 \equiv n\hat{\alpha}_0 \rightarrow p_0 \equiv \frac{1}{2} \left( \frac{W(1)^2 - 1}{\int W(x)^2 dx} \right) \text{ in distribution.} \quad (29.22)$$

The density<sup>7</sup> of the distribution of  $p_0$  is displayed in Figure 29.1, which clearly shows that the distribution involved is nonnormal and asymmetric, with a fat left tail.

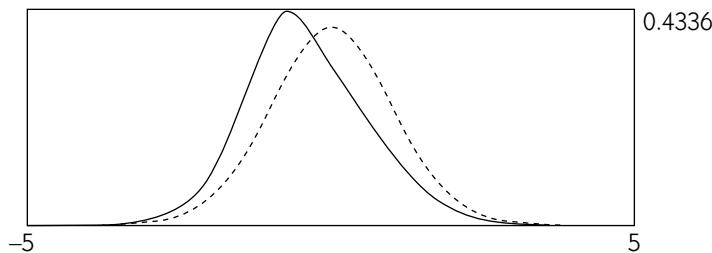
Also the limiting distribution of the usual  $t$ -test statistic of the null hypothesis  $\alpha_0 = 0$  is nonnormal. First, observe that due to (29.10), (29.22), and Lemma 1, the residual sum of squares (RSS) of the regression (29.2) under the unit root hypothesis is:

$$\text{RSS} = \sum_{t=1}^n (\Delta y_t - \hat{\alpha}_0 y_{t-1})^2 = \sum_{t=1}^n u_t^2 - (n\hat{\alpha}_0)^2 (1/n^2) \sum_{t=1}^n y_{t-1}^2 = \sum_{t=1}^n u_t^2 + O_p(1). \quad (29.23)$$

Hence  $\text{RSS}/(n-1) = \sigma^2 + O_p(1/n)$ . Therefore, similarly to (29.14) and (29.22), the Dickey–Fuller  $t$ -statistic  $\hat{\tau}_0$  involved satisfies:

$$\begin{aligned} \hat{\tau}_0 &\equiv n\hat{\alpha}_0 \frac{\sqrt{(1/n^2)\sum_{t=1}^n y_{t-1}^2}}{\sqrt{\text{RSS}/(n-1)}} = \frac{(W_n(1)^2 - 1)/2}{\sqrt{\int W_n(x)^2 dx}} + o_p(1) \rightarrow \\ \tau_0 &\equiv \frac{(W(1)^2 - 1)/2}{\sqrt{\int W(x)^2 dx}} \text{ in distribution.} \end{aligned} \quad (29.24)$$

Note that the unit root tests based on the statistics  $\hat{p}_0 \equiv n\hat{\alpha}_0$  and  $\hat{\tau}_0$  are left-sided: under the alternative of stationarity,  $-2 < \alpha_0 < 0$ , we have  $\text{plim}_{n \rightarrow \infty} \hat{\alpha}_0 = \alpha_0 < 0$ , hence  $\hat{p}_0 \rightarrow -\infty$  in probability at rate  $n$ , and  $\hat{\tau}_0 \rightarrow -\infty$  in probability at rate  $\sqrt{n}$ .



**Figure 29.2** Density of  $\tau_0$  compared with the standard normal density (dashed curve)

The nonnormality of the limiting distributions  $p_0$  and  $\tau_0$  is no problem, though, as long one is aware of it. The distributions involved are free of nuisance parameters, and asymptotic critical values of the unit root tests  $\hat{p}_0$  and  $\hat{\tau}_0$  can easily be tabulated, using Monte Carlo simulation. In particular,

$$P(\tau_0 \leq -1.95) = 0.05, \quad P(\tau_0 \leq -1.62) = 0.10, \quad (29.25)$$

(see Fuller, 1996, p. 642), whereas for a standard normal random variable  $e$ ,

$$P(e \leq -1.64) = 0.05, \quad P(e \leq -1.28) = 0.10. \quad (29.26)$$

In Figure 29.2 the density of  $\tau_0$  is compared with the standard normal density. We see that the density of  $\tau_0$  is shifted to left of the standard normal density, which causes the difference between (29.25) and (29.26). Using the left-sided standard normal test would result in a type 1 error of about twice the size: compare (29.26) with

$$P(\tau_0 \leq -1.64) \approx 0.09, \quad P(\tau_0 \leq -1.28) \approx 0.18 \quad (29.27)$$

### 3 THE GAUSSIAN AR(1) CASE WITH INTERCEPT UNDER THE ALTERNATIVE OF STATIONARITY

If under the stationarity hypothesis the AR(1) process has an intercept, but not under the unit root hypothesis, the AR(1) model that covers both the null and the alternative is:

$$\Delta y_t = \alpha_0 + \alpha_1 y_{t-1} + u_t, \quad \text{where } \alpha_0 = -c\alpha_1. \quad (29.28)$$

If  $-2 < \alpha_1 < 0$ , then the process  $y_t$  is stationary around the constant  $c$ :

$$y_t = -c\alpha_1 + (1 + \alpha_1)y_{t-1} + u_t = \sum_{j=0}^{\infty} (1 + \alpha_1)^j (-c\alpha_1 + u_{t-j}) = c + \sum_{j=0}^{\infty} (1 + \alpha_1)^j u_{t-j}, \quad (29.29)$$

hence  $E(y_t^2) = c^2 + (1 - (1 + \alpha_1)^2)^{-1}\sigma^2$ ,  $E(y_t y_{t-1}) = c^2 + (1 + \alpha_1)(1 - (1 + \alpha_1)^2)^{-1}\sigma^2$ , and

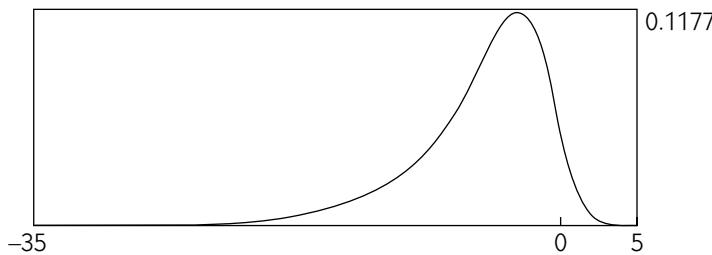


Figure 29.3 Density of  $\rho_1$

$$\operatorname{plim}_{n \rightarrow \infty} \hat{\alpha}_0 = \frac{E(y_t y_{t-1})}{E(y_{t-1}^2)} - 1 = \frac{\alpha_1}{1 + (c/\sigma)^2(1 - (1 + \alpha_1)^2)}, \quad (29.30)$$

which approaches zero if  $c^2/\sigma^2 \rightarrow \infty$ . Therefore, the power of the test  $\hat{\rho}_0$  will be low if the variance of  $u_t$  is small relative to  $[E(y_t)]^2$ . The same applies to the  $t$ -test  $\hat{\tau}_0$ . We should therefore use the OLS estimator of  $\alpha_1$  and the corresponding  $t$ -value in the regression of  $\Delta y_t$  on  $y_{t-1}$  with intercept.

Denoting  $\bar{y}_{-1} = (1/n)\sum_{t=1}^n y_{t-1}$ ,  $\bar{u} = (1/n)\sum_{t=1}^n u_t$ , the OLS estimator of  $\alpha_1$  is:

$$\hat{\alpha}_1 = \alpha_1 + \frac{\sum_{t=1}^n u_t y_{t-1} - n\bar{u}\bar{y}_{-1}}{\sum_{t=1}^n y_{t-1}^2 - n\bar{y}_{-1}^2}. \quad (29.31)$$

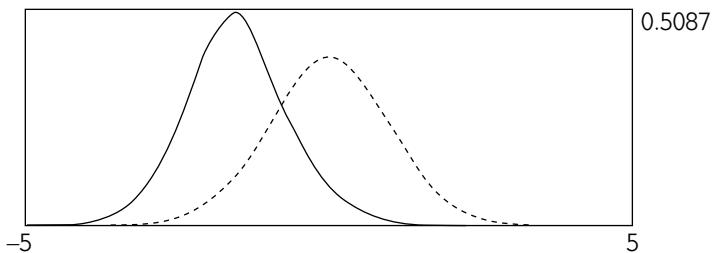
Since by (29.8),  $\sqrt{n}\bar{u} = \sigma W_n(1)$ , and under the null hypothesis  $\alpha_1 = 0$  and the maintained hypothesis (29.3),

$$\bar{y}_{-1}/\sqrt{n} = \frac{1}{n\sqrt{n}} \sum_{t=1}^n S_{t-1} = \sigma \int W_n(x) dx, \quad (29.32)$$

where the last equality follows from (29.11) with  $m = 1$ , it follows from Lemma 1, similarly to (29.14) and (29.22) that

$$\begin{aligned} \hat{\rho}_1 &\equiv n\hat{\alpha}_1 = \frac{(1/2)(W_n(1)^2 - 1) - W_n(1)\int W_n(x) dx}{\int W_n(x)^2 dx - (\int W_n(x) dx)^2} + o_p(1) \\ &\rightarrow \rho_1 \equiv \frac{(1/2)(W(1)^2 - 1) - W(1)\int W(x) dx}{\int W(x)^2 dx - (\int W(x) dx)^2} \text{ in distribution.} \end{aligned} \quad (29.33)$$

The density of  $\rho_1$  is displayed in Figure 29.3. Comparing Figures 29.1 and 29.3, we see that the density of  $\rho_1$  is farther left of zero than the density of  $\rho_0$ , and has a fatter left tail.



**Figure 29.4** Density of  $\tau_1$  compared with the standard normal density (dashed curve)

As to the  $t$ -value  $\hat{\tau}_1$  of  $\alpha_1$  in this case, it follows similarly to (29.24) and (29.33) that under the unit root hypothesis,

$$\hat{\tau}_1 \rightarrow \tau_1 \equiv \frac{(1/2)(W(1)^2 - 1) - W(1)\int W(x)dx}{\sqrt{\int W(x)^2 dx - (\int W(x)dx)^2}} \text{ in distribution.} \quad (29.34)$$

Again, the results (29.33) and (29.34) do not hinge on assumption (29.3).

The distribution of  $\tau_1$  is even farther away from the normal distribution than the distribution of  $\tau_0$ , as follows from comparison of (29.26) with

$$P(\tau_1 \leq -2.86) = 0.05, \quad P(\tau_1 \leq -2.57) = 0.1 \quad (29.35)$$

See again Fuller (1996, p. 642). This is corroborated by Figure 29.4, where the density of  $\tau_1$  is compared with the standard normal density.

We see that the density of  $\tau_1$  is shifted even more to the left of the standard normal density than in Figure 29.2, hence the left-sided standard normal test would result in a dramatically higher type 1 error than in the case without an intercept: compare

$$P(\tau_1 \leq -1.64) \approx 0.46, \quad P(\tau_1 \leq -1.28) \approx 0.64 \quad (29.36)$$

with (29.26) and (29.27).

#### 4 GENERAL AR PROCESSES WITH A UNIT ROOT, AND THE AUGMENTED DICKEY–FULLER TEST

The assumption made in Sections 2 and 3 that the data-generating process is an AR(1) process, is not very realistic for macroeconomic time series, because even after differencing most of these time series will still display a fair amount of dependence. Therefore we now consider an AR( $p$ ) process:

$$y_t = \beta_0 + \sum_{j=1}^p \beta_j y_{t-j} + u_t, \quad u_t \sim \text{iid } N(0, \sigma^2). \quad (29.37)$$

By recursively replacing  $y_{t-j}$  by  $\Delta y_{t-j} + y_{t-1-j}$  for  $j = 0, 1, \dots, p-1$ , this model can be written as

$$\Delta y_t = \alpha_0 + \sum_{j=1}^{p-1} \alpha_j \Delta y_{t-j} + \alpha_p y_{t-p} + u_t, \quad u_t \sim \text{iid } N(0, \sigma^2), \quad (29.38)$$

where  $\alpha_0 = \beta_0$ ,  $\alpha_j = \sum_{i=1}^j \beta_i - 1$ ,  $j = 1, \dots, p$ . Alternatively and equivalently, by recursively replacing  $y_{t-p+j}$  by  $y_{t-p+j+1} - \Delta y_{t-p+j+1}$  for  $j = 0, 1, \dots, p-1$ , model (29.37) can also be written as

$$\Delta y_t = \alpha_0 + \sum_{j=1}^{p-1} \alpha_j \Delta y_{t-j} + \alpha_p y_{t-1} + u_t, \quad u_t \sim \text{iid } N(0, \sigma^2), \quad (29.39)$$

where now  $\alpha_j = -\sum_{i=1}^j \beta_i$ ,  $j = 1, \dots, p-1$ ,  $\alpha_p = \sum_{i=1}^p \beta_i - 1$ .

If the AP( $p$ ) process (29.37) has a unit root, then clearly  $\alpha_p = 0$  in (29.38) and (29.39). If the process (29.37) is stationary, i.e. all the roots of the lag polynomial  $1 - \sum_{i=1}^p \beta_i L^i$  lie outside the complex unit circle, then  $\alpha_p = \sum_{i=1}^p \beta_i - 1 < 0$  in (29.38) and (29.39).<sup>8</sup> The unit root hypothesis can therefore be tested by testing the null hypothesis  $\alpha_p = 0$  against the alternative hypothesis  $\alpha_p < 0$ , using the  $t$ -value  $\hat{\tau}_p$  of  $\alpha_p$  in model (29.38) or model (29.39). This test is known as the augmented Dickey–Fuller (ADF) test.

We will show now for the case  $p = 2$ , with intercept under the alternative, i.e.

$$\Delta y_t = \alpha_0 + \alpha_1 \Delta y_{t-1} + \alpha_2 y_{t-2} + u_t, \quad u_t \sim \text{iid } N(0, \sigma^2), \quad t = 1, \dots, n. \quad (29.40)$$

that under the unit root (without drift)<sup>9</sup> hypothesis the limiting distribution of  $n\hat{\alpha}_p$  is proportional to the limiting distribution in (29.33), and the limiting distribution of  $\hat{\tau}_p$  is the same as in (29.34).

Under the unit root hypothesis, i.e.  $\alpha_0 = \alpha_2 = 0$ ,  $|\alpha_1| < 1$ , we have

$$\begin{aligned} \Delta y_t &= \alpha_1 \Delta y_{t-1} + u_t = (1 - \alpha_1 L)^{-1} u_t = (1 - \alpha_1)^{-1} u_t + [(1 - \alpha_1 L)^{-1} - (1 - \alpha_1)^{-1}] u_t \\ &= (1 - \alpha_1)^{-1} u_t - \alpha_1 (1 - \alpha_1)^{-1} (1 - \alpha_1 L)^{-1} (1 - L) u_t = (1 - \alpha_1)^{-1} u_t + v_t - v_{t-1}, \end{aligned} \quad (29.41)$$

say, where  $v_t = -\alpha_1 (1 - \alpha_1)^{-1} (1 - \alpha_1 L)^{-1} u_t = -\alpha_1 (1 - \alpha_1)^{-1} \sum_{j=0}^{\infty} \alpha_1^j u_{t-j}$  is a stationary process. Hence:

$$\begin{aligned} y_t / \sqrt{n} &= y_0 / \sqrt{n} + v_t / \sqrt{n} - v_0 / \sqrt{n} + (1 - \alpha_1)^{-1} (1 / \sqrt{n}) \sum_{j=1}^t u_j \\ &= y_0 / \sqrt{n} + v_t / \sqrt{n} - v_0 / \sqrt{n} + \sigma (1 - \alpha_1)^{-1} W_n(t/n) \end{aligned} \quad (29.42)$$

and therefore, similarly to (29.6), (29.7), and (29.32), it follows that

$$(1/n) \sum_{t=1}^n y_{t-1} / \sqrt{n} = \sigma (1 - \alpha_1)^{-1} \int W_n(x) dx + o_p(1), \quad (29.43)$$

$$(1/n^2) \sum_{t=1}^n y_{t-1}^2 = \sigma^2 (1 - \alpha_1)^{-2} \int W_n(x)^2 dx + o_p(1), \quad (29.44)$$

$$\begin{aligned} (1/n) \sum_{t=1}^n u_t y_{t-1} &= (1/n) \sum_{t=1}^n u_t \left( (1 - \alpha_1)^{-1} \sum_{j=1}^{t-1} u_j + y_0 + v_{t-1} - v_0 \right) \\ &= (1 - \alpha_1)^{-1} (1/n) \sum_{t=1}^n u_t \sum_{j=1}^{t-1} u_j + (y_0 - v_0)(1/n) \sum_{t=1}^n u_t + (1/n) \sum_{t=1}^n u_t v_{t-1} \\ &= \frac{(1 - \alpha_1)^{-1} \sigma^2}{2} (W_n(1)^2 - 1) + o_p(1). \end{aligned} \quad (29.45)$$

Moreover,

$$\operatorname{plim}_{n \rightarrow \infty} (1/n) \sum_{t=1}^n \Delta y_{t-1} = E(\Delta y_t) = 0, \quad \operatorname{plim}_{n \rightarrow \infty} (1/n) \sum_{t=1}^n (\Delta y_{t-1})^2 = E(\Delta y_t)^2 = \sigma^2 / (1 - \alpha_1^2) \quad (29.46)$$

and

$$\begin{aligned} (1/n) \sum_{t=1}^n y_{t-1} \Delta y_{t-1} &= (1/n) \sum_{t=1}^n (\Delta y_{t-1})^2 + (1/n) \sum_{t=1}^n y_{t-2} \Delta y_{t-1} \\ &= (1/n) \sum_{t=1}^n (\Delta y_{t-1})^2 + \frac{1}{2} \left( (1/n) \sum_{t=1}^n y_{t-1}^2 - (1/n) \sum_{t=1}^n y_{t-2}^2 - (1/n) \sum_{t=1}^n (\Delta y_{t-1})^2 \right) \\ &= \frac{1}{2} \left( (1/n) \sum_{t=1}^n (\Delta y_{t-1})^2 + y_{n-1}^2/n - y_1^2/n \right) \\ &= \frac{1}{2} (\sigma^2 / (1 - \alpha_1^2) + \sigma^2 (1 - \alpha_1)^{-2} W_n(1)^2) + o_p(1) \end{aligned} \quad (29.47)$$

hence

$$(1/n) \sum_{t=1}^n y_{t-1} \Delta y_{t-1} / \sqrt{n} = O_p(1/\sqrt{n}). \quad (29.48)$$

Next, let  $\hat{\alpha} = (\hat{\alpha}_0, \hat{\alpha}_1, \hat{\alpha}_2)^T$  be the OLS estimator of  $\alpha = (\alpha_0, \alpha_1, \alpha_2)^T$ . Under the unit root hypothesis we have

$$\begin{pmatrix} \sqrt{n} \hat{\alpha}_0 \\ \sqrt{n} (\hat{\alpha}_1 - \alpha_1) \\ n \hat{\alpha}_2 \end{pmatrix} = \sqrt{n} D_n \hat{\Sigma}_{xx}^{-1} \hat{\Sigma}_{xu} = (D_n^{-1} \hat{\Sigma}_{xx} D_n^{-1})^{-1} \sqrt{n} D_n^{-1} \hat{\Sigma}_{xu}, \quad (29.49)$$

where

$$D_n = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & \sqrt{n} \end{pmatrix}, \quad (29.50)$$

$$\hat{\Sigma}_{xx} = \begin{pmatrix} 1 & (1/n)\sum_{t=1}^n \Delta y_{t-1} & (1/n)\sum_{t=1}^n y_{t-1} \\ (1/n)\sum_{t=1}^n \Delta y_{t-1} & (1/n)\sum_{t=1}^n (\Delta y_{t-1})^2 & (1/n)\sum_{t=1}^n y_{t-1} \Delta y_{t-1} \\ (1/n)\sum_{t=1}^n y_{t-1} & (1/n)\sum_{t=1}^n y_{t-1} \Delta y_{t-1} & (1/n)\sum_{t=1}^n y_{t-1}^2 \end{pmatrix}, \quad (29.51)$$

and

$$\hat{\Sigma}_{xu} = \begin{pmatrix} (1/n)\sum_{t=1}^n u_t \\ (1/n)\sum_{t=1}^n u_t \Delta y_{t-1} \\ (1/n)\sum_{t=1}^n u_t y_{t-1} \end{pmatrix}. \quad (29.52)$$

It follows from (29.43) through (29.48) that

$$D_n^{-1} \hat{\Sigma}_{xx} D_n^{-1} = \begin{pmatrix} 1 & 0 & \sigma(1 - \alpha_1)^{-1} \int W_n(x) dx \\ 0 & \sigma^2 / (1 - \alpha_1^2) & 0 \\ \sigma(1 - \alpha_1)^{-1} \int W_n(x) dx & 0 & \sigma^2(1 - \alpha_1)^{-2} \int W_n(x)^2 dx \end{pmatrix} + o_p(1), \quad (29.53)$$

hence, using the easy equality

$$\begin{pmatrix} 1 & 0 & a \\ 0 & b & 0 \\ a & 0 & c \end{pmatrix}^{-1} = \frac{1}{c - a^2} \begin{pmatrix} c & 0 & -a \\ 0 & b^{-1}(c - a^2) & 0 \\ -a & 0 & 1 \end{pmatrix},$$

it follows that

$$\begin{aligned} (D_n^{-1} \hat{\Sigma}_{xx} D_n^{-1})^{-1} &= \frac{\sigma^{-2}(1 - \alpha_1)^2}{\int W_n(x)^2 dx - (\int W_n(x) dx)^2} \\ &\times \begin{pmatrix} \sigma^2(1 - \alpha_1)^{-2} \int W_n(x)^2 dx & 0 & -\sigma(1 - \alpha_1)^{-1} \int W_n(x) dx \\ 0 & \frac{\int W_n(x)^2 dx - (\int W_n(x) dx)}{(1 - \alpha_1^2)(1 - \alpha_1)^2} & 0 \\ -\sigma(1 - \alpha_1)^{-1} \int W_n(x) dx & 0 & 1 \end{pmatrix} \\ &+ o_p(1). \end{aligned} \quad (29.54)$$

Moreover, it follows from (29.8) and (29.45) that

$$\sqrt{n}D_n^{-1}\hat{\Sigma}_{uu} = \begin{pmatrix} \sigma W_n(1) \\ (1/\sqrt{n})\sum_{t=1}^n u_t \Delta y_{t-1} \\ \sigma^2(1 - \alpha_1)^{-2}(W_n(1)^2 - 1)/2 \end{pmatrix} + o_p(1). \quad (29.55)$$

Combining (29.49), (29.54) and (29.55), and using Lemma 1, it follows now easily that

$$\frac{n\hat{\alpha}_2}{1 - \alpha_1} = \frac{\frac{1}{2}(W_n(1)^2 - 1) - W_n(1)\int W_n(x)dx}{\int W_n(x)^2 dx - (\int W_n(x)dx)^2} + o_p(1) \rightarrow \rho_1 \text{ in distribution}, \quad (29.56)$$

where  $\rho_1$  is defined in (29.33). Along the same lines it can be shown:

**Theorem 1.** Let  $y_t$  be generated by (29.39), and let  $\hat{\alpha}_p$  be the OLS estimator of  $\alpha_p$ . Under the unit root hypothesis, i.e.  $\alpha_p = 0$  and  $\alpha_0 = 0$ , the following hold: If model (29.39) is estimated without intercept, then  $n\hat{\alpha}_p \rightarrow (1 - \sum_{j=1}^{p-1} \alpha_j)\rho_0$  in distribution, where  $\rho_0$  is defined in (29.22). If model (29.39) is estimated with intercept, then  $n\hat{\alpha}_p \rightarrow (1 - \sum_{j=1}^{p-1} \alpha_j)\rho_1$  in distribution, where  $\rho_1$  is defined in (29.33). Moreover, under the stationarity hypothesis,  $\text{plim}_{n \rightarrow \infty} \hat{\alpha}_p = \alpha_p < 0$ , hence  $\text{plim}_{n \rightarrow \infty} \hat{\alpha}_p = -\infty$ , provided that in the case where the model is estimated without intercept this intercept,  $\alpha_0$ , is indeed zero.

Due to the factor  $1 - \sum_{j=1}^{p-1} \alpha_j$  in the limiting distribution of  $n\hat{\alpha}_p$  under the unit root hypothesis, we cannot use  $n\hat{\alpha}_p$  directly as a unit root test. However, it can be shown that under the unit root hypothesis this factor can be consistently estimated by  $1 - \sum_{j=1}^{p-1} \hat{\alpha}_j$ , hence we can use  $n\hat{\alpha}_p / |1 - \sum_{j=1}^{p-1} \hat{\alpha}_j|$  as a unit root test statistic, with limiting distribution given by (29.22) or (29.33). The reason for the absolute value is that under the alternative of stationarity the probability limit of  $1 - \sum_{j=1}^{p-1} \hat{\alpha}_j$  may be negative.<sup>10</sup>

The actual ADF test is based on the  $t$ -value of  $\alpha_p$ , because the factor  $1 - \sum_{j=1}^{p-1} \alpha_j$  will cancel out in the limiting distribution involved. We will show this for the AR(2) case.

First, it is not too hard to verify from (29.43) through (29.48), and (29.54), that the residual sum of squares RSS of the regression (29.40) satisfies:

$$\text{RSS} = \sum_{t=1}^n u_t^2 + O_p(1). \quad (29.57)$$

This result carries over to the general AR( $p$ ) case, and also holds under the stationarity hypothesis. Moreover, under the unit root hypothesis it follows easily from (29.54) and (29.57) that the OLS standard error,  $s_2$ , say, of  $\hat{\alpha}_2$  in model (29.40) satisfies:

$$\begin{aligned}
ns_2 &= \sqrt{\frac{(\text{RSS}/(n-3))\sigma^{-2}(1-\alpha_1)^2}{\int W_n(x)^2 dx - (\int W_n(x)dx)^2}} + o_p(1) \\
&= \frac{1-\alpha_1}{\sqrt{\int W_n(x)^2 dx - (\int W_n(x)dx)^2}} + o_p(1),
\end{aligned} \tag{29.58}$$

hence it follows from (29.56) that the  $t$ -value  $\hat{t}_2$  of  $\hat{\alpha}_2$  in model (29.40) satisfies (29.34). Again, this result carries over to the general AR( $p$ ) case:

**Theorem 2.** Let  $y_t$  be generated by (29.39), and let  $\hat{t}_p$  be  $t$ -value of the OLS estimator of  $\alpha_p$ . Under the unit root hypothesis, i.e.  $\alpha_p = 0$  and  $\alpha_0 = 0$ , the following hold: If model (29.39) is estimated without intercept, then  $\hat{t}_p \rightarrow \tau_0$  in distribution, where  $\tau_0$  is defined in (29.24). If model (29.39) is estimated with intercept, then  $\hat{t}_p \rightarrow \tau_1$  in distribution, where  $\tau_1$  is defined in (29.34). Moreover, under the stationarity hypothesis,  $\text{plim}_{n \rightarrow \infty} \hat{t}_p / \sqrt{n} < 0$ , hence  $\text{plim}_{n \rightarrow \infty} \hat{t}_p = -\infty$ , provided that in the case where the model is estimated without intercept this intercept,  $\alpha_0$ , is indeed zero.

## 5 ARIMA PROCESSES, AND THE PHILLIPS–PERRON TEST

The ADF test requires that the order  $p$  of the AR model involved is finite, and correctly specified, i.e. the specified order should not be smaller than the actual order. In order to analyze what happens if  $p$  is misspecified, suppose that the actual data generating process is given by (29.39) with  $\alpha_0 = \alpha_2 = 0$  and  $p > 1$ , and that the unit root hypothesis is tested on the basis of the assumption that  $p = 1$ . Denoting  $e_t = u_t/\sigma$ , model (29.39) with  $\alpha_0 = \alpha_2 = 0$  can be rewritten as

$$\Delta y_t = \left( \sum_{j=0}^{\infty} \gamma_j L^j \right) e_t = \gamma(L) e_t, \quad e_t \sim \text{iid } N(0, 1), \tag{29.59}$$

where  $\gamma(L) = \alpha(L)^{-1}$ , with  $\alpha(L) = 1 - \sum_{j=1}^{p-1} \alpha_j L^j$ . This data generating process can be nested in the auxiliary model

$$\Delta y_t = \alpha_0 + \alpha_1 y_{t-1} + u_t, \quad u_t = \gamma(L) e_t, \quad e_t \sim \text{iid } N(0, 1). \tag{29.60}$$

We will now determine the limiting distribution of the OLS estimate  $\hat{\alpha}_1$  and corresponding  $t$ -value  $\hat{t}_1$  of the parameter  $\alpha_1$  in the regression (29.60), derived under the assumption that the  $u_t$ s are independent, while in reality (29.59) holds.

Similarly to (29.41) we can write  $\Delta y_t = \gamma(1)e_t + v_t - v_{t-1}$ , where  $v_t = [(\gamma(L) - \gamma(1))/(1-L)]e_t$  is a stationary process. The latter follows from the fact that by construction the lag polynomial  $\gamma(L) - \gamma(1)$  has a unit root, and therefore contains a factor  $1 - L$ . Next, redefining  $W_n(x)$  as

$$W_n(x) = (1/\sqrt{n}) \sum_{t=1}^{[nx]} e_t \quad \text{if } x \in [n^{-1}, 1], \quad W_n(x) = 0 \quad \text{if } x \in [0, n^{-1}), \tag{29.61}$$

it follows similarly to (29.42) that

$$y_t/\sqrt{n} = y_0/\sqrt{n} + v_t/\sqrt{n} - v_0/\sqrt{n} + \gamma(1)W_n(t/n), \quad (29.62)$$

hence

$$y_n/\sqrt{n} = \gamma(1)W_n(1) + O_p(1/\sqrt{n}), \quad (29.63)$$

and similarly to (29.43) and (29.44) that

$$\bar{y}_{-1}/\sqrt{n} = \frac{1}{n} \sum_{t=1}^n y_{t-1}/\sqrt{n} = \gamma(1) \int W_n(x)dx + o_p(1), \quad (29.64)$$

and

$$\frac{1}{n^2} \sum_{t=1}^n y_{t-1}^2 = \gamma(1)^2 \int W_n(x)^2 dx + o_p(1). \quad (29.65)$$

Moreover, similarly to (29.6) we have

$$\begin{aligned} \frac{1}{n} \sum_{t=1}^n (\Delta y_t) y_{t-1} &= \frac{1}{2} \left( y_n^2/n - y_0^2/n - \frac{1}{n} \sum_{t=1}^n (\Delta y_t)^2 \right) \\ &= \frac{1}{2} \left( \gamma(1)^2 W_n(1)^2 - \frac{1}{n} \sum_{t=1}^n (\gamma(L)e_t)^2 \right) + o_p(1) \\ &= \gamma(1)^2 \frac{1}{2} (W_n(1) - \lambda) + o_p(1), \end{aligned} \quad (29.66)$$

where

$$\lambda = \frac{E(\gamma(L)e_t)^2}{\gamma(1)^2} = \frac{\sum_{j=0}^{\infty} \gamma_j^2}{\left( \sum_{j=0}^{\infty} \gamma_j \right)^2}. \quad (29.67)$$

Therefore, (29.33) now becomes:

$$\begin{aligned} n\hat{\alpha}_1 &= \frac{(1/2)(W_n(1)^2 - \lambda) - W_n(1) \int W_n(x)dx}{\int W_n(x)^2 dx - (\int W_n(x)dx)^2} + o_p(1) \rightarrow \\ &\rho_1 + \frac{0.5(1 - \lambda)}{\int W(x)^2 dx - (\int W(x)dx)^2} \end{aligned} \quad (29.68)$$

in distribution, and (29.34) becomes:

$$\begin{aligned}\hat{\tau}_1 &= \frac{(1/2)(W_n(1)^2 - \lambda) - W_n(1)\int W_n(x)dx}{\sqrt{\int W_n(x)^2 dx - (\int W_n(x)dx)^2}} + o_p(1) \rightarrow \\ \tau_1 &+ \frac{0.5(1 - \lambda)}{\sqrt{\int W(x)^2 dx - (\int W(x)dx)^2}}\end{aligned}\quad (29.69)$$

in distribution. These results carry straightforwardly over to the case where the actual data generating process is an ARIMA process  $\alpha(L)\Delta y_t = \beta(L)e_t$ , simply by redefining  $\gamma(L) = \beta(L)/\alpha(L)$ .

The parameter  $\gamma(1)^2$  is known as the long-run variance of  $u_t = \gamma(L)e_t$ :

$$\sigma_L^2 = \lim_{n \rightarrow \infty} \text{var} \left[ (1/\sqrt{n}) \sum_{t=1}^n u_t \right] = \gamma(1)^2 \quad (29.70)$$

which in general is different from the variance of  $u_t$  itself:

$$\therefore \sigma_u^2 = \text{var}(u_t) = E(u_t^2) = E \left( \sum_{j=0}^{\infty} \gamma_j e_{t-j} \right)^2 = \sum_{j=0}^{\infty} \gamma_j^2. \quad (29.71)$$

If we would know  $\sigma_L^2$  and  $\sigma_u^2$ , and thus  $\lambda = \sigma_u^2/\sigma_L^2$ , then it follows from (29.64), (29.65), and Lemma 1, that

$$\frac{\sigma_L^2 - \sigma_u^2}{(1/n^2) \sum_{t=1}^n (y_{t-1} - \bar{y}_{-1})^2} \rightarrow \frac{1 - \lambda}{\int W(x)^2 dx - (\int W(x)dx)^2} \text{ in distribution.} \quad (29.72)$$

It is an easy exercise to verify that this result also holds if we replace  $y_{t-1}$  by  $y_t$  and  $\bar{y}_{-1}$  by  $\bar{y} = (1/n) \sum_{t=1}^n y_t$ . Therefore, it follows from (29.68) and (29.72) that:

**Theorem 3.** (Phillips–Perron test 1) Under the unit root hypothesis, and given consistent estimators  $\hat{\sigma}_L^2$  and  $\hat{\sigma}_u^2$  of  $\sigma_L^2$  and  $\sigma_u^2$ , respectively, we have

$$\hat{Z}_1 = n \left( \hat{\alpha}_1 - \frac{(\hat{\sigma}_L^2 - \hat{\sigma}_u^2)/2}{(1/n) \sum_{t=1}^n (y_t - \bar{y})^2} \right) \rightarrow \rho_1 \text{ in distribution.} \quad (29.73)$$

This correction of (29.68) has been proposed by Phillips and Perron (1988) for particular estimators  $\hat{\sigma}_L^2$  and  $\hat{\sigma}_u^2$ , following the approach of Phillips (1987) for the case where the intercept  $\alpha_0$  in (29.60) is assumed to be zero.

It is desirable to choose the estimators  $\hat{\sigma}_L^2$  and  $\hat{\sigma}_u^2$  such that under the stationarity alternative,  $\text{plim}_{n \rightarrow \infty} \hat{Z}_1 = -\infty$ . We show now that this is the case if we choose

$$\hat{\sigma}_u^2 = \frac{1}{n} \sum_{t=1}^n \hat{u}_t^2, \quad \text{where } \hat{u}_t = \Delta y_t - \hat{\alpha}_0 - \hat{\alpha}_1 y_{t-1}, \quad (29.74)$$

and  $\hat{\sigma}_L^2$  such that  $\bar{\sigma}_L^2 = \text{plim}_{n \rightarrow \infty} \hat{\sigma}_L^2 \geq 0$  under the alternative of stationarity.

First, it is easy to verify that  $\hat{\sigma}_u^2$  is consistent under the null hypothesis, by verifying that (29.57) still holds. Under stationarity we have  $\text{plim}_{n \rightarrow \infty} \hat{\alpha}_1 = \text{cov}(y_t, y_{t-1}) / \text{var}(y_t) - 1 = \alpha_1^*$ , say,  $\text{plim}_{n \rightarrow \infty} \hat{\alpha}_0 = -\alpha_1^* E(y_t) = \alpha_0^*$ , say, and  $\text{plim}_{n \rightarrow \infty} \hat{\sigma}_u^2 = (1 - (\alpha_1^* + 1)^2) \text{var}(y_t) = \sigma_{\star}^2$ , say. Therefore,

$$\text{plim}_{n \rightarrow \infty} \hat{Z}_1/n = -0.5(\alpha_1^{*2} + \bar{\sigma}_L^2 / \text{var}(y_t)) < 0. \quad (29.75)$$

Phillips and Perron (1988) propose to estimate the long-run variance by the Newey-West (1987) estimator

$$\hat{\sigma}_L^2 = \hat{\sigma}_u^2 + 2 \sum_{i=1}^m [1 - i/(m+1)](1/n) \sum_{t=i+1}^n \hat{u}_t \hat{u}_{t-i}, \quad (29.76)$$

where  $\hat{u}_t$  is defined in (29.74), and  $m$  converges to infinity with  $n$  at rate  $o(n^{1/4})$ . Andrews (1991) has shown (and we will show it again along the lines in Bierens, 1994) that the rate  $o(n^{1/4})$  can be relaxed to  $o(n^{1/2})$ . The weights  $1 - j/(m+1)$  guarantee that this estimator is always positive. The reason for the latter is the following. Let  $u_t^* = u_t$  for  $t = 1, \dots, n$ , and  $u_t^* = 0$  for  $t < 1$  and  $t > n$ . Then,

$$\begin{aligned} \hat{\sigma}_L^{*2} &\equiv \frac{1}{n} \sum_{t=1}^{n+m} \left( \frac{1}{\sqrt{m+1}} \sum_{j=0}^m u_{t-j}^* \right)^2 = \frac{1}{m+1} \sum_{j=0}^m \frac{1}{n} \sum_{t=1}^{n+m} u_{t-j}^{*2} + 2 \frac{1}{m+1} \sum_{j=0}^{m-1} \sum_{i=1}^{m-j} \frac{1}{n} \sum_{t=1}^{n+m} u_{t-j}^* u_{t-j-i}^* \\ &= \frac{1}{m+1} \sum_{j=0}^m \frac{1}{n} \sum_{t=1-j}^{n+m-j} u_t^{*2} + 2 \frac{1}{m+1} \sum_{j=0}^{m-1} \sum_{i=1}^{m-j} \frac{1}{n} \sum_{t=1-j}^{n+m-j} u_t^* u_{t-i}^* \\ &= \frac{1}{n} \sum_{t=1}^n u_t^2 + 2 \frac{1}{m+1} \sum_{j=0}^{m-1} \sum_{i=1}^{m-j} \frac{1}{n} \sum_{t=i+1}^n u_t u_{t-i} \\ &= \frac{1}{n} \sum_{t=1}^n u_t^2 + 2 \frac{1}{m+1} \sum_{i=1}^m (m+1-i) \frac{1}{n} \sum_{t=i+1}^n u_t u_{t-i} \end{aligned} \quad (29.77)$$

is positive, and so is  $\hat{\sigma}_L^2$ . Next, observe from (29.62) and (29.74) that

$$\hat{u}_t = u_t - \sqrt{n} \hat{\alpha}_1 \gamma(1) W_n(t/n) - \hat{\alpha}_1 v_t + \hat{\alpha}_1 (v_0 - y_0) - \hat{\alpha}_0. \quad (29.78)$$

Since

$$E |(1/n) \sum_{t=1+i}^n u_t W_n((t-i)/n)| \leq \sqrt{(1/n) \sum_{t=1+i}^n E(u_t^2)} \sqrt{(1/n) \sum_{t=1+i}^n E(W_n((t-i)/n)^2)} = O(1),$$

it follows that  $(1/n) \sum_{t=1+i}^n u_t W_n((t-i)/n) = O_p(1)$ . Similarly,  $(1/n) \sum_{t=1+i}^n u_{t-i} W_n(t/n) = O_p(1)$ . Moreover,  $\hat{\alpha}_1 = O_p(1/n)$ , and similarly, it can be shown that  $\hat{\alpha}_0 = O_p(1/\sqrt{n})$ . Therefore, it follows from (29.77) and (29.78) that

$$\hat{\sigma}_L^2 - \hat{\sigma}_L^{*2} = O_p(1/n) + O_p \left( \sum_{i=1}^m [1 - i/(m+1)]/\sqrt{n} \right) = O_p(1/n) + O_p(m/\sqrt{n}). \quad (29.79)$$

A similar result holds under the stationarity hypothesis. Moreover, substituting  $u_t = \sigma_L^2 e_t + v_t - v_{t-1}$ , and denoting  $e_t^* = e_t$ ,  $v_t^* = v_t$  for  $t = 1, \dots, n$ ,  $v_t^* = e_t^* = 0$  for  $t < 1$  and  $t > n$ , it is easy to verify that under the unit root hypothesis,

$$\begin{aligned}\hat{\sigma}_L^{*2} &= \frac{1}{n} \sum_{t=1}^{n+m} \left( \sigma_L \frac{1}{\sqrt{m+1}} \sum_{j=0}^m e_{t-j}^* + \frac{v_t^* - v_{t-m}^*}{\sqrt{m+1}} \right)^2 \\ &= \sigma_L^2 \frac{1}{n} \sum_{t=1}^{n+m} \left( \frac{1}{\sqrt{m+1}} \sum_{j=0}^m e_{t-j}^* \right)^2 + 2\sigma_L \frac{1}{n} \sum_{t=1}^{n+m} \left( \frac{1}{\sqrt{m+1}} \sum_{j=0}^m e_{t-j}^* \right) \left( \frac{v_t^* - v_{t-m}^*}{\sqrt{m+1}} \right) \\ &\quad + \frac{1}{n} \sum_{t=1}^{n+m} \left( \frac{v_t^* - v_{t-m}^*}{\sqrt{m+1}} \right)^2 = \sigma_L^2 + O_p(\sqrt{m/n}) + O_p(1/\sqrt{m}) + O_p(1/m). \quad (29.80)\end{aligned}$$

A similar result holds under the stationarity hypothesis. Thus:

**Theorem 4.** Let  $m$  increase with  $n$  to infinity at rate  $o(n^{1/2})$ . Then under both the unit root and stationarity hypothesis,  $\text{plim}_{n \rightarrow \infty} (\hat{\sigma}_L^{*2} - \hat{\sigma}_L^{*2}) = 0$ . Moreover, under the unit root hypothesis,  $\text{plim}_{n \rightarrow \infty} \hat{\sigma}_L^{*2} = \sigma_L^2$ , and under the stationarity hypothesis,  $\text{plim}_{n \rightarrow \infty} \hat{\sigma}_L^{*2} > 0$ . Consequently, under stationarity, the Phillips–Perron test satisfies  $\text{plim}_{n \rightarrow \infty} \hat{Z}_1/n < 0$ .

Finally, note that the advantage of the PP test is that there is no need to specify the ARIMA process under the null hypothesis. It is in essence a nonparametric test. Of course, we still have to specify the Newey–West truncation lag  $m$  as a function of  $n$ , but as long as  $m = o(\sqrt{n})$ , this specification is asymptotically not critical.

## 6 UNIT ROOT WITH DRIFT VS. TREND STATIONARITY

Most macroeconomic time series in (log) levels have an upwards sloping pattern. Therefore, if they are (covariance) stationary, then they are stationary around a deterministic trend. If we would conduct the ADF and PP tests in Sections 4 and 5 to a linear trend stationary process, we will likely accept the unit root hypothesis, due to the following. Suppose we conduct the ADF test under the hypothesis  $p = 1$  to the trend stationary process  $y_t = \beta_0 + \beta_1 t + u_t$ , where the  $u_t$ s are iid  $N(0, \sigma^2)$ . It is a standard exercise to verify that then  $\text{plim}_{n \rightarrow \infty} n\hat{\alpha}_1 = 0$ , hence the ADF and PP tests in sections 4 and 5 have no power against linear trend stationarity!

Therefore, if one wishes to test the unit root hypothesis against linear trend stationarity, then a trend term should be included in the auxiliary regressions (29.39) in the ADF case, and in (29.60) in the PP case: Thus the ADF regression (29.39) now becomes

$$\Delta y_t = \alpha_0 + \sum_{j=1}^{p-1} \alpha_j \Delta y_{t-1} + \alpha_p y_{t-1} + \alpha_{p+1} t + u_t, \quad u_t \sim \text{iid } N(0, \sigma_2) \quad (29.81)$$

where the null hypothesis of a unit root with drift corresponds to the hypothesis  $\alpha_p = \alpha_{p+1} = 0$ , and the PP regression becomes:

$$\Delta y_t = \alpha_0 + \alpha_1 y_{t-1} + \alpha_2 t + u_t, \quad u_t = \gamma(L)e_t, \quad e_t \sim \text{iid } N(0, 1). \quad (29.82)$$

The asymptotic null distributions of the ADF and PP tests for the case with drift are quite similar to the ADF test without an intercept. The difference is that the Wiener process  $W(x)$  is replaced by the de-trended Wiener process:

$$W^{**}(x) = W(x) - 4 \int W(z)dz + 6 \int zW(z)dz + 6 \left( \int W(z)dz - 2 \int zW(z)dz \right)x$$

After some tedious but not too difficult calculations it can be shown that effectively the statistics  $n\hat{\alpha}_p/(1 - \sum_{j=1}^p \alpha_j)$  and  $\hat{t}_p$  are asymptotically equivalent to the Dickey–Fuller tests statistics  $\hat{\rho}_0$  and  $\hat{t}_0$ , respectively, applied to de-trended time series.

**Theorem 5.** Let  $y_t$  be generated by (29.81), and let  $\hat{\alpha}_p$  and  $\hat{t}_p$  be the OLS estimator and corresponding  $t$ -value of  $\alpha_p$ . Under the unit root with drift hypothesis, i.e.  $\alpha_p = \alpha_{p+1} = 0$ , we have  $n\hat{\alpha}_p \rightarrow (1 - \sum_{j=1}^{p-1} \alpha_j)\rho_2$  and  $\hat{t}_p \rightarrow \tau_2$  in distribution, where

$$\rho_2 = \frac{1}{2} \left( \frac{W^{**}(1) - 1}{\int W^{**}(x)^2 dx} \right), \quad \tau_2 = \frac{1}{2} \left( \frac{W^{**}(1) - 1}{\sqrt{\int W^{**}(x)^2 dx}} \right).$$

Under the trend stationarity hypothesis,  $\text{plim}_{n \rightarrow \infty} \hat{\alpha}_p = \alpha_p < 0$ , hence  $\text{plim}_{n \rightarrow \infty} \hat{t}_p / \sqrt{n} < 0$ .

The densities of  $\rho_2$  and  $\tau_2$  (the latter compared with the standard normal density), are displayed in Figures 29.5 and 29.6, respectively. Again, these densities are farther to the left, and heavier left-tailed, than the corresponding densities displayed in Figures 29.1–29.4. The asymptotic 5 percent and 10 percent critical values of the Dickey–Fuller  $t$ -test are:

$$P(\tau_2 < -3.41) = 0.05, \quad P(\tau_2 < -3.13) = 0.10$$

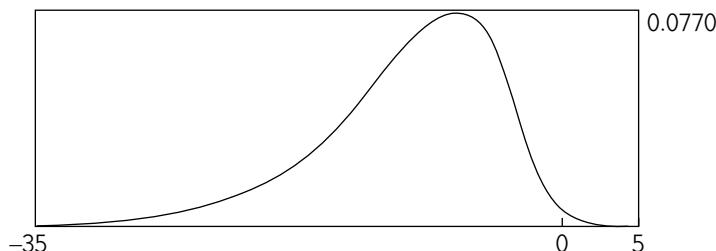
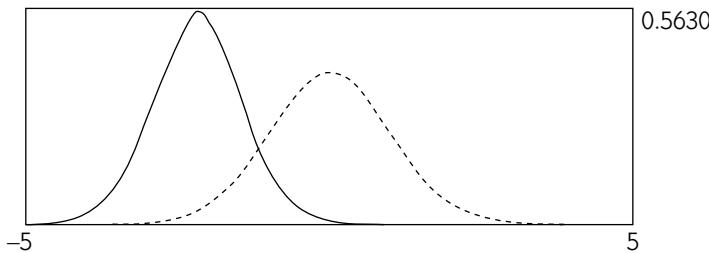


Figure 29.5 Density of  $\rho_2$



**Figure 29.6** Density of  $\tau_2$  compared with the standard normal density (dashed curve)

Moreover, comparing (29.26) with

$$P(\tau_2 \leq -1.64) \approx 0.77, \quad P(\tau_2 \leq -1.28) \approx 0.89,$$

we see that the standard normal tests at the 5 percent and 10 percent significance level would reject the correct unit root with drift hypothesis with probabilities of about 0.77 and 0.89, respectively!

A similar result as in Theorem 5 can be derived for the PP test, on the basis of the OLS estimator of  $\alpha_1$  in, and the residuals  $\hat{u}_t$  of, the auxiliary regression (29.82):

**Theorem 6.** (Phillips–Perron test 2) Let  $\hat{r}_t$  be the residuals of the OLS regression of  $y_t$  on  $t$  and a constant, and let  $\hat{\sigma}_u^2$  and  $\hat{\sigma}_L^2$  be as before, with the  $\hat{u}_t$ s the OLS residuals of the auxiliary regression (29.82). Under the unit root with drift hypothesis,

$$\hat{Z}_2 = n \left( \hat{\alpha}_1 - \frac{(\hat{\sigma}_L^2 - \hat{\sigma}_u^2)/2}{(1/n)\sum_{t=1}^n \hat{r}_t^2} \right) \rightarrow \rho_2 \text{ in distribution,} \quad (29.83)$$

whereas under trend stationarity  $\text{plim}_{n \rightarrow \infty} \hat{Z}_2/n < 0$ .

## 7 CONCLUDING REMARKS

In the discussion of the ADF test we have assumed that the lag length  $p$  of the auxiliary regression (29.81) is fixed. It should be noted that we may choose  $p$  as a function of the length  $n$  of the time series involved, similarly to the truncation width of the Newey–West estimator of the long-run variance in the Phillips–Perron test. See Said and Dickey (1984).

We have seen that the ADF and Phillips–Perron tests for a unit root against stationarity around a constant have almost no power if the correct alternative is linear trend stationarity. However, the same may apply to the tests discussed in Section 6 if the alternative is trend stationarity with a broken trend. See Perron (1988, 1989, 1990), Perron and Vogelsang (1992), and Zivot and Andrews (1992), among others.

All the tests discussed so far have the unit root as the null hypothesis, and (trend) stationarity as the alternative. However, it is also possible to test the other

way around. See Bierens and Guo (1993), and Kwiatkowski *et al.* (1992). The latter test is known as the KPSS test.

Finally, note that the ADF and Phillips–Perron tests can easily be conducted by various econometric software packages, for example TSP, EViews, RATS, and *EasyReg*.<sup>11</sup>

## Notes

- \* The useful comments of three referees are gratefully acknowledged.
- 1 See Chapter 26 on spurious regression in this volume. This phenomenon can easily be demonstrated by using my free software package *EasyReg*, which is downloadable from website <http://econ.la.psu.edu/~hbierens/EASYREG.HTM> (Click on “Tools”, and then on “Teaching tools”).
- 2 See Chapter 30 on cointegration in this volume.
- 3 The reason for changing the subscript of  $\alpha$  from 1 in (29.1) to 0 is to indicate the number of other parameters at the right-hand side of the equation. See also (29.39).
- 4 Recall that the notation  $o_p(a_n)$ , with  $a_n$  a deterministic sequence, stands for a sequence of random variables or vectors  $x_n$ , say, such that  $\text{plim}_{n \rightarrow \infty} x_n/a_n = 0$ , and that the notation  $O_p(a_n)$  stands for a sequence of random variables or vectors  $x_n$  such that  $x_n/a_n$  is stochastically bounded:  $\forall \varepsilon \in (0, 1) \exists M \in (0, \infty) : \sup_{n \geq 1} P(|x_n/a_n| > M) < \varepsilon$ . Also, recall that convergence in distribution implies stochastic boundedness.
- 5 The Borel sets in  $\mathbb{R}$  are the members of the smallest  $\sigma$ -algebra containing the collection  $\mathcal{C}$ , say, of all half-open intervals  $(-\infty, x]$ ,  $x \in \mathbb{R}$ . Equivalently, we may also define the Borel sets as the members of the smallest  $\sigma$ -algebra containing the collection of open subsets of  $\mathbb{R}$ . A collection  $\mathcal{F}$  of subsets of a set  $\Omega$  is called a  $\sigma$ -algebra if the following three conditions hold:  $\Omega \in \mathcal{F}$ ;  $A \in \mathcal{F}$  implies that its complement also belongs to  $\mathcal{F}$ ;  $\Omega \setminus A \in \mathcal{F}$  (hence, the empty set  $\emptyset$  belongs to  $\mathcal{F}$ );  $A_n \in \mathcal{F}$ ,  $n = 1, 2, 3, \dots$ , implies  $\bigcup_{n=1}^{\infty} A_n \in \mathcal{F}$ . The smallest  $\sigma$ -algebra containing a collection  $\mathcal{C}$  of sets is the intersection of all  $\sigma$ -algebras containing the collection  $\mathcal{C}$ .
- 6 Under the assumption that  $e_t$  is iid  $N(0, 1)$ ,

$$P(\max_{1 \leq t \leq n} |e_t| \leq \varepsilon \sqrt{n}) = \left( 1 - 2 \int_{\varepsilon \sqrt{n}}^{\infty} \frac{\exp(-x^2/2)}{\sqrt{2\pi}} dx \right)^n \rightarrow 1$$

for arbitrary  $\varepsilon > 0$ .

- 7 This density is actually a kernel estimate of the density of  $\hat{\rho}_0$  on the basis of 10,000 replications of a Gaussian random walk  $y_t = y_{t-1} + e_t$ ,  $t = 0, 1, \dots, 1,000$ ,  $y_0 = 0$  for  $t < 0$ . The kernel involved is the standard normal density, and the bandwidth  $h = c.s10,000^{-1/5}$ , where  $s$  is the sample standard error, and  $c = 1$ . The scale factor  $c$  has been chosen by experimenting with various values. The value  $c = 1$  is about the smallest one for which the kernel estimate remains a smooth curve; for smaller values of  $c$  the kernel estimate becomes wobbly. The densities of  $\rho_1$ ,  $\tau_1$ ,  $\rho_2$ , and  $\tau_2$  in Figures 29.2–29.6 have been constructed in the same way, with  $c = 1$ .
- 8 To see this, write  $1 - \sum_{j=1}^p \beta_j L^j = \prod_{j=1}^p (1 - \rho_j L)$ , so that  $1 - \sum_{j=1}^p \beta_j = \prod_{j=1}^p (1 - \rho_j)$ , where the  $1/\rho_j$ s are the roots of the lag polynomial involved. If root  $1/\rho_j$  is real valued, then the stationarity condition implies  $-1 < \rho_j < 1$ , so that  $1 - \rho_j > 0$ . If some roots are complex-valued, then these roots come in complex-conjugate pairs, say  $1/\rho_1 = a + i.b$  and  $1/\rho_2 = a - i.b$ , hence  $(1 - \rho_1)(1 - \rho_2) = (1/\rho_1 - 1)(1/\rho_2 - 1)\rho_1\rho_2 = ((a - 1)^2 + b^2)/(a^2 + b^2) > 0$ .
- 9 In the sequel we shall suppress the statement “without drift.” A unit root process is from now on by default a unit root without drift process, except if otherwise indicated.

- 10 For example, let  $p = 2$  in (29.37) and (29.39). Then  $\alpha_1 = -\beta_1$ , hence if  $\beta_1 < -1$  then  $1 - \alpha_1 < 0$ . In order to show that  $\beta_1 < -1$  can be compatible with stationarity, assume that  $\beta_1^2 = 4\beta_2$ , so that the lag polynomial  $1 - \beta_1 L - \beta_2 L^2$  has two common roots  $-2/|\beta_1|$ . Then the AR(2) process involved is stationary for  $-2 < \beta_1 < -1$ .
- 11 The most important difference with other econometric software packages is that *EasyReg* is free. See footnote 1. *EasyReg* also contains my own unit root tests, Bierens (1993, 1997), Bierens and Guo (1993), and the KPSS test.

## References

- Andrews, D.W.K. (1991). Heteroskedasticity and autocorrelation consistent covariance matrix estimators. *Econometrica* 59, 817–58.
- Bierens, H.J. (1993). Higher order sample autocorrelations and the unit root hypothesis. *Journal of Econometrics* 57, 137–60.
- Bierens, H.J. (1997). Testing the unit root hypothesis against nonlinear trend stationarity, with an application to the price level and interest rate in the U.S. *Journal of Econometrics* 81, 29–64.
- Bierens, H.J. (1994). *Topics in Advanced Econometrics: Estimation, Testing and Specification of Cross-Section and Time Series Models*. Cambridge: Cambridge University Press.
- Bierens, H.J., and S. Guo (1993). Testing stationarity and trend stationarity against the unit root hypothesis. *Econometric Reviews* 12, 1–32.
- Billingsley, P. (1968). *Convergence of Probability Measures*. New York: John Wiley.
- Dickey, D.A., and W.A. Fuller (1979). Distribution of the estimators for autoregressive times series with a unit root. *Journal of the American Statistical Association* 74, 427–31.
- Dickey, D.A., and W.A. Fuller (1981). Likelihood ratio statistics for autoregressive time series with a unit root. *Econometrica* 49, 1057–72.
- Fuller, W.A. (1996). *Introduction to Statistical Time Series*. New York: John Wiley.
- Green, W. (1997). *Econometric Analysis*. Upper Saddle River, NJ: Prentice Hall.
- Hogg, R.V., and A.T. Craig (1978). *Introduction to Mathematical Statistics*. London: Macmillan.
- Kwiatkowski, D., P.C.B. Phillips, P. Schmidt, and Y. Shin (1992). Testing the null of stationarity against the alternative of a unit root. *Journal of Econometrics* 54, 159–78.
- Newey, W.K., and K.D. West (1987). A simple positive definite heteroskedasticity and autocorrelation consistent covariance matrix. *Econometrica* 55, 703–8.
- Perron, P. (1988). Trends and random walks in macroeconomic time series: further evidence from a new approach. *Journal of Economic Dynamics and Control* 12, 297–332.
- Perron, P. (1989). The great crash, the oil price shock and the unit root hypothesis. *Econometrica* 57, 1361–402.
- Perron, P. (1990). Testing the unit root in a time series with a changing mean. *Journal of Business and Economic Statistics* 8, 153–62.
- Perron, P., and T.J. Vogelsang (1992). Nonstationarity and level shifts with an application to purchasing power parity. *Journal of Business and Economic Statistics* 10, 301–20.
- Phillips, P.C.B. (1987). Time series regression with a unit root. *Econometrica* 55, 277–301.
- Phillips, P.C.B., and P. Perron (1988). Testing for a unit root in time series regression. *Biometrika* 75, 335–46.
- Said, S.E., and D.A. Dickey (1984). Testing for unit roots in autoregressive-moving average of unknown order. *Biometrika* 71, 599–607.
- Zivot, E., and D.W.K. Andrews (1992). Further evidence on the great crash, the oil price shock, and the unit root hypothesis. *Journal of Business and Economic Statistics* 10, 251–70.

CHAPTER THIRTY

# Cointegration

*Juan J. Dolado, Jesús Gonzalo,  
and Francesc Marmol*

## 1 INTRODUCTION

A substantial part of economic theory generally deals with long-run equilibrium relationships generated by market forces and behavioral rules. Correspondingly, most empirical econometric studies entailing time series can be interpreted as attempts to evaluate such relationships in a dynamic framework.

At one time, conventional wisdom was that in order to apply standard inference procedures in such studies, the variables in the system needed to be stationary since the vast majority of econometric theory is built upon the assumption of stationarity. Consequently, for many years econometricians proceeded as if stationarity could be achieved by simply removing deterministic components (e.g. drifts and trends) from the data. However, stationary series should at least have constant unconditional mean and variance over time, a condition which hardly appears to be satisfied in economics, even after removing those deterministic terms.

Those problems were somehow ignored in applied work until important papers by Granger and Newbold (1974) and Nelson and Plosser (1982) alerted many to the econometric implications of nonstationarity and the dangers of running *nonsense* or *spurious* regressions; see, e.g. Chapter 26 by Granger in this volume for further details. In particular, most of the attention focused on the implications of dealing with integrated variables which are a specific class of nonstationary variables with important economic and statistical properties. These are derived from the presence of unit roots which give rise to stochastic trends, as opposed to pure deterministic trends, with innovations to an integrated process being permanent rather than transitory.

The presence of, at least, a unit root in economic time series is implied in many economic models. Among them, there are those based on the rational use of available information or the existence of very high adjustment costs in some

markets. Interesting examples include future contracts, stock prices, yield curves, exchange rates, money velocity, hysteresis theories of unemployment, and, perhaps the most popular, the implications of the permanent income hypothesis for real consumption under rational expectations.

Statisticians, in turn, following the influential approach by Box and Jenkins (1970), had advocated transforming integrated time series into stationary ones by successive differencing of the series before modelization. Therefore, from their viewpoint, removing unit roots through differencing ought to be a prerequisite for regression analysis. However, some authors, notably Sargan (1964), Hendry and Mizon (1978) and Davidson *et al.* (1978), *inter alia*, started to criticize on a number of grounds the specification of dynamic models in terms of differenced variables only, especially because of the difficulties in inferring the long-run equilibrium from the estimated model. After all, if deviations from that equilibrium relationship affect future changes in a set of variables, omitting the former, i.e. estimating a differenced model, should entail a misspecification error. However, for some time it remained to be well understood how both variables in differences and levels could coexist in regression models.

Granger (1981), resting upon the previous ideas, solved the puzzle by pointing out that a vector of variables, all of which achieve stationarity after differencing, could have linear combinations which are stationary in levels. Later, Granger (1986) and Engle and Granger (1987) were the first to formalize the idea of integrated variables sharing an equilibrium relation which turned out to be either stationary or have a lower degree of integration than the original series. They denoted this property by *cointegration*, signifying comovements among trending variables which could be exploited to test for the existence of equilibrium relationships within a fully dynamic specification framework. Notice that the notion of "equilibrium" used here is that of a state to which a dynamic system tends to converge over time after any of the variables in the system is perturbed by a shock. In economics, the strength of attraction to such a state depends on the actions of a market or on government intervention. In this sense, the basic concept of cointegration applies in a variety of economic models including the relationships between capital and output, real wages and labor productivity, nominal exchange rates and relative prices, consumption and disposable income, long- and short-term interest rates, money velocity and interest rates, price of shares and dividends, production and sales, etc. In particular, Campbell and Shiller (1987) have pointed out that a pair of integrated variables that are related through a present value model, as it is often the case in macroeconomics and finance, must be cointegrated.

In view of the strength of these ideas, a burgeoning literature on cointegration has developed over the last decade. In this chapter we will explore the basic conceptual issues and discuss related econometric techniques, with the aim of offering an introductory coverage of the main developments in this new field of research. Section 2 provides some preliminaries on the implications of cointegration and the basic estimation and testing procedures in a single equation framework, when variables have a single unit root. In Section 3, we extend the previous techniques to more general multivariate setups, introducing those system-based

approaches to cointegration which are now in common use. Section 4, in turn, presents some interesting developments on which the recent research on cointegration has been focusing. Finally, Section 5 draws some concluding remarks.

Nowadays, the interested reader, who wants to deepen beyond the introductory level offered here, could find a number of textbooks (e.g. Banerjee *et al.*, 1993; Johansen, 1995; Hatanaka, 1996; Maddala and Kim, 1998) and surveys (e.g. Engle and Granger, 1991; Watson, 1994) on cointegration where more general treatments of the relevant issues covered in this chapter are presented. Likewise, there are now many software packages that support the techniques discussed here (e.g. Gauss-COINT, E-VIEWS and PC-FIML).

## 2 PRELIMINARIES: UNIT ROOTS AND COINTEGRATION

### 2.1 Some basic concepts

A well known result in time series analysis is Wold's (1938) decomposition theorem which states that a stationary time series process, after removal of any deterministic components, has an infinite moving average (MA) representation which, under some technical conditions (absolute summability of the MA coefficients), can be represented by a finite autoregressive moving average (ARMA) process.

However, as mentioned in the introduction, many time series need to be appropriately differenced in order to achieve stationarity. From this comes the definition of integration: a time series is said to be integrated of order  $d$ , in short  $I(d)$ , if it has a stationary, invertible, non-deterministic ARMA representation after differencing  $d$  times. A white noise series and a stable first-order autoregressive AR(1) process are well known examples of  $I(0)$  series, a random walk process is an example of an  $I(1)$  series, while accumulating a random walk gives rise to an  $I(2)$  series, etc.

Consider now two time series  $y_{1t}$  and  $y_{2t}$  which are both  $I(d)$  (i.e. they have compatible long-run properties). In general, any linear combination of  $y_{1t}$  and  $y_{2t}$  will be also  $I(d)$ . However, if there exists a vector  $(1, -\beta)'$ , such that the linear combination

$$z_t = y_{1t} - \alpha - \beta y_{2t} \quad (30.1)$$

is  $I(d - b)$ ,  $d \geq b > 0$ , then, following Engle and Granger (1987),  $y_{1t}$  and  $y_{2t}$  are defined as cointegrated of order  $(d, b)$ , denoted  $y_t = (y_{1t}, y_{2t})' \sim CI(d, b)$ , with  $(1, -\beta)'$  called the cointegrating vector.

Several features in (30.1) are noteworthy. First, as defined above, cointegration refers to a linear combination of nonstationary variables. Although theoretically it is possible that nonlinear relationships may exist among a set of integrated variables, the econometric practice about this more general type of cointegration is less developed (see more on this in Section 4). Second, note that the cointegrating vector is not uniquely defined, since for any nonzero value of  $\lambda$ ,  $(\lambda, -\lambda\beta)'$  is also a cointegrating vector. Thus, a normalization rule needs to be used; for example,

$\lambda = 1$  has been chosen in (30.1). Third, all variables must be integrated of the same order to be candidates to form a cointegrating relationship. Notwithstanding, there are extensions of the concept of cointegration, called *multicointegration*, when the number of variables considered is larger than two and where the possibility of having variables with different order of integration can be addressed (see, e.g. Granger and Lee, 1989). For example, in a trivariate system, we may have that  $y_{1t}$  and  $y_{2t}$  are I(2) and  $y_{3t}$  is I(1); if  $y_{1t}$  and  $y_{2t}$  are CI(2, 1), it is possible that the corresponding combination of  $y_{1t}$  and  $y_{2t}$  which achieves that property be itself cointegrated with  $y_{3t}$  giving rise to an I(0) linear combination among the three variables. Fourth, and most important, most of the cointegration literature focuses on the case where variables contain a single unit root, since few economic variables prove in practice to be integrated of higher order. If variables have a strong seasonal component, however, there may be unit roots at the seasonal frequencies, a case that we will briefly consider in Section 4; see Chapter 30 by Ghysels, Osborn, and Rodrigues in this volume for further details. Hence, the remainder of this chapter will mainly focus on the case of CI(1, 1) variables, so that  $z_t$  in (30.1) is I(0) and the concept of cointegration mimics the existence of a long-run equilibrium to which the system converges over time. If, e.g., economic theory suggests the following long-run relationship between  $y_{1t}$  and  $y_{2t}$ ,

$$y_{1t} = \alpha + \beta y_{2t}, \quad (30.2)$$

then  $z_t$  can be interpreted as the equilibrium error (i.e. the distance that the system is away from the equilibrium at any point in time). Note that a constant term has been included in (30.1) in order to allow for the possibility that  $z_t$  may have nonzero mean. For example, a standard theory of spatial competition argues that arbitrage will prevent prices of similar products in different locations from moving too far apart even if the prices are nonstationary. However, if there are fixed transportation costs from one location to another, a constant term needs to be included in (30.1).

At this stage, it is important to point out that a useful way to understand cointegrating relationships is through the observation that CI(1, 1) variables must share a set of stochastic trends. Using the example in (30.1), since  $y_{1t}$  and  $y_{2t}$  are I(1) variables, they can be decomposed into an I(1) component (say, a random walk) plus an irregular I(0) component (not necessarily white noise). Denoting the first components by  $\mu_{it}$  and the second components by  $u_{it}$ ,  $i = 1, 2$ , we can write

$$y_{1t} = \mu_{1t} + u_{1t} \quad (30.3)$$

$$y_{2t} = \mu_{2t} + u_{2t}. \quad (30.3')$$

Since the sum of an I(1) process and an I(0) process is always I(1), the previous representation must characterize the individual stochastic properties of  $y_{1t}$  and  $y_{2t}$ . However, if  $y_{1t} - \beta y_{2t}$  is I(0), it must be that  $\mu_{1t} = \beta \mu_{2t}$ , annihilating the I(1) component in the cointegrating relationship. In other words, if  $y_{1t}$  and  $y_{2t}$  are

CI(1, 1) variables, they must share (up to a scalar) the same stochastic trend, say  $\mu_t$ , denoted as *common trend*, so that  $\mu_{1t} = \mu_t$  and  $\mu_{2t} = \beta\mu_t$ . As before, notice that if  $\mu_t$  is a common trend for  $y_{1t}$  and  $y_{2t}$ ,  $\lambda\mu_t$  will also be a common trend implying that a normalization rule is needed for identification. Generalizing the previous argument to a vector of cointegration and common trends, then it can be proved that if there are  $n - r$  common trends among the  $n$  variables, there must be  $r$  cointegrating relationships. Note that  $0 < r < n$ , since  $r = 0$  implies that each series in the system is governed by a different stochastic trend and that  $r = n$  implies that the series are I(0) instead of I(1). These properties constitute the core of two important dual approaches toward testing for cointegration, namely, one that tests directly for the number of cointegrating vectors ( $r$ ) and another which tests for the number of common trends ( $n - r$ ). However, before explaining those approaches in more detail (see Section 3), we now turn to another useful representation of CI(1, 1) systems which has proved very popular in practice.

Engle and Granger (1987) have shown that if  $y_{1t}$  and  $y_{2t}$  are cointegrated CI(1, 1), then there must exist a so-called *vector error correction model* (VECM) representation of the dynamic system governing the joint behavior of  $y_{1t}$  and  $y_{2t}$  over time, of the following form

$$\Delta y_{1t} = \theta_{10} + \theta_{11}z_{t-1} + \sum_{i=1}^{p_1} \theta_{12,i} \Delta y_{1,t-i} + \sum_{i=1}^{p_2} \theta_{13,i} \Delta y_{2,t-i} + \varepsilon_{1t}, \quad (30.4)$$

$$\Delta y_{2t} = \theta_{20} + \theta_{21}z_{t-1} + \sum_{i=1}^{p_3} \theta_{22,i} \Delta y_{1,t-i} + \sum_{i=1}^{p_4} \theta_{23,i} \Delta y_{2,t-i} + \varepsilon_{2t}, \quad (30.4')$$

where  $\Delta$  denotes the first-order time difference (i.e.  $\Delta y_t = y_t - y_{t-1}$ ) and where the lag lengths  $p_i$ ,  $i = 1, \dots, 4$  are such that the innovations  $\varepsilon_t = (\varepsilon_{1t}, \varepsilon_{2t})'$  are iid  $(0, \Sigma)$ . Furthermore, they proved the converse result that a VECM generates cointegrated CI(1, 1) series as long as the coefficients on  $z_{t-1}$  (the so-called *loading or speed of adjustment parameters*) are not simultaneously equal to zero.

Note that the term  $z_{t-1}$  in equations (30.4) and (30.4') represents the extent of the disequilibrium levels of  $y_1$  and  $y_2$  in the previous period. Thus, the VECM representation states that changes in one variable not only depends on changes of the other variables and its own past changes, but also on the extent of the disequilibrium between the levels of  $y_1$  and  $y_2$ . For example, if  $\beta = 1$  in (30.1), as many theories predict when  $y_{1t}$  and  $y_{2t}$  are taken in logarithmic form, then if  $y_1$  is larger than  $y_2$  in the past ( $z_{t-1} > 0$ ), then  $\theta_{11} < 0$  and  $\theta_{21} > 0$  will imply that, everything else equal,  $y_1$  would fall and  $y_2$  would rise in the current period, implying that both series adjust toward its long-run equilibrium. Notice that both  $\theta_{11}$  and  $\theta_{21}$  cannot be equal to zero. However, if  $\theta_{11} < 0$  and  $\theta_{21} = 0$ , then all of the adjustment falls on  $y_1$ , or vice versa if  $\theta_{11} = 0$  and  $\theta_{21} > 0$ . Note also that the larger are the speed of adjustment parameters (with the right signs), the greater is the convergence rate toward equilibrium. Of course, at least one of those terms must be nonzero, implying the existence of Granger causality in cointegrated systems in at least one direction; see Chapter 32 by Lütkepohl in this volume for

the formal definition of causality. Hence, the appeal of the VECM formulation is that it combines flexibility in dynamic specification with desirable long-run properties: it could be seen as capturing the transitional dynamics of the system to the long-run equilibrium suggested by economic theory (see, e.g. Hendry and Richard, 1983). Further, if cointegration exists, the VECM representation will generate better forecasts than the corresponding representation in first-differenced form (i.e. with  $\theta_{11} = \theta_{21} = 0$ ), particularly over medium- and long-run horizons, since under cointegration  $z_t$  will have a finite forecast error variance whereas any other linear combination of the forecasts of the individual series in  $y_t$  will have infinite variance; see Engle and Yoo (1987) for further details.

## 2.2 Estimation and testing for cointegration in a single equation framework

Based upon the VECM representation, Engle and Granger (1987) suggest a two-step estimation procedure for single equation dynamic modeling which has become very popular in applied research. Assuming that  $y_t \sim I(1)$ , then the procedure goes as follows:

1. First, in order to test whether the series are cointegrated, the *cointegration regression*

$$y_{1t} = \alpha + \beta y_{2t} + z_t \quad (30.5)$$

is estimated by ordinary least squares (OLS) and it is tested whether the *cointegrating residuals*  $\hat{z}_t = y_{1t} - \hat{\alpha} - \hat{\beta} y_{2t}$  are  $I(1)$ . To do this, for example, we can perform a Dickey–Fuller test on the residual sequence  $\{\hat{z}_t\}$  to determine whether it has a unit root. For this, consider the autoregression of the residuals

$$\Delta \hat{z}_t = \rho_1 \hat{z}_{t-1} + \varepsilon_t, \quad (30.6)$$

where no intercept term has been included since the  $\{\hat{z}_t\}$ , being residuals from a regression equation with a constant term, have zero mean. If we can reject the null hypothesis that  $\rho_1 = 0$  against the alternative  $\rho_1 < 0$  at a given significance level, we can conclude that the residual sequence is  $I(0)$  and, therefore, that  $y_{1t}$  and  $y_{2t}$  are  $CI(1, 1)$ . It is noteworthy that for carrying out this test it is not possible to use the Dickey–Fuller tables themselves since  $\{\hat{z}_t\}$  are a generated series of residuals from fitting regression (30.5). The problem is that the OLS estimates of  $\alpha$  and  $\beta$  are such that they minimize the residual variance in (30.5) and thus prejudice the testing procedure toward finding stationarity. Hence, larger (in absolute value) critical levels than the standard Dickey–Fuller ones are needed. In this respect, MacKinnon (1991) provides appropriate tables to test the null hypothesis  $\rho_1 = 0$  for any sample size and also when the number of regressors in (30.5) is expanded from one to several variables. In general, if the  $\{\hat{\varepsilon}_t\}$  sequence exhibits serial correlation, then an augmented Dickey–Fuller (ADF) test should be used, based this time on the extended autoregression

$$\Delta \hat{z}_t = p_1 \hat{z}_{t-1} + \sum_{i=1}^p \zeta_i \Delta \hat{z}_{t-i} + \varepsilon_t, \quad (30.6')$$

where again, if  $p_1 < 0$ , we can conclude that  $y_{1t}$  and  $y_{2t}$  are CI(1, 1). Alternative versions of the test on  $\{\hat{z}_t\}$  being I(1) versus I(0) can be found in Phillips and Ouliaris (1990). Banerjee *et al.* (1998), in turn, suggest another class of tests based this time on the direct significance of the loading parameters in (30.4) and (30.4') where the  $\beta$  coefficient is estimated alongside the remaining parameters in a single step using nonlinear least squares (NLS).

If we reject that  $\hat{z}_t$  are I(1), Stock (1987) has shown that the OLS estimate of  $\beta$  in equation (30.5) is *super-consistent*, in the sense that the OLS estimator  $\hat{\beta}$  converges in probability to its true value  $\beta$  at a rate proportional to the inverse of the sample size,  $T^{-1}$ , rather than at  $T^{-1/2}$  as is the standard result in the ordinary case where  $y_{1t}$  and  $y_{2t}$  are I(0). Thus, when  $T$  grows, convergence is much quicker in the CI(1, 1) case. The intuition behind this remarkable result can be seen by analyzing the behavior of  $\hat{\beta}$  in (30.5) (where the constant is omitted for simplicity) in the particular case where  $z_t \sim \text{iid } (0, \sigma_z^2)$ , and that  $\theta_{20} = \theta_{21} = 0$  and  $p_3 = p_4 = 0$ , so that  $y_{2t}$  is assumed to follow a simple random walk

$$\Delta y_{2t} = \varepsilon_{2t}, \quad (30.7)$$

or, integrating (30.7) backwards with  $y_{20} = 0$ ,

$$y_{2t} = \sum_{i=1}^t \varepsilon_{2i}, \quad (30.7')$$

with  $\varepsilon_{2t}$  possibly correlated with  $z_t$ . In this case, we get  $\text{var}(y_{2t}) = t \text{ var}(\varepsilon_{21}) = t \sigma_z^2$ , exploding as  $T \uparrow \infty$ . Nevertheless, it is not difficult to show that  $T^{-2} \sum_{t=1}^T y_{2t}^2$  converges to a random variable. Similarly, the cross-product  $T^{-1/2} \sum_{t=1}^T y_{2t} z_t$  will explode, in contrast to the stationary case where a simple application of the central limit theorem implies that it is asymptotically normally distributed. In the I(1) case,  $T^{-1} \sum_{t=1}^T y_{2t} z_t$  converges also to a random variable. Both random variables are functionals of Brownian motions which will be denoted henceforth, in general, as  $f(B)$ . A Brownian motion is a zero-mean normally distributed continuous (a.s.) process with independent increments, i.e. loosely speaking, the continuous version of the discrete random walk; see Phillips (1987), and Chapter 29 by Bierens in this volume for further details.

Now, from the expression for the OLS estimator of  $\beta$ , we obtain

$$\hat{\beta} - \beta = \frac{\sum_{t=1}^T y_{2t} z_t}{\sum_{t=1}^T y_{2t}^2}, \quad (30.8)$$

and, from the previous discussion, it follows that

$$T(\hat{\beta} - \beta) = \frac{T^{-1} \sum_{t=1}^T y_{2t} z_t}{T^{-2} \sum_{t=1}^T y_{2t}^2} \quad (30.9)$$

is asymptotically (as  $T \uparrow \infty$ ) the ratio of two non-degenerate random variables that in general, is not normally distributed. Thus, in spite of the super-consistency, standard inference cannot be applied to  $\hat{\beta}$  except in some restrictive cases which are discussed below.

2. After rejecting the null hypothesis that the cointegrating residuals in equation (30.5) are I(1), the  $\hat{z}_{t-1}$  term is included in the VECM system and the remaining parameters are estimated by OLS. Indeed, given the super-consistency of  $\hat{\beta}$ , Engle and Granger (1987) show that their asymptotic distributions will be identical to using the true value of  $\beta$ . Now, all the variables in (30.4) and (30.4') are I(0) and conventional modeling strategies (e.g. testing the maximum lag length, residual autocorrelation or whether either  $\theta_{11}$  or  $\theta_{21}$  is zero, etc.) can be applied to assess model adequacy; see Chapter 32 by Lütkepohl in this volume for further details.

In spite of the beauty and simplicity of the previous procedure, however, several problems remain. In particular, although  $\hat{\beta}$  is super-consistent, this is an asymptotic result and thus biases could be important in finite samples. For instance, assume that the rates of convergence of two estimators are  $T^{-1/2}$  and  $10^{10}T^{-1}$ . Then, we will need huge sample sizes to have the second estimator dominating the first one. In this sense, Monte Carlo experiments by Banerjee *et al.* (1993) showed that the biases could be important particularly when  $z_t$  and  $\Delta y_{2t}$  are highly serially correlated and they are not independent. Phillips (1991), in turn, has shown analytically that in the case where  $y_{2t}$  and  $z_t$  are independent at all leads and lags, the distribution in (30.9) as  $T$  grows behaves like a Gaussian distribution (technically is a *mixture of normals*) and, hence, the distribution of the  $t$ -statistic of  $\beta$  is also asymptotically normal. For this reason, Phillips and Hansen (1990) have developed an estimation procedure which corrects for the previous bias while achieving-asymptotic normality. The procedure, denoted as a *fully modified ordinary least squares estimator* (FM-OLS), is based upon a correction to the OLS estimator given in (30.8) by which the error term  $z_t$  is conditioned on the whole process  $\{\Delta y_{2t}, t = 0, \pm 1, \dots\}$  and, hence, orthogonality between regressors and disturbance is achieved by construction. For example, if  $z_t$  and  $\varepsilon_{2t}$  in (30.5) and (30.7) are correlated white noises with  $\gamma = E(z_t \varepsilon_{2t})/\text{var}(\varepsilon_{2t})$ , the FM-OLS estimator of  $\beta$ , denoted  $\hat{\beta}_{\text{FM}}$ , is given by

$$\hat{\beta}_{\text{FM}} = \frac{\sum_{t=1}^T y_{2t}(y_{1t} - \hat{\gamma}\Delta y_{2t})}{\sum_{t=1}^T y_{2t}^2}, \quad (30.10)$$

where  $\hat{\gamma}$  is the empirical counterpart of  $\gamma$  obtained from regressing the OLS residuals  $\hat{z}_t$  on  $\Delta y_{2t}$ . When  $z_t$  and  $\Delta y_{2t}$  follow more general processes, the FM-OLS estimator of  $\beta$  is similar to (30.10) except that further corrections are needed in its numerator. Alternatively, Saikkonen (1991) and Stock and Watson (1993) have shown that, since  $E(z_t | \{\Delta y_{2t}\}) = h(L)\Delta y_{2t}$ , where  $h(L)$  is a two-sided filter in the lag operator  $L$ , regression of  $y_{1t}$  on  $y_{2t}$  and leads and lags of  $\Delta y_{2t}$  (suitably truncated), using either OLS or GLS, will yield an estimator of  $\beta$  which is asymptotically equivalent to the FM-OLS estimator. The resulting estimation approach is known as *dynamic OLS* (respectively GLS) or DOLS (respectively, DGLS).

### 3 SYSTEM-BASED APPROACHES TO COINTEGRATION

Whereas in the previous section we confined the analysis to the case where there is at most a single cointegrating vector in a bivariate system, this setup is usually quite restrictive when analyzing the cointegrating properties of an  $n$ -dimensional vector of I(1) variables where several cointegration relationships may arise. For example, when dealing with a trivariate system formed by the logarithms of nominal wages, prices, and labor productivity, there may exist two relationships, one determining an employment equation and another determining a wage equation. In this section we survey some of the popular estimation and testing procedures for cointegration in this more general multivariate context, which will be denoted as system-based approaches.

In general, if  $y_t$  now represents a vector of  $n$  I(1) variables its Wold representation (assuming again no deterministic terms) is given by

$$\Delta y_t = C(L)\varepsilon_t, \quad (30.11)$$

where now  $\varepsilon_t \sim \text{nid}(0, \Sigma)$ ,  $\Sigma$  being the covariance matrix of  $\varepsilon_t$  and  $C(L)$  an  $(n \times n)$  invertible matrix of polynomial lags, where the term “invertible” means that  $|C(L)| = 0$  has all its roots strictly larger than unity in absolute value. If there is a cointegrating  $(n \times 1)$  vector,  $\beta' = (\beta_{11}, \dots, \beta_{nn})$ , then, premultiplying (30.11) by  $\beta'$  yields

$$\beta' \Delta y_t = \beta' [C(1) + \tilde{C}(L)\Delta] \varepsilon_t, \quad (30.12)$$

where  $C(L)$  has been expanded around  $L = 1$  using a first-order Taylor expansion and  $\tilde{C}(L)$  can be shown to be an invertible lag matrix. Since the cointegration property implies that  $\beta' y_t$  is I(0), then it must be that  $\beta' C(1) = 0$  and hence  $\Delta (= 1 - L)$  will cancel out on both sides of (30.12). Moreover, given that  $C(L)$  is invertible, then  $y_t$  has a vector autoregressive representation such that

$$A(L)y_t = \varepsilon_t, \quad (30.13)$$

where  $A(L)C(L) = (1 - L)I_n$ ,  $I_n$  being the  $(n \times n)$  identity matrix. Hence, we must have that  $A(1)\tilde{C}(1) = 0$ , implying that  $A(1)$  can be written as a linear combination of the elements  $\beta$ , namely,  $A(1) = \alpha\beta'$ , with  $\alpha$  being another  $(n \times 1)$  vector. In the

same manner, if there were  $r$  cointegrating vectors ( $0 < r < n$ ), then  $A(1) = B\Gamma'$ , where  $B$  and  $\Gamma$  are this time  $(n \times r)$  matrices which collect the  $r$  different  $\alpha$  and  $\beta$  vectors. Matrix  $B$  is known as the *loading matrix* since its rows determine how many cointegrating relationships enter each of the individual dynamic equations in (30.13). Testing the rank of  $A(1)$  or  $C(1)$ , which happen to be  $r$  and  $n - r$ , respectively, constitutes the basis of the following two procedures.

### 3.1 The Johansen's method

Johansen (1995) develops a maximum likelihood estimation procedure based on the so-called *reduced rank regression method* that, as the other methods to be later discussed, presents some advantages over the two-step regression procedure described in the previous section. First, it relaxes the assumption that the cointegrating vector is unique, and, second, it takes into account the short-run dynamics of the system when estimating the cointegrating vectors. The underlying intuition behind Johansen's testing procedure can be easily explained by means of the following example. Assume that  $y_t$  has a VAR(1) representation, that is,  $A(L)$  in (30.13) is such that  $A(L) = I_n - A_1 L$ . Hence, the VAR(1) process can be reparameterized in the VECM representation as

$$\Delta y_t = (A_1 - I_n)y_{t-1} + \varepsilon_t. \quad (30.14)$$

If  $A_1 - I_n = -A(1) = 0$ , then  $y_t$  is I(1) and there are no cointegrating relationships ( $r = 0$ ), whereas if  $\text{rank}(A_1 - I_n) = n$ , there are  $n$  cointegrating relationships among the  $n$  series and hence  $y_t \sim I(0)$ . Thus, testing the null hypothesis that the number of cointegrating vectors ( $r$ ) is equivalent to testing whether  $\text{rank}(A_1 - I_n) = r$ . Likewise, alternative hypotheses could be designed in different ways, e.g. that the rank is  $(r + 1)$  or that it is  $n$ .

Under the previous considerations, Johansen (1995) deals with the more general case where  $y_t$  follows a VAR( $p$ ) process of the form

$$y_t = A_1 y_{t-1} + A_2 y_{t-2} + \dots + A_p y_{t-p} + \varepsilon_t, \quad (30.15)$$

which, as in (30.4) and (30.4'), can be rewritten in the ECM representation

$$\Delta y_t = D_1 \Delta y_{t-1} + D_2 \Delta y_{t-2} + \dots + D_{p-1} \Delta y_{t-p+1} + D y_{t-1} + \varepsilon_t. \quad (30.16)$$

Where  $D_i = -(A_{i+1} + \dots + A_p)$ ,  $i = 1, 2, \dots, p - 1$ , and  $D = (A_1 + \dots + A_p - I_n) = -A(1) = -B\Gamma'$ . To estimate  $B$  and  $\Gamma$ , we need to estimate  $D$  subject to some identification restriction since otherwise  $B$  and  $\Gamma$  could not be separately identified. Maximum likelihood estimation of  $D$  goes along the same principles of the basic partitioned regression model, namely, the regressand and the regressor of interest ( $\Delta y_t$  and  $y_{t-1}$ ) are regressed by OLS on the remaining set of regressors ( $\Delta y_{t-1}, \dots, \Delta y_{t-p+1}$ ) giving rise to two matrices of residuals denoted as  $\hat{e}_0$  and  $\hat{e}_1$  and the regression model  $\hat{e}_{ot} = \hat{D}\hat{e}_{1t} + \text{residuals}$ . Following the preceding discussion, Johansen (1995) shows that testing for the rank of  $\hat{D}$  is equivalent to test for the number of

canonical correlations between  $\hat{e}_0$  and  $\hat{e}_1$  that are different from zero. This can be conducted using either of the following two test statistics

$$\lambda_{tr}(r) = -T \sum_{i=r+1}^n \ln(1 - \hat{\lambda}_i) \quad (30.17)$$

$$\lambda_{max}(r, r+1) = -T \ln(1 - \hat{\lambda}_{r+1}), \quad (30.18)$$

where the  $\hat{\lambda}_i$ s are the eigenvalues of the matrix  $S_{10}S_{00}^{-1}S_{01}$  with respect to the matrix  $S_{11}$ , ordered in decreasing order ( $1 > \hat{\lambda}_1 > \dots > \hat{\lambda}_n > 0$ ), where  $S_{ij} = T^{-1}\sum_{t=1}^T \hat{e}_{it}\hat{e}_{jt}'$ ,  $i, j = 0, 1$ . These eigenvalues can be obtained as the solution of the determinantal equation

$$|\lambda S_{11} - S_{10}S_{00}^{-1}S_{01}| = 0. \quad (30.19)$$

The statistic in (30.17), known as the *trace statistic*, tests the null hypothesis that the number of cointegrating vectors is less than or equal to  $r$  against a general alternative. Note that, since  $\ln(1) = 0$  and  $\ln(0)$  tends to  $-\infty$ , it is clear that the trace statistic equals zero when all the  $\hat{\lambda}_i$ s are zero, whereas the further the eigenvalues are from zero the more negative is  $\ln(1 - \hat{\lambda}_i)$  and the larger is the statistic. Likewise, the statistic in (30.18), known as the *maximum eigenvalue statistic*, tests a null of  $r$  cointegrating vectors against the specific alternative of  $r + 1$ . As above, if  $\hat{\lambda}_{r+1}$  is close to zero, the statistic will be small. Further, if the null hypothesis is not rejected, the  $r$  cointegrating vectors contained in matrix  $\Gamma$  can be estimated as the first  $r$  columns of matrix  $\hat{V} = (\hat{v}_1, \dots, \hat{v}_n)$  which contains the eigenvectors associated to the eigenvalues in (30.19) computed as

$$(\lambda_i S_{11} - S_{10}S_{00}^{-1}S_{01})\hat{v}_i = 0, \quad i = 1, 2, \dots, n$$

subject to the length normalization rule  $\hat{V}'S_{11}\hat{V} = I_n$ . Once  $\Gamma$  has been estimated, estimates of the  $B$ ,  $D_i$ , and  $\Sigma$  matrices in (30.16) can be obtained by inserting  $\hat{\Gamma}$  in their corresponding OLS formulae which will be functions of  $\Gamma$ .

Osterwald-Lenum (1992) has tabulated the critical values for both tests using Monte Carlo simulations, since their asymptotic distributions are multivariate  $f(B)$  which depend upon: (i) the number of nonstationary components under the null hypothesis ( $n - r$ ) and (ii) the form of the vector of deterministic components,  $\mu$  (e.g. a vector of drift terms), which needs to be included in the estimation of the ECM representation where the variables have nonzero means. Since, in order to simplify matters, the inclusion of deterministic components in (30.16) has not been considered so far, it is worth using a simple example to illustrate the type of interesting statistical problems that may arise when taking them into account. Suppose that  $r = 1$  and that the unique cointegrating vector in  $\Gamma$  is normalized to be  $\beta' = (1, \beta_{22}, \dots, \beta_{nn})$ , while the vector of speed of adjustment parameters, with which the cointegrating vector appears in each of the equations for the  $n$  variables, is  $\alpha' = (\alpha_{11}, \dots, \alpha_{nn})$ . If there is a vector of drift terms  $\mu' = (\mu_1, \dots, \mu_n)$  such that they satisfy the restrictions  $\mu_i = \alpha_{11}\mu_1$  (with  $\alpha_{11} = 1$ ), it then

follows that all  $\Delta y_{it}$  in (30.16) are expected to be zero when  $y_{1,t-1} + \beta_{22}y_{2,t-1} + \dots + \beta_{nn}y_{n,t-1} + \mu_1 = 0$  and, hence, the general solution for each of the  $\{y_{it}\}$  processes, when integrated, will not contain a time trend. Many other possibilities, like, for example, allowing for a linear trend in each variable but not in the cointegrating relations, may be considered. In each case, the asymptotic distribution of the cointegration tests given in (30.17) and (30.18) will differ, and the corresponding sets of simulated critical values can be found in the reference quoted above. Sometimes, theory will guide the choice of restrictions; for example, if one is considering the relation between short-term and long-term interest rates, it may be wise to impose the restriction that the processes for both interest rates do not have linear trends and that the drift terms are restricted to appear in the cointegrating relationship interpreted as the “term structure.” However, in other instances one may be interested in testing alternative sets of restrictions on the way  $\mu$  enters the system; see, e.g. Chapter 32 by Lütkepohl in this volume for further details.

In that respect, the Johansen’s approach allows to test restrictions on  $\mu$ ,  $B$ , and  $\Gamma$  subject to a given number of cointegrating relationships. The insight to all these tests, which turn out to have asymptotic chi-square distributions, is to compare the number of cointegrating vectors (i.e. the number of eigenvalues which are significantly different from zero) both when the restrictions are imposed and when they are not. Since if the true cointegration rank is  $r$ , only  $r$  linear combinations of the variables are stationary, one should find that the number of cointegrating vectors does not diminish if the restrictions are not binding and vice versa. Thus, denoting by  $\hat{\lambda}_i$  and  $\lambda_i^*$  the set of  $r$  eigenvalues for the unrestricted and restricted cases, both sets of eigenvalues should be equivalent if the restrictions are valid. For example, a modification of the trace test in the form

$$T \sum_{i=1}^r [\ln(1 - \lambda_i^*) - \ln(1 - \hat{\lambda}_i)] \quad (30.20)$$

will be small if the  $\lambda_i^*$ s are similar to the  $\hat{\lambda}_i$ s, whereas it will be large if the  $\lambda_i^*$ s are smaller than the  $\hat{\lambda}_i$ s. If we impose  $s$  restrictions, then the above test will reject the null hypothesis if the calculated value of (30.20) exceeds that in a chi-square table with  $r(n - s)$  degrees of freedom.

Most of the existing Monte Carlo studies on the Johansen methodology point out that dimension of the data series for a given sample size may pose particular problems since the number of parameters of the underlying VAR models grows very large as the dimension increases. Likewise, difficulties often arise when, for a given  $n$ , the lag length of the system,  $p$ , is either over- or under-parameterized. In particular, Ho and Sorensen (1996) and Gonzalo and Pitarakis (1999) show, by numerical methods, that the cointegrating order will tend to be overestimated as the dimension of the system increases relative to the time dimension, while serious size and power distortions arise when choosing too short and too long a lag length, respectively. Although several degrees of freedom adjustments to improve the performance of the test statistics have been advocated (see, e.g. Reinsel and Ahn, 1992), researchers ought to have considerable care when using

the Johansen estimator to determine cointegration order in high dimensional systems with small sample sizes. Nonetheless, it is worth noticing that a useful approach to reduce the dimension of the VAR system is to rely upon exogeneity arguments to construct smaller conditional systems as suggested by Ericsson (1992) and Johansen (1992a). Equally, if the VAR specification is not appropriate, Phillips (1991) and Saikkonen (1992) provide efficient estimation of cointegrating vectors in more general time series settings, including vector ARMA processes.

### 3.2 Common trends representation

As mentioned above, there is a dual relationship between the number of cointegrating vectors ( $r$ ) and the number of common trends ( $n - r$ ) in an  $n$ -dimensional system. Hence, testing for the dimension of the set of “common trends” provides an alternative approach to testing for the cointegration order in a VAR//VECM representation. Stock and Watson (1988) provide a detailed study of this type of methodology based on the use of the so-called Beveridge–Nelson (1981) decomposition. This works from the Wold representation of an I(1) system, which we can write as in expression (30.11) with  $C(L) = \sum_{j=0}^{\infty} C_j L^j$ ,  $C_0 = I_n$ . As shown in expression (30.12),  $C(L)$  can be expanded as  $C(L) = C(1) + \tilde{C}(L)(1 - L)$ , so that, by integrating (30.11), we get

$$y_t = C(1)Y_t + \tilde{w}_t, \quad (30.21)$$

where  $\tilde{w}_t = \tilde{C}(L)\varepsilon_t$  can be shown to be covariance stationary, and  $Y_t = \sum_{i=1}^t \varepsilon_i$  is a latent or unobservable set of random walks which capture the I(1) nature of the data. However, as above mentioned, if the cointegration order is  $r$ , there must be an  $(r \times n)$   $\Gamma$  matrix such that  $\Gamma' C(1) = 0$  since, otherwise,  $\Gamma' y_t$  would be I(1) instead of I(0). This means that the  $(n \times n)$   $C(1)$  matrix cannot have full rank. Indeed, from standard linear algebra arguments, it is easy to prove that the rank of  $C(1)$  is  $(n - r)$ , implying that there are only  $(n - r)$  independent common trends in the system. Hence, there exists the so-called *common trends representation* of a cointegrated system, such that

$$y_t = \Phi y_t^c + \tilde{w}_t, \quad (30.22)$$

where  $\Phi$  is an  $n \times (n - r)$  matrix of loading coefficients such that  $\Gamma' \Phi = 0$  and  $y_t^c$  is an  $(n - r)$  vector random walk. In other words,  $y_t$  can be written as the sum of  $(n - r)$  common trends and an I(0) component. Thus, testing for  $(n - r)$  common trends in the system is equivalent to testing for  $r$  cointegrating vectors. In this sense, Stock and Watson’s (1988) testing approach relies upon the observation that, under the null hypothesis, the first-order autoregressive matrix of  $y_t^c$  should have  $(n - r)$  eigenvalues equal to unity, whereas, under the alternative hypothesis of higher cointegration order, some of those eigenvalues will be less than unity. It is worth noticing that there are other alternative strategies to identify the set of common trends,  $y_t^c$ , which do not impose a vector random walk structure. In

particular, Gonzalo and Granger (1995), using arguments embedded in the Johansen's approach, suggest identifying  $y_t^c$  as linear combinations of  $y_t$  which are not caused in the long-run by the cointegration relationships  $\Gamma'y_{t-1}$ . These linear combinations are the orthogonal complement of matrix  $B$  in (30.16),  $y_t^c = B_\perp y_t$ , where  $B_\perp$  is an  $(n \times (n - r))$  full ranked matrix, such that  $B'B_\perp = 0$ , that can be estimated as the last  $(n - r)$  eigenvectors of the second moments matrix  $S_{01}S_{11}^{-1}S_{10}$  with respect to  $S_{00}$ . For instance, when some of the rows of matrix  $B$  are zero, the common trends will be linear combinations of those I(1) variables in the system where the cointegrating vectors do not enter into their respective adjustment equations. Since common trends are expressed in terms of observable variables, instead of a latent set of random walks, economic theory can again be quite useful in helping to provide useful interpretation of their role. For example, the rational expectations version of the permanent income hypothesis of consumption states that consumption follows a random walk whilst saving (disposable income minus consumption) is I(0). Thus, if the theory is a valid one, the cointegrating vector in the system formed by consumption and disposable income should be  $\beta' = (1, -1)$  and it would only appear in the second equation (i.e.  $\alpha' = (0, \alpha_{22})$ ), implying that consumption should be the common trend behind the nonstationary behavior of both variables.

To give a simple illustration of the conceptual issues discussed in the previous two sections, let us consider the following Wold (MA) representation of the bivariate I(1) process  $y_t = (y_{1t}, y_{2t})'$ ,

$$(1 - L) \begin{pmatrix} y_{1t} \\ y_{2t} \end{pmatrix} = (1 - 0.2L)^{-1} \begin{pmatrix} 1 - 0.6L & 0.8L \\ 0.2L & 1 - 0.6L \end{pmatrix} \begin{pmatrix} \varepsilon_{1t} \\ \varepsilon_{2t} \end{pmatrix}.$$

Evaluating  $C(L)$  at  $L = 1$  yields

$$C(1) = \begin{pmatrix} 0.5 & 1 \\ 0.25 & 0.5 \end{pmatrix},$$

so that  $\text{rank } C(1) = 1$ . Hence,  $y_t \sim CI(1, 1)$ . Next, inverting  $C(L)$ , yields the VAR representation

$$\begin{pmatrix} 1 - 0.6L & -0.8L \\ -0.2L & 1 - 0.6L \end{pmatrix} \begin{pmatrix} y_{1t} \\ y_{2t} \end{pmatrix} = \begin{pmatrix} \varepsilon_{1t} \\ \varepsilon_{2t} \end{pmatrix}$$

where

$$A(1) = \begin{pmatrix} 0.4 & -0.8 \\ -0.2 & 0.4 \end{pmatrix},$$

so that  $\text{rank } A(1) = 1$  and

$$A(1) = \begin{pmatrix} 0.4 \\ -0.2 \end{pmatrix} (1, -2) = \alpha \beta'.$$

Hence, having normalized on the first element, the cointegrating vector is  $\beta' = (1, -2)$ , leading to the following VECM representation of the system

$$(1 - L) \begin{pmatrix} y_{1t} \\ y_{2t} \end{pmatrix} = \begin{pmatrix} -0.4 \\ 0.2 \end{pmatrix} (1, -2) \begin{pmatrix} y_{1,t-1} \\ y_{2,t-1} \end{pmatrix} + \begin{pmatrix} \varepsilon_{1t} \\ \varepsilon_{2t} \end{pmatrix}.$$

Next, given  $C(1)$  and normalizing again on the first element, it is clear that the common factor is  $y_t^c = \sum_{i=1}^t \varepsilon_{1i} + 2\sum_{i=1}^t \varepsilon_{2i}$ , whereas the loading vector  $\Phi$  and the common trend representation would be as follows

$$\begin{pmatrix} y_{1t} \\ y_{2t} \end{pmatrix} = \begin{pmatrix} 0.5 \\ 0.25 \end{pmatrix} y_t^c + \tilde{w}_t.$$

Notice that  $\beta' y_t$  eliminates  $y_t^c$  from the linear combination which achieves cointegration. In other words,  $\Phi$  is the orthogonal complement of  $\beta$  once the normalization criteria has been chosen.

Finally, to examine the effects of drift terms, let us add a vector  $\mu = (\mu_1, \mu_2)'$  of drift coefficients to the VAR representation. Then, it is easy to prove that  $y_{1t}$  and  $y_{2t}$  will have a linear trends with slopes equal to  $\mu_1/2 + \mu_2$  and  $\mu_1/4 + \mu_2/2$ , respectively. When  $2\mu_1 + \mu_2 \neq 0$  the data will have linear trends, whereas the cointegrating relationship will not have them, since the linear combination in  $\beta$  annihilates the individual trends for any  $\mu_1$  and  $\mu_2$ .

The interesting case arises when the restriction  $2\mu_1 + \mu_2 = 0$  holds, since now the linear trend is purged from the system, leading to the restricted ECM representation

$$(1 - L) \begin{pmatrix} y_{1t} \\ y_{2t} \end{pmatrix} = \begin{pmatrix} -0.4 \\ 0.2 \end{pmatrix} (1, -2, -\mu_1^*) \begin{pmatrix} y_{1,t-1} \\ y_{2,t-1} \\ 1 \end{pmatrix} + \begin{pmatrix} \varepsilon_{1t} \\ \varepsilon_{2t} \end{pmatrix},$$

where  $\mu_1^* = \mu_1/0.4$ .

#### 4 FURTHER RESEARCH ON COINTEGRATION

Although the discussion in the previous sections has been confined to the possibility of cointegration arising from linear combinations of I(1) variables, the literature is currently proceeding in several interesting extensions of this standard setup. In the sequel we will briefly outline some of those extensions which have drawn a substantial amount of research in the recent past.

## 4.1 Higher order cointegrated systems

The statistical theory of  $I(d)$  systems with  $d = 2, 3, \dots$ , is much less developed than the theory for the  $I(1)$  model, partly because it is uncommon to find time series, at least in economics, whose degree of integration higher than two, partly because the theory is quite involved as it must deal with possibly multicointegrated cases where, for instance, linear combinations of levels and first differences can achieve stationarity. We refer the reader to Haldrup (1999) for a survey of the statistical treatment of  $I(2)$  models, restricting the discussion in this chapter to the basics of the  $CI(2, 2)$  case.

Assuming, thus, that  $y_t \sim CI(2, 2)$ , with Wold representation given by

$$(1 - L)^2 y_t = C(L) \varepsilon_t, \quad (30.23)$$

then, by means of a Taylor expansion, we can write  $C(L)$  as

$$C(L) = C(1) - C^*(1)(1 - L) + \tilde{C}(L)(1 - L)^2,$$

with  $C^*(1)$  being the first derivative of  $C(L)$  with respect to  $L$ , evaluated at  $L = 1$ . Following the arguments in the previous section,  $y_t \sim CI(2, 2)$  implies that there exists a set of cointegrating vectors such that  $\Gamma' C(1) = \Gamma' C^*(1) = 0$ , from which the following VECM representation can be derived

$$A^*(L)(1 - L)^2 y_t = -B_1 \Gamma'_1 y_{t-1} - B_2 \Gamma'_2 \Delta y_{t-1} + \varepsilon_t \quad (30.24)$$

with  $A^*(0) = I_n$ . Johansen (1992b) has developed the maximum likelihood estimation of this class of models, which, albeit more complicated than in the  $CI(1, 1)$  case, proceeds along similar lines to those discussed in Section 3.

Likewise, there are systems where the variables have unit roots at the seasonal frequencies. For example, if a seasonally integrated variable is measured every half-a-year, then it will have the following Wold representation

$$(1 - L^2) y_t = C(L) \varepsilon_t. \quad (30.25)$$

Since  $(1 - L^2) = (1 - L)(1 + L)$ , the  $\{y_t\}$  process could be cointegrated by obtaining linear combinations which eliminate the unit root at the zero frequency,  $(1 - L)$ , and/or at the seasonal frequency,  $(1 + L)$ . Assuming that  $\Gamma_1$  and  $\Gamma_2$  are sets of cointegrating relationships at each of the two above mentioned frequencies, Helleberg *et al.* (1990) have shown that the VECM representation of the system this time will be

$$A^*(L)(1 - L^2) y_t = -B_1 \Gamma'_1 \Delta y_{t-1} - B_2 \Gamma'_2 (y_{t-1} + y_{t-2}) + \varepsilon_t, \quad (30.26)$$

with  $A^*(0) = I_n$ . Notice that if there is no cointegration in  $(1 + L)$ ,  $\Gamma_2 = 0$  and the second term in the right-hand side of (30.26) will vanish, whereas lack of cointegration in  $(1 - L)$  implies  $\Gamma_1 = 0$  and the first term will disappear. Similar arguments

can be used to obtain VECM representations for quarterly or monthly data with seasonal difference operators of the form  $(1 - L^4)$  and  $(1 - L^{12})$ , respectively.

## 4.2 Fractionally cointegrated systems

As discussed earlier in this chapter, one of the main characteristics of the existence of unit roots in the Wold representation of a time series is that they have “long memory,” in the sense that shocks have permanent effects on the levels of the series so that the variance of the levels of the series explodes. In general, it is known that if the differencing filter  $(1 - L)^d$ ,  $d$  being now a real number, is needed to achieve stationarity, then the coefficient of  $\varepsilon_{i,j}$  in the Wold representation of the  $I(d)$  process has a leading term  $j^{d-1}$  (e.g. the coefficient in an  $I(1)$  process is unity, since  $d = 1$ ) and the process is said to be *fractionally integrated of order d*. In this case, the variance of the series in levels will explode at the rate  $T^{2d-1}$  (e.g. at the rate  $T$  when  $d = 1$ ) and then all that is needed to have this kind of long memory is a degree of differencing  $d > 1/2$ .

Consequently, it is clear that a wide range of dynamic behavior is ruled out a priori if  $d$  is restricted to integer values and that a much broader range of cointegration possibilities are entailed when fractional cases are considered. For example, we could have a pair of series which are  $I(d_1)$ ,  $d_1 > 1/2$ , which cointegrate to obtain an  $I(d_0)$  linear combination such that  $0 < d_0 < 1/2$ . A further complication arises in this case if the various integration orders are not assumed to be known and need to be estimated for which frequency domain regression methods are normally used. Extensions of least squares and maximum likelihood methods of estimation and testing for cointegration within this more general framework can be found in Jeganathan (1996), Marmol (1998) and Robinson and Marinucci (1998).

## 4.3 Nearly cointegrated systems

Even when a vector of time series is  $I(1)$ , the size of the unit root in each of the series could be very different. For example, in terms of the common trend representation of a bivariate system discussed above, it could well be the case that  $y_{1t} = \phi_1 y_t^c + \tilde{w}_{1t}$  and  $y_{2t} = \phi_2 y_t^c + \tilde{w}_{2t}$  are such that  $\phi_1$  is close to zero and that  $\phi_2$  is large. Then  $y_{1t}$  will not be different from  $\tilde{w}_{1t}$  which is an  $I(0)$  series while  $y_{2t}$  will be clearly  $I(1)$ . The two series are cointegrated, since they share a common trend. However, if we regress  $y_{1t}$  on  $y_{2t}$ , i.e. we normalize the cointegrating vector on the coefficient of  $y_{1t}$ , the regression will be nearly unbalanced, namely, the regressand is almost  $I(0)$  whilst the regressor is  $I(1)$ . In this case, the estimated coefficient on  $y_{2t}$  will converge quickly to zero and the residuals will resemble the properties of  $y_{1t}$ , i.e. they will look stationary. Thus, according to the Engle and Granger testing approach, we will often reject the null of no cointegration. By contrast, if we regress  $y_{2t}$  on  $y_{1t}$ , now the residuals will resemble the  $I(1)$  properties of the regressand and we will often reject cointegration. Therefore, normalization plays a crucial role in least squares estimation of cointegrating vectors in nearly cointegrated systems. Consequently, if one uses the static regression approach to

estimate the cointegrating vector, it follows from the previous discussion that it is better to use the “less integrated” variable as the regressand. Ng and Perron (1997) have shown that these problems remain when the equations are estimated using more efficient methods like FM-OLS and DOLS, while the Johansen’s methodology provides a better estimation approach, since normalization is only imposed on the length of the eigenvectors.

#### 4.4 Nonlinear error correction models

When discussing the role of the cointegrating relationship  $z_t$  in (30.3) and (30.3'), we motivated the EC model as the disequilibrium mechanism that leads to the particular equilibrium. However, as a function of an I(0) process is generally also I(0), an alternative more general VECM model has  $z_{t-1}$  in (30.3) and (30.3') replaced by  $g(z_{t-1})$  where  $g(z)$  is a function such that  $g(0) = 0$  and  $E[g(z)]$  exists. The function  $g(z)$  is such that it can be estimated nonparametrically or by assuming a particular parametric form. For example, one can include  $z^+ = \max\{0, z_t\}$  and  $z^- = \min\{0, z_t\}$  separately into the model or large and small values of  $z$  according to some prespecified threshold in order to deal with possible sign or size asymmetries in the dynamic adjustment. Further examples can be found in Granger and Teräsvirta (1993). The theory of nonlinear cointegration models is still fairly incomplete, but nice applications can be found in Gonzalez and Gonzalo (1998) and Balke and Fomby (1997).

#### 4.5 Structural breaks in cointegrated systems

The parameters in the cointegrating regression model (30.5) may not be constant through time. Gregory and Hansen (1995) developed a test for cointegration allowing for a structural break in the intercept as well as in the slope of model (30.5). The new regression model now looks like

$$y_{1t} = \alpha_1 + \alpha_2 D(t_0) + \beta_1 y_{2t} + \beta_2 y_{2t} D(t_0) + z_t, \quad (30.27)$$

where  $D(t_0)$  is a dummy variable such that  $D(t_0) = 0$  if  $0 < t \leq t_0$  and  $D(t_0) = 1$  if  $t_0 < t \leq T$ . The test for cointegration is conducted by testing for unit roots (for instance, with an ADF test) on the residuals  $\hat{z}_t$  for each  $t_0$ . Gregory and Hansen propose and tabulate the critical values of the test statistic

$$ADF^* = \inf_{1 < t_0 < T} \{ADF(t_0)\}.$$

The null hypothesis of no cointegration and no structural break is rejected if the statistic  $ADF^*$  is smaller than the corresponding critical value. In this case the structural break will be located at time  $t^*$  where the inf of the ADF test is obtained. The work of Gregory and Hansen is opening an extensive research on analyzing the stability of the parameters of multivariate possibly cointegrated systems models like the VECM in (30.16). Further work in this direction can be found in Hansen and Johansen (1993), Quintos (1994), Juhl (1997), and Arranz and Escribano (2000).

## 5 CONCLUDING REMARKS

The considerable gap in the past between the economic theorist, who had much to say about equilibrium but relatively less to say about dynamics and the econometrician whose models concentrated on the short-run dynamics disregarding the long-run equilibrium, has been bridged by the concept of cointegration. In addition to allowing the data to determine the short-run dynamics, cointegration suggest that models can be significantly improved by including long-run equilibrium conditions as suggested by economic theory. The generic existence of such long-run relationships, in turn, should be tested using the techniques discussed in this chapter to reduce the risk of finding spurious conclusions.

The literature on cointegration has greatly enhanced the existing methods of dynamic econometric modeling of economic time series and should be considered nowadays as a very valuable part of the practitioner's toolkit.

## References

- Arranz, M.A., and A. Escribano (2000). Cointegration testing under structural breaks: A robust extended error correction model. *Oxford Bulletin of Economics and Statistics* 62, 23–52.
- Balke, N., and T. Fomby (1997). Threshold cointegration. *International Economic Review* 38, 627–45.
- Banerjee, A., J.J. Dolado, J.W. Galbraith, and D.F. Hendry (1993). *Co-integration, Error Correction and the Econometric Analysis of Non-stationary Data*, Oxford: Oxford University Press.
- Banerjee, A., J.J. Dolado, and R. Mestre (1998). Error-correction mechanism tests for cointegration in a single-equation framework. *Journal of Time Series Analysis* 19, 267–84.
- Beveridge, S., and C.R. Nelson (1981). A new approach to decomposition of economic time series into permanent and transitory components with particular attention to measurement of the "Business Cycle". *Journal of Monetary Economics* 7, 151–74.
- Box, G.E.P., and G.M. Jenkins (1970). *Time Series Analysis: Forecasting and Control*, San Francisco: Holden Day.
- Campbell, J.Y., and R.J. Shiller (1987). Cointegration and tests of present value models. *Journal of Political Economy* 95, 1062–88.
- Davidson, J.E.H., D.F. Hendry, F. Srba, and S. Yeo (1978). Econometric modelling of the aggregate time series relationships between consumer's expenditure and income in the United Kingdom. *Economic Journal* 88, 661–92.
- Engle, R.F., and C.W.J. Granger (1987). Co-integration and error correction: Representation, estimation and testing. *Econometrica* 55, 251–76.
- Engle, R.F. and C.W.J. Granger (eds.) (1991). *Long-run Economic Relationships: Readings in Cointegration*. Oxford: Oxford University Press.
- Engle, R.F., and B.S. Yoo (1987). Forecasting and testing in cointegrated systems. *Journal of Econometrics* 35, 143–59.
- Ericsson, N.R. (1992). Cointegration, exogeneity and policy analysis: An overview. *Journal of Policy Modeling* 14, 424–38.
- Gonzalez, M., and J. Gonzalo (1998). Inference in threshold error correction model. *Discussion Paper*, Universidad Carlos III de Madrid.
- Gonzalo, J., and C.W.J. Granger (1995). Estimation of common long memory components in cointegrated systems. *Journal of Business and Economic Statistics* 13, 27–36.

- Gonzalo, J., and J.-Y. Pitarakis (1999). Dimensionality effect in cointegrated systems. In Granger Festschrift edited by R. Engle and H. White (eds.). Oxford: Oxford University Press.
- Granger, C.W.J. (1981). Some properties of time series data and their use in econometric model specification. *Journal of Econometrics* 23, 121–30.
- Granger, C.W.J. (1986). Developments in the study of cointegrated economic variables. *Oxford Bulletin of Economics and Statistics* 48, 213–28.
- Granger, C.W.J., and T.H. Lee (1989). Multicointegration. *Advances in Econometrics* 8, 71–84.
- Granger, C.W.J., and P. Newbold (1974). Spurious regressions in econometrics. *Journal of Econometrics* 2, 111–20.
- Granger, C.W.J., and T. Teräsvirta (1993). *Modelling Nonlinear Economic Relationships*. Oxford: Oxford University Press.
- Gregory, A.W., and B.E. Hansen (1995). Residual-based tests for cointegration in models with regime shifts. *Journal of Econometrics* 70, 99–126.
- Haldrup, N. (1999). A review of the econometric analysis of I(2) variables. In L. Oxley and M. McAleer (eds.) *Practical Issues in Cointegration Analysis*. Oxford: Blackwell.
- Hansen, H., and S. Johansen (1993). Recursive estimation in cointegrated VAR models. *Preprint 1*, Institute of Mathematical Statistics, University of Copenhagen.
- Hatanaka, M. (1996). *Time Series-Based Econometrics*. Oxford: Oxford University Press.
- Hendry, D.F., and G.E. Mizon (1978). Serial correlation as a convenient simplification not a nuisance: A comment on a study of the demand for money by the Bank of England. *Economic Journal* 88, 549–63.
- Hendry, D.F., and J.-F. Richard (1983). The econometric analysis of economic time series (with discussant). *International Statistical Review* 51, 111–63.
- Hylleberg, S., R.F. Engle, C.W.J. Granger, and B.S. Yoo (1990). Seasonal integration and cointegration. *Journal of Econometrics* 44, 215–28.
- Ho, M. and B. Sorensen (1996). Finding cointegration rank in high dimensional systems using the Johansen test: An illustration using data based on Monte Carlo simulations. *Review of Economics and Statistics* 78, 726–32.
- Jeganathan, P. (1996). On asymptotic inference in cointegrated time series with fractionally integrated errors. *Discussion Paper*, University of Michigan.
- Johansen, S. (1992a). A representation of vector autoregressive processes integrated of order 2. *Econometric Theory* 8, 188–202.
- Johansen, S. (1992b). Cointegration in partial systems and the efficiency of single-equation analysis. *Journal of Econometrics* 52, 389–402.
- Johansen, S. (1995). *Likelihood-Based Inference in Cointegrated Vector Auto-regressive Models*. Oxford: Oxford University Press.
- Juhl, T., (1997), Likelihood ratio tests for cointegration in the presence of multiple breaks. *Discussion Paper*, University of Pennsylvania.
- MacKinnon, J.G. (1991). Critical values for cointegration tests. In R.F. Engle and C.W.J. Granger (eds.) *Long-run Economic Relationships: Readings in Cointegration*. Oxford: Oxford University Press.
- Maddala, G.S., and I.M. Kim (1998) *Unit Roots, Cointegration and Structural Change*. Cambridge: Cambridge University Press.
- Marmol, F. (1998). Spurious regression theory with nonstationary fractionally integrated processes. *Journal of Econometrics* 84, 232–50.
- Nelson, C.R., and C.I. Plosser (1982). Trends and random walks in macroeconomic time series: Some evidence and implications. *Journal of Monetary Economics* 10, 139–62.
- Ng, S., and P. Perron (1997). Estimation and inference in nearly unbalanced nearly cointegrated systems. *Journal of Econometrics* 79, 53–81.

- Osterwald-Lenum, M. (1992). A note with quantiles of the asymptotic distribution of the maximum likelihood cointegration rank test statistics. *Oxford Bulletin of Economics and Statistics* 54, 461–72.
- Phillips, P.C.B. (1987). Time series regression with a unit root. *Econometrica* 55, 277–301.
- Phillips, P.C.B. (1991). Optimal inference in cointegrated systems. *Econometrica* 59, 283–306.
- Phillips, P.C.B., and B.E. Hansen (1990). Statistical inference in instrumental variables regression with I(1) processes. *Review of Economic Studies* 57, 99–125.
- Phillips, P.C.B., and S. Ouliaris (1990). Asymptotic properties of residual based tests for cointegration. *Econometrica* 58, 165–94.
- Quintos, C. (1994). Rank constancy tests in cointegrating regressions. Discussion Paper, Yale University.
- Reinsel, G.C., and S.K. Ahn (1992). Vector autoregressive models with unit roots and reduced rank structure: Estimation, likelihood ratio test, and forecasting. *Journal of Time Series Analysis* 13, 353–75.
- Robinson, P., and D. Marinucci (1998). Semiparametric frequency domain analysis of fractional cointegration. Discussion Paper, London School of Economics.
- Saikkonen, P. (1991). Asymptotically efficient estimation of cointegrated regressions. *Econometric Theory* 7, 1–21.
- Saikkonen, P. (1992). Estimation and testing of cointegrated systems by an autoregressive approximation. *Econometric Theory* 8, 1–27.
- Sargan, J.D. (1964). Wages and prices in the United Kingdom: A study in econometric methodology. In D.F. Hendry and K.F. Wallis (eds.) *Econometrics and Quantitative Economics*. Oxford: Blackwell.
- Stock, J.H. (1987). Asymptotic properties of least squares estimators of cointegrating vectors. *Econometrica* 55, 277–302.
- Stock, J.H., and M.W. Watson (1988). Testing for common trends. *Journal of the American Statistical Association* 83, 1097–107.
- Stock, J.H., and M.W. Watson (1993). A simple estimator of cointegrating vectors in higher order integrated systems. *Econometrica* 61, 783–820.
- Watson, M.W. (1994). Vector autoregressions and cointegration. In Engle, R.F. and D.L. McFadden (eds.) *Handbook of Econometrics IV*, New York: Elsevier.
- Wold, H. (1938). *A Study in the Analysis of Stationary Time Series*, Stockholm: Almqvist and Wiksell.

CHAPTER THIRTY - ONE

# Seasonal Nonstationarity and Near-Nonstationarity\*

*Eric Ghysels, Denise R. Osborn,  
and Paulo M.M. Rodrigues\**

## 1 INTRODUCTION

Over the last three decades there has been an increasing interest in modeling seasonality. Progressing from the traditional view that the seasonal pattern is a nuisance which needed to be removed, it is now considered to be an informative feature of economic time series which should be modeled explicitly (see for instance Ghysels (1994) for a review).

Since the seminal work by Box and Jenkins (1970), the stochastic properties of seasonality have been a major focus of research. In particular, the recognition that the seasonal behavior of economic time series may be varying and changing over time due to the presence of seasonal unit roots (see for example Canova and Ghysels (1994), Hylleberg (1994), Hylleberg, Jørgensen, and Sørensen (1993) and Osborn (1990)), has led to the development of a considerable number of testing procedures (*inter alia*, Canova and Hansen (1995), Dickey, Hasza, and Fuller (1984), Franses (1994), Hylleberg, Engle, Granger, and Yoo (1990) and Osborn, Chui, Smith, and Birchenhall (1988)).

In this chapter, we review the properties of stochastic seasonal nonstationary processes, as well as the properties of several seasonal unit root tests. More specifically, in Section 2 we analyze the characteristics of the seasonal random walk and generalize our discussion for seasonally integrated ARMA (autoregressive moving average) processes. Furthermore, we also illustrate the implications that can emerge when nonstationary stochastic seasonality is posited as deterministic.

In Section 3 we consider the asymptotic properties of the seasonal unit root test procedures proposed by Dickey *et al.* (1984) and Hylleberg *et al.* (1990). Section 4 generalizes most of the results of Section 3 by considering the behavior of the test procedures in a near seasonally integrated framework. Finally, Section 5 concludes the chapter.

To devote a chapter on the narrow subject of seasonal nonstationarity deserves some explanation. Seasonal time series appear nonstationary, a feature shared by many economic data recorded at fixed time intervals. Whether we study so-called seasonal adjusted data or raw series, the question of seasonal unit roots looms behind the univariate models being used. The standard seasonal adjustment programs like X-11 and X-12/ARIMA involve removal of seasonal unit roots (for details of seasonal adjustment programs see for instance Findley *et al.* (1998) or Ghysels and Osborn (2000)). Removing such roots may be unwarranted if they are not present and may cause statistical problems such as non-invertible MA roots (see Maravall, 1993, for further discussion). When unadjusted series are considered, then the question of seasonal unit roots is a basic issue of univariate time series model specification. Since we are particularly interested in how the asymptotic results for conventional unit root processes generalize to the seasonal context we will use the large sample distribution theory involving Brownian motion representations.

## 2 PROPERTIES OF SEASONAL UNIT ROOT PROCESSES

The case of primary interest in the context of seasonal unit roots occurs when the process  $y_t$  is nonstationary and annual differencing is required to induce stationarity. This is often referred to as *seasonal integration*. More formally:

**Definition 1.** The nonstationary stochastic process  $y_t$ , observed at  $\mathbb{S}$  equally spaced time intervals per year, is said to be seasonally integrated of order  $d$ , denoted  $y_t \sim SI(d)$ , if  $\Delta_{\mathbb{S}}^d y_t = (1 - L^{\mathbb{S}})^d y_t$  is a stationary, invertible ARMA process.

Therefore, if first order annual differencing renders  $y_t$  a stationary and invertible process, then  $y_t \sim SI(1)$ . The simplest case of such a process is the seasonal random walk, which will be the focus of analysis throughout most of this chapter. We refer to  $\mathbb{S}$  as the number of seasons per year for  $y_t$ .

### 2.1 The seasonal random walk

The seasonal random walk is a seasonal autoregressive process of order 1, or SAR(1), such that

$$y_t = y_{t-\mathbb{S}} + \varepsilon_t, \quad t = 1, 2, \dots, T \tag{31.1}$$

with  $\varepsilon_t \sim iid(0, \sigma^2)$ . Denoting the season in which observation  $t$  falls as  $s_t$ , with  $s_t = 1 + (t - 1) \bmod \mathbb{S}$ , backward substitution for lagged  $y_t$  in this process implies that

$$y_t = y_{s_t-\mathbb{S}} + \sum_{j=0}^{n_t-1} \varepsilon_{t-\mathbb{S}j}, \quad (31.2)$$

where  $n_t = 1 + [(t - 1)/\mathbb{S}]$  and  $[.]$  represents the greatest integer less or equal to  $(t - 1)/\mathbb{S}$ . As noted by Dickey *et al.* (1984) and emphasized by Osborn (1993), the random walk in this case is defined in terms of the disturbances for the specific season  $s_t$  only, with the summation over the current disturbance  $\varepsilon_t$  and the disturbance for this season in the  $n_t - 1$  previous years of the observation period. The term  $y_{s_t-\mathbb{S}} = y_{t-n_t\mathbb{S}}$  refers to the appropriate starting value for the process. Equation (31.1) is, of course, a generalization of the conventional nonseasonal random walk. Note that the unconditional mean of  $y_t$  from (31.2) is

$$E(y_t) = E(y_{s_t-\mathbb{S}}). \quad (31.3)$$

Thus, although the process (31.1) does not explicitly contain deterministic seasonal effects, these are implicitly included when  $E(y_{s_t-\mathbb{S}})$  is nonzero and varies over  $s_t = 1, \dots, \mathbb{S}$ . In their analysis of seasonal unit roots, Dickey *et al.* (1984) separate the  $y_t$  corresponding to each of the  $\mathbb{S}$  seasons into distinct series. Notationally, this is conveniently achieved using two subscripts, the first referring to the season and the second to the year. Then

$$y_t = y_{s+\mathbb{S}(n-1)} = y_{sn}, \quad (31.4)$$

where  $s_t$  and  $n_t$  are here written as  $s$  and  $n$  for simplicity of notation. Correspondingly  $\mathbb{S}$  disturbance series can be defined as

$$\varepsilon_t = \varepsilon_{s_t+\mathbb{S}(n-1)} = \varepsilon_{sn}. \quad (31.5)$$

Using these definitions, and assuming that observations are available for precisely  $N$  ( $N = T/\mathbb{S}$ ) complete years, then (31.1) can be written as

$$y_{sn} = y_{s,0} + \sum_{j=1}^n \varepsilon_{sj} \quad s = 1, \dots, \mathbb{S} \quad \text{and} \quad n = 1, \dots, N \quad (31.6)$$

which simply defines a random walk for each season  $s = 1, \dots, \mathbb{S}$ . Because the disturbances  $\varepsilon_t$  of (31.1) are uncorrelated, the random walks defined by (31.6) for the  $\mathbb{S}$  seasons of the year are also uncorrelated. Thus, any linear combination of these processes can itself be represented as a random walk. The accumulation of disturbances allows the differences to wander far from the mean over time, giving rise to the phenomenon that “summer may become winter.”

## 2.2 More general processes

To generalize the above discussion, weakly stationary autocorrelations can be permitted in the SI(1) process. That is, (31.1) can be generalized to the seasonally integrated ARMA process:

$$\phi(L)\Delta_{\mathbb{S}}y_t = \theta(L)\varepsilon_t, \quad t = 1, 2, \dots, T, \quad (31.7)$$

where, as before,  $\varepsilon_t \sim \text{iid}(0, \sigma^2)$ , while the polynomials  $\phi(L)$  and  $\theta(L)$  in the lag operator  $L$  have all roots outside the unit circle. It is, of course, permissible that these polynomials take the multiplicative form of the seasonal ARMA model of Box and Jenkins (1970). Inverting the stationary autoregressive polynomial and defining  $z_t = \phi(L)^{-1}\theta(L)\varepsilon_t$ , we can write (31.7) as:

$$\Delta_{\mathbb{S}}y_t = z_t, \quad t = 1, \dots, T. \quad (31.8)$$

The process superficially looks like the seasonal random walk, namely (31.1). There is, however, a crucial difference in that  $z_t$  here is a stationary, invertible ARMA process. Nevertheless, performing the same substitution for lagged  $y_t$  as above leads to the corresponding result, which can be written as

$$y_{sn} = y_{s,0} + \sum_{j=1}^n z_{sj} \quad s = 1, \dots, \mathbb{S} \quad \text{and} \quad n = 1, \dots, N. \quad (31.9)$$

As in (31.6), (31.9) implies that there are  $\mathbb{S}$  distinct unit root processes, one corresponding to each of the seasons. The important distinction is that these processes in (31.9) may be autocorrelated and cross-correlated. Nevertheless, it is only the stationary components which are correlated.

Defining the observation and (weakly stationary) disturbance vectors for year  $n$  as  $\mathbf{Y}_n = (y_{1n}, \dots, y_{\mathbb{S}n})'$  and  $\mathbf{Z}_n = (z_{1n}, \dots, z_{\mathbb{S}n})'$  respectively, the vector representation of (31.9) is:

$$\Delta\mathbf{Y}_n = \mathbf{Z}_n, \quad n = 1, \dots, N. \quad (31.10)$$

The disturbances here follow a stationary vector ARMA process

$$\Phi(L)\mathbf{Z}_n = \Theta(L)\mathbf{E}_n. \quad (31.11)$$

It is sufficient to note that  $\Phi(L)$  and  $\Theta(L)$  are appropriately defined  $\mathbb{S} \times \mathbb{S}$  polynomial matrices in  $L$  with all roots outside the unit circle and  $\mathbf{E}_n = (\varepsilon_{1n}, \dots, \varepsilon_{\mathbb{S}n})'$ . The seasonal difference of (31.7) is converted to a first difference in (31.10) because  $\Delta\mathbf{Y}_n = \mathbf{Y}_n - \mathbf{Y}_{n-1}$  defines an annual (that is, seasonal) difference of the vector  $\mathbf{Y}_t$ . Now, in (31.10) we have a vector ARMA process in  $\Delta\mathbf{Y}_n$ , which is a vector ARIMA process in  $\mathbf{Y}_n$ . In the terminology of Engle and Granger (1987), the  $\mathbb{S}$  processes in the vector  $\mathbf{Y}_t$  cannot be cointegrated if this is the data generating process (DGP). Expressed in a slightly different way, if the process is written in terms of the level  $\mathbf{Y}_n$ , the vector process will contain  $\mathbb{S}$  unit roots due to the presence of the factor  $\Delta = 1 - L$  in each of the equations. Therefore, the implication drawn from the seasonal random walk of (31.1) that any linear combination of the separate seasonal series is itself an I(1) process carries over to this case too.

For the purpose of this chapter, only the simple seasonal random walk case will be considered in the subsequent analysis. It should, however, be recognized that the key results extend to more general seasonally integrated processes.

## 2.3 Asymptotic properties

Consider the DGP of the seasonal random walk with initial values  $y_{-\mathbb{S}+s} = \dots = y_0 = 0$ . Using the notation of (31.6), the following  $\mathbb{S}$  independent partial sum processes (PSPs) can be obtained:

$$S_{sn} = \sum_{j=1}^n \varepsilon_{sj} \quad s = 1, \dots, \mathbb{S}, \quad n = 1, \dots, N \quad (31.12)$$

where  $n$  represents the number of years of observations to time  $t$ . From the functional central limit theorem (FCLT) and the continuous mapping theorem (CMT) the appropriately scaled PSPs in (31.12) converge as  $N \rightarrow \infty$  to

$$\frac{1}{\sqrt{N}} S_{sn} \Rightarrow \sigma W_s(r), \quad (31.13)$$

where  $\Rightarrow$  indicates convergence in distribution, while  $W_s(r)$ ,  $s = 1, \dots, \mathbb{S}$  are independent standard Brownian motions. Furthermore, the following Lemma collecting the relevant convergence results for seasonal unit root processes of periodicity  $\mathbb{S}$  can be stated:

**Lemma 1.** Assuming that the DGP is the seasonal random walk in (31.1) with initial values equal to zero,  $\varepsilon_t \sim \text{iid}(0, \sigma^2)$  and  $T = \mathbb{S}N$ , then from the CMT, as  $T \rightarrow \infty$ ,

$$(a) \quad T^{-1/2} y_{t-k} \Rightarrow \mathbb{S}^{-1/2} \sigma L^k W_s$$

$$(b) \quad T^{-3/2} \sum_{t=1}^T y_{t-k} \Rightarrow \mathbb{S}^{-3/2} \sigma \sum_{s=1}^{\mathbb{S}} \int_0^1 W_s dr$$

$$(c) \quad T^{-2} \sum_{t=1}^T y_{t-i} y_{t-k} \Rightarrow \mathbb{S}^{-2} \sigma^2 \sum_{s=1}^{\mathbb{S}} \int_0^1 W_s (L^{k-i} W_s) dr \quad k \geq i$$

$$(d) \quad T^{-1} \sum_{t=1}^T y_{t-k} \varepsilon_t \Rightarrow \mathbb{S}^{-1} \sigma^2 \sum_{s=1}^{\mathbb{S}} \int_0^1 (L^k W_s) dW_s$$

where  $k = 1, \dots, \mathbb{S}$ ,  $W_s(r)$  ( $s = 1 + (t - 1) \bmod \mathbb{S}$ ) are independent standard Brownian motions,  $L$  is the lag operator which shifts the Brownian motions between seasons ( $L^k W_s = W_{s-k}$  with  $W_{s-k} = W_{\mathbb{S}+s-k}$  for  $s - k \leq 0$ ) and  $W_s = W_s(r)$  for simplicity of notation.

It is important to note the circular property regarding the rotation of the  $W_k$ , so that after  $\mathbb{S}$  lags of  $y_t$  the same sum of  $\mathbb{S}$  integrals emerges. The Lemma is established in Osborn and Rodrigues (1998).

## 2.4 Deterministic seasonality

A common practice is to attempt the removal of seasonal patterns via seasonal dummy variables (see, for example, Barsky and Miron, 1989; Beaulieu and Miron, 1991; Osborn, 1990). The interpretation of the seasonal dummy approach is that seasonality is essentially deterministic so that the series is stationary around seasonally varying means. The simplest deterministic seasonal model is

$$y_t = \sum_{s=1}^S \delta_{st} m_s + \varepsilon_t \quad (31.14)$$

where  $\delta_{st}$  is the seasonal dummy variable which takes the value 1 when  $t$  falls in season  $s$  and  $\varepsilon_t \sim \text{iid}(0, \sigma^2)$ . Typically,  $y_t$  is a first difference series in order to account for the zero frequency unit root commonly found in economic time series. When a model like (31.14) is used, the coefficient of determination ( $R^2$ ) is often computed as a measure of the strength of the seasonal pattern. However, as Abeysinghe (1991, 1994) and Franses, Hylleberg, and Lee (1995) indicate, the presence of seasonal unit roots in the DGP will have important consequences for  $R^2$ .

To illustrate this issue, take the seasonal random walk of (31.1) as the DGP and assume that (31.14) is used to model the seasonal pattern. As is well known, the OLS estimates of  $m_s$ ,  $s = 1, \dots, S$  are simply the mean values of  $y_t$  in each season. Thus, using the notation of (31.4),

$$\hat{m}_s = \frac{1}{N} \sum_{t=1}^T \delta_{st} y_t = \frac{1}{N} \sum_{t=1}^N y_{sn} \quad (31.15)$$

where (as before)  $T$  and  $N$  are the total number of observations and the total number of complete years of observations available, respectively, and it is again assumed for simplicity that  $T = SN$ . As noted by Franses *et al.* (1995), the estimated seasonal intercepts diverge under the seasonal random walk DGP. In particular, the appropriately scaled  $\hat{m}_s$  converges to a normal random variable

$$N^{-1/2} \hat{m}_s = N^{-3/2} \sum_{t=1}^T \delta_{st} y_t \Rightarrow \sigma \int_0^1 W_s(r) dr = N(0, \sigma^2/3), \quad s = 1, \dots, S \quad (31.16)$$

where the latter follows from Banerjee *et al.* (1993, pp. 43–5) who show that  $\int_0^1 W_s(r) dr = N(0, 1/3)$ . For this DGP, the  $R^2$  from (31.14) has the non-degenerate asymptotic distribution,<sup>1</sup>

$$R^2 = \frac{\sum_{t=1}^T (\hat{y}_t - \bar{y})^2}{\sum_{t=1}^T (y_t - \bar{y})^2} \Rightarrow \frac{\sum_{s=1}^S \left( \int_0^1 W_s(r) dr \right)^2 - \frac{1}{S} \left[ \int_0^1 \left( \sum_{s=1}^S W_s(r) \right) dr \right]^2}{\sum_{s=1}^S \int_0^1 W_s^2(r) dr - \frac{1}{S} \left[ \int_0^1 \left( \sum_{s=1}^S W_s(r) \right) dr \right]^2}. \quad (31.17)$$

Consequently, high values for this statistic are to be anticipated, as concluded by Franses *et al.* These are spurious in the sense that the DGP contains no deterministic seasonality since  $E(y_t) = 0$  when the starting values for (31.1) are zero. Hence a high value of  $R^2$  when (31.14) is estimated does not constitute evidence in favor of deterministic seasonality.

### 3 TESTING THE SEASONAL UNIT ROOT NULL HYPOTHESIS

In this section we discuss the test procedures proposed by Dickey *et al.* (1984) and Hylleberg, Engle, Granger, and Yoo (HEGY) (1990) to test the null hypothesis of seasonal integration. It should be noted that while there are a large number of seasonal unit root tests available (see, for example, Rodrigues (1998) for an extensive survey), casual observation of the literature shows that the HEGY test is the most frequently used procedure in empirical work. For simplicity of presentation, throughout this section we assume that augmentation of the test regression to account for autocorrelation is unnecessary and that presample starting values for the DGP are equal to zero.

#### 3.1 The Dickey–Hasza–Fuller test

The first test of the null hypothesis  $y_t \sim SI(1)$  was proposed by Dickey, Hasza, and Fuller (DHF) (1984), as a direct generalization of the test proposed by Dickey and Fuller (1979) for a nonseasonal AR(1) process. Assuming that the process is known to be a SAR(1), then the DHF test can be parameterized as

$$\Delta_{\mathbb{S}} y_t = \alpha_{\mathbb{S}} y_{t-\mathbb{S}} + \varepsilon_t. \quad (31.18)$$

The null hypothesis of seasonal integration corresponds to  $\alpha_{\mathbb{S}} = 0$ , while the alternative of a stationary stochastic seasonal process implies  $\alpha_{\mathbb{S}} < 0$ . The appropriately scaled least squares bias obtained from the estimation of  $\alpha_{\mathbb{S}}$  under the null hypothesis is

$$T\hat{\alpha}_{\mathbb{S}} = \frac{\frac{1}{T} \sum_{t=1}^T y_{t-\mathbb{S}} \varepsilon_t}{\frac{1}{T^2} \sum_{t=1}^T y_{t-\mathbb{S}}^2} \quad (31.19)$$

and the associated  $t$ -statistic is

$$t_{\hat{\alpha}_{\mathbb{S}}} = \frac{\frac{1}{T} \sum_{t=1}^T y_{t-\mathbb{S}} \varepsilon_t}{\tilde{\sigma} \left[ \frac{1}{T^2} \sum_{t=1}^T y_{t-\mathbb{S}}^2 \right]^{\frac{1}{2}}}, \quad (31.20)$$

where  $\tilde{\sigma}$  is the usual degrees of freedom corrected estimator of  $\sigma$ . Similarly to the usual Dickey–Fuller approach, the test is typically implemented using (31.20). Using the results in (c) and (d) of Lemma 1, it is straightforward to establish that (31.19) and (31.20) converge to

$$\frac{T}{\mathbb{S}} \hat{\alpha}_{\mathbb{S}} \Rightarrow \frac{\sum_{s=1}^{\mathbb{S}} \int_0^1 W_s(r) dW_s(r)}{\sum_{s=1}^{\mathbb{S}} \int_0^1 W_s^2(r) dr} \quad (31.21)$$

and

$$t_{\hat{\alpha}_{\mathbb{S}}} \Rightarrow \frac{\sum_{s=1}^{\mathbb{S}} \int_0^1 W_s(r) dW_s(r)}{\left[ \sum_{s=1}^{\mathbb{S}} \int_0^1 W_s^2(r) dr \right]^{\frac{1}{2}}}, \quad (31.22)$$

respectively. Note that  $\tilde{\sigma}^2 \xrightarrow{P} \sigma^2$ .

The asymptotic distribution of the DHF statistic given by (31.22) is non-standard, but is of similar type to the Dickey–Fuller  $t$ -distribution. Indeed, it is precisely the Dickey–Fuller  $t$ -distribution in the special case  $\mathbb{S} = 1$ , when the test regression (31.18) is the usual Dickey–Fuller test regression for a conventional random walk. It can also be seen from (31.22) that the distribution for the DHF  $t$ -statistic depends on  $\mathbb{S}$ , that is on the frequency with which observations are made within each year. On the basis of Monte Carlo simulations, DHF tabulated critical values of  $\frac{T}{\mathbb{S}} \hat{\alpha}_{\mathbb{S}}$  and  $t_{\hat{\alpha}_{\mathbb{S}}}$  for various  $T$  and  $\mathbb{S}$ . Note that the limit distributions presented as functions of Brownian motions can also be found in Chan (1989), Boswijk and Franses (1996) and more recently in Osborn and Rodrigues (1998). To explore the dependence on  $\mathbb{S}$  a little further, note first that

$$\int_0^1 W_s(r) dW_s(r) = \frac{1}{2} \{ [W_s(1)]^2 - 1 \}, \quad (31.23)$$

where  $[W_s(1)]^2$  is  $\chi^2(1)$  (see, for example, Banerjee *et al.*, 1993, p. 91). The numerator of (31.22) involves the sum of  $\mathbb{S}$  such terms which are mutually independent and hence

$$\begin{aligned} \sum_{s=1}^{\mathbb{S}} \int_0^1 W_s(r) dW_s(r) &= \frac{1}{2} \sum_{s=1}^{\mathbb{S}} \{ [W_s(1)]^2 - 1 \} \\ &= \frac{1}{2} \{ \chi^2(\mathbb{S}) - \mathbb{S} \}, \end{aligned} \quad (31.24)$$

which is half the difference between a  $\chi^2(\mathbb{S})$  statistic and its mean of  $\mathbb{S}$ . It is well known that the Dickey–Fuller  $t$ -statistic is not symmetric about zero. Indeed,

Fuller (1996, p. 549) comments that asymptotically the probability of (in our notation)  $\hat{\alpha}_1 < 0$  is 0.68 for the nonseasonal random walk because  $\Pr[\chi^2(1) < 1] = 0.68$ . In terms of (31.22), the denominator is always positive and hence  $\Pr[\chi^2(\mathbb{S}) < \mathbb{S}]$  dictates the probability that  $t_{\hat{\alpha}_s}$  is negative. With a seasonal random walk and quarterly data,  $\Pr[\chi^2(4) < 4] = 0.59$ , while in the monthly case  $\Pr[\chi^2(12) < 12] = 0.55$ . Therefore, the preponderance of negative test statistics is expected to decrease as  $\mathbb{S}$  increases. As seen from the percentiles tabulated by DHF, the dispersion of  $t_{\hat{\alpha}_s}$  is effectively invariant to  $\mathbb{S}$ , so that the principal effect of an increasing frequency of observation is a reduction in the asymmetry of this test statistic around zero.

### 3.2 Testing complex unit roots

Before proceeding to the examination of the procedure proposed by Hylleberg *et al.* (1990) it will be useful to consider some of the issues related to testing complex unit roots, because these are an intrinsic part of any SI(1) process.

The simplest process which contains a pair of complex unit roots is

$$y_t = -y_{t-2} + u_t, \quad (31.25)$$

with  $u_t \sim \text{iid}(0, \sigma^2)$ . This process has  $\mathbb{S} = 2$  and, using the notation identifying the season  $s$  and year  $n$ , it can be equivalently written as

$$y_{sn} = -y_{s,n-1} + u_{sn} \quad s = 1, 2. \quad (31.26)$$

Notice that the seasonal patterns reverse each year. Due to this alternating pattern, and assuming  $y_0 = y_{-1} = 0$ , it can be seen that

$$y_t = S_{sn}^* = \sum_{i=0}^{n-1} (-1)^i u_{s,n-i} = -S_{s,n-1}^* + u_{sn}, \quad (31.27)$$

where, in this case,  $n = \left[ \frac{t+1}{2} \right]$ . Note that  $S_{sn}^*$  ( $s = 1, 2$ ) are independent processes, one corresponding to each of the two seasons of the year. Nevertheless, the nature of the seasonality implied by (31.25) is not of the conventional type in that  $S_{sj}^*$  (for given  $s$ ) tends to oscillate as  $j$  increases. Moreover, it can be observed from (31.27) that aggregation of the process over full cycles of two years annihilates the nonstationarity as  $S_{s,n-1}^* + S_{sn}^* = u_{sn}$ . To relate these  $S_{sn}^*$  to the  $\mathbb{S}$  independent random walks of (31.6), let  $\varepsilon_{sj} = (-1)^j u_{sj}$  which (providing the distribution of  $u_t$  is symmetric) has identical properties. Then

$$S_{sn}^* = \begin{cases} \sum_{j=1}^n (-1)^{j+1} u_{sj} = -\sum_{j=1}^n \varepsilon_{sj} = -S_{jn}, & n \text{ odd} \\ \sum_{j=1}^n (-1)^j u_{sj} = \sum_{j=1}^n \varepsilon_{sj} = S_{jn}, & n \text{ even} \end{cases} \quad (31.28)$$

where  $S_{jn}$  is defined in (31.12). Analogously to the DHF test, the unit root process (31.25) may be tested through the  $t$ -ratio for  $\hat{\alpha}_2^*$  in

$$(1 + L^2)y_t = \alpha_2^* y_{t-2} + u_t. \quad (31.29)$$

The null hypothesis is  $\alpha_2^* = 0$  with the alternative of stationarity implying  $\alpha_2^* > 0$ . Then, assuming  $T = 2N$ , under the null hypothesis

$$T\hat{\alpha}_2^* = \frac{T^{-1} \sum_{t=1}^T y_{t-2} u_t}{T^{-2} \sum_{t=1}^T y_{t-2}^2} = \frac{(2N)^{-1} \sum_{s=1}^2 \sum_{j=1}^N S_{s,j-1}^* (S_{s,j}^* + S_{s,j-1}^*)}{(2N)^{-2} \sum_{s=1}^2 \sum_{j=1}^N (S_{s,j-1}^*)^2}. \quad (31.30)$$

and

$$t(\hat{\alpha}_2^*) = \frac{\sum_{t=1}^T y_{t-2} u_t}{\tilde{\sigma} \left[ \sum_{t=1}^T y_{t-2}^2 \right]^{\frac{1}{2}}} = \frac{(2N)^{-1} \sum_{s=1}^2 \sum_{j=1}^N S_{s,j-1}^* (S_{s,j}^* + S_{s,j-1}^*)}{\tilde{\sigma} \left[ (2N)^{-2} \sum_{s=1}^2 \sum_{j=1}^N (S_{s,j-1}^*)^2 \right]^{\frac{1}{2}}}. \quad (31.31)$$

If, for further expositional clarity, we assume that  $N$  is even, then using (31.28), we have

$$\begin{aligned} \sum_{j=1}^N S_{s,j-1}^* (S_{s,j}^* + S_{s,j-1}^*) &= \sum_{i=1}^{N/2} [S_{s,2i-2}^* (S_{s,2i-1}^* + S_{s,2i-2}^*) + S_{s,2i-1}^* (S_{s,2i}^* + S_{s,2i-1}^*)] \\ &= \sum_{i=1}^{N/2} [S_{s,2i-2}^* (-S_{s,2i-1} + S_{s,2i-2}) - S_{s,2i-1}^* (S_{s,2i} - S_{s,2i-1})] \\ &= - \sum_{j=1}^N S_{s,j-1}^* (S_{s,j}^* - S_{s,j-1}^*). \end{aligned}$$

Thus, there is a “mirror image” relationship between the numerator of (31.30) and (31.31) compared with that of (31.19) and (31.20) with  $\mathbb{S} = 2$ . The corresponding denominators are identical as  $(S_{sj}^*)^2 = S_{sj}^2$ . Thus, by applying similar arguments as in the proof of Lemma 1:

$$\frac{T}{2} \hat{\alpha}_2^* \Rightarrow - \frac{\sum_{s=1}^2 \int_0^1 W_s(r) dW_s(r)}{\sum_{s=1}^2 \int_0^1 [W_s(r)]^2 dr} \quad (31.32)$$

and

$$t_{\hat{\alpha}_2^*} \Rightarrow -\frac{\sum_{s=1}^2 \int_0^1 W_s(r) dW_s(r)}{\left\{ \sum_{s=1}^2 \int_0^1 [W_s(r)]^2 dr \right\}^{\frac{1}{2}}}, \quad (31.33)$$

which can be compared with (31.21) and (31.22) respectively. This mirror image property of these test statistics has also been shown by Chan and Wei (1988) and Fuller (1996, pp. 553–4). One important practical consequence of (31.33) is that with a simple change of sign, the DHF tables with  $\mathbb{S} = 2$  apply to the case of testing  $\alpha_2^* = 0$  in (31.29). Under the assumed DGP (31.25), we may also consider testing the null hypothesis  $\alpha_1^* = 0$  against the alternative  $\alpha_1^* \neq 0$  in

$$(1 + L^2)y_t = \alpha_1^* y_{t-1} + u_t. \quad (31.34)$$

The test here is not, strictly speaking, a unit root test, since the unit coefficient on  $L^2$  in (31.34) implies that the process contains two roots of modulus one, irrespective of the value of  $\alpha_1^*$ . Rather, the test of  $\alpha_1^* = 0$  is a test of the null hypothesis that the process contains a half-cycle every  $\mathbb{S} = 2$  periods, and hence a full cycle every four periods. The appropriate alternative hypothesis is, therefore, two-sided. For this test regression,

$$T\hat{\alpha}_1^* = \frac{T^{-1} \sum_{t=1}^T y_{t-1} u_t}{T^{-2} \sum_{t=1}^T y_{t-1}^2}.$$

Again referring to (31.27) and (31.28), we can see that

$$T\hat{\alpha}_1^* = \frac{(2N)^{-1} \sum_{j=1}^N [-S_{2,j-1}(S_{1,j} - S_{1,j-1}) + S_{1,j}(S_{2,j} - S_{2,j-1})]}{(2N)^{-2} \sum_{j=1}^N (S_{1,j-1}^2 + S_{2,j}^2)}. \quad (31.35)$$

Thus, (31.35) converges to,

$$\frac{T}{2}\hat{\alpha}_1^* \Rightarrow \frac{\int_0^1 W_1(r) dW_2(r) - \int_0^1 W_2(r) dW_1(r)}{\sum_{s=1}^2 \int_0^1 [W_s(r)]^2 dr}, \quad (31.36)$$

and consequently,

$$t_{\alpha_1^*} \Rightarrow \frac{\int_0^1 W_1(r)dW_2(r) - \int_0^1 W_2(r)dW_1(r)}{\left\{ \sum_{s=1}^2 \int_0^1 [W_s(r)]^2 dr \right\}^{1/2}}. \quad (31.37)$$

Indeed, the results for the distributions associated with the test statistics in (31.29) and (31.34) continue to apply for the test regression

$$(1 + L^2)y_t = \alpha_1^* y_{t-1} + \alpha_2^* y_{t-2} + \varepsilon_t \quad (31.38)$$

because the regressors  $y_{t-1}$  and  $y_{t-2}$  can be shown to be asymptotically orthogonal (see, for instance, Ahtola and Tiao (1987) or Chan and Wei (1988) for more details).

### 3.3 The Hylleberg–Engle–Granger–Yoo test

It is well known that the seasonal difference operator  $\Delta_S = 1 - L^S$  can always be factorized as

$$1 - L^S = (1 - L)(1 + L + L^2 + \dots + L^{S-1}). \quad (31.39)$$

Hence, (31.39) indicates that an SI(1) process always contains a conventional unit root and a set of  $S - 1$  seasonal unit roots. The approach suggested by Hylleberg *et al.* (1990), commonly known as HEGY, examines the validity of  $\Delta_S$  through exploiting (31.39) by testing the unit root of 1 and the  $S - 1$  separate nonstationary roots on the unit circle implied by  $1 + L + \dots + L^{S-1}$ . To see the implications of this factorization, consider the case of quarterly data ( $S = 4$ ) where

$$\begin{aligned} 1 - L^4 &= (1 - L)(1 + L + L^2 + L^3) \\ &= (1 - L)(1 + L)(1 + L^2). \end{aligned} \quad (31.40)$$

Thus,  $\Delta_4 = 1 - L^4$  has four roots on the unit circle,<sup>2</sup> namely 1 and  $-1$  which occur at the 0 and  $\pi$  frequencies respectively, and the complex pair  $\pm i$  at the frequencies  $\frac{\pi}{2}$  and  $\frac{3\pi}{2}$ . Hence, in addition to the conventional unit root, the quarterly case implies three seasonal unit roots, which are  $-1$  and the complex pair  $\pm i$ .

Corresponding to each of the three factors of (31.40), using a Lagrange approximation, HEGY suggest the following linear transformations:

$$y_{(1),t} = (1 + L)(1 + L^2)y_t = y_t + y_{t-1} + y_{t-2} + y_{t-3} \quad (31.41)$$

$$y_{(2),t} = -(1 - L)(1 + L^2)y_t = -y_t + y_{t-1} - y_{t-2} + y_{t-3} \quad (31.42)$$

$$y_{(3),t} = -(1 - L)(1 + L)y_t = -y_t + y_{t-2}. \quad (31.43)$$

By construction, each of the variables in (31.41) to (31.43) accepts all the factors of  $\Delta_4$  except one. That is,  $y_{(1),t}$  assumes the factors  $(1 + L)$  and  $(1 + L^2)$ ,  $y_{(2),t}$  assumes  $(1 - L)$  and  $(1 + L^2)$ , while  $y_{(3),t}$  assumes  $(1 - L)$  and  $(1 + L)$ . The test regression for quarterly data suggested by HEGY has the form:

$$\Delta_4 y_t = \pi_1 y_{(1),t-1} + \pi_2 y_{(2),t-1} + \pi_3 y_{(3),t-2} + \pi_4 y_{(3),t-1} + \varepsilon_t, \quad t = 1, 2, \dots, T \quad (31.44)$$

where  $y_{(1),t}$ ,  $y_{(2),t}$ , and  $y_{(3),t}$  are defined in (31.41), (31.42), and (31.43), respectively. Note that these regressors are asymptotically orthogonal by construction. The two lags of  $y_{(3),t}$  arise because the pair of complex roots  $\pm i$  imply two restrictions on a second order polynomial  $1 + \phi_1 L + \phi_2 L^2$ , namely  $\phi_1 = 0$  and  $\phi_2 = 1$  (see Section 3.2). The overall null hypothesis  $y_t \sim SI(1)$  implies  $\pi_1 = \pi_2 = \pi_3 = \pi_4 = 0$  and hence  $\Delta_4 y_t = \varepsilon_t$  as for the DHF test. The HEGY regression (31.44) and the associated asymptotic distributions can be motivated by considering the three factors of  $\Delta_4 = (1 - L)(1 + L)(1 + L^2)$  one by one. Through the variable  $y_{(1),t}$ , we may consider the DGP

$$y_{(1),t} = y_{(1),t-1} + \varepsilon_t. \quad (31.45)$$

Therefore, when  $y_t$  is generated from a seasonal random walk,  $y_{(1),t}$  has the properties of a conventional random walk process and hence, with initial values equal to zero,

$$y_{(1),t} = \sum_{j=0}^{t-1} \varepsilon_{t-j}. \quad (31.46)$$

Since  $\Delta_1 y_{(1),t} = \Delta_4 y_t$ , the Dickey–Fuller test regression for the DGP (31.45) is

$$\Delta_4 y_t = \pi_1 y_{(1),t-1} + \varepsilon_t, \quad (31.47)$$

where we test  $\pi_1 = 0$  against  $\pi_1 < 0$ . Considering

$$T \hat{\pi}_1 = \frac{T^{-1} \sum_{t=1}^T y_{(1),t-1} \varepsilon_t}{T^{-2} \sum_{t=1}^T y_{(1),t-1}^2} = \frac{T^{-1} \sum_{t=1}^T (y_{t-1} + y_{t-2} + y_{t-3} + y_{t-4}) \varepsilon_t}{T^{-2} \sum_{t=1}^T (y_{t-1} + y_{t-2} + y_{t-3} + y_{t-4})^2}, \quad (31.48)$$

then from Lemma 1 and (31.13) it can be observed that under the seasonal random walk null hypothesis

$$T^{-1} \sum_{t=1}^T (y_{t-1} + y_{t-2} + y_{t-3} + y_{t-4}) \varepsilon_t \Rightarrow \frac{\sigma^2}{4} \left\{ \int_0^1 W_{(1)}(r) dW_{(1)}(r) \right\} \quad (31.49)$$

and

$$T^{-2} \sum_{t=1}^T (y_{t-1} + y_{t-2} + y_{t-3} + y_{t-4})^2 \Rightarrow \frac{\sigma^2}{16} \int_0^1 4W_{(1)}^2(r) dr, \quad (31.50)$$

where  $W_{(1)}(r) = \sum_{s=1}^4 W_s(r)$ . Substituting (31.49) and (31.50) into (31.48) gives:

$$T\hat{\pi}_1 \Rightarrow \frac{\int_0^1 W_{(1)}(r) dW_{(1)}(r)}{\int_0^1 [W_{(1)}(r)]^2 dr}. \quad (31.51)$$

The associated  $t$ -statistic, which is commonly used to test for the zero frequency unit root, can be expressed as

$$t_{\hat{\pi}_1} \Rightarrow \frac{\int_0^1 W_1^*(r) dW_1^*(r)}{\left\{ \int_0^1 [W_1^*(r)]^2 dr \right\}^{1/2}} \quad (31.52)$$

where  $W_1^*(r) = W_{(1)}(r)/2$ . Division by 2 is undertaken here so that  $W_1^*(r)$  is standard Brownian motion, whereas  $W_{(1)}(r)$  is not. Therefore, (31.52) is the conventional Dickey–Fuller  $t$ -distribution, tabulated by Fuller (1996).

Similarly, based on (31.42), the seasonal random walk (31.1) implies

$$-(1 + L)y_{(2),t} = \varepsilon_t. \quad (31.53)$$

Notice the “bounce back” effect in (31.53) which implies a half cycle for  $y_{(2),t}$  every period and hence a full cycle every two periods. Also note that (31.53) effectively has the same form as (31.26). Testing the root of  $-1$  implied by (31.53) leads to a test of  $\phi_2 = 1$  against  $\phi_2 < 1$  in

$$-(1 + \phi_2 L)y_{(2),t} = \varepsilon_t.$$

Equivalently, defining  $\pi_2 = \phi_2 - 1$  and again using (31.42) yields:

$$\Delta_4 y_t = \pi_2 y_{(2),t-1} + \varepsilon_t, \quad (31.54)$$

with null and alternative hypotheses  $\pi_2 = 0$  and  $\pi_2 < 0$ , respectively. Under the null hypothesis, and using analogous reasoning to Section 3.2 combined with Lemma 1, we obtain

$$T\hat{\pi}_2 \Rightarrow \frac{\int_0^1 W_{(2)}(r)dW_{(2)}(r)}{\int_0^1 [W_{(2)}(r)]^2 dr} \quad (31.55)$$

and

$$t_{\hat{\pi}_2} \Rightarrow \frac{\int_0^1 W_2^*(r)dW_2^*(r)}{\left\{ \int_0^1 [W_2^*(r)]^2 dr \right\}^{1/2}}, \quad (31.56)$$

where the Brownian motion  $W_{(2)}(r) = W_1(r) - W_2(r) + W_3(r) - W_4(r)$  is standardized as  $W_2^*(r) = W_{(2)}(r)/2$ . Like (31.52), (31.56) is the conventional Dickey–Fuller distribution tabulated by Fuller (1996). It is important to note that, as indicated by Chan and Wei (1988) and Fuller (1996), the distributions of the least squares bias and corresponding  $t$ -statistic when the DGP is an AR(1) with a  $-1$  root are the “mirror image” of those obtained when testing the conventional random walk. However, in (31.55) and (31.56), this mirror image is incorporated through the design of the HEGY test regression in that the linear transformation of  $y_{(2),t}$  is defined with a minus sign as  $-(1 - L)(1 + L^2)$ .

Finally, from (31.43) it follows that  $y_t \sim SI(1)$  as in (31.1) with  $S = 4$  implies

$$-(1 + L^2)y_{(3),t} = \varepsilon_t. \quad (31.57)$$

This process implies a “bounce back” after two periods and a full cycle after four. This process has the complex root form identical to (31.25). Hence, the results presented for that process carry over directly for this case. Noting again that  $-(1 + L^2)y_{(3),t} = \Delta_4 y_t$ , we can test  $\phi_3 = 1$  and  $\phi_4 = 0$  in

$$-(1 + \phi_4 L + \phi_3 L^2)y_{(3),t} = \varepsilon_t$$

through the regression

$$\Delta_4 y_t = \pi_3 y_{(3),t-2} + \pi_4 y_{(3),t-1} + \varepsilon_t \quad (31.58)$$

with  $\pi_3 = \phi_3 - 1$  and  $\pi_4 = -\phi_4$ . Testing against stationarity implies null and alternative hypotheses of  $\pi_3 = 0$  and  $\pi_3 < 0$ . However, while  $\pi_4 = 0$  is also indicated under the null hypothesis, the alternative here is  $\pi_4 \neq 0$ . The reasoning for this two-sided alternative is precisely that for the test regression (31.34) and (31.58) has the same form as (31.38). Therefore, using similar arguments to Section 3.2, and noting that the “mirror image” property discussed there is incorporated through the minus sign in the definition of  $y_{(3),t}$ , it can be seen that

$$t_{\hat{\pi}_3} \Rightarrow \frac{\int_0^1 W_3^*(r) dW_3^*(r) + \int_0^1 W_4^*(r) dW_4^*(r)}{\left\{ \int_0^1 [W_3^*(r)]^2 dr + \int_0^1 [W_4^*(r)]^2 dr \right\}^{1/2}} \quad (31.59)$$

and

$$t_{\hat{\pi}_4} \Rightarrow \frac{\int_0^1 W_4^*(r) dW_3^*(r) - \int_0^1 W_3^*(r) dW_4^*(r)}{\left\{ \int_0^1 [W_3^*(r)]^2 dr + \int_0^1 [W_4^*(r)]^2 dr \right\}^{1/2}}, \quad (31.60)$$

where  $W_3^*(r) = [W_1(r) - W_3(r)]/\sqrt{2}$  and  $W_4^*(r) = [W_2(r) - W_4(r)]/\sqrt{2}$  are independent standard Brownian motions. Note that the least squares bias  $T\hat{\pi}_3$  and  $T\hat{\pi}_4$  can also be obtained from (31.32) and (31.36).

HEGY suggest that  $\pi_3$  and  $\pi_4$  might be jointly tested, since they are both associated with the pair of nonstationary complex roots  $\pm i$ . Such joint testing might be accomplished by computing  $F(\hat{\pi}_3 \cap \hat{\pi}_4)$  as for a standard F-test, although the distribution will not, of course, be the standard F-distribution. Engle, Granger, Hylleberg, and Lee (1993) show that the limiting distribution of  $F(\hat{\pi}_3 \cap \hat{\pi}_4)$  is identical to that of  $\frac{1}{2}[t_{\hat{\pi}_3}^2 + t_{\hat{\pi}_4}^2]$ , where the two individual components are given in (31.59) and (31.60). More details can be found in Smith and Taylor (1998) or Osborn and Rodrigues (1998).

Due to the asymptotic orthogonality of the regressors in (31.47), (31.54) and (31.58), these can be combined into the single test regression (31.44) without any effect on the asymptotic properties of the coefficient estimators.

### 3.4 Extensions to the HEGY approach

Ghysels, Lee, and Noh (1994), or GLN, consider further the asymptotic distribution of the HEGY test statistics for quarterly data and present some extensions. In particular, they propose the joint test statistics  $F(\hat{\pi}_1 \cap \hat{\pi}_2 \cap \hat{\pi}_3 \cap \hat{\pi}_4)$  and  $F(\hat{\pi}_2 \cap \hat{\pi}_3 \cap \hat{\pi}_4)$ , the former being an overall test of the null hypothesis  $y_t \sim SI(1)$  and the latter a joint test of the seasonal unit roots implied by the summation operator  $1 + L + L^2 + L^3$ . Due to the two-sided nature of all F-tests, the alternative hypothesis in each case is that one or more of the unit root restrictions is not valid. Thus, in particular, these tests should not be interpreted as testing seasonal integration against stationarity for the process. From the asymptotic independence of  $t_{\hat{\pi}_i}$ ,  $i = 1, \dots, 4$ , it follows that  $F(\hat{\pi}_1 \cap \hat{\pi}_2 \cap \hat{\pi}_3 \cap \hat{\pi}_4)$  has the same asymptotic distribution as  $\frac{1}{4}\sum_{i=1}^4(t_{\hat{\pi}_i})^2$ , where the individual asymptotic distributions are given by (31.52), (31.56), (31.59) and (31.60). Hence,  $F(\hat{\pi}_1 \cap \hat{\pi}_2 \cap \hat{\pi}_3 \cap \hat{\pi}_4)$  is asymptotically distributed as the simple average of the squares of each of two Dickey–Fuller distributions, a DHF distribution with  $S = 2$  and (31.60). It is straightforward to see that a similar

expression results for  $F(\hat{\pi}_2 \cap \hat{\pi}_3 \cap \hat{\pi}_4)$ , which is a simple average of the squares of a Dickey–Fuller distribution, a DHF distribution with  $\mathbb{S} = 2$  and (31.60).

GLN also observe that the usual test procedure of Dickey and Fuller (DF) (1979) can validly be applied in the presence of seasonal unit roots. However, this validity only applies if the regression contains sufficient augmentation. The essential reason derives from (31.39), so that the SI(1) process  $\Delta_{\mathbb{S}}y_t = \varepsilon_t$  can be written as

$$\Delta_1 y_t = \alpha_1 y_{t-1} + \phi_1 \Delta_1 y_{t-1} + \dots + \phi_{\mathbb{S}-1} \Delta_1 y_{t-\mathbb{S}+1} + \varepsilon_t \quad (31.61)$$

with  $\alpha_1 = 0$  and  $\phi_1 = \dots = \phi_{\mathbb{S}-1} = -1$ . With (31.61) applied as a unit root test regression,  $t_{\hat{\alpha}_1}$  asymptotically follows the usual DF distribution, as given in (31.52). See Ghysels, Lee, and Siklos (1993), Ghysels *et al.* (1994) and Rodrigues (2000a) for a more detailed discussion.

Beaulieu and Miron (1993) and Franses (1991) develop the HEGY approach for the case of monthly data.<sup>3</sup> This requires the construction of at least seven transformed variables, analogous to  $y_{(1),t}$ ,  $y_{(2),t}$ , and  $y_{(3),t}$  used in (31.41) to (31.43), and the estimation of twelve coefficients  $\pi_i$  ( $i = 1, \dots, 12$ ). Beaulieu and Miron present the asymptotic distributions, noting that the  $t$ -type statistics corresponding to the two real roots of +1 and -1 each have the usual Dickey–Fuller form, while the remaining coefficients correspond to pairs of complex roots. In the Beaulieu and Miron parameterization, each of the five pairs of complex roots leads to a  $t_{\hat{\pi}_i}$  with a DHF distribution (again with  $\mathbb{S} = 2$ ) and a  $t_{\hat{\pi}_i}$  with the distribution (31.60).

Although both Beaulieu and Miron (1993) and Franses (1991) discuss the use of joint F-type statistics for the two coefficients corresponding to a pair of complex roots, neither considers the use of the F-tests as in Ghysels *et al.* (1994) to test the overall  $\Delta_{12}$  filter or the eleven seasonal unit roots. Taylor (1998) supplies critical values for these overall joint tests in the monthly case.

Kunst (1997) takes an apparently different approach to testing seasonal integration from that of HEGY, but his approach is easily seen to be related to that of DHF. Kunst is primarily concerned with the distribution of a joint test statistic. Although apparently overlooked by Kunst, it is easy to see that his joint test is identical to the joint test of all coefficients which arises in the HEGY framework. Ghysels and Osborn (2000) and Osborn and Rodrigues (1998) discuss in detail the equivalence between the tests proposed by Kunst and prior existing tests. Also, comparison of the percentiles tabulated by Ghysels *et al.* (1994) and Kunst with  $\mathbb{S} = 4$  are effectively identical.<sup>4</sup> Naturally, these results carry over to the monthly case, with the HEGY F-statistic of Taylor (1998) being equivalent to that of Kunst with  $\mathbb{S} = 12$ . Kunst does, however, provide critical values for other cases, including  $\mathbb{S} = 7$ , which is relevant for testing the null hypothesis that a daily series is seasonally integrated at a period of one week.

### 3.5 Multiple tests and levels of significance

It is notable that many tests of the seasonal unit root null hypothesis involve tests on multiple coefficients. In particular, for the application of the HEGY test (31.44),

Hylleberg *et al.* (1990) recommend that one-sided tests of  $\pi_1$  and  $\pi_2$  should be applied, with  $(\pi_3, \pi_4)$  either tested sequentially or jointly. The rationale for applying one-sided tests for  $\pi_1$ ,  $\pi_2$ , and  $\pi_3$  is that it permits a test against stationarity, which is not the case when a joint F-type test is applied. Thus, the null hypothesis is rejected against stationarity only if the null hypothesis is rejected for each of these three tests. Many applied researchers have followed HEGY's advice, apparently failing to recognize the implications of this strategy for the overall level of significance for the implied joint test of  $\pi_1 = \pi_2 = \pi_3 = \pi_4 = 0$ .

Let us assume that separate tests are applied to  $\pi_1$  and  $\pi_2$ , with a joint test applied to  $(\pi_3, \pi_4)$ , with each of these three tests applied at the same level of significance,  $\alpha$ . Conveniently, these tests are mutually independent, due to the asymptotic orthogonality of the regressors, as discussed in Section 3.3. Therefore, the overall probability of not rejecting the SI(1) null hypothesis when it is true is  $(1 - \alpha)^3 \approx 1 - 3\alpha$  for  $\alpha$  small. Thus, with  $\alpha = .05$ , the implied level of significance for the overall test is  $1 - .95^3 = .14$ , or approximately three times that of each individual test. With monthly data the issue is even more important of course.

In conclusion, the impact of multiple tests must be borne in mind when applying seasonal unit root tests. To date, however, these issues have received relatively little attention in this literature.

#### 4 NEAR SEASONAL INTEGRATION

As noted in Section 3.1 for the DHF test,  $\Pr[t_{\hat{\alpha}_s} < 0] = \Pr[\chi^2(\mathbb{S}) < \mathbb{S}]$  seems to be converging to 1/2 as  $\mathbb{S}$  increases. However, for the periodicities typically considered this probability always exceeds 1/2. This phenomenon indicates that a standard normal distribution may not be a satisfactory approximation when the characteristic root is close to 1 and the sample size is moderate, as Chan and Wei (1987) point out. It is also a well established fact that the power of unit root tests is quite poor when the parameter of interest is in the neighborhood of unity (see, for example, Evans and Savin (1981, 1984) and Perron (1989)). This suggests a distributional gap between the standard distribution typically assumed under stationarity and the function of Brownian motions obtained when the DGP is a random walk. To close this gap, a new class of models have been proposed, which allow the characteristic root of a process to be in the neighborhood of unity. This type of process is often called near integrated. Important work concerning near integration in a conventional AR(1) process includes Bobkoski (1983), Cavanagh (1986), Phillips (1987, 1988), Chan and Wei (1987), Chan (1988, 1989), and Nabeya and Perron (1994). In the exposition of the preceding sections, it has been assumed that the DGP is a special case of

$$y_t = \phi_{\mathbb{S}} y_{t-\mathbb{S}} + \varepsilon_t \quad (31.62)$$

with  $\phi_{\mathbb{S}} = 1$  and  $y_{-\mathbb{S}+1} = \dots = y_0 = 0$ . In this section we generalize the results by considering a class of processes characterized by an autoregressive parameter  $\phi_{\mathbb{S}}$  close to 1.

Analogously to the conventional near integrated AR(1), a noncentrality parameter  $c$  can be considered such that

$$\phi_{\mathbb{S}} = e^{c/N} \approx 1 + \frac{c}{N} = 1 + \frac{\mathbb{S}c}{T}. \quad (31.63)$$

This characterizes a near seasonally integrated process, which can be locally stationary ( $c < 0$ ), locally explosive ( $c > 0$ ), or a conventional seasonal random walk ( $c = 0$ ). This type of near seasonally integrated process has been considered by Chan (1988, 1989), Perron (1992), Rodrigues (2000b), and Tanaka (1996). Similarly to the seasonal random walk, when the DGP is given by (31.62) and (31.63), and assuming that the observations are available for exactly  $N$  ( $N = T/\mathbb{S}$ ) complete years, then

$$S_{sn} = \sum_{j=0}^{n-1} e^{\frac{jc}{N}} \varepsilon_{s,n-j} = \sum_{j=1}^n e^{(n-j)\frac{c}{N}} \varepsilon_{s,j} \quad s = 1, \dots, \mathbb{S}. \quad (31.64)$$

This indicates that each season represents a near integrated process with a common noncentrality parameter  $c$  across seasons. One of the main features of a process like (31.62) with  $\phi_{\mathbb{S}} = e^{c/N}$ , is that

$$\frac{1}{N^{1/2}} y_{sn} = \frac{1}{N^{1/2}} S_{sn} \Rightarrow \sigma^2 J_{sc}(r), \quad s = 1, \dots, \mathbb{S}, \quad (31.65)$$

where  $S_{sn}$  is the PSP corresponding to season  $s$  and  $J_{sc}(r)$  is an Ornstein–Uhlenbeck process and not a Brownian motion as in the seasonal random walk case. Note that, as indicated by, for example, Phillips (1987) or Perron (1992), this diffusion process is generated by the stochastic differential equation

$$dJ_{sc}(r) = cJ_{sc}(r)dr + dW_s(r), \quad (31.66)$$

so that

$$J_{sc}(r) = W_s(r) + c \int_0^1 e^{(r-v)c} W_s(v) dv \quad (31.67)$$

and  $J_{sc}(0) = 0$ .

Applying results given by Phillips (1987), and following analogous steps to those underlying Section 3.1, yields:

$$\frac{T}{\mathbb{S}} (\hat{\phi}_{\mathbb{S}} - \phi_{\mathbb{S}}) = \frac{\sum_{s=1}^{\mathbb{S}} \int_0^1 J_{sc}(r) dW_s(r)}{\sum_{s=1}^{\mathbb{S}} \int_0^1 J_{sc}^2(r) dr}, \quad (31.68)$$

where  $J_{sc}(r)$  and  $W_s(r)$ ,  $s = 1, \dots, S$ , are independent Ornstein–Uhlenbeck processes and standard Brownian motions, respectively. Similarly, the  $t$ -statistic converges to

$$t_{(\hat{\phi}_S - \phi_S)} = \frac{\sum_{s=1}^S \int_0^1 J_{sc}(r) dW_s(r)}{\left[ \sum_{s=1}^S \int_0^1 J_{sc}^2(r) dr \right]^{\frac{1}{2}}}. \quad (31.69)$$

A more detailed analysis appears in Chan (1988, 1989), Perron (1992), Rodrigues (2000b) and Tanaka (1996). The result in (31.69) is the asymptotic power function for the DHF  $t$ -test. It is straightforward to observe that the distributions in (31.21) and (31.22) are particular cases of (31.68) and (31.69) respectively with  $c = 0$ .

The examination of the HEGY procedure in a near seasonally integrated framework is slightly more involved. As indicated by Rodrigues (2000b),  $(1 - (1 + \frac{c}{N})L^4)$  can be approximated by,

$$\left[ 1 - \left( 1 + \frac{c}{4N} + O(N^{-2}) \right) L \right] \left[ 1 + \left( 1 + \frac{c}{4N} + O(N^{-2}) \right) L \right] \times \left[ 1 + \left( 1 + \frac{c}{2N} + O(N^{-2}) \right) L^2 \right]. \quad (31.70)$$

The results provided by Jeganathan (1991), together with the orthogonality of the regressors in the HEGY test regression, yield the distributions of the HEGY statistics in the context of a near seasonally integrated process. Rodrigues (2000b) establishes the limit results for the HEGY test regression. One important result also put forward by Rodrigues (2000b) is that the distributions are still valid when we allow different noncentrality parameters for each factor in (31.70).

## 5 CONCLUSION

We have considered only the simple seasonal random walk case, which was used to present the general properties of seasonally integrated processes. It should be noted, however, that the effect of nonzero initial values and drifts on the distributions of the seasonal unit root test statistics can easily be handled substituting the standard Brownian motions by demeaned or detrended independent Brownian motions.

Among other issues not considered are the implications of autocorrelation and mean shifts for unit root tests. The first is discussed in detail in Ghysels *et al.* (1994), Hylleberg (1995), and Rodrigues and Osborn (1999). It is known that strong MA components can distort the power of these procedures. To a certain extent, however, these distortions can be corrected by augmenting the test regression with lags of the dependent variable.

The negative impact of mean shifts on the unit root test procedures, was noted by Ghysels (1991). Recently, Smith and Otero (1997) and Franses and Vogelsang

(1998) have shown, using artificial data, that the HEGY test is strongly affected by seasonal mean shifts. This led Franses and Vogelsang to adapt the HEGY test so as to allow for deterministic mean shifts (Smith and Otero also present relevant critical values for the HEGY procedure in this context).

## Notes

- \* We would like to thank three referees for their valuable comments.
- 1 The proof of the result appears in the Appendix to the companion working paper Ghysels, Osborn, and Rodrigues (1999).
- 2 Notice that the unit roots of a monthly seasonal random walk are:

$$1, -1, \pm i, -\frac{1}{2}(1 \pm \sqrt{3}i), \frac{1}{2}(1 \pm \sqrt{3}i), -\frac{1}{2}(\sqrt{3} \pm i), \frac{1}{2}(\sqrt{3} \pm i).$$

The first is, once again, the conventional nonseasonal, or zero frequency, unit root. The remaining 11 seasonal unit roots arise from the seasonal summation operator  $1 + L + L^2 + \dots + L^{11}$  and result in nonstationary cycles with a maximum duration of one year. As can be observed, this monthly case implies five pairs of complex roots on the unit circle.

- 3 The reparameterization of the regressors proposed for monthly data by Beaulieu and Miron (1993) is typically preferred because, in contrast to that of Franses (1991), the constructed variables are asymptotically orthogonal.
- 4 Once again, due to the definition of his F-type statistic, the Kunst (1997) percentiles have to be divided by 4 to be comparable with those of Ghysels *et al.* (1994). In the monthly case, the Kunst values have to be divided by 12 for comparison with Taylor (1998). Since these percentiles are obtained from Monte Carlo simulations, they will not be identical across different studies.

## References

- Abeysinghe, T. (1991). Inappropriate use of seasonal dummies in regression. *Economic Letters* 36, 175–9.
- Abeysinghe, T. (1994). Deterministic seasonal models and spurious regressions. *Journal of Econometrics* 61, 259–72.
- Ahtola, J., and G.C. Tiao (1987). Distributions of least squares estimators of autoregressive parameters for a process with complex roots on the unit circle. *Journal of Time Series Analysis* 8, 1–14.
- Banerjee, A., J. Dolado, J.W. Galbraith, and D. Hendry (1993). *Cointegration, Error-Correction and the Econometric Analysis of Nonstationary Data*. Oxford: Oxford University Press.
- Barsky, R.B., and J.A. Miron (1989). The seasonal cycle and the business cycle. *Journal of Political Economy* 97, 503–35.
- Beaulieu, J.J., and J.A. Miron (1991). The seasonal cycle in U.S. manufacturing. *Economics Letters* 37, 115–18.
- Beaulieu, J.J., and J.A. Miron (1993). Seasonal unit roots in aggregate U.S. data. *Journal of Econometrics* 55, 305–28.
- Bobkoski, M.J. (1983). Hypothesis testing in nonstationary time series. Ph.D. thesis, University of Wisconsin, Madison.
- Boswijk, H.P., and P.H. Franses (1996). Unit roots in periodic autoregressions. *Journal of Time Series Analysis* 17, 221–45.

- Box, G.E.P., and G.M. Jenkins (1970). *Time Series Analysis: Forecasting and Control*. San Francisco: Holden-Day.
- Cavanagh, C. (1986). Roots local to unity. Manuscript, Department of Economics, Harvard University, Cambridge, MA.
- Canova, F., and E. Ghysels (1994). Changes in seasonal patterns: are they cyclical? *Journal of Economic Dynamics and Control* 18, 1143–71.
- Canova, F., and B.E. Hansen (1995). Are seasonal patterns constant over time? a test for seasonal stability. *Journal of Business and Economic Statistics* 13, 237–52.
- Chan, N.H. (1988). The parameter inference for nearly nonstationary time series. *Journal of the American Statistical Association* 83, 857–62.
- Chan, N.H. (1989). On the nearly nonstationary seasonal time series. *Canadian Journal of Statistics* 17, 279–84.
- Chan, N.H., and C.Z. Wei (1987). Asymptotic inference for nearly nonstationary AR(1) processes. *Annals of Statistics* 15, 1050–63.
- Chan, N.H., and C.Z. Wei (1988). Limiting distributions of least squares estimates of unstable autoregressive processes. *Annals of Statistics* 16, 367–401.
- Dickey, D.A., D.P. Hasza, and W.A. Fuller (1984). Testing for unit roots in seasonal time series. *Journal of the American Statistical Association* 79, 355–67.
- Dickey, D.A., and W.A. Fuller (1979). Distribution of the estimators for autoregressive time series with a unit root. *Journal of the American Statistical Association* 74, 427–31.
- Engle, R.F., and C.W.J. Granger (1987). Cointegration and error correction: representation, estimation and testing. *Econometrica* 55, 251–76.
- Engle, R.F., C.W.J. Granger, S. Hylleberg, and H.S. Lee (1993). Seasonal cointegration: the Japanese consumption function. *Journal of Econometrics* 55, 275–98.
- Evans, G.B.A., and N.E. Savin (1981). Testing for unit roots: 1. *Econometrica* 49, 753–79.
- Evans, G.B.A., and N.E. Savin (1984). Testing for unit roots: 2. *Econometrica* 52, 1241–69.
- Findley, D.F., B.C. Monsell, W.R. Bell, M.C. Otto, and B.-C. Chen (1998). New capabilities and methods of the X-12-ARIMA seasonal-adjustment program. *Journal of Business and Economic Statistics* 16, 127–52.
- Franses, P.H. (1991). Seasonality, nonstationarity and forecasting monthly time series. *Journal of Forecasting* 7, 199–208.
- Franses, P.H. (1994). A multivariate approach to modelling univariate seasonal time series. *Journal of Econometrics* 63, 133–51.
- Franses, H.P., S. Hylleberg, and H.S. Lee (1995). Spurious deterministic seasonality. *Economics Letters* 48, 249–56.
- Franses, P.H., and T.J. Vogelsang (1998). Testing for seasonal unit roots in the presence of changing seasonal means. *Review of Economics and Statistics* 80, 231–40.
- Fuller, W.A. (1996). *Introduction to Statistical Time Series*, 2nd edn. New York: John Wiley.
- Ghysels, E. (1991). On seasonal asymmetries and their implications on deterministic and stochastic models of seasonality. Mimeo, C.R.D.E., Université de Montréal.
- Ghysels, E. (1994). On the economics and econometrics of seasonality. In C.A. Sims (ed.) *Advances in Econometrics*, pp. 257–316. Cambridge: Cambridge University Press.
- Ghysels, E., H.S. Lee, and P.L. Siklos (1993). On the (mis)specification of seasonality and its consequences: an empirical investigation with US data. *Empirical Economics* 18, 747–60.
- Ghysels, E., H.S. Lee, and J. Noh (1994). Testing for unit roots in seasonal time series: some theoretical extensions and a Monte Carlo investigation. *Journal of Econometrics* 62, 415–42.
- Ghysels, E., and D.R. Osborn (2000). *The Econometric Analysis of Seasonal Time Series*. Cambridge: Cambridge University Press.
- Ghysels, E., D.R. Osborn, and P.M.M. Rodrigues (1999). Seasonal nonstationarity and near-nonstationarity. Discussion paper 99s-05, CIRANO, available at <http://ftp.cirano.umontreal.ca/pub/publication/99s-05.pdf.zip>.

- Hylleberg, S. (1994). Modelling seasonal variation in nonstationary time series analysis and cointegration. In C. Hargreaves (ed.) *Non-Stationary Time Series Analysis and Cointegration*, pp. 153–178. Oxford: Oxford University Press.
- Hylleberg, S. (1995). Tests for seasonal unit roots: general to specific or specific to general?. *Journal of Econometrics* 69, 5–25.
- Hylleberg, S., C. Jørgensen, and N.K. Sørensen (1993). Seasonality in macroeconomic time series. *Empirical Economics* 18, 321–35.
- Hylleberg, S., R.F. Engle, C.W.J. Granger, and B.S. Yoo (1990). Seasonal integration and cointegration. *Journal of Econometrics* 44, 215–38.
- Kunst, R.M. (1997). Testing for cyclical non-stationarity in autoregressive processes. *Journal of Time Series Analysis* 18, 325–30.
- Jeganathan, P. (1991). On the asymptotic behavior of least-squares estimators in AR time series with roots near the unit circle. *Econometric Theory* 7, 269–306.
- Maravall, A. (1993). Stochastic linear trends: models and estimators. *Journal of Econometrics* 56, 5–38.
- Nabeya, S., and P. Perron (1994). Local asymptotic distribution related to the AR(1) model with dependent errors. *Journal of Econometrics* 62, 229–64.
- Osborn, D.R. (1990). A survey of seasonality in UK macroeconomic variables. *International Journal of Forecasting* 6, 327–36.
- Osborn, D.R. (1993). Discussion on “Seasonal Cointegration: The Japanese Consumption Function”. *Journal of Econometrics* 55, 299–303.
- Osborn, D.R., A.P.L. Chui, J.P. Smith, and C.R. Birchenhall (1988). Seasonality and the order of integration for consumption. *Oxford Bulletin of Economics and Statistics* 50, 361–77.
- Osborn, D.R., and P.M.M. Rodrigues (1998). The asymptotic distributions of seasonal unit root tests: a unifying approach. University of Manchester, School of Economic Studies Discussion Paper Series No. 9811.
- Perron, P. (1989). The calculation of the limiting distribution of the least-squares estimator in a near-integrated model. *Econometric Theory* 5, 241–55.
- Perron, P. (1992). The limiting distribution of the least-squares estimator in nearly integrated seasonal models. *Canadian Journal of Statistics* 20, 121–34.
- Phillips, P.C.B. (1987). Towards a unified asymptotic theory for autoregression. *Biometrika* 74, 535–47.
- Phillips, P.C.B. (1988). Regression theory for near-integrated time series. *Econometrica* 56, 1021–43.
- Rodrigues, P.M.M. (1998). Inference in seasonal nonstationary processes. Unpublished Ph.D. thesis, School of Economic Studies, University of Manchester.
- Rodrigues, P.M.M. (2000a). A note on the application of the DF test to seasonal data. *Statistics and Probability Letters* 47, 171–5.
- Rodrigues, P.M.M. (2000b). Near seasonal integration. *Econometric Theory*, forthcoming.
- Rodrigues, P.M.M., and D.R. Osborn (1999). Performance of seasonal unit root tests for monthly data. *Journal of Applied Statistics* 26, 985–1004.
- Smith, J., and J. Otero (1997). Structural breaks and seasonal integration. *Economics Letters* 56, 13–19.
- Smith, R.J., and A.M.R. Taylor (1998). Additional critical values and asymptotic representations for seasonal unit root tests. *Journal of Econometrics* 85, 269–88.
- Tanaka, K. (1996). *Time Series Analysis: Nonstationarity and Noninvertible Distribution Theory*. New York: John Wiley.
- Taylor, A.M.R. (1998). Testing for unit roots in monthly time series. *Journal of Time Series Analysis* 19, 349–68.

---

CHAPTER THIRTY-TWO

# Vector Autoregressions

*Helmut Lütkepohl\**

## 1 INTRODUCTION

The last 60 years have witnessed a rapid development in the field of econometrics. In the 1940s and 1950s the foundations were laid by the Cowles Commission researchers for analyzing econometric simultaneous equations models. Once the basic statistical theory was available many such models were constructed for empirical analysis. The parallel development of computer technology in the 1950s and 1960s has resulted in simultaneous equations models of increasing size with the accompanying hope that more detailed models would result in better approximations to the underlying data generation mechanisms. It turned out, however, that increasing the number of variables and equations of the models did not generally lead to improvements in performance in terms of forecasting, for instance. In fact, in some forecast comparisons univariate time series models were found to be superior to large scale econometric models. One explanation of this failure of the latter models is their insufficient representation of the dynamic interactions in a system of variables.

The poor performance of standard macroeconometric models in some respects resulted in a critical assessment of econometric simultaneous equations modeling as summarized by Sims (1980) who advocated using vector autoregressive (VAR) models as alternatives. In these models all variables are often treated as a priori endogenous and allowance is made for rich dynamics. Restrictions are imposed to a large extent by statistical tools rather than by prior beliefs based on uncertain theoretical considerations. Although VAR models are, now, standard instruments in econometric analyses, it has become apparent that certain types of interpretations and economic investigations are not possible without incorporating nonstatistical a priori information. Therefore, so-called *structural* VAR models are now often used in practice. Moreover, the invention of cointegration by Granger

(1981) and Engle and Granger (1987) has resulted in specific parameterizations which support the analysis of the cointegration structure. The cointegrating relations are often interpreted as the connecting link to the relations derived from economic theory. Therefore they are of particular interest in an analysis of a set of time series variables.

In the following I will first discuss some of the related models which are now in common use. I will then consider estimation and specification issues in Sections 3 and 4, respectively. Possible uses of VAR models are presented in Section 5. Conclusions and extensions are considered in Section 6. Nowadays a number of books are available which treat modern developments in VAR modeling and dynamic econometric analysis more generally in some detail. Surveys of vector autoregressive modeling include Watson (1994) and Lütkepohl and Breitung (1997).

## 2 VAR MODELS

### 2.1 Characteristics of variables

The characteristics of the variables involved determine to some extent which model is a suitable representation of the data generation process (DGP). For instance, the trending properties of the variables and their seasonal fluctuations are of importance in setting up a suitable model. In the following a variable is called *integrated* of order  $d$  ( $I(d)$ ) if stochastic trends or unit roots can be removed by differencing the variable  $d$  times (see also Chapter 29 by Bierens in this volume). In the present chapter it is assumed that all variables are at most  $I(1)$  if not otherwise stated so that, for any time series variable  $y_{kt}$  it is assumed that  $\Delta y_{kt} \equiv y_{kt} - y_{k,t-1}$  has no stochastic trend. Note, however, that  $\Delta y_{kt}$  may still have deterministic components such as a polynomial trend and a seasonal component whereas seasonal unit roots are excluded. Note also that a variable without a stochastic trend or unit root is sometimes called  $I(0)$ . In other words, a variable is  $I(0)$  if its stochastic part is stationary. A set of  $I(1)$  variables is called *cointegrated* if a linear combination exists which is  $I(0)$ . Occasionally it is convenient to consider systems with both  $I(1)$  and  $I(0)$  variables. In this case the concept of cointegration is extended by calling any linear combination which is  $I(0)$  a cointegration relation although this terminology is not in the spirit of the original definition because it can result in a linear combination of  $I(0)$  variables being called a cointegration relation. Further discussion of cointegration may also be found in Chapter 30 by Dolado, Gonzalo, and Marmol in this volume.

As mentioned earlier, we allow for deterministic polynomial trends. For these terms we assume for convenience that they are at most linear. In other words, we exclude higher order polynomial trend terms. For practical purposes this assumption is not a severe limitation.

### 2.2 Alternative models and model representations

Given a set of  $K$  time series variables  $y_t = (y_{1t}, \dots, y_{Kt})'$ , the basic VAR model is of the form

$$y_t = A_1 y_{t-1} + \dots + A_p y_{t-p} + u_t, \quad (32.1)$$

where  $u_t = (u_{1t}, \dots, u_{Kt})'$  is an unobservable zero-mean independent white noise process with time invariant positive definite covariance matrix  $E(u_t u_t') = \Sigma_u$  and the  $A_i$  are  $(K \times K)$  coefficient matrices. This model is often briefly referred to as a VAR( $p$ ) process because the number of lags is  $p$ .

The process is *stable* if

$$\det(I_K - A_1 z - \dots - A_p z^p) \neq 0 \text{ for } |z| \leq 1. \quad (32.2)$$

Assuming that it has been initiated in the infinite past, it generates stationary time series which have time invariant means, variances, and covariance structure. If the determinantal polynomial in (32.2) has roots for  $z = 1$  (i.e. unit roots), then some or all of the variables are I(1) and they may also be cointegrated. Thus, the present model is general enough to accommodate variables with stochastic trends. On the other hand, it is not the most suitable type of model if interest centers on the cointegration relations because they do not appear explicitly in the VAR version (32.1). They are more easily analyzed by reparameterizing (32.1) to obtain the so-called *vector error correction model* (VECM):

$$\Delta y_t = \Pi y_{t-1} + \Gamma_1 \Delta y_{t-1} + \dots + \Gamma_{p-1} \Delta y_{t-p+1} + u_t, \quad (32.3)$$

where  $\Pi = -(I_K - A_1 - \dots - A_p)$  and  $\Gamma_i = -(A_{i+1} + \dots + A_p)$  for  $i = 1, \dots, p-1$ . This representation of the process is obtained from (32.1) by subtracting  $y_{t-1}$  from both sides and rearranging terms. Because  $\Delta y_t$  does not contain stochastic trends by our assumption that all variables can be at most I(1), the term  $\Pi y_{t-1}$  is the only one which includes I(1) variables. Hence,  $\Pi y_{t-1}$  must also be I(0). Thus, it contains the cointegrating relations. The  $\Gamma_j$  ( $j = 1, \dots, p-1$ ) are often referred to as the short-term or short-run parameters while  $\Pi y_{t-1}$  is sometimes called long-run or long-term part. The model in (32.3) will be abbreviated as VECM( $p$ ) because  $p$  is the largest lag of the levels  $y_t$  that appears in the model. To distinguish the VECM from the VAR model the latter is sometimes called the levels version. Of course, it is also possible to determine the  $A_j$  levels parameter matrices from the coefficients of the VECM as  $A_1 = \Gamma_1 + \Pi + I_K$ ,  $A_i = \Gamma_i - \Gamma_{i-1}$  for  $i = 2, \dots, p-1$ , and  $A_p = -\Gamma_{p-1}$ .

If the VAR( $p$ ) process has unit roots, that is,  $\det(I_K - A_1 z - \dots - A_p z^p) = 0$  for  $z = 1$ , the matrix  $\Pi$  is singular. Suppose it has rank  $r$ , that is,  $\text{rank}(\Pi) = r$ . Then it is well known that  $\Pi$  can be written as a product  $\Pi = \alpha \beta'$ , where  $\alpha$  and  $\beta$  are  $(K \times r)$  matrices with  $\text{rank}(\alpha) = \text{rank}(\beta) = r$ . Premultiplying an I(0) vector by some matrix results again in an I(0) process. Hence, premultiplying  $\Pi y_{t-1} = \alpha \beta' y_{t-1}$  by  $(\alpha' \alpha)^{-1} \alpha'$  shows that  $\beta' y_{t-1}$  is I(0) and, therefore, contains the cointegrating relations. Hence, there are  $r = \text{rank}(\Pi)$  linearly independent cointegrating relations among the components of  $y_t$ . The matrices  $\alpha$  and  $\beta$  are not unique so that there are many possible  $\beta$  matrices which contain the cointegrating relations or linear transformations of them. Consequently, cointegrating relations with economic content cannot be extracted purely from the observed time series. Some nonsample information is required to identify them uniquely.

Special cases included in (32.3) are I(0) processes for which  $r = K$  and systems that have a stable VAR representation in first differences. In the latter case,  $r = 0$  and the term  $\Pi y_{t-1}$  disappears in (32.3). These boundary cases do not represent cointegrated systems in the usual sense. There are also other cases where no cointegration in the original sense is present although the model (32.3) has a cointegrating rank strictly between 0 and  $K$ . Still it is convenient to include these cases in the present framework because they can be accommodated easily as far as estimation and inference are concerned.

In practice the basic models (32.1) and (32.3) are usually too restrictive to represent the main characteristics of the data. In particular, deterministic terms such as an intercept, a linear trend term or seasonal dummy variables may be required for a proper representation of the data. There are two ways to include deterministic terms. The first possibility is to represent the observed variables  $y_t$  as a sum of a deterministic term and a stochastic part,

$$y_t = \mu_t + x_t, \quad (32.4)$$

where  $\mu_t$  is the deterministic part and  $x_t$  is a stochastic process which may have a VAR or VECM representation as in (32.1) or (32.3), that is,  $x_t = A_1 x_{t-1} + \dots + A_p x_{t-p} + u_t$  or  $\Delta x_t = \Pi x_{t-1} + \Gamma_1 \Delta x_{t-1} + \dots + \Gamma_{p-1} \Delta x_{t-p+1} + u_t$ . In that case, if  $\mu_t$  is a linear trend term, that is,  $\mu_t = \mu_0 + \mu_1 t$ , then  $y_t$  has a VAR( $p$ ) representation of the form

$$y_t = v_0 + v_1 t + A_1 y_{t-1} + \dots + A_p y_{t-p} + u_t, \quad (32.5)$$

where  $v_0 = -\Pi \mu_0 + (\sum_{j=1}^p j A_j) \mu_1$  and  $v_1 = -\Pi \mu_1$ . In other words,  $v_0$  and  $v_1$  satisfy a set of restrictions. Note, however, that if (32.5) is regarded as the basic model without restrictions for  $v_i$ ,  $i = 0, 1$ , the model can in principle generate quadratic trends if I(1) variables are included, whereas in (32.4) with a deterministic term  $\mu_t = \mu_0 + \mu_1 t$  a linear trend term is permitted only. The fact that in (32.4) a clear partitioning of the process in a deterministic and a stochastic component is available is sometimes advantageous in theoretical derivations. Also, in practice, it may be possible to subtract the deterministic term first and then focus the analysis on the stochastic part which usually contains the behavioral relations. Therefore this part is often of primary interest in econometric analyses. Of course, a VECM( $p$ ) representation equivalent to (32.5) also exists.

In practice, these representations with possibly additional deterministic terms may still not be general enough. At times one may wish to include stochastic exogenous variables on top of the deterministic part. A fairly general VECM form which includes all these terms is

$$\Delta y_t = \Pi y_{t-1} + \Gamma_1 \Delta y_{t-1} + \dots + \Gamma_{p-1} \Delta y_{t-p+1} + C D_t + B z_t + u_t, \quad (32.6)$$

where the  $z_t$  are exogenous variables,  $D_t$  contains all regressors associated with deterministic terms, and  $C$  and  $B$  are parameter matrices.

All the models we have presented so far are reduced form models in that they do not include instantaneous relations between the endogenous variables  $y_t$ . In

practice it is often desirable to model the contemporaneous relations as well and therefore it is useful to consider a structural form

$$\Gamma_0^* \Delta y_t = \Pi^* y_{t-1} + \Gamma_1^* \Delta y_{t-1} + \dots + \Gamma_{p-1}^* \Delta y_{t-p+1} + C^* D_t + B^* z_t + v_t, \quad (32.7)$$

where  $v_t$  is a  $(K \times 1)$  zero mean white noise process with covariance matrix  $\Sigma_v$  and the  $\Pi^*$ ,  $\Gamma_j^*$  ( $j = 0, \dots, p-1$ ),  $C^*$  and  $B^*$  are structural form parameter matrices. The reduced form corresponding to the structural model (32.7) is given in (32.6) with  $\Gamma_j = (\Gamma_0^*)^{-1} \Gamma_j^*$  ( $j = 1, \dots, p-1$ ),  $C = (\Gamma_0^*)^{-1} C^*$ ,  $\Pi = (\Gamma_0^*)^{-1} \Pi^*$ ,  $B = (\Gamma_0^*)^{-1} B^*$  and  $u_t = (\Gamma_0^*)^{-1} v_t$ . Of course, a number of restrictions are usually imposed on the general forms of our models. These restrictions are important at the estimation stage which will be discussed next.

### 3 ESTIMATION

Because estimation of some of the special case models is particularly easy these cases will be considered in more detail in the following. We begin with the levels VAR representation (32.1) under the condition that no restrictions are imposed. Then estimation of the VECM (32.3) is treated and finally more general model variants are discussed.

#### 3.1 Estimation of unrestricted VARs and VECMs

Given a sample  $y_1, \dots, y_T$  and presample values  $y_{-p+1}, \dots, y_0$ , the  $K$  equations of the VAR (32.1) may be estimated separately by least squares (LS) without losing efficiency relative to generalized LS (GLS) approaches. In fact, in this case LS is identical to GLS. Under standard assumptions, the LS estimator  $\hat{A}$  of  $A = [A_1 : \dots : A_p]$  is consistent and asymptotically normally distributed (see, e.g., Lütkepohl, 1991),

$$\sqrt{T} \text{vec}(\hat{A} - A) \xrightarrow{d} N(0, \Sigma_{\hat{A}}) \quad \text{or, more intuitively, } \text{vec}(\hat{A}) \xrightarrow{d} N(\text{vec}(A), \Sigma_{\hat{A}}/T). \quad (32.8)$$

Here  $\text{vec}$  denotes the column stacking operator which stacks the columns of a matrix in a column vector,  $\xrightarrow{d}$  signifies convergence in distribution and  $\xrightarrow{a}$  indicates "asymptotically distributed as".

Although this result also holds for I(1) cointegrated systems (see Sims, Stock, and Watson, 1990; Lütkepohl, 1991, ch. 11) it is important to note that in this case the covariance matrix  $\Sigma_{\hat{A}}$  is singular whereas it is nonsingular in the usual I(0) case. In other words, if there are integrated or cointegrated variables, some estimated coefficients or linear combinations of coefficients converge with a faster rate than  $\sqrt{T}$ . Therefore, the usual  $t$ -,  $\chi^2$ -, and F-tests for inference regarding the VAR parameters may not be valid in this case (see, e.g. Toda and Phillips, 1993). Although inference problems may arise in VAR models with I(1) variables, there are also many unproblematic cases. Dolado and Lütkepohl (1996) show that if all variables are I(1) or I(0) and if a null hypothesis is considered which does not

restrict elements of each of the  $A_i$  ( $i = 1, \dots, p$ ) the usual tests have their standard asymptotic properties. For example, if the VAR order  $p \geq 2$ , the  $t$ -ratios have their usual asymptotic standard normal distributions (see also Toda and Yamamoto (1995) for a related result).

If the white noise process  $u_t$  is normally distributed (Gaussian) and the process  $y_t$  is  $I(0)$ , then the LS estimator is identical to the maximum likelihood (ML) estimator conditional on the initial values. It is also straightforward to include deterministic terms such as polynomial trends in the model (32.1). In this case the asymptotic properties of the VAR coefficients remain essentially the same as in the case without deterministic terms (Sims *et al.*, 1990).

If the cointegrating rank of the system under consideration is known and one wishes to impose a corresponding restriction, working with the VECM form (32.3) is convenient. If the VAR order is  $p = 1$  the estimators may be obtained by applying reduced rank regression (RRR) to  $\Delta y_t = \Pi y_{t-1} + u_t$  subject to  $\text{rank}(\Pi) = r$ . The approach is easily extended to higher order VAR processes as well (see Johansen, 1995). Under Gaussian assumptions the ML estimators conditional on the presample values may, in fact, be obtained in this way. However, in order to estimate the matrices  $\alpha$  and  $\beta$  in  $\Pi = \alpha\beta'$  consistently, it is necessary to impose identifying restrictions. Without such restrictions only the product  $\alpha\beta' = \Pi$  can be estimated consistently. If uniqueness restrictions are imposed it can be shown that  $T(\hat{\beta} - \beta)$  and  $\sqrt{T}(\hat{\alpha} - \alpha)$  converge in distribution (Johansen, 1995). Hence, the estimator of  $\beta$  converges with the fast rate  $T$  and is therefore sometimes called *super-consistent*. In contrast, the estimator of  $\alpha$  converges with the usual rate  $\sqrt{T}$ . The estimators of  $\Gamma = [\Gamma_1 : \dots : \Gamma_{p-1}]$  and  $\Pi$  are consistent and asymptotically normal under general assumptions. The asymptotic distribution of  $\hat{\Gamma}$  is nonsingular so that standard inference may be used for the short-term parameters  $\Gamma_j$ . On the other hand, the asymptotic distribution of  $\hat{\Pi}$  is singular if  $r < K$ . This result is due to two forces. On the one hand, imposing the rank constraint in estimating  $\Pi$  restricts the parameter space and, on the other hand,  $\Pi$  involves the cointegrating relations which are estimated super-consistently.

It is perhaps interesting to note that an estimator of  $A$  can be computed via the estimates of  $\Pi$  and  $\Gamma$ . That estimator has the advantage of imposing the cointegrating restrictions on the levels version of the VAR process. However, its asymptotic distribution is the same as in (32.8) where no restrictions have been imposed in estimating  $A$ .

### 3.2 Estimation of restricted models and structural forms

Efficient estimation of a general structural form model such as (32.7) with restrictions on the parameter matrices is more complicated. Of course, identifying restrictions are necessary for consistent estimation. In practice, various over-identifying restrictions are usually available, typically in the form of zero restrictions on  $\Gamma_j^*$  ( $j = 0, \dots, p - 1$ ),  $C^*$  and  $B^*$ . In addition, there may be a rank restriction for  $\Pi^*$  given by the number of cointegrating relations. Alternatively,

$\Pi^*$  may be replaced by the product  $\alpha^*\beta^*$ , if identifying restrictions are available for the cointegrating relations and/or the loading matrix  $\alpha^*$ . Restrictions for  $\alpha^*$  are typically zero constraints, meaning that some cointegrating relations are excluded from some of the equations of the system. In some cases it is possible to estimate  $\beta^*$  in a first stage, for example, using a reduced form procedure which ignores some or all of the structural restrictions on the short-term parameters. Let the estimator be  $\hat{\beta}^*$ . Because the estimators of the cointegrating parameters converge at a better rate than the estimators of the short-term parameters they may be treated as fixed in a second-stage procedure for the structural form. In other words, a systems estimation procedure may be applied to

$$\Gamma_0^* \Delta y_t = \alpha^* \hat{\beta}^* y_{t-1} + \Gamma_1^* \Delta y_{t-1} + \dots + \Gamma_{p-1}^* \Delta y_{t-p+1} + C^* D_t + B^* z_t + \hat{v}_t. \quad (32.9)$$

If only exclusion restrictions are imposed on the parameter matrices in this form, standard three-stage LS or similar methods may be applied which result in estimators of the short-term parameters with the usual asymptotic properties. Important results on estimating models with integrated variables are due to Phillips and his co-workers (e.g. Phillips, 1987, 1991).

If deterministic variables are to be included in the cointegration relations this requires a suitable reparameterization of the model. Such reparameterizations for intercepts and linear trend terms are presented in Section 4.2, where tests for the cointegrating rank are discussed. In that context a proper treatment of deterministic terms is of particular importance. Therefore, a more detailed discussion is deferred to Section 4.2. In a subsequent analysis of the model the parameters of the deterministic terms are often of minor interest and therefore the properties of the corresponding estimators are not treated in detail here (see, however, Sims *et al.*, 1990).

## 4 MODEL SPECIFICATION AND MODEL CHECKING

### 4.1 Choosing the model order

Unrestricted VAR models usually involve a substantial number of parameters which in turn results in rather imprecise estimators. Therefore, it is desirable to impose restrictions that reduce the dimensionality of the parameter space. Such restrictions may be based on economic theory or other nonsample information and on statistical procedures. Of course, for structural models nonsample information is required for imposing identifying constraints. On top of that there may be further overidentifying constraints on the basis of a priori knowledge.

Tests are common statistical procedures for detecting possible restrictions. For example, *t*-ratios and F-tests are available for this purpose. These tests retain their usual asymptotic properties if they are applied to the short-run parameters in a VECM whereas problems may arise in the levels VAR representation as explained in the previous section. A particular set of restrictions where such problems occur is discussed in more detail in Section 5.2. In case of doubt it may be preferable to work on the VECM form. This form also makes it easy to test

restrictions on the cointegration vectors (see Chapter 30 by Dolado, Gonzalo, and Marmol in this volume).

Because the cointegrating rank  $r$  is usually unknown when the choice of  $p$  is made, it is useful to focus on the VAR form (32.1) at this stage. Various model selection criteria are available that can be used in this context. In practice, it is not uncommon to start from a model with some prespecified maximum lag length, say  $p_{\max}$ , and apply tests sequentially, eliminating one or more variables in each step until a relatively parsimonious representation with significant parameter estimates has been found. Instead of sequential tests one may alternatively choose the lag length or determine exclusion restrictions by model selection procedures. For example, for determining the VAR order, the general approach is to fit VAR( $m$ ) models with orders  $m = 0, \dots, p_{\max}$  and choose an estimator of the order  $p$  which minimizes a criterion such as

$$\text{AIC}(m) = \log \det(\tilde{\Sigma}_u(m)) + \frac{2}{T} m K^2,$$

(see Akaike, 1974);

$$\text{HQ}(m) = \log \det(\tilde{\Sigma}_u(m)) + \frac{2 \log \log T}{T} m K^2$$

proposed by Hannan and Quinn (1979); or

$$\text{SC}(m) = \log \det(\tilde{\Sigma}_u(m)) + \frac{\log T}{T} m K^2$$

due to Schwarz (1978). Here  $\det(\cdot)$  denotes the determinant,  $\log$  is the natural logarithm and  $\tilde{\Sigma}_u(m) = T^{-1} \sum_{t=1}^T \hat{u}_t \hat{u}'_t$  is the residual covariance matrix estimator for a model of order  $m$ . The term  $\log \det(\tilde{\Sigma}_u(m))$  measures the fit of a model with order  $m$ . Since there is no correction for degrees of freedom in the covariance matrix estimator the log determinant decreases (or at least does not increase) when  $m$  increases. Note that the sample size is assumed to be held constant and, hence, the number of presample values set aside for estimation is determined by the maximum order  $p_{\max}$ . The last terms in the criteria penalize large VAR orders. In each case the estimator  $\hat{p}$  of  $p$  is chosen to be the order which minimizes the desired criterion so that the two terms in the sum on the right-hand sides are balanced optimally.

The AIC criterion asymptotically overestimates the order with positive probability whereas the last two criteria estimate the order consistently ( $\text{plim } \hat{p} = p$  or  $\hat{p} \rightarrow p$  a.s.) under quite general conditions, if the actual DGP has a finite VAR order and the maximum order  $p_{\max}$  is larger than the true order. These results not only hold for I(0) processes but also for I(1) processes with cointegrated variables (Paulsen, 1984). Denoting the orders selected by the three criteria by  $\hat{p}(\text{AIC})$ ,  $\hat{p}(\text{HQ})$ , and  $\hat{p}(\text{SC})$ , respectively, the following relations hold even in small samples of fixed size  $T \geq 16$  (see Lütkepohl, 1991, chs 4 and 11):  $\hat{p}(\text{SC}) \leq \hat{p}(\text{HQ}) \leq \hat{p}(\text{AIC})$ .

Model selection criteria may also be used for identifying single coefficients that may be replaced by zero or other exclusion restrictions. After a model has been set up, a series of checks may be employed to confirm the model's adequacy. Some such checks will be mentioned briefly in a subsequent section. Before that issue is taken up, procedures for specifying the cointegrating rank will be reviewed.

## 4.2 Specifying the cointegrating rank

In practice, the cointegrating rank  $r$  is also usually unknown. It is commonly determined by a sequential testing procedure based on likelihood ratio (LR) type tests. Because for a given cointegrating rank Gaussian ML estimates for the reduced form VECM are easy to compute, as mentioned in Section 3.1, LR test statistics are also easily available. The following sequence of hypotheses may be considered:

$$H_0(r_0) : \text{rank}(\Pi) = r_0 \quad \text{versus} \quad H_1(r_0) : \text{rank}(\Pi) > r_0, \quad r_0 = 0, \dots, K - 1. \quad (32.10)$$

The testing sequence terminates if the null hypothesis cannot be rejected for the first time. If the first null hypothesis,  $H_0(0)$ , cannot be rejected, a VAR process in first differences is considered. At the other end, if all the null hypotheses can be rejected, the process is assumed to be I(0) in levels.

In principle, it is also possible to investigate the cointegrating rank by testing hypotheses  $H_0 : \text{rank}(\Pi) \leq r_0$  versus  $H_1 : \text{rank}(\Pi) > r_0$ ,  $r_0 = K - 1, \dots, 1, 0$ , that is, by testing in reverse order from the largest to the smallest rank. The cointegrating rank is then chosen as the last  $r_0$  for which  $H_0$  is not rejected. From a theoretical viewpoint such a procedure has a drawback, however. Strictly speaking the critical values for the tests to be discussed in the following apply for the situation that  $\text{rank}(\Pi)$  is equal to  $r_0$  and not smaller than  $r_0$ . Moreover, starting with the smallest rank as in (32.10) means to test the most restricted model first. Thus, a less restricted model is considered only if the data are strongly in favor of removing the restrictions.

Although, under Gaussian assumptions, LR tests can be used here, it turns out that the limiting distribution of the LR statistic under  $H_0(r_0)$  is non-standard. It depends on the difference  $K - r_0$  and on the deterministic terms included in the DGP. In particular, the deterministic trend terms in the DGP have an impact on the null distribution of the LR tests. Therefore, LR type tests have been derived under different assumptions regarding the deterministic trend parameters. Fortunately, the limiting null distributions do not depend on the short-term dynamics if the latter are properly specified and, hence, critical values for LR type tests have been tabulated for different values of  $K - r_0$  under alternative assumptions for deterministic trend terms.

In this context it turns out that the model (32.4), where the deterministic and stochastic parts are separated, is a convenient point of departure. Therefore we consider the model

$$y_t = \mu_0 + \mu_1 t + x_t \quad (32.11)$$

**Table 32.1** Models and LR type tests

<i>Assumption for deterministic term</i>	<i>Model</i>	<i>Reference</i>
$\mu_0$ arbitrary $\mu_1 = 0$	$\Delta y_t = v_0 + \Pi y_{t-1} + \sum_{j=1}^{p-1} \Gamma_j \Delta y_{t-j} + u_t$	Johansen (1995)
	$\Delta y_t = [\Pi : v_0] \begin{bmatrix} y_{t-1} \\ 1 \end{bmatrix} + \sum_{j=1}^{p-1} \Gamma_j \Delta y_{t-j} + u_t$	Johansen and Juselius (1990)
	$\Delta y_t = \Pi(y_{t-1} - \tilde{\mu}_0) + \sum_{j=1}^{p-1} \Gamma_j \Delta y_{t-j} + u_t$	Saikkonen and Luukkonen (1997)
$\mu_0$ arbitrary $\mu_1 \neq 0, \beta' \mu_1 = 0$	$\Delta y_t = v_0 + \Pi y_{t-1} + \sum_{j=1}^{p-1} \Gamma_j \Delta y_{t-j} + u_t$	Johansen (1995)
	$\Delta y_t - \hat{\mu}_1 = \Pi(y_{t-1} - \hat{\mu}_0) + \sum_{j=1}^{p-1} \Gamma_j (\Delta y_{t-j} - \hat{\mu}_1) + u_t$	Saikkonen and Lütkepohl (2000b)
$\mu_0, \mu_1$ arbitrary	$\Delta y_t = v + [\Pi : v_1] \begin{bmatrix} y_{t-1} \\ t-1 \end{bmatrix} + \sum_{j=1}^{p-1} \Gamma_j \Delta y_{t-j} + u_t$	Johansen (1995)
	$\Delta y_t = v_0 + v_1 t + \Pi y_{t-1} + \sum_{j=1}^{p-1} \Gamma_j \Delta y_{t-j} + u_t$	Perron and Campbell (1993)
	$\Delta y_t - \hat{\mu}_1 = \Pi(y_{t-1} - \hat{\mu}_0 - \hat{\mu}_1(t-1)) + \sum_{j=1}^{p-1} \Gamma_j (\Delta y_{t-j} - \hat{\mu}_1) + u_t$	Saikkonen and Lütkepohl (2000a) Lütkepohl and Saikkonen (2000)

with

$$\Delta x_t = \Pi x_{t-1} + \Gamma_1 \Delta x_{t-1} + \dots + \Gamma_{p-1} \Delta x_{t-p+1} + u_t. \quad (32.12)$$

It is easy to see that the process  $y_t$  has a VECM representation

$$\begin{aligned}
 \Delta y_t &= v_0 + v_1 t + \Pi y_{t-1} + \sum_{j=1}^{p-1} \Gamma_j \Delta y_{t-j} + u_t \\
 &= v + [\Pi : v_1] \begin{bmatrix} y_{t-1} \\ t-1 \end{bmatrix} + \sum_{j=1}^{p-1} \Gamma_j \Delta y_{t-j} + u_t \\
 &= v + \Pi^+ y_{t-1}^+ + \sum_{j=1}^{p-1} \Gamma_j \Delta y_{t-j} + u_t,
 \end{aligned} \quad (32.13)$$

where  $v_0$  and  $v_1$  are as defined below (32.5),  $v = v_0 + v_1$ ,  $\Pi^+ = [\Pi : v_1]$  and  $y_{t-1}^+ = [y_{t-1}' : t-1]'$ . Depending on the assumptions for  $\mu_0$  and  $\mu_1$ , different LR type tests can be obtained in this framework by appropriately restricting the parameters of the deterministic part and using RR regression techniques. An overview is given in Table 32.1 which is adopted from Table 1 of Hubrich, Lütkepohl and Saikkonen (2001) where more details on the tests may be found.

For instance, if  $\mu_1 = 0$  and  $\mu_0$  is unrestricted, a nonzero mean term is accommodated whereas a deterministic linear trend term is excluded by assumption. Three variants of LR type tests have been considered in the literature for this situation plus a number of asymptotically equivalent modifications. As can be seen from Table 32.1, the three statistics can be obtained easily from VECMs. The first test is obtained by dropping the  $v_1 t$  term in (32.13) and estimating the intercept term in the VECM in unrestricted form and hence, the estimated model may generate linear trends. The second test enforces the restriction that there is no linear deterministic trend in computing the test statistic by absorbing the intercept into the cointegration relations. Finally, in the third test the mean term  $\mu_0$  is estimated in a first step and is subtracted from  $y_t$ . Then the estimation procedure with rank restriction for  $\Pi$  is applied to (32.12) with  $x_t$  replaced by  $\tilde{x}_t = y_t - \hat{\mu}_0$ . A suitable estimator  $\tilde{\mu}_0$  is proposed by Saikkonen and Luukkonen (1997) who also give the asymptotic distribution of the resulting test statistic under the null hypothesis. It is shown in Saikkonen and Lütkepohl (1999) that the latter test can have considerably more local power than the other two LR tests. Thus, based on local power it is the first choice if  $\mu_1 = 0$ .

If  $\mu_0$  is arbitrary,  $\mu_1 \neq 0$  and  $\beta' \mu_1 = 0$ , at least one of the variables has a deterministic linear trend because  $\mu_1 \neq 0$ , whereas the cointegration relations do not have a linear trend due to the constraint  $\beta' \mu_1 = 0$ . The resulting tests are perhaps the most frequently used ones for determining the cointegrating rank in applied work. It may be worth emphasizing, however, that for the  $(K \times r)$  matrix  $\beta$  to satisfy  $\beta' \mu_1 = 0$ ,  $\mu_1 \neq 0$  implies that  $r < K$ . Hence, if a trend is known to be present then it should also be allowed for under the alternative and consequently even under the alternative the rank must be smaller than  $K$ . In other words, in the present setting only tests of null hypotheses  $\text{rank}(\Pi) = r_0 < K - 1$  make sense. This result is a consequence of the fact that a linear trend is assumed in at least one of the variables ( $\mu_1 \neq 0$ ) whereas a stable model with an intercept cannot generate a linear trend. Two different LR type tests are available for this case.

In the third case, both  $\mu_0$  and  $\mu_1$  are unrestricted so that the variables and the cointegrating relations may have a deterministic linear trend. Three different LR type tests and some asymptotically equivalent relatives have been proposed for this situation. Again, these test statistics can be obtained conveniently via the VECMs using the techniques mentioned in Section 3. The first model is set up in such a way so as to impose the linearity of the trend term. The second model includes the trend term in unrestricted form. As mentioned earlier, in principle such a model can generate quadratic trends. Finally, the last test in Table 32.1 is again based on prior trend adjustment and estimation of the resulting VECM for the trend adjusted variables. The trend parameters are again estimated in a first step by a generalized LS procedure. Critical values for all these tests may be found in the references given in Table 32.1.

Instead of the pair of hypotheses in (32.10) one may alternatively test  $H_0(r_0) : \text{rank}(\Pi) = r_0$  versus  $H_1^*(r_0) : \text{rank}(\Pi) = r_0 + 1$ . LR tests for this pair of hypotheses were also pioneered by Johansen and are known as *maximum eigenvalue tests*. They can be applied for all the different cases listed in Table 32.1. They also have

non-standard limiting distributions. Critical values can be found in the literature cited in the foregoing.

A comprehensive survey of the properties of LR type tests for the cointegrating rank as well as a substantial number of other tests that have been proposed in the literature is given by Hubrich *et al.* (2001). We refer the interested reader to that article for further details.

### 4.3 Model checking

Once a model has been specified and estimated its adequacy is usually checked with a range of tests and other statistical procedures. Many of these model checking tools are based on the residuals of the final model. Some of them are applied to the residuals of individual equations and others are based on the full residual vectors. Examples of specification checking tools are visual inspection of the plots of the residuals and their autocorrelations. In addition, autocorrelations of squared residuals may be considered to check for possible autoregressive conditional heteroskedasticity (ARCH). Although it may be quite insightful to inspect the autocorrelations visually, formal statistical tests for remaining residual autocorrelation should also be applied. Such tests are often based on LM (Lagrange multiplier) or Portmanteau statistics. Moreover, normality tests of the Lomnicki–Jarque–Bera type may be applied to the residuals (see, e.g. Lütkepohl, 1991; Doornik and Hendry, 1997).

If model defects are detected at the checking stage this is usually regarded as an indication of the model being a poor representation of the DGP and efforts are made to find a better representation by adding other variables or lags to the model, by including nonlinear terms or changing the functional form, by modifying the sampling period or getting other data.

## 5 USES OF VECTOR AUTOREGRESSIVE MODELS

When an adequate model for the DGP of a system of variables has been found it may be used for forecasting and economic analysis. Different tools have been proposed for the latter purpose. For instance, there has been an extensive discussion of how to analyze causal relations between the variables of a system of interest. In this section forecasting VAR processes will be discussed first. Forecasting in more general terms is discussed in Chapter 27 by Stock in this volume. In subsection 5.2 the concept of Granger-causality will be introduced which is based on forecast performance. It has received considerable attention in the theoretical and empirical literature. In subsection 5.3 impulse responses are considered. They may also be regarded as instruments for analyzing causal relations between variables. Finally, forecast error variance decompositions and policy analysis are discussed in subsections 5.4 and 5.5, respectively.

### 5.1 Forecasting VAR processes

Neglecting deterministic terms and exogenous variables the levels VAR form (32.1) is particularly convenient to use in forecasting the variables  $y_t$ . Suppose the

$u_t$  are generated by an independent rather than just uncorrelated white noise process. Then the optimal (minimum MSE) one-step forecast in period  $T$  is the conditional expectation,

$$y_{T+1|T} = E(y_{T+1} | y_T, y_{T-1}, \dots) = A_1 y_T + \dots + A_p y_{T+1-p}. \quad (32.14)$$

Forecasts for larger horizons  $h > 1$  may be obtained recursively as

$$y_{T+h|T} = A_1 y_{T+h-1|T} + \dots + A_p y_{T+h-p|T}, \quad (32.15)$$

where  $y_{T+j|T} = y_{T+j}$  for  $j \leq 0$ . The corresponding forecast errors are

$$y_{T+h} - y_{T+h|T} = u_{T+h} + \Phi_1 u_{T+h-1} + \dots + \Phi_{h-1} u_{T+1}, \quad (32.16)$$

where it is easy to see by successive substitution that  $\Phi_s = \sum_{j=1}^s \Phi_{s-j} A_j$  ( $s = 1, 2, \dots$ ) with  $\Phi_0 = I_K$  and  $A_j = 0$  for  $j > p$  (see Lütkepohl, 1991, sec. 11.3). Hence, the forecasts are unbiased, that is, the forecast errors have expectation 0 and their MSE matrix is

$$\Sigma_y(h) = E\{(y_{T+h} - y_{T+h|T})(y_{T+h} - y_{T+h|T})'\} = \sum_{j=0}^{h-1} \Phi_j \Sigma_u \Phi_j'. \quad (32.17)$$

For any other  $h$ -step forecast with MSE matrix  $\Sigma_y^*(h)$ , say, the difference  $\Sigma_y^*(h) - \Sigma_y(h)$  is a positive semidefinite matrix. This result relies on the assumption that  $u_t$  is independent white noise, i.e.  $u_t$  and  $u_s$  are independent for  $s \neq t$ . If  $u_t$  is uncorrelated white noise and not necessarily independent over time, these forecasts are just best linear forecasts in general (see Lütkepohl, 1991, sec. 2.2.2).

The forecast MSEs for integrated processes are generally unbounded as the horizon  $h$  goes to infinity. Thus the forecast uncertainty increases without bounds for forecasts of the distant future. This contrasts with the case of  $I(0)$  variables for which the forecast MSEs are bounded by the unconditional covariance  $\Sigma_y$  of  $y_t$ . This means, in particular, that forecasts of cointegration relations have bounded MSEs even for horizons approaching infinity. The corresponding forecast intervals reflect this property as well. Assuming that the process  $y_t$  is Gaussian, that is,  $u_t \sim \text{iid } N(0, \Sigma_u)$ , the forecast errors are also multivariate normal. This result may be used to set up forecast intervals in the usual way.

In practice the parameters of a VAR process are usually estimated. Denoting by  $\hat{y}_{T+h|T}$  the forecast based on estimated coefficients corresponding to  $y_{T+h|T}$  the associated forecast error is

$$y_{T+h} - \hat{y}_{T+h|T} = [y_{T+h} - y_{T+h|T}] + [y_{T+h|T} - \hat{y}_{T+h|T}].$$

At the forecast origin  $T$  the first term on the right-hand side involves future residuals only whereas the second term involves present and past variables only, provided only past variables have been used for estimation. Consequently, if  $u_t$  is independent white noise, the two terms are independent. Moreover, under

standard assumptions, the difference  $y_{T+h|T} - \hat{y}_{T+h|T}$  is small in probability as the sample size used for estimation gets large. Hence, the forecast error covariance matrix in this case is

$$\begin{aligned}\Sigma_g(h) &= E\{[y_{T+h} - \hat{y}_{T+h|T}][y_{T+h} - \hat{y}_{T+h|T}]'\} \\ &= \Sigma_y(h) + o(1),\end{aligned}$$

where  $o(1)$  denotes a term which approaches zero as the sample size tends to infinity. Thus, for large samples the estimation uncertainty may be ignored in evaluating the forecast precision and setting up forecast intervals. In small samples, including a correction term is preferable, however. In this case the precision of the forecasts will depend on the precision of the estimators. Hence, if precise forecasts are desired, it is a good strategy to look for precise parameter estimators.

## 5.2 Granger-causality analysis

### THE CONCEPT

The causality concept introduced by Granger (1969) is perhaps the most widely discussed form of causality in the econometrics literature. Granger defines a variable  $y_{1t}$  to be causal for another time series variable  $y_{2t}$  if the former helps predicting the latter. Formally, denoting by  $y_{2,t+h|\Omega_t}$  the optimal  $h$ -step predictor of  $y_{2t}$  at origin  $t$  based on the set of all the relevant information in the universe  $\Omega_t$ ,  $y_{1t}$  may be defined to be Granger-noncausal for  $y_{2t}$  if and only if

$$y_{2,t+h|\Omega_t} = y_{2,t+h|\Omega_t \setminus \{y_{1,s} | s \leq t\}}, \quad h = 1, 2, \dots \quad (32.18)$$

Here  $\Omega_t \setminus \mathcal{A}$  denotes the set containing all elements of  $\Omega_t$  which are not in the set  $\mathcal{A}$ . In other words,  $y_{1t}$  is not causal for  $y_{2t}$  if removing the past of  $y_{1t}$  from the information set does not change the optimal forecast for  $y_{2t}$  at any forecast horizon. In turn,  $y_{1t}$  is Granger-causal for  $y_{2t}$  if the equality in (32.18) is violated for at least one  $h$  and, thus, a better forecast of  $y_{2t}$  is obtained for some forecast horizon by including the past of  $y_{1t}$  in the information set. If  $\Omega_t = \{(y_{1,s}, y_{2,s})' | s \leq t\}$  and  $(y_{1t}, y_{2t})'$  is generated by a bivariate VAR( $p$ ) process,

$$\begin{bmatrix} y_{1t} \\ y_{2t} \end{bmatrix} = \sum_{i=1}^p \begin{bmatrix} \alpha_{11,i} & \alpha_{12,i} \\ \alpha_{21,i} & \alpha_{22,i} \end{bmatrix} \begin{bmatrix} y_{1,t-i} \\ y_{2,t-i} \end{bmatrix} + u_t, \quad (32.19)$$

then (32.18) is easily seen to be equivalent to

$$\alpha_{21,i} = 0, \quad i = 1, 2, \dots, p. \quad (32.20)$$

Of course, Granger-causality can also be investigated in the framework of the VECM (see, e.g. Mosconi and Giannini, 1992).

Economic systems usually consist of more than two relevant variables. Hence, it is desirable to extend the concept of Granger-causality to higher dimensional systems. Different possible extensions have been considered (see, e.g. Lütkepohl, 1993; Dufour and Renault, 1998). One possible generalization assumes that the vector of all variables,  $y_t$ , is partitioned into two subvectors so that  $y_t = (y'_{1t}, y'_{2t})'$ . Then the definition in (32.18) may be used for the two subvectors,  $y_{1t}$ ,  $y_{2t}$ , rather than two individual variables. If  $\Omega_t = \{y_s | s \leq t\}$  and  $y_t$  is a VAR process of the form (32.19), where the  $\alpha_{kh,i}$  are now matrices of appropriate dimensions, the restrictions for noncausality are the same as in the bivariate case so that  $y_{1t}$  is Granger-noncausal for  $y_{2t}$  if  $\alpha_{21,i} = 0$  for  $i = 1, \dots, p$  (Lütkepohl, 1991, sec. 2.3.1).

This approach is not satisfactory if interest centers on a causal relation between two variables within a higher dimensional system because a set of variables being causal for another set of variables does not necessarily imply that each member of the former set is causal for each member of the latter set. Therefore it is of interest to consider causality of  $y_{1t}$  to  $y_{2t}$  if there are further variables in the system. In this context, different causality concepts have been proposed which are most easily explained in terms of the three-dimensional VAR process

$$y_t = \begin{bmatrix} y_{1t} \\ y_{2t} \\ y_{3t} \end{bmatrix} = \sum_{i=1}^p \begin{bmatrix} \alpha_{11,i} & \alpha_{12,i} & \alpha_{13,i} \\ \alpha_{21,i} & \alpha_{22,i} & \alpha_{23,i} \\ \alpha_{31,i} & \alpha_{32,i} & \alpha_{33,i} \end{bmatrix} \begin{bmatrix} y_{1,t-i} \\ y_{2,t-i} \\ y_{3,t-i} \end{bmatrix} + u_t. \quad (32.21)$$

Within this system causality of  $y_{1t}$  for  $y_{2t}$  is sometimes checked by testing

$$H_0 : \alpha_{21,i} = 0, \quad i = 1, \dots, p. \quad (32.22)$$

These restrictions are not equivalent to (32.18), however. They are equivalent to equality of the one-step forecasts,  $y_{2,t+1|\Omega_t} = y_{2,t+1|\Omega_t \setminus \{y_{1,s} | s \leq t\}}$ . The information in past  $y_{1t}$  may still help improving the forecasts of  $y_{2t}$  more than one period ahead if (32.22) holds (Lütkepohl, 1993). Intuitively, this happens because there may be indirect causal links, e.g.  $y_{1t}$  may have an impact on  $y_{3t}$  which in turn may affect  $y_{2t}$ . Thus, the definition of noncausality corresponding to the restrictions in (32.22) is not in line with an intuitive notion of the term. For higher dimensional processes the definition based on (32.18) results in more complicated nonlinear restrictions for the VAR coefficients. Details are given in Dufour and Renault (1998).

## TESTING FOR GRANGER-CAUSALITY

Wald tests are standard tools for testing restrictions on the coefficients of VAR processes because the test statistics are easy to compute in this context. Unfortunately, they may have non-standard asymptotic properties if the VAR contains I(1) variables. In particular, Wald tests for Granger-causality are known to result in nonstandard limiting distributions depending on the cointegration properties of the system and possibly on nuisance parameters (see Toda and Phillips, 1993).

Dolado and Lütkepohl (1996) and Toda and Yamamoto (1995) point out a simple way to overcome the problems with these tests in the present context. As mentioned in Section 3.1, the non-standard asymptotic properties of the standard tests on the coefficients of cointegrated VAR processes are due to the singularity of the asymptotic distribution of the LS estimators. Hence, the idea is to get rid of the singularity by fitting a VAR process whose order exceeds the true order. It can be shown that this device leads to a nonsingular asymptotic distribution of the relevant coefficients, overcoming the problems associated with standard tests and their complicated non-standard limiting properties.

More generally, as mentioned in Section 3.1, Dolado and Lütkepohl (1996) show that whenever the elements in at least one of the complete coefficient matrices  $A_i$  are not restricted at all under the null hypothesis, the Wald statistic has its usual limiting  $\chi^2$ -distribution. Thus, if elements from all  $A_i$ ,  $i = 1, \dots, p$ , are involved in the restrictions as, for instance, in the noncausality restrictions in (32.20) or (32.22), simply adding an extra (redundant) lag in estimating the parameters of the process, ensures standard asymptotics for the Wald test. Of course, if the true DGP is a VAR( $p$ ) process, then a VAR( $p + 1$ ) with  $A_{p+1} = 0$  is also an appropriate model. The test is then performed on the  $A_i$ ,  $i = 1, \dots, p$ , only.

For this procedure to work it is not necessary to know the cointegration properties of the system. Thus, if there is uncertainty with respect to the integration properties of the variables an extra lag may simply be added and the test may be performed on the lag augmented model to be on the safe side. Unfortunately, the procedure is not fully efficient due to the redundant parameters. A generalization of these ideas to Wald tests for nonlinear restrictions representing, for instance, other causality definitions, is discussed by Lütkepohl and Burda (1997).

### 5.3 Impulse response analysis

Tracing out the effects of shocks in the variables of a given system may also be regarded as a type of causality analysis. If the process  $y_t$  is I(0), it has a Wold moving average (MA) representation

$$y_t = \Phi_0 u_t + \Phi_1 u_{t-1} + \Phi_2 u_{t-2} + \dots, \quad (32.23)$$

where  $\Phi_0 = I_K$  and the  $\Phi_s$  may be computed recursively as in (32.16). The coefficients of this representation may be interpreted as reflecting the responses to impulses hitting the system. The  $(i, j)$ th elements of the matrices  $\Phi_s$ , regarded as a function of  $s$ , trace out the expected response of  $y_{i,t+s}$  to a unit change in  $y_{jt}$  holding constant all past values of  $y_t$ . Since the change in  $y_{it}$  given  $\{y_{t-1}, y_{t-2}, \dots\}$  is measured by the innovation  $u_{it}$ , the elements of  $\Phi_s$  represent the impulse responses of the components of  $y_t$  with respect to the  $u_t$  innovations. In the presently considered I(0) case,  $\Phi_s \rightarrow 0$  as  $s \rightarrow \infty$ . Hence, the effect of an impulse is transitory as it vanishes over time. These impulse responses are sometimes called *forecast error impulse responses* because the  $u_t$  are the one-step ahead forecast errors.

Although the Wold representation does not exist for nonstationary cointegrated processes it is easy to see that the  $\Phi_s$  impulse response matrices can be computed

in the same way as in (32.16) (Lütkepohl, 1991, ch. 11; Lütkepohl and Reimers, 1992). In this case the  $\Phi_s$  may not converge to zero as  $s \rightarrow \infty$  and, consequently, some shocks may have permanent effects. Assuming that all variables are I(1), it is also reasonable to consider the Wold representation of the stationary process  $\Delta y_t$ ,

$$\Delta y_t = \Xi_0 u_t + \Xi_1 u_{t-1} + \Xi_2 u_{t-2} + \dots, \quad (32.24)$$

where  $\Xi_0 = I_K$  and  $\Xi_j = \Phi_j - \Phi_{j-1}$  ( $j = 1, 2, \dots$ ). Again, the coefficients of this representation may be interpreted as impulse responses. Because  $\Phi_s = \sum_{j=0}^s \Xi_j$ ,  $s = 1, 2, \dots$ , the  $\Phi_s$  may be regarded as accumulated impulse responses of the representation in first differences.

A critique that has been raised against forecast error impulse responses is that the underlying shocks are not likely to occur in isolation if the components of  $u_t$  are not instantaneously uncorrelated, that is, if  $\Sigma_u$  is not diagonal. Therefore, in many applications the innovations of the VAR are orthogonalized using a Cholesky decomposition of the covariance matrix  $\Sigma_u$ . Denoting by  $P$  a lower triangular matrix such that  $\Sigma_u = PP'$ , the orthogonalized shocks are given by  $\varepsilon_t = P^{-1}u_t$ . Hence, in the stationary case we get from (32.23),

$$y_t = \Psi_0 \varepsilon_t + \Psi_1 \varepsilon_{t-1} + \dots, \quad (32.25)$$

where  $\Psi_i = \Phi_i P$  ( $i = 0, 1, 2, \dots$ ). Here  $\Psi_0 = P$  is lower triangular so that an  $\varepsilon$  shock in the first variable may have an instantaneous effect on all the other variables as well, whereas a shock in the second variable cannot have an instantaneous impact on  $y_{1t}$  but only on the remaining variables and so on.

Since many matrices  $P$  exist which satisfy  $PP' = \Sigma_u$ , using this approach is to some extent arbitrary. Even if  $P$  is found by a lower triangular Choleski decomposition, choosing a different ordering of the variables in the vector  $y_t$  may produce different shocks. Hence, the effects of a shock may depend on the way the variables are arranged in the vector  $y_t$ . In view of this difficulty, Sims (1981) recommends trying various triangular orthogonalizations and checking the robustness of the results with respect to the ordering of the variables. He also recommends using a priori hypotheses about the structure if possible. The resulting models are known as *structural VARs*. They are of the general form (32.7). In addition, the residuals may be represented as  $v_t = R\varepsilon_t$ , where  $R$  is a fixed  $(K \times K)$  matrix and  $\varepsilon_t$  is a  $(K \times 1)$  vector of structural shocks with covariance matrix  $E(\varepsilon_t \varepsilon_t') = \Sigma_\varepsilon$ . Usually it is assumed that  $\Sigma_\varepsilon$  is a diagonal matrix so that the structural shocks are instantaneously uncorrelated. The relation to the reduced form residuals is given by  $\Gamma_0^* u_t = R\varepsilon_t$ .

In recent years, different types of identifying restrictions were considered (see, e.g. Watson (1994) and Lütkepohl and Breitung (1997) for discussions). The aforementioned triangular system is a special case of such a class of structural models with  $P = \Gamma_0^{*-1}R$ . Obviously, identifying restrictions are required to obtain a unique structural representation. In the early literature, linear restrictions on  $\Gamma_0^*$  or  $R$  were used to identify the system (e.g. Pagan, 1995). Later Blanchard and Quah

(1989), King, Plosser, Stock, and Watson (1991), Gali (1992) and others introduced nonlinear restrictions. To motivate the nonlinear constraints it is useful to consider the moving average representation (32.24) and write it in terms of the structural residuals:

$$\Delta y_t = \Theta_0 \varepsilon_t + \Theta_1 \varepsilon_{t-1} + \Theta_2 \varepsilon_{t-2} + \dots, \quad (32.26)$$

where  $\Theta_s = \Xi_s \Gamma_0^{*-1} R$  ( $s = 0, 1, \dots$ ). The long run impact of the structural shocks on  $y_t$  is given by  $\lim_{n \rightarrow \infty} \partial y_{t+n} / \partial \varepsilon'_t = \lim_{n \rightarrow \infty} \Phi_n \Gamma_0^{*-1} R = \sum_{s=0}^{\infty} \Theta_s \equiv \bar{\Theta}$ . If the shock  $\varepsilon_{jt}$  has a transitory effect on  $y_{it}$ , then the  $(i, j)$ th element of  $\bar{\Theta}$  is zero. Hence, the restriction that  $\varepsilon_{jt}$  does not affect  $y_{it}$  in the long run may be written as the nonlinear constraint

$$e_i' \bar{\Theta} e_j = e_i' (I_K + \Xi_1 + \Xi_2 + \dots) \Gamma_0^{*-1} R e_j = 0.$$

Here  $e_i$  ( $e_j$ ) is the  $i$ th ( $j$ th) column of the identity matrix. It can be shown that for a cointegrated system with cointegrating rank  $r$ , the matrix  $\bar{\Theta}$  has rank  $n - r$  so that there exist  $n - r$  shocks with permanent effects (e.g. Engle and Granger, 1987).

Imposing this kind of nonlinear restrictions in the estimation procedure requires that nonlinear procedures are used. For instance, generalized methods of moments (GMM) estimation may be applied (see Watson, 1994). If an estimator  $\hat{\alpha}$ , say, of the VAR coefficients summarized in the vector  $\alpha$  is available, estimators of the impulse responses may be obtained as  $\hat{\phi}_{ij,h} = \phi_{ij,h}(\hat{\alpha})$ . Assuming that  $\hat{\alpha}$  has a normal limiting distribution, the  $\hat{\phi}_{ij,h}$  are also asymptotically normally distributed. However, due to the nonlinearity of the functional relationship, the latter limiting distribution may be singular. Moreover, the asymptotic covariance matrix of  $\hat{\alpha}$  may also be singular if there are constraints on the coefficients or, as mentioned earlier, if there are I(1) variables. Therefore, standard asymptotic inference for the impulse response coefficients may fail.

In practice, bootstrap methods are often used to construct confidence intervals (CIs) for impulse responses because these methods occasionally lead to more reliable small sample inference than asymptotic theory (e.g. Kilian, 1998). Moreover, the analytical expressions of the asymptotic variances of the impulse response coefficients are rather complicated. Using the bootstrap for setting up CIs, the precise expressions of the variances are not needed and, hence, deriving the analytical expressions explicitly can be avoided. Unfortunately, the bootstrap does not necessarily overcome the problems due to the aforementioned singularity in the asymptotic distribution. In other words, in these cases bootstrap CIs may not have the desired coverage. For a critical discussion see Benkowitz, Lütkepohl, and Neumann (2000).

## 5.4 Forecast error variance decomposition

In practice forecast error variance decompositions are also popular tools for interpreting VAR models. Expressing the  $h$ -step forecast error from (32.16) in terms of

the orthogonalized impulse responses  $\varepsilon_t = (\varepsilon_{1t}, \dots, \varepsilon_{Kt})' = P^{-1}u_t$  from (32.25), where  $P$  is a lower triangular matrix such that  $PP' = \Sigma_u$ , gives

$$y_{T+h} - y_{T+h|T} = \Psi_0 \varepsilon_{T+h} + \Psi_1 \varepsilon_{T+h-1} + \dots + \Psi_{h-1} \varepsilon_{T+1}.$$

Denoting the  $(i, j)$ th element of  $\Psi_n$  by  $\psi_{ij,n}$ , the  $k$ th element of the forecast error vector becomes

$$y_{k,T+h} - y_{k,T+h|T} = \sum_{n=0}^{h-1} (\psi_{k1,n} \varepsilon_{1,T+h-n} + \dots + \psi_{kK,n} \varepsilon_{K,T+h-n}).$$

Using the fact that the  $\varepsilon_{kt}$  are contemporaneously and serially uncorrelated and have unit variances by construction, it follows that the corresponding forecast error variance is

$$\sigma_k^2(h) = \sum_{n=0}^{h-1} (\psi_{k1,n}^2 + \dots + \psi_{kK,n}^2) = \sum_{j=1}^K (\psi_{kj,0}^2 + \dots + \psi_{kj,h-1}^2).$$

The term  $(\psi_{kj,0}^2 + \dots + \psi_{kj,h-1}^2)$  is interpreted as the contribution of variable  $j$  to the  $h$ -step forecast error variance of variable  $k$ . This interpretation makes sense if the  $\varepsilon_{it}$  can be interpreted as shocks in variable  $i$ . Dividing the above terms by  $\sigma_k^2(h)$  gives the percentage contribution of variable  $j$  to the  $h$ -step forecast error variance of variable  $k$ ,

$$w_{kj}(h) = (\psi_{kj,0}^2 + \dots + \psi_{kj,h-1}^2) / \sigma_k^2(h).$$

These quantities, computed from estimated parameters, are often reported for various forecast horizons. Clearly, their interpretation as forecast error variance components may be criticized on the same grounds as orthogonalized impulse responses because they are based on the latter quantities.

## 5.5 Policy analysis

If there are exogenous variables in the system (32.7), the model may also be used directly for policy analysis. In other words, if a policy maker affects the values or properties of  $z_t$  the effect on the endogenous variables may be investigated within the conditional model (32.7). If the policy maker sets the values of  $z_t$  the effect of such an action can be analyzed by considering the resulting dynamic effects on the endogenous variables similar to an impulse response analysis. In general, if  $z_t$  represents stochastic variables, it is more natural to think of policy actions as changes in the distribution of  $z_t$ . For instance, a policy maker may shift the mean of  $z_t$ . Again, such changes can be analyzed in the context of our extended VAR models. For details see, for example, Hendry and Mizon (1998).

## 6 CONCLUSIONS AND EXTENSIONS

Since the publication of Sims' (1980) critique of classical econometric modeling, VAR processes have become standard tools for macroeconometric analyses. A brief introduction to these models, their estimation, specification, and analysis has been provided. Special attention has been given to cointegrated systems. Forecasting, causality, impulse response, and policy analysis are discussed as possible uses of VAR models. In some of the discussion exogenous variables and deterministic terms are explicitly allowed for and, hence, the model class is generalized slightly relative to standard pure VAR processes.

There are now different software packages that support VAR analyses. For example, PcFiml (see Doornik and Hendry, 1997) and EVIEWS may be used. Furthermore, packages programmed in GAUSS exist which simplify a VAR analysis (see, e.g. Haase *et al.*, 1992).

In practice, further model generalizations are often useful. For instance, to obtain a more parsimonious parameterization allowing for MA terms as well and, hence, considering the class of vector autoregressive moving average processes may be desirable (see Hannan and Deistler, 1988; Lütkepohl and Poskitt, 1996). Generalizations of the concept of cointegration may be found in Chapter 30 by Dolado, Gonzalo, and Marmol in this volume. Especially for financial time series modeling the conditional second moments is sometimes of primary interest. Multivariate ARCH type models that can be used for this purpose are, for instance, discussed by Engle and Kroner (1995). Generally, nonlinearities of unknown functional form may be treated nonparametrically, semiparametrically, or semi-nonparametrically. A large body of literature is currently developing on these issues.

### Note

\* I thank Jörg Breitung and Moses Salau for helpful comments on an earlier draft of this chapter and the Deutsche Forschungsgemeinschaft, SFB 373, as well as the European Commission under the Training and Mobility of Researchers Programme (contract No. ERBFMRXCT980213) for financial support. An extended version of this chapter with more explanations, examples and references is available from the internet at <http://sfb.wiwi.hu-berlin.de> in subdirectory bub/papers/sfb373.

### References

- Akaike, H. (1974). A new look at the statistical model identification. *IEEE Transactions on Automatic Control* AC-19, 716–23.
- Benkwitz, A., H. Lütkepohl, and M.H. Neumann (2000). Problems related to confidence intervals for impulse responses of autoregressive processes. *Econometric Reviews* 19, 69–103.
- Blanchard, O., and D. Quah (1989). The dynamic effects of aggregate demand and supply disturbances. *American Economic Review* 79, 655–73.
- Dolado, J.J., and H. Lütkepohl (1996). Making Wald tests work for cointegrated systems. *Econometric Reviews* 15, 369–86.

- Doornik, J.A., and D.F. Hendry (1997). *Modelling Dynamic Systems Using PcFiml 9.0 for Windows*. London: International Thomson Business Press.
- Dufour, J.-M., and E. Renault (1998). Short run and long run causality in time series: Theory, *Econometrica* 66, 1099–125.
- Engle, R.F., and C.W.J. Granger (1987). Cointegration and error correction: Representation, estimation and testing, *Econometrica* 55, 251–76.
- Engle, R.F., and K.F. Kroner (1995). Multivariate simultaneous generalized GARCH. *Econometric Theory* 11, 122–50.
- Gali, J. (1992). How well does the IS-LM model fit postwar U.S. data. *Quarterly Journal of Economics* 107, 709–38.
- Granger, C.W.J. (1969). Investigating causal relations by econometric models and cross-spectral methods. *Econometrica* 37, 424–38.
- Granger, C.W.J. (1981). Some properties of time series data and their use in econometric model specification. *Journal of Econometrics* 16, 121–30.
- Haase, K., H. Lütkepohl, H. Claessen, M. Moryson, and W. Schneider (1992). *MuLTi: A Menu-Driven GAUSS Program for Multiple Time Series Analysis*. Kiel, Germany: Universität Kiel.
- Hannan, E.J., and M. Deistler (1988). *The Statistical Theory of Linear Systems*. New York: John Wiley.
- Hannan, E.J., and B.G. Quinn (1979). The determination of the order of an autoregression. *Journal of the Royal Statistical Society B* 41, 190–5.
- Hendry, D.F., and G.E. Mizon (1998). Exogeneity, causality, and co-breaking in economic policy analysis of a small econometric model of money in the UK. *Empirical Economics* 23, 267–94.
- Hubrich, K., H. Lütkepohl, and P. Saikkonen (2001). A review of systems cointegration tests. *Econometric Reviews*, 20, 247–318.
- Johansen, S. (1995). *Likelihood Based Inference in Cointegrated Vector Autoregressive Models*. Oxford: Oxford University Press.
- Johansen, S., and K. Juselius (1990). Maximum likelihood estimation and inference on cointegration – with applications to the demand for money. *Oxford Bulletin of Economics and Statistics* 52, 169–210.
- Kilian, L. (1998). Small-sample confidence intervals for impulse response functions. *Review of Economics and Statistics* 80, 218–30.
- King, R.G., C.I. Plosser, J.H. Stock, and M.W. Watson (1991). Stochastic trends and economic fluctuations. *American Economic Review* 81, 819–40.
- Lütkepohl, H. (1991). *Introduction to Multiple Time Series Analysis*. Berlin: Springer-Verlag.
- Lütkepohl, H. (1993). Testing for causation between two variables in higher dimensional VAR models. In H. Schneeweiss and K.F. Zimmermann (eds.) *Studies in Applied Econometrics*, pp. 75–91. Heidelberg: Physica-Verlag.
- Lütkepohl, H., and J. Breitung (1997). Impulse response analysis of vector autoregressive processes. In C. Heij, H. Schumacher, B. Hanzon, and C. Praagman (eds.) *System Dynamics in Economic and Financial Models*. Chichester: John Wiley.
- Lütkepohl, H., and M.M. Burda (1997). Modified Wald tests under nonregular conditions. *Journal of Econometrics* 78, 315–32.
- Lütkepohl, H., and D.S. Poskitt (1996). Specification of echelon form VARMA models. *Journal of Business and Economic Statistics* 14, 69–79.
- Lütkepohl, H., and H.-E. Reimers (1992). Impulse response analysis of cointegrated systems. *Journal of Economic Dynamics and Control* 16, 53–78.
- Lütkepohl, H., and P. Saikkonen (2000). Testing for the cointegrating rank of a VAR process with a time trend. *Journal of Econometrics* 95, 177–98.

- Mosconi, R., and C. Giannini (1992). Non-causality in cointegrated systems: Representation, estimation and testing. *Oxford Bulletin of Economics and Statistics* 54, 399–417.
- Pagan, A. (1995). Three econometric methodologies: An update. In L. Oxley, D.A.R. George, C.J. Roberts, and S. Sayer (eds.) *Surveys in Econometrics* Oxford: Basil Blackwell.
- Paulsen, J. (1984). Order determination of multivariate autoregressive time series with unit roots. *Journal of Time Series Analysis* 5, 115–27.
- Perron, P., and J.Y. Campbell (1993). A note on Johansen's cointegration procedure when trends are present. *Empirical Economics* 18, 777–89.
- Phillips, P.C.B. (1987). Time series regression with a unit root. *Econometrica* 55, 277–301.
- Phillips, P.C.B. (1991). Optimal inference in cointegrated systems. *Econometrica* 59, 283–306.
- Saikkonen, P., and H. Lütkepohl (1999). Local power of likelihood ratio tests for the cointegrating rank of a VAR process. *Econometric Theory* 15, 50–78.
- Saikkonen, P., and H. Lütkepohl (2000a). Trend adjustment prior to testing for the cointegrating rank of a VAR process. *Journal of Time Series Analysis*, 21, 435–56.
- Saikkonen, P., and H. Lütkepohl (2000b). Testing for the cointegrating rank of a VAR process with an intercept. *Econometric Theory* 16, 373–406.
- Saikkonen, P., and R. Luukkonen (1997). Testing cointegration in infinite order vector autoregressive processes. *Journal of Econometrics* 81, 93–126.
- Schwarz, G. (1978). Estimating the dimension of a model. *Annals of Statistics* 6, 461–4.
- Sims, C.A. (1980). Macroeconomics and reality. *Econometrica* 48, 1–48.
- Sims, C.A. (1981). An autoregressive index model for the U.S. 1948–1975. In J. Kmenta and J.B. Ramsey (eds.) *Large-Scale Macro-Econometric Models*. pp. 283–327. Amsterdam: North-Holland.
- Sims, C.A., J.H. Stock, and M.W. Watson (1990). Inference in linear time series models with some unit roots. *Econometrica* 58, 113–44.
- Toda, H.Y., and P.C.B. Phillips (1993). Vector autoregressions and causality. *Econometrica* 61, 1367–93.
- Toda, H.Y., and T. Yamamoto (1995). Statistical inference in vector autoregressions with possibly integrated processes. *Journal of Econometrics* 66, 225–50.
- Watson, M.W. (1994). Vector autoregressions and cointegration. In: R.F. Engle and D.L. McFadden (eds.) *Handbook of Econometrics*, Volume 4. New York: Elsevier.

# Index

- accelerated hazard model, 457, 461  
acceptance region, 496, 510  
adaptive estimation, 92  
additive regressions, 430  
aggregation, 121, 335, 410, 428, 492, 608, 663  
Akaike information criterion (AIC), 78, 79, 272, 566, 685  
almost sure convergence, 203, 204, 206  
AR(1) process, 33, 36–7, 63–8, 71, 74–7, 80, 195, 363, 379, 516, 560, 589, 592–8, 609–12, 618, 620, 636, 661, 669, 672–7  
ARIMA model, 66, 78, 590, 625, 627, 629, 656, 658, 676  
artificial neural networks, 572, 584  
artificial orthogonalization, 272–3  
artificial regression, 17–26, 28–9, 31–7, 184, 295, 324, 328, 379–80, 547  
asymptotic efficiency, 321, 397, 405  
asymptotic normality, 24, 224–6, 237, 315, 321, 323, 374, 376, 405, 432–3, 436, 439–41, 443, 551, 641  
asymptotic theory, 201, 227, 235, 240, 243, 251, 304, 376, 508, 548, 556, 677, 695  
asymptotic variance, 47, 138, 189, 190, 193, 233, 240, 241, 242, 246, 249, 252, 323, 376, 387, 397, 551, 695  
asymptotically unbiased, 202, 205, 206  
asymptotically, distribution free, 176  
attenuation, 164  
Augmented Dickey–Fuller test (ADF), 176, 611, 621, 624–5, 629–32, 639, 651  
autocorrelation, 111, 116, 310–12, 314, 316, 322–3, 326–9  
autocorrelation function, 64, 429, 462  
autocorrelation, residual, 641  
autocorrelation spatial, *see* spatial autocorrelation  
autocovariances, 314  
autoregressive conditional duration (ACD), 459–63  
autoregressive conditional heteroskedastic (ARCH), 83, 117, 193, 197, 496, 510, 515, 539, 563, 566, 581, 689, 697  
autoregressive moving average (ARMA), 64–9, 74, 76–8, 80–1, 219, 461, 571, 572, 581, 583, 589–90, 600–2, 604, 607, 625–6, 629, 636, 646, 655–8, 676  
Bayesian inference, 93, 95–8, 467, 470, 491–2, 521–4, 527, 529, 532–4, 536  
Bayesian Information Criteria (BIC), 78, 272, 566, 568–71, 574–6, 579  
BDS test, 197  
best linear unbiased (BLUE), 83, 112, 116, 258, 260, 605  
best linear unbiased predictor (BLUP), 260  
block diagonal, 105, 111, 189, 350  
Bootstrap, 120, 254, 307–9, 442, 517–19, 554  
Borel-measurable, 212, 217, 220  
Brownian motion, 616, 640, 656, 659, 662, 668, 669, 672–4  
business cycle, 162, 444, 570, 578, 582, 675

- canonical correlations, 644  
Cauchy distribution, 43  
causation, 698  
central limit theorem (CLT), 132, 192, 202, 221–7, 229, 237, 239, 250, 313, 315, 326, 328, 640  
central limit theorem, functional, 610, 659  
characteristic function, 208, 211, 213–14  
characteristic roots, 259–60; *see also* eigenvalue  
characteristic vector, 128, 259; *see also* eigenvector  
Chebyshev, 205  
chi-squared distribution, 133, 540  
Cholesky decomposition, 70, 392, 512–13, 694  
classical assumptions, 103, 124  
closed form, 81, 236, 238, 302, 338, 377, 398, 405  
coefficient of determination, 421, 660; *see also*  $R^2$   
coherency, 469  
cointegrated systems, 638, 649–52, 654, 681–2, 697–9  
cointegrating, 181, 578, 609, 636–9, 641–51, 654, 679–81, 683–9, 695, 698–9  
cointegrating rank, 681, 683–6, 688–9, 695, 698–9  
cointegrating residuals, 639, 641  
cointegrating vector, 578, 609, 636, 638, 642–51, 654  
cointegration, 161, 561, 563–4, 577–8, 583–4, 591, 607, 611, 632, 635–9, 642, 645–54, 676–81, 684–5, 688, 690, 692–3, 697–9  
cointegration regression, 561, 639, 651, 654  
collinearity, 103, 182, 253, 256–78, 389; *see also* multicollinearity  
combined regression, 295, 430, 435, 440, 442  
common factor, 65, 173, 606, 648  
common trends, 638, 646–7, 654  
comparative fit index (CFI), 177  
concave function, 139, 553  
concentrated likelihood, 69, 74, 75, 88, 130, 320–1, 390  
condition index, 261–3, 265  
conditional moment tests, 60, 92, 199, 243, 381, 443  
conditional probability, 245, 284  
conditional symmetry, 189  
conditioning, 94, 98, 192, 268, 274, 276–9, 281–2, 284, 345, 356, 420, 596, 599, 606–7, 609  
conditioning set, 282  
conditioning variables, 192, 281–2, 284, 420  
confidence interval, 85, 109, 111, 117, 120, 142, 233, 237, 240, 308, 354, 434, 495, 505, 515, 525, 534, 554, 695, 697, 698  
consistency of GMM, 246  
consistent adjusted least squares (CALS), 167, 179  
consistent and asymptotically normal, 192, 249, 356–7, 360, 362, 374, 405, 546, 682, 683  
consistent test, 200, 244  
consistent uniformly asymptotically normal (CUAN), 133, 246  
consumption function, 275, 307, 676  
contemporaneous correlation, 103, 107, 111–12, 249, 517  
continuity points, 207–8, 210, 216, 228, 614  
continuous distribution, 32, 339, 506  
continuous mapping theorem, 615–16, 659  
convergence in distribution, 45, 208–11, 213, 610, 614–15, 632, 659, 682  
convergence in probability, 203–7, 209–10, 213, 289  
convergence, almost sure, 203–4, 206  
convergence, weak, 208, 210–11, 615  
correlation coefficient, 138, 389, 587–8  
correlogram, 587–90  
count data, 92, 200, 331–2, 334–7, 341–4, 346–8, 362, 369, 460  
covariance stationarity, 322  
Cox-test, 299  
criterion function, 17–20, 22–3, 25–6, 28–33, 36–7  
critical region, 41–2, 44–5, 61, 291, 414, 496–7, 506–7, 509–10  
critical value, 75–6, 78, 187, 254, 266, 269, 293, 304, 354, 413, 442, 494, 497, 542–3, 548, 552, 559, 618, 630, 644–5, 651, 662, 671, 675, 677, 686  
cumulative distribution function (c.d.f.), 34, 207–9, 319, 335, 368, 549  
curse of dimensionality, 430, 433  
data, binary, 335–6, 342  
data, censored, 332

- data, count, 92, 200, 331–2, 334–7, 341–4, 346–8, 362, 369, 460  
 data envelopment analysis (DEA), 521, 525, 534–5  
 data generating process (DGP), 52, 145, 282–4, 286, 340–1, 356, 496, 586, 625–7, 658–61, 665, 667–8, 672–3, 679, 685–6, 689, 693  
 data, panel, 92, 100, 118, 121, 166, 171–2, 177–9, 198, 230, 310–11, 314–15, 317–18, 324, 328–9, 331–2, 342–6, 349–51, 356, 358–9, 362–6, 372–3, 380–1, 386, 438, 442–3, 454, 492, 518, 521, 528–30, 533–7, 554  
 degenerate distribution, 207–8, 221, 226  
 degenerate random variable, 641  
 degrees of freedom, 96, 133, 135–6, 139–40, 169–70, 176, 244, 290–1, 346, 349, 434, 529–30, 566, 645, 662, 685  
 determinant, 70, 265, 319, 328, 445, 577, 644, 680, 685  
 deterministic trend, 560, 590, 629, 634, 686, 688  
 diagnostic testing, 180–7, 197–8, 328, 347, 438  
 difference equation, 81, 592, 609  
 direct search, 321  
 discrete choice models, 92, 308, 318, 323, 332, 363, 380–2, 395, 466–7, 470, 492, 495  
 distribution, bivariate, 388, 594, 597  
 distribution, continuous, 32, 339, 506  
 distribution, gamma, 283, 344, 535  
 distribution, Gaussian, 141, 338, 474, 641  
 distribution, multinomial, 48, 369  
 distribution, normal, *see* normal distribution  
 distribution, uniform, *see* uniform distribution  
 distribution, Weibull, *see* Weibull distribution  
 dummy variable, 87, 112, 318, 368, 369, 501, 526, 529, 598, 651  
 dummy variable, seasonal, 80, 660, 681  
 duration data, 285, 335, 337, 444–5, 458, 459, 462  
 duration dependence, 380, 447–8, 452, 583  
 dynamic completeness, 193–6  
 dynamic model, 78, 107, 145, 318, 350, 353, 363, 365, 446, 517, 608, 635, 639  
 dynamic multinomial choice, 467–8, 488  
 dynamic optimization, 467, 476  
 dynamic panel data models, 345, 363–4  
 dynamic programming, 468, 478, 492–3  
 dynamic regression, 68, 77–8, 80, 108  
 efficiency, 38, 92–3, 100, 103, 117, 119, 141, 242, 253, 322, 344, 353, 378, 381, 393–4, 397, 401–3, 405, 407–9, 502, 504, 518, 520, 521, 524–9, 531, 533–7, 605, 612, 653, 682  
 eigenvalue, 135, 170, 259–62, 276, 688  
 eigenvector, 272  
 elasticity, 334, 424–5, 527  
 ellipsoid bounds, 168  
 EM algorithm, 119  
 empirical distribution, 280, 291, 302, 304, 435, 437, 439, 552–4  
 encompassing approach, 200, 280, 296, 299  
 encompassing test, 288, 298–9  
 endogeneity, 178, 316, 321  
 endogenous variable, 123–4, 127, 138–9, 141–3, 157, 173–4, 316, 393, 399, 402, 408, 498–9, 515, 681, 696  
 equality restrictions, 23, 543–7  
 equilibrium condition, 652  
 ergodicity, 219, 238, 591, 599  
 error components, *see* variance components, 76, 106, 116, 321, 324, 362–3, 365, 443  
 error correction model (ECM), 643–4, 648, 651–2  
 errors in variables, 160, 177, 179  
 estimation, two-stage, 108, 387–9, 391, 393–6, 401, 408  
 estimation, two-step, 242, 639  
 Euclidean space, 202, 615  
 Euler equations, 296  
 exact restrictions, 145, 269–70  
 excess zeros, 336, 339–40  
 exogeneity, 198, 372, 646, 652  
 exogeneity, strict, 195  
 experimental design, 299, 470  
 exponential distribution, 283, 448, 450, 460, 464, 522, 532, 535  
 exponential regression, 451, 452  
 exponential smoothing, 571, 590  
 exponentially weighted moving average, 571, 580  
 factor analysis, 159–60, 173, 175  
 factorization, 149, 259, 666

- feasible generalized least squares (FGLS), 103–5, 108–11, 114–15, 120, 129, 320, 322, 414
- finite mixture models, 339, 345
- first difference, 66, 172, 356, 358, 570, 573, 649, 658, 660, 681, 686, 694
- first-order condition, 22, 28, 29, 31, 34, 54, 74, 236, 239, 245, 273, 320, 333, 336, 341, 451, 600
- fixed effects, 318, 345, 358–9, 361, 364, 372–3, 528–9, 531, 534
- fixed regressors, 182, 502
- forecast, errors, 564, 566–7, 575, 580, 690, 693
- forecast, feasible, 564
- forecast, multivariate, 563, 576–9
- forecast, optimal, 572
- forecast, out of sample, 571, 574, 580
- forecast, univariate, 563, 571–2, 579, 583
- fractionally integrated process, 66, 192, 462, 561, 653
- Frisch–Waugh–Lovell theorem (FWL), 24–5, 27
- F*-test, 185, 256, 299–300, 499, 514, 545, 670, 672, 682, 684
- full information, 122, 124, 127, 129–30
- fully modified ordinary least squares, 641
- functional central limit theorem, 610, 659
- gamma distribution, 283, 344, 535
- GARCH, 193, 296, 308, 461, 698
- GAUSS, 269, 345, 377, 380, 540, 549, 697, 698
- Gaussian distribution, 141, 338, 474, 641
- Gauss–Newton regression (GNR), 19–23, 25–9, 31, 35–6
- generalized inverse, 298
- generalized least squares, feasible (FGLS), 103–5, 108–11, 114–15, 120, 129, 320, 322, 414
- generalized linear models (GLM), 275–6, 278, 342, 345–6
- generalized method of moments (GMM), 16, 28, 32, 36, 136, 176, 227, 230, 231–7, 241–50, 252, 254–5, 296, 309, 314–15, 322, 326, 329, 342, 355–6, 373, 378, 695
- generalized minimum distance, 394
- Geweke–Hajivassiliou–Keane (GHK), 377, 379, 380, 392, 393
- Gibbs–Sampler, 97, 114, 377–9, 470, 473, 476, 478, 482, 492, 521, 523, 526–7, 532, 536
- global identification, 144, 147, 159–60, 234, 238
- goodness of fit, 61, 178–9, 434, 436, 441
- gradient, 18–20, 25–6, 30, 33, 191, 275, 372, 394, 402, 467
- Granger causality, 638
- grouped data, 369
- hazard functions, 446, 449, 453, 457
- hazard rate, 450–1, 453, 463
- Hessian matrix, 130
- heterogeneity, 111, 198, 311, 315, 337, 339, 344, 345, 349, 353, 366, 373, 446, 450, 451–3, 463–4, 521, 586, 589–90, 593, 597–8
- heteroskedasticity, *see* homoskedasticity, 16, 29, 31–2, 35–7, 82–3, 90–3, 98–100, 104–6, 110, 112, 186–91, 193–200, 240, 253, 311, 313–14, 321–3, 326–7, 329, 341, 372, 387–8, 395, 401, 405, 429–30, 434, 439, 451, 464, 495, 510, 516, 539, 563, 582, 597, 633, 689
- heteroskedasticity autocorrelation consistent covariance (HACC), 240, 245
- heteroskedasticity multiplicative, 87, 99, 275
- Hilbert space, 609
- homokurtosis, 189, 190, 196–7
- homoskedasticity, *see* heteroskedasticity, 39, 55, 60, 90, 185–9, 191, 193–6, 336, 352, 595, 603, 604
- hypothesis test, 16, 23, 25, 31, 38, 52, 58, 62, 66, 78, 83, 85–6, 97, 116, 180, 199, 229, 231, 252, 254, 277, 280, 285, 287–90, 293–4, 301, 309, 369, 372, 377, 414, 421, 429, 443, 494, 496, 539, 556, 609
- I(1) process, 558–60, 637, 647, 650, 654, 658, 685
- I(d) process, 650
- identically distributed, 85, 123, 125, 145, 163, 201, 217, 360, 433, 591, 606
- identification, global, 144, 147, 159–60, 234, 238
- identification, local, 78, 147, 151, 155, 159, 232, 234–5, 253
- identification, partial, 155

- identification, problem, 64, 159, 166, 174, 314, 522  
 identifying restrictions, 166, 173, 236, 237, 241, 243, 683, 684, 694  
 implicit function theorem, 151–2, 155–6  
 impulse responses, 689, 693–7  
 independent and identically distributed (i.i.d.), 22, 123, 182, 197, 488  
 independent sequence, 195, 247  
 indicator function, 203, 304, 358–9, 374, 471, 522, 524, 573  
 indirect least squares, 128  
 individual effects, 256, 353, 356–8, 360, 363, 373, 528, 532, 536  
 inequality hypotheses, 538  
 influential observations, 265–6  
 information criteria, 78, 305, 565–6, 571, 577, 580  
 information matrix, 26–7, 35–7, 47, 50, 89, 93, 145, 147–9, 180, 276, 359, 369–72, 379–81, 390, 394, 398, 546–7  
 information matrix test (IM), 27, 36–7, 180, 371, 381  
 instrumental variable estimator (IV), 29, 122, 125–6, 128–9, 131–7, 139, 141–2, 168–70, 172, 179, 248–9, 254, 392, 394, 399, 401–2, 498–500, 513, 654  
 integrated, process, 198, 200, 634, 658, 673–4, 690, 699  
 invariance, 40, 48, 56, 106, 119, 229, 514, 516, 544, 547  
 Jacobian determinant, 325  
 Jacobian matrix, 28, 145, 150, 152–3, 155–6, 158  
 Johansen methodology, 645  
 joint probability, 52, 148, 280, 370  
 just-identified, 141, 235  
 Kalman filter, 563  
*k*-class estimators, 136–7, 140, 143  
 kernel estimator, 400–1, 431, 441  
 kernel function, 358, 396, 399  
 Kronecker product, 103, 351  
 Kullback–Leibler information criterion (KLIC), 280, 300, 301, 303  
 kurtosis, 500  
 lag, operator (L), 66, 461–2, 601, 610, 642, 658–9  
 lag, polynomial, 642  
 Lagrange multiplier (LM) test, 42, 54–5, 60–1, 86, 90–1, 108, 114, 117, 327, 395, 441, 517, 539, 543, 689  
 latent variable model, 162, 406, 409, 466  
 latent variables, 113, 119, 161–2, 174–5, 364, 395, 454, 472  
 law of iterated expectations, 181, 186, 218  
 law of large numbers (LLN), 216–21, 316, 613  
 law of large numbers, uniform, 228–9  
 leading economic indicators, 577–8, 583  
 least absolute deviations (LAD), 405  
 least squares, generalized, 83–6, 88, 92, 96, 98–9, 103, 126, 129, 175, 270, 275, 320, 341, 351, 393, 408, 605  
 least squares, indirect, 128  
 least squares, nonlinear, 16–23, 30–1, 70, 191, 192, 229, 273–5, 322, 343, 345, 396, 565, 571, 573, 640  
 least squares, ordinary, 16, 20, 53, 70, 86, 90, 103, 125–6, 163, 182, 216–17, 232, 394, 413, 457, 557–8, 571, 639  
 least squares, predictor, 260  
 least squares, residuals, 272  
 least squares, three-stage, *see* 3SLS  
 least squares, two-stage, 329  
 least squares, two-stage, nonlinear, 16–18, 20, 70, 191–2, 229, 273–5, 322, 343, 345, 396, 565, 571, 573, 640  
 least squares, weighted, *see* weighted least square  
 least variance ratio, 128  
 Lebesgue measure, 151, 153–4  
 likelihood equations, 451  
 likelihood function, 39, 53, 59, 64, 77–8, 80, 93, 107, 127–8, 269, 271, 320, 335, 341, 354–6, 369, 373–4, 376, 378, 390, 392, 393, 399, 473, 514, 523–4, 527, 535, 543, 596, 600  
 likelihood ratio test, 41, 58, 86, 110, 116, 323, 516, 654, 699  
 limited dependent variables, 375, 381, 395, 407–8, 492, 518  
 limited information instrumental variable (LIVE), 125–6, 130–1, 133  
 limited information maximum likelihood (LIML), 127–8, 132–3, 135–8, 140–3, 170  
 linear approximation, 296  
 linear dependence, 260, 262

- linear probability model, 342, 367–8  
 linear process, 219, 227, 229  
 linear restrictions, 108–10, 116, 118, 268,  
   511, 513–14, 543, 694  
 LISREL, 175–6, 179  
 loading, 175, 638, 640, 643, 646, 648, 684  
 local alternatives, 42–4, 49, 181, 188,  
   299–300, 441  
 local identification, 78, 147, 151, 155, 159,  
   232, 234–5, 253  
 locally most powerful, 47, 61  
 location-scale model, 500  
 logistic distribution, 358, 368, 373, 377  
 logit model, 34, 192, 199, 285, 302–3,  
   356–8, 368, 370, 373, 375, 390, 420  
 logit model, ordered, 35, 37
- marginal likelihood estimation, 74  
 Markov Chain Monte Carlo (MCMC), 95,  
   97–8, 356, 466–7, 492  
 Markov models, 344, 467, 572  
 Markov switching models, 572  
 martingale, 218, 226–7, 240, 253, 591, 595,  
   603–4  
 martingale difference, 218, 226–7, 240, 253,  
   595, 603–4  
 matrix, commutation, 167  
 matrix, diagonal, 29–30, 90, 223–4, 259,  
   412, 431, 694  
 matrix, Hessian, 130  
 matrix, identity, 29, 103, 289, 351, 522, 642,  
   695  
 matrix, inverse, 137  
 matrix, nonsingular, 32, 158, 166, 168, 502  
 matrix, positive definite, 110, 141, 176, 235,  
   239, 250  
 matrix, positive semi-definite, 235  
 matrix, projection, 24, 29, 76, 237, 244  
 matrix, singular, 58, 105, 244, 245  
 matrix, symmetric, 259  
 maximum entropy, 644  
 maximum likelihood estimator (MLE),  
   41, 48, 50, 54, 89–90, 119, 218, 220, 227,  
   245–7, 275, 303, 333, 336, 354–6, 362,  
   380, 390, 394, 397, 406, 441, 451, 512,  
   514, 525, 577  
 maximum likelihood, full information  
   (FIML), 130, 134, 137, 636  
 maximum likelihood, limited information,  
   *see* LIML
- maximum likelihood, quasi, *see* quasi  
   maximum likelihood  
 maximum score, 358, 361–2, 364, 381  
 mean squared error (MSE), 139–40,  
   268–71, 379, 430, 432, 434–6, 562, 567,  
   690  
 mean value theorem, 148  
 measurable, 214, 218–19, 227–8  
 measurement error, 123, 159–60, 162–6,  
   168, 169–72, 175, 178–9, 362–4, 410–11,  
   415–18, 420, 522, 532, 534, 570  
 median, 42, 140, 180, 375, 524, 535  
 method of moments (MM), 109, 170, 178,  
   230–1, 236, 241, 245, 322, 325, 378, 392  
 method of scoring, 275  
 metric, 314, 364, 614  
 Metropolis algorithm, 94, 95  
 misspecification, 82, 92, 115, 180, 184,  
   187, 190–2, 198, 244, 246, 324, 340, 391,  
   394–5, 429, 470, 476, 565, 588, 595–7,  
   604, 635  
 mixed estimation, 270  
 mixture, 294, 337, 339, 346–7, 406, 540–1  
 mixture of normals, 641  
 model, dynamic, 78, 107, 145, 318, 350,  
   353, 363, 365, 446, 517, 608, 635, 639  
 model, functional, 164, 167, 170  
 model, selection criteria, 80, 305, 307, 352,  
   685  
 model, structural, 36, 163, 166–7, 178,  
   468–9, 493, 682, 684, 694  
 modified count models, 339  
 modulus, 665  
 moment equations, 157–8, 392, 402  
 Monte Carlo, 92, 99–100, 108, 110–11, 114,  
   118–19, 299, 306–7, 328–30, 353–4, 375,  
   379, 389, 392, 408, 470, 476, 488, 492–5,  
   497, 499, 501, 504, 505–6, 508–11,  
   516–17, 531, 536, 540, 542, 548, 549, 552,  
   554, 583, 618, 641, 644–5, 653, 662,  
   675–6  
 moving average process (MA), 37, 64–6,  
   79–80, 114, 219, 296, 308, 313, 571, 590,  
   598, 601, 633, 636, 655, 693, 695  
 moving average process, first-order, 64  
 multicointegration, 637  
 multicollinearity, 267, 278, 349, 388–9, 391  
 multinomial choice models, 309, 378  
 multinomial distribution, 48, 369  
 multinomial probit model, 379–80, 492

- multiple indicators-multiple causes (MIMIC), 173–5  
 multiple regression, 163–5, 264, 441  
 multiplicative heteroskedasticity, 87, 99, 275  
 multivariate forecasting, 563, 576–9  
 multivariate normal distribution, 319, 378  
 near seasonally integrated framework, 656, 674  
 nearly cointegrated systems, 650, 653  
 negative binomial distribution, 338  
 negative duration dependence, 446–7, 450, 452–3  
 nested, 37, 199, 284, 287–94, 297, 299–300, 307, 371, 390–1, 441, 491, 515, 517, 555, 606, 625  
 nondeterministic, 636  
 nonlinear dynamic models, 230  
 nonlinear forecasting, 572  
 nonlinear least squares (NLS), 16–23, 30–1, 70, 191–2, 229, 273–5, 322, 343, 345, 396, 565, 571, 573–4, 640  
 nonlinear regression, 17–19, 21–2, 29, 35, 119, 192, 196, 273–4, 285, 521  
 nonlinear restrictions, 56, 60, 545, 692–3, 695  
 nonnested hypothesis, 25, 33, 280, 287–9, 291, 293, 296, 300, 305, 307–9  
 nonparametric regression, 396–7, 402, 430–1, 433–5, 442–3  
 nonsample information, 267–8, 270, 276, 680, 684  
 norm, 202, 206, 215, 614  
 normal distribution, *see* Gaussian distribution, 32, 138, 140, 149–50, 175, 186, 223, 239, 241, 335, 368–9, 371, 387, 389, 391–2, 394, 408, 449, 465, 482, 503, 512–13, 524, 530, 548, 552, 572, 620  
 normal distribution, multivariate, 319, 378  
 normal distribution, singular, 222, 252  
 normal distribution, standard, 32, 293, 319, 464, 498, 552, 672, 683  
 normal equations, 125, 258  
 normality test, 500–1, 504, 508, 510, 517, 689  
 normalization, 35, 123–4, 128, 174, 223, 379, 400, 636, 638, 644, 648, 650, 651  
 normed fit index, 177  
 nuisance parameters, 49, 74, 80, 98, 184, 187, 330, 427, 495–7, 503–4, 508, 511, 514–17, 540, 618, 692  
 null model, 176–7, 184–5, 191, 194, 196, 198, 287, 304, 439  
 numerical optimization, 236, 276, 600  
 observational equivalence, 144  
 omitted variables, 83, 183, 185, 187, 372, 395, 411, 416–7  
 omnibus test, 180, 190, 197  
 one-sided alternative, 52, 539  
 one-step efficient, 16, 20, 23  
 optimal tests, 38, 40, 77  
 optimization estimator, 409  
 order condition for identification, 400  
 order statistic, 391, 404, 500, 549–51  
 orthogonality, 136, 163, 259, 354, 378, 380, 401, 451, 641, 669, 672, 674  
 orthogonal complement, 237, 647–8  
 outer product of the gradient (OPG), 25–8, 34, 36, 372  
 overdispersion, 332, 335–8, 341, 343, 346, 462  
 overidentification, 139–40, 394, 400, 408  
 overidentified, 131, 497  
 overidentifying restrictions, 133, 232–3, 237, 241, 243–5, 253, 469  
 panel data, 92, 100, 118, 121, 166, 171–2, 177–9, 198, 230, 310–11, 314–15, 317–18, 324, 328–9, 331–2, 342–6, 349–51, 356, 358–9, 362–6, 372–3, 380–1, 386, 438, 442–3, 454, 492, 518, 521, 528–30, 533–7, 554  
 parameter space, 17, 146, 155–6, 159, 185, 198, 219, 231, 233–4, 274, 291, 315, 321, 375–6, 468, 496–7, 509, 514, 546, 564, 594, 603–4, 683–4  
 parameters of interest, 93, 263, 281, 299  
 parametric models, 98, 142, 144–5, 159, 161, 197, 344, 383, 407, 409, 467, 562  
 partial correlations, 417  
 partial sum process, 659  
 partially parametric models, 341, 343  
 partitioned regression, 643  
 perfect multicollinearity, 68  
 perpendicular, 164, 169  
 Phillips curve, 577, 579–80  
 pivotal statistic, 495, 504–7, 510, 514  
 plim, 17, 25, 125–7, 132–4, 163–7, 169, 175–6, 203, 612, 617, 624, 627–32, 685  
 point optimal tests, 77

- Poisson distribution, 333, 460  
Poisson process, 337, 459–60, 514  
polynomial distributed lag, 268  
positive definite, 106, 110, 112, 123, 132–3,  
  138, 141, 163–4, 176, 235, 238–9, 250,  
  259, 314, 512, 633, 680  
positive duration dependence, 447  
positive semi-definite, 164, 199, 235, 240,  
  252, 254, 329  
posterior probability, 93, 95, 306, 352–3  
power function, 38, 42, 44, 79, 542, 546, 674  
precision, 260, 266, 268–9, 276, 473, 524,  
  532, 534, 691  
predetermined, 254, 365, 445, 505  
prediction, 260, 288, 311, 349, 352, 421–2,  
  428, 434, 566, 582–3  
prediction, interval, 566  
predictive least squares, 577, 584  
pretest, 568–70, 574–6, 583  
principal components, 127, 271–2, 277  
prior distribution, 145, 295, 359, 473  
prior information, 60, 93, 154, 158, 167,  
  271, 420, 523–4, 533, 579  
probabilistic framework, 589, 591  
probability density function, 40, 136–7,  
  281, 297, 305, 426, 500  
probability, joint, 52, 148, 280, 370  
probability limit, 21, 25, 27, 126, 137, 146,  
  164–5, 168–9, 217, 282, 292, 624; *see also*  
  plim  
probability measure, 204, 208, 212, 614,  
  615  
probability, prior, 270  
probit model, 34, 117, 296, 318, 322–4,  
  326–7, 330, 358–9, 366, 368–9, 372–3,  
  375, 377–80, 390–1  
probit model, multinomial, 379–80, 492  
projection matrix, 24, 29, 76, 237, 244  
proportional hazard models, 456  
pseudo maximum likelihood, 376
- quadratic form, 24, 80, 141, 183, 187, 188,  
  190, 244, 322  
quadratic programming, 540–1  
qualitative response models, 366–7, 371–2,  
  381  
quantile, 464, 541, 549–51  
quasi maximum likelihood, 192, 282, 336,  
  338, 341, 346  
quasi-Newton, 19
- random coefficient models, 410–28, 352  
random effects, 106, 318, 345, 352–3,  
  363–4, 373, 379, 528, 531–4  
random number generator, 476  
random sample, 187, 399, 450  
random sampling, 61, 182, 383  
random variable, discrete, 335  
random walk, 66, 95, 528, 558–9, 561,  
  572, 609, 632–3, 636–7, 640, 646–7, 653,  
  655–60, 662–3, 667–8, 672–5  
rank condition, 159, 235, 400  
rational expectations, 254, 309, 467–8, 470,  
  476, 635, 647  
real numbers, 215, 225  
reduced form, 124–7, 130, 141–2, 159,  
  174, 316, 321, 393–4, 399–400, 497, 511,  
  681–2, 684, 686, 694  
reduced rank regression, 173–4, 177, 643,  
  683  
regime change, 539  
regression sum of squares, 128  
regression, exponential, 451–2  
regression, multiple, 163–5, 264, 441  
regression, reverse, 165, 178  
regression, spurious, 557–61, 590, 605, 609,  
  611, 632, 634, 675  
regular point, 147–9, 153–6, 158–9  
regularity conditions, 26, 39, 130, 134,  
  149, 169, 175, 181, 185, 194, 198, 218,  
  231, 238–9, 241–2, 244, 248, 250, 315,  
  321, 339, 362, 375, 390, 473, 498, 501,  
  523–4, 539, 546, 553, 566  
relative noncentrality index, 177  
reparameterization, 86, 358–9, 675, 684  
repeated measurements, 171  
residual sum of squares, 434, 617, 624  
restricted least squares (RLS), 268–9, 272,  
  276  
restrictions, equality, 23, 543, 546–7  
restrictions, identifying, 166, 173, 236–7,  
  241, 243, 683–4, 694  
restrictions, linear, 108–10, 116, 118, 268,  
  511, 513–14, 543, 694  
restrictions, overidentifying, 133, 232–3,  
  237, 241, 243–5, 253, 469  
reverse regression, 165, 178  
ridge regression, 277  
root, stationary AR(p), 68  
root-*n* consistent, 17, 20–1, 23–5, 343, 358,  
  362, 365, 375

- sample correlogram, 587  
 sample selection, 92, 99, 111, 335, 359–61, 364, 372, 381, 383–7, 389–405, 407–9, 492  
 sample selectivity, 350, 359–60, 372, 407  
 sample space, 39, 233, 267, 496  
 sampling theory, 83, 92–3, 96–7, 100  
 score function, 42, 44–5, 47–8, 56–7, 246, 249, 375, 398–9  
 score test, 36, 38–9, 42, 48–9, 54–6, 59–61, 100, 254, 327, 371, 379, 394–5, 441–2  
 score vector, 394, 398  
 scoring, method of, 275  
 search algorithm, 467  
 seasonal autoregressive, 656–61  
 seasonal integration, 656, 661, 667, 669, 671, 677  
 seasonality, 588, 591, 655, 660–1, 663, 676–7  
 seasonality adjustment, 563, 656  
 seasonality, unit root test, 655–6, 661, 672, 674, 677  
 seemingly unrelated regressions (SUR), 93, 101–2, 104–21, 314, 317–18, 321, 327, 365, 501, 517, 519, 555  
 selectivity bias, 361, 384, 387, 394, 407–9  
 self-selection, 383–7, 407, 409, 492  
 semiparametric, 92, 198–9, 253, 343, 358, 360, 362, 365, 367, 372, 375–6, 381–3, 388, 395–407, 409, 429–30, 438, 441–3, 446, 456, 554  
 semiparametric efficiency bound, 376, 402  
 semiparametric estimation, 367, 383, 388, 395–7, 400, 403–5  
 sequence, independent, 195, 247  
 sequential tests, 685  
 serial correlation, 25, 52, 59–60, 62–3, 65–6, 75–6, 79–80, 110, 117, 194–6, 199, 343–4, 346, 365, 373, 378–9, 395, 414, 425, 435, 441–2, 461–2, 563, 608, 639  
 serially uncorrelated, 28, 32, 193, 240, 419, 571, 577, 696  
 sgn function, 358  
 significance level, 505, 554, 631, 639  
 significance level, nominal, 547  
 significance test, 443  
 simulated likelihood, 378, 392, 408, 467  
 simulated maximum likelihood (SML), 338, 392–3, 466–7, 515  
 simulated moments, 378, 392, 408, 493  
 simulated samples, 495, 506, 509–10  
 simulated scores, 378  
 simulation approach, 280, 309  
 simultaneity, 140, 316, 323  
 simultaneous equations, 62, 122–4, 133–4, 142–3, 145, 157, 159–61, 170, 175, 275, 342, 398–400, 406, 408, 513, 517, 545, 555, 577, 607, 678  
 singular-value decomposition, 259  
 size of test, 42  
 skewness, 140, 394, 500  
 small sample properties, 118, 139, 143, 170–1, 179, 299, 372  
 smooth test, 38, 46, 56  
 smooth transition autoregression, 572–3  
 spatial autocorrelation, 111, 116, 310–12, 314, 316, 322–3, 326–9  
 spatial autoregressive, 312–16, 319–23, 329–30  
 spatial dependence, 54, 311, 315–19, 321, 326–8  
 spatial econometrics, 54, 59, 311, 314–15, 325, 327–8  
 spatial error model, 316, 320–2, 325  
 spatial heterogeneity, 310–11, 327  
 spatial lag operator, 312  
 spatial moving average (SMA), 312, 315  
 spatial multiplier, 313, 316, 319  
 spatial weights matrix, 312–13, 317, 319, 321, 325  
 specification analysis, 308, 362  
 specification test, 25, 36–7, 60, 187, 198–200, 306, 311, 327, 346, 366, 371–2, 375, 381, 395, 440, 441–2, 495, 516  
 spurious correlation, 587–8, 591  
 spurious regression, 557–61, 590, 605, 609, 611, 632, 634, 675  
 state-space models, 488  
 stationarity, 63–5, 219, 313–15, 461, 468, 559, 589–91, 595, 597–8, 600–2, 606, 611, 617–18, 624, 627–9, 631–6, 639, 649–50, 656, 664, 669, 672, 677  
 stationarity, covariance, 322  
 stationarity, trend, 611, 629–31, 633  
 statistical adequacy, 585, 607  
 statistical model, 17, 230, 492, 585, 588, 590, 591–5, 597, 602, 697  
 Stein-rule, 114, 269, 271  
 step function, 397, 615  
 stochastic dominance, 539, 548, 550, 552, 555–6

- stochastic frontier models, 443, 521, 522–3, 526, 534–7  
stochastic trend, 570, 634–8, 679–80  
stochastic volatility duration, 465  
strict stationarity, 194  
structural break, 564, 567, 583, 651  
structural equation, 123–6, 137, 143, 161, 174–6, 393–4, 399–402, 418, 497–8, 515  
structural form, 142, 159, 682–4  
Student's-*t* distribution, 597  
sufficient statistic, 149, 356, 359, 373  
sum of squares, 70, 394, 413, 434  
sum of squares, explained, 22, 35, 128  
sum of squares, residual, 434, 617, 624  
sum of squares, total, 25  
super-consistency, 641  
super-consistent, 640–1, 683  
superefficient, 397  
supremum, 397  
survivor functions, 452
- t* ratio, 255  
target model, 176  
temporal independence, 586  
test for ARCH, 197, 496  
test for cointegration, 651, 654  
test for functional form, 198, 430, 443  
three-stage least squares, *see* 3SLS  
time trend, 68, 343, 571, 574–7, 645, 698  
time-varying parameters models, 344  
Tobit model, 406–7, 409  
total sum of squares, 25  
trace statistic, 644  
trace test, 645  
triangular array, 223–5, 253, 315  
trimming function, 401–2, 405  
truncated distribution, 270  
truncated regression model, 92, 405  
two-stage least squares, *see* 2SLS  
two-step estimation, 242, 639
- unbiased, 44–5, 47, 57, 84, 108, 140, 146, 160, 182, 257–8, 316, 377, 433, 605, 690  
underidentified, 144, 175  
uniform distribution, 46, 204, 389  
uniformly most powerful, 38, 40, 44–5, 76  
union, 111, 298–9, 538, 540, 541, 548, 552, 554
- unit circle, 65, 461, 571, 582, 621, 658, 666, 675, 677  
unit roots, 181, 192, 591, 607, 610, 633–5, 637, 649–51, 654–8, 660, 663, 666, 669, 671, 675–7, 679–80, 699
- VAR, 194, 575, 577, 579–80, 583, 590, 602–4, 606–7, 643, 645–8, 653, 678–86, 689–90, 692–9  
VAR, bivariate case, 464, 691  
variable addition statistics, 184  
variance components, 521, 539, 696  
variance decomposition, 257, 262, 276, 689, 695  
variance, inflation factor, 264  
VARMA model, 698  
VECM, 638–9, 641, 643, 646, 648–51, 680–4, 686–8, 691  
vector autoregression, 194, 561, 564, 699; *see also* VAR  
vector error correction model, *see* VECM  
vector space, 262, 496  
vector space, Euclidean, 202, 206, 271, 615  
von Neuman, 75–6, 553
- wage equation, 470, 472, 474–5, 478, 482–3, 485, 491–2, 642  
Wald test, 48, 56, 58, 60, 498–500, 505, 513–14, 544, 547, 555, 692–3, 697–8  
weak instruments, 141, 143, 179, 255, 354, 519  
weak law of large numbers, 192, 250  
Weibull distribution, 285  
weighted least squares (WLS), 92, 99, 100, 367, 369  
weighting matrix, 28, 231, 233, 235–6, 241–3, 247–8, 342  
white noise, 65, 464, 557–8, 560–1, 592, 598, 636–7, 641, 680, 682–3, 689–90  
White test, 189  
Wiener process, 559, 616, 630  
within estimator, 530  
Wold decomposition, 599, 600
- 2SLS, *see* two-stage least squares, 126–9, 131–42, 169, 178, 248–9, 321, 323, 325, 327, 498–9  
3SLS, *see* three-stage least squares, 129–31, 133–4, 137, 321, 499